2018-04-01

The Genome Sequence of Gossypium herbaceum (A1), a Domesticated Diploid Cotton

Alex J. Freeman
Brigham Young University

The Genome Sequence of *Gossypium herbaceum* (A_1), a Domesticated Diploid Cotton

Alex J Freeman

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Craig Coleman, Chair
Joshua Udall
Peter J. Maughan
John Kauwe

Department of Plant and Wildlife Sciences

Brigham Young University

ABSTRACT

The Genome Sequence of *Gossypium herbaceum* (A_1), a Domesticated Diploid Cotton

Alex J Freeman
Department of Plant and Wildlife Sciences, BYU
Master of Science

Gossypium herbaceum is a species of cotton native to Africa and Asia. As part of a larger effort to investigate structural variation in assorted diploid and polyploid cotton genomes we have sequenced and assembled the genome of *G. herbaceum*. Cultivated *G. herbaceum* is an A_1-genome diploid from the Old World (Africa) with a genome size of approximately 1.7 Gb. Long range information is essential in constructing a high-quality assembly, especially when the genome is expected to be highly repetitive. Here we present a quality draft genome of *G. herbaceum* (*cv*. Wagad) using a multi-platform sequencing strategy (PacBio RS II, Dovetail Genomics, Phase Genomics, BioNano Genomics). PacBio RS II (60X) long reads were *de novo* assembled using the CANU assembler. Illumina sequence reads generated from the PROXIMO library method from Phase Genomics, and BioNano high-fidelity whole genome maps were used to further scaffolding. Finally, the assembly was polished using PILON. This multi-platform long range sequencing strategy will help greatly in attaining high quality *de novo* reconstructions of genomes. This assembly will be used towards comparative analysis with *G. arboreum*, which is also a domesticated A_2-genome diploid. Not only will this provide a quality reference genome for *G. herbaceum,* it also provides an opportunity to assess recent technologies such as Dovetail Genomics, Phase Genomics, and Bionano Genomics. The *G. herbaceum* genome sequence serves as an example to the plant genomics community for those who have an interest in using multi-platform sequencing technologies for *de novo* genome sequencing.

Keywords: *Gossypium, G. herbaceum,* cotton, Pacific Biosciences, draft sequence assembly, proximity guided assembly

ACKNOWLEDGEMENTS

I would like to thank my good friends and coworkers, particularly the undergraduates, who have helped all along the way in this project. Their assistance has been invaluable in facilitating and performing hundreds of Bionano high molecular weight DNA extractions, watering the cotton plants in the greenhouse and helping to manage the large quantities of data we generated over years of working on this project.

I also want to express gratitude for my graduate committee here at Brigham Young University. Through unforeseen circumstances necessitating an official change in committee membership they have helped me by giving insight into the true value of this project and its potential. They have all been supportive and helpful as I've progressed towards graduation and future publication of this manuscript.

I would like to thank my original mentor and committee chair, Dr. Joshua Udall, for giving me the opportunity to work with him here at BYU on this project. I have increased my critical thinking capacity, developed a stronger character, and deepened my work ethic, all valuable skills because of the work I have done here at BYU and by continuing my education.

I also thank my good friend Sarah Webb for encouraging me along the way, for helping me to manage stress when things felt out of control, and for being willing to edit my thesis. The hikes, spontaneous trips to hot springs, and various other activities helped to keep me sane and focus my mind when work needed to get done.

TABLE OF CONTENTS

TITLE PAGE i

ABSTRACT ii

ACKNOWLEDGEMENTS iii

TABLE OF CONTENTS iv

LIST OF FIGURES v

LIST OF TABLES vii

INTRODUCTION 1

METHODS & MATERIALS 3

Sample Collection, DNA Isolation & Library Preparation 3

Genome Sequence Data Generation 3

Genome Assembly 4

Bionano Physical Map Generation and Integration 5

RESULTS 9

Genome Sequence Assembly 9

Bionano Genome Assembly 10

Manual Integration of Bionano and Hi-C 11

Genome Annotation 12

Minimap2 Comparative Analysis 14

AN OVERVIEW OF GENOMICS 21

LITERATURE CITED 24

FIGURES 38

TABLES 46

LIST OF FIGURES

Figure 1: In A, an orientation change involving a block of two contigs can be seen in box 1. A contig that needs to be moved and then re-oriented can be seen in box 2. In B, we have the corrected PG pseudomolecule sequence from A.
.....38

Figure 2: Simple Hi-C based inversion error corrected by Bionano physical maps. Seen in A and D a .bed file representing the contigs as placed and oriented by Phase Genomics Hi-C. In B and E, we have the pseudomolecule. In C and F, we have the Bionano maps. The red boxes indicate a contig correctly placed by Hi-C, and confirmed by Bionano maps in C and F. The orientation is incorrect, however, which is represented by the inversion in ABC. After correcting the orientation in the Hi-C data, and re-generating the sequence fasta, we improved the sequence assembly, represented in DEF.
.....39

Figure 3: A PG generated group ordering file, indicating contig number in column 1, contig ID in column 2, orientation in column 3, and log link liklihood in column 4.
.....39

Figure 4: Assembly Workflow. Square boxes represent data or post-analysis data. Circles represent processes and programs ran. PB RSII, Pacific Biosciences RSII long read sequencing technology. Hi-C, Phase Genomics Hi-C data collection, titled PROXIMO. PBJELLY, gap filling software. BNG, Bionano Genomics. Pilon, error correction of PB long read data with Illumina short read data. BUSCO, gene space quality check. MAKER-P, genome annotation. GenSAS, online genome annotation unifying many separate genome annotation programs.
.....40

Figure 5: Pairwise heat map showing log-likelihood of contig placement. The X and Y axis are each contig of every pseudomolecule laid end to end. The darker the red of the long link density the more interactions each contig has with its neighboring contigs. The diagonal axis represents the alignment of each sequence to itself. “Dots” of red, exemplified by a blue circle, outside the diagonal axis represent regions which have high frequency of interactions between chromosomes. Regions consistently along the “end” or “center” of a pseudomolecule likely represent telomeres or centromeres, as indicated by black arrows. White space along the center of the diagonal represent individual contigs with sufficient length to be seen, at this level, by the naked eye. One particular contig is indicated with a blue arrow.
.....41

Figure 6: Exemplification of Bionano motif repeat identified with Bionano contigs. A: Geneious *in silico* digestion identifying BSSSI nick sites. B: PG Assembly contig with *in silico* digestion with BSSSI nick sites. C: Bionano contigs aligning to the motif repeat area. D: Magnified selection of the motif repeat.
.....42

Figure 7: Repeat Modeler and Repeat Masker results from the GenSAS annotation.
.....43

Figure 8: Minimap2 sequence to sequence alignments. Alignment of *G. herbaceum* to *G. raimondii*, top left. Alignment of *G. herbaceum* to *G. hirsutum* A_T, top right. Alignment of *G. hirsutum* D_T, to *G. raimondii,* bottom left. Alignment of *G. herbaceum* to *G. hirsutum* D_T, bottom right.
.....44

Figure 9: Minimap2 sequence to sequence alignments. Alignment of *G. herbaceum* to *G. arboreum*, left. Alignment of *G. arboreum* to *G. hirsutum* A_T, right.
.....45

LIST OF TABLES

Table 1: Assembly Statistics of the draft genome assembly at various stages along the assembly process. Statistics assessed using GAEMER basic statistics.
.....46

Table 2: Bionano Physical Map Statistics
.....47

Table 3: BUSCO Statistics
.....47

Table 4: Total Manual Edits. Edits made to PG scaffolded pseudomolecules during manual integration of Bionano and Hi-C data. Total edits made per chromosome listed as well as total contigs previously unscaffolded, which were incorporated into psuedomolecules via manual integration.
.....48

Table 5: Effects of Manual Integration on Pseudomolecules. Total size changes effected by manual scaffolding of previously unscaffolded contigs. A total of 36 contigs were scaffolded into the pseudomolecuels.
.....48

The Genome Sequence of *Gossypium herbaceum* (A_1), a Domesticated Diploid Cotton

Alex J Freeman[a], Joshua Udall, Craig Coleman[a], Peter Maughan[a], John Kauwe[a],
[a]Department of Wildlife Sciences, Brigham Young University, Provo, UT
National Science Foundation

INTRODUCTION

Cotton is an economically essential international crop worldwide, with over 12.6 million acres being utilized for fiber and cottonseed production in the United States alone [1]. The genus originated from a paleo-hexaploid (n=13) and has diversified into eight sub-genomes ranging from A through G, and K, totaling over 45 diploid and 7 tetraploid species [2][3]. Genome sizes range from approximately 880 Mb in the D genome species to 2,500 Mb in the K genome species [1][3][5]. The African native A genome species diverged from the Mexican native D genome species approximately 5~10 million years ago (MYA). Between 1~2 MYA these species formed an interspecific hybrid which led to the generation of the AD genome tetraploid. The major cotton fiber producing species is the tetraploid *G. hirsutum* (AD_1) [6][7], and a small amount of tetraploid *G. barbadense* (AD_2), known as Pima cotton [7], is also cultivated for cotton fiber production [7]. *G. herbaceum*, levant cotton [20], or African cotton is still a locally cultivated A-genome species and produces a small percentage of cotton tonnage in arid regions of India. In addition to fiber production, seeds of diploid and tetraploid cotton are also used for cottonseed oil production, and the husk and kernel of processed seeds are used as meal for livestock [12][14]. The two species whose seeds are most used to produce cottonseed oil are *G. hirsutum* and *G. herbaceum.*

The existence of extant diploid A- and D-genome species and extant tetraploid AD-genome species provides an excellent opportunity for studying polyploidization and genome evolution [8][9][3][10] and how polyploidization can lead to increased expression of desirable agronomic traits, as evidenced by tetraploid cotton [7]. The genome of tetraploid cotton (*G. hrisutum* and *G. barbadense*) has two subgenomes of 13 chromosomes (A_T and D_T, where 'D_T' refers to the D-genome in the tetraploid nucleus). The diploid genome of *G. raimondii* is more closely related to the D_T genome of the tetraploids than any of the other D-genome species. A high-quality reference genome of this D-genome species has been published [11] and used as a proxy reference for several studies of the evolutionary history of tetraploid cotton [12][13]. The diploid A-genome of *G. herbaceum* (A_1) is arguably more closely related to the A_T subgenome of the tetraploids than the other A-genome species *G. arboreum.* However, controversy remains today regarding which A-genome species is most closely related to the A_T, with proponents supporting both *G. arboreum* [14][15] and *G. herbaceum* [16][17][18][19]. Some recent studies suggest that both A-genome species are equally divergent from the A_T [20][21]. A draft sequence assembly has been published of the diploid A-genome species *G. arboreum* [14]. Although the general academic consensus suggests that *G. herbaceum* is the closest related diploid A-genome species to the A_T, a genome sequence has not yet been published.

A high-quality genome sequence of *G. herbaceum* is necessary to better study how structural variations affect genome evolution after polyploidization, using the A- and D-genome cottons in comparison to the AD tetraploids, and allows an investigation of domestication of the A-genome. In this study, we report a genome sequence assembly of A_1-genome species cotton which can be used to further evolutionary comparative analysis research, cotton research, and cottonseed oil research worldwide. A combination of Pacific Biosciences long read data, Dovetail genomics Hi-C scaffolding data, Phase Genomics Hi-C scaffolding data, and Bionano

Genomics physical mapping data was used in the assembly process. This combination of data produced a validated assembly that can contribute to comparative genomic analysis and demonstrated the newer scaffolding technologies of Phase Genomics, Dovetail genomics and BioNano Genomics.

METHODS & MATERIALS

Sample Collection, DNA Isolation & Library Preparation

Plant tissue for *G. herbaceum,* accession Wagad, was grown at Brigham Young University Greenhouse. Young tissues were collected and DNA was extracted through the CTAB method [22]. In our use of this method, tissue was lyophilized, ground in liquid nitrogen to disrupt membranes, resuspended in buffer and incubated with a lysis solution at 60^{o} C for 30 minutes, treated with RNAse, treated with chloroform to separate DNA from insoluble particles, precipitated for removal of salts and rehydrated in TE buffer. DNA was then shipped to the National Center for Genome Resources and NovaGene for library preparation and sequencing.

Genome Sequence Data Generation

Pacific Biosciences RSII Sequencing systems were used to generate ~60x PacBio long read data. The PacBio system uses a technique called single molecule sequencing to "read" pulses of light as individual fluorescent nucleotides are incorporated onto the DNA strand [23]. This reaction occurs at real time on a SMRT chip using proprietary polymerases and chemistry. P6/C4 chemistry (polymerase generation 6, chemistry generation 4) was used to collect our PacBio data [24] (Base Pairs: 99,104,937,685; Read Length N50 13.2k; Mean Read Length

8.87k). Libraries were generated using PacBio's standard protocol including BluePippin™ size-selection. Illumina mate-pair libraries were also created for fragment lengths of 180bp, 4kbp, and 9kbp. In addition, ~133x Illumina sequencing data was also generated.

