Genetic Map Construction and Functional Characterization of Genes within the Segregation Distortion Regions (SDRs) in the F2:3 Populations Derived from Wild Cotton Species of the D Genome

Joy Nyangasi KIRUNGU
State Key Laboratory of Cotton Biology/institute of cotton Research Chinese Academy of Agricultural Sciences
Richard Odongo MAGWANGA
State Key Laboratory of Cotton Biology/ Institute of Cotton research Chinese Academy of Agricultural Sciences
Margaret Linyerera SHIRAKU
State key laboratory of Cotton Biology/ Institute of cotton research Chinese Academy of Agricultural Sciences
LU Pu
State key Laboratory of Cotton Biology/ Insstitute of Cotton Research Chinese Academy of Agricultural Sciences
Teame Gereziher MEHARI
State Key laboratory of Cotton Biology/ Institute of Cotton Research Chinese Academy of Agricultural Sciences
XU Yuanchao
State Key laboratory of cotton Biology/ Institute of Cotton research Chinese Academy of Gricultural Sciences
HOU Yuqing
State Key laboratory of Cotton Biology/ Institute of Cotton research Chinese Academy of Agricultural Sciences
Stephen Gaya AGONG
Jaramogi Oginga Odinga University of Science and Technology School of Biological and Physical Sciences
ZHOU Yun
School of Life Sciences Henan University
CAI Xiaoyan
State Key Laboratory of Cotton Biology/ Institute of Cotton Research Chinese Academy of Agricultural Sciences
Zhongli Zhou (zhonglizhou@163.com)
Chinese Academy of Agricultural Sciences Cotton Research Institute https://orcid.org/0000-0002-1900-5798
WANG Kunbo
State key Laboratory of Cotton Biology/ Institute of Cotton Research Chinese Academy of Agricultural Sciences
LIU Fang
State Key Laboratory of Cotton Biology/ Institute of Cotton Research Chinese Academy of Agricultural Sciences

Research

Keywords: Genetic Map, Segregation Distortion Region, Cis-regulatory elements, Genes, miRNA

DOI: https://doi.org/10.21203/rs.3.rs-30612/v3

Abstract

Background: Segregation distortion (SD) is a phenomenon common among stable or segregating populations, and the principle behind it still puzzles many researchers. The $F_{2:3}$ progenies developed from the wild cotton species of the D genomes were used to investigate the possible plant transcription factors within the segregation distortion regions (SDRs). A consensus map was developed between two maps from the four D genome, map A derived from $F_{2:3}$ progenies of *Gossypium klotzschianum* and *G. davidsonii* while Map B from *G. thurberi* and *G. trilobum* $F_{2:3}$ generations. In each map, 188 individual plants were used.

Results: The consensus linkage map had 1 492 markers across the 13 linkage groups; with a map size of 1467.445 cM and an average marker distance of 1.037 0 cM. Chromosome $D_5$02 had the highest percentage of SD with 58.621%, followed by Chromosome $D_5$07 with 47.887%. Six thousand and thirty- eight genes were mined within the SDRs on chromosome $D_5$02 and $D_5$07 of the consensus map. Within chromosome $D_5$02 and $D_5$07, 2,308 and 3 730 genes were mined, respectively, and were found to belong to 1 117 domains out of which 622 domains were common across the two chromosomes. Moreover, the first 9 domains were members of the plant resistance genes (R genes), while Pkinase; Protein kinase domain (PF00069) was the dominant group with 188 genes. Further analysis on the dominant domains revealed that 287 miRNAs were found to target various genes, such as the gr-miR398, gra-miR5207, miR164a, miR164b, miR164c among others, which have been found to target top-ranked stress-responsive transcription factors such as *NAC* genes. Moreover, some of the stress- responsive *cis*-regulatory elements were also detected. Furthermore, RNA profiling of the genes from the dominant family showed that higher numbers of genes were highly upregulated under salt and osmotic stress conditions, and also they were highly expressed at different stages of fiber development.

Conclusion: The results indicated the critical role of the SDRs in the evolution of significant genes in plants.

Background

Segregation distortion (SD) is described as a deviation from the expected Mendelian ratio within a segregating population due to various segregating distorters (Anhalt et al. 2008). Some of the factors that may lead to SDs include gametic and zygotic selections, non-homologous chromosome recombination, gene transfer, environmental agents, mapping population, marker types and genetic transmission (Mello et al. 1991). During the construction of genetic maps, it has been observed that some alleles in chromosomal regions skew from the normal Mendelian ratio. These alleles tend to cluster at segments of the chromosome, and these regions are referred to as the segregation distortion region (SDR) (Lu et al. 2002).

Research has shown that SD could bring errors in the marker order and map distances in the linkage map and thus reduce the accuracy of the maps(Yuan et al. 2019). However genes of significance have been mined within the SDR regions, for instance, the gene for crown rot resistance in wheat was identified within the SDR (Bovill et al. 2006), while the gene responsible for stem rust tolerance, was detected in the SDR on chromosome 2B in wheat (Tsilo et al. 2008). Moreover, SD has been observed in a variety of populations of organisms including insects (Sandler and Golic 1985), plants (Yuan et al. 2019), and mammals (Kumari et al. 1992).

Higher frequencies of occurrence of the SDR have been found in populations developed through interspecific as compared with intraspecific crosses (Dai et al. 2017), for example in rice more SDRs were detected in the double haploid compared to the $F_{2:3}$ populations developed from the same intraspecific cross (Xu et al. 1997; Wu et al. 2010), thirty-six SDRs were detected on 20 chromosomes in recombinant inbred lines in tetraploid cotton (Jamshed et al. 2016). Further evidence points out that the genes associated with zygotic and gametic selection could be responsible for SD (Manrique-Carpintero et al. 2016).

The use of molecular markers is preferred in the genotyping of populations because they are less influenced by phenotype and are significant in the study of SD (Zhang et al. 2013). The most used molecular marker in the analysis of SD is the simple sequence repeat (SSR); it has been widely used in the study of SD in the majority of plants and animals (Cheng et al. 2016; Wang et al. 2019). Several studies on SDs have been conducted in several plant species, including rice (Reflinur et al. 2014; Yang et al. 2014), maize (Lu et al. 2002; Wang et al. 2012), wheat (Kumar et al. 2007), barley (Liu et al. 2011), soybean (Liu et al.

2000), rapeseed (Yang et al. 2006), cotton (Wu et al. 2003; Amudha et al. 2012), and other plants. In the analysis of SD in the $F_{2:3}$ population of *Aegilops tauschii,* it was observed that some regions had skewed ratios towards particular alleles in the chromosomes (Fans et al. 1998).

The studies conducted in cotton showed that the majority of the SDs were mainly skewed towards the male parent rather than the female population, as was observed on chromosome 18 (Dai et al. 2017). However, in all the studies conducted to unravel the mystery of SDs in cotton, no experiment has been undertaken to explore the SDs in the $F_{2:3}$ population derived from the diploid wild cotton parental lines. The latest attempt to explore the SDs in the wild cotton progenitors involved a backcross population developed between *G. hirsutum* as the recurrent parent and *G. mustelinum* as the donor cultivar (Chandnani et al. 2017). And therefore, to explore the phenomena of the SDs in wild cotton progenitors, an interspecific population between *G. klotzschianum* and *G. davidsonii*, and between *G. thurberi and G. trilobum* were developed. The four parental lines were primarily selected because of their diverse genetic traits and broader ecological niches. The four parental lines used in the construction of the genetic maps are known to have traits for resistance to bacterial blight (*G. davidsonii*) (Zhang et al. 2016), sucking pests such as aphids (*G klotzschianum*) (Wei et al. 2017), *Fusarium wilt*, silver leaf whitefly and cotton bollworm resistance (*G. thurberi*) (Natwick 2006), *Verticillium wilt* (*G. trilobum*) (Dong et al. 2019). A total of 188 individuals were genotyped using SSR markers, primarily focusing on the exploitation of the genetic mechanism of the SD in severely distorted chromosome $D_5$02 and chromosome $D_5$07. The analysis of the SD from the genetic maps constructed from the diploid cotton of the D genome was conducted. The first map was then generated from two closely related parents, *G. klotzschianum* and *G. davidsonii* (Kirungu et al. 2018) and the second map developed from *G. thurberi and G. trilobum* (Li et al. 2018), in either of the maps, the $F_{2:3}$ population used, the genotypic data from the two maps were combined to generate the consensus map, and the consensus map was generated by using the two maps. The only available maps developed from the wild cotton species of the D genome. The focus was on chromosome $D_5$02 and chromosome $D_5$07 which showed severe distortions of markers from the two maps. Moreover, the marker segregation and genes within the SDRs were mined and analyzed. The genes mined within the SDR and understanding their roles will be significant in elucidating the role played by segregation distortion, and will help in improving the elite cultivated cotton germplasms with ever-shrinking genetic base and significantly lower adaptive mechanisms to various abiotic and biotic stress factors.

Materials And Methods

Parental materials

The two genetic maps were generated from an interspecific population obtained from the four parental lines. The first genetic map (Map A) was constructed from the $F_{2:3}$ population derived from the self-pollinating F_1 population of *G. klotzschianum* (female parent) and *G. davidsonii* (male parent). Similarly, the second genetic map (Map B) was constructed from $F_{2:3}$ populations derived from *G. thurberi* (female parent) and *G. trilobum* (male parent). A total of 188 progenies were used as the mapping population. The $F_{2:3}$ progenies from the four parental lines were developed and grown in the wild cotton nurseries, managed by the Institute of Cotton Research, Chinese Academy of Agricultural Sciences (ICR, CAAS), located in Sanya, Hainan province, China. The development of the $F_{2:3}$ progenies followed a similar pattern as described by Magwanga et al. (Magwanga et al. 2020) in the development of the backcross progenies between *G. tomentosum* (donor male parental line) and *G. hirsutum* (recurrent female parental line).

Molecular Markers Genotyping

Total DNA was extracted from the $F_{2:3}$ progenies and their parental lines using the CTAB method (Zhang et al. 2000b). Polymerase chain reaction (PCR) was conducted. The amplified PCR products were electrophoresed on non-denaturing 10% polyacrylamide gel electrophoresis in the 1×TBE buffer, and the gels were then visualized after silver staining (Huang et al. 2018). The primers used were the SWU markers which were developed by Southwest University in China, hence the acronym SWU. In the construction of the genetic map A, a total of 12 560 SWU markers were screened of which 1000 markers were found to be polymorphic. Out of the 1 000 polymorphic markers, 728 markers were mapped and generated the 13 linkage groups,

designated as chromosome $D_5$01 to $D_5$13. In the second genetic map, map B 12 560 SWU markers were screened, of which 996 markers were polymorphic, and only 849 polymorphic markers were mapped onto the 13 linkage groups. For the construction of consensus map, 1 492 polymorphic markers were applied to generate the genetic map, after removing the duplicated markers. The details of the markers and their sequences are shown in Supplementary Table S1

Linkage Map Construction and Determination of the Segregation Distortion of Molecular Markers

Markers with less than 5% missing data were used in the mapping of the linkage groups in the three maps (Coulton et al. 2020). The Joinmap 4.0 mapping tool was applied with a recombination frequency of 0.40, and a LOD score of 3.0, any LOD above 2.5 is known to be above the noise level (Faleiro et al. 2003). The Kosambi mapping function was used to convert the recombination frequencies to map distances. The linkage groups were then constructed using Mapchart 2.3 software (Voorrips 2002). The consensus map was constructed by merging the two individual data sets. Maps were drawn using MapChart 2.2 (Voorrips 2002)

Segregation Distortion Analysis

Segregation distortion (SD) within the mapping population was determined when the genotypic ratios deviated significantly from the expected Mendelian expectation (Reflinur et al. 2014). A Chi-square ($\chi 2$) test was performed for each marker to assess whether it significantly deviated from Mendelian segregation ratios. The markers showing segregation distortion were indicated by asterisks. The level of distortion was determined as follows *$P < 0.05$, **$P < 0.01$, ***$P < 0.00$, ****$P < 0.001$, *****$P < 0.0005$, ******$P < 0.0001$, *******$P < 0.00005$ in which *******$P < 0.00005$ denoted the highly distorted markers. The Chi-square test was used to calculate the distortion of each marker.

Annotation of Genes at The Segregation Distortion Regions (SDRs) and The Analysis of Phylogenetic Tree

Sequences corresponding to the SSR markers were identified by BLASTN to the cotton ESTs with an E ≤ 1 e-15 and were annotated using BLASTX (NCBI, Bethesda, MD, USA). The four genotypes *Gossypiumklotzschianum*, *G. davidsonii*, *G. thurberi* and *G. trilobum* have not been sequenced, the D_5, *Gossypium raimondii* was used as the reference genome. A similar method has been used to explore the genetic variation among the BC_2F_2 genotypes developed from *Gossypium hirsutum* as the recurrent parent and *Gossypium tomentosum* as the donor parent (Magwanga et al. 2018b). The mined genes within this SDR that belonged to the two most abundant subfamilies, the probable protein types and the Serine/threonine-protein kinase were then analyzed for their properties and function. A phylogenetic tree was constructed and, the multiple sequence alignments of all the proteins were done by Clustal omega, MEGA 7.0 software (Kumar et al. 2016). The neighboring method (NJ) was used with a bootstrap value of 1 000 replications, and other parameters were applied as per the default set up, as previously used in the analysis of the phylogenetic relationships of the LEA proteins in cotton (Magwanga et al. 2018b). Transcriptional response elements of genes for the two major subfamilies were predicted using an online tool, the PLACE database (http://www.dna.affrc.go.jp/PLACE/ signals can.html) (Higo et al. 1999). The genes targeted by miRNAs were predicted by searching 5' and 3´ untranslated regions (UTRs) and the coding sequences (CDS) of all the genes for their complementary sequences for the cotton miRNAs using the psRNATarget server (http://plantgrn.noble.org/psRNATarget /function).

Gene Ontology (GO) Annotation

Analysis of GO annotation was conducted using Blast2GO PRO software version 4.1.1 (https://www.blast2go.com). The GO annotations described the hierarchal roles of the genes and their products; it entailed three independent ontological terms, the molecular function (MF), biological process (BP), and cellular component (CC) (Langfelder and Horvath 2008; Magwanga et al. 2018c). The protein sequences of the dominant gene domains were obtained within the SDR regions and subsequently analyzed through Blast2GO as previously applied in the analysis of the *LEA* genes in cotton (Magwanga et al. 2018b).

RNA and RT-qPCR validation of key genes harbored within the SDR regions

Based on the previous work by our research team, *Gossypium raimondii* (D5), *Gossypium thurberi*, and *Gossypium trilobum* were profiled under biotic stress conditions, in which the plants were exposed to *Verticillium dahliae* infection (Dong et al. 2019). The

genes which were harbored within the SDR were also prominently expressed, and majorities were members of the Probable Protein Types and the Serine/Threonine-Protein Kinase. Moreover, the denovo sequencing of the *Gossypiumklotzschianum*, and *G. davidsonii* revealed a similar pattern (the data yet to be published). The highly upregulated genes were further validated under abiotic stress conditions, in which the seedlings of *G. klotzschianum, G. davidsonii, G. thurberi,* and *G. trilobum* at three leaf stage were exposed to drought and salt stress by exposing the seedings to 15% of Polyethylene glycol 6000 (PEG6000) and 250 mM NaCl, respectively. The leaf tissues were then harvested for RNA extraction at 0h, 1h, 3h, 6h and 12h of post-stress exposure. RNA extraction, purification, and RT-qPCR analysis were carried out as described by Lu et al (Lu et al. 2018). Cotton *GrActin* was applied as the reference gene.

Results

Linkage Map Construction

The first map was developed from the $F_{2:3}$ population between *G. klotzschianum* and *G. davidsonii*, a total of 728 polymorphic markers were used. The total map length was 1 480.23 cM, with an average marker interval of 2.182 cM (Kirungu et al. 2018). This map was designated as map A. The second map, designated as map B, was derived by genotyping the $F_{2:3}$ population developed between *G. thurberi* and *G. trilobum*, and 849 polymorphic markers were used in the linkage map construction. The map size was 1 012.46 cM with an average marker distance of 1.193 cM. In both maps, it was observed that chromosome number two also annotated as $D_5$02 had the least map size of 82.908 cM and 28.665 cM in map A and map B, respectively. Interestingly in both the maps, chromosome $D_5$02 had a smaller map size but with the highest percentage of SD (Table 1). Similar results have been observed in other linkage maps in cotton (Yu et al. 2011; Li et al. 2016).

The consensus map was constructed by merging two data sets from the two genetic maps. A total of 1 492 markers, were mapped onto the 13 linkage groups encompassing the 13 chromosomes, and only 85 markers remained unlinked. The diploid cotton species has 13 chromosomes, while the tetraploid cotton species has 52 chromosomes (Mendoza et al. 2013; Magwanga et al. 2018a). This work was based on the diploid cotton species of the D genome. The consensus map size was 1 467.445 cM with an average marker distance of 1.037cM. Even though the map size was relatively smaller than map A, the marker interval was low, which improved the precision of the consensus map. From the consensus map, we observed that Chromosome $D_5$02 had the highest percentage of SD with 58.621%, followed by Chromosome $D_5$07 with 47.887%. Chromosome $D_5$01 had the highest number of markers with 143 markers, while Chromosome $D_5$02 had the least number of markers of 58 (Table 1). Most of the markers mapped on the consensus map were found to be contributed by map B rather than map A. A total of 797 markers from map B were mapped on the consensus map accounting for 53.41% while only 695 markers (46.58%) were from map A. The chromosome with the highest number of markers was Chromosome $D_5$01 with 143 markers while the chromosome with the least number of markers was Chromosome $D_5$02 with only 58 markers (Fig. 1)

Segregation Distortion (SD) Analysis

In map A, out of the 728 markers mapped, 159 markers were distorted accounting for 22.2 %, and the highest SD was observed in Chromosome $D_5$02 with 76.087 % followed by Chromosome $D_5$07 with 40.698 %. The SDRs were located on Chromosome $D_5$02, $D_5$05, $D_5$07, and $D_5$08. Chromosome $D_5$02 had the largest SDR, while Chromosome $D_5$07 had the highest number of SDR.

