2018-06-01

Exploration of the Gossypium raimondii Genome Using Bionano Genomics Physical Mapping Technology

Christopher Jon Hanson
Brigham Young University

Exploration of the *Gossypium raimondii* Genome Using Bionano Genomics Physical Mapping Technology

Christopher Jon Hanson

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

Master of Science

Peter J. Maughan, Chair
David E. Jarvis
Steven M. Johnson

Department of Plant and Wildlife Sciences

Brigham Young University

ABSTRACT

Exploration of the *Gossypium raimondii* Genome Using
Bionano Genomics Physical Mapping Technology

Christopher Jon Hanson
Department of Plant and Wildlife Sciences, BYU
Master of Science

Cotton is a crop with a large global economic impact as well as a large, complex genome. Most industrial cotton production is from two tetraploid species (*Gossypium hirsutum* L. *and Gossypium barbadense* L.) which contain two subgenomes, specifically the A_T and D_T subgenomes. The D_T subgenome is nearly half the size of the A_T subgenome in tetraploid cotton and is closely related to an extant D-genome *Gossypium* species, *G. raimondii* Ulbr. Characterization of the structural variants present in diploid D-genome should provide greater insight into the evolution of the D_T subgenome in the tetraploid cotton. Bionano (BNG) optical mapping uses patterns of fluorescent labels inserted at specific endonuclease sites to create physical maps of the genomes which can then be examined for structural variation. To develop optical maps in *G. raimondii*, we first developed a *de novo* PacBio long read sequence assembly of *G. raimondii*. This sequence assembly consisted of 2,379 contigs, an average contig length of 413 Kb and a contig N50 of 4.9 Mb. Using BNG technology, we developed two optical maps of the diploid D genome of *G. raimondii*. One was created using the Nt.BssSI endonuclease and one with the Nt.BspQI endonuclease. Using the BNG optical maps, the PacBio assembly was hybrid scaffolded into 100 scaffolds (+ 5 unscaffolded contigs) with an average scaffold length of 7.5 Mb and a scaffold N50 of 13.1 Mb. A comparison between the Nt. BssSI BNG optical map and the two sequence assemblies identified 3,195 structural variants. These were used to validate the accuracy of the reference sequence of *G. raimondii* and structural variants were used to create a new phylogeny of nine major cotton species.

Keywords**:** cotton, *Gossypium raimondii*, Bionano Genomics, physical mapping, structural variants

ACKNOWLEDGEMENTS

I would like to thank Joshua Udall for all of his help in training me to be the scientist I am today. It was due to him that I was able to take on this master's program. I would also like to thank Jeff Maughan for pushing me and helping me create a finished product that I am proud of.

I would like to thank my family and especially my wife Mackenzie for all the love and support that she has given throughout my entire schooling. She has made so many things possible for me that may not have otherwise been possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The hybridization of diploid A-genome and D-genome cotton happened approximately 1.2 million years ago [1]. This hybridization created an AD-genome allotetraploid species (Figure 1). About 4,000 years ago, Old world, diploid cotton (*Gossypium arboreum* L. *and G. herbaceum* L.) and New world, tetraploid cottons (*G. barbadense* L. *and G. hirsutum* L.) were domesticated [2], with the domestication process being primarily guided by the selection of longer fibers and disease/drought resistance. This selective breeding has undoubtedly influenced the evolution of the cotton genome. A-genome diploid cotton has fibers that are of a suitable length for spinning, whereas D-genome diploid cottons do not produce any useful fibers. The A-genome diploids (G. *herbaceum* and G. *arboreum*) were domesticated in the Old World, whereas tetraploid cotton (G. *hirsutum*) was domesticated in the New World [3]. Tetraploid cotton produces much longer and higher quality fibers than either the A or Dgenome alone, suggesting that the polyploidization event was essential for fiber improvement.

The genomes of A-diploid, D-diploid and tetraploid cotton species have been studied [4], and most of the genomes in the *Gossypium* genus have annotated genome assemblies. While genome sequences for different cotton species exist, most of them are draft assemblies constructed using short read sequencing technologies [5-7]. While useful in cotton genomic research, these genome assemblies could be more useful if the assembled genomes were more complete (contiguous), with improved accuracy. Improved cotton genomes would further illuminate the evolutionary history of cotton, enhance the rate of cotton improvement in breeding programs, facilitate the discovery of

new genes or other genomic regions of interest, and help researchers better understand the role of genomic rearrangements in evolution.

The only draft assembly of *G. raimondii* was reported by Paterson et. al (2013) [6] using shotgun and BAC-end Sanger sequences. This assembly was fragmented, consisting of 1,033 scaffolds, spanning 760 Mb (95% of total genome length), presumably due to the repeat content of the genome and the inability of the short reads to span the repeats. Third generation sequencing technologies, including Pacific Biosciences (PacBio) sequencing, produce extremely long sequence reads that can span the large repeats characteristic of many plant genomes. The ability to span large repeats results in significantly improved continuity (high N50) which more accurately reflects the structure of the genome.

While 3rd generation sequencing technologies dramatically improve continuity, assemblies can be further improved when scaffolded with BNG physical maps. Accurately assembled genomes should have highly overlapped alignments between their sequence and physical maps [7]. BNG physical mapping technology generates molecule maps that are approximately 500X the size of an average Illumina read and 150X the size of an average Sanger read. Individual BNG molecules are easily assembled into a physical assembly of the genome based on overlapping fingerprint patterns [9-14] [7]. Alignment of the sequence-based assembly with BNG physical assembly can then be used to scaffold sequence contigs – often times into Mb size scaffolds. Besides verification (good sequence assemblies should have high congruency with individual BNG maps), BNG physical assemblies can also be used to identify structural variation between genomes [8]. Structural variants (SVs) are relatively

large sections, generally larger than 1 Kb in size, of a chromosome that has been rearranged (e.g., balanced translocations, inversions, and copy number variants) [9]. Using BNG technology, physical assemblies can be generated with reasonable ease and efficiency that enable accurate SV identification [17-19]. SVs are hypothesized to have played a crucial role in the evolution of cotton, as well as in other species, including African cichlid fishes [10], *Drosophila melanogaster* [11] and even humans [12]. Understanding and correctly categorizing these SVs will greatly improve our understanding of the evolutionary history that shaped the modern cotton genome.

Here we report the *de novo* assembly of a BNG optical map of *G. raimondii* and a PacBio based *de novo* assembly of *G. raimondii*. We report the identification of large scale SVs within the *G. raimondii* genome by comparing the physical assembly to both our PacBio assembly and a previously assembled *G. raimondii* reference sequence [13]. The physical maps were used to validate all three of the previously mentioned assemblies (Nt. BssSI BNG physical map, PacBio de novo assembly and the Paterson Sanger reference), by comparing similarities and differences in SVs between them. Finally, a phylogeny based on SVs of nine major species of cotton was generated and compared to a currently accepted cotton phylogeny. Because no technology prior to BNG optical mapping has been capable of delivering such a comprehensive analysis of SV and general genome structure, this paper represents the first report of a phylogeny of species in the Gossypium genus created solely using SVs.

MATERIALS AND METHODS

Plant Material, DNA Extraction and Bionano Data Collection

Fresh, young leaf tissue was collected from *G. raimondii* plants grown in the BYU research greenhouse. These were grown in basic potting soil with a natural day length. Two grams of tissue were collected and fixed in 2% formaldehyde solution and roughly chopped with a razor blade. The tissue was then blended with an immersion blender and the slurry was filtered sequentially through 100 µm and 40 µm filters. Nuclei were pelleted and washed with a series of centrifugations followed by the addition of fresh buffer. The nuclei were separated from the suspension using a Percoll (Sigma-Aldrich GE17-5445-01) density gradient and then washed and cast into 2% agarose plugs. The nuclear membrane was lysed in the agarose plugs using a proprietary Bionano lysing agent (Bionano Genomics, San Diego, cat#80003, part#20255) followed by four washes using Bionano 5X wash buffer (Bionano Genomics, San Diego, cat#80003, part#20256) and five washes using TE buffer. The agarose plugs were melted and drop dialyzed using TE buffer [14].

Extracted DNA was nicked, labelled, and repaired according to BNG protocols [15]. In brief, single-stranded nicks were created using the endonuclease Nt.BspQI (NEB cat#R0644) or Nt.BssSI (NEB cat# R0681) which recognize seven and six base pair recognition sites, respectively. Fluorescently labeled nucleotides were incorporated into the cut sites using DNA polymerase I (NEB cat# M0273S) and repaired using *Taq* DNA ligase I (NEB cat #M0208). After labeling the nicked sites with fluorescent nucleotides, the DNA backbone was stained with DAPI to enable visualization and length estimates by the Irys™ instrument. Before running the Irys instrument for

molecule detection, an optimization protocol, described in the Irys instrument user manual [16], was run to optimize the voltage used and the time spent concentrating the DNA prior to moving it into the nanochannels (Figure 2).

Physical Map Assembly and Initial SV Identification

The data from the Irys instrument was passed to the BNG IrysSolve™ assembler. The initial assembly *p*-value was set to $1.25X10^{-8}$ while the extension and refinement *p*-values were set to $1.25X10^{-9}$. Molecules shorter than 150 Kb or with fewer than eight fluorescent labels were discarded. The computer program SVcompare™, included in the package with IrysSolve, was then used to call SVs relative to a reference sequence. SV location, size and type was reported as a bed file, which was visualized on IrysView's™ graphical interface (Figure 3).

Generation of PacBio data

Fresh leaf tissue from the *G. raimondii* plant from the BYU greenhouse was freeze-dried prior to DNA extractions. A CTAB extraction method was used to extract high molecular weight (HMW) DNA from this tissue [17]. PacBio sequencing libraries were prepped at the DNA sequencing center at Brigham Young University (Provo, UT) according to the manufacturer's standard protocols [18], using Sequel 2.1 chemistry with SMRTbell libraries developed from Template Prep Kit 1.0. In short, the process includes shearing the DNA, concentrating and purifying it, and doing a blunt-end ligation. Fragments larger than 19 Kb were size selected using a Blue Pippin (Sage

Science, Beverly, MA). After size selection the DNA was purified once more and damaged sites were repaired. It was then loaded onto PacBio SMRTcells™ and sequenced on the PacBio Sequel™ instrument.

PacBio Sequence and Hybrid Assemblies

The PacBio reads were assembled using the CANU v.1.4 *de novo* assembler [19] with the expected genome size set to 750 Mb, corMhapSensitivity=normal and corOutCoverage=40. For hybrid assembly, the BNG Hybrid Scaffold™ assembler (default parameters and not cutting at conflicting sites) uses the BNG physical map assembly with an *in silco* digested sequence assembly (assembled by Canu) to output a "hybrid scaffold assembly" that incorporates both NGS contigs and BNG contigs. Hybrid scaffold assembly should combine contigs, thus reducing the total number of scaffolds (L50) in the final assembly while increasing the overall size of the scaffolds (N50).

Verification of Bionano, PacBio and Paterson Sanger Assemblies of G. raimondii using SVs

The accuracy of the BNG and PacBio assemblies were validated using SVcompare. SVcompare was used independently of the IrysSolve assembler to call a list of SVs in the BNG map aligned to the *de novo* PacBio assembly. This process outputs a .bed file which lists all the SVs and their locations. BEDtools [20], bedintersect was used to output a new bed file that contained all of the SVs that are shared between the two comparisons (e.g., BNG map to the Paterson Sanger sequence vs. BNG map to the *de novo* PacBio assembly). An important thing to note is that the Paterson Sanger

reference and the PacBio de novo sequence assembly, were made using the same accession go G. raimondii (D5-3) and the Bionano physical map was not made using that same accession. Because these three assemblies were made with three different technologies, a high number of similar SVs between the assemblies would suggest that all three assemblies are correctly assembled (Figure 4).

BNG Maps From Other Cotton Species

The BNG maps of seven other cotton species were provided by Joshua A. Udall, (Brigham Young University, Provo, Utah, USA). These included an A-genome diploid species [*G. herbaceum* L. (wagad)] and six tetraploid species [*G. tomentosum* Nutt ex Seem*, G. barbadense* L*, G. hirsutum* L. (including Acala Maxxa and TM1), *G. stevensii SG*, *G. darwinii G. Watt*, and *G. mustelinum* Miers ex G.Watt. The same protocol as described previously, using Nt.BssSI, was used to generate these maps.

Creation of a Novel Phylogeny of Nine Prominent Cotton Species

Regions containing SVs among the nine cotton species were compared using an in house developed R script [21] using *G. hirsutum* (TM1) as the reference for comparison. The R script created bins of 5 Kb regions in the genomes and then compared them to the reference to determine if the regions contained structural similarity. A binary (0,1) structural similarity matrix based on genome structure was developed for each bin of the genome across all species relative to the TM1 reference map. Phylogenetic signatures were determined from the matrix using the phylogeny

building package [22] in R. This package used an agglomerative hierarchical clustering method of creating the phylogenetic tree. Each species began in its own cluster and then the similarity (and dissimilarity) between each species was tested in a pairwise manner. The most similar clusters were joined iteratively until the final product was one large cluster, or as we see it, one phylogenetic tree.

RESULTS

Bionano Physical Map Assembly

A total of 533 Gb of data in 4,796,436 individual molecules was collected from the Irys instrument for the D5 *G. raimondii* optical map created with Nt.BssSI. The average length and N50 of these molecules was 111 Kb and 142 Kb, respectively. The average fluorescent labels per 100 Kb of DNA was 9.9. The assembled Nt.BssSI BNG physcial map assembly consisted of 1,020 contigs, spanning 737.6 Mb. Average contig length was 723 Kb, contig N50 was 936 Kb and maximum contig length was 5.6 Mb. After alignment to the Paterson Sanger assembly, the BNG contigs spanned 634 Mb of the 752 Mb sequence assembly (Table 1). The *G. raimondii* map, created with Nt.BspQI, was generated from a total of 230.5 Gb of data in 1,175,238 individual molecules. The average length of molecules in the Nt.BspQI dataset was 196 Kb and the molecule N50 was 210 Kb, with an average of 6.0 fluorescent labels per 100 Kb of DNA. The assembled Nt.BspQI BNG physical map assembly consisted of 413 contigs, spanning of 240.9 GB. The average contig length was 1.8 Mb, with a contig N50 of 2.6 Mb with the largest contig spanning 9.8 Mb. After alignment to the Paterson

Sangernce assembly, the Nt.BspQI-based contigs spanned 563.3 Mb of the 752 Mb sequence assembly (Table 2).

PacBio G. raimondii Reference Assembly and Hybrid Scaffolding

The PacBio Sequel instrument yielded 2,707,851 reads (32.7 Gb), equivalent to 43.7X coverage of the G. rainmondii genome. The average read length was 14 Kb and the read N50 was 22.5 Kb across all five cells. These reads were assembled into 2,379 contigs that spanned 982.7 Mb. The average contig size was 413 Kb, contig N50 was 4.9 Mb, N90 was 117Kb, L50 and L90 were 53 and 526 respectively (Table 3).

Hybrid scaffolding with the Nt.BssSI physical map resulted in 106 scaffolds with an average hybrid scaffold length of 7.3 Mb and a scaffold N50 of 13.1 Mb. The total length of the hybrid assembly was 778.1.5 Mb, which is close to the expected size of the D genome (752 Mb; Paterson et al.). The hybrid assembly was then re-scaffolded with the Nt.BspQI map which further reduced to the number of scaffolds to 43, with an average scaffold length and N50 of 18.3 Mb and 25.8 Mb, respectively. The max scaffold length in this assembly was 37 Mb. The total length of the hybrid scaffold assembly was 785 Mb without unscaffolded PacBio contigs and 795 Mb when including all unscaffolded PacBio contigs. Each of these lengths is closer to the expected size of the D genome than that of the previously assembled sequence (Table 4).

Structural Variation Identification

After *de novo* assembly of the BNG molecules, IrysSolve was used to align the BNG assemblies to the previously reported [6] Paterson Sanger assembly of *G. raimondii* that has been labeled *in silico* with the same endonucleases as those used in the BNG assemblies (Nt.BssSI and Nt.BspQI). A total of 611 Mb (81%) and 563 Mb (75%) of the Nt.BssSI and Nt.BspQI BNG maps, respectively, aligned to the *G. raimondii* reference genome sequence. Within the Nt.BssSI aligned regions, SvCompare (a subprogram of IrysSolve) called 46 deletions, 449 insertions, 126 translocations, 189 other complex SVs and 373 end SVs which are defined as unaligned regions of at least five label sites and 50 Kb at one end of the genome map. The average size of insertions identified was 12.5 Kb, average deletion size was 9.8 Kb, and average translocation size was 5.5 Mb. Within the Nt.BspQI aligned regions, SVCompare identified 101 deletions, 611 insertions, 150 translocations, one inversion, 1,007 other complex SVs and 143 end SVs. The average size of insertions was 13.7 Kb, average deletion size was 7.7 Kb, and average translocation size was 3.1Mb (table 5).

The BNG physical assemblies were then aligned to the *in silico* labeled hybrid PacBio assembly described earlier (post hybrid scaffolding) in order to call SVs based on the PacBio assembly. 97% of bionano contigs aligned to the PacBio sequence. In the PacBio aligned regions, SVCompare identified 782 deletions, 300 insertions, 21 translocations, one inversion, 88 other complex SVs and 165 end SVs. The average size of insertions was 5.0 Kb, average deletion size was 2.0 Kb, and average translocation size was 16.2 Mb. BedIntersect, a subtool of bedtools [20], was used to

identify common SV features between the BNG physical map assemblies and the sequence assemblies (*de novo* PacBio sequence and Paterson Sanger sequence. In total, 777 SVs were identified that were shared between the comparisons, including 288 insertions and 489 deletions (Table 7). The average size of the shared structural variants was 23 Kb with the largest structural variant spanning 667 Kb and the smallest SV spanning just over 2 Kb.

Phylogeny Based on the Similarities and Differences in SVs Between Species

BNG physical assemblies created for *G. herbaceum (wagad)], G. tomentosum, G. barbadense, G. hirsutum (including Acala Maxxa an TM1), G. stevensii, G. darwinii, and G. mustelinum* were graciously provided by Joshua A. Udall (Brigham Young University, Utah). The physical map created for *G. herbaceum* consisted of 1,838 contigs, spanning 1.5 Gb. The average contig length was 852 Kb, and the contig N50 was 1.2 Mb. The physical map created for *G. hirsutum* (Maxxa) consisted of 2,087 contigs, spanning 2.3 Gb. The average contig length was 1.1 Mb, and the contig N50 was 1.6 Mb. The map created for *G. hirsutum* (TM1) consisted of 2,196 contigs, spanning 2.2 Gb. The average contig length was 995 Kb, and the contig N50 was 1.3 Mb. The map created for *G. barbadense* consisted of 1,711 contigs, spanning 2.1 Gb. The average contig length was 1.2 Mb, and the contig N50 was 1.8 Mb. The map created for *G. tomentosum* consisted of 3,176 contigs, spanning 1.9 Gb. The average contig length was 618 Kb, and the contig N50 was 740 Kb. The *G. darwinii* map consisted of 4,939 contigs, spanning 2.0 Gb. The average contig length was 406 Kb, and the contig N50 was 473 Kb. The *G. stevensii* (wilkes) map consisted of 3,246

contigs, spanning 2.1 Gb. The average contig length was 666 Kb, and the contig N50 was 846 Kb. The *G. mustelinum* map consisted of 4,622 contigs, spanning 1.8 Gb. The average contig length was 397 Kb, and the contig N50 was 438 Kb (Table 6).

Using the matrix described in Materials and Methods, a phylogeny was created (Figure 6). Out of the 43,402 regions compared between the *G. hirsutum* reference genome (TM1) and the other species, *G. raimondii* had 2,057 matching regions, *G. herbaceum* had 7,664 matching regions, *G. mustelinum* had 21,115 matching regions, *G. darwinii* had 27,751 matching regions*, G tomentosum* had 29,530 matching regions, *G. barbadense* had 30,272 matching regions, *G. stevensii* had 34,964 matching regions and *G. hirsutum* (Acala Maxxa) had 35,411 matching regions. The two diploid species of cotton are obviously separated from the rest of the species on the phylogeny. The remaining six tetraploid species are within the same clade and mostly reflect the interspecies relationships identified in previously made phylogenies [23]. In this novel phylogeny the *G. hirsutum* (TM1) BNG map was used as the reference and the remaining eight BNG maps were compared to it to identify the amount of structural homology between them. The total number of bins that were created for this comparison was 43,402. Taking the number of bins similar between species and dividing it by the total number of bins gives the homology percentage. When compared with this TM1 map, *G. herbaceum* (wagad) showed 17.6% homology, *G. raimondii* showed 4.7% homology, *G. mustelinum* showed 48.7% homology, *G. darwinii* showed 63.9% homology, *G. tomentosum* showed 68.0% homology, *G. barbadense* showed 69.8% homology, *G. stevensii* (wilkes) showed 80.6% homology and *G. hirsutum* (maxxa) showed 82.0% homology (Table 8).

DISCUSSION

Improved Assembly of the G. raimondii Genome

Previous to the generation in this report of a physical map of *G. raimondii*, very few plant genomes had been mapped using the BNG method. Plant cell walls and other unknown secondary metabolites make it difficult to extract enough HMW DNA to accurately create a useful physical map. Despite these challenges, we were able to create a high quality physical map of *G. raimondii* as well as several additional cotton species.

The previously published Paterson reference sequence for *G. raimondii* was created using Sanger sequencing technology. While the quality of this initial assembly was excellent, short read (<1 Kb) based assemblies are often highly fragmented as they cannot efficiently span the size of repeat elements found within many plant genomes (cotton included). Scaffolding of the contigs derived from short reads can be accomplished using mate pair libraries (i.e., 2-3 Kb and 6-8 Kb mate pairs); however, in genomes with a significant fraction of repetitive elements, misassembly of the scaffolds is a real concern. Here we report the development of an improved *G. rainmondii* genome assembly using PacBio SMRT sequencing followed by hybrid sequence assembly using a BNG physical assembly. PacBio sequencing does have a higher error rate when calling bases, but this is overcome with deep read coverage [24]. Our assembly produced a highly contiguous assembly with an N50 of nearly 7 Mb. Such highly contiguous assemblies allow for the first time the investigation of large scale SVs in the cotton genome. Moreover, when the PacBio sequence assembly was used in

conjunction with the BNG physical map, we were able to analyze large-scale regions of the genome for SV as well as to perform assembly validation and accuracy checks.

Hybrid Assembly Improvements

One of the major advantages of the BNG technology is its utility in scaffolding sequence assemblies to improve their continuity [25][26]. The molecule maps (>150 Kb) generated using BNG technologies can span large repeat regions of the genome which are otherwise intractable with conventional sequencing technologies. Such so-called "hybrid assembles" often dramatically improve overall contiguity as shown from improved N50 and L50 statistics. Not only can the hybrid scaffolder improve the assembly by scaffolding contigs across repetitive regions, but it can also identify misassemblies in the existing NGS assembly (e.g., satellite repeats that the NGS assembly has collapsed). Such areas of conflict can be simply annotated or auto-reorganized to produced more parsimonious assemblies. Hybrid scaffolding of the G. *rainmondii* assembly reduced the number of contigs significantly (from 2,379 contigs to 43 scaffolds + 5 unscaffolded contigs), and N50 and average contig length were both dramatically increased (6.8 Mb to 25.8 Mb and 3.5 Mb to 18.3 Mb respectively). It is worth noting why the increase in N50 is not proportional to the increase in contig number. The Hybrid Scaffolding software first determines if each read has enough matching labels for its length to be useful. If it does not, the contig is not included in the scaffolding process. Many of the original small contigs were thrown out before scaffolding. At the end of the second round of scaffolding 100 contigs had been used

from the original Pacbio assembly. This also explains why, while the scaffolding software was making joins between contigs the length of the overall assembly went down.

Structural Variants as a Tool for Genomic Research

SVs are large-scale variations in the genome either between species or between individuals within a species. SVs are also one of the largest source of genetic variation within species [27]. We show that SVs are also useful in determining the validity of an assembly. More specifically, the concordance between physical (e.g., BNG) and NGS assemblies for the placement of SVs provides evidence of a correctly assembled genome. Relative to the BNG physical assembly, we identified over 777 SVs in the Paterson Sanger and PacBio assemblies, of which 75% were common between the assemblies. Because the SVs were detected by one physical mapping technology (BNG) copmared to two different sequencing technologies (PacBio and Sanger), they are highly supported and unlikely to be spurious. Because the comparison of the *G. raimondii Paterson* Sanger sequence to the *G. raimondii* BNG map was 75% similar to the comparison of the *G. raimondii* PacBio sequence to the *G. raimondii* BNG assembly, and because the two sequence assemblies were made using the same accession of *G. raimondii*, the BNG assembly seems to be an accurate representation of the genomic structure of the D genome of cotton. It also indicates that the two sequence assemblies are likely structurally accurate. The fact that the PacBio and Paterson Sanger assemblies are not 100% identical relative to SVs is likely a reflection of the shorter N50 of the contigs derived from the Paterson Sanger assembly.

Creation of a Novel Phylogeny Based on Structural Homology

SVs play a large part in evolution and speciation [28]. Oftentimes, species within a single genus differ only by one or a few large-scale changes in genome structure. Indeed, one of the earliest events associated with speciation are postzygotic barriers, which are often large SVs (e.g., translocations, inversions, and chromosome fusions or fissions) which disrupt mitotic and meiotic division leading to hybrid breakdown and sterility [29]. Similar to single nucleotide polymorphisms (SNPs), SVs are conserved over time and between related species and can be a target of natural selection [9]. Thus, SVs are potentially valuable for discerning phylogenetic relationships among related species. While several researchers have reported phylogenies for *Gossypium* based on genetic and phenotypic markers, here we report the first phylogeny for the genus using large-scale genomic structural variants, identified through physical map assemblies using BNG technology. While there are slight differences in the placement of the species in the SV-based phylogeny, the phylogeny is highly similar to those based on sequence polymorphism and/or morphological markers. The similarity between the SV and sequence phylogenies suggests that structural variants can be used to accurately build phylogenies and that they may represent a secondary method for validating existing sequence-based phylogenies. Indeed, the differences found in SV-based phylogenies may provide, upon further research, additional and potentially substanital insight into evolution and speciation.

LITERATURE CITED

1. Wendel, J. F., Clