Lead Discovery in Cotton

Geo Velikkakam James

860115-864-010

8th March, 2010

Wageningen University
&
KeyGene N.V.

Master's Thesis in

Lead Discovery in Cotton

Wageningen University & KeyGene N. V.

Geo Velikkakam James

860115-864-010

8th March, 2010

Supervisors

Wageningen University:	KeyGene N.V.:
Prof. Dr. Richard Visser	Dr. Edwin van der Vossen

Table of Contents

Summary

Cotton (*Gossypium* genus) fiber is a unicellular seed trichome, a natural source of textile fiber. From the ancient time, constant selection has been made to improve this trait. QTL mapping showed diverged results due to the lack of repeated experiments and also due to the complexity of assigning linkage group in cotton. The goal of this project was to integrate all genetic and genomic data available in cotton to find candidate genes or genome locus for fiber trait. Integration of genetic maps and sequence information yields more inclusive picture of what is available in cotton, thus facilitates to understand the pattern of organization of genetic loci behind this trait. We have set up a relational database called Cotton Lead Discovery Database (CLDDB) by using the available public data, which enables cross talking of data. This information is illustrated in a local Gbrowser to aid effective visualization. A lead discovery scheme was established by considering available data sets. As a part of lead discovery scheme colinearity was established between QTL rich region of cotton chromosome 14 and *Arabidopsis thalina* genome, with the help of ColinerScan software, by comparing sequence, annotation and expression pattern of cotton fiber trait.

1. Introduction

Cotton (*Gossypium* genus) is the world's most important natural textile fiber source. Cotton fibers are used in innumerable commodities, ranging from textile fabrics, home furnishings, and medical supplies to automobile breaks. More than 150 countries are involved in cotton trading with an estimated economic impact of approximately $500 billion per year worldwide (National Cotton Council 2009, http://www.cotton.org/). Apart from this, cottonseed, a secondary product of cotton, is widely used for food oil, animal feeds and to produce industrial materials such as soap. Cottonseed oil is ranked fifth in the production and consumption volume among all the vegetable oils in past decades. Cottonseed oil is one of the few oils accepted for reducing saturated fat intake (National Cottonseed Products Association 2009, http://www.cottonseed.com/ publications/facts.asp). More than 80 countries around the globe produce cotton, out of which China, India and United States together provide two-third of the world's cotton production (United States Department of Agriculture 2009, http://ers.usda.gov/ briefing/cotton/).

The genus *Gossypium* occurs naturally throughout tropical and subtropical regions, and includes more than 50 species in which 4 species are independently domesticated. The *Gossypium* genus belongs to the Malvaceae family, a family of flowering plants. Out of 50 species, ~45 are diploid with chromosome number 2n = 26 and five are allotetraploid with 2n=4x=52. Within four domesticated diverged *Gossypium* species, *G. hirsutum* and *G. barbadense* are from America and *G. arboretum* and *G. herbaceum* are from Africa–Asia. The diploid *Gossypium* species are grouped into eight genome groups designated A through G and K, depending on the chromosome pairing affinities. Despite the fact that all diploid species have the same chromosome number (n = 13), their haploid genome size varies more than three fold, ranging from 885 Mb in D genome species to 2,572 Mb in K genome species, making this crop complex (Senchina et al. 2003) (Figure 1).

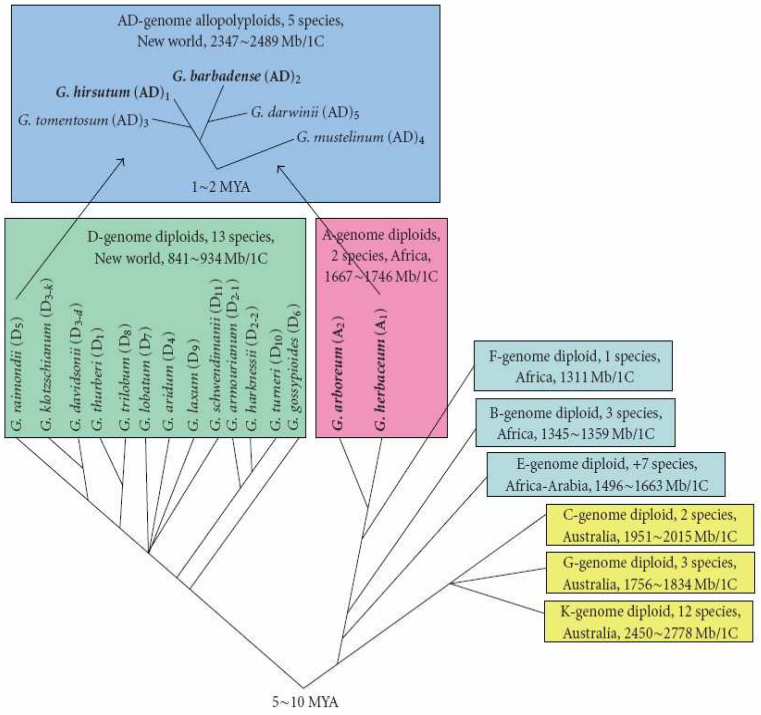

Figure 1: Phylogeny and evolution of *Gossypium* species (Zhang et al. 2008).

As shown in the Figure 1, there are four major lineages of diploid species. Each of the genome groups corresponds to a natural lineage, which is mostly a geographical region or continent. From a geographical point of view, genome C, G and K belong to Australia, genome D originates from America and lineage A, B and F which is close to lineage E belongs to Africa-Arabia (Jonathan F. Wendel et al. 2009)

Since four cultivated *Gossypium* species show morphological differences, especially in fiber and plant architecture, they are accepted in different regions for cultivation. *Gossypium hirsutum* is the first species in production with a contribution of over 90% of the world's cotton. *G. hirsutum* widely known as upland cotton or Mexican cotton, is well known for early maturation and high yields to produce long staple cotton. *G. barbadense*, also known as American Pima, Sea Island cotton or Egyptian cotton, contributes 8% of world's total cotton production to get the second position in production, and is bred for fibers with superior length, strength and fineness (Rong et al. 2005). One of the most attracting features of *G. barbadense* is its extra long staple cotton which is greatly accepted in the textile field. Both of these allotetraploid species have a genome constitution of AD, which is diverged from a common ancestor ~6-11 million

years ago (MYA) and have been reunited in a common nucleus ~1.1-1.9 MYA with a cytoplasm closely related to diploid A genome (Senchina et al. 2003). The remaining cultivated species having a diploid A genome are; *G. herbaceum*, also known as Levant Cotton and *G. arboretum*, also known as Tree Cotton, together produce 2% of the world's cotton.

Apart from the economic importance, cotton is taken as a model system for several important biological studies like genome size evolution, polyploidization and single-celled (trichome) biological processes. Cotton fibers not only serve as a natural fiber for industries but also as the most exaggerated plant cell type which is single celled, making them more interesting for the study of cell fate. However, studies to understand the genomics of cotton are much less as compared to the other major crops like rice, wheat and maize. Nevertheless, information is significantly increasing at present because of more funding in cotton research (Zhang et al. 2008).

1.1 Cotton genomic research

A large variety of molecular markers have been used to construct cotton genetic maps, starting from traditional markers like RFLP, RAPD and AFLP to the modern sequence based markers like SNPs. Without any surprise, the first cotton genetic map was constructed by using RFLP markers, having 705 loci in an interspecific cross between *G. hirsutum x G. barbadense* (Reinisch et al. 1994). Rong et al. (2004) extended this map to a larger extent to construct a map comprised of 2584 loci covering all 13 homologous chromosomes of allotetraploid cotton, which is widely used as a reference map today. In total, more than 25 genetic maps (Table 1) have been published based on intra and interspecific crosses, in which most of them were interspecific crosses. Because of their abundance, user-friendliness and co-dominant nature, Simple Sequence Repeats (SSR) or microsatellite markers got extra attention and are used more frequently in genetic map construction. Currently more than 5000 SSR markers are publicly available and this resource is sufficient to satisfy the requirement of cotton genome mapping and marker assisted selection (MAS) (Reddy et al. 2001). In order to nullify the drawbacks and also to overcome the scarcity of single marker types, there were efforts to incorporate different type of markers to construct dense genetic maps (Lacape et al. 2003; Mei et al. 2004).