It was observed that the alleles in the SDR region were skewed towards a particular parental line, like in Chromosome $D_5$02 towards the female parent (*G. klotzschianum*), and Chromosome $D_5$07 towards the heterozygosity(Kirungu et al. 2018). In the second genetic map B, there was a slightly lower number of distorted markers, with only 135 accounting for 15.783%, and the highest segregation distortions were observed in Chromosome $D_5$02 and Chromosome $D_5$07 with 42.857 % and 38.333%, respectively (Table 1). Chromosomes that had the SDRs were $D_5$01, $D_5$02, $D_5$06, $D_5$07, $D_5$09, $D_5$10, and $D_5$11. Moreover, the largest SDR was located on Chromosome $D_5$02, while Chromosome $D_5$07 had the highest number of SDR.

In the consensus map, the highest SDs were located on Chromosome $D_5$02 and $D_5$07, with distortion percentages of 58.621% and 47.887%, respectively. Similarly, the two chromosomes had the largest SDRs, as shown in Fig. 2. The largest SDR was

located on Chromosome $D_5$02-2 and was skewed toward the female parents while SDR located on Chromosome $D_5$02-1 was skewed towards the heterozygous. Chromosome $D_5$07 had the highest number of SDRs with a total of five SDRs, and all the SDR were skewed towards the heterozygotes except for the SDR located on Chromosome $D_5$07-1, which was skewed towards the female parents. The majority of the SDRs were skewed towards the heterozygotes; similar results were observed in the analysis of SDRs in tetraploid cotton, more specifically on the chromosome 18 (Dai et al. 2017), rice (Wu et al. 2010), and wheat (Fans et al. 1998). Based on the individual maps, the SDs were skewed towards the female compared to the male parent, the results obtained were in agreement with the study conducted on an interspecific F_2 population in which the segregated distorted markers were skewed towards the female parent Li et al. 2007.

Annotation of Genes at SDR

We conducted a blast search, and a total of 6, 038 genes were mined within the SDR region in Chromosome $D_5$02 and Chromosome $D_5$07 (Supplementary Table S2). The proportions of the genes between the two chromosomes were 2 308 genes in Chromosome $D_5$02 and 3 730 genes in Chromosome $D_5$07. These genes were further grouped according to their domain, in which a total of 1 117 domains were obtained. There are 622 domains which were shared between Chromosome $D_5$02 and Chromosome $D_5$07; the largest domain was the PF00069 (Pkinase; Protein kinase domain) with a total of 188 genes, followed by PF13855 (LRR_8; Leucine-rich repeat) with 132 genes and the third was PF07714 (Pkinase_Tyr; Protein tyrosine kinase) with a total of 108 genes. The genes in the three main domains were highly correlated with abiotic stress responsiveness. The genes located within the largest 12 domains were analyzed. Out of the 12 domains 9 domains were found to contain 'members of the resistant genes (R group of genes), these domains include Protein kinase domain; LRR_8; Leucine-rich repeat; Protein tyrosine kinase domain; NB-ARC domain; LRRNT_2; Leucine-rich repeat N-terminal domain; Pentatricopeptide repeat (PPR); Pentatricopeptide repeat (PPR_2) repeat family; Cytochromes P450 (CYPs); Myb-like DNA-binding domain and RNA recognition motif (RRM, RBD, or RNP domain) (Table 2 and Table 3).

Analysis of the Physiochemical Properties and Structures of the Genes Obtained from the Dominant Domain the(Protein kinase domain) Mined within the SDR in Chromosome $D_5$02 and Chromosome $D_5$07

The dominant domain was the Protein kinase domain (PF00069). It has been widely studied; for instance, it was found to be the dominant domain in the analysis of the genes conserved between the two upland cotton, *G. hirsutum* and its wild relative *G. tomentosum* (Magwanga et al. 2018b). We, therefore, explored the genes which belonged to this domain. The physiochemical properties of these genes showed significant variations; the molecular weight ranged between 10.351 kDa and 134.232 kDa, the charge was between -27 and 40; Isoelectric point (*pI*) values were between 4.375 and 10.382; the GRAVY values ranged between -0.721 and 0 251 while their protein lengths ranged between 611 aa and 12 310 aa (Supplementary Table S3). The GRAVY values were all below zero, indicating that these genes were mainly hydrophilic. the Protein kinase domain contained 28 different subfamilies. The subfamily with the highest number of genes was Probable types with a total of 64 genes, which included members such as the Probable inactive receptor kinase (4 genes); Probable leucine-rich repeat receptor-like serine (21), Probable L-type lectin-domain containing receptor kinase (3 genes); Probable receptor-like protein kinase (25 genes) among others (Supplementary Table S4).

The two most abundant subfamilies, the probable protein types and the Serine/threonine-protein kinase were further analyzed, by looking into their classification based on the phylogenetic tree analysis. The genes were found to be grouped into five clades, with clade 2 being the majority (Fig. 3). The most interesting concept is that the members within clade 3 had a percentage bootstrap similarity value of 100%. The majority of these genes have previously been found to be highly correlated to biotic stress tolerance; for instance, 11 genes such as *Gorai.007G335000*, *Gorai.002G039900*, *Gorai.002G040100*, *Gorai.002G041100*, *Gorai.002G041200*, *Gorai.002G041800*, *Gorai.002G042100*, *Gorai.002G047500*, *Gorai.002G047900*, *Gorai.007G182500* and *Gorai.007G334900* are homologous to an Arabidopsis gene, *At5g39020*, which has a functional role in leaf senescence during viral infection in Arabidopsis (Espinoza et al. 2007). Moreover, the remaining genes were homologous to an Arabidopsis gene, *At1g67000*, which plays a more significant role in salt stress pathways. It was also was found to be highly upregulated in the roots under salt stress conditions (Ma et al. 2006).

Cis-regulatory Elements Analyses of the Major Two Subfamilies: The Probable Protein Types and the Serine/Threonine-Protein Kinase

We examined the two major subfamilies to determine if there could be any of the regulatory elements related to either abiotic or biotic stress factors. *Cis*-regulatory elements are known to enhance the functions of the genes (Tümpel et al. 2006). In the analysis of the cis-elements, all the genes were found to be associated with either abiotic or biotic stress-responsive *cis*-regulatory elements; for instance, ARFAT with the sequence “TGTCTC” was found to be associated with 87 genes which function as ABA and auxin responsiveness. ABA is a plant phytohormone that is vital for plants' response towards stress (Trivedi et al. 2016). Other *cis*-regulatory elements predicted were CBFHV with a role in dehydration-responsive element / cold acclimation, DRECRTCOREAT functioning as activators that function in drought-, high-salt- and cold-responsive gene, lastly ABRELATERD1 with a function in early responsive to dehydration, AGMOTIFNTMYB2 induced by various stress such as wounding or elicitor treatment among others (Fig. 4 and Supplementary Table S5). The *cis*-regulatory elements detected such as ABRE have previously been found to associate with top-ranked plant stress-responsive transcription factors such as the NAC, MYB (Nakashima et al. 2009).

miRNA Prediction for the Major Two Subfamilies; The probable protein types and the Serine/threonine-protein kinase

In the prediction analysis of the miRNA targeting the various genes obtained for the two major subfamilies, a total of 287 miRNAs were found to target 91 genes (Supplementary Table 5). The high miRNA targets detected for these genes showed that the genes obtained from the SDR on chromosome $D_5$02 and chromosome $D_5$07 have a significant role in various biological processes within the plant. The highest level of miRNA target was observed for the following genes: *Gorai.002G039900* (6 miRNAs), *Gorai.002G041100* (9 miRNAs), *Gorai.002G114100* (9 miRNAs), *Gorai.002G133000* (7 miRNAs), *Gorai.002G134400* (8 miRNAs), *Gorai.007G244000* (9 miRNAs), *Gorai.007G271300* (10 miRNAs) among the rest. The miRNA's targets were observed to be very high, with a single gene being targeted by a minimum of two to a maximum of 10 miRNAs. Some of the miRNAs detected were gra-miR172a and gra-miR172b all found to target *Gorai.007G059900* which is a member of the serine/threonine-protein kinase. The SAPK2 mined within the SDR located on chromosome $D_5$07 has been found to have a function in fiber development in cotton (Abdurakhmonov et al. 2008). Moreover, miR398 has been extensively studied and found to have a role in enhancing abiotic stress tolerance in plants; for instance, gr-miR398 was found to be upregulated in plants exposed to water deficit conditions, and thus found to be responsible for enhancing tolerance towards oxidative stress, water deficit, salt stress, abscisic acid stress, ultraviolet stress, copper and phosphate deficiency, high sucrose and bacterial infection (Jia et al. 2009; Lu et al. 2010; Pashkovskii et al. 2010). The same miRNA was found to target *Gorai.007G335000* a member of the probable receptor-like protein kinase mined within the SDR on chromosome $D_5$07.

