Genetic Map Construction and Functional Characterization of Genes within the Segregation Distortion Regions (SDRs) in the $F_{2:3}$ Generation Derived from Wild Cotton Species of the D Genome

Joy Nyangasi Kirungu [Corresp., Equal first author, 1], **Richard Odongo Magwanga** [Equal first author, 2], **Margaret Linyerera Shiraku** [3], **Pu Lu** [1], **Teame Gereziher Mehari** [1], **Yang Xiu** [1], **Zhongli Zhou** [1], **Yuanchao Xu** [1], **Yuqing Hou** [1], **Pengcheng Li** [1], **Yun Zhou** [4], **Xiaoyan Cai** [1], **Kunbo Wang** [1], **Fang Liu** [1]

[1] Cotton Biology, State Key Laboratory of Cotton Biology /Institute of Cotton Research, Chinese Academy of Agricultural Science, Anyang, Henan 455000, China., Anyang, Henan, China

[2] Cotton biology, State Key Laboratory of Cotton Biology/ Institute of Cotton Research, Chinese Academy of 15 Agricultural Sciences, Anyang, Henan 455000, China, Anyang, Henan, China

[3] Cotton Biology, State Key Laboratory of Cotton Biology/ Institute of Cotton Research, Chinese Academy of 15 Agricultural Sciences, Anyang, Henan 455000, China, Anyang, Henan, China

[4] Laboratory of Plant Stress Biology, School of Life Science, Henan University/State Key Laboratory of Cotton Biology/Henan Key Laboratory of Plant Stress Biology, Kaifeng, Henan 475004, China, Keifang, China

Corresponding Author: Joy Nyangasi Kirungu
Email address: nyangasijoy@yahoo.com

Segregation distortion (SD) is a phenomenon common among stable or segregating populations, and the principle behind it is still an issue that puzzles many researchers. An $F_{2:3}$ generations developed from the wild cotton species of the D genomes was applied to investigate the possible plant transcription factors within the segregation distortion regions (SDRs). We constructed a consensus map between two maps in the D genome, map A; *Gossypium klotzschianum* and *Gossypium davidsonii* and Map B; *Gossypium thurberi* and *Gossypium trilobum*. The two maps were developed from 188 $F_{2:3}$ populations for each map, a total of 1492 markers, were linked to the 13 linkage groups. The consensus linkage map size was 1467.445 cM with an average marker distance of 1.0370cM. Chr02 had the highest percentage of segregation distortion with 58.621% followed by Chr07 with 47.887%. A total of 6,038 genes were mined within the segregation distortion regions (SDR region) of Chr02 and Chr07 with 2,308 gene in Chr02 and 3,730 genes in Chr07, we obtained a total of 1,117 domains within the SDR with a total of 622 domains shared between the two chromosomes, the first 9 domains all belonged to the plant resistance genes (R genes), the largest domain was PF00069 with a total of 188 genes. A total of 287 miRNAs were found to target the various genes, such as 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, the genes were found to be regulated by various stress responsive *cis*-regulatory elements. RNA

profiling showed that higher numbers of genes were highly upregulated in abiotic and different fiber development stages. The result shows that the SDR regions could be playing an important role in the evolution of significant genes in plants.

Genetic Map Construction and Functional Characterization of Genes within the Segregation Distortion Regions (SDRs) in the $F_{2:3}$ Generation Derived from Wild Cotton Species of the D Genome

Joy Nyangasi Kirungu [1, ‡], Richard Odongo Magwanga[1, 2, ‡], Margaret Linyerera Shiraku1, Pu Lu1, Teame Gereziher Mehari1, Xiu Yang1, Zhongli Zhou1, Yuanchao Xu1, Yuqing Hou1, Pengcheng Li1, 3, , Yun Zhou3, Xiaoyan Cai1*, Kunbo Wang 1* and Fang Liu1, 4*

[1] State Key Laboratory of Cotton Biology /Institute of Cotton Research, Chinese Academy of Agricultural Science, Anyang, Henan 455000, China.

[2] School of biological and physical sciences (SBPS), Jaramogi Oginga Odinga University of Science and Technology (JOOUST), P.O Box 210-40601, Bondo-Kenya

[3] School of Life Science, Henan University/State Key Laboratory of Cotton Biology/Henan Key Laboratory of Plant Stress Biology, Kaifeng, Henan 475004, China.

[4]School of Agricultural Sciences, Zhengzhou University, Zhengzhou Henan, 450001, China

Corresponding Author:

1. Kunbo Wang

State Key Laboratory of Cotton Biology /Institute of Cotton Research, Chinese Academy of Agricultural Science, Anyang, Henan 455000, China.

Email address: wkbcri@cricaas.com.cn

2. Liu Fang

State Key Laboratory of Cotton Biology /Institute of Cotton Research, Chinese Academy of Agricultural Science, Anyang, Henan 455000, China.

Email address: liufcri@caas.com

Abstract

Segregation distortion (SD) is a phenomenon common among stable or segregating populations, and the principle behind it is still an issue that puzzles many researchers. An $F_{2:3}$ generations developed from the wild cotton species of the D genomes was applied to investigate the possible plant transcription factors within the segregation distortion regions (SDRs). We constructed a consensus map between two maps in the D genome, map A; *Gossypium klotzschianum* and *Gossypium davidsonii* and Map B; *Gossypium thurberi* and *Gossypium trilobum*. The two maps were developed from 188 $F_{2:3}$ populations for each map, a total of 1492 markers, were linked to the 13 linkage groups. The consensus linkage map size was 1467.445 cM with an average marker distance of 1.0370cM. Chr02 had the highest percentage of segregation distortion with 58.621% followed by Chr07 with 47.887%. A total of 6,038 genes were mined within the segregation distortion regions (SDR region) of Chr02 and Chr07 with 2,308 gene in Chr02 and 3,730 genes in Chr07, we obtained a total of 1,117 domains within the

SDR with a total of 622 domains shared between the two chromosomes, the first 9 domains all belonged to the plant resistance genes (R genes), the largest domain was PF00069 with a total of 188 genes. A total of 287 miRNAs were found to target the various genes, such as 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, the genes were found to be regulated by various stress responsive *cis*-regulatory elements. RNA profiling showed that higher numbers of genes were highly upregulated in abiotic and different fiber development stages. The result shows that the SDR regions could be playing an important role in the evolution of significant genes in plants.

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

Introduction

Segregation distortion (SD) is described as a deviation from the expected Mendelian ratio within a segregating population caused by various segregating distorter's (Anhalt et al., 2008). Some of the factors that could lead to SDs include gametic and zygotic selections, non-homologous chromosome recombination, gene transfer, environmental agents, mapping population, marker types and genetic transmission (Xian-Liang, Xue-Zhen & Tian-Zhen, 2006). In the construction of genetic map some chromosomal regions, exhibit segregating alleles skewing from their expected Mendelian ratios towards a particular alleles, this alleles tend to cluster at small genomic regions within a chromosome, these regions are referred to as Segregation distortion regions (SDR) this phenomenon is widespread in both plants and animals (Lu, Romero-Severson & Bernardo, 2002a). Research shows that SD could bring errors in the marker order and map distances in the linkage map construction hence affecting the accuracy of the maps (Xian-Liang, Xue-Zhen & Tian-Zhen, 2017). However gene of significance have been mined within the SDR regions, for example, the gene for crown rot resistance in wheat was identified within the SDR (Bovill et al., 2006), while the *Sr36* gene responsible for stem rust tolerance, its locus was detected in the SDR on chromosome 2B in wheat (Tsilo, Jin & Anderson, 2008). Segregation distortion has been observed in a wide variety of populations of organisms including insects (Sandler & Golic, 1985), plants (Xian-Liang, Xue-Zhen & Tian-Zhen, 2006), and mammals (Kumari, Srikumari & Valenzuela, 1992). Several studies have pointed out that the SDR could be harboring some significant genes with profound roles in organisms life (Manrique-Carpintero et al., 2016). Moreover, higher frequencies of SDR have been found in populations developed through interspecific as opposed to intraspecific crosses, as was found in rice in which 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). Further evidence points at the genes associated with zygotic and gametic selection could be responsible for SD (Manrique-Carpintero et al., 2016). Thus SD is suggested to be a powerful evolutionary force, being it could be responsible for the alleviation of population divergence leading to speciation (McDermott & Noor, 2010). The use of molecular markers has been preferred in the genotyping of a population 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); they have been widely used in the study of SD in the majority of plants and animals (Froelicher et al., 2000). Several studies on SDs have been conducted in plants, including rice (Reflinur et al., 2014; Yang et al., 2014), maize (Lu,

Romero-Severson & Bernardo, 2002b; Wang et al., 2012), Wheat (Kumar, Gill & Faris, 2007), Barley (Liu et al., 2011), soybean (Liu, Wu & Chen, 2000), rape seed (Yang et al., 2006) and in cotton (Zhang, Guo & Wang, 2000; Wu et al., 2003; Amudha et al., 2012), among other plants. Fans Laddomada et al (Fans, Laddomada & Gill, 1998) performed an analysis of SD in the $F_{2:3}$ population of *Aegilops tauschii* using molecular markers, he observed that some regions had skewed ratios to particular alleles in the chromosomes. The regions that showed adverse distortion was on chromosome 5D. In the evaluation of the SDs in cotton, the majority of the SDs were mainly skewed towards the male rather than the female population and on chromosome 18 (Dai et al., 2017). In all the studies conducted in unraveling the mystery of SDs in cotton, no reports have been made on exploring the SDs in the $F_{2:3}$ population derived from diploid wild cotton parental lines. The latest attempt to explore the wild cotton progenitors was the development of a backcross population between *G. hirsutum* and *G. mustelinum* (Chandnani et al., 2017). In a previous study, our laboratory developed an $F_{2:3}$ population and constructed an interspecific genetic linkage map that included 728 loci on 13 chromosomes using the $F_{2:3}$ population of 188 individuals (Kirungu et al., 2018). A total of 159 loci were found to be distorted, with chromosome 2 and chromosome 7 harboring the highest number of SDRs. However, little is known about the genetic mechanism of SD on chromosome 2 and chromosome 7 in cotton. To explore the phenomena of the SDs, we further developed an interspecific population and genotyped 188 individuals with SSR markers, primarily focusing on the exploitation of the genetic mechanism of SD in severely distorted chromosome 2 and chromosome 7. In this study we analyzed SD from three genetic maps constructed from diploid cotton in the D genome, the first map was developed from two closely related parents; *G. klotzschianum* and *G. davidsonii* (Kirungu et al., 2018), and the second map was developed from *G. thurberi and G. trilobum* (Li et al., 2018) both maps were developed from the $F_{2:3}$ population and the third map was a consensus map developed from the two genetic map. Currently, these are the only maps developed from the diploid cotton in the D genome, we specifically focused on chromosome 2 and chromosome 7 that showed severe distortions of markers. We investigated marker segregation and subsequently mined genes in the segregation distortion regions (SDRs). This study will provide an insight into the relationship between the segregating markers and the genes mined within the SDR of the genetic maps.