Table 1: The list of cotton genetic maps used for this study.

map ID	MapName	Genome	Cross	IntegratedToMapID	Female	Male	Reference
1	Rong 2004	AD	Gh/Gb		Palmeri	K-101	(Rong et al. 2004)
2	Lacape 2005	AD	Gh/Gb		Guazuncho 2	VH8	(Lacape et al. 2005)
3	Lacape 2003	AD	Gh/Gb	2	Guazuncho 2	VH8	(Lacape et al. 2003)
4	Frelichowski 2006	AD	Gh/Gb		TM1	3-79	(Frelichowski et al. 2006)
5	Rong 2005	AD	Gh/Gb	1	Palmeri	K-101	(Rong et al. 2005)
6	Reinisch 1994	AD	Gh/Gb	1	Palmeri	K-101	(Reinisch et al. 1994)
7	Nguyen 2004	AD	Gh/Gb	2	Guazuncho 2	VH8	(Nguyen et al. 2004)
8	Park 2005	AD	Gh/Gb	4	Guazuncho2	VH8	(Park et al. 2005)
9	Lacape 2009	AD	Gh/Gb		Guazuncho2	VH8	(Lacape et al. 2009)
10	Saranga 2004	AD	Gh/Gb		Sic'on	F-177	(Saranga et al. 2004)
11	Paterson 2003	AD	Gh/Gb	10	Sic'on	F-177	(Paterson et al. 2003)
12	Saranga 2001	AD	Gh/Gb	10	Sic'on	F-177	(Saranga et al. 2001)
13	Yu 2007	AD	Gh/Gb		CCRI-36	Hai-7124	(Yu et al. 2007)
14	He 2007	AD	Gh/Gb		Handan-208	Pima-90	(He et al. 2007)
15	Lin 2005	AD	Gh/Gb	14	Handan-208	Pima-90	(Lin et al. 2005)
16	He 2005	AD	Gh/Gb	14	Handan-208	Pima-90	(He et al. 2005)
17	Zhang 2008	AD	Gh/Gb		Emian-22	3-79	(Zhang et al. 2008)
18	Guo 2007	AD	Gh/Gb		TM-1	Hai-7124	(Guo et al. 2007)
19	Draye 2005	AD	Gh/Gb		Tamcot-2111	Pima S-6	(Draye et al. 2005)
20	Chee 2005	AD	Gh/Gb	19	Tamcot-2111	Pima S-6	(Chee et al. 2005)
21	Han 2006	AD	Gh/Gb		TM1	Hai7124	(Han et al. 2006)
22	Wu 2009	AD	Gh/Gh		HS46	Marcabuc ag8US-1-88	(Wu et al. 2009)
23	Shen 2007	AD	Gh/Gh		7235	TM1	(Shen et al. 2007)
24	Zhang 2005	AD	Gh/Gh		Yumian	T586	(Zhang et al. 2005)
25	Guo 2008	AD	Gh/Gh// Gh/Gh		Simian3/Sumian12	Zhong4133/8891	(Guo et al. 2008)
26	Chee 2005b	AD	Gh/Gb	19	Tamcot-2111	Pima S-6	(Chee et al. 2005b)

Genetic maps not only serve to understand the evolution and organization of the cotton genome but also provide anchor points to locate genes or loci which influence quantitative and qualitative traits. Though biometrical genetics revealed the cumulative effect of genetic loci involved in quantitative traits, tagging those genetic loci by finding DNA markers closely linked to each trait, makes the crop improvement much faster and precise. Once the genomic region is identified, which has a role in the trait of interest, map based cloning is possible to harden the improvement process. In cotton, 32 cDNA libraries have been published from various tissues, mainly from different stages of fiber development, collected from various genotypes. These EST sequences from different libraries served as a good source to generate PCR based in-gene markers. Since these ESTs are developed from different species, this enables to compare the transcriptome among the species.

1.2 Fiber trait

Cotton fiber is the soft hairy fiber derived from the outer epidermis of ovules that grow in a form known as boll/ovary, around the seed. Fiber quality is one of the most important traits in cotton crop improvement. Importance of the cotton fiber is not only as a major source of fiber but also as an exceptional single-cell model to study fundamental processes in plant, like cell elongation and cellulose biosynthesis. The highly elongated structure and chemical composition makes cotton fibers suitable to study plant cell elongation and cell wall biogenesis (Kim & Triplett 2001). Unlike any other plant cell walls, cotton fiber cell walls don't contain lignin. Cotton fiber is composed of nearly pure cellulose and has a large central vacuole that becomes prominent quite early in the development.

The long fiber, which is spinnable into yarn and is required for the textile industry, is known as lint and the shorter fiber around the seed is known as fuzz. Lint fibers usually initiate on the day of anthesis, and fuzz fibers develop a few days later. In general, A genome species produce both lint and fuzz, where as D genome species produce very few lint fibers. Surprisingly the cultivated allotetraploid (AADD) cotton species produce more abundant and higher quality fibers than its diploid AA and DD genome progenitor, showing a strong selection on polyploid cotton for fiber quality (Lee et al. 2007).

The development process of fiber is near-synchronous in each developing ovule and among ovules within each ovary (boll). Each ovule contains approximately 21,000 lint fiber cells and a single ovary contains about one-half million synchronously elongating cells (Lee et al. 2007). This developmental process is divided into four discrete but overlapping stages; Initiation, Elongation/primary cell wall synthesis (PCW), secondary cell wall synthesis (SCW) and maturation leading to mature fibers (Kim & Triplett 2001) (Figure 2).

Initiation of fiber development is conveniently timed relative to the number of days post-anthesis (DPA). A negative DPA number indicates days before fiber cell emergence. Fiber cell initials usually come out on the day of anthesis and initials continue to grow rapidly without cell division for 16-25 days via a diffuse-growth mechanism that directs polarized growth of developing fiber cells. The process of fiber cell elongation is associated with strong cell turgor pressure and plasmodesmatal dynamics (Ruan et al. 2001). The elongated fiber cells may reach a length of nearly 6 cm, or one-third the height of an Arabidopsis plant. Although it is hard to differentiate fiber elongation stage from prior and subsequent stages, it is comparatively well studied and generally defined to be from 5 to 20 DPA. During the next stage (SCW), the β-1,4-glucan chain that forms the cellulose microfibrils of the secondary wall are synthesized and arranged helically around the growing fiber with periodic changes in the deposition angle. Successive layers of cellulose are deposited until the wall is 3 to 4 µm thick (Kim & Triplett 2001). Finally, fiber cells mature from 50 to 60 DPA when cotton bolls open and the long and mature (lint) fiber can be detached from the seeds. The maturation stage of fiber has been studied less because of the low protein and nucleotide availability.

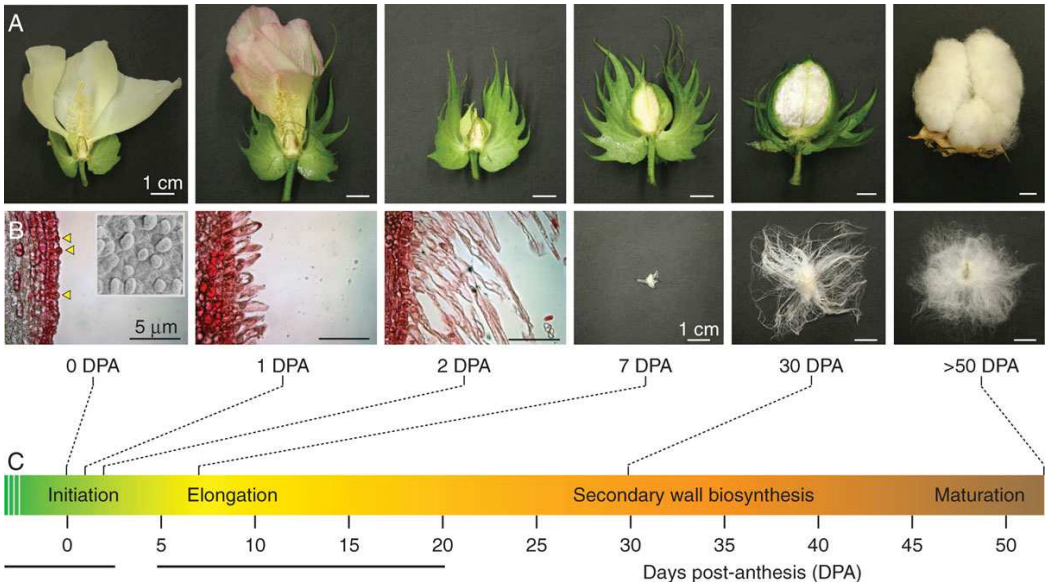

Figure 2: Cotton fiber developmental stages. (A) Cotton boll and fiber development. (B) and (C) Fiber development is shown over developmental stages (Lee et al. 2007)

The improvement of cotton fiber quality has become more important because of cotton demand and changes in spinning technology (Shen et al. 2007). Fiber quality is evaluated by different criteria, which mainly includes length components, strength, elongation, fiber uniformity, color components, maturity parameters and fineness (Lacape et al. 2005). There are different studies to tag economically important fiber related quantitative trait loci (QTL) in cotton by using the populations derived from inter/intra specific crosses mainly from the cultivated *G. hirsutum* and *G. barbadense.* This complex quantitative trait was driven by multiple loci interactions which tend to exhibit continuous variations in segregating populations and is also affected by environment effects. Moreover negative genetic correlation between fiber quality and lint yield has been a major problem for the improvement of cotton which is partially solved by backcrossing and intermating to generate desired recombinants (Shen et al. 2007). The lack of consistent QTL across studies makes this improvement process more complex. Different attempts have been tried to find consistent QTLs by carrying out meta-analysis and QTL mapping in different environments (Shen et al. 2007; Rong et al. 2007). The fact that only 10% of the total QTL set studied in meta-analysis showed the lack of consistency underlines the complexity of the trait.

With the help of functional genomics, few cotton fibers related genes have been identified. Mutations in two genes result in a complete loss of trichomes. The first of these is GLABRA1 (GL1), which encodes R2R3-type MYB related transcription factor and the second is TRANSPARENT TESTA GLABRA1 (TTG1). In Arabidopsis, apart from trichome formation, TTG1 is involved in mainly pathways and has pleiotropic effects including the control of synthesis of anthocyanin pigments, the production of seed coat mucilage and the development of root hairs (Humphries et al. 2005). From biochemical and physiological studies, ethylene biosynthesis is found to be significantly upregulated during the fiber growth period. Apart from this the cellulose and non-cellulose polysaccharide biosynthesis are also upregulated (Xu et al. 2007).