GO Annotation of the Major Two Subfamilies; The probable protein types and the Serine/threonine-protein kinase of the Dominant Gene Domains

In the analysis of the GO terms, a total of 188 genes were found to have GO terms, in which a high number of genes were found to be involved in biological process (BP), with functions such as regulation of the biological process, response to stimulus, single-organism process, metabolic process, and cellular process, in relation to cellular component (CC), four major functions were detected, namely the cell, cell part, membrane part, and membrane while in molecular functions (MF), and only two functions were observed, binding and catalytic activity (Fig. 5). Some unique observations were made in some of the genes found within the SDR regions; for instance, *Gorai.002G14960* (BRASSINOSTEROID INSENSITIVE 1-like) was found to have 20 GO functions, with 3 cellular component functions, namely endosome (C: GO:0005768), plasma membrane (C: GO:0005886) and integral component of membrane (C: GO:0016021). Five molecular functions were; protein serine/threonine kinase activity (F: GO:0004674), steroid binding (F: GO:0005496), ATP binding (F: GO:0005524), protein homodimerization activity (F: GO:0042803), and protein heterodimerization activity (F: GO:0046982). A very high number of biological processes were observed microtubule bundle formation (P: GO:0001578), protein phosphorylation (P: GO:0006468), skotomorphogenesis (P: GO:0009647), detection of brassinosteroid stimulus (P: GO:0009729), brassinosteroid mediated signaling pathway (P: GO:0009742), positive regulation of flower development (P: GO:0009911), response to UV-B (P: GO:0010224), pollen exine formation (P: GO:0010584), leaf development (P: GO:0048366), anthers wall tapetum cell differentiation (P: GO:0048657),

negative regulation of cell death (P: GO:0060548) and regulation of seedling development (P: GO:1900140). Other genes harbored a range of GO functions from three to 10 different functions (Fig. 6 and Supplementary Table S6).

RNA Sequence Data Analysis Profiled under Abiotic Stress Conditions and in Different Fiber Developmental Stages

By the fact that the two major subfamilies were found to be targeted by stress-specific miRNAs and even found to be associated with some known *cis*-regulatory elements, we undertook to investigate if the genes would have any varying expression under drought, salt and even different stages of fiber development. Genes were then obtained from the Denovo sequenced data. The raw data for the RNA sequencing were transformed into log 2 and used in the construction of the heat map. The RNA expression analysis showed that the genes were categorized into three groups, with group 1 members exhibiting higher expression analysis at different fiber development stages (Fig. 7). The majority of the highly upregulated genes were obtained from the SDR regions in chromosome $D_5$07, such as *Gorai.007G283900* (Serine/threonine-protein kinase Nek2), *Gorai.007G186000* (Probable inactive receptor kinase At1g48480), *Gorai.007G053000* (Serine/threonine-protein kinase SRK2I), *Gorai.007G285300* (Serine/threonine-protein kinase WNK1), *Gorai.007G235600* (Genome polyprotein), *Gorai.007G247600* (Serine/threonine-protein kinase ppk15) and *Gorai.007G308900*. It is interesting to note that the gene which was highly upregulated in various stages of fiber development, was also found to be targeted by gr-miR164a, and the same miRNA has been found to target the NAC transcription factor family (Xie et al. 2000). Moreover, mutant Arabidopsis lacking ath-miR164c was found to exhibit a slight defect in carpel fusion (Baker et al. 2005). In addition, miR164a,b,c has been found to have a regulatory role in the expression of CUP-SHAPED COTYLEDON1 (CUC1) and CUC2, which encode key transcriptional regulators involved in organ boundary specification (Huang et al. 2012). These previous findings show that the gene found to be targeted by miR164a/b/c could be playing an essential role in fiber development.

Under abiotic stress conditions, genes exhibited differential expression, with group 3 members exhibiting significantly higher expression under salt, cold and drought stress conditions. Some of the genes which were highly expressed include *Gorai.007G167300* (Probable serine/threonine-protein kinase WNK11), *Gorai.007G247600* (Serine/threonine-protein kinase ppk15), *Gorai.007G186000* (Probable inactive receptor kinase At1g48480), *Gorai.002G102000* (Serine/threonine-protein kinase D6PKL2), *Gorai.002G115600* (Serine/threonine-protein kinase CDL1), *Gorai.007G295100* (Serine/threonine-protein kinase CDL1), *Gorai.007G157300* (Serine/threonine-protein kinase MHK), *Gorai.007G287200* (Probable serine/threonine-protein kinase *At1g54610*), *Gorai.007G322800* (Probable serine/threonine-protein kinase *At1g09600*), *Gorai.007G078700* (Probable receptor-like protein kinase *At5g15080*) and *Gorai.007G020100* (Serine/threonine-protein kinase fray2). Among the highly expressed genes, Gorai.007G167300 was targeted by gra-miR398. Gorai.007G247600 was found to be targeted by gra-miR5207; miR398 is the first plant miRNA reported miRNA to be down-regulated by oxidative stresses. It has been intensively studied and found to be important in the regulatory process of copper homeostasis, in response to abiotic stresses such as heavy metals toxicity, sucrose, and heat, in addition to having a role in biotic stresses through the down-regulation of the expression of Cu/Zn-superoxide dismutase (CSD) (Sunkar 2006; Lu et al. 2010; Pashkovskii et al. 2010). The result shows that the SDR regions could be vital in the evolution of some of the significant genes required for the survival of the plants.

RT-qPCR Validation of the Selected Genes within the SDR Regions of Chromosome $D_5$02 and $D_5$07 under Drought and Salt Stress Conditions

Thirty genes were profiled on the leaf tissues of the four parental lines under drought and salt stress conditions. The genes exhibited three types of expressions across the four parental lines; however, more genes were found to be highly upregulated in the leaves of *G. klotzschianum* and *G. thurberi* compared with *G. davidsonii* and *G. trilobum* (Fig. 8A-D). The results obtained were in agreement to previous findings which have shown that *G. thurberi* is more tolerant to both biotic stress conditions, more so to *Verticillium dahliae* which is a fungal pathogen causing *Verticillium wilt*, a terminal disease to various crops (Dong et al. 2019). Moreover, in the study carried by Cai et al. (Cai et al. 2019) revealed that *G. thurberi* was highly tolerant to cold stress compared with *G.trilobum*. Furthermore, Kirungu et al. (Kirungu et al. 2018) found that *G. klotzschianum* harbored more beneficial traits compared with *G. davidsonii*.

Discussion

Genetic maps have become significantly important in understanding markers, breeding, association genetics, map-assisted gene cloning, gene mining, and mapping of quantitative trait loci (QTLs) (Golestan Hashemi et al. 2015). In our study, we integrated two genetic maps from the D genome of the diploid cotton with a mapping size of 188 $F_{2:3}$ population. The first genetic map (Map A) was composed of a genetic cross between *G. klotzschianum* (female parent) and *G. davidsonii* (male parent) while the second genetic map (Map B) was developed from *G. thurberi* (female parent) and *G. trilobum* (male parent). Map B had a higher number of markers linked and a smaller average distance as compared with map A. map B had a smaller average distance between markers as compared with map A. This map could play a fundamental role in the analysis of QTLs. In the construction of the consensus map; more markers were contributed by map B as compared with map A. Inconsistencies of marker order including the translocation or inversions between individual markers in consensus maps were observed especially on markers that were closely linked together as observed in the SDR region of Chromosome $D_5$02-2. Similar results were observed in the consensus map of flax seed (Cloutier et al. 2012).

The segregation distortion among the three maps ranged from 15.783% to 22.2% with map B having the highest percentage and map A having the lowest percentage. Segregation distorted markers have previously been studied in various plants (Takumi et al. 2013). The study of segregation distortion is of significant because distorted markers may be linked to genes or traits of interest, these genes may be beneficial or lethal to the organism. Therefore, it's important to include the segregation distortion markers in the construction of genetic maps since the exclusion of such markers could cause biasness of the data and result in the loss of significant genetic information. In our study we examined the trend of segregation distortion within Chromosome $D_5$02 and Chromosome $D_5$07. We observed that the two chromosomes had the highest segregation distorted markers.In contrats, chromosome $D_5$02 had the least mapped markers with a higher percentage of segregation distortion ranging between 42.857% and 76.087% in the three genetic maps. Similar results have been observed in cotton (Li et al. 2007; Khan et al. 2016; Shang et al. 2016). The two chromosomes also showed SDR which was skewed towards a specific allele. These SDRs may be due to, pre- or post-zygotic selection and chromosome loss or rearrangements.

From the analysis of the genes located on the dominant domain of Protein kinase, we observed that 29 genes were not disrupted by introns (intronless); intronless genes contain a single exon and do not contain introns from its beginning to the end neither in its UTR or CDS regions (Yan et al. 2016). The intronless genes are known to promote the efficiency of transcription initiation and elongation in spliced genes (Sakharkar et al. 2006). Their Isoelectric point (*pI*) values ranged from both acidic to basic proteins. The *pI* values are known to affect the solubility of protein molecules; hence proteins are less soluble when the pH of the solution is at its isoelectric point (Dawes et al. 1994). All of the proteins were observed to have a GRAVY value less than zero, indicating that they were hydrophilic. Hydrophilic proteins have a high solubility; hence these proteins could be playing a significant role in desiccation tolerance (Hundertmark and Hincha 2008), and also aid in enzymatic activities involved in the biochemical processes.