Materials & Methods

4.1 Parental Materials

The first genetic map (Map A) was developed from $F_{2:3}$ population derived from self-pollinating of the F_1 generation from two parental lines; *G. klotzschianum* (female parent) and *G. davidsonii* (male parent). The second genetic map (Map B) was developed $F_{2:3}$ populations developed from two parental lines *G. thurberi* (female parent) and *G. trilobum* (male parent). A total of 188 individuals were used in the analysis for each of the mapping population, the populations were developed at the National Wild Cotton Nursery in Sanya, Hainan Island, China.

4.2 Molecular Markers Genotyping

DNA extraction from the $F_{2:3}$ populations and their parents was conducted by the CTAB method with minor modifications (Zhang, Stewart & Mac, 2000). Polymerase chain reaction (PCR) was conducted and the amplified PCR products were electrophoresed on non denaturing 10% polyacrylamide gel electrophoresis in the 1×TBE buffer after silver staining the gels were

visualized (Kinter & Sherman, 2000). In map A a total of 1,000 polymorphic primers were obtained from the total 12,560 screened markers of these 994 polymorphic primers were used in the construction of linkage groups. In map B 12,560 markers were screened of which we obtained 996 polymorphic loci and 849 polymorphic markers were linked to the 13 linkage groups.

4.3 Linkage Map Construction and Segregation Distortion of Markers

The markers with less than 5% missing data were used in the formation of the two linkage maps and formation of a consensus map, this was done by the use of Joinmap 4.0 with a recombination frequency of 0.40 and a LOD score of 2.5 (Faleiro et al., 2003). The Kosambi mapping function was used to convert the recombination frequencies to map distances. The linkage groups were then drawn using Mapchart 2.3 Software (Voorrips, 2002). The markers that deviated significantly from Mendelian ratio 1:2:1 were regarded as segregation distorted markers. Segregation distortion region (SDR) was detected if closely linked markers distorted significantly in the $F_{2:3}$ population. The consensus map was constructed by merging the two individual data sets as described by (Li et al., 2015). The other parameters were set similar to the previously constructed individual map. Maps were drawn using MapChart 2.2 (Voorrips, 2002)

4.4 Segregation Distortion Analysis

Distorted segregation, deviation of the expected Mendelian ration for either recessive or dominant parental cross of 1 homozygous: 1 heterozygous and or 1 homozygous dominant: 2 heterozygous: 1 homozygous recessive was determined by a Chi-square test for each SSR markers. A Chi-square ($\chi 2$) test was performed for each marker to assess whether it significantly deviated from Mendelian segregation ratios (1:1). The markers showing segregation distortion were indicated by asterisks (*$P < 0.05$, **$P < 0.01$, ***$P < 0.00$, ****$P < 0.001$, *****$P < 0.0005$, ******$P < 0.0001$, *******$P < 0.00005$. The Chi-square test was used to calculate the distortion of each of the marker and the reason for segregation analyzed.

4.5 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 E ≤ 1 e-15 and were annotated using BLASTX (NCBI, Bethesda, MD, USA). The mined genes within this SDR were then analyzed for their properties and function. A phylogenetic tree was constructed, the multiple sequence alignments of all the proteins were done by Clustal omega, MEGA 7.0 software (Kumar, Stecher & Tamura, 2016). The neighboring method (NJ) was used with bootstrap value of 1000 replications, other parameters were applied as per the default set up, has been previously used in the analysis of the phylogeny of the LEA proteins in cotton (Magwanga et al., 2018b) Transcriptional response elements of genes of the two major subfamilies promoters were predicted using an online tool the PLACE database (http://www.dna.affrc.go.jp/PLACE/ signals can.html) (Higo et al., 1999). In addition, the genes targeted by miRNAs were predicted by searching 5′ and 3´ UTRs and the CDS of all the genes for complementary sequences of the cotton miRNAs using the psRNATarget server with default parameters (http://plantgrn.noble.org/psRNATarget/function) (Dai & Zhao, 2011).

4.6. Gene Ontology (GO) Annotation

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

Results

2.1 Linkage Map Construction

In the population derived between the *G. klotzschianum* and *G. davidsonii* population, the map was developed from 728 polymorphic markers. The total map length was 1,480.23 cM and 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*, 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 2 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 2 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 mapping populations, a total of 1,492 markers, were linked to the 13 linkage groups encompassing the 13 chromosomes, only 85 markers remained unlinked. The consensus map size was 1,467.445 cM with an average marker distance of 1.037cM. even though the map size was relatively lower compared to map A, the marker interval was lower, which improved the precision of the consensus map From the consensus map we noted that Chr02 had the highest percentage of segregation distortion with 58.621% followed by Chr07 with 47.887%. Chr01 had the highest number of markers with 143 markers while Chr02 had the least number of markers of 58 markers (Table 1). Most of the markers mapped on the consensus were found to be contributed to a larger extent by map B rather than map A markers. A total of 997 markers from map B were mapped on the consensus map accounting for 66.82% while only 495 markers (33.18%) were from map A. The chromosome with the highest number of markers was Chr01 with 143 markers while the chromosome with the least number of markers was Chr02 with only 58 markers (Figure 1)

2.2 Common Markers within the Two Individual Maps, Developed between *G. Klotzschianum* and *G davidsonii* (Map A) and (Map B) developed from *G. thurberi* and *G. trilobum*

There were 82 SSR markers that overlapped between two genetic maps; map A and B they were mainly located in Chr04, Chr05, Chr07, Chr08, Chr09, Chr10 Chr11, Chr12, and Chr13, (Figure 2) Which means the common SSR region could be a conservative region in D genomes evolution. There were inconsistencies in marker position in the consensus map with the reference to the two maps, Map A and B, this could be attributed to structural variations within the cotton D genome, mapping populations, or genotyping technical error, similar results were observed in the consensus genetic map of the Apple genome (Khan et al., 2012). The construction of the consensus map was made possible by the presence of multiple common markers across the two maps, map A and map B.

2.3 Segregation Distortion (SD) Analysis

In map A, out of the 728 markers mapped 159 markers were distorted accounting for 22.2 %, the highest SD was observed in Chr02 with 76.087 % followed by Chr07 with 40.698 %. The SDR was located on Chr02, Chr05, Chr07, and Chr08. Chr02 had the largest SDR while Chr07 had the highest number of SDR, it was noted that the alleles in the SDR region were skewed towards a particular parental allele like in Chr02 they were skewed towards the female parent (*G. klotzschianum*), while Chr07 was skewed towards the heterozygosity (Kirungu et al., 2018). In the second genetic map B there was a slightly a lower number of distorted markers, only 135 of the total mapped markers were distorted accounting for 15.783%, highest segregation distortions were observed in Chr02 and Chr07 with 42.857 % and 38.333%, respectively (Table 1). Chromosomes with the SDRs were Chr01, Chr02, Chr06, Chr07, Chr09, Chr10, and Chr11. Moreover, the largest SDR were located on Chr02 while Chr07 had the highest number of SDR.

In the consensus map, the highest SDs were also observed to be located on Chr02 and Chr07, with the percentage level of distortion of 58.621% and 47.887%, respectively, similarly, the two chromosomes had the largest SDRs as shown in (Figure 3). The largest SDR was located Chr2-2 and was skewed toward the female parents while SDR Chr02-1 was skewed towards the heterozygous. Chr07 had the highest number of SDRs with a total of five (5) SDRs, with all the SDR skewed towards the heterozygotes except for the SDR Chr07-1, which was skewed towards the female parents. 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 in wheat (Fans, Laddomada & Gill, 1998).

We analyzed the overlapped markers within Chr07, two markers,swu16562, and swu16586 showed distortion between the three genetic maps. Further analysis of the two markers revealed that swu16562 coded for two genes which were At4g27220 (Probable disease resistance protein) and *Gorai.007G355900* (an uncharacterized gene) while swu16586 coded for seven (7) genes *LHP1* (Chromo domain-containing protein), *SIGB* (RNA polymerase sigma factor), *TFCA* (Tubulin-folding cofactor A), CYP89A2 (Cytochrome P450 89A2) (2 genes) and two uncharacterized genes (Gorai.007G347400 and Gorai.007G347500).