In this study, we investigated the public resources in cotton, containing genetic and genomic information related to fiber development and quality. We integrated this data to enable cross talking to find the candidate genes/genome loci for fiber traits. This is done by establishing colinearity in QTL rich regions of cotton with Arabidopsis genome and investigating gene expression patterns.

2. Materials and Methods

2.1 Cotton resources

The availability of public cotton genetic and genomics data is becoming more substantial. Efforts have been put in to increase the accessibility and integration of data generated by different groups. Development of new molecular markers, construction of linkage maps and QTL mapping, BAC library construction, physical mapping and cotton genome sequencing are a few examples of it.

The public Cotton Marker Database (CMD) is one of the initial attempts to integrate all publicly available cotton SSR and SNP marker data. This web-based database provides centralized access not only to the marker information but also regarding the different microsatellite projects and the data mining tool suite. For the comfortable view and comparison of maps, cMap tool has been adopted which enables a web based user friendly view of genetic maps. For cotton, there are mainly 3 web services which provides more or less complete collection of genetic maps available in cotton. Even though the base view of all cMaps are the same there are additional informations and integrations of different genetic maps which make each sources different (Table 2). In addition to cMap, TropGene has a web tool to query for both markers and QTLs in different genetic map. The IntegratedMap display from the public Cotton Diversity Database (CDD) is another attempt to bring all data under the same roof. The CDD has included OxfordGrid for the cotton genome in order to have a comparative overview of genetic maps. Despite the fact that there is no physical map available in the public databases, the Plant Genome Mapping Lab (PGML) site provides web interface to query hybridization data of probes in their six BAC libraries and also a ftp repository containing band files of the *G. raimondii* BAC library which is used during this study to build a draft physical map (Section 2.2.2).

Apart from NCBI, cotton genomic and transcriptomic data can also be accessed by other plant dedicated databases like plantGDB and the DFCI cotton gene index. So far in all the databases, ESTs from *G. hirsutum* and *G. barbadense* are used to make assemblies, bringing upto 25000 Unigenes. Even though EST data for the species *G. arboreum* (tree cotton) has a count of 41,768 ESTs, this hasn't been assembled to make

contigs/unigenes. The Program for Assembling and Viewing ESTs (PAVE) from the Arizona Genomic Institute is a program for viewing EST assemblies. This tool has an informative visualization of assemblies with base quality checks. The genome sequencing project of *G. raimondi* and *G. hirsutum* is on the way. This sequence information is expected in due time.

Table 2: A Quick reference table for cotton public resources.

Name*	**URL**
General	
JGI	www.jgi.doe.gov
PGML	www.plantgenome.uga.edu/
Wendal lab	www.eeob.iastate.edu/faculty/WendelJ/
PVP Accessions	www.ars-grin.gov/cgi-bin/npgs/html/pvp.pl?Cotton
Iowa state University	www.plantgenomesecrets.org/story-cotton
CEGC	http://cottonevolution.info/
NSF Cotton	www.cottongenomics.org/
Plant Database	http://plants.usda.gov/
CUGI	http://www.genome.clemson.edu/
Database	
CMD	www.cottonmarker.org
CDD	http://cotton.agtec.uga.edu
TropGene DB	http://tropgenedb.cirad.fr/en/cotton.html
CottonDB	www.cottondb.org
NCBI	www.ncbi.nlm.nih.gov/
PlantGDB	www.plantgdb.org/
Arizona GCL	www.agcol.arizona.edu
DFCI	http://compbio.dfci.harvard.edu/tgi/plant.html
PLEXdb	http://www.plexdb.org/index.php
Tool/Resources	
PAVE	www.agcol.arizona.edu/cgi-bin/pave/Cotton/index.cgi
PUT	www.plantgdb.org/prj/ESTCluster/progress.php
IntegratedMap	http://cotton.agtec.uga.edu/map/cottonmapdisp.aspx?lg=D06

Note* : JGI: Joint Genome Institute; PGML: The Plant Genome Mapping Laboratory; USDA : United State Department of Agricultural; CEGC: Comparative Evolutionary Genomics of Cotton; NSF: National Science Foundation; CUGI: Clemson University Genomic Institute; CMD: Cotton Marker Database; CDD: The Cotton Diversity Database; CottonDB: Cotton Genome Database; NCBI: National Center for Biotechnology Information; PlantGDB: Plant Genome Database; Arizona GCL: Arizona Genomic Computational Laboratory; DFCI: Computational Biology and Functional Genomic Laboratory; PLEXdb: Plant Expression Database; PAVE: Program for Assembling and Viewing ESTs; PUT: PlantGDB-assembled unique transcripts.

2.2 Data collection and integration

2.2.1 Integration of genetic maps

For crops like cotton, which doesn't have much sequence and genomic information available in public databases, the most common and effective way to start with is mining data from publications. Thorough investigation was carried out on different marker projects, genetic maps and literatures available in cotton. This allowed understanding the background information of different marker sets and their usage in genetic maps thereby avoiding redundancy. Genetic maps having fiber related markers were collected from different public sources (Table 2). As each public database has its own different criteria to integrate maps, which were not clear, only optimum sources were used for the different genetic maps for this study (Figure 3). IntegratedMap display for cotton was explicitly used for all the genetic maps and probe sequence information from Rong et al. (2004; 2005). Marker names and genetic map data format were standardized to a general format and original names were stored as alias. For example: marker MUSB810, MUSB0810a and MUSB0810 were standardized to MUSB0810. This standardized name was used throughout the project. These collections of genetic maps were stored in the local Cotton Lead Discovery Database (CLDDB, Section 2.3.3) along with the information of each map and its root information to enable back referencing in the future. Since genetic maps with a minimal number of loci and minimal coverage are not included in any of the public sources, QTLs and flanking marker data were extracted from the original literature source.

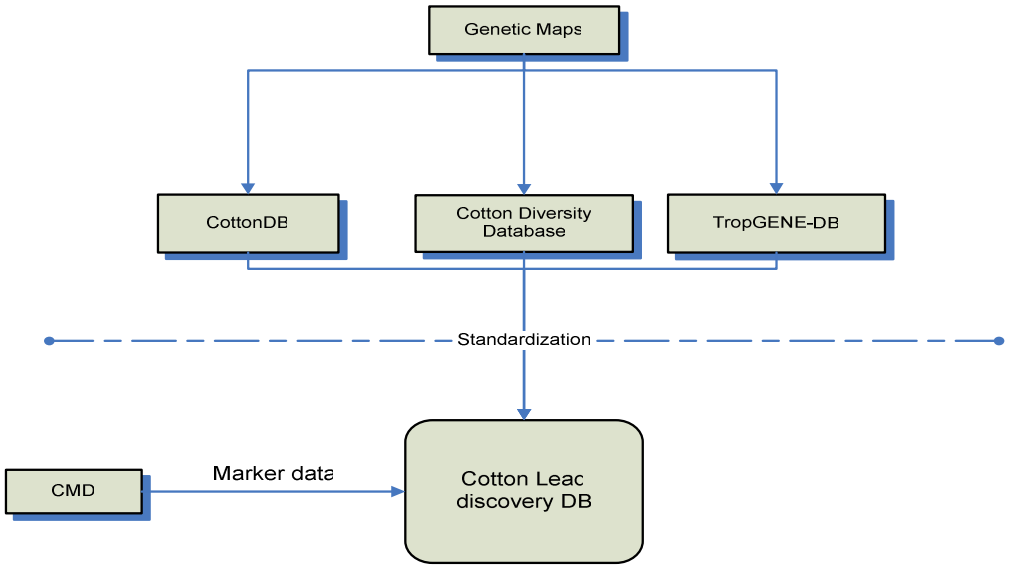

Figure 3: Schematic overview of the integration of cotton genetic maps and corresponding sequences.

2.2.2 Construction of draft physical map

Band file information of *G. raimondii* was downloaded from the PGML site and a preliminary version of a physical map was made by using the default settings of the FingerPrinted Contigs (FPC) software (http://www.agcol.arizona.edu/software/fpc/). Because of the broken internal links between IntegratedMap to BAC clone and also because of the lack of a bulk data retrieval system for probe hybridization data from the Cotton Genome Database, an ad hoc Perl script was written to automize downloading of all these probe hybridization data which were used in the genetic map of Rong et al. (2004; 2005). Raw HTML files from the script were parsed to extract tables and then converted into ACE format which were used by FPC to add the probe data to the physical map and to realign contigs.

2.2.3 Cotton Lead Discovery database (CLDDB)

The Cotton Lead Discovery database (CLDDB) (Appendix figure I) served as a data-warehouse containing most of the publicly available cotton genetic and genomic data. Apart from cotton data, limited Arabidopsis data has been included to enable comparative genomics. The main tables and there functions are as follows:

MapData: serves as a pool of genetic map collections. This is one of the main tables in this database.

MapInfo: Stores details about each maps stored in the database. This table consists of the type of parents used, dependency with other genetic maps and a reference to the original publication which is linked to the main literature database.

QTL: Information about QTL's discussed in literature which is not included in the genetic maps file directly.