The analysis of the genes mined within the SDR of chromosome $D_5$02 and chromosome $D_5$07, revealed that the dominant domain was the Pkinase gene family, with a Pfam number of PF00069. There were so many genes within this domain, it was technically impossible to analyze all of them, and thus, we determined the dominant subfamily, and further analyzed two of them. The two major dominant subfamilies were the probable kinases and the serine/threonine kinases genes. These domains have been widely studied in both plants and animals (Jun et al. 2015). In the cotton plant, overexpression of *GbRLK*, a putative receptor-like kinase gene, has been found to confer tolerance to *Verticillium wilt*, a plant disease that is known to cause massive losses in cotton production regions (Jun et al. 2015).

Similarly, overexpression of the *GbRLK* gene isolated from *G. barbadense* has been found to confer drought and salt stress tolerance in transgenic Arabidopsis plants (Zhao et al. 2013). The detection of these genes within the SDR regions demonstrates the significant role played by the SDRs in the evolution or synthesis of vital proteins with a profound role in enhancing tolerance levels of plants to various abiotic and biotic stress factors. The main genes found to be located within the SDR in the consensus map were the R genes. This group of genes is known to play an integral role in signaling during pathogen recognition; hence assist in the activation of plant defense mechanisms.

The R genes work in coordination with other domains to bring combinatorial variations in signal response specificity to pathogens. Moreover, the R genes are mainly associated with proteins that identify specific pathogen effectors, known as avirulence proteins, which work in a particluar genes. These genes are known to have a gene-to-gene interaction between an organism and its pathogens (Rouxel and Balesdent 2010). These genes were segregating within the SDR in synchrony intending to help in plant defense mechanisms, these mechanisms are involved in a series of enzymatic activities within the proteins. From the recent analysis, it has been observed that the proteins encoded by resistance genes (R genes) display modular domain structures and require several dynamic interactions between specific domains to perform their function (Wang et al. 2016), hence a very close interaction of these genes at SDR. In a study conducted on determining significant QTLs for drought stress tolerance, the majority of the marker loci co-localized with known QTLs for blast tolerance or NBS-LRR disease resistance genes were located within the regions of significantly distortion levels (Dixit et al. 2014). Similar observation on Bangladeshi rice landrace Capsule in relation to salt stress tolerance (Rahman et al. 2019). The four parental lines used in the construction of the genetic map are known to contain traits for resistance to bacterial blight (*G. davidsonii*), sucking pests such as aphids (*G. klotzschianum*), *Fusarium wilt*, silver leaf whitefly and cotton bollworm resistance (*G. thurberi*), *Verticillium wilt* (*G. trilobum*). This explains the reasons for a large number of plant resistant genes (R genes) detected within the SDR regions in chromosome $D_5$02 and $D_5$07.

The carrying out insilico analysis of the genes obtained within the SDR regions, the *cis*- regulatory elements, miRNA and GO analysis showed that the R genes could be playing a significant role within the plant. Recent evidence indicates that plant miRNAs play a role in biotic and abiotic stress responses (Sunkar et al. 2007). In the analysis of the genes obtained within the SDR regions, several miRNAs were found to target several genes; for instance, miR157a and miR157b were found to a single gene *Gorai.007G063800*, a member of the serine/threonine-protein kinase. The same miRNA family was found to be the most abundant, followed by miR156, miR166, and miR168, with variation within each family in Pomegranate. This fruit has enormous importance in human health mainly because of its antioxidant properties, it does accumulate a high amount of anthocyanins in skin and arils (Saminathan et al. 2016). The antioxidant enzymes are important to plants in reducing the deleterious effects of reactive oxygen species (ROS). When plants are exposed to stress, the production and elimination of the ROS process altered leading to excessive accumulation of ROS within the cell resulting in oxidative stress. The association of miR157 to induction of antioxidant enzymes, showed that these genes within the SDR are critical for plants

The various *Cis*-regulatory elements (CREs) targeting the genes within the SDRs, were found to perform a myriad of CREs with diverse functions. More specifically it is geared towards enhancing plants tolerance to various environmental stresses; for instance, ABREATCONSENSUS targets not only the stress-responsive genes but also those involved in transportation such as the nitrate transporter (NRT) genes as evident in poplar plant (Aichi et al. 2006; Bai et al. 2013). The results obtained for the CREs were further augmented by GO annotation. The various genes obtained within the SDR regions were found to be playing an integral in all the three GO functional annotations. In cellular component (CC), functions such as an integral component of membrane (GO: 0016021), cortical microtubule (GO: 0055028) among others were detected. The integrity of the cell membrane is important because the membrane is the communicating channel between intra and extracellular environments, and any damage to the cell membrane affects various biological processes such as osmosis, thus affecting cell water retention. The detection of these cellular component roles showed that the genes found in the SDR regions have a function in maintaining cell membrane stability, and therefore enhancing the delicate osmotic balance within the cell. Moreover, an integral component of the membrane was a function found to be unanimous with the *LEA* genes (Magwanga et al. 2018b).

Several gametophytic and zygotic barriers causing deviation of allele frequencies from Mendelian ratios have been reported in several plants such as rice (Wang et al. 2009). Therefore detection of SDRs in the two populations developed from the two wild parental lines is a common feature more so among the $F_{2:3}$ populations. It is assumed based on Mendelian law that there is an equal probability of transmission of alleles from either parent during sexual reproduction, but this has not been the case in several studies, being there tend to be phenomena referred to as the preferential transmission of alleles or genotypes known as segregation distortion (SD) (Nadeau 2017). The evolution of segregation distortion may have profound evolutionary implications. From previous studies the bulk pollen sequencing indicated a rapid evolution of segregation distortion (Corbett-Detig et al. 2019). SD has been described as powerful evolutionary tools that could lead to speciation (Liberman and Feldman

1982). SDR has been observed not only among the controlled population but also among the natural population (McLaughlin and Malik 2017). The results from the two maps and their consensus showed that SDs are a common feature in segregating population and could be used to mine genes of significance that could be introgressed into the already cultivated species.

Conclusions

The use of genetic map analysis has become increasingly significant in understanding, markers assisted selection, gene mining and gene cloning. However, intensive investigation of genes located within the SDR has not been widely studied. In our research we examined the only two interspecific maps developed in the D genome of the diploid cotton. We constructed a consensus map from the two genetic maps and noted that in all the three maps $D_5$02 and $D_5$07 had the highest of SD, and hence we mined the genes within the SDR of $D_5$02 and $D_5$07 to find out if there were genes of significance that could be segregating within this region. A total of 2 308 genes in $D_5$02 and 3 730 genes in $D_5$07, were mined within the SDR, these genes were grouped into 1 117 domains of which 622 domains were shared between the two chromosomes. We further observed that the 12 largest domains had a significant role in the plant defense mechanism of which 9 out of the 12 domains belonged to the resistant genes (R group of genes) the largest domain was PF00069 with a total of 188 genes. We analyzed for the properties of these genes, the largest subdomain being the Serine/threonine-protein kinase. The analysis of the genes within the SDR revealed that genes that performed similar roles clustered together within the SDR. These genes have similar feature being hydrophilic,the study of these genes will provide future researchers with an understanding of the significance of genes within the SDR and the role of the consensus map in mining these genes.

Declarations

Acknowledgements

We are indebted to the entire cotton biology research team for their immense support and technical assistance in the course of this research work.

Availability of data:

All files supporting the findings are included within the manuscripts as figures, tables, and supplementary files.

Author Contributions

Kirungu JN, Magwanga RO, Wang K, and Liu F: Conceptualization of the concept, Kirungu JN, Magwanga RO, Wang K, Shiraku ML, Mehari TG, and Liu F Data curation, Kirungu JN, Magwanga RO, Wang K, and Liu F; Formal analysis, Kirungu JN, Magwanga RO, Wang K, and Liu F; Funding acquisition, Kirungu JN, Magwanga RO, Wang K, Liu F, Zhou Z, Pu L, Xu Y, Hou Y , Zhou Y, Cai X, Agong SG, Wang K and Liu F; Resources, Kirungu JN, Magwanga RO, Wang K, Agong SG and Liu F; Validation, Kirungu JN and Magwanga RO; Wrote the original draft, Kirungu JN and Magwanga RO; reviewed & edited the final manuscript, All authors approved the final manuscript.

Funding

This research program was financially sponsored by the 'National Key Research and Development Plan (2016YFD0100306) and the National Natural Science Foundation of China (31671745, 31530053).

Availability of data and materials:

All files supporting the findings are included within the manuscripts as figures, tables and supplementary files.

Ethics approval and consent to participate

No ethical nor consent to participate in this research was sought.

Consent for publication

No consent to publish the work was sort.

Competing interests

The authors declare no form of competing interest.