Genome Assembly

The CANU2 assembler was run with map sensitivity set to normal, a minimum read length of 2000, and a minimum overlap of 800. This resulted in an assembly with a contig N50 of 315kb and 9280 contigs. Fresh tissue and the assembly were sent to Phase Genomics (PG) where they generated 13X Illumina sequencing data and used their patent-pending PROXIMO high-throughput chromosome conformation capture (HI-C) technology to organize the contigs into pseudomolecules. The pseudomolecules represent the 13 chromosomes of cotton with contigs ordered according to the highest likelihood of where each contig should be placed. In summary, the data collection and scaffolding process consisted of cells being fixed with formaldehyde, and cell membranes being disrupted. Fixed DNA was then digested with HINDIII. Sticky ends are biotinylated and proximity ligated, forming chimeric reads. These chimeras are enriched and the DNA is sheared at 300-500 bp. Libraries are generated using the chimeric long-distance interacting molecules and sequenced. Reads are mapped back to assembly contigs, and the frequency of interactions between individual contigs is used to generate a log link likelihood. The log link likelihood indicates how proximal or distant contigs are in relation to each other contig, and is used to scaffold contigs into pseudomolecules. Scaffolding assembly contigs using PG Hi-C proximity data resulted in a scaffold number reduction from 9,280 to 1,086, with the scaffold N50 increasing to 126 Mb.

Bionano Physical Map Generation and Integration

A *G. herbaceum* plant was dark treated for at least 18 hours prior to tissue collection for Bionano high molecular weight (HWM) DNA extraction. 0.5 grams of the youngest tissue was harvested and subsequently fixed with 2% formaldehyde for 20 minutes. The tissue was washed for 30 minutes, with 3 10-minute washes. Tissue was then blended with the Qiagen tissueruptor [25] 5 times at intervals of 30 seconds. The lysed leaf tissue was filtered with a 100 micron filter, and again with a 40 micron filter. The nuclei-leaf debris was then taken through a series of centrifugation steps with proprietary Bionano buffers [26] to isolate pure nuclei from the sample. Nuclei were then embedded in 2% low melting agarose and lysed. During this step, proteinase K was also used to remove unwanted proteins. Agarose plugs containing raw HMW DNA were then treated with RNAse. The plugs were then treated with agarase to free the DNA from the plugs. Raw DNA was placed on Millipore filters floated on top of pH 8 TE buffer to remove free floating agar molecules. DNA quantity was measured with the Qubit 2.0 fluorometer.

DNA aliquots of sufficient concentration were then processed with Bionano Genomics' Nick Label Repair Stain (NLRS) protocol, generating DNA which was ready to load onto an IRYS v2 chip for imaging. The NLRS protocol consist of nicking DNA with a modified restriction endonuclease, which only cuts one strand of the DNA instead of both. DNA is then treated with green fluorescent dideoxy ribonucleotides and TAQ polymerase. A random quantity of base pairs are removed and replaced by the TAQ polymerase. The incorporation of green fluorescent nucleotides creates labels on the DNA molecules which can be imaged. The DNA strands are then treated with DNA ligase, which repairs the remaining nicks on the DNA strands, and then the reaction is quenched and treated with DNA stain to counterstain the backbone of the molecule blue. Labeled and stained DNA is then ready to be loaded onto an IRYS V2 chip for

imaging. The imaging process consists of DNA passing through pillars and into microchannels to linearize it in preparation for entry into nanochannels where it will be imaged. The DNA is electrophoresed into the nanochannels and immobilized with equal and opposite current from both the "forward" and "backwards" directions. A laser is used to excite the fluorescent molecules of the labels and stains and images are taken. The contrasting blue molecules with green labels are what the IRYS software detects during image processing post data collection.

A total of 140X coverage of *G. herbaceum,* Wagad accession, was collected through this process, and a total of 4 IRYS chips were used to collect the data. The data was then assembled using Bionano Solve. At the start of this process raw molecules were filtered based on length, with molecules shorter than 100 kb and longer than 500 kb being excluded from the assembly. The final Bionano assembly had a total size of 1566 Mb, which is 93.9% of the estimated 1667 Mb. It had a total of 1838 Bionano contigs with an N50 of 1.20 Mb. The Bionano assembly was aligned to the CANU-PG-PBJELLY sequence using Bionano software. This Bionano to sequence assembly comparison aligned 89% of the Bionano maps to the sequence assembly.

The Bionano physical map was then integrated with the PG scaffolds using a manual visual-inspection based approach. Hybrid Scaffold, Bionano Genomics' map-assembly integration software, yielded unsatisfactory results. When the Hybrid Scaffold results were analyzed with Bionano Access, a web-based browser used to visualize the alignments between the Phase Genomics generated pseudomolecules and the Bionano map, many regions were identified that were not corrected by Hybrid Scaffold which had sufficient evidence to merit correction. As such it was deemed necessary to perform a manual integration of the two data sets. There were a total of 934 edits deemed necessary to effect in the PG pseudomolecules, based on the sequence alignment to the Bionano map.

The first class of edits effected was orientation correction. As seen in figure 2, many small contigs in the PG pseudomolecules were identified that needed to be inverted, and many small groups of contigs as seen in figure 1, that needed to be inverted as blocks. This indicates Hi-C was accurately able to locate where these contigs belonged, but unable to correctly assess their orientation. The orientation corrections involving one single contig were simple to make and consisted of changing the orientation score in the third column of the PG generated group ordering files (Figure 3). Orientation corrections involving two contigs or more were slightly more complicated to effect, as it included changing the order of the contigs involved, and then changing the orientation score for each contig (Figures 1). These edits were made on a visual inspection basis, and if the Bionano-PG pseudomolecule alignment indicated an orientation change was necessary, it was performed. These orientation corrections used the Bionano contigs as a method of correcting PG scaffolding and, therefore, it was assumed that Bionano was more correct in near every instance. However, when a Bionano contig had very little alignment to the PG sequence assembly, it was assumed that the Bionano contig was an erroneous or chimeric contig and was thrown out. In total, 374 contig orientation changes in the PG scaffolds.

The next two classes of edits effected included moving contigs from one location to another in the pseudomolecule, and if necessary changing the orientation of the contig. These edits were broken up into two classes. This is because the Bionano-PG pseudomolecule alignment frequently identified contigs that should be moved only short distances and less frequently identified contigs that needed to be moved longer distances. It was decided to make these two types of edits distinct as a measure to indicate how accurately PG was able to place contigs.

As PG uses log-likelihood ranks based on contig-contig interaction frequencies to order contigs into pseudomolecules it was decided decided to measure how far each contig was moved in relation to total contigs per chromosome, instead, of a physical base pair difference. This more accurately represents how far contigs were moved and how accurately PG was able to place contigs than a base pair scale, as some regions of the pseudomolecules contain less than a few million base pairs but hundreds of contigs. Contigs that were moved "short" distances are classified as having moved less than 10% of the total amount of contigs in that pseudomolecule. Any contigs that were moved more than 10% of the total amount of contigs in their pseudomolecule are identified as having moved long distances. When determining if a contig should be moved or not, the Bionano-PG pseudomolecule alignment and the PG log likelihood score were assessed to determine if an edit should be made. When a Bionano contig indicated a contig needed to be moved, the fourth column of the group ordering file would be used (Figure 3) to determine if the contig should be moved. As this is a novel approach to genome scaffolding, it was decided to use a value of 60 to determine if a move should be made, meaning that if the log likelihood was less than 60 the Bionano contig was trusted to be more correct and a change was effected, and if the log likelihood was greater than 60 the PG contig placement would be trusted and no change made. A total of 407 short distance edits were made and 117 long distance corrections were made.

The fourth class of edits involved scaffolding contigs that PG did not place into the pseudomolecules, which Bionano alignment evidence could accurately place. After PG scaffolding there were 1073 small contigs remaining unscaffolded. Some of these contigs had Bionano-sequence alignment. It was found that the Bionano contigs mapping to the pseudomolecules frequently had small gaps of alignment, indicating the PG pseudomolecules were missing sequence data. By comparing the Bionano contigs aligning to small unscaffolded

sequence contigs, contigs were identified which had sufficient Bionano alignment to warrant manually scaffolding these contigs into the PG pseudomolecules. A total of 36 previously unscaffolded contigs were scaffolded in this manner.

RESULTS

Genome Sequence Assembly

The CANU [27] assembler was used to generate a *de novo* PacBio assembly of *G. herbaceum*, Wagad accession, using 60X PacBio long read coverage. The assembly had a total length of 1.6 Gb, including gaps (95% expected size), with a scaffold N50 of 315 kb (Table 1). The scaffold count totaled 9,280. To further improve genome scaffolding quality ~13X Seq. Coverage of PG Proximo^{TM} Hi-C data was also generated [28]. Of course, incorporation of PG Hi-C data yielded significant improvements to the genome scaffold N50, and the majority of contigs were arranged into 13 pseudomolecules. Total size remained the same, but the scaffold number was reduced from 9,280 to 1,086 and scaffold N50 increased to 126 Mb. To represent the genome assembly, a pairwise heatmap constructed from the scaffolded log-likelihoods illustrates chromosome contiguity and repeats that are likely telomeres and centromeres (Figure 5). After PG incorporation gaps were filled with PBJELLY2, greatly improving contig metrics. The number of contigs were reduced from 9,280 to 5,484, doubling the contig N50 from 315 kb to 685 kb. It also reduced the number of scaffolds to 1,058 and increased their N50 to 129 Mb. Gap filling increased total assembly length to 1.6 Gb (approximately 97% of the estimated 1.7 Gb). After gap-filling with PacBio reads, the assembly was corrected with Illumina short reads using PILON [29] to correct for base errors. (Figure 4).

Bionano Genome Assembly

A Bionano physical map assembly was generated using Bionano IRYS. The assembly had a total length of 1.6 Gb without gaps (95% expected size), included 1,842 individual Bionano contigs, and an N50 of 1.195 Mb (Table 2). The overall alignment rate between the Bionano contigs and the scaffolded sequences was 89%. Bionano's Hybrid Scaffold software was run to integrate the Bionano assembly and the PG assembly. Hybrid Scaffold was unable to correct the orientation of small contigs PG had placed correctly with incorrect orientation. As such we decided to manually integrate the two assemblies. The improvements yielded by the hybrid scaffolding attempt were insufficient to warrant progressing with the hybrid scaffolded assembly, as it would have added a layer of complexity into the manual integration of the two assemblies, which was more promising than the Hybrid Scaffold results. The manual integration yielded significant improvements to the genome sequence assembly in terms of correcting the orientation of contigs (Figure 2).

During assembly of Bionano molecules, we identified a striking repetitive pattern of nick sites that spanned 50,000 – 150,000 kb depending on the Bionano contig (Figure 6). The main repeat consisted of three BssSI nick sites, approximately 5000 bps in width that repeat 10-30 times at one location on each of the thirteen chromosomes. We performed a variety of tests to better characterize these genomic regions. Sequence contigs of the repetitive Bionano regions were searched for genes or other matching annotated sequence patterns. Blast results indicated that this region had no known genes or gene families, ruling out a variety of possibilities from nucleolus organizer regions (NOR) to high repeat gene families. Depth of coverage analysis, which used minimap2 to align raw PacBio reads to individual chromosomes, showed spikes of coverage in areas not coincident with our Bionano nick repeat.

We also mapped these motif repeat containing contigs to all available PacBio *Gossypium* sequence assemblies and identified the nick site repeat in every sequence assembly. Additionally, and of greater importance, when Bionano maps containing the nick repeat was mapped to the PacBio assembly of tropical durian fruit, *Durio zibethinus* [30], a close relative of *Gossypium,* we also identified sequence contigs containing the repetitive nick sites similarly spaced to those we found in A1. This is quite remarkable as it indicates the motif repeat has been conserved in species separated by 60-77 million years. After comparing the Bionano nick repeat to *Theobroma cacao,* the next closest relative of the cotton-durian fruit ancestry, we were unable to identify the motif repeat in the sequence assembly. After performing the test in *T. cacao,* we additionally performed the same comparison with *A. thaliana,* and *Brassica juncea* cultivar tumida. These species are some of the closest related angiosperms with sequence assemblies incorporating PacBio long read single molecule data. The Bionano nick site repeat was absent in both of these species [31].

Manual Integration of Bionano and Hi-C

Using Bionano Access to visualize Bionano alignments, and a PG generated .bed file indicating start and stop locations of each contig and gap in the superscaffolded pseudomolecules, we manually corrected the order and orientation of many contigs incorrectly scaffolded by PG (Figure 2). We also scaffolded 36 previously unscaffolded contigs into the PG Hi-C pseudomolecules, decreasing the total contig number and increasing the scaffold N50 (Table 4). This manual integration of the Bionano assembly and the Hi-C generated pseudomolecules yielded additional improvements to the overall genome assembly. The total size increase of the pseudomolecules was 7.29 Mb (Table 5), and the percentage of Bionano

contigs aligning to the CANU-PG-PBJELLY-BNG sequence assembly increased from 89.0% to 90.3%. Manual integration of the Bionano contigs with the PG pseudomolecules represented the final adjustments to the nucleotide positions prior to genome annotation.