EST_marker: contains information of NAU, HAU, MGHES markers with their Accession number and sequence information from CMD

Rong_2005: All marker and probe information of the Rong et al. (2004; 2005) genetic map.

Unigene: Unigene sequence information from NCBI was stored in this table by converting fasta format to tabular format. Additional information of NCBI's annotation was also added by parsing the raw information downloaded from NCBI for each cotton unigene.

EST_Unigene_converter: This table consists of information regarding EST composition in unigene from NCBI.

ColinearScan: This table contains the result from ColinearScan software. An ad hoc parser was written in AWK to parse the block file from ColinearScan to tabular format.

Arab_Gene_ref: This Table contains reference Gene models from Arabidopsis downloaded from TAIR (The Arabidopsis Information Resource 2009, www. arabidopsis.org).

EST_library: Contain information about all EST libraries and their expression patterns from NCBI.

There were quite a number of queries made which fulfilled the need of daily use of the database including queries which take input from the user in order to search specific regions of chromosomes of interest and give full information including marker positions, sequence and unigenes in that region. A local cotton Gbrowser was made from this database to facilitate an easy access and informative visualization of data. Blast hits to the Arabidopsis protein database were also included in the Gbrowser. Depending on the chromosome number of Arabidopsis, these blast hits were colored to understand the pattern of distribution in Arabidopsis chromosomes.

2.3 Lead discovery for fiber quality in cotton

QTL study: The distribution of fiber related QTLs across the cotton genome was studied. Data from the QTL table was organized based on chromosome and trait. Since there was no standard way of labeling QTL's, a new field with standard name for each trait was introduced into the table. By doing so, it became easier to compare QTLs from different literature sources. Consistency of QTL's for similar traits in nearby regions in the same linkage group from different maps were studied. This study helped to identify QTL rich chromosomes in general.

Lead discovery scheme: Since cotton doesn't have much sequence information; finding candidate genes for fiber related trait was mainly done through a comparative study with Arabidopsis. In order to establish link between cotton and Arabidopsis, a sequence similarity search was done with the help of BLAST and later their annotations were compared to strengthen the similarity. A keyword list of genes and protein families, which have a role in fiber development and quality; proved by experiments was made (Appendix table I). In later stages of our study, both keywords and expression data of EST's which belong to candidate genes, were used to reinforce the functional association of candidate genes with fiber trait (Figure 4).

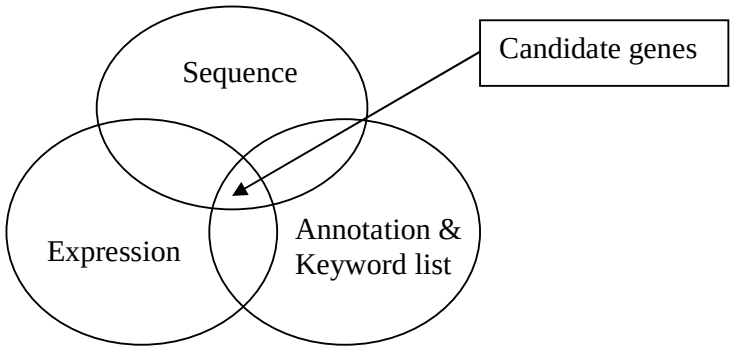

Figure 4: Lead discovery scheme.

ColinearScan (Wang et al. 2006) was used for finding colinear region between cotton and Arabidopsis. ColinearScan is a free software, distributed under GNU Public Licence v2 with a statistical model for colinearity prediction between or within chromosomes. Like all other sequence similarity programs, ColinearScan also uses the power of dynamic programming in order to detect the colinearity. This software works in Linux

environment, mainly consisting of 4 separate programs called get_pairs, repeat_mask, max_gap and block_scan.

With the help of Cotton Lead Discovery Database, QTL rich regions of chromosomes were identified. To identify candidate genes in those regions which have a putative role in fiber development and quality, a comparative study was done with Arabidopsis by finding colinear regions in Arabidopsis with the help of ColinearScan. The entire procedure can be divided into following steps

- Finding anchor points between cotton and Arabidopsis: sequence derived markers from a QTL rich region of cotton chromosome 14 (excluding SSR marker sequence) were blasted against the Arabidopsis protein database from TAIR as of Januvary, 2010. The default matrix BLOSUM 62 and a cut off value E= 0.00001 were used in all BLAST searches. High scoring hits per marker were used as anchor points.
- The list of anchor points from the blast result were used as input data for ColinearScan to find colinear blocks (Figure 5). Chromosome length file was made for both cotton and Arabidopsis to calculate max_gap value by ColinearScan and these values were used as threshold for the final analysis by ColinearScan to find colinear blocks (Appendix table II). A score of 75 was used as a threshold for get_pairs to the parse blast output. In repeat_mask, anchor pairs repeated more than 5 times were masked.
- The resulting block files from ColinearScan for each chromosome pairs were analyzed. Blocks having a higher number of anchor points within the least genomic distance were selected for manual editing.
- During manual editing, additional linking points (blast hits) which were screened out during the first step, but having E value less than 0.0001, were added into the suitable positions in the block.
- In every block, all Arabidopsis genes which are present within the block range were screened by using the fiber related keyword list. A Chi-square test was performed to check the significance of finding fiber related genes (defined by keywords). At the same time EST expression patterns of cotton anchoring unigenes were also examined to identify the correlation of expression with fiber trait (Figure 6). Blocks having higher number of anchor/linking points between

Cotton-Arabidopsis and which met the above mentioned criteria, were selected as putative colinear blocks containing fiber related genes.

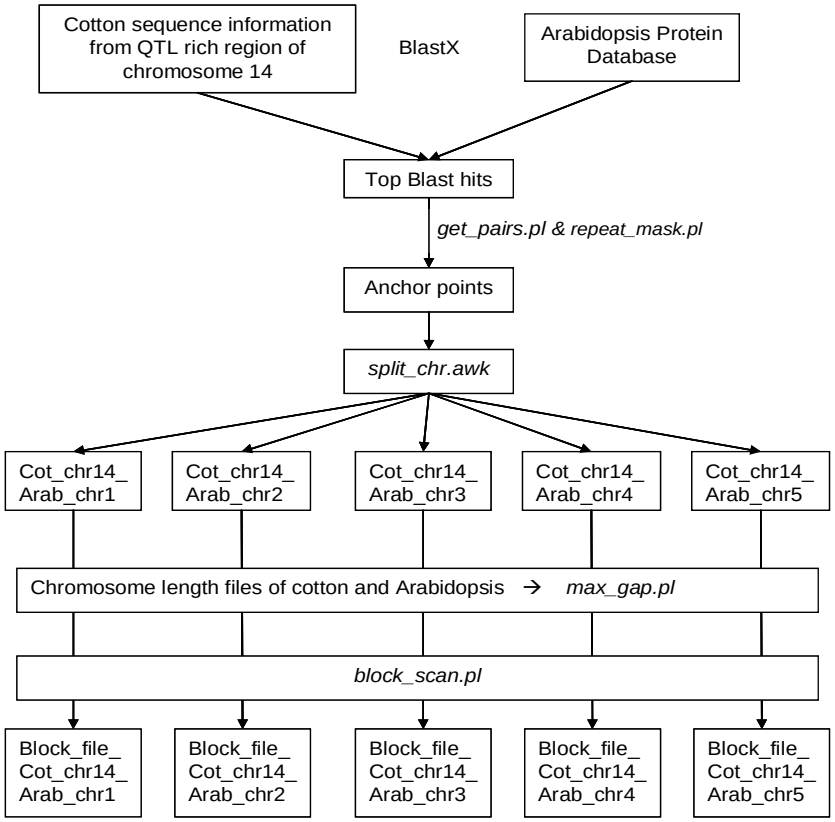

Figure 5: Work flow of ColinearScan software. get_pairs.pl, repeat_mask.pl, max_gap.pl and block_scan.pl are the perl script from ColinearScan package. Ad hoc AWK script was written (split_chr.awk) to fix ColinearScan package into the lead discovery scheme.

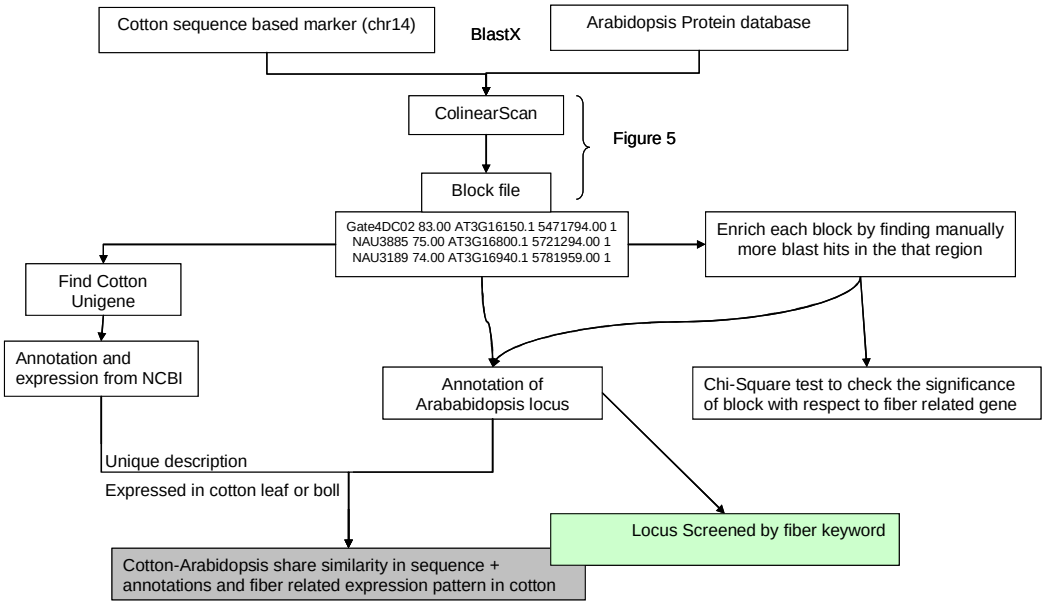

Figure 6: Complete workflow of lead discovery in cotton for fiber related trait.