Abbreviations

SDR: segregation distortion region; GO: Gene ontology; NRT: nitrate transporter; ROS: reactive oxygen species; cM: centiMorgan; QTL: quantitative trait loci; CRE: Cis- regulatory elements; PPR: Pentatricopeptide repeat; CYPs: Cytochromes P450; CC: cellular component; MF: Molecular function; LEA: Late Embryogenesis Abundant proteins

References

1. Abdurakhmonov IY, Devor EJ, Buriev ZT, et al. Small RNA regulation of ovule development in the cotton plant, *G. hirsutum* L. BMC Plant Biol. 2008; 8: 93. doi: 10.1186/1471-2229-8-93
2. Aichi M, Yoshihara S, Yamashita M, et al. Characterization of the nitrate-nitrite transporter of the major facilitator superfamily (the *nrtP* gene product) from the cyanobacterium *Nostoc punctiforme* strain ATCC 29133. Biosci Biotechnol Biochem. 2006; 70:2682–89. doi: 10.1271/bbb.60286
3. Amudha J, Balasubramani G, Malathi VG, et al. Segregation pattern of gene expression in cotton leaf curl virus-resistant transgenics. Arch Phytopathol Plant Prot. 2012; 45:487–98. doi: 10.1080/03235408.2011.587987
4. Anhalt UCM, Heslop-Harrison PJS, Byrne S, et al. Segregation distortion in Lolium: evidence for genetic effects. Theor Appl Genet. 2008;117:297–306. doi: 10.1007/s00122-008-0774-7
5. Bai H, Euring D, Volmer K, et al. The nitrate transporter (NRT) gene family in poplar. PLoS One. 2013; 8: e72126. doi: 10.1371/journal.pone.0072126
6. Baker CC, Sieber P, Wellmer F, et al. The early extra petals1 mutant uncovers a role for microRNA miR164c in regulating petal number in Arabidopsis. Curr Biol. 2005; 15:303–15. doi: 10.1016/j.cub.2005.02.017
7. Bovill WD, Ma W, Ritter K, et al. Identification of novel QTL for resistance to crown rot in the doubled haploid wheat population “W21MMT70” x “Mendos.” Plant Breed. 2006;125:538–543. doi: 10.1111/j.1439-0523.2006.01251.x
8. Cai X, Magwanga RO, Xu Y, et al. Comparative transcriptome , physiological and biochemical analyses reveal response mechanism mediated by *CBF4* and *ICE2* in enhancing cold stress tolerance in *Gossypium thurberi*. AoB plants. 2019; 1–17. doi: 10.1093/aobpla/plz045
9. Chandnani R, Wang B, Draye X, et al. Segregation distortion and genome-wide digenic interactions affect transmission of introgressed chromatin from wild cotton species. Theor Appl Genet. 2017; 130:2219–30. doi: 10.1007/s00122-017-2952-y
10. Cheng J, Zhao Z, Li B, et al. A comprehensive characterization of simple sequence repeats in pepper genomes provides valuable resources for marker development in Capsicum. Sci Rep. 2016; 1–12. doi: 10.1038/srep18919
11. Cloutier S, Ragupathy R, Miranda E, et al. Integrated consensus genetic and physical maps of flax (*Linum usitatissimum* L.). Theor Appl Genet. 2012; 125:1783–95. doi: 10.1007/s00122-012-1953-0
12. Corbett-Detig R, Medina P, Frérot H, et al. Bulk pollen sequencing reveals rapid evolution of segregation distortion in the male germline of Arabidopsis hybrids. Evol Lett. 2019; 3:93–103. doi: 10.1002/evl3.96
13. Coulton A, Przewieslik-allen AM, Burridge AJ, et al. Segregation distortion : Utilizing simulated genotyping data to evaluate statistical methods. PLoS One. 2020; 15(2): e0228951. doi: 10.1371/journal.pone.0228951
14. Dai B, Guo H, Huang C, et al. Identification and Characterization of Segregation Distortion Loci on Cotton Chromosome 18. Front Plant Sci. 2017; 7: 2037. doi: 10.3389/fpls.2016.02037
15. Dawes H, Boyes S, Keene J, et al. Protein Instability of Wines: Influence of Protein Isolelectric Point. Am J Enol Vitic. 1994; 45:319–326

16. Dixit S, Huang BE, Sta Cruz MaT, et al. QTLs for tolerance of drought and breeding for tolerance of abiotic and biotic stress: An integrated approach. PLoS ONE. 2014; 9 (10): e109574. https://doi.org/10.1371/journal.pone.0109574. '
17. Dong Q, Magwanga R, Cai X, et al. RNA-Sequencing, Physiological and RNAi Analyses Provide Insights into the Response Mechanism of the ABC-Mediated Resistance to *Verticillium dahliae* Infection in Cotton. Genes (Basel). 2019; 10:110. doi: 10.3390/genes10020110
18. Espinoza C, Medina C, Somerville S, et al. Senescence-associated genes induced during compatible viral interactions with grapevine and Arabidopsis. J Exp Bot. 2007; 58:3197–212. doi: 10.1093/jxb/erm165
19. Faleiro FG, Schuster I, Ragagnin VA, et al. Characterization of recombinant inbred lines and QTL mapping associated to the cycle and yield of common bean. Pesqui Agropecu Bras. 2003; 38:1387–1397
20. Fans JD, Laddomada B, and Gill BS. Molecular mapping of segregation distortion loci in *Aegilops tauschii*. Genetics. 1998; 149:319–327
21. Golestan Hashemi FS, Rafii MY, Ismail MR, et al. The genetic and molecular origin of natural variation for the fragrance trait in an elite Malaysian aromatic rice through quantitative trait loci mapping using SSR and gene-based markers. Gene. 2015; 555:101–107. doi: 10.1016/j.gene.2014.10.048
22. Higo K, Ugawa Y, Iwamoto M, et al. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. 1999; 27:297–300
23. Huang L, Deng X, Li R, et al. A Fast Silver Staining Protocol Enabling Simple and Efficient Detection of SSR Markers using a Non-denaturing Polyacrylamide Gel. J Vis Exp. . 2018: e57192 doi: 10.3791/57192
24. Huang T, Lopez-Giraldez F, Townsend JP, et al. RBE controls microRNA164 expression to effect floral organogenesis. Development. 2012; 139:2161–2169. doi: 10.1242/dev.075069
25. Hundertmark M, Hincha DK. LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. BMC Genomics. 2008; 9:118. doi: 10.1186/1471-2164-9-118
26. Jamshed M, Jia F, Gong J, et al. Identification of stable quantitative trait loci (QTLs) for fiber quality traits across multiple environments in *Gossypium hirsutum* recombinant inbred line population. BMC Genomics. 2016; 17:1–13. doi: 10.1186/s12864-016-2560-2
27. Jia X, Wang WX, Ren L, et al. Differential and dynamic regulation of miR398 in response to ABA and salt stress in *Populus tremula* and *Arabidopsis thaliana*. Plant Mol Biol. 2009; 71:51–59. doi: 10.1007/s11103-009-9508-8
28. Jun Z, Zhang Z, Gao Y, et al. Overexpression of GbRLK, a putative receptor-like kinase gene, improved cotton tolerance to *Verticillium wilt*. Sci Rep. 2015; 5:. doi: 10.1038/srep15048
29. Khan MKR, Chen H, Zhou Z, et al. Genome Wide SSR High Density Genetic Map Construction from an Interspecific Cross of *Gossypium hirsutum* × *Gossypium tomentosum*. Front Plant Sci. 2016; 7:. doi: 10.3389/fpls.2016.00436
30. Kirungu JN, Deng Y, Cai X, et al. Simple sequence repeat (SSR) genetic linkage map of D genome diploid cotton derived from an interspecific cross between *Gossypium davidsonii* and *Gossypium klotzschianum*. Int J Mol Sci. 2018; 19:. doi: 10.3390/ijms19010204
31. Kumar S, Gill BS, and Faris JD. Identification and characterization of segregation distortion loci along chromosome 5B in tetraploid wheat. Mol Genet Genomics. 2007; 278:187–96. doi: 10.1007/s00438-007-0248-7
32. Kumar S, Stecher G, and Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33:1870–74. doi: 10.1093/molbev/msw054
33. Kumari JR, Srikumari CR, Valenzuela CY. ABO segregation distortion in Visakhapatnam, India. Anthropol Anz. 1992; 50:307–14
34. Langfelder P, and Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9:. doi: 10.1186/1471-2105-9-559
35. Li P, Kirungu JN, Lu H, et al. SSR-Linkage map of interspecific populations derived from *Gossypium trilobum* and *Gossypium thurberi* and determination of genes harbored within the segregating distortion regions. PLoS One. 2018; 13:e0207271. doi: 10.1371/journal.pone.0207271