Genome Annotation

The genome sequence annotation server (GenSAS) [32][33] was used to annotate the A_1 genome sequence. First, repeats were identified and masked within GenSAS using Repeat Masker and Repeat Modeler. Repeat regions were found to comprise 76% of the draft genome. There were a very large quantity of unknown repeats, totaling 60% of the *G. herbaceum* draft genome repeats, followed by Gypsy repeats (29%) and then Copia (4.1%). Many other repeat families and classes were also identified with relatively low frequency (Figure 7). These results are in general accordance with other repeat distributions of the Malvaceae family [30]. When compared to the A_2 genome cotton species, the *G. herbacuem* genome has undergone repeat deletion in both the Gypsy and Copia classes, as *G. arboreum* was reported to contain 55.8% Gypsy and 5.5% Copia. The large discrepancy between the closely related cotton species could be due to true evolutionary divergence, repeat misidentifications, or misassembly of repeat regions in either assembly. Both Repeat Masker and Repeat Modeler predicted a large quantity of unknown repeats. Repeat Modeler predicted 60% "Unknown" repeats, and Repeat Masker predicted 62% "Simple" repeats. These two prediction algorithms each predicted a relative abundance of Gypsy and Copia repeat elements. This indicates that the repetitive elements in the assembly were accurately identified, though many of them may not have been labeled. It is possible that *G. herbaceum* contains many repeat classes not yet named, or that Repeat Modeler and Repeat Masker were too conservative with labeling repeats.

Second, GenSAS was used to run Augustus [34], BLAT [35], GeneMarkES [36], Genscan [37], GlimmerM [38], PASA [39], and SNAP [40] for gene prediction. In addition, an independent annotation effort was performed using Maker [41]. All of the programs run by GenSAS, excluding PASA, were unable to accurately predict genes in the sequence assembly. Subsequently, it was determined to use the Maker annotation by itself, and then incorporate PASA in a later refinement step. The final gene predictions included 28,273 genes, 39,518 mRNA sequences, 244,936 exons and 227,530 coding sequences. Coding sequences are defined by PASA as the altering of protein coding sequences which lead to untranslated regions of exons.

Third, GenSAS was used to run BlastP [42], BlastP with SwissProt [42][43], InterProScan [44], Pfam [45], SignalP [46], and TargetP [47] for functional annotation. 24,775 genes were annotated by InterProScan, and 22,279 (89.9%) were identified and named. BlastsP against the SwissProt curated database identified and named 8,549 genes in the functional annotation.

To confirm the accuracy of our Maker/PASA genome annotation, we ran a BUSCO analysis [48]. BUSCO is an independent analysis of genome assembly, gene space, and transcriptome completeness. It uses a set of genes under single-copy selection pressure as a standard against which new genome sequence assemblies can be measured. The absence of many genes in a genome sequence assembly being compared to the BUSCO standard can indicate that errors took place, either in the sequencing or assembly of that genome sequence. For our genome sequence assembly, BUSCO predicted a total of 1,336 out of the 1,440 (92.8%) highly conserved genes in the Embryophyta gene set. 1,218 (84.6%) of these were identified as complete with a single copy in the genome, 118 (8.2%) were identified as complete with multiple copies, and an additional 29 genes (2.0%) were identified as fragmented (Table 3). As the genes utilized by

BUSCO for gene space analysis are under selection pressure to maintain a single copy, the majority of genomes tested should have few duplicated genes. The high percentage of duplicated genes can indicate haplotigs that were unsuccessfully merged in the assembly process. However, due to the recent whole genome duplication in the *Gossypium* lineage, we commonly see high percentages of duplicated genes, such as the 12.2% in *G. arboreum,* the A_2 genome cotton and 11.5% in *G. raimondii*, the D5 cotton species [30]. In contrast, *Theobroma cacao*, which has not undergone a recent WGD, has a duplicated gene percentage of 1.2%. This suggests that the duplicated genes identified by BUSCO are not misassemblies but rather separate and unique copies of highly conserved genes from the Embryophta gene set.

Minimap2 Comparative Analysis

By comparing our A_1 sequence assembly using minimap2 to the A_2-, D_5-, A_T-, and D_T-genomes in the *Gossypium* genus, we assessed the overall correctness of scaffolding contigs with PG Hi-C data. Minimap2 is a versatile pairwise alignment program used to compare sequences. It can compare reads to references including PacBio, Oxford nanopore, and Illumina reads, find overlaps, assembly-to-assembly, and full-genome alignments. We used the assembly-to-assembly pairwise alignment function. Previous experiments (unpublished data) using minimap2 indicated that all genomes of the cotton genus, though separated by 5-10 MYA, are sufficiently related so that two high quality genomes of *Gossypium* should show clear synteny and colinearly along all 13 pseudomolecules, with occasional inversions and translocations if present in the genome sequence assembly (Figure 8). The comparisons between A_1 and the other species indicate that the sequence assembly contains some misassembles even though it contains a high scaffold N50 and accurate gene space annotation (Figure 8). An A_1 to D_5 comparison displays a

lack of sequence homology between chromosomes and reveals many regions which have very little homology between the two genomes. To test if the misassemblies reside within the A_1 sequence assembly or the D_5 assembly, we compared A_1 to A_T, and D to D_T. The D to D_T minimap2 alignment showed a clear homology between the two assemblies with a few regions clearly showing large scale inversions. The A_1 to A_T minimap2 alignment shows clear evidence that the two assemblies have major regions on each chromosome which do not share any homology and are vastly misassembled.

We additionally compared A_1 to A_2, and A_2 to A_T to compare genome quality and to potentially find regions where the diploid A-genome species are the same, but differ when compared to the A_T-genome. We found that the "high-quality" A_2 assembly appears to have little to no pseudomolecule homology with either the A_1 genome or the A_T sub genome. As seen in Figure 9, the minimap2 comparison shows that the 13 pseudomolecules in the A_2 assembly were almost completely randomly scaffolded. The attempt to find regions of synteny between the two diploid A-genomes and where they diverge from the A_T cannot be undertaken due to the lack of correctness in the assembly of the A_2 genome.

DISCUSSION

Polyploidization events are strong drivers of evolution and speciation [8]. After polyploidization, genes may evolve new functions or regulatory mechanisms. As selection pressure for each new gene copy is reduced, mutations may arise which can lead to repurposing of genes. These mutations occur at random and many are deleterious and selected against in successive generations. Some are beneficial and increase progeny fitness. To understand how genome polyploidization can generate new species and phenotypes, large scale comparative genomic analysis must be undertaken. Analysis regarding how structural variations affect

evolution and how polyploidization affects speciation are areas of research that have yet to be explored in depth. To undertake such an analysis sequence for every major branch of a genus would be required. By including genome sequences for multiple branches of the genus and including multiple sequences from the same branch, a study can become very robust and contribute significant knowledge of cotton genome evolution. The addition of our draft sequence to the pool of *Gossypium* genome sequences facilitates such a large-scale study of the cotton genus currently being undertaken.

Our experience with Bionano Genomics, Dovetail Genomics, and Phase Genomics contributes valuable experience with newer scaffolding technologies to the genomics community. Prior to PG scaffolding we attempted to integrate Bionano Genomics' physical maps and our contigs; however, the results were not promising. We were unable to satisfactorily scaffold the sequence contigs with Bionano physical maps due to the low contig N50 and map N50 before scaffolding with PG. After scaffolding the contigs with PG data, we successfully integrated Bionano maps with the pseudomolecules to improve the assembly quality.

These results appeared highly promising. However, once we began running minimap2 [49], we saw the sequence comparisons between our *G. herbaceum* genome sequence and other cotton species to be highly discontinuous. By comparing high quality genome sequences of *Gossypium* (unpublished data) we know that all *Gossypium* genomes have enough synteny and colinearly to generate minimap2 dot plots that appear linear with a few key inversions and/or translocations. Comparisons between *G. herbaceum* and the *G. raimondii* revealed large regions of multiple pseudomolecules that are incorrectly scaffolded. The majority of these regions are unlikely to be correct representations of structural rearrangements, though a few may be actual genome rearrangements. Comparison of the D to D_T tetraploid subgenome (Figure 8) shows

highly similar sequence homology with a few clearly identifiable regions where inversions and other structural rearrangements have taken place. We hoped to see similar results when comparing the A to A_T subgenome, however, the alignment appears drastically different. We further aligned the A-genome to the D_T subgenome for robustness and again saw large regions of the sequence assembly which are highly discontinuous.

When examining the quality of the genome using only the high scaffold N50, high BUSCO identified gene percentage and high alignment percentage between the genome sequence and the Bionano physical maps, the genome appears to be of high quality. However, the low contig N50 has a very strong negative side effect when combined with the PG scaffolding data. PG generated 13 pseudomolecules the approximate size of each of the 13 *G. herbaceum* chromosomes. The subcentromeric regions of each pseudomolecule are filled with many small contigs that PG was unable to correctly place on the pseudomolecule. Hi-C based techniques capture intrachromosomal interactions and use this information to scaffold contigs into pseudomolecules approximating the actual chromosome. The technique also captures interchromosomal interactions, which can confuse contig placement. Minimap evidence suggests that the majority of contigs are placed on the correct pseudomolecules, but in incorrect and, at times, apparently random locations.

This has two implications when considering genome assembly. First, companies such as Phase Genomics will arrange contigs into pseudomolecules according to the number of input chromosomes identified, whether or not each contig truly belongs there. Second, and more importantly, high contig N50 is crucial to receiving a good assembly from Hi-C based approaches. Hi-C techniques are based on proximity interactions between contigs. Larger contigs have a higher probability of having more interactions with neighboring contigs, which makes

them easier to correctly scaffold. Smaller contigs have a lower probability of having high frequencies of intrachromosomal interactions with nearby contigs and a reduced probability of being correctly placed in a pseudomolecule. Additionally, interchromosomal interactions may cause contigs to be placed on pseudomolecules to which they do not belong. Even if contigs are placed on the correct pseudomolecule, placement will be less precise and more guesswork as the low signal to noise ratio will decrease accuracy.

In accordance with this, having a high contig N50 is very important when sequencing a genome, as having a high contig N50 reduces the likelihood of Hi-C based approaches having any negative impact on sequence assembly quality. Frequently, the region of a pseudomolecule which displays the most misalignments is near the middle of each pseudomolecule. This indicates that PG was more successful in scaffolding telomeric and subtelomeric contigs and less successful in scaffolding subcentromeric and pericentromeric regions. The scaffolding difficulty could be due to an innate weakness in Hi-C proximity capture techniques, weaknesses in PacBio sequencing, or an increased frequency of unmerged haplotigs nearing the centromeres. Unmerged haplotigs in pericentromeric regions would decrease local contig N50, subsequently increasing the difficulty of correct contig placement.

For a future draft of this genome sequence assembly, to increase the contig N50, we sequenced additional fresh tissue of *G. herbaceum* with the new Pacific Biosciences Sequel. We generated an additional 18X coverage and are in the process of incorporating the data into the sequence assembly presented here of 60X coverage. We are currently working on refining the CANU assembly parameters and input data sets to produce a sequence assembly with a higher N50. While the N50 increased in subsequent draft assemblies incorporating this additional data, we expect the raw contig N50 can be further increased by further fine-tuning the input

parameters of the CANU assembler. We believe that the higher contig N50 in the next versions of this sequence assembly will lead to better integration of all data types and will provide a sequence assembly with more correct contig placement and increased sequence homology when compared to other species of *Gossypium*. This future draft genome will provide an even higher quality reference which can be used to further probe the relationships between evolution and polyploidy.

Though the current draft genome has regions of contig placement which are not perfectly placed within the pseudomolecules, the gene space is excellent, and the quality of the sequence assembly is still useful. This sequence assembly is a valuable tool for future research on cotton genomics and for cottonseed oil production research. The discovery of a Bionano nick site repeat sequence previously unidentified is of particular interest. We were unable to find the repeat in *T. cacao*. There are two possible explanations for this. First, the repeat is not found in *T. cacao*, and the repeat originated sometime after the *T. cacao – Gossypium/Durio* divergence, potentially before the whole genome duplication event that marks the *Gossypium* and *Durio* divergence [30]. Second, the *T. cacao* genome assembly was unable to capture the motif repeat sequence due to the "short" read length of the data used, as it was generated with a combination of Sanger, Roche 454 pyrosequencing, and Illumina read pairs [50].This repeat was identified in *D. zibethinus,* which diverged from the *Gossypium* genus over 60 MYA and indicates that this region is of some importance for an as of yet unidentified characteristic of plant physiology. Though the depth of coverage analysis suggests this region is not centromeric, testing with FISH probes designed from the putative repeat region combined with CEN FISH probes would conclusively confirm or reject the hypothesis that this region, and these unique repeats, are centromeric.

The fact that these repeats were only identifiable with Bionano physical maps additionally contributes to the scientific community. Genomic maps have been used for decades, mainly with the intent of facilitating genome assembly and BAC placement. More recently, with the advent of optical mapping technologies such as Bionano Genomics and Nabsys [51], genomic maps have been utilized for detecting structural variants and identifying how they contribute to a variety of human diseases [52][53][54]. Here we have identified motif repeats which are invisible to traditional repeat identification software but are identified by physical mapping. Although we are currently unable to elucidate the purpose of the motif repeat, it has now been identified. This motif repeat can be a study focus for future research and additional studies can be undertaken to see if similar repeats can be found in other eukaryotic genomes.