3. Results and discussion

3.1 Lack of standard and consistent data in cotton

Finding leads in a crop like cotton with a huge genome size with multiple genome and lacking genomic information is a challenge. Various research groups from all over the world have generated genetic and genomics data and made specialized tools and databases for cotton. However, the low priority and deficiency in maintaining these data reduces the accessibility of it. One of the best examples for this is the migration process of the old cotton database (http://algodon.tamu.edu/cdb/index-old.php) to the new CottonDB (http://cottondb.org) which hasn't established many links. More than 25 maps have been published in cotton, by using different mapping populations (Rong et al. 2004; Frelichowski et al. 2006; He et al. 2007). Although these studies covered both A and D genome by using intra- and interspecific crosses, there are only a limited number of shared markers between different genetic maps published by different groups. Another major constrain in integrating genetic maps in cotton is the inconsistency of QTLs. Some part of this inconsistency is because of the difficulty to assign linkage groups in cotton. Even though Wang et al. (2006) solved this problem partially by assigning three common unknown linkage groups to corresponding chromosomes by FISH analysis, there are still considerable challenges in this area. The inconsistency of reported QTLs can be overcome by the replicated experiments in different environmental conditions. Attempts like IntegratedMap, to overcome these issues in cotton by developing integrated genetic maps, were quite successful. However, the lack of updating, maintenance and documentation reduces the usability of this tool. Differences between visualization and the original data, also questioned the integrity of the tool. Nevertheless tools like this will encourage the same strategies and give room to start improving the data.

Our attempt to collect all publicly available genetic maps and corresponding marker sequences with less redundancy, helped to get an overview of the information available on cotton. Standardized marker names allowed bringing genetic maps from different cMap data and sequence data from CMD together.

3.2 CLDDB: The data warehouse

Cotton Lead Discovery Database (CLDDB) is a relational database which served as an integrated platform of all public data for this study. The standardized nomenclature of loci across the tables made the accessibility of data easier and permits cross referring (Appendix figure I). The QTLs mined from literature completed the QTL collection which are not given in any cMap data and thus facilitate the QTL studies not only across the chromosomes but also within a chromosome to distinguish consistent QTLs. Apart from that, this QTL mining gave additional information about QTL effects, additive and dominant effects and flanking markers. It was found that the distribution of fiber related (including some of the yield related) QTLs among the genome, is not even. Although there is no significant difference between A and D genome in the amount of QTLs, the D genome has a slightly higher count than the A genome (275/222) with a difference of 53 (24%) (Figure 7). This observation from our studies agree with the most common concept that although the diploid D genome was derived from an ancestor which doesn't produce a spinnable fiber in tetraploid cotton, the D genome carries more QTLs which influence fiber quality (Rong 2007 , 692 ; Park et al. 2005).

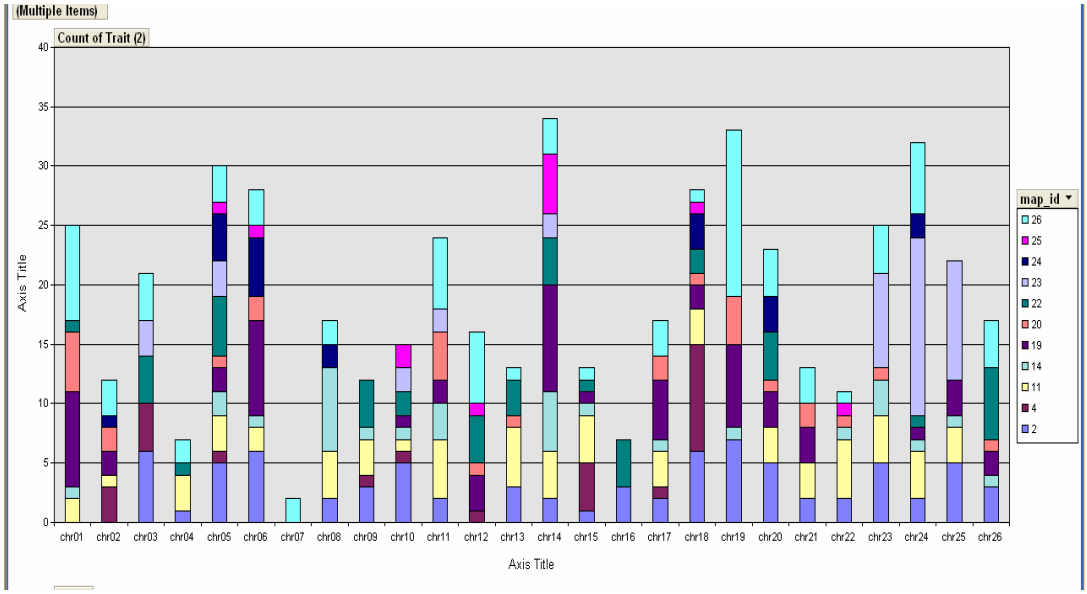

Figure 7: Fiber related QTL distribution across the cotton chromosomes. Each color indicates the different genetic maps. Refer appendix table III for the detailed count per chromosome per each trait.

Out of 26 chromosomes in cotton, chromosome 19, 14, 24 and 15 had a QTL count more than 30. Since cotton Chromosome 14 has a single larger QTL rich region, this region was taken as a starting point to establish the lead discovery scheme. The same lead discovery scheme can be followed for other QTL rich regions to find candidate genes.

3.3 Cotton Gbrowser and physical map

Apart from CLDDB, a Cotton Gbrowser which enables easier visualization of data and a draft version of the physical map, are the other two cotton tools/resources made during the study. This first cotton Gbrowser contained all marker information along with their best blast hit to the Arabidopsis protein database. Apart from that the QTL feature in the Gbrowser made it possible to distinguish QTL rich regions within chromosomes by simply visualizing a global view of chromosomes (Figure 8). One or more QTL clustered regions were found on most of the chromosomes. A closer investigation in these clustered regions might explain the molecular basis of corresponding traits. At present to do so, shortage of sequence is the major barrier which will be removed in the near future by the availability of genome sequences.

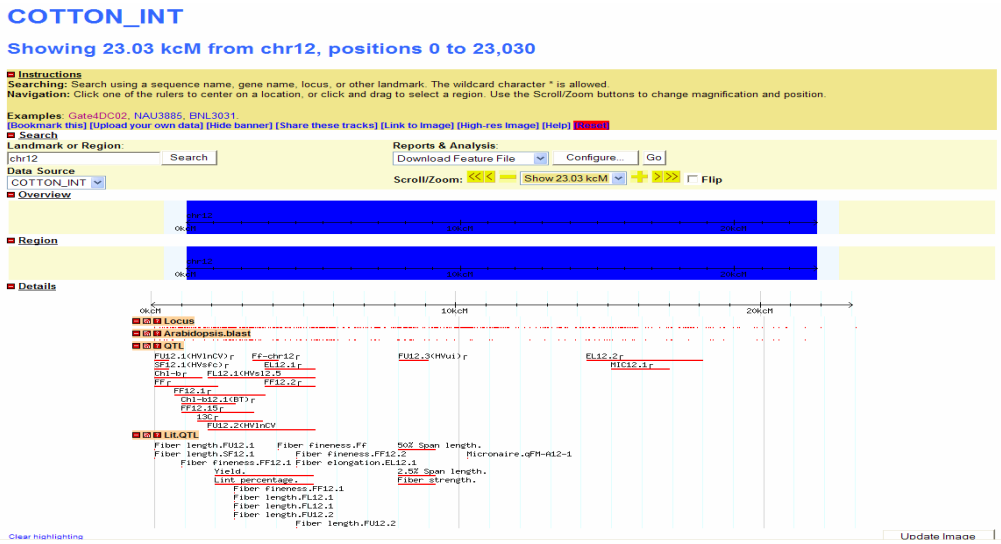

Figure 8: The Gbrowser view of cotton chromosome 12. The integrated genetic map is used as a ruler (cM).

The current draft version of *G. raimondii* physical map consists of 12540 contigs and 40105 singletons (Figure 9). Even though addition of hybridization data from 1645 marker probes were helped to glue a few contigs in the physical map, availability of BAC end sequences or more probe hybridization data will help to complete the physical map by assembling more contigs to a longer one. Addition of chromosome specific makers within this draft physical map will permit not only the assignment of chromosome numbers to the contigs but also facilitate the linking of the physical map to the genomic sequence, since they don't have multiple loci in the genome. By considering the genome size and the complexity of cotton genome, ambiguity can be foreseen while mapping back these short marker sequences into genomic sequences once it is available. Dedicated programs like BALT (Kent 2002) and Gmap (Wu & Watanabe 2005) would be handy to solve this puzzle with the help of EST sequences. Approaches like expression profiling without genome sequence information by next generation sequencing (Birzele et al. 2010) will be an alternative solution to improve cotton genome knowledge.

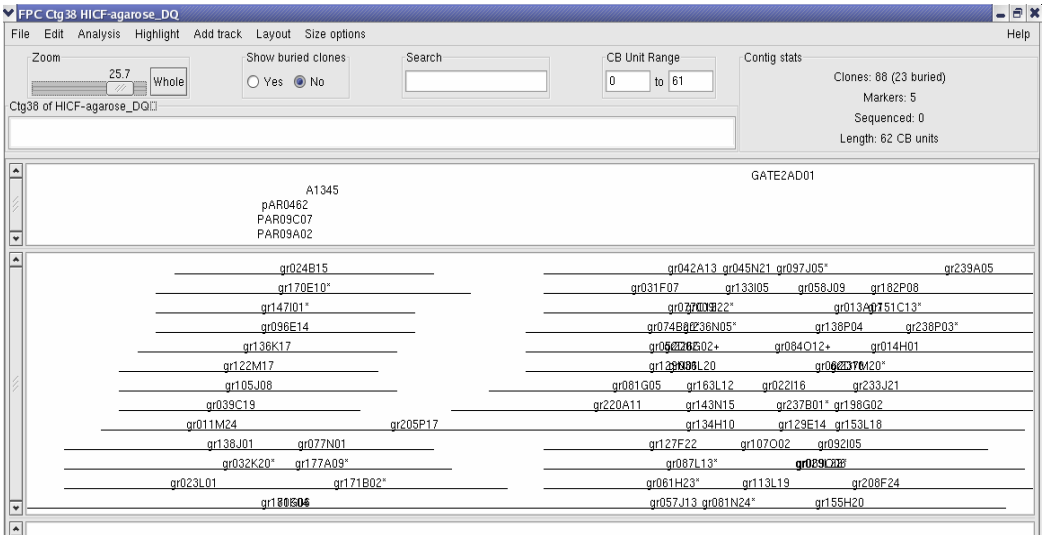

Figure 9: A screenshot of *G. raimondii* physical map.

3.4 Colinearity study between cotton and Arabidopsis

Colinearity between cotton chromosome 14 with all 5 chromosomes of Arabidopsis was established. ColinearScan was used to analyze the Blast results to find

colinear blocks. There are two main advantages of ColinearScan that makes it suitable for this study; the first one is the flexibility to handle both genetic and sequence maps simultaneously. Since cotton doesn't have any other maps than genetic maps, ColinearScan nullifies this disadvantage by using two different threshold maximum gap (max_gap) distance values. Secondly the threshold value (max_gap) used to define consecutive points in other software, was default. However, in ColinearScan, this threshold was calculated statistically on the basis of chromosome length. Since this threshold value is very important and has to change case by case, this approach is found to be more precise. ColinearScan can only handle one chromosome pair at a time. This drawback of ColinearScan was filled out by incorporating a custom made script called split_chr.awk, in the ColinearScan package, which splits the chromosome pairs from the blast results according to the chromosome number.

There are mainly two things which play a vital role in finding the colinear blocks with the help of ColinerScan; the input data to ColinearScan that is the blast hits and the threshold value (max_gap value). Adjusting these two parameters to find the optimum result is crucial. This is one of the reasons manual editing is so important to find out the missing links back to the block.

3.5 Lead discovery for fiber quality in cotton

A scheme for lead discovery in cotton for fiber trait was established by considering the available genetic and genomic data (Figure 4, 5 and 6). Even though lack of sequences in cotton had an effect on the lead discovery project, use of the Arabidopsis genome helped to cancel out this drawback upto a certain point. Instead of using cotton unigenes, sequence information of cotton genetic markers were used as a starting point for the colinearity search. There were two reasons why unigenes were not used as a starting point; first of all, not all cotton ESTs are used in unigene assembly. Only ESTs from *G. hirsutum* and *G. raimondii* are used in NCBI unigene construction. *G. arboreum* has considerable EST sequences from which some of the marker sets have been designed, but these numbers are just not good enough to fulfill NCBI standard to make unigenes. Thus, raw cotton EST sequences were downloaded for this study along with unigene and its composition file, to represent the complete set of available data. Secondly it is always easy to trace back the unigene, to which ESTs of interest belong to. So once the putative

block of colinearity was established, the unigenes which belong to that particular region could always be found in order to expand the sequence information of that site.

By following the established lead discovery scheme for cotton fiber traits, on cotton chromosome 14, colinear blocks which have a putative role in fiber development or quality were identified. One of the manually edited blocks has 14 linking points between cotton and Arabidopsis, covering 98 cM in cotton genetic maps (Appendix table IV). Since the pathways behind fiber development are still unclear and the genes which are highly expressed during fiber development are coding for phytohormones, energy/carbohydrate metabolism and secondary cell wall synthesis (Hovav et al. 2008; Lee et al. 2007) which represent major functional groups in plant development, it is hard to define the correlation between candidate genes and fiber. The same problem has been faced during the definition of the keyword list. The keyword list used for this study is vast enough not to miss any blast hits, which may have a correlation to fiber trait. But this has to be revised with more stringent criteria, once the knowledge about the mechanism behind fiber development is known.

Although we succeeded to find blocks of colinearity between cotton chromosome 14 and Arabidopsis, it was found not to be strong enough to adopt the sequence information from Arabidopsis to cotton. This is not surprising because since both plants are diverged evolutionarily by more than 100 MYA (Van de Peer et al. 2009), and the Arabidopsis genome size is seven times smaller than the smallest genome of cotton. In fact more sequence information of cotton may give a clear picture. Designing primers on these candidate regions of Arabidopsis and testing them in cotton will give a hint about the percentage of transferability of Arabidopsis sequence to cotton.

.

References

Birzele, F. et al., 2010. Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. *Nucl. Acids Res.*, gkq116.

Chee, P., Draye, X., Jiang, C., Decanini, L., Delmonte, T.A. et al., 2005. Molecular dissection of interspecific variation between Gossypium hirsutum and Gossypium barbadense (cotton) by a backcross-self approach: I. Fiber elongation. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 111(4), 757-763.

Chee, P., Draye, X., Jiang, C., Decanini, L., Delmonte, T.A. et al., 2005b. Molecular dissection of phenotypic variation between Gossypium hirsutum and Gossypium barbadense (cotton) by a backcross-self approach: III. Fiber length. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 111(4), 772-781.

Draye, X. et al., 2005. Molecular dissection of interspecific variation between Gossypium hirsutum and G. barbadense (cotton) by a backcross-self approach: II. Fiber fineness. *TAG Theoretical and Applied Genetics*, 111(4), 764-771.

Frelichowski, J.E. et al., 2006. Cotton genome mapping with new microsatellites from Acala 'Maxxa' BAC-ends. *Molecular Genetics and Genomics: MGG*, 275(5), 479-491.

Guo, W. et al., 2007. A Microsatellite-Based, Gene-Rich Linkage Map Reveals Genome Structure, Function and Evolution in Gossypium. *Genetics*, 176(1), 527-541.

Guo, W. et al., 2008. A preliminary analysis of genome structure and composition in Gossypium hirsutum. *BMC Genomics*, 9, 314.

Han, Z. et al., 2006. Characteristics, development and mapping of Gossypium hirsutum derived EST-SSRs in allotetraploid cotton. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 112(3), 430-439.

He, C. et al., 2005. Expression of an Arabidopsis vacuolar sodium/proton antiporter gene in cotton improves photosynthetic performance under salt conditions and increases fiber yield in the field. *Plant & Cell Physiology*, 46(11), 1848-1854.

He, D. et al., 2007. QTL mapping for economic traits based on a dense genetic map of cotton with PCR-based markers using the interspecific cross of Gossypium hirsutum × Gossypium barbadense. *Euphytica*, 153(1), 181-197.

Hovav, R. et al., 2008. The Evolution of Spinnable Cotton Fiber Entailed Prolonged Development and a Novel Metabolism. *PLoS Genet*, 4(2), e25.

Humphries, J.A. et al., 2005. Two WD-repeat genes from cotton are functional homologues of the Arabidopsis thaliana TRANSPARENT TESTA GLABRA1 (TTG1) gene. *Plant Molecular Biology*, 57(1), 67-81.

Kent, W.J., 2002. BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4), 656-664.

Kim, H.J. & Triplett, B.A., 2001. Cotton Fiber Growth in Planta and in Vitro. Models for Plant Cell Elongation and Cell Wall Biogenesis. *Plant Physiol.*, 127(4), 1361-1366.

Lacape, J.M. et al., 2003. A combined RFLP-SSR-AFLP map of tetraploid cotton based on a Gossypium hirsutum x Gossypium barbadense backcross population. *Genome / National Research Council Canada = Génome / Conseil National De Recherches Canada*, 46(4), 612-626.

Lacape, J.M. et al., 2009. A new interspecific, Gossypium hirsutum x G. barbadense, RIL population: towards a unified consensus linkage map of tetraploid cotton. *Theor Appl Genet*, 119(2), 281–292.

Lacape, J.M. et al., 2005. QTL Analysis of Cotton Fiber Quality Using Multiple Gossypium hirsutum x Gossypium barbadense Backcross Generations. *Crop Sci*, 45(1), 123-140.

Lee, J.J., Woodward, A.W. & Chen, Z.J., 2007. Gene expression changes and early events in cotton fibre development. *Annals of Botany*, 100(7), 1391-1401.

Lin, Z. et al., 2005. Linkage map construction and mapping QTL for cotton fibre quality using SRAP, SSR and RAPD. *Plant Breeding*, 124(2), 180-187.

Mei, M. et al., 2004. Genetic mapping and QTL analysis of fiber-related traits in cotton (Gossypium). *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 108(2), 280-291.

Nguyen, T.B. et al., 2004. Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 109(1), 167-175.

Park, Y.H. et al., 2005. Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. *Molecular Genetics and Genomics: MGG*, 274(4), 428-441.

Paterson, A.H. et al., 2003. QTL analysis of genotype x environment interactions affecting cotton fiber quality. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 106(3), 384-396.

Reddy, O.U.K. et al., 2001. New dinucleotide and trinucleotide microsatellite marker

resources for cotton genome research. *Journal of Cotton Science*, 5(2), 103-113.

Reinisch, A.J. et al., 1994. A detailed RFLP map of cotton, Gossypium hirsutum x Gossypium barbadense: chromosome organization and evolution in a disomic polyploid genome. *Genetics*, 138(3), 829-847.

Rong, J. et al., 2007. Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. *Genetics*, 176(4), 2577-2588.

Rong, J. et al., 2005. Genetic mapping and comparative analysis of seven mutants related to seed fiber development in cotton. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 111(6), 1137-1146.

Rong, J. et al., 2004. A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium). *Genetics*, 166(1), 389-417.

Ruan, Y., Llewellyn, D. & Furbank R., 2001. The control of single-celled cotton fiber elongation by developmentally reversible gating of plasmodesmata and coordinated expression of sucrose and k(þ) transporters and expansin. *The Plant Cell*, 13, 47–60.

Saranga, Y. et al., 2004. Genetic dissection of cotton physiological responses to arid conditions and their inter-relationships with productivity. *Plant, Cell & Environment*, 27(3), 263-277.

Saranga, Y. et al., 2001. Genomic Dissection of Genotype × Environment Interactions Conferring Adaptation of Cotton to Arid Conditions. *Genome Research*, 11(12), 1988-1995.

Saranga, Y. et al., 2004. Genetic dissection of cotton physiological responses to arid conditions and their inter-relationships with productivity. *Plant, Cell & Environment*, 27(3), 263-277.

Senchina, D.S. et al., 2003. Rate Variation Among Nuclear Genes and the Age of Polyploidy in Gossypium. *Mol Biol Evol*, 20(4), 633-643.

Shen, X. et al., 2007. Genetic mapping of quantitative trait loci for fiber quality and yield trait by RIL approach in Upland cotton. *Euphytica*, 155(3), 371-380.

Van de Peer, Y. et al., 2009. The flowering world: a tale of duplications. *Trends in Plant Science*, 14(12), 680-688

Wang, K. et al., 2006. Complete assignment of the chromosomes of Gossypium hirsutum L. by translocation and fluorescence in situ hybridization mapping. *TAG. Theoretical and Applied Genetics*,113(1), 73-80.

Wang, X. et al., 2006. Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. *BMC Bioinformatics*, 7, 447-447

Wendel, J.F. et al., 2009. Evolution and Natural History of the Cotton Genus. In *Genetics and Genomics of Cotton*. pp. 1-20. Available at: http://dx.doi.org/10.1007/978-0-387-70810-2_1 [Accessed January 10, 2010].

Wu, J. et al., 2009. Quantitative analysis and QTL mapping for agronomic and fiber traits in an RI population of upland cotton. *Euphytica*, 165(2), 231-245.

Wu, T.D. & Watanabe, C.K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9), 1859-1875.

Xu, Y., Li, H. & Zhu, Y., 2007. Molecular Biological and Biochemical Studies Reveal New Pathways Important for Cotton Fiber Development. *Journal of Integrative Plant Biology*, 49(1), 69-74.

Yu, J. et al., 2007. High-density Linkage Map of Cultivated Allotetraploid Cotton Based on SSR, TRAP, SRAP and AFLP Markers. *Journal of Integrative Plant Biology*, 49(5), 716-724.

Zhang, H. et al., 2008. Recent Advances in Cotton Genomics. *International Journal of Plant Genomics*, 2008.

Zhang, J., Turley, R.B. & Stewart, J.M., 2008. Comparative analysis of gene expression between CMS-D8 restored plants and normal non-restoring fertile plants in cotton by differential display. *Plant Cell Reports*, 27(3), 553-561.

Zhang, Z.S. et al., 2005. Construction of a genetic linkage map and QTL analysis of fiber-related traits in upland cotton (Gossypium hirsutum L.). *Euphytica*, 144(1), 91-99.

Appendix

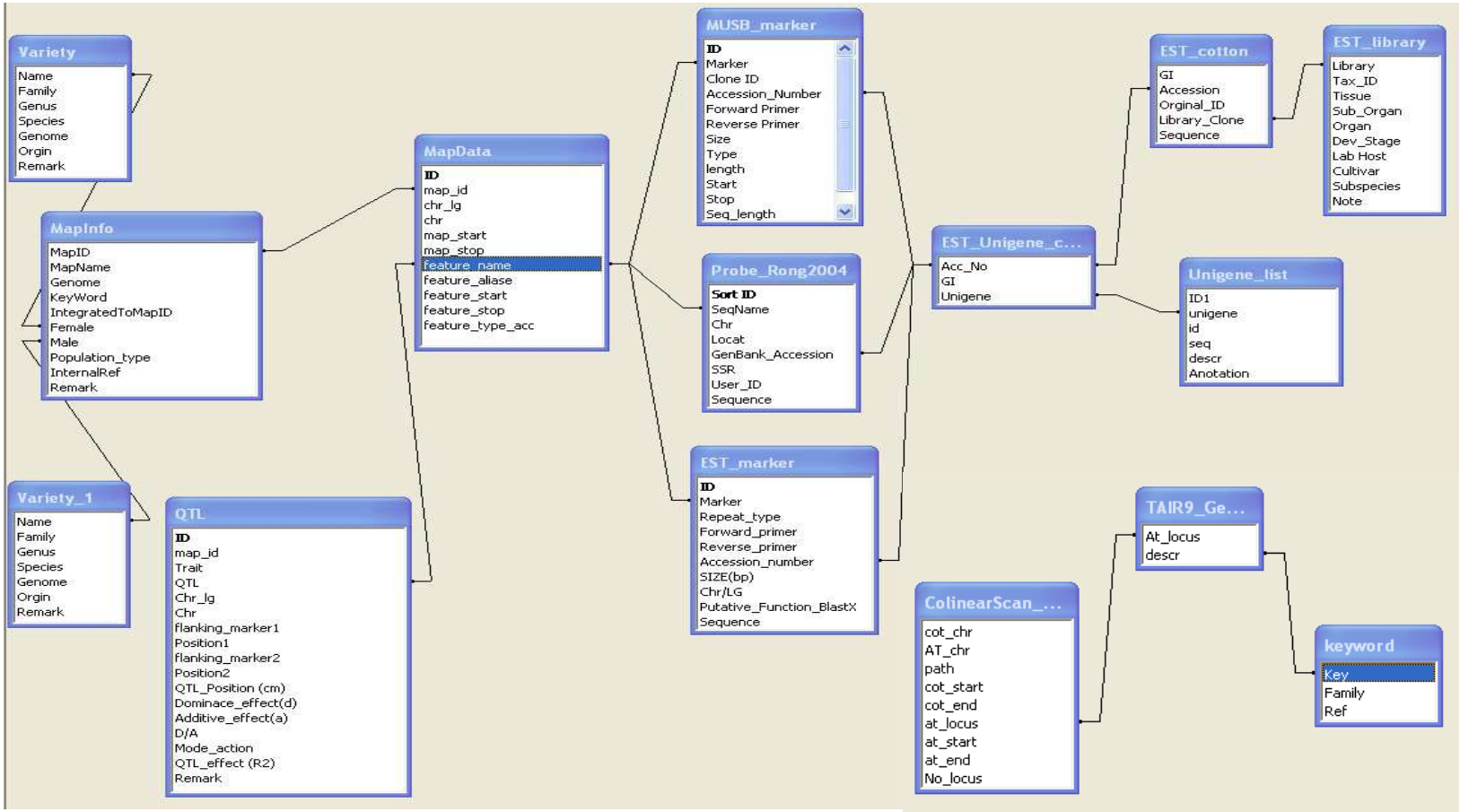

Figure I: Structure and relationship of Cotton Lead Discovery Database (CLDDB)

Table I: A list of keywords used to screen the blast results.

Keyword list		
bHLH	RD22	Acyltransferase
GLABR	cell wall	sucrose
R2R3	Fatty acid elongase	Vacuolar
myb	Lipid transfer	phosphoenolpyruvate
leucine zipper	glycine rich	PEPCase
HD-ZIP	Protodermal	xyloglucan
WD repeat	Ethylene	XTH
TTG	brassinosteroid	glucanases
NOECK	Tubulin	EGases
MIXTA	cytoskeleton	proline-rich
Auxin	actin binding	expansin
gibberellin	ACT1	fiber
BRASSINAZOLE	Cellulo	Kelch repeat
BZR	Cellulase	GTPase
Fatty acid elongase	serine carboxypeptidase	RAS protein
abscisic acid	MAP	calmodulin
ABA	Kinesin	calcium
cytokinin	G-protein	calreticulin
Glycuronosyltrasferase	ketoacyl-CoA	ER lumen

Table II: ColinearScan result from cotton chromosome 14 with Arabidopsis.