We then looked at the gene features, protein statistics, and transcript features. Two of the cytochrome P450 genes (CYP89A2) were intronless which showed that they could possibly be important in the evolution of vital tissue-specific roles in plants (Yan et al., 2014), while two genes were highly disrupted by introns (*At4g27220* and *SIGB*) each with 8 exons. Their molecular weights ranged between 12.859 kDa and 252.737 kDa, the Isoelectric Point (pI) ranged from 4.563 and 9.897. All the grand average of hydropathy (GRAVY) values were negative indicating that these genes were mainly hydrophilic in nature (Table 2). All the genes belonged to different domain except two genes which were members of the same domain, Cytochrome P450 (CYPs), with the CYPS type of CYP89A2. We noted that all the genes have a functional role in the coordination of the plant defense mechanism against the effects of both biotic and abiotic stress factors. For instance, the At4g27220 is found to be a member of the NB-ARC domain containing disease resistance protein family that are known to causes stabilization of the active conformation, and subsequent activation of immune signaling pathways by the exchange of bound adenosine diphosphate (ADP) for adenosine triphosphate (ATP) (Pal, Chakrabarti & Basak, 2007). LHP1 act as an activator and a repressor of transcription at enzyme metabolic pathways (Agostoni et al., 1995), SIGB acts as an activator of transcription factor and positively regulate plant defense against necrotrophic pathogens (Kanamaru et al., 1999). TFCA is important for defense against fungal pathogens by guiding secretion towards the penetration

site (Abe & Hashimoto, 2005) and the CYP89A2 causes defense against the root-infecting fungal pathogens, they also act as antimicrobials that induce the expression of defense genes (Morant et al., 2003)

2.4 Annotation of Genes at SDR

We conducted a blast search and we were able to mine a total of 6,038 genes within the SDR region in Chr02 and Chr07 (Supplementary Table S1). The proportions of the genes between the two chromosomes were 2,308 gene in Chr02 and 3,730 genes in Chr07. We further grouped these genes according to their domain, in which a total of 1,117 domains were obtained, out of which 622 domains were shared between the two chromosomes, the largest domain being 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. being that the main three domains found for the genes mined within the SDR of the two chromosomes, chr02 and chr07, were highly correlated to stress responsiveness. We analyzed the genes located within the largest 12 domain. More significantly, all the 9 out of the 12 domains were found to contain genes which were 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. (a.k.a. RRM, RBD, or RNP domain) (Table 3). The R genes are mainly associated with proteins that identify specific pathogen effectors, known as avirulence proteins, they work in specific genes. They are known to have a gene-to-gene interaction between an organism and its pathogens (Rouxel & Balesdent, 2010). Hence these genes were segregating within the SDR in synchrony with an aim of helping in plant defense mechanisms, these mechanisms involve 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. The four parental lines used in the construction of the genetic map are known to harbor 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 2 and chromosome 7.

2.5 Analysis of the Physiochemical Properties and Structures of the Genes Obtained from the Dominant Domain Mined within the SDR in Chr02 and Chr07

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., 2018a). 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, charge was between -27 and 40, Isoelectric Point (pl) values were between 4.375 and 10.382 while GRAVY values ranged from -0.721 to 0 251 while their protein lengths ranged from 611 aa to 12,310 aa (Supplementary Table S2). The GRAVY values were all below zero indicating that these genes were mainly hydrophilic in

nature. this 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 S3).

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 (5) clades, with clade 2 being the majority (Figure 4). The most interesting concept is that the members within clade 3 had percentage bootstrap similarity value of 100%. Majority of these genes have been previously 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*, that 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 has been found to be playing a greater role in salt stress pathways, being it was found to be highly upregulated in the roots under salt stress conditions (Ma, Gong & Bohnert, 2006).

2.6 *Cis*-Regulatory Elements Analyses of the Major Two Subfamilies of the Dominant Gene Domains

We examined the two major subfamilies in order to determine if there could be any of the regulatory elements in relation to either abiotic or biotic stress factors. *Cis*-regulatory elements 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 and which functions as ABA and auxin responsiveness. ABA is a plant phytohormone which is vital for plants response towards stress (Trivedi, Gill & Tuteja, 2016). Other *cis*-regulatory element predicted were CBFHV with a role in dehydration-responsive element / cold acclimation, DRECRTCOREAT functions as activators that function in drought-, high-salt- and cold-responsive gene, lastly ABRELATERD1 with a function in early responsive to dehydration, AGMOTIFNTMYB2 with is induced by various stress such as wounding or elicitor treatment among others (Figure 5 and Supplementary Table S4). The *cis*-regulatory elements detected such as ABRE has been previously found to associate with top-ranked plant stress-responsive transcription factors such as the NAC, MYB (Nakashima, Ito & Yamaguchi-Shinozaki, 2009).

2.7 miRNA Prediction for the Major Two Subfamilies of the Dominant Gene Domains

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 2 and chromosome 7 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 miRNAs 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 SAPK2 mined within the SDR located on chromosome 7 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 7.

2.8 GO Annotation of the Major Two Subfamilies 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 high number of genes were found to be involved in biological process (BP), with functions such as biological regulation, regulation of 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), only two functions were observed, binding and catalytic activity (Figure 6). A 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) an integral component of membrane (C:GO:0016021). Five molecular functions which 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). Similar 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 (Supplementary Table S6)

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

Being 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. We, therefore, blasted each of the protein sequences into NCBI to obtain homologous genes from upland cotton, *G. hirsutum*, the homolog genes were used to mine the RNA expression profiling under drought, cold, salt and different stages of fiber

development from the cotton functional genome database (https://cottonfgd.org/search/). 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 under different fiber development stages (Figure 7). Majority of the highly upregulated genes were those obtained from the SDR regions in chromosome 7, 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, the same miRNA has been found to target 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 important 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 stresses condition. 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 found to be 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, and has been intensively studied and found to be important in regulation process in 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.

Discussion

Genetic maps have become significantly important in understanding markers, breeding, association genetics, and 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 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 to map A. map B had a smaller average distance between markers as compared to map A, this map could play a fundamental role in the analysis of QTLs. In the construction of the consensus map we observed more markers were contributed by map B as compared to 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 SDR region of Chr02-2 similar results were observed in the consensus map of flax seed (Cloutier et al., 2012).

The segregation distortion between 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 been previously been studied in various plants (Schon et al., 1991; Lu, Romero-Severson & Bernardo, 2002a; Matsushita et al., 2003; Baumbach et al., 2012; Takumi et al., 2013), the study of segregation distortion is 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 since the exclusion of such markers could bias the data and result in the loss of some important genetic information. In our study we examined the trend of segregation distortion within Chr02 and Chr07, we observed the two chromosomes had highest segregation distortion markers, Chr02 had the least mapped markers but had the highest percentage of segregation distortion ranging between 42.857% and 76.087% in the three genetic maps similar results have been observed in cotton (Li, Lin & Zhang, 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 on the dominant domain 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 (*pl*) values ranged from both acidic and basic proteins, the pl 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 in nature. Hydrophilic proteins have a high solubility, hence these proteins could be playing significant role in desiccation tolerance (Hundertmark & Hincha, 2008), and also aid in enzymatic activities involved in the biochemical processes.

In the analysis of the genes mined within the SDR of chr02 and chr07, the dominant domain was the Pkinase gene family, with Pfam number of PF00069. Since the genes within this domain were so many, it was technically impossible to analyze all of them, and thus, we determined the dominant subfamily, out of which two were further analyzed. The two major dominant subfamilies were the probable kinases and the serine/threonine kinases genes. These domain has been widely studied both in plants and animals (Jun et al., 2015). In the cotton plant, overexpression of *GbRLK*, a putative receptor-like kinase gene have been found to confer tolerance to *Verticillium wilt*, a plant disease which is known to cause massive losses in cotton production regions (Jun et al., 2015). Similarly, overexpression of *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 demonstrate 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 are 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.

In 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, a number of 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 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 becomes 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

s. In the analysis of the various *Cis*- regulatory elements (CREs) targeting the genes within the SDRs, we found myriad of CREs with diverse functions, more specifically geared towards enhancing plants tolerance to various environmental stresses, for instance ABREATCONSENSUS has been found to be targeting 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 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 being the membrane is the communicating channel between intra and extracellular environments, any damage to the cell membrane affects various biological process such as osmosis, and thus cell water retention, the detection of these cellular component roles showed that the genes found at the SDR regions has 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 a number of plants such as rice (Wang et al., 2009), and therefore detection of SDRs in the two populations developed from the two wild parental lines is a common feature more so among the $F_{2:3}$ generations. 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 the preferential transmission of alleles or genotypes known as segregation distortion (SD) (Nadeau, 2017). SDR has been observed not only among the controlled population but also among the natural population (McLaughlin & Malik, 2017). The results from the two maps and

their consensus shows 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 analysis of gene 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 both the three maps Chr02 and Chr07 had the highest of SD, and hence we mined the genes within the SDR of Chr02 and Chr07 to find out if there were genes of significance that could be segregated within this regions. A total of 2,308 genes in Chr02 and 3,730 genes in Chr07, were mined within the SDR, these genes were grouped into 1117 domains of which 622 domain shared between the two chromosomes. We further analyzed that the 12 largest domain had a significant role in 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 role clustered together within the SDR these genes have similar feature such they were all hydrophilic in nature the study of these genes will provide future researcher in understanding of significance of genes within the SDR, and the role of the consensus map in mining these genes

Acknowledgments

We are indebted to Prof Liu Fang for his immense support, together with Prof Wang Kunbo, for the financial and moral support in the course of this research work.

References

Abdurakhmonov IY, Devor EJ, Buriev ZT, Huang L, Makamov A, Shermatov SE, Bozorov T, Kushanov FN, Mavlonov GT, Abdukarimov A. 2008. Small RNA regulation of ovule development in the cotton plant, *G. hirsutum* L. *BMC Plant Biology* 8. DOI: 10.1186/1471-2229-8-93.

Abe T, Hashimoto T. 2005. Altered microtubule dynamics by expression of modified α-tubulin protein causes right-handed helical growth in transgenic Arabidopsis plants. *Plant Journal* 43:191–204. DOI: 10.1111/j.1365-313X.2005.02442.x.

Agostoni E, Hill F, Wittmann C, Tobler H, Mueller F. 1995. Cec-1: a chromo domain containing protein in C elegans. *International C. elegans Meeting*.

Aichi M, Yoshihara S, Yamashita M, Maeda S, Nagai K, Omata T. 2006. Characterization of the nitrate-nitrite transporter of the major facilitator superfamily (the nrtP gene product) from the cyanobacterium Nostoc punctiforme strain ATCC 29133. *Bioscience, biotechnology, and biochemistry* 70:2682–2689. DOI: 10.1271/bbb.60286.