It is possible that these repeats are Bionano artifacts. However, most Bionano artifacts are generated by DNA molecules getting "stuck" in the IRYS nanochannels. In successive rounds of DNA imaging, more and more strands of DNA are pulled into the nanochannels. If stuck molecules are present, the new strands are pulled onto the end of the stuck DNA molecules, creating *in vitro* very long DNA chimeric molecules which are imaged repeatedly, generating map artifacts. These strands appear correct to the assembly programs and link together different sections of the genome which may or may not be close to one another. This phenomenon is easy to overcome. During the assembly process, small molecules are routinely eliminated from the data. This is done for all Bionano physical map assemblies. If the line of code which selects molecules longer than length X is copied and modified, it is possible to then eliminate any molecules above a certain threshold as well. We have found that by removing any molecules longer than 500kb we drastically decrease total map number and increase map N50. We suspect that DNA molecules above this threshold are primarily molecules which are chimeras and only

hinder map assembly. As we have eliminated any molecules below 100kb and above 500kb, we successfully removed any Bionano map artifacts. Due to this, we believe that the motif repeat pattern of Bionano nick sites identified is indeed a real repeat motif in the DNA sequence.

AN OVERVIEW OF GENOMICS

Genomics is a relatively young scientific field, originating in the 1990's with the human genome project [55]. With the advent of sanger sequencing in 1977 [56], "next generation sequencing" platforms in the early 2000's [57], and modern long-read sequencing technologies such as PacBio [58] and Oxford Nanopore [59] the genomics field is rapidly progressing towards sequencing hundreds and even thousands of complete genomes. It was realized early on that sequence data alone is not sufficient to assemble a genome sequence assembly. There are many regions in each genome that are too complex to assemble using even today's impressive PacBio long-read data. These regions are primarily composed of repeats and include centromeric repeats, telomeric repeats, transposable elements such as gypsy and copia repeats, and many more repeat classes both classified and unclassified [60]. To be able to generate an assembly that is as correct as possible, while understanding that current technologies will not be able to resolve some of the more complex repeat regions, additional technologies were developed to correctly order, orient, and scaffold contigs. These include genetic mapping technologies, physical mapping technologies, and more recently Hi-C based interaction technologies.

Traditional genetic mapping, or linkage mapping technologies were among one of the first techniques used to scaffold sequence contigs. This technique consists of identifying many unique patterns in the DNA sequence and identifying which chromosome the sequence comes from. Once that information is obtained, these genetic maps can be used to link individual

contigs to chromosomes [61]. Another technology which was used early on in modern day genetics was physical mapping. In the beginning, this consisted of time consuming and expensive bacterial artificial chromosomes (BACs) [62]. More recently, these techniques have been replaced by more efficient and cost effective optical mapping technologies, such as Bionano Genomics [63]. In addition to improved physical mapping technology, Hi-C based scaffolding approaches have recently further increased the ability of genomicists to more correctly assembly a genome sequence. Hi-C based techniques utilize innate chromosome organization and DNA-DNA interactions by cross-linking DNA that is interacting in vivo. These locations are then sequenced, and data can be generated by measure of how many interactions a contig has with other contigs, indicating where contigs belong on a scaffold [64].

Combinations of these technologies are leading to increasingly high-quality genome sequence assemblies. Today, genomes are being sequenced with contig N50s reaching into the Mb scale, with scaffold N50s approaching chromosome level lengths [10][11]. Additionally, the long read length of PacBio and Oxford nanopore technologies are allowing for unparalleled resolution and characterization of complex repeat regions which have hindered forward progress of genomics in the past.

Multiple high-quality genome sequence assemblies have recently been published using PacBio long read sequences with some form of Hi-C data to scaffold contigs, and occasionally additional scaffolding technology such as Bionano Genomics physical mapping. These assemblies have chromosome length scaffold N50s and N90s, showcasing the incredible capacity of PacBio, Hi-C based approaches, and physical mapping [64][65][30]. The future of the genomics field continues to shift in this direction, meaning genomes are being sequenced with

increasing frequency and incorporating long read sequence data with at least one form of scaffolding data.

Genome assembly quality will continue to increase as technologies improve and assembly and scaffolding algorithms are further refined. The applications of the genomics field are rapidly expanding. In addition to sequencing genomes for studies on evolution, speciation, crop improvement, or livestock improvement, scientists and medical professionals are continuing to rapidly expand genomics into increasingly more important studies of disease and personalized medicine. One emerging trend in the future of genomics is personalized medicine. The cost of DNA sequencing has consistently plummeted since it's invention. Where the original human genome cost 3 billion dollars to sequence, it now costs roughly $1000. This cost is not for a *de novo* genome sequence with PacBio and Hi-C, but rather an Illumina-only based assembly, which uses one of the many high-quality reference genome sequences of *Homo sapiens* to ensure correct assembly. As the cost of genome sequencing has been so reduced, it is increasingly common for individuals to have their genome sequenced for medical purposes. Despite the cost reduction of genome sequencing, there are still very few tools that can effectively use personal genomic data. The future of genomics will rely heavily on the development of novel software which can utilize the data we can currently collect much faster than we can analyze.

LITERATURE CITED

1. Meyers, Leslie: USDA, Economics, S. and M. I. S. “Cotton and Wool Outlook” (2018): Available at http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1281

2. Wendel, J., Brubaker, C., and Seelanan, T. “The Origin and Evolution of Gossypium” *Physiology of Cotton* (2010): 1–18. doi:10.1007/978-90-481-3195-2_1, Available at http://www.springerlink.com/index/10.1007/978-90-481-3195-2_1

3. Page, J. T., Huynh, M. D., Liechty, Z. S., Grupp, K., Stelly, D., Hulse, A. M., Ashrafi, H., Deynze, A. Van, Wendel, J. F., and Udall, J. A. “Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-Sequencing” *G3 (Bethesda)* 3, no. 10 (2013): 1809–1818. doi:10.1534/g3.113.007229, Available at http://g3journal.org/lookup/doi/10.1534/g3.113.007229

4. Grover, C. E., Gallagher, J. P., Jareczek, J. J., Page, J. T., Udall, J. A., Gore, M. A., and Wendel, J. F. “Re-Evaluating the Phylogeny of Allopolyploid Gossypium L.” *Molecular Phylogenetics and Evolution* 92, (2015): 45–52. doi:10.1016/j.ympev.2015.05.023

5. Gallagher, J. P., Grover, C. E., Rex, K., Moran, M., and Wendel, J. F. “A New Species of Cotton from Wake Atoll, *Gossypium Stephensii* (Malvaceae)” *Systematic Botany* 42, no. 1 (2017): 115–123. doi:10.1600/036364417X694593, Available at http://www.ingentaconnect.com/content/10.1600/036364417X694593

6. Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., Zhang, J., Saski, C. A., Scheffler, B. E., Stelly, D. M., Hulse-Kemp, A. M., Wan, Q., Liu, B., Liu, C., Wang, S., Pan, M., Wang, Y., Wang, D., Ye, W., Chang, L., Zhang, W., Song, Q., Kirkbride, R. C., Chen, X., Dennis, E.,

Llewellyn, D. J., Peterson, D. G., Thaxton, P., Jones, D. C., Wang, Q., Xu, X., Zhang, H., Wu, H., Zhou, L., Mei, G., Chen, S., Tian, Y., Xiang, D., Li, X., Ding, J., Zuo, Q., Tao, L., Liu, Y., Li, J., Lin, Y., Hui, Y., Cao, Z., Cai, C., Zhu, X., Jiang, Z., Zhou, B., Guo, W., Li, R., and Chen, Z. J. "Sequencing of Allotetraploid Cotton (Gossypium Hirsutum L. Acc. TM-1) Provides a Resource for Fiber Improvement." *Nature biotechnology* 33, no. 5 (2015): 531–537. doi:10.1038/nbt.3207, Available at http://www.nature.com/doifinder/10.1038/nbt.3207

7. Yuan, D., Tang, Z., Wang, M., Gao, W., Tu, L., Jin, X., Chen, L., He, Y., Zhang, L., Zhu, L., Li, Y., Liang, Q., Lin, Z., Yang, X., Liu, N., Jin, S., Lei, Y., Ding, Y., Li, G., Ruan, X., Ruan, Y., and Zhang, X. "The Genome Sequence of Sea-Island Cotton (Gossypium Barbadense) Provides Insights into the Allopolyploidization and Development of Superior Spinnable Fibres" *Sci Rep* 5, no. October (2015): 17662. doi:10.1038/srep17662, Available at http://www.nature.com/articles/srep17662

8. Otto, S. P. "The Evolutionary Consequences of Polyploidy." *Cell* 131, no. 3 (2007): 452–62. doi:10.1016/j.cell.2007.10.022, Available at http://www.ncbi.nlm.nih.gov/pubmed/17981114

9. Salman-Minkov, A., Sabath, N., and Mayrose, I. "Whole-Genome Duplication as a Key Factor in Crop Domestication" *Nature Plants* 2, no. 8 (2016): 16115. doi:10.1038/nplants.2016.115, Available at http://www.nature.com/articles/nplants2016115

10. Desai, A., Chee, P. W., Rong, J., May, O. L., and Paterson, A. H. "Chromosome Structural Changes in Diploid and Tetraploid A Genomes of *Gossypium*" *Genome* 49, no. 4 (2006): 336–345. doi:10.1139/g05-116, Available at http://www.nrcresearchpress.com/doi/10.1139/g05-116

11. Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., Yue, Z., Cong, L., Shang, H., Zhu, S., Zou, C., Li, Q., Yuan, Y., Lu, C., Wei, H., Gou, C., Zheng, Z., Yin, Y., Zhang, X., Liu, K.,

Wang, B., Song, C., Shi, N., Kohel, R. J., Percy, R. G., Yu, J. Z., Zhu, Y.-X., Wang, J., and Yu, S. “The Draft Genome of a Diploid Cotton Gossypium Raimondii” *Nature Genetics* 44, no. 10 (2012): 1098–1103. doi:10.1038/ng.2371, Available at http://www.nature.com/doifinder/10.1038/ng.2371

12. Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., Llewellyn, D., Showmaker, K. C., Shu, S., Udall, J., Yoo, M., Byers, R., Chen, W., Doron-Faigenboim, A., Duke, M. V., Gong, L., Grimwood, J., Grover, C., Grupp, K., Hu, G., Lee, T., Li, J., Lin, L., Liu, T., Marler, B. S., Page, J. T., Roberts, A. W., Romanel, E., Sanders, W. S., Szadkowski, E., Tan, X., Tang, H., Xu, C., Wang, J., Wang, Z., Zhang, D., Zhang, L., Ashrafi, H., Bedon, F., Bowers, J. E., Brubaker, C. L., Chee, P. W., Das, S., Gingle, A. R., Haigler, C. H., Harker, D., Hoffmann, L. V., Hovav, R., Jones, D. C., Lemke, C., Mansoor, S., Rahman, M. ur, Rainville, L. N., Rambani, A., Reddy, U. K., Rong, J., Saranga, Y., Scheffler, B. E., Scheffler, J. A., Stelly, D. M., Triplett, B. A., Deynze, A. Van, Vaslin, M. F. S., Waghmare, V. N., Walford, S. A., Wright, R. J., Zaki, E. A., Zhang, T., Dennis, E. S., Mayer, K. F. X., Peterson, D. G., Rokhsar, D. S., Wang, X., and Schmutz, J. “Repeated Polyploidization of Gossypium Genomes and the Evolution of Spinnable Cotton Fibres” *Nature* 492, no. 7429 (2012): 423–427. doi:10.1038/nature11798, Available at http://www.nature.com/articles/nature11798

13. Liu, X., Zhao, B., Zheng, H.-J., Hu, Y., Lu, G., Yang, C.-Q., Chen, J.-D., Chen, J.-J., Chen, D.-Y., Zhang, L., Zhou, Y., Wang, L.-J., Guo, W.-Z., Bai, Y.-L., Ruan, J.-X., Shangguan, X.-X., Mao, Y.-B., Shan, C.-M., Jiang, J.-P., Zhu, Y.-Q., Jin, L., Kang, H., Chen, S.-T., He, X.-L., Wang, R., Wang, Y.-Z., Chen, J., Wang, L.-J., Yu, S.-T., Wang, B.-Y., Wei, J., Song, S.-C., Lu, X.-Y., Gao, Z.-C., Gu, W.-Y., Deng, X., Ma, D., Wang, S., Liang, W.-H., Fang, L., Cai, C.-P., Zhu, X.-F., Zhou, B.-L., Jeffrey Chen, Z., Xu, S.-H., Zhang, Y.-G., Wang, S.-Y., Zhang, T.-Z.,

Zhao, G.-P., and Chen, X.-Y. “Gossypium Barbadense Genome Sequence Provides Insight into the Evolution of Extra-Long Staple Fiber and Specialized Metabolites.” *Scientific reports* 5, (2015): 14139. doi:10.1038/srep14139, Available at http://www.ncbi.nlm.nih.gov/pubmed/26420475

14. Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., Li, Q., Ma, Z., Lu, C., Zou, C., Chen, W., Liang, X., Shang, H., Liu, W., Shi, C., Xiao, G., Gou, C., Ye, W., Xu, X., Zhang, X., Wei, H., Li, Z., Zhang, G., Wang, J., Liu, K., Kohel, R. J., Percy, R. G., Yu, J. Z., Zhu, Y.-X., Wang, J., and Yu, S. “Genome Sequence of the Cultivated Cotton Gossypium Arboreum” *Nature Genetics* 46, no. 6 (2014): 567–572. doi:10.1038/ng.2987, Available at http://www.nature.com/doifinder/10.1038/ng.2987

15. Lu, C., Zou, C., Zhang, Y., Yu, D., Cheng, H., Jiang, P., Yang, W., Wang, Q., Feng, X., Prosper, M. A., Guo, X., and Song, G. “Development of Chromosome-Specific Markers with High Polymorphism for Allotetraploid Cotton Based on Genome-Wide Characterization of Simple Sequence Repeats in Diploid Cottons (Gossypium Arboreum L. and Gossypium Raimondii Ulbrich).” *BMC genomics* 16, no. 1 (2015): 55. doi:10.1186/s12864-015-1265-2, Available at http://www.ncbi.nlm.nih.gov/pubmed/25652321

16. Kulkarni, V. N., Khadi, B. M., Maralappanavar, M. S., Deshapande, L. A., and Narayanan, S. S. “The Worldwide Gene Pools of Gossypium Arboreum L. and G. Herbaceum L., and Their Improvement” *Genetics and Genomics of Cotton* (2009): 69–97. doi:10.1007/978-0-387-70810-2_4, Available at http://link.springer.com/10.1007/978-0-387-70810-2_4

17. Wendel, Jonathan F. Chronn, R. C. “Polyploidy and the Evolutionary History of Cotton” *Advances in Agronomy* 78, no. 139 (2003): 139–186. Available at https://lib.dr.iastate.edu/bot_pubs/?utm_source=lib.dr.iastate.edu%2Fbot_pubs%2F23&utm_med

ium=PDF&utm_campaign=PDFCoverPages

18. Cui, X., Liu, F., Liu, Y., Zhou, Z., Zhao, Y., Wang, C., Wang, X., Cai, X., Wang, Y., Meng, F., Peng, R., and Wang, K. "Construction of Cytogenetic Map of Gossypium Herbaceum Chromosome 1 and Its Integration with Genetic Maps." *Molecular cytogenetics* 8, no. 1 (2015): 2. doi:10.1186/s13039-015-0106-y, Available at http://www.ncbi.nlm.nih.gov/pubmed/25628758

19. Xu, Q., Xiong, G., Li, P., He, F., Huang, Y., Wang, K., Li, Z., and Hua, J. "Analysis of Complete Nucleotide Sequences of 12 Gossypium Chloroplast Genomes: Origin and Evolution of Allotetraploids." *PloS one* 7, no. 8 (2012): e37128. doi:10.1371/journal.pone.0037128, Available at http://www.ncbi.nlm.nih.gov/pubmed/22876273

20. Wendel, J. F., Brubaker, C., Alvarez, I., Cronn, R., and Stewart, J. M. "Evolution and Natural History of the Cotton Genus" *Genetics and Genomics of Cotton* (2009): 3–22. doi:10.1007/978-0-387-70810-2_1, Available at http://link.springer.com/10.1007/978-0-387-70810-2_1

21. Renny-Byfield, S., Page, J. T., Udall, J. A., Sanders, W. S., Peterson, D. G., Arick, M. A., Grover, C. E., Wendel, J. F., and Wendel, J. F. "Independent Domestication of Two Old World Cotton Species." *Genome biology and evolution* 8, no. 6 (2016): 1940–7. doi:10.1093/gbe/evw129, Available at http://www.ncbi.nlm.nih.gov/pubmed/27289095

22. Ops Diagnostics. "CTAB Protocol for the Isolation of DNA from Plant Tissues" Available at https://opsdiagnostics.com/notes/protocols/ctab_protocol_for_plants.htm

23. Rhoads, A. and Au, K. F. "PacBio Sequencing and Its Applications" *Genomics, Proteomics & Bioinformatics* 13, no. 5 (2015): 278–289. doi:10.1016/J.GPB.2015.08.002, Available at https://www.sciencedirect.com/science/article/pii/S1672022915001345

24. Pacific Biosciences. "New Chemistry Boosts Average Read Length to 10 Kb – 15 Kb for

PacBio® RS II - PacBio" *Pacific Biosciences Blog* (2014): Available at https://www.pacb.com/blog/new-chemistry-boosts-average-read/

25. QIAGEN. "TissueRuptor II" Available at https://www.qiagen.com/us/shop/automated-solutions/sample-disruption/tissueruptor-ii/#orderinginformation

26. Bionano Genomics. "Bionano Prep Kits" Available at https://bionanogenomics.com/products/bionano-prep-kits/

27. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. "Canu: Scalable and Accurate Long-Read Assembly via Adaptivek-Mer Weighting and Repeat Separation." *Genome research* 27, no. 5 (2017): 722–736. doi:10.1101/gr.215087.116, Available at http://www.ncbi.nlm.nih.gov/pubmed/28298431

28. Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., Lee, J., Lam, E. T., Liachko, I., Sullivan, S. T., Burton, J. N., Huson, H. J., Nystrom, J. C., Kelley, C. M., Hutchison, J. L., Zhou, Y., Sun, J., Crisà, A., Ponce de León, F. A., Schwartz, J. C., Hammond, J. A., Waldbieser, G. C., Schroeder, S. G., Liu, G. E., Dunham, M. J., Shendure, J., Sonstegard, T. S., Phillippy, A. M., Tassell, C. P. Van, and Smith, T. P. L. "Single-Molecule Sequencing and Chromatin Conformation Capture Enable de Novo Reference Assembly of the Domestic Goat Genome" *Nature Genetics* 49, no. 4 (2017): 643–650. doi:10.1038/ng.3802, Available at http://www.nature.com/doifinder/10.1038/ng.3802

29. Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement" *PLoS ONE* 9, no. 11 (2014): e112963. doi:10.1371/journal.pone.0112963, Available at

http://dx.plos.org/10.1371/journal.pone.0112963

30. Teh, B. T., Lim, K., Yong, C. H., Ng, C. C. Y., Rao, S. R., Rajasegaran, V., Lim, W. K., Ong, C. K., Chan, K., Cheng, V. K. Y., Soh, P. S., Swarup, S., Rozen, S. G., Nagarajan, N., and Tan, P. "The Draft Genome of Tropical Fruit Durian (Durio Zibethinus)" *Nature Genetics* 49, no. 11 (2017): 1633–1641. doi:10.1038/ng.3972, Available at http://www.nature.com/doifinder/10.1038/ng.3972

31. Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G., and Soltis, D. E. "Phylogenetic Analysis of 83 Plastid Genes Further Resolves the Early Diversification of Eudicots." *Proceedings of the National Academy of Sciences of the United States of America* 107, no. 10 (2010): 4623–8. doi:10.1073/pnas.0907801107, Available at http://www.ncbi.nlm.nih.gov/pubmed/20176954

32. Humann, J., Lee, T., Ficklin, S., Cheng, C.-H., Hough, H., Jung, S., Wegrzyn, J., Neale, D., and Main, D. "Plant and Animal Genome XXVI Conference (January 13 - 17, 2018)" *A Web-Based Platform for Structural and Functional Annotation and Curation of Genomes* (2018): P0090. Available at https://pag.confex.com/pag/xxvi/meetingapp.cgi/Paper/28580

33. Lee, T., Peace, C., Jung, S., Zheng, P., Main, D., and Cho, I. "GenSAS — An Online Integrated Genome Sequence Annotation Pipeline" *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)* (2011): 1967–1973. doi:10.1109/BMEI.2011.6098712, Available at http://ieeexplore.ieee.org/document/6098712/

34. Stanke, M., Tzvetkova, A., and Morgenstern, B. "AUGUSTUS at EGASP: Using EST, Protein and Genomic Alignments for Improved Gene Prediction in the Human Genome." *Genome biology* 7 Suppl 1, no. Suppl 1 (2006): S11.1-8. doi:10.1186/gb-2006-7-s1-s11,

Available at http://www.ncbi.nlm.nih.gov/pubmed/16925833

35. Kent, W. J. “BLAT--the BLAST-like Alignment Tool.” *Genome research* 12, no. 4 (2002): 656–64. doi:10.1101/gr.229202, Available at http://www.ncbi.nlm.nih.gov/pubmed/11932250

36. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., and Borodovsky, M. “Gene Identification in Novel Eukaryotic Genomes by Self-Training Algorithm” *Nucleic Acids Research* 33, no. 20 (2005): 6494–6506. doi:10.1093/nar/gki937, Available at https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki937

37. Burge, C. and Karlin, S. “Prediction of Complete Gene Structures in Human Genomic DNA” *Journal of Molecular Biology* 268, no. 1 (1997): 78–94. doi:10.1006/jmbi.1997.0951, Available at http://www.ncbi.nlm.nih.gov/pubmed/9149143

38. Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. “Improved Microbial Gene Identification with GLIMMER.” *Nucleic acids research* 27, no. 23 (1999): 4636–41. Available at http://www.ncbi.nlm.nih.gov/pubmed/10556321

39. Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., White, O., and White, O. “Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies.” *Nucleic acids research* 31, no. 19 (2003): 5654–66. doi:10.1093/NAR/GKG770, Available at http://www.ncbi.nlm.nih.gov/pubmed/14500829

40. Korf, I. “Gene Finding in Novel Genomes” *BMC Bioinformatics* 5, no. 1 (2004): 59. doi:10.1186/1471-2105-5-59, Available at http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-59

41. Holt, C. and Yandell, M. “MAKER2: An Annotation Pipeline and Genome-Database

Management Tool for Second-Generation Genome Projects" *BMC Bioinformatics* 12, no. 1 (2011): 491. doi:10.1186/1471-2105-12-491, Available at http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-491

42. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. "Basic Local Alignment Search Tool" *Journal of Molecular Biology* 215, no. 3 (1990): 403–410. doi:10.1016/S0022-2836(05)80360-2, Available at http://www.ncbi.nlm.nih.gov/pubmed/2231712

43. Bairoch, A. and Apweiler, R. "The SWISS-PROT Protein Sequence Database and Its Supplement TrEMBL in 2000." *Nucleic acids research* 28, no. 1 (2000): 45–8. Available at http://www.ncbi.nlm.nih.gov/pubmed/10592178

44. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. "InterProScan: Protein Domains Identifier." *Nucleic acids research* 33, no. Web Server issue (2005): W116-20. doi:10.1093/nar/gki442, Available at http://www.ncbi.nlm.nih.gov/pubmed/15980438

45. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. "Pfam: The Protein Families Database." *Nucleic acids research* 42, no. Database issue (2014): D222-30. doi:10.1093/nar/gkt1223, Available at http://www.ncbi.nlm.nih.gov/pubmed/24288371

46. Dyrløv Bendtsen, J., Nielsen, H., Heijne, G. von, and Brunak, S. "Improved Prediction of Signal Peptides: SignalP 3.0" *Journal of Molecular Biology* 340, no. 4 (2004): 783–795. doi:10.1016/j.jmb.2004.05.028, Available at http://www.ncbi.nlm.nih.gov/pubmed/15223320

47. Emanuelsson, O., Nielsen, H., Brunak, S., and Heijne, G. von. "Predicting Subcellular

Localization of Proteins Based on Their N-Terminal Amino Acid Sequence" *Journal of Molecular Biology* 300, no. 4 (2000): 1005–1016. doi:10.1006/jmbi.2000.3903, Available at http://www.ncbi.nlm.nih.gov/pubmed/10891285

48. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs" *Bioinformatics* 31, no. 19 (2015): 3210–3212. doi:10.1093/bioinformatics/btv351, Available at https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv351

49. Li, H. "Minimap2: Versatile Pairwise Alignment for Nucleotide Sequences" (2017): Available at http://arxiv.org/abs/1708.01492

50. Motamayor, J. C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone, D., Cornejo, O., Findley, S. D., Zheng, P., Utro, F., Royaert, S., Saski, C., Jenkins, J., Podicheti, R., Zhao, M., Scheffler, B. E., Stack, J. C., Feltus, F. A., Mustiga, G. M., Amores, F., Phillips, W., Marelli, J. P., May, G. D., Shapiro, H., Ma, J., Bustamante, C. D., Schnell, R. J., Main, D., Gilbert, D., Parida, L., and Kuhn, D. N. "The Genome Sequence of the Most Widely Cultivated Cacao Type and Its Use to Identify Candidate Genes Regulating Pod Color." *Genome biology* 14, no. 6 (2013): r53. doi:10.1186/gb-2013-14-6-r53, Available at http://www.ncbi.nlm.nih.gov/pubmed/23731509