Cot_chr	Arab_chr	Path	Cot_marker	Cot_locus	At_locus	At_region	At_direction
chr14	chr1	1th	pGH699	104	AT1G03687.2	918308	1
chr14	chr1	1th	Gate3BE11	87	AT1G09290.1	3001682	-1
chr14	chr1	1th	NAU2312	76	AT1G11950.1	4034747	-1
chr14	chr1	1th	A1148	71	AT1G18270.1	6283634	-1
chr14	chr1	1th	NAU3485	70	AT1G21200.1	7421483	1
chr14	chr1	1th	NAU3913	55	AT1G25450.1	8938679	-1
chr14	chr1	1th	Gate1CB10	26	AT1G26580.1	9185620	1
chr14	chr1	2th	NAU3120	61	AT1G14650.1	5028077	1
chr14	chr1	2th	NAU3485	70	AT1G21200.1	7421483	1
chr14	chr1	2th	pGH551	73	AT1G26580.1	9185620	1
chr14	chr1	2th	Gate1CD07	82	AT1G27970.1	9746921	1
chr14	chr1	3th	NAU2336	52	AT1G68530.1	25712881	-1
chr14	chr1	3th	NAU3308	64	AT1G75080.1	28185709	1
chr14	chr1	3th	pGH699	104	AT1G76390.1	28655914	1
chr14	chr1	3th	pAR0815	107	AT1G76490.1	28695801	1
chr14	chr2	1th	NAU3903	110	AT2G38370.1	16072184	-1
chr14	chr2	1th	M16-161	105	AT2G39460.1	16475049	1
chr14	chr2	1th	MUCS105	74	AT2G43790.1	18138477	1
chr14	chr2	1th	Coau1J10	70	AT2G46960.1	19292295	-1
chr14	chr2	1th	NAU2929	66	AT2G46960.1	19292295	-1
chr14	chr3	1th	Gate4DC02	83	AT3G16150.1	5471794	1
chr14	chr3	1th	NAU3885	75	AT3G16800.1	5721294	1
chr14	chr3	1th	NAU3189	74	AT3G16940.1	5781959	1
chr14	chr3	1th	NAU5421	73	AT3G17460.1	5976949	-1

chr14	chr3	1th	pAR01-22	63	AT3G18710.1	6434234	-1
chr14	chr3	1th	G1012	34	AT3G22880.1	8097948	-1
chr14	chr3	1th	pAR0043	26	AT3G25040.1	9124479	1
chr14	chr3	1th	pAR0582	0	AT3G25920.1	9491268	-1
chr14	chr3	2th	NAU5467	90	AT3G06080.1	1835462	-1
chr14	chr3	2th	NAU3189	74	AT3G06590.1	2054647	-1
chr14	chr3	2th	NAU3225	38	AT3G07610.1	2426148	1
chr14	chr3	2th	NAU3209	20	AT3G07880.1	2514175	1
chr14	chr3	2th	pAR0582	0	AT3G16150.1	5471794	1
chr14	chr3	3th	NAU2987	61	AT3G02820.1	611573	1
chr14	chr3	3th	ESTS154	72	AT3G05530.1	1603540	1
chr14	chr3	3th	NAU3189	74	AT3G06590.1	2054647	-1
chr14	chr3	3th	NAU3885	75	AT3G06810.1	2146534	1
chr14	chr3	4th	Gate1BE06	58	AT3G16860.1	5759643	-1
chr14	chr3	4th	pAR01-22	63	AT3G18710.1	6434234	-1
chr14	chr3	4th	NAU5499	96	AT3G25040.1	9124479	1
chr14	chr3	4th	pAR0955	104	AT3G25920.1	9491268	-1

Table III: Summary of cotton fiber related QTLs. Assignment of unknown linkage group on the basis of Wang et al. (2006)

	Boll size	Boll weight	Bolls/plant	Color	Fiber elongation	Fiber fineness	Fiber length	Fiber strength	Fiber yellowness	Lint index	Lint percent	Lint yield	Maturity	Micronaire	Perimeter	Seed/boll	uniformity	Wall thickness	Weight fitness	Grand Total
Chr	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL	No. of QTL
chr01					5	8	9	2				1								25
chr02					2	3	6	1												12
chr03				1	1	4	6	2			4		1	1				1		21
chr04						2	2	1			1						1			7
chr05					3	8	7	4			1	3	1	2				1		30
chr06				2	3	11	6		1		1			2			2			28
chr07							2													2
chr08				1	1	1	4	4	1	1		2		1			1			17
chr09				1	1	2	3	1	1		1		1					1		12
chr10	2			1	2	3	2	1			1			1	1				1	15
chr11					7	2	7	2	1			2		1			2			24
chr12					1	4	8	1			1			1						16
chr13				1	1	2	4	2	1			1		1						13
chr14					2	8	7	2	1		3	2		5	1	1	2		1	35
chr15		1			3	3	1	3				1					1			13
chr16		1		1		1		2				1	1							7
chr17				1	3	3	3	1	1	1				4						17
chr18				1	3	3	9	6	1		1		1				2	1		28
chr19				2	5	7	14			1				3			1			33
chr20					4	5	8	2			1						2		1	23
chr21					2	4	4	2	1											13
chr22					3	1	1	2	1					1			2			11
chr23	2			1	4	1	8	4		1		1		3						25
chr24	4		2	1		4	7	7				2		4			1			32
chr25				1		5	4	4	1		4	1		2						22
chr26		1		1	2	1	7	1			1	2		1						17
Grand Total	8	3	2	16	58	96	139	57	11	4	20	19	5	33	2	1	17	4	3	498

Table IV: Colinear block after manual edition.

Cot_marker	Cot_position	Unigene	At_locus	At_region	At_direction	Colinear path	At_annotation
pAR0043	26		AT1G04270.1	1141852	-1	1st path	Symbols: RPS15 | RPS15 (CYTOSOLIC RIBOSOMAL PROTEIN S15); structural constituent of ribosome
Coau4N12	32		AT1G05560.1	1645674	-1	1st path	UGT1, UGT75B1 | UGT75B1 (UDP-GLUCOSYLTRANSFERASE 75B1); UDP-glucose:4-aminobenzoate acylglucosyltransferase
pAR0451	35		AT1G07400.1	2275148	1	1st path	17.8 kDa class I heat shock protein (HSP17.8-CI) | chr1:2275148-2275621 FORWARD
NAU3225	38		AT1G11950.1	4034747	-1	1st path	transcription factor | chr1:4034747-4038310 REVERSE
NAU3120	61		AT1G14650.1	5028077	1	1st path	SWAP (Suppressor-of-White-APricot)/surp domain-containing protein / ubiquitin family protein | chr1:5028077-5030520 FORWARD
A1148	56,68,71,104		AT1G18270.1	6283634	-1	2nd path	ketose-bisphosphate aldolase class-II family protein | chr1:6283634-6293772 REVERSE
NAU3308	64		AT1G19350.1	6688841	1	1st path	BES1, BZR2 | BES1 (BRI1-EMS-SUPPRESSOR 1); protein binding / transcription factor/ transcription regulator
NAU3913	55		AT1G19440.1	6729119	1	Blast	KCS4 | KCS4 (3-KETOACYL-COA SYNTHASE 4); acyltransferase/ catalytic/ transferase, transferring acyl groups other than amino-acyl groups

NAU3485	70		AT1G21200.1	7421483	1	2th	transcription factor | chr1:7421483-7422814 FORWARD
Gate4AD09	68		AT1G21400.1	7493492	1	1st path	2-oxoisovalerate dehydrogenase, putative / 3-methyl-2-oxobutanoate dehydrogenase
NAU2336	69	Ghi.9938	AT1G25450.1	8938679	-1	2nd path	KCS5, CER60 | KCS5 (3-KETOACYL-COA SYNTHASE 5); fatty acid elongase | chr1:8938679-8940282 REVERSE"
pGH551	73		AT1G26580.1	9185620	1	1st path	myb family transcription factor / ELM2 domain-containing protein (TAIR:AT2G03470.1)
Gate1CD07	82		AT1G27310.1	9484615	-1	1st path	NTF2A | NTF2A (NUCLEAR TRANSPORT FACTOR 2A); Ran GTPase binding / protein transporter | chr1:9484615-9485790 REVERSE
NAU5499	96	Ghi.4886	AT1G29330.1	10258580	-1	1st path	ATERD2, AERD2, ERD2 | ERD2 (ENDOPLASMIC RETICULUM RETENTION DEFECTIVE 2); KDEL sequence binding / receptor | chr1:10258580-10260906 REVERSE"
BNL3932	124		AT1G30450.1	10762905	1	1st path	CCC1, ATCCC1, HAP5 | CCC1 (CATION-CHLORIDE CO-TRANSPORTER 1); cation:chloride symporter/ sodium:potassium:chloride symporter | chr1:10762905-10769061 FORWARD"