Amudha J, Balasubramani G, Malathi VG, Monga D, Kranthi KR. 2012. Segregation pattern of gene expression in cotton leaf curl virus-resistant transgenics. *Archives of Phytopathology and Plant Protection* 45:487–498. DOI: 10.1080/03235408.2011.587987.

Anhalt UCM, Heslop-Harrison PJS, Byrne S, Guillard a, Barth S. 2008. Segregation distortion in Lolium: evidence for genetic effects. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 117:297–306. DOI: 10.1007/s00122-008-0774-7.

568 Bai H, Euring D, Volmer K, Janz D, Polle A. 2013. The nitrate transporter (NRT) gene family in
569 poplar. *PloS one* 8. DOI: 10.1371/journal.pone.0072126.
570 Baker CC, Sieber P, Wellmer F, Meyerowitz EM. 2005. The early extra petals1 mutant uncovers
571 a role for microRNA miR164c in regulating petal number in Arabidopsis. *Current Biology*
572 15:303–315. DOI: 10.1016/j.cub.2005.02.017.
573 Baumbach J, Rogers JP, Slattery RA, Narayanan NN, Xu M, Palmer RG, Bhattacharyya MK,
574 Sandhu D. 2012. Segregation distortion in a region containing a male-sterility, female-
575 sterility locus in soybean. *Plant Science* 195:151–156. DOI: 10.1016/j.plantsci.2012.07.003.
576 Bovill WD, Ma W, Ritter K, Collard BCY, Davis M, Wildermuth GB, Sutherland MW. 2006.
577 Identification of novel QTL for resistance to crown rot in the doubled haploid wheat
578 population "W21MMT70" x "Mendos." *Plant Breeding* 125:538–543. DOI:
579 10.1111/j.1439-0523.2006.01251.x.
580 Chandnani R, Wang B, Draye X, Rainville LK, Auckland S, Zhuang Z, Lubbers EL, May OL,
581 Chee PW, Paterson AH. 2017. Segregation distortion and genome-wide digenic interactions
582 affect transmission of introgressed chromatin from wild cotton species. *Theoretical and*
583 *Applied Genetics* 130:2219–2230. DOI: 10.1007/s00122-017-2952-y.
584 Cloutier S, Ragupathy R, Miranda E, Radovanovic N, Reimer E, Walichnowski A, Ward K,
585 Rowland G, Duguid S, Banik M. 2012. Integrated consensus genetic and physical maps of
586 flax (*Linum usitatissimum* L.). *Theoretical and Applied Genetics* 125:1783–1795. DOI:
587 10.1007/s00122-012-1953-0.
588 Dai B, Guo H, Huang C, Ahmed MM, Lin Z. 2017. Identification and Characterization of
589 Segregation Distortion Loci on Cotton Chromosome 18. *Frontiers in Plant Science* 7. DOI:
590 10.3389/fpls.2016.02037.
591 Dai X, Zhao PX. 2011. PsRNATarget: A plant small RNA target analysis server. *Nucleic Acids*
592 *Research* 39. DOI: 10.1093/nar/gkr319.
593 Dawes H, Boyes S, Keene J, Heatherbell D. 1994. Protein Instability of Wines: Influence of
594 Protein Isolelectric Point. *Am. J. Enol. Vitic.* 45:319–326.
595 Espinoza C, Medina C, Somerville S, Arce-Johnson P. 2007. Senescence-associated genes
596 induced during compatible viral interactions with grapevine and Arabidopsis. *Journal of*
597 *Experimental Botany* 58:3197–3212. DOI: 10.1093/jxb/erm165.
598 Faleiro FG, Schuster I, Ragagnin VA, Cruz CD, Correa RX, Moreira MA, de Barros EG. 2003.
599 Characterization of recombinant inbred lines and QTL mapping associated to the cycle and
600 yield of common bean. *Pesquisa Agropecuaria Brasileira* 38:1387–1397.
601 Fans JD, Laddomada B, Gill BS. 1998. Molecular mapping of segregation distortion loci in
602 Aegilops tauschii. *Genetics* 149:319–327.
603 Froelicher Y, Luro F, Ollitrault P, ISC Congress. 9. Orlando E-U 2000-12-03/2000-12-07; I-IS
604 of C. 2000. Analysis of meiotic behavior of the tetraploid Clausena excavata species by
605 molecular marker segregation studies. In: *In : 9th ISC Congress, 3-7 December 2000,*
606 *Orlando (Etats-Unis) / International Society of Citriculture. - s.l. : s.n., 2000*. 116.
607 Golestan Hashemi FS, Rafii MY, Ismail MR, Mohamed MTM, Rahim HA, Latif MA, Aslani F.
608 2015. The genetic and molecular origin of natural variation for the fragrance trait in an elite
609 Malaysian aromatic rice through quantitative trait loci mapping using SSR and gene-based
610 markers. *Gene* 555:101–107. DOI: 10.1016/j.gene.2014.10.048.
611 Higo K, Ugawa Y, Iwamoto M, Korenaga T. 1999. Plant *cis*-acting regulatory DNA elements
612 (PLACE) database: 1999. *Nucleic Acids Research* 27:297–300. DOI: 10.1093/nar/27.1.297.
613 Huang T, Lopez-Giraldez F, Townsend JP, Irish VF. 2012. RBE controls microRNA164

expression to effect floral organogenesis. *Development* 139:2161–2169. DOI: 10.1242/dev.075069.

Hundertmark M, Hincha DK. 2008. LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics* 9:118. DOI: 10.1186/1471-2164-9-118.

Jia X, Wang WX, Ren L, Chen QJ, Mendu V, Willcut B, Dinkins R, Tang X, Tang G. 2009. Differential and dynamic regulation of miR398 in response to ABA and salt stress in *Populus tremula* and *Arabidopsis thaliana*. *Plant Molecular Biology* 71:51–59. DOI: 10.1007/s11103-009-9508-8.

Jun Z, Zhang Z, Gao Y, Zhou L, Fang L, Chen X, Ning Z, Chen T, Guo W, Zhang T. 2015. Overexpression of GbRLK, a putative receptor-like kinase gene, improved cotton tolerance to *Verticillium wilt*. *Scientific Reports* 5. DOI: 10.1038/srep15048.

Kanamaru K, Fujiwara M, Seki M, Katagiri T, Nakamura M, Mochizuki N, Nagatani a, Shinozaki K, Tanaka K, Takahashi H. 1999. Plastidic RNA polymerase sigma factors in Arabidopsis. *Plant & cell physiology* 40:832–42.

Khan MKR, Chen H, Zhou Z, Ilyas MK, Wang X, Cai X, Wang C, Liu F, Wang K. 2016. Genome Wide SSR High Density Genetic Map Construction from an Interspecific Cross of *Gossypium hirsutum* × *Gossypium tomentosum*. *Frontiers in Plant Science* 7. DOI: 10.3389/fpls.2016.00436.

Khan MA, Han Y, Zhao YF, Troggio M, Korban SS. 2012. A Multi-Population Consensus Genetic Map Reveals Inconsistent Marker Order among Maps Likely Attributed to Structural Variations in the Apple Genome. *PLoS ONE* 7. DOI: 10.1371/journal.pone.0047864.

Kinter M, Sherman N. 2000. Basic Polyacrylamide Gel Electrophoresis. In: *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. 117–146.

Kirungu JN, Deng Y, Cai X, Magwanga RO, Zhou Z, Wang X, Wang Y, Zhang Z, Wang K, Liu F. 2018. Simple sequence repeat (SSR) genetic linkage map of D genome diploid cotton derived from an interspecific cross between *Gossypium davidsonii* and *Gossypium klotzschianum*. *International Journal of Molecular Sciences* 19. DOI: 10.3390/ijms19010204.

Kumar S, Gill BS, Faris JD. 2007. Identification and characterization of segregation distortion loci along chromosome 5B in tetraploid wheat. *Molecular Genetics and Genomics* 278:187–196. DOI: 10.1007/s00438-007-0248-7.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* 33:1870–1874. DOI: 10.1093/molbev/msw054.

Kumari JR, Srikumari CR, Valenzuela CY. 1992. ABO segregation distortion in Visakhapatnam, India. *Anthropologischer Anzeiger; Bericht uber die biologisch-anthropologische Literatur* 50:307–314.

Li C, Bai G, Chao S, Wang Z. 2015. A high-density SNP and SSR consensus map reveals segregation distortion regions in wheat. *BioMed Research International* 2015. DOI: 10.1155/2015/830618.

Li X, Jin X, Wang H, Zhang X, Lin Z. 2016. Structure, evolution, and comparative genomics of tetraploid cotton based on a high-density genetic linkage map. *DNA Research* 23:283–293. DOI: 10.1093/dnares/dsw016.

Li P, Kirungu JN, Lu H, Magwanga RO, Lu P, Cai X, Zhou Z, Wang X, Hou Y, Wang Y, Xu Y,

Peng R, Cai Y, Zhou Y, Wang K, Liu F. 2018. 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* 13:e0207271. DOI: 10.1371/journal.pone.0207271.

Li W, Lin Z, Zhang X. 2007. A Novel Segregation Distortion in Intraspecific Population of Asian Cotton (*Gossypium arboretum* L.) Detected by Molecular Markers. *Journal of Genetics and Genomics* 34:634–640. DOI: 10.1016/S1673-8527(07)60072-1.

Liu F, Wu X-L, Chen S-Y. 2000. Segregation distortion of molecular markers in recombinant inbred populations in soybean (*G. max*). *Acta Genetica Sinica* 27.

Liu X, You J, Guo L, Liu X, He Y, Yuan J, Liu G, Feng Z. 2011. Genetic Analysis of Segregation Distortion of SSR Markers in F_2 Population of Barley. *Journal of Agricultural Science* 3:172–177. DOI: 10.5539/jas.v3n2p172.

Lu Y, Feng Z, Bian L, Xie H, Liang J. 2010. miR398 regulation in rice of the responses to abiotic and biotic stresses depends on CSD1 and CSD2 expression. Functional Plant Biology 38:44–53. DOI: 10.1071/FP10178.

Lu H, Romero-Severson J, Bernardo R. 2002a. Chromosomal regions associated with segregation distortion in maize. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 105:622–628. DOI: 10.1007/s00122-002-0970-9.

Lu H, Romero-Severson J, Bernardo R. 2002b. Chromosomal regions associated with segregation distortion in maize. *Theoretical and Applied Genetics* 105:622–628. DOI: 10.1007/s00122-002-0970-9.

Ma S, Gong Q, Bohnert HJ. 2006. Dissecting salt stress pathways. In: *Journal of Experimental Botany*. 1097–1107. DOI: 10.1093/jxb/erj098.

Magwanga RO, Lu P, Kirungu JN, Diouf L, Dong Q, Hu Y, Cai X, Xu Y, Hou Y, Zhou Z, Wang X, Wang K, Liu F. 2018a. 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. *International Journal of Molecular Sciences* 19. DOI: 10.3390/ijms19061614.

Magwanga RO, Lu P, Kirungu JN, Lu H, Wang X, Cai X, Zhou Z, Zhang Z, Salih H, Wang K, Liu F. 2018b. Characterization of the late embryogenesis abundant (LEA) proteins family and their role in drought stress tolerance in upland cotton. *BMC Genetics* 19. DOI: 10.1186/s12863-017-0596-1.

Manrique-Carpintero NC, Coombs JJ, Veilleux RE, Buell CR, Douches DS. 2016. Comparative Analysis of Regions with Distorted Segregation in Three Diploid Populations of Potato. *G3: Genes|Genomes|Genetics* 6:2617–2628. DOI: 10.1534/g3.116.030031.

Matsushita S, Iseki T, Fukuta Y, Araki E, Kobayashi S, Osaki M, Yamagishi M. 2003. Characterization of segregation distortion on chromosome 3 induced in wide hybridization between indica and japonica type rice varieties. *Euphytica* 134:27–32. DOI: 10.1023/A:1026182312730.

McDermott SR, Noor MAF. 2010. The role of meiotic drive in hybrid male sterility. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365:1265–1272. DOI: 10.1098/rstb.2009.0264.

McLaughlin RN, Malik HS. 2017. Genetic conflicts: the usual suspects and beyond. *The Journal of Experimental Biology* 220:6–17. DOI: 10.1242/jeb.148148.

Morant M, Bak S, Møller BL, Werck-Reichhart D. 2003. Plant cytochromes P450: Tools for pharmacology, plant protection and phytoremediation. *Current Opinion in Biotechnology*

14:151–162. DOI: 10.1016/S0958-1669(03)00024-7.
Nadeau JH. 2017. Do gametes woo? Evidence for their nonrandom union at fertilization. *Genetics* 207:369–387. DOI: 10.1534/genetics.117.300109.
Nakashima K, Ito Y, Yamaguchi-Shinozaki K. 2009. Transcriptional regulatory networks in response to abiotic stresses in Arabidopsis and grasses. *Plant physiology* 149:88–95. DOI: 10.1104/pp.108.129791.
Pal A, Chakrabarti A, Basak J. 2007. New motifs within the NB-ARC domain of R proteins: Probable mechanisms of integration of geminiviral signatures within the host species of Fabaceae family and implications in conferring disease resistance. *Journal of Theoretical Biology* 246:564–573. DOI: 10.1016/j.jtbi.2007.01.013.
Pashkovskii PP, Ryazanskii SS, Radyukina NL, Gvozdev VA, Kuznetsov V V. 2010. MIR398 and expression regulation of the cytoplasmic Cu/Zn-superoxide dismutase gene in Thellungiella halophila plants under stress conditions. *Russian Journal of Plant Physiology* 57:707–714. DOI: 10.1134/S1021443710050146.
Reflinur, Kim B, Jang SM, Chu S-H, Bordiya Y, Akter MB, Lee J, Chin JH, Koh H-J. 2014. Analysis of segregation distortion and its relationship to hybrid barriers in rice. *Rice (New York, N.Y.)* 7:3. DOI: 10.1186/s12284-014-0003-8.
Rouxel T, Balesdent M-H. 2010. Avirulence Genes. In: *Encyclopedia of Life Sciences*. DOI: 10.1002/9780470015902.a0021267.
Sakharkar KR, Sakharkar MK, Culiat CT, Chow VTK, Pervaiz S. 2006. Functional and evolutionary analyses on expressed intronless genes in the mouse genome. *FEBS Letters* 580:1472–1478. DOI: 10.1016/j.febslet.2006.01.070.
Saminathan T, Bodunrin A, Singh N V., Devarajan R, Nimmakayala P, Jeff M, Aradhya M, Reddy UK. 2016. Genome-wide identification of microRNAs in pomegranate (*Punica granatum* L.) by high-throughput sequencing. *BMC Plant Biology* 16. DOI: 10.1186/s12870-016-0807-3.
Sandler L, Golic K. 1985. Segregation distortion in drosophila. *Trends in Genetics* 1:181–185. DOI: 10.1016/0168-9525(85)90074-5.
Schon CC, Hayes PM, Blake TK, Knapp SJ. 1991. Gametophytic Selection in a Winter X Spring Barley Cross. *Genome* 34:918–922. DOI: 10.1139/g91-141.
Shang L, Wang Y, Wang X, Liu F, Abduweli A, Cai S. 2016. Genetic Analysis and QTL Detection on Fiber Traits Using Two Recombinant Inbred Lines and Their Backcross Populations in Upland Cotton. 6:2717–2724. DOI: 10.1534/g3.116.031302.
Sunkar R. 2006. Posttranscriptional Induction of Two Cu/Zn Superoxide Dismutase Genes in Arabidopsis Is Mediated by Downregulation of miR398 and Important for Oxidative Stress Tolerance. *THE PLANT CELL ONLINE* 18:2051–2065. DOI: 10.1105/tpc.106.041673.
Sunkar R, Chinnusamy V, Zhu J, Zhu JK. 2007. Small RNAs as big players in plant abiotic stress responses and nutrient deprivation. *Trends in Plant Science* 12:301–309. DOI: 10.1016/j.tplants.2007.05.001.
Takumi S, Motomura Y, Iehisa JCM, Kobayashi F. 2013. Segregation distortion caused by weak hybrid necrosis in recombinant inbred lines of common wheat. *Genetica* 141:463–470. DOI: 10.1007/s10709-013-9745-2.
Trivedi DK, Gill SS, Tuteja N. 2016. Abscisic Acid (ABA): Biosynthesis, Regulation, and Role in Abiotic Stress Tolerance. In: *Abiotic Stress Response in Plants*. 315–326.
Tsilo TJ, Jin Y, Anderson JA. 2008. Diagnostic microsatellite markers for the detection of stem rust resistance gene Sr36 in diverse genetic backgrounds of wheat. *Crop Science* 48:253–

261. DOI: 10.2135/cropsci2007.04.0204.
Tümpel S, Cambronero F, Wiedemann LM, Krumlauf R. 2006. Evolution of cis elements in the differential expression of two Hoxa2 coparalogous genes in pufferfish (*Takifugu rubripes*). *Proceedings of the National Academy of Sciences of the United States of America* 103:5419–5424. DOI: 10.1073/pnas.0600993103.
Voorrips RE. 2002. MapChart: software for the graphical presentation of linkage maps and QTLs. *The Journal of heredity* 93:77–78. DOI: 10.1093/jhered/93.1.77.
Wang G, He QQ, Xu ZK, Song RT. 2012. High segregation distortion in maize B73 x teosinte crosses. *Genetics and molecular research : GMR* 11:693–706. DOI: 10.4238/2012.March.19.3.
Wang S, Tan Y, Tan X, Zhang Z, Wen J, Kou S. 2009. Segregation distortion detected in six rice F_2 populations generated from reciprocal hybrids at three altitudes. *Genetics Research* 91:345–353. DOI: 10.1017/S0016672309990176.
Wang X, Yang B, Li K, Kang Z, Cantu D, Dubcovsky J. 2016. A Conserved *Puccinia striiformis* Protein Interacts with Wheat NPR1 and Reduces Induction of Pathogenesis-Related Genes in Response to Pathogens. *MPMI* 29:977–98910. DOI: 10.1094/MPMI-10-16-0207-R.
Wu YP, Ko PY, Lee WC, Wei FJ, Kuo SC, Ho SW, Hour AL, Hsing YI, Lin YR. 2010. Comparative analyses of linkage maps and segregation distortion of two F_2 populations derived from japonica crossed with indica rice. *Hereditas* 147:225–236. DOI: 10.1111/j.1601-5223.2010.02120.x.
Wu JH, Zhang XL, Luo XL, Tian YC. 2003. [Inheritance and segregation of transformants in cotton with two types of insect-resistant genes]. *Yi Chuan Xue Bao* 30:631–636.
Xian-Liang S, Xue-Zhen S, Tian-Zhen Z. 2006. Segregation Distortion and Its Effect on Genetic Mapping in Plants. *Chinese Journal of Agricultural Biotechnology* 3:163–169. DOI: 10.1079/CJB2006110.
Xian-Liang S, Xue-Zhen S, Tian-Zhen Z. 2017. Segregation distortion and its effect on genetic mapping in plants. *Journal of Agricultural Biotechnology* 14:286–292. DOI: 10.1079/CJB2006110.
Xie Q, Frugis G, Colgan D, Chua NH. 2000. Arabidopsis NAC1 transduces auxin signal downstream of TIR1 to promote lateral root development. *Genes and Development* 14:3024–3036. DOI: 10.1101/gad.852200.
Xu Y, Zhu L, Xiao J, Huang N, McCouch SR. 1997. Chromosomal regions associated with segregation distortion of molecular markers in F_2, backcross, doubled haploid, and recombinant inbred populations in rice (*Oryza sativa* L.). *Molecular and General Genetics* 253:535–545. DOI: 10.1007/s004380050355.
Yan H, Dai X, Feng K, Ma Q, Yin T. 2016. IGDD: A database of intronless genes in dicots. *BMC Bioinformatics* 17. DOI: 10.1186/s12859-016-1148-9.
Yan H, Zhang W, Lin Y, Dong Q, Peng X, Jiang H, Zhu S, Cheng B. 2014. Different evolutionary patterns among intronless genes in maize genome. *Biochemical and Biophysical Research Communications* 449:146–150. DOI: 10.1016/j.bbrc.2014.05.008.
Yang RC, Thiagarajah MR, Bansal VK, Stringam GR, Rahman MH. 2006. Detecting and estimating segregation distortion and linkage between glufosinate tolerance and blackleg resistance in *Brassica napus* L. *Euphytica* 148:217–225. DOI: 10.1007/s10681-005-9003-5.
Yang C, Wang Z, Yang X, Lu BR. 2014. 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. *Journal of Systematics and*

Evolution 52:466–476. DOI: 10.1111/jse.12078.

Yu Y, Yuan D, Liang S, Li X, Wang X, Lin Z, Zhang X. 2011. Genome structure of cotton revealed by a genome-wide SSR genetic map constructed from a BC1 population between *Gossypium hirsutum* and G. barbadense. *BMC Genomics* 12:15. DOI: 10.1186/1471-2164-12-15.

Zhang BH, Guo TL, Wang QL. 2000. Inheritance and segregation of exogenous genes in transgenic cotton. *Journal of Genetics* 79:71–75. DOI: 10.1007/BF02728948.

Zhang J, Stewart J, Mac. 2000. Economical and rapid method for extracting cotton genomic DNA. *J. Cotton Sci* 4:193–201.

Zhang Y, Wang L, Xin H, Li D, Ma C, Ding X, Hong W, Zhang X. 2013. Construction of a high-density genetic map for sesame based on large scale marker development by specific length amplified fragment (SLAF) sequencing. *BMC Plant Biology* 13. DOI: 10.1186/1471-2229-13-141.

Zhao J, Gao Y, Zhang Z, Chen T, Guo W, Zhang T. 2013. A receptor-like kinase gene (GbRLK) from *Gossypium barbadense* enhances salinity and drought-stress tolerance in Arabidopsis. *BMC Plant Biology* 13:110. DOI: 10.1186/1471-2229-13-110.

Figure 1(on next page)

Consensus Genetic linkage map of cotton in D genome representing 13 linkage groups, developed from map A (*G. klotzschianum* and G*. davidsonii*) and map B (*G. thurberi* and *G. trilobum*). Markers in green font represent map B while mark

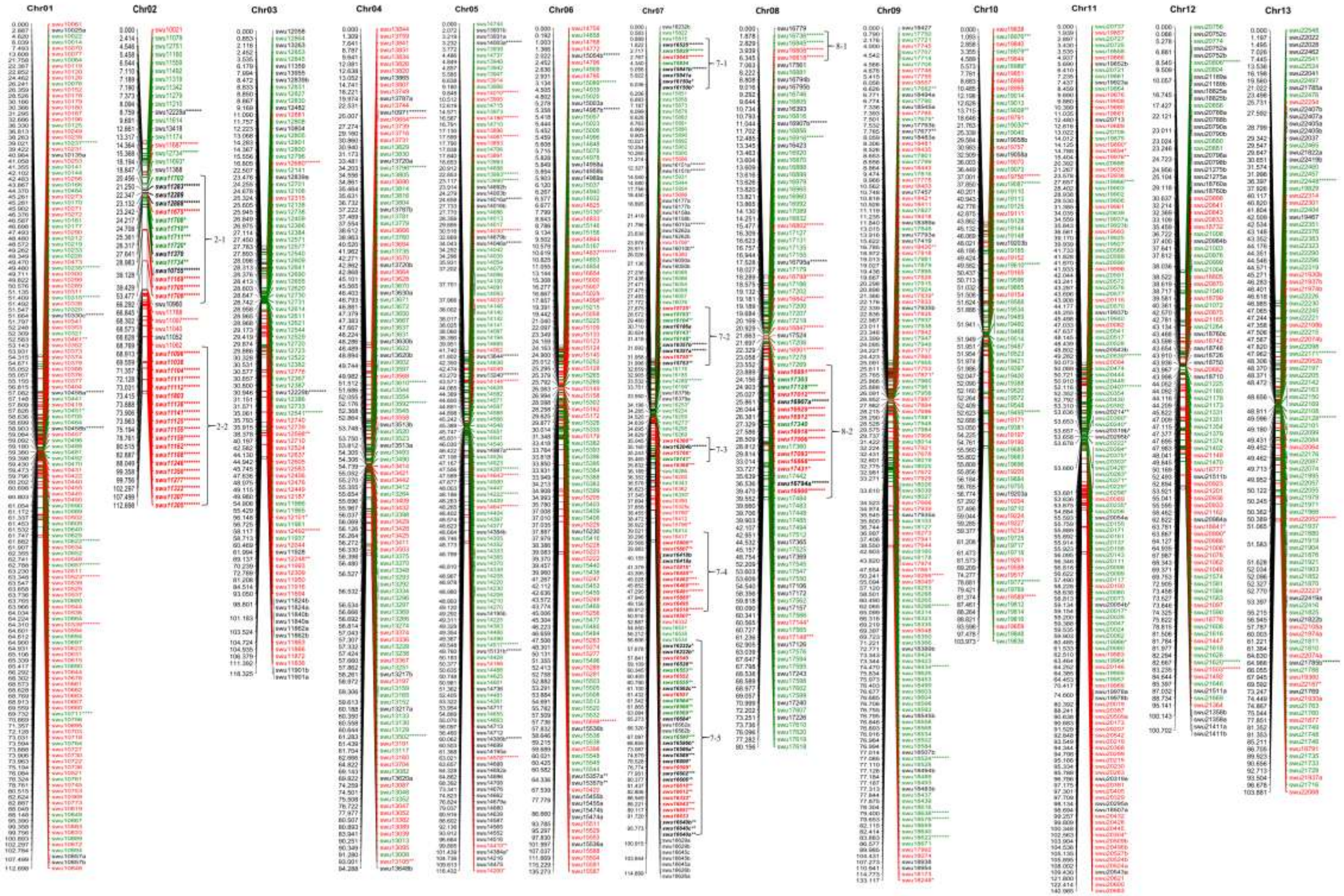


Figure 2(on next page)

Common markers between genetic map A and B

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27845v1 | CC BY 4.0 Open Access | rec: 10 Jul 2019, publ: 10 Jul 2019

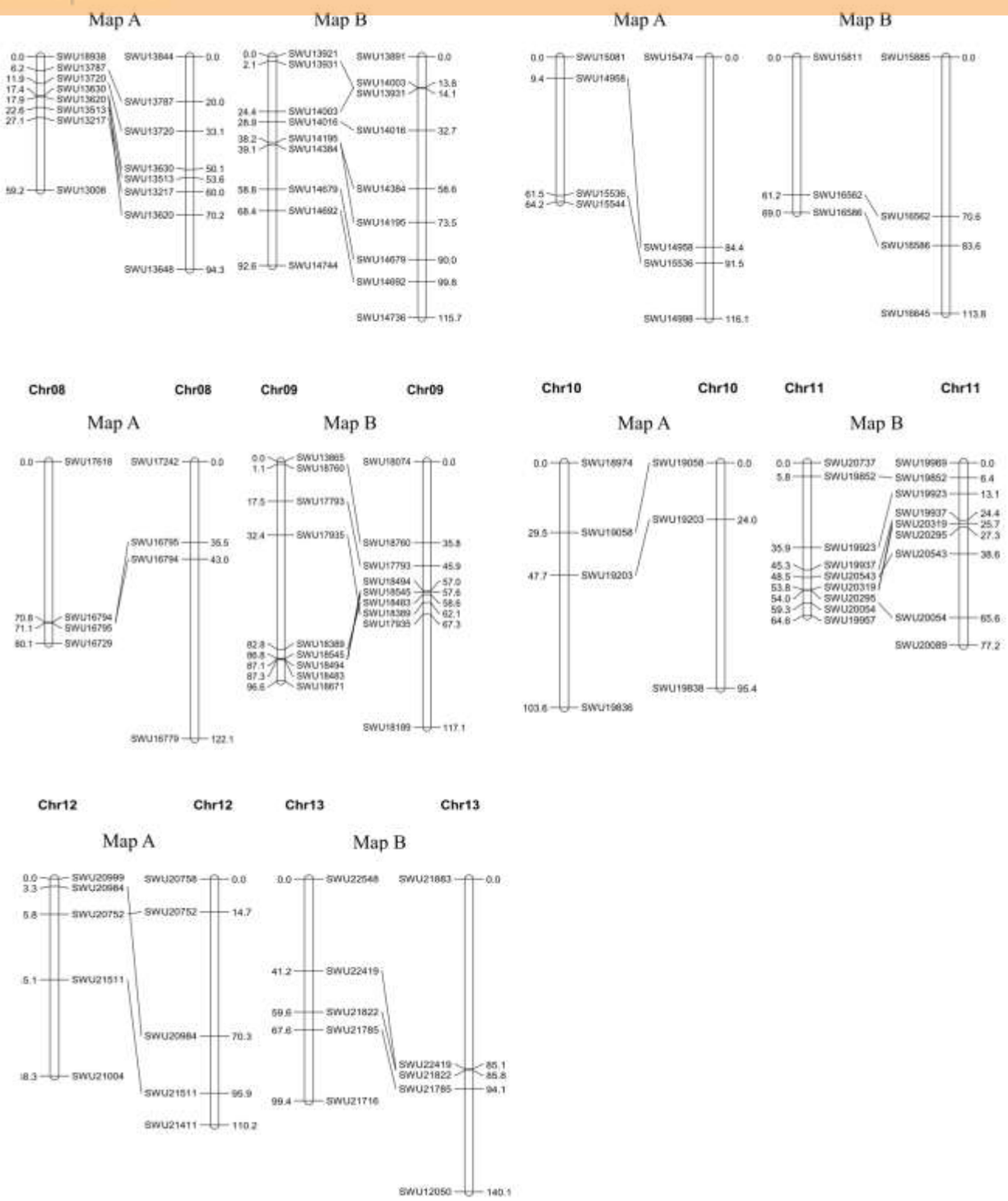


Figure 3(on next page)

Segregation Distortion region (SDR) within Chr02 and Chr07 in Map A, Consensus map and Map B; Markers in green font represent map B while markers in red font representing map A. the markers in black represent markers translocated from other chromosomes wi

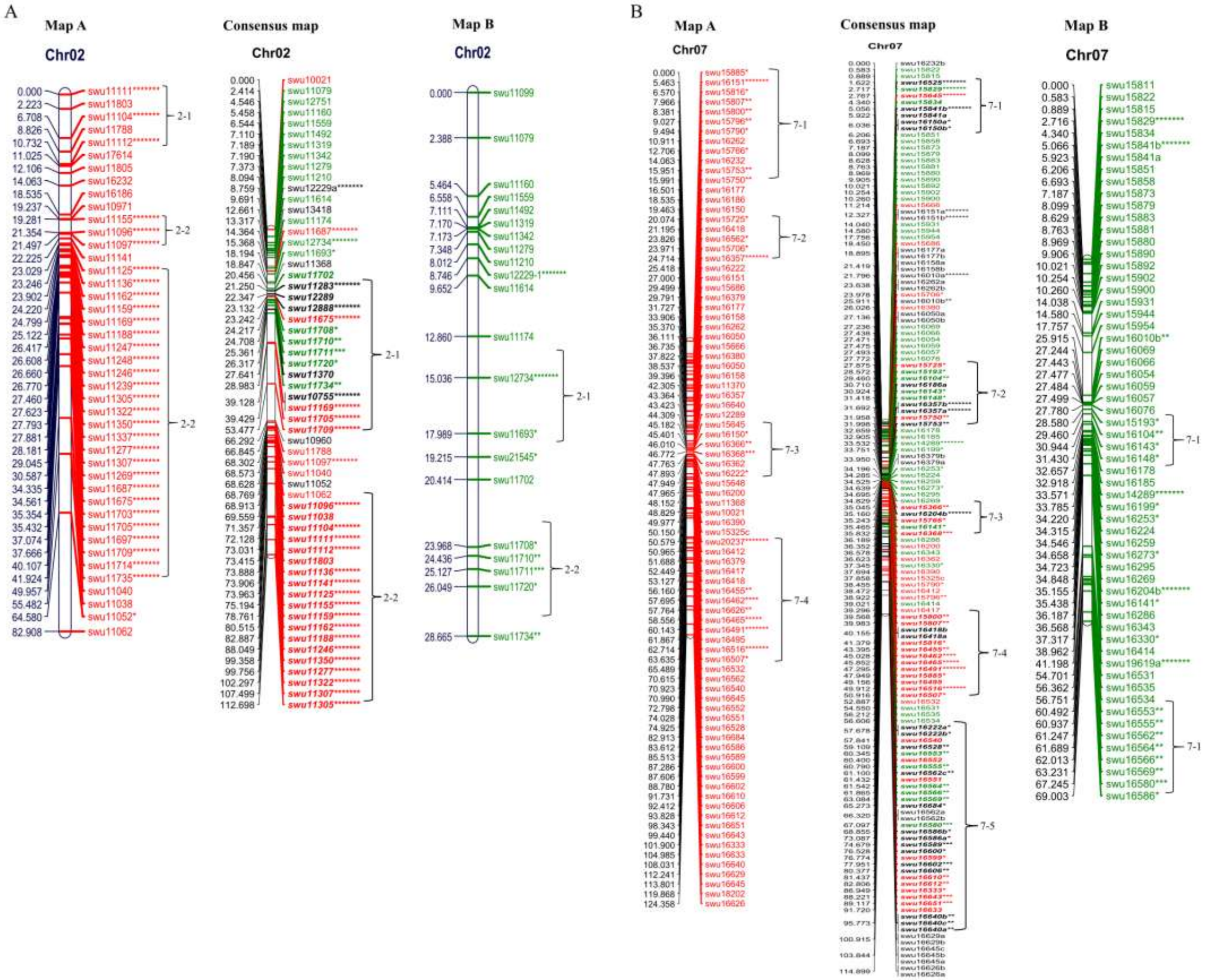


Figure 4(on next page)

Phylogenetic tree analysis of the genes found to be the most abundant sub family of the dominant domain, Pkinases mined within the SDR regions in chromosome 2 and chromosome 7.

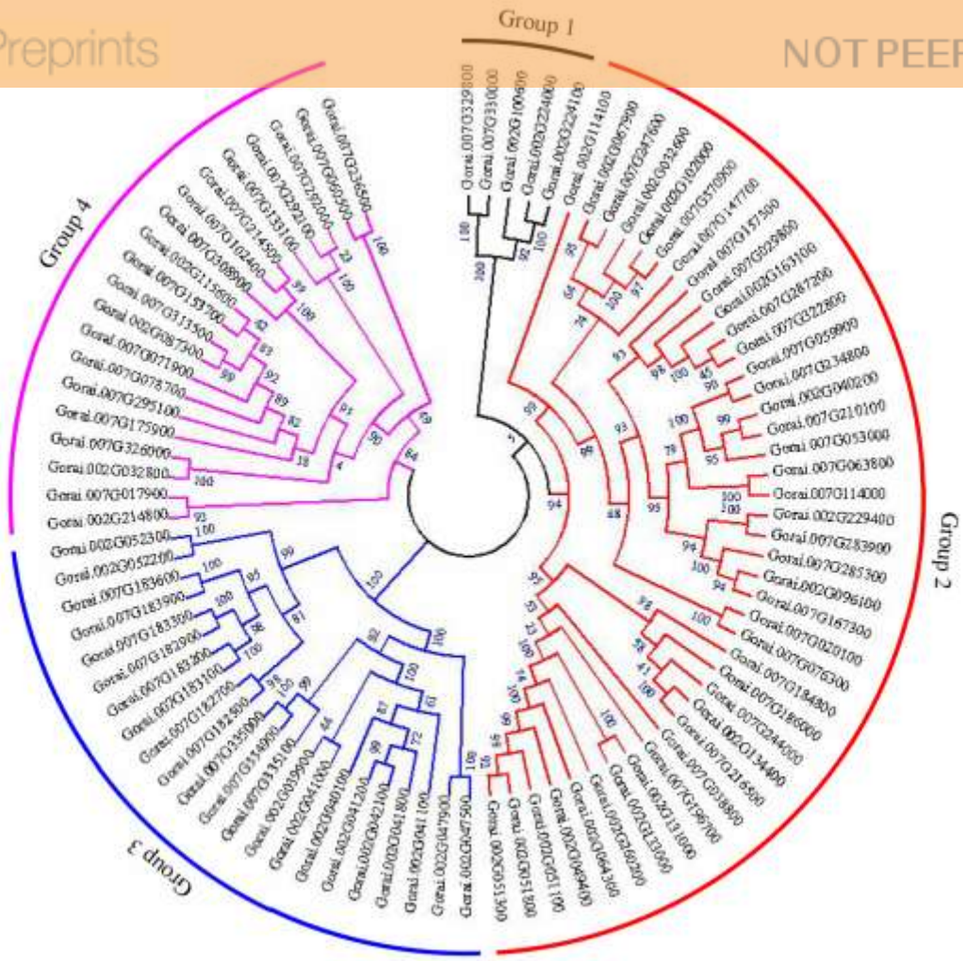


Figure 5(on next page)

Average number of the cis-promoters ABREATCONSENSUS (YACGTGGC), CBFHV (RYCGAC), DRECRTCOREAT (RCCGAC), ARR1AT (NGATT) and others in promoter region of *Gossypium raimondii* genes from the two major subfamilies of the dominant gene domain mined within

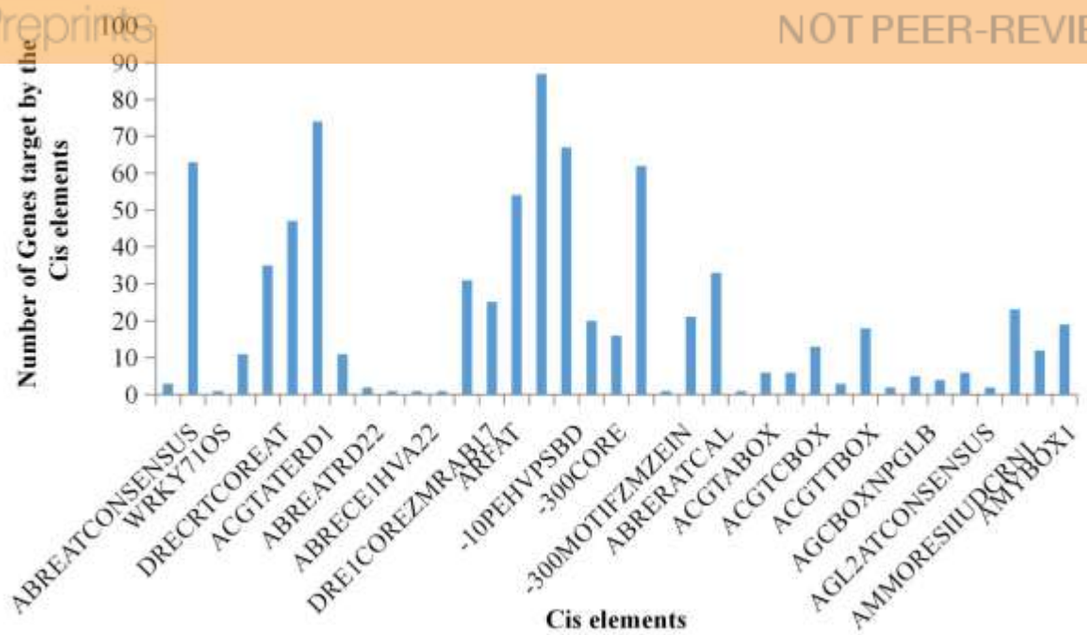
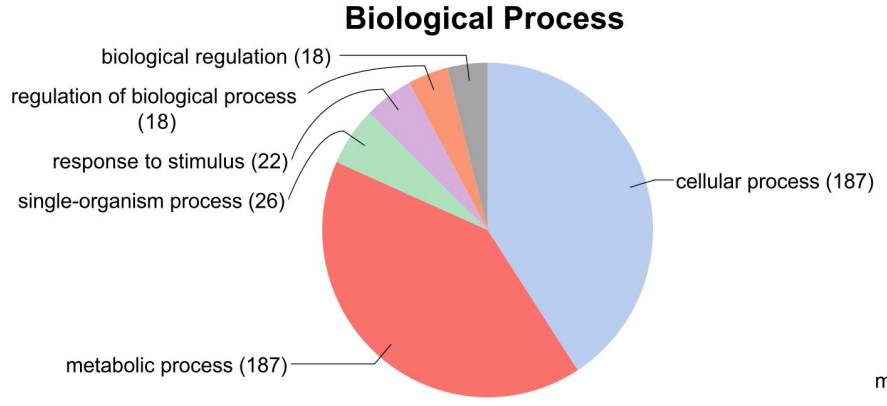
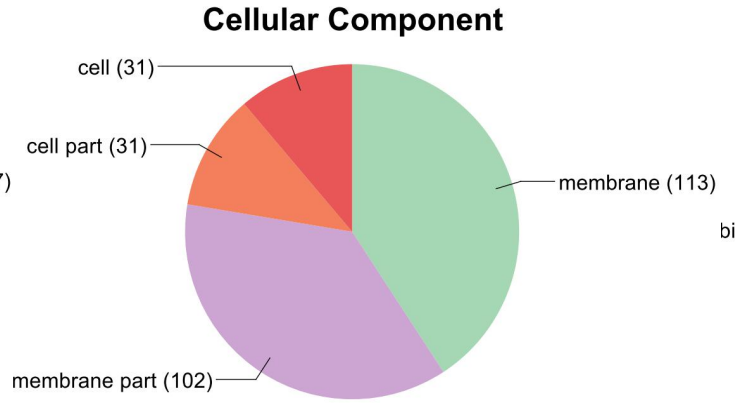
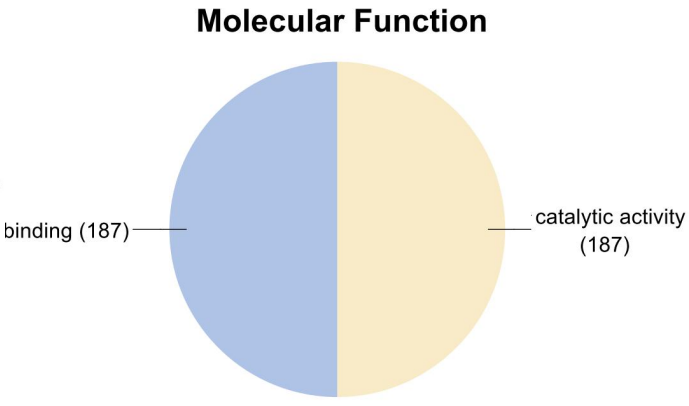


Figure 6(on next page)

Gene Ontology (GO) annotation results for the genes obtained within the SDR regions of chromosome 2 and 7. GO analysis of the 186 protein sequences predicted for their involvement in biological processes (BP), molecular functions (MF) and cellular compone







PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27845v1 | CC BY 4.0 Open Access | rec: 10 Jul 2019, publ: 10 Jul 2019

Figure 7(on next page)

Differential expression of the two major subfamilies under drought, salt, cold and fibre development. The heat map was visualized using the MeV_4_ 9_0 program. Yellow and blue indicate high and low levels of expression, respectively. (A) Heat map showing

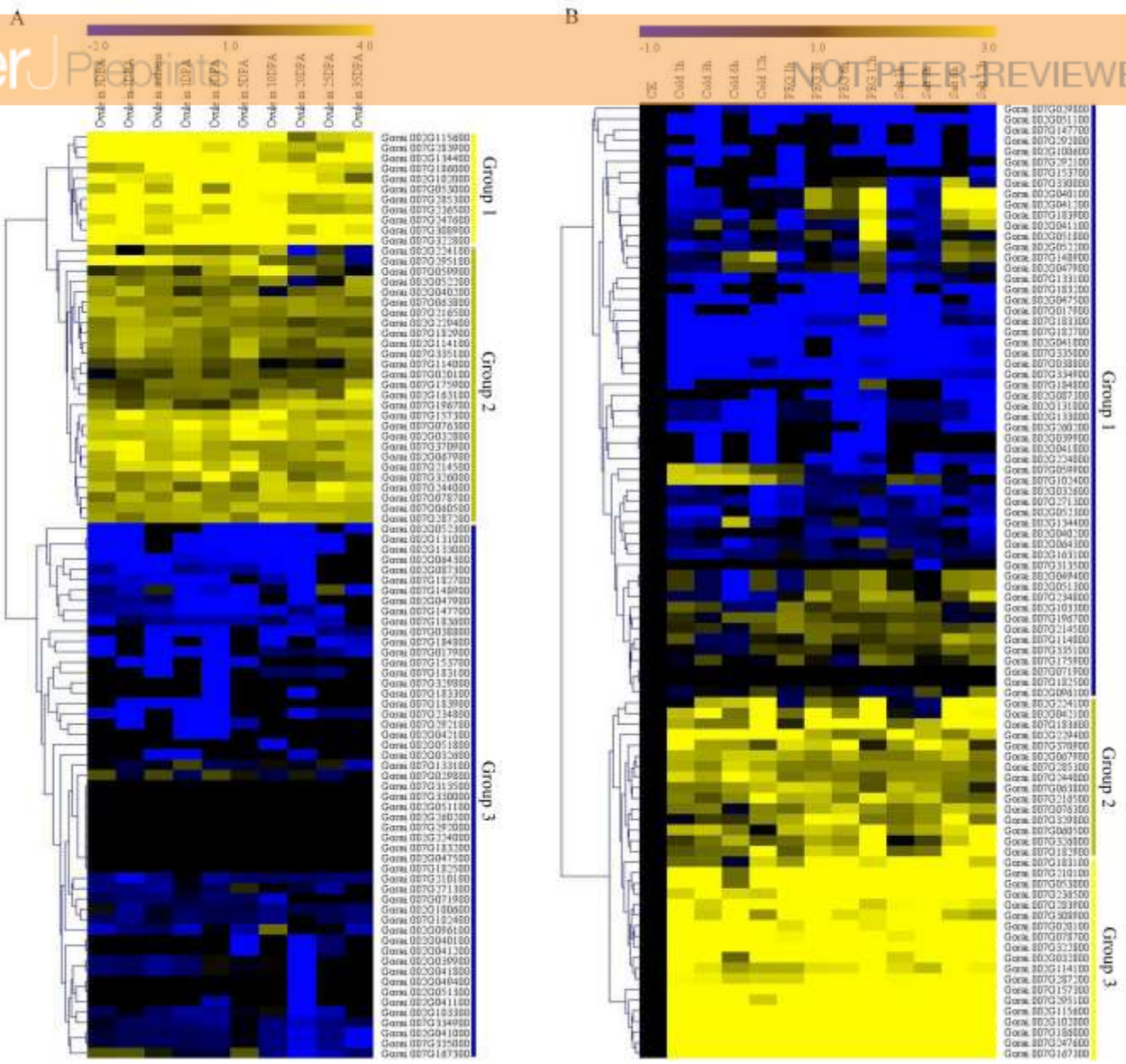


Table 1(on next page)

Mapping statistics for the two individual maps and the consensus genetic maps of diploid cotton in the D Genome Map A represent map A; *G klotzschianum* and *G davidsonii*, Map B; *G thurberi* and *G trilobum* while Consens Map represent t

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

Chr	Marker Per Chr			Average Distance (cM)			Map Size (cM)			Number Of Sl	
	Map A	Map B	Consens Map	Map A	Map B	Consens Map	Map A	Map B	Consens Map	Map A	Map
Chr01	89	60	143	1.304	1.713	0.788	116.045	102.761	112.698	3	12
Chr02	44	21	58	1.884	1.365	1.943	82.908	28.665	112.698	35	10
Chr03	45	56	94	2.59	1.136	1.259	116.528	63.601	118.325	8	5
Chr04	56	70	123	1.997	0.846	0.767	111.846	59.229	94.288	2	6
Chr05	49	89	136	2.361	1.04	0.856	115.671	92.563	116.432	17	8
Chr06	58	73	125	2.001	0.88	1.082	116.045	64.213	135.273	5	8
Chr07	86	60	142	1.446	1.15	0.809	124.358	69.003	114.899	35	23
Chr08	49	64	99	2.492	1.251	0.81	122.13	80.053	80.156	25	5
Chr09	69	93	141	1.697	1.038	0.944	117.06	96.559	133.117	12	10
Chr10	34	58	84	2.998	1.786	1.238	101.93	103.563	103.973	2	12
Chr11	63	82	140	1.806	0.788	1.007	113.801	64.604	140.985	5	23
Chr12	49	41	100	2.301	2.153	1.007	112.739	88.288	100.72	6	2
Chr13	37	82	107	3.491	1.212	0.971	129.164	99.356	103.881	4	11
Totals	728	849	1492	2.182	1.193	1.037	1480.23	1012.458	1467.445	159	135

Table 2(on next page)

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

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(on next page)

Distribution of genes of the 12 largest domains within Chr02 and Chr07 in the consensus map

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27845v1 | CC BY 4.0 Open Access | rec: 10 Jul 2019, publ: 10 Jul 2019

1 Table 1: Distribution of genes of the 12 largest domains within Chr02 and Chr07 in the consensus map

PF number	Domain	Chr02 genes	Chr07 genes	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

2