36. Li W, Lin Z, Zhang X. A Novel Segregation Distortion in Intraspecific Population of Asian Cotton (*Gossypium arboretum* L.) Detected by Molecular Markers. J Genet Genomics. 2007; 34:634–40. doi: 10.1016/S1673-8527 (07) 60072-1
37. Li X, Jin X, Wang H, et al. Structure, evolution, and comparative genomics of tetraploid cotton based on a high-density genetic linkage map. DNA Res. 2016;23: 283–293. doi: 10.1093/dnares/dsw016
38. Liberman U, and Feldman MW. On the evolution of fluctuating segregation distortion. Theor Popul Biol. 1982; 21:301–317. doi: 10.1016/0040-5809(82)90020-X
39. Liu F, Wu X-L, and Chen S-Y. Segregation distortion of molecular markers in recombinant inbred populations in soybean (*G. max*). Acta Genet Sin. 2000; 27(10):883-7. Chinese. PMID: 11192432.
40. Liu X, You J, Guo L, et al. Genetic Analysis of Segregation Distortion of SSR Markers in F_2 Population of Barley. J Agric Sci. 2011; 3:172–77. doi: 10.5539/jas.v3n2p172
41. Lu H, Romero-Severson J, and Bernardo R. Chromosomal regions associated with segregation distortion in maize. Theor Appl Genet. 2002; 105:622–28. doi: 10.1007/s00122-002-0970-9
42. Lu P, Magwanga RO, Lu H, et al. A novel G-protein-coupled receptors gene from upland cotton enhances salt stress tolerance in transgenic Arabidopsis. Genes (Basel). 2018; 9(4):209 :. doi: 10.3390/genes9040209
43. Lu Y, Feng Z, Bian L, et al. miR398 regulation in rice of the responses to abiotic and biotic stresses depends on CSD1 and CSD2 expression. Functional Plant Biol. 2010; 38:44–53. doi.org/10.1071/FP10178
44. Ma S, Gong Q, Bohnert HJ. Dissecting salt stress pathways. In: Journal of Experimental Botany. 2006; 57 (5): 1097–1107. doi.org/10.1093/jxb/erj098
45. Magwanga, R.O., Lu, P., Kirungu, J.N. et al. Identification of QTLs and candidate genes for physiological traits associated with drought tolerance in cotton. J Cotton Res. 2020; 3, 3. https://doi.org/10.1186/s42397-020-0043-0
46. Magwanga R, Lu P, Kirungu J, et al. GBS Mapping and Analysis of Genes Conserved between *Gossypium tomentosum* and *Gossypium hirsutum* Cotton Cultivars that Respond to Drought Stress at the Seedling Stage of the BC_2F_2 Generation. Int J Mol Sci. 2018a; 19:1614. doi: 10.3390/ijms19061614
47. Magwanga RO, Lu P, Kirungu JN, et al. Characterization of the late embryogenesis abundant (LEA) proteins family and their role in drought stress tolerance in upland cotton. BMC Genet. 2018b; 19:. doi: 10.1186/s12863-017-0596-1
48. Magwanga RO, Lu P, Kirungu JN, et al. Identification of Cotton Cyclin Dependent Kinase (CDK) Genes and Overexpression of *Gh _ D12G2017 (CDKF4)* Confer Drought and Salt Stress Tolerance in Transgenic Arabidopsis. Int J Mol Sci. 2018c; 19(9): 2625.
49. Manrique-Carpintero NC, Coombs JJ, Veilleux RE, et al. Comparative Analysis of Regions with Distorted Segregation in Three Diploid Populations of Potato. G3: Genes|Genomes|Genetics. 2016; 6:2617–28. doi: 10.1534/g3.116.030031
50. McLaughlin RN, Malik HS. Genetic conflicts: the usual suspects and beyond. J Exp Biol. 2017; 220:6–17. doi: 10.1242/jeb.148148
51. Mello CC, Kramer JM, Stinchcomb D, et al. Efficient gene transfer in *C.elegans*: extrachromosomal maintenance and integration of transforming sequences. EMBO J. 1991; 10 (12): 3959–70. PMID: 1935914; PMCID: PMC453137.'
52. Mendoza CP, Ulloa M, Abdurakhmonov IY, et al. Genetic diversity and population structure of cotton (*Gossypium* spp.) of the New World assessed by SSR markers Genetic diversity and population structure of cotton (*Gossypium* spp.) of the New World assessed by SSR markers. Botany. 2013; 11:54. doi: 10.1139/cjb-2012-0192
53. Nadeau JH. Do gametes woo? Evidence for their nonrandom union at fertilization. Genetics. 2017; 207:369–387. doi.org/10.1534/genetics.117.300109
54. Nakashima K, Ito Y, Yamaguchi-Shinozaki K. Transcriptional regulatory networks in response to abiotic stresses in Arabidopsis and grasses. Plant Physiol. 2009; 149:88–95. doi: 10.1104/pp.108.129791
55. Natwick E. Resistance to silverleaf whitefly , *Bemisia argentifolii* (Hem. , Aleyrodidae), in *Gossypium thurberi* , a wild cotton. .. Resistance to silverleaf whitefly , *Bemisia argentifolii* (Hem, Aleyrodidae), in *Gossypium thurberi* , a wild cotton species.J Applied entomology. 2006; doi: 10.1111/j.1439-0418.2006.01083.x

56. Pashkovskii PP, Ryazanskii SS, Radyukina NL, et al. MIR398 and expression regulation of the cytoplasmic Cu/Zn-superoxide dismutase gene in *Thellungiella halophila* plants under stress conditions. Russ J Plant Physiol. 2010; 57:707–14. doi: 10.1134/S1021443710050146
57. Rahman, M.A., Thomson, M.J., De Ocampo, M. et al. Assessing trait contribution and mapping novel QTL for salinity tolerance using the Bangladeshi rice landrace Capsule. *Rice*. 2019; 12, 63. https://doi.org/10.1186/s12284-019-0319-5
58. Reflinur, Kim B, Jang SM, et al. Analysis of segregation distortion and its relationship to hybrid barriers in rice. Rice (N Y). 2014; 7:3. doi: 10.1186/s12284-014-0003-8
59. Rouxel T, Balesdent M-H. Avirulence Genes. In: Encyclopedia of Life Sciences. 2010. doi.org/10.1002/9780470015902.a0021267
60. Sakharkar KR, Sakharkar MK, Culiat CT, et al. Functional and evolutionary analyses on expressed intronless genes in the mouse genome. FEBS Lett. 2006; 580:1472–78. doi: 10.1016/j.febslet.2006.01.070
61. Saminathan T, Bodunrin A, Singh N V., et al. Genome-wide identification of microRNAs in pomegranate (*Punica granatum* L.) by high-throughput sequencing. BMC Plant Biol. 2016; 16: (1):122. doi: 10.1186/s12870-016-0807-3
62. Sandler L, Golic K. Segregation distortion in drosophila. Trends Genet. 1985; 1:181–185. doi.org/10.1016/0168-9525(85)90074-5
63. Shang L, Wang Y, Wang X, et al. Genetic Analysis and QTL Detection on Fiber Traits Using Two Recombinant Inbred Lines and Their Backcross Populations in Upland Cotton. G3 (Bethesda). 2016; 6:2717–24. doi: 10.1534/g3.116.031302
64. Sunkar R. Posttranscriptional Induction of Two Cu/Zn Superoxide Dismutase Genes in Arabidopsis Is Mediated by Downregulation of miR398 and Important for Oxidative Stress Tolerance. Plant cell online. 2006; 18:2051–2065. doi: 10.1105/tpc.106.041673
65. Sunkar R, Chinnusamy V, Zhu J, et al. Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. Trends Plant Sci. 2007; 12:301–309. doi: 10.1016/j.tplants.2007.05.001
66. Takumi S, Motomura Y, Iehisa JCM, et al. Segregation distortion caused by weak hybrid necrosis in recombinant inbred lines of common wheat. Genetica. 2013; 141:463–70. doi: 10.1007/s10709-013-9745-2
67. Trivedi DK, Gill SS, Tuteja N. Edited by Narendra Tuteja and Sarvajeet S. Gill. Abscisic Acid (ABA): Biosynthesis, Regulation, and Role in Abiotic Stress Tolerance. In: Abiotic Stress Response in Plants. Wiley-VCH Verlag GmbH & Co. KGaA, Boschstr. 12, 69469 Weinheim, Germany. 2016; pp 315–326. https://doi.org/10.1002/9783527694570.ch15
68. Tsilo TJ, Jin Y, Anderson JA. Diagnostic microsatellite markers for the detection of stem rust resistance gene Sr36 in diverse genetic backgrounds of wheat. Crop Sci. 2008; 48:253–61. doi: 10.2135/cropsci2007.04.0204
69. Tümpel S, Cambronero F, Wiedemann LM, et al. Evolution of cis elements in the differential expression of two Hoxa2 coparalogous genes in pufferfish (*Takifugu rubripes)*. Proc Natl Acad Sci U S A. 2006; 103:5419–24. doi: 10.1073/pnas.0600993103
70. Voorrips RE. MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered. 2002; 93:77–78. doi: 10.1093/jhered/93.1.77
71. Wang G, He QQ, Xu ZK, et al. High segregation distortion in maize B73 x teosinte crosses. Genet Mol Res. 2012; 11:693–706. doi: 10.4238/2012.March.19.3
72. Wang S, Tan Y, Tan X, et al. Segregation distortion detected in six rice F_2 populations generated from reciprocal hybrids at three altitudes. Genet Res (Camb). 2009; 91:345–53. doi: 10.1017/S0016672309990176
73. Wang X, Yang B, Li K, et al. A Conserved *Puccinia striiformis* Protein Interacts with Wheat NPR1 and Reduces Induction of Pathogenesis-Related Genes in Response to Pathogens. Mol Plant Microbe Interact. 2016; 29: 977-89. doi: 10.1094/MPMI-10-16-0207-R
74. Wang X, Zhang Y, Qiao L, et al. Comparative analyses of simple sequence repeats (SSRs) in 23 mosquito species genomes : Identification , characterization and distribution (*Diptera : Culicidae*). J insect science. 2019; 607–19. doi: 10.1111/1744-7917.12577

75. Wei Y, Xu Y, Lu P, et al. Salt stress responsiveness of a wild cotton species (*Gossypium klotzschianum*) based on transcriptomic analysis. PLoS One. 2017; 26;12(5):. doi: 10.1371/journal.pone.0178313
76. Wu JH, Zhang XL, Luo XL, et al. Inheritance and segregation of transformants in cotton with two types of insect-resistant genes. Yi Chuan Xue Bao. 2003; 30:631–36. Chinese. PMID: 14579531.
77. Wu YP, Ko PY, Lee WC, et al. Comparative analyses of linkage maps and segregation distortion of two F_2 populations derived from japonica crossed with indica rice. Hereditas. 2010; 147:225–36. doi: 10.1111/j.1601-5223.2010.02120.x
78. Xie Q, Frugis G, Colgan D, et al. Arabidopsis NAC1 transduces auxin signal downstream of TIR1 to promote lateral root development. Genes Dev. 2000; 14:3024–36. doi: 10.1101/gad.852200
79. Xu Y, Zhu L, Xiao J, et al. Chromosomal regions associated with segregation distortion of molecular markers in F_2, backcross, doubled haploid, and recombinant inbred populations in rice (*Oryza sativa* L.). Mol Gen Genet. 1997; 253:535–45. doi: 10.1007/s004380050355
80. Yan H, Dai X, Feng K, et al. IGDD: A database of intronless genes in dicots. BMC Bioinformatics. 2016; 17, 289. https://doi.org/10.1186/s12859-016-1148-9 . Yang C, Wang Z, Yang X, et al. Segregation distortion affected by transgenes in early generations of rice crop-weed hybrid progeny: Implications for assessing potential evolutionary impacts from transgene flow into wild relatives. J Syst Evol. 2014; 52:466–76. doi: 10.1111/jse.12078
81. Yang RC, Thiagarajah MR, Bansal VK, et al. Detecting and estimating segregation distortion and linkage between glufosinate tolerance and blackleg resistance in *Brassica napus* L. Euphytica. 2006; 148:217–25. doi: 10.1007/s10681-005-9003-5
82. Yu Y, Yuan D, Liang S, et al. Genome structure of cotton revealed by a genome-wide SSR genetic map constructed from a BC_1 population between *Gossypium hirsutum* and *G. barbadense*. BMC Genomics. 2011; 12:15. doi: 10.1186/1471-2164-12-15
83. Yuan JZ, Peng N, Feng CJ, et al. Effect of marker segregation distortion on high density linkage map construction and QTL mapping in Soybean (*Glycine max* L.). Heredity (Edinb). 2019; 579–92. doi: 10.1038/s41437-019-0238-7
84. Zhang BH, Guo TL, Wang QL. Inheritance and segregation of exogenous genes in transgenic cotton. J Genet. 2000a; 79:71–75. doi: 10.1007/BF02728948
85. Zhang F, Zhu G, Du L, et al. Genetic regulation of salt stress tolerance revealed by RNA-Seq in cotton diploid wild species, *Gossypium davidsonii*. Sci Rep. 2016; 6:20582. doi: 10.1038/srep20582
86. Zhang J, Stewart J, and Mac. Economical and rapid method for extracting cotton genomic DNA. J Cott Sci. 2000b; 4:193–201
87. Zhang Y, Wang L, Xin H, et al. Construction of a high-density genetic map for sesame based on large scale marker development by specific length amplified fragment (SLAF) sequencing. BMC Plant Biol. 2013; 13:. doi: 10.1186/1471-2229-13-141
88. Zhao J, Gao Y, Zhang Z, et al. A receptor-like kinase gene (GbRLK) from *Gossypium barbadense* enhances salinity and drought-stress tolerance in Arabidopsis. BMC Plant Biol. 2013; 13:110. doi: 10.1186/1471-2229-13-110

Additional Files

Table S1: Details of primers used in this research

Table S2: Details of primers used for the RT-qPCR validation of the 30 selected genes within the SDR regions on chromosome $D_5$02 and $D_5$07

Table S3: Genes within the dominant domain

Table S4: Genes mined within the SDR of chromosome $D_5$02 and chromosome $D_5$07

Table S5: *Cis*-regulatory promoter elements identified for the genes obtained within the SDR regions

Table S6: miRNA targets prediction

Table S7: GO annotation for the genes obtained within the SDR regions of chromosome $D_5$02 and $D_5$07.

eins

Tables

Table 1: Mapping statistics for the two individual maps and the consensus genetic maps of diploid cotton in the D Genome

Chr	Marker numbers per chromosome			Average distance /cM			Map size /cM			Number of SD			Average SD /%		
	Map A	Map B	Consensus Map	Map A	Map B	Consensus Map	Map A	Map B	Consensus Map	Map A	Map B	Consens Map	Map A	Map B	Consensus Map
$D_5$01	89	60	143	1.304	1.713	0.788	116.045	102.761	112.698	3	12	13	3.371	20	9.091
$D_5$02	44	21	58	1.884	1.365	1.943	82.908	28.665	112.698	35	10	34	76.087	42.857	58.621
$D_5$03	45	56	94	2.59	1.136	1.259	116.528	63.601	118.325	8	5	7	17.778	8.929	7.447
$D_5$04	56	70	123	1.997	0.846	0.767	111.846	59.229	94.288	2	6	7	3.571	8.571	5.691
$D_5$05	49	89	136	2.361	1.04	0.856	115.671	92.563	116.432	17	8	25	34.694	8.989	18.382
$D_5$06	58	73	125	2.001	0.88	1.082	116.045	64.213	135.273	5	8	9	8.621	10.959	7.200
$D_5$07	86	60	142	1.446	1.15	0.809	124.358	69.003	114.899	35	23	68	40.698	38.333	47.887
$D_5$08	49	64	99	2.492	1.251	0.81	122.13	80.053	80.156	25	5	27	51.02	7.813	27.273
$D_5$09	69	93	141	1.697	1.038	0.944	117.06	96.559	133.117	12	10	15	16.216	10.753	10.638
$D_5$10	34	58	84	2.998	1.786	1.238	101.93	103.563	103.973	2	12	11	5.882	20.69	13.095
$D_5$11	63	82	140	1.806	0.788	1.007	113.801	64.604	140.985	5	23	25	7.937	28.049	17.857
$D_5$12	49	41	100	2.301	2.153	1.007	112.739	88.288	100.72	6	2	6	12.245	4.878	6.000
$D_5$13	37	82	107	3.491	1.212	0.971	129.164	99.356	103.881	4	11	7	10.526	13.415	6.542
Totals	728	849	1492	2.182	1.193	1.037	1480.23	1012.458	1467.445	159	135	254	22.2	15.783	18.133

Table 2: Characteristics of the genes found within the two common markers; swu16562 and swu16586 between the three genetic maps

Gene ID	Gene name	Description	Molecular weight /kDa	Charge	pI	GRAVY value	Domain list	Domain
Gorai.007G355900	NA	NA	26.649	-12	4.563	-0.408	-	
Gorai.007G356000	At4g27220	Probable disease resistance protein At4g27220	252.737	-10	6.175	-0.127	PF00931	NB-ARC domain
Gorai.007G347200	LHP1	Chromo domain-containing protein LHP1	48.046	-13.5	4.855	-1.049	PF00385	Chromatin organization modifier
Gorai.007G347300	SIGB	RNA polymerase sigma factor sigB	64.627	15.5	9.115	-0.54	PF00140	Sigma-70 factor, region 1.2
Gorai.007G347400	NA	NA	16.507	17.5	9.897	-0.956	-	
Gorai.007G347500	NA	NA	47.568	14.5	9	-0.157	PF06219	Protein of unknown function (DUF1005)
Gorai.007G347600	TFCA	Tubulin-folding cofactor A	12.859	-5	4.781	-0.821	PF02970	Tubulin binding cofactor A
Gorai.007G347700	CYP89A2	Cytochrome P450 89A2	58.73	10	8.563	-0.074	PF00067	Cytochromes P450 (CYPs)
Gorai.007G347800	CYP89A2	Cytochrome P450 89A2	58.775	15.5	9.358	-0.151	PF00067	Cytochromes P450 (CYPs)

Table 3: Distribution of genes of the 12 largest domains within $D_5$02 and $D_5$07 in the consensus map

PF number	Domain	$D_5$02 Gene number	$D_5$07 Gene number	Total genes per domain
PF00069	Protein kinase domain	71	117	188
PF13855	LRR_8; Leucine rich repeat	52	78	130
PF07714	Protein tyrosine kinase domain	55	53	108
PF00931	NB-ARC domain	15	82	97
PF08263	LRRNT_2; Leucine rich repeat N-terminal domain	31	58	89
PF00560	LRR_1; Leucine Rich Repeat	24	61	85
PF01535	Pentatricopeptide repeat (PPR)	26	51	77
PF13041	Pentatricopeptide repeat (PPR_2) repeat family	26	49	75
PF00067	Cytochromes P450 (CYPs)	29	32	61
PF00249	Myb-like DNA-binding domain	15	41	56
PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	25	29	54
PF13639	zf-RING_2; Ring finger domain	15	36	51
Total		**384**	**687**	**1071**