51. Kaiser, M. D., Davis, J. R., Grinberg, B. S., Oliver, J. S., Sage, J. M., Seward, L., and Bready, B. "Automated Structural Variant Verification In Human Genomes Using Single-Molecule Electronic DNA Mapping" *bioRxiv* (2017): 140699. doi:10.1101/140699, Available at https://www.biorxiv.org/content/early/2017/05/22/140699

52. Mak, A. C. Y., Lai, Y. Y. Y., Lam, E. T., Kwok, T.-P., Leung, A. K. Y., Poon, A., Mostovoy, Y., Hastie, A. R., Stedman, W., Anantharaman, T., Andrews, W., Zhou, X., Pang, A. W. C., Dai, H., Chu, C., Lin, C., Wu, J. J. K., Li, C. M. L., Li, J.-W., Yim, A. K. Y., Chan, S., Sibert, J., Džakula, Ž., Cao, H., Yiu, S.-M., Chan, T.-F., Yip, K. Y., Xiao, M., and Kwok, P.-Y. "Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays." *Genetics* 202, no. 1 (2016): 351–62. doi:10.1534/genetics.115.183483, Available at http://www.ncbi.nlm.nih.gov/pubmed/26510793

53. Cao, H., Hastie, A. R., Cao, D., Lam, E. T., Sun, Y., Huang, H., Liu, X., Lin, L., Andrews, W., Chan, S., Huang, S., Tong, X., Requa, M., Anantharaman, T., Krogh, A., Yang, H., Cao, H., and Xu, X. "Rapid Detection of Structural Variation in a Human Genome Using Nanochannel-Based Genome Mapping Technology" *GigaScience* 3, no. 1 (2014): 34. doi:10.1186/2047-217X-3-34, Available at https://academic.oup.com/gigascience/article-lookup/doi/10.1186/2047-217X-3-34

54. Barseghyan, H., Tang, W., Wang, R. T., Almalvez, M., Segura, E., Bramble, M. S., Lipson, A., Douine, E. D., Lee, H., Délot, E. C., Nelson, S. F., and Vilain, E. "Next-Generation Mapping: A Novel Approach for Detection of Pathogenic Structural Variants with a Potential Utility in Clinical Diagnosis" *Genome Medicine* 9, no. 1 (2017): 90. doi:10.1186/s13073-017-0479-0, Available at http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0479-0

55. Genome.gov. "A Brief History of the Human Genome Project - National Human Genome Research Institute (NHGRI)" *National Human Genome Research Institute* (2012): Available at https://www.genome.gov/12011239/a-brief-history-of-the-human-genome-project/

56. Sanger, F., Nicklen, S., and Coulson, A. R. "DNA Sequencing with Chain-Terminating

Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74, no. 12 (1977): 5463–7. Available at http://www.ncbi.nlm.nih.gov/pubmed/271968

57. Bennett, S. "Solexa Ltd" *Pharmacogenomics* 5, no. 4 (2004): 433–438. doi:10.1517/14622416.5.4.433, Available at http://www.futuremedicine.com/doi/abs/10.1517/14622416.5.4.433

58. Ferrarini, M., Moretto, M., Ward, J. A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., Cavalieri, D., Velasco, R., Cestaro, A., and Sargent, D. J. "An Evaluation of the PacBio RS Platform for Sequencing and de Novo Assembly of a Chloroplast Genome." *BMC genomics* 14, (2013): 670. doi:10.1186/1471-2164-14-670, Available at http://www.ncbi.nlm.nih.gov/pubmed/24083400

59. Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., and Studholme, D. J. "Assessing the Performance of the Oxford Nanopore Technologies MinION" *Biomolecular Detection and Quantification* 3, (2015): 1–8. doi:10.1016/J.BDQ.2015.02.001, Available at https://www.sciencedirect.com/science/article/pii/S2214753515000224

60. Llorens, C., Futami, R., Covelli, L., Dominguez-Escriba, L., Viu, J. M., Tamarit, D., Aguilar-Rodriguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G. P., Maumus, F., Munoz-Pomer, A., Sempere, J. M., Latorre, A., and Moya, A. "The Gypsy Database (GyDB) of Mobile Genetic Elements: Release 2.0" *Nucleic Acids Research* 39, no. Database (2011): D70–D74. doi:10.1093/nar/gkq1061, Available at https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq1061

61. Rafalski, J. A. "Novel Genetic Mapping Tools in Plants: SNPs and LD-Based Approaches" *Plant Science* 162, no. 3 (2002): 329–333. doi:10.1016/S0168-9452(01)00587-8, Available at

https://www.sciencedirect.com/science/article/pii/S0168945201005878

62. Mun, J.-H., Kwon, S.-J., Yang, T.-J., Kim, H.-S., Choi, B.-S., Baek, S., Kim, J., Jin, M., Kim, J. A., Lim, M.-H., Lee, S., Kim, H.-I., Kim, H., Lim, Y., and Park, B.-S. "The First Generation of a BAC-Based Physical Map of Brassica Rapa" *BMC Genomics* 9, no. 1 (2008): 280. doi:10.1186/1471-2164-9-280, Available at http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-9-280

63. Luo, M. C., Deal, K. R., Murray, A., Zhu, T., Hastie, A. R., Stedman, W., Sadowski, H., and Saghbini, M. "Optical Nano-Mapping and Analysis of Plant Genomes" *Methods in Molecular Biology* 1429, (2016): 103–117. doi:10.1007/978-1-4939-3622-9_9, Available at http://link.springer.com/10.1007/978-1-4939-3622-9_9

64. Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., Lee, J., Lam, E. T., Liachko, I., Sullivan, S. T., Burton, J. N., Huson, H. J., Nystrom, J. C., Kelley, C. M., Hutchison, J. L., Zhou, Y., Sun, J., Crisà, A., Ponce de León, F. A., Schwartz, J. C., Hammond, J. A., Waldbieser, G. C., Schroeder, S. G., Liu, G. E., Dunham, M. J., Shendure, J., Sonstegard, T. S., Phillippy, A. M., Tassell, C. P. Van, and Smith, T. P. L. "Single-Molecule Sequencing and Chromatin Conformation Capture Enable de Novo Reference Assembly of the Domestic Goat Genome" *Nature Genetics* 49, no. 4 (2017): 643–650. doi:10.1038/ng.3802, Available at http://www.nature.com/articles/ng.3802

65. Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmöckel, S. M., Li, B., Borm, T. J. A., Ohyanagi, H., Mineta, K., Michell, C. T., Saber, N., Kharbatia, N. M., Rupper, R. R., Sharp, A. R., Dally, N., Boughton, B. A., Woo, Y. H., Gao, G., Schijlen, E. G. W. M., Guo, X., Momin, A. A., Negrão, S., Al-Babili, S., Gehring, C., Roessner, U., Jung, C., Murphy, K., Arold, S. T., Gojobori, T., Linden, C. G. van der, Loo, E. N. van, Jellen, E. N., Maughan, P. J., and Tester, M.

"The Genome of Chenopodium Quinoa" *Nature* 542, no. 7641 (2017): 307–312. doi:10.1038/nature21370, Available at http://www.nature.com/doifinder/10.1038/nature21370

FIGURES

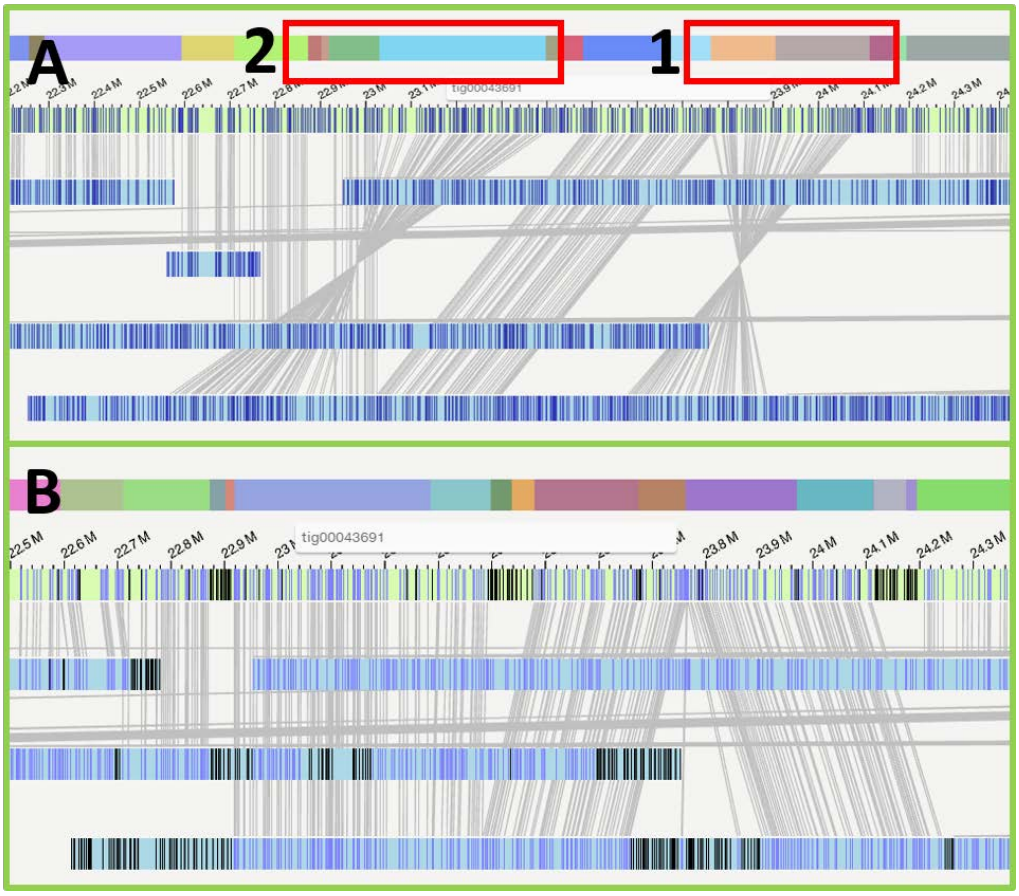

Figure 1: In A, an orientation change involving a block of two contigs can be seen in box 1. A contig that needs to be moved and then re-oriented can be seen in box 2. In B, we have the corrected PG pseudomolecule sequence from A.

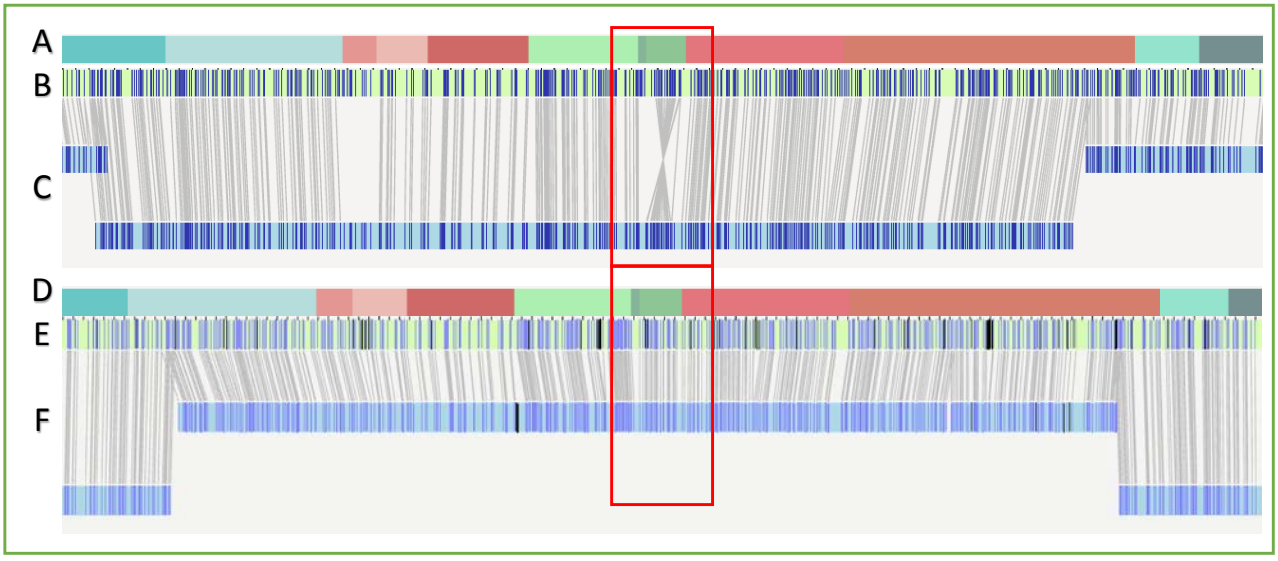

Figure 2: Simple Hi-C based inversion error corrected by Bionano physical maps. Seen in A and D a .bed file representing the contigs as placed and oriented by Phase Genomics Hi-C. In B and E, we have the pseudomolecule. In C and F, we have the Bionano maps. The red boxes indicate a contig correctly placed by Hi-C, and confirmed by Bionano maps in C and F. The orientation is incorrect, however, which is represented by the inversion in ABC. After correcting the orientation in the Hi-C data, and re-generating the sequence fasta, we improved the sequence assembly, represented in DEF.

749	tig00049110	1	7.35612	.
323	tig00013888	1	121.451	.
39	tig00001877	0	40.8089	.
675	tig00048001	1	1.21307	.
46	tig00002131	1	0.334039	
455	tig00043691	1	149.004	.
454	tig00043689	1	22.9583	.
456	tig00043693	0	3.87406	.
457	tig00043695	0	4.44861	.
453	tig00043687	1	20.6712	.
452	tig00043685	1	12.0337	.
126	tig00005138	0	60.25	.
545	tig00045230	0	38.7148	.
154	tig00006695	1	6.51379	.
130	tig00005292	1	7.72355	.
19	tig00000959	0	0.0361116	
327	tig00013945	1	0.553459	
219	tig00009265	0	9.68717	.

Figure 3: A PG generated group ordering file, indicating contig number in column 1, contig ID in column 2, orientation in column 3, and log link liklihood in column 4.

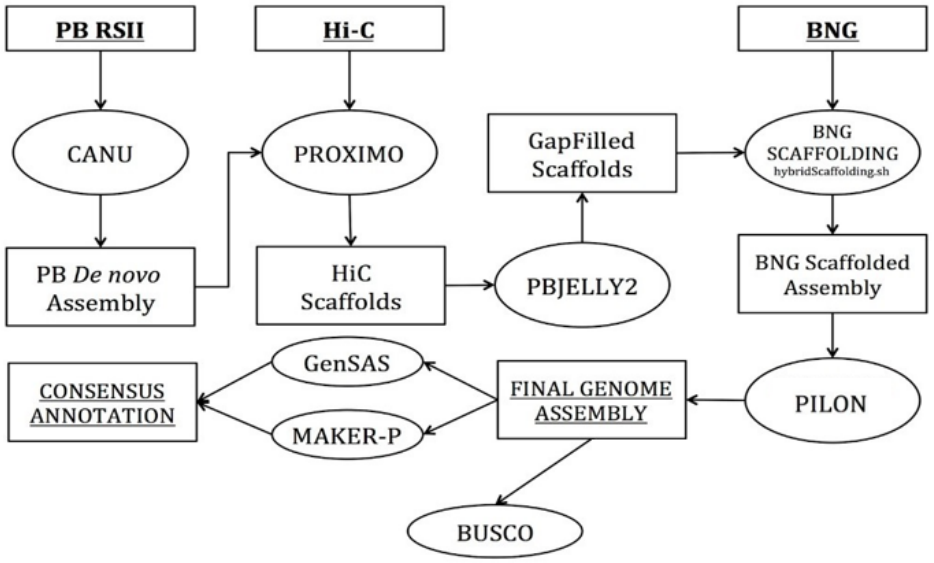

Figure 4: Assembly Workflow. Square boxes represent data or post-analysis data. Circles represent processes and programs ran. PB RSII, Pacific Biosciences RSII long read sequencing technology. Hi-C, Phase Genomics Hi-C data collection, titled PROXIMO. PBJELLY, gap filling software. BNG, Bionano Genomics. PILON, error correction of PB long read data with Illumina short read data. BUSCO, gene space quality check. MAKER-P, genome annotation. GenSAS, online genome annotation unifying many separate genome annotation programs.

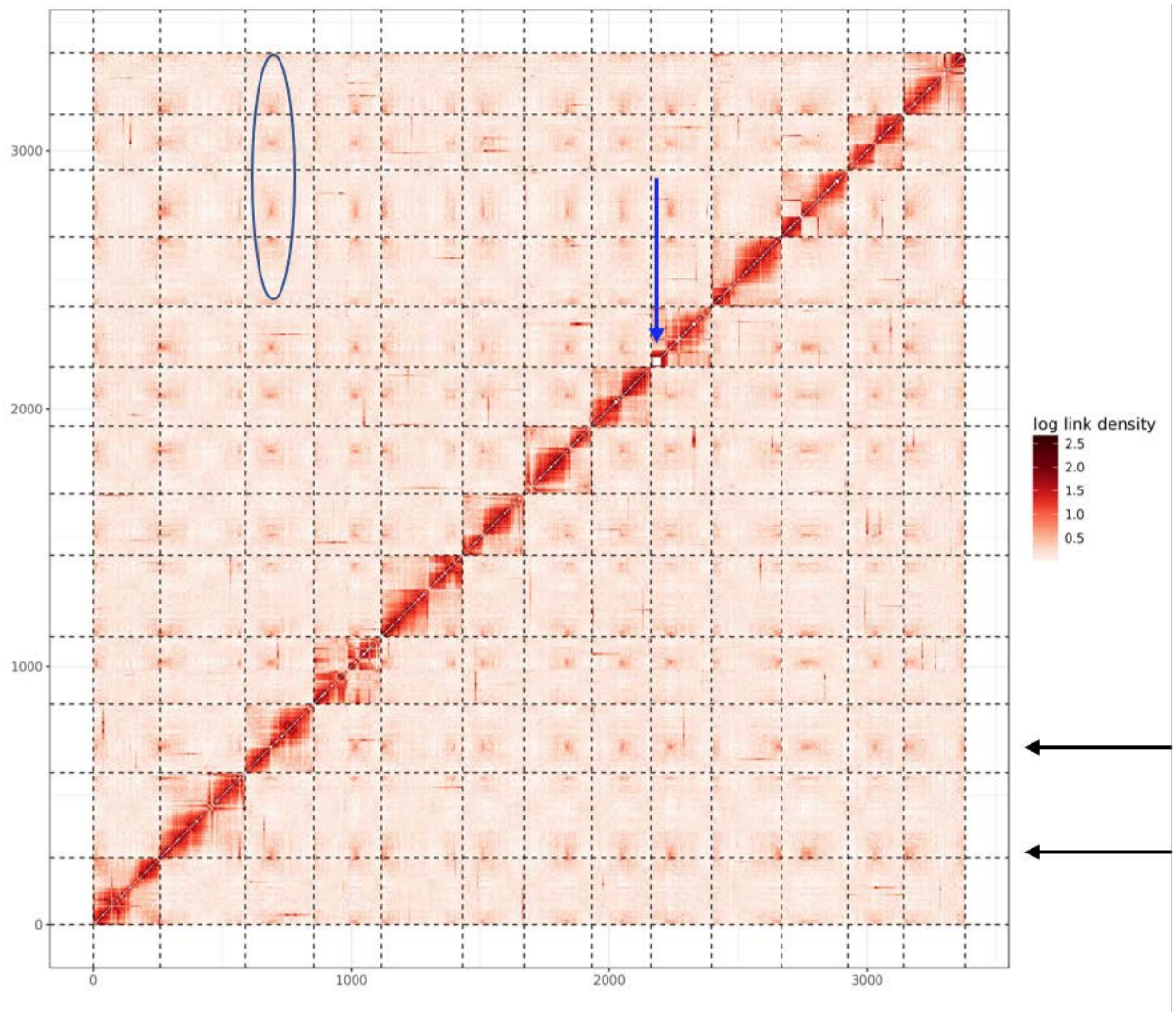

Figure 5: Pairwise heat map showing log-likelihood of contig placement. The X and Y axis are each contig of every pseudomolecule laid end to end. The darker the red of the long link density the more interactions each contig has with its neighboring contigs. The diagonal axis represents the alignment of each sequence to itself. “Dots” of red, exemplified by a blue circle, outside the diagonal axis represent regions which have high frequency of interactions between chromosomes. Regions consistently along the “end” or “center” of a pseudomolecule likely represent telomeres or centromeres, as indicated by black arrows. White space along the center of the diagonal represent individual contigs with sufficient length to be seen, at this level, by the naked eye. One particular contig is indicated with a blue arrow.

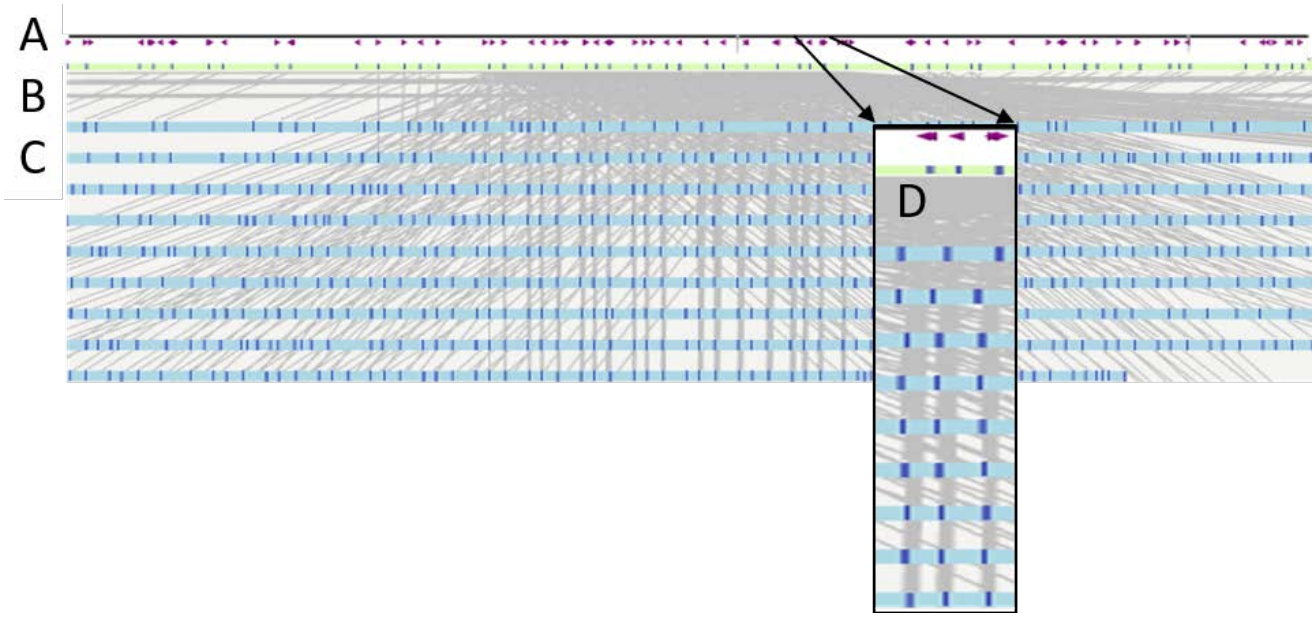

Figure 6: Exemplification of Bionano motif repeat identified with Bionano contigs. A: Geneious *in silico* digestion identifying BSSSI nick sites. B: PG Assembly contig with *in silico* digestion with BSSSI nick sites. C: Bionano contigs aligning to the motif repeat area. D: Magnified selection of the motif repeat.

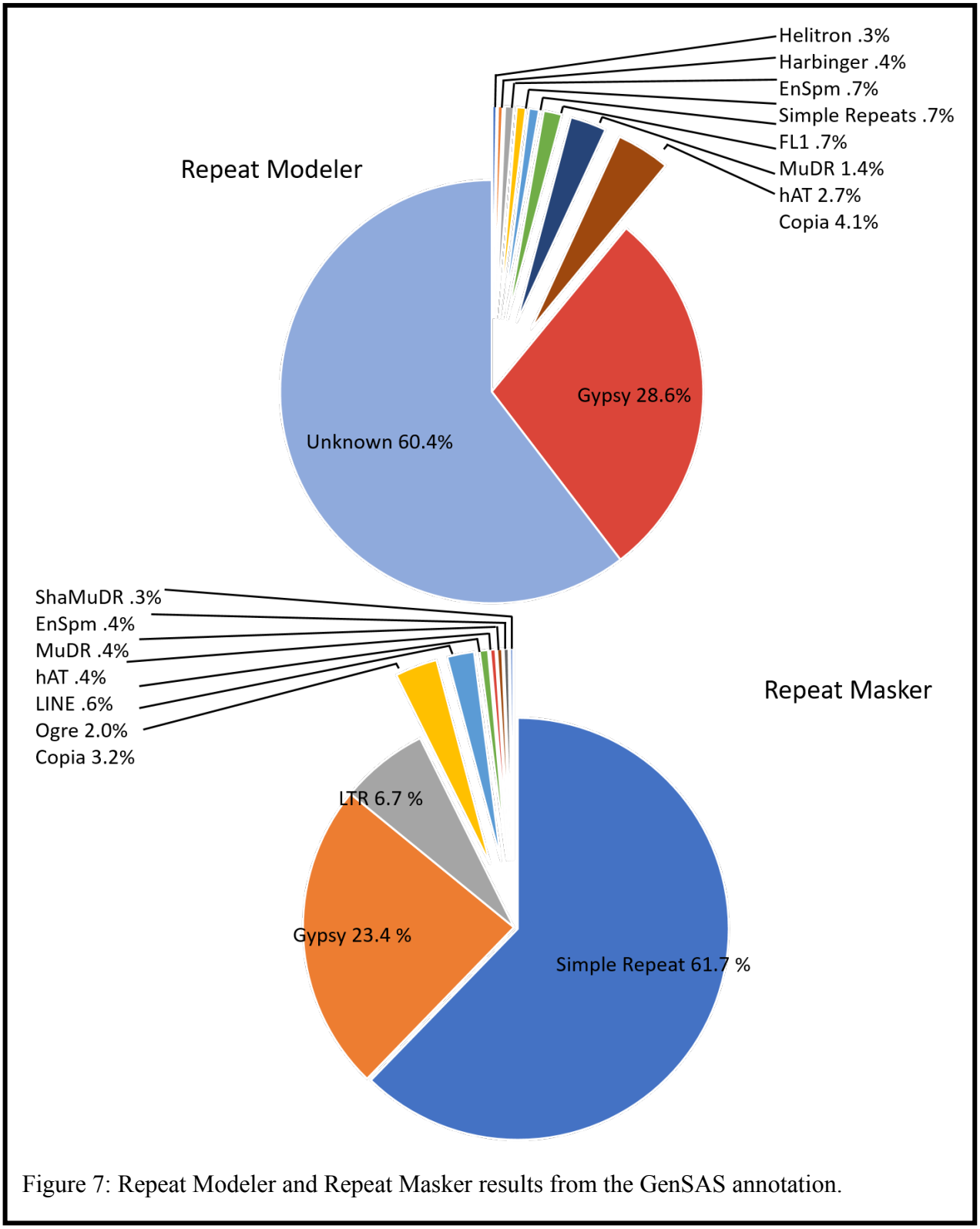

Figure 7: Repeat Modeler and Repeat Masker results from the GenSAS annotation.

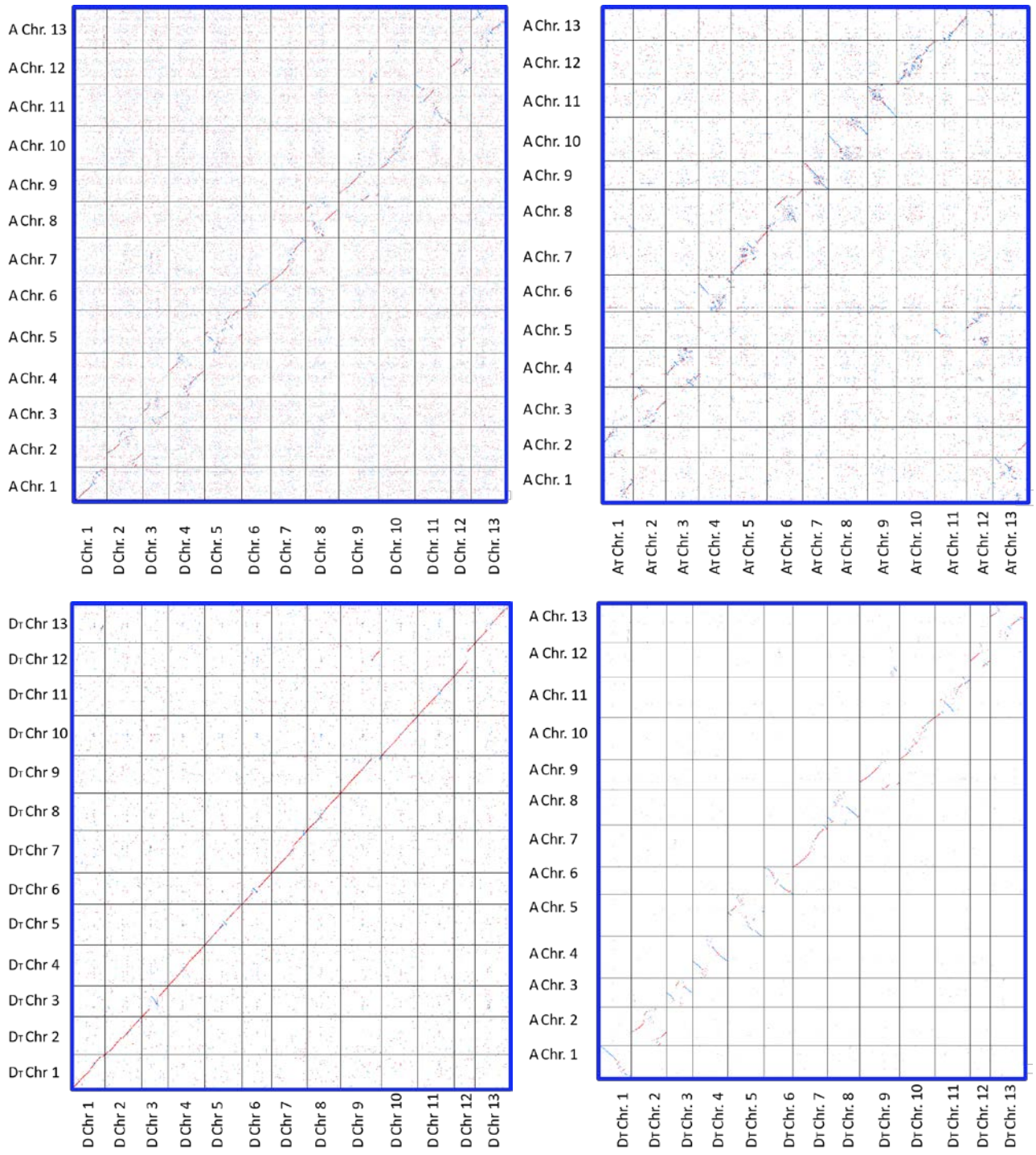

Figure 8: Minimap2 sequence to sequence alignments. Alignment of *G. herbaceum* to *G. raimondii*, top left. Alignment of *G. herbaceum* to *G. hirsutum* A_T, top right. Alignment of *G. hirsutum* D_T, to *G. raimondii,* bottom left. Alignment of *G. herbaceum* to *G. hirsutum* D_T, bottom right.

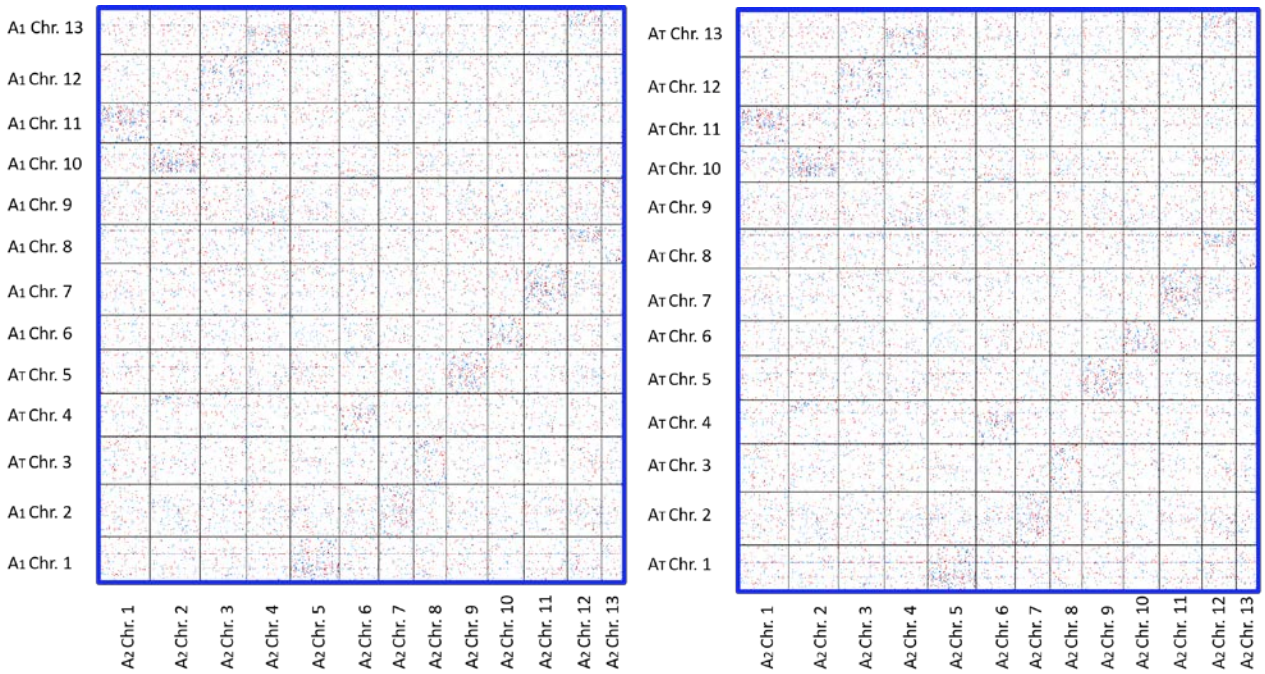

Figure 9: Minimap2 sequence to sequence alignments. Alignment of *G. herbaceum* to *G. arboreum*, left. Alignment of *G. arboreum* to *G. hirsutum* A_T, right.

TABLES

Table 1: Assembly Statistics of the draft genome assembly at various stages along the assembly process. Statistics assessed using GAEMER basic statistics.

Assembler	CANU	CANU-PG	CANU-PG-PBJELLY2	CANU-PG-PBJELLY2-PILON	CANU-PG-MANUGAL-INTEGRATION
Contigs	9,280	9,280	5,484	5,462	9,259
Max Contig	6,756,708	6,756,708	6,757,302	6,760,572	6,756,708
Mean Contig	171,321	171,321	295,809	297,103	171,482
Contig N50	315,162	315,162	684,931	688,517	314,989
Contig N90	78,589	78,589	157,958	158,537	78,583
Total Contig Length	1,589,858,884	1,589,858,884	1,622,219,146	1,622,775,370	1,587,752,003
Assembly GC	35.04	35.04	35.1	35	35.04
Scaffolds	9,280	1,086	1,058	1,058	1,029
Max Scaffold	6,756,708	138,011,914	141,045,733	141,119,079	139,474,330
Mean Scaffold	171,321	1,464,713	1,533,456	1,533,976	1,543,805
Scaffold N50	315,162	126,778,845	129,674,467	129,721,115	127,491,032
Scaffold N90	78,589	96,373,437	98,292,082	98,327,974	96,444,436
Total Scaffold Length	1,589,858,884	1,590,678,284	1,622,396,686	1,622,946,093	1,588,575,003
Captured Gaps	0	8,194	4,426	4,404	8,230
Mean Gap	0	100	40	39	100
Gap N50	0	100	50	55	100
Total Gap Length	0	819,400	177,540	170,723	823,000

Table 2: Bionano Physical Map Statistics

Number of Genome Maps	1842
Total Genome Map Length (Mb)	1569.623
Average Genome Map Length (Mb)	0.852
Median Genome Map Length (Mb)	0.633
Genome Map n50 (Mb)	1.195
Total Reference Length (Mb)	1579.424
Total Genome Map Length / Reference Length	0.994
Number of Genome Maps which Align	1806 (0.98)
Total Aligned Length (Mb)	1462.786
Total Aligned Length / Ref Length	0.926
Total Unique Aligned Length (Mb)	1425.824
Total Unique Aligned Length / Reference Length	0.903

Table 3: BUSCO Statistics

BUSCO 2.0 beta 4	
embryophyta_odb9 (Creation date: 2016-02-13, number of species: 30, number of BUSCOs: 1440)	
Summarized benchmarking in BUSCO notation for Wagad genome assembly	
BUSCO mode: genome	
C:92.8%[S:84.6%,D:8.2%],F:2.0%,M:5.2%,n:1440	
1,336	Complete BUSCOs (C)
1,218	Complete and single-copy BUSCOs (S)
118	Complete and duplicated BUSCOs (D)
29	Fragmented BUSCOs (F)
75	Missing BUSCOs (M)
1,440	Total BUSCO groups searched

Table 4: Total Manual Edits. Edits made to PG scaffolded pseudomolecules during manual integration of Bionano and Hi-C data. Total edits made per chromosome listed as well as total contigs previously unscaffolded which were incorporated into psuedomolecules via manual integration.

Chromosome	# Edits Made	Scaffolded (previously unscaffolded contigs)
A1-1	46	3
A1-2	50	1
A1-3	101	4
A1-4	54	1
A1-5	61	1
A1-6	75	8
A1-7	52	2
A1-8	115	3
A1-9	52	0
A1-10	93	2
A1-11	86	2
A1-12	99	8
A1-13	50	1
Total	934	36

Table 5: Effects of Manual Integration on Pseudomolecules. Total size changes effected by manual scaffolding of previously unscaffolded contigs. A total of 36 contigs were scaffolded into the pseudomolecuels.

Chr.	Edited Chr. Length	Original Chr. Length	Difference in bp	Difference in Mb
A1-1	127,491,132	126,778,945	712,187	0.712187
A1-2	96,444,536	96,373,537	70,999	0.070999
A1-3	137,780,711	137,028,703	752,008	0.752008
A1-4	100,555,661	100,417,012	138,649	0.138649
A1-5	113,840,674	113,769,675	70,999	0.070999
A1-6	139,474,430	137,497,694	1,976,736	1.976736
A1-7	104,484,132	103,955,798	528,334	0.528334
A1-8	138,135,069	137,812,063	323,006	0.323006
A1-9	90,398,184	90,398,184	0	0
A1-10	133,339,774	133,115,148	224,626	0.224626
A1-11	138,196,879	137,989,025	207,854	0.207854
A1-12	115,947,866	113,711,559	2,236,307	2.236307
A1-13	123,191,655	123,140,091	51,564	0.051564
Total	1,559,280,703	1,551,987,434	7,293,269	7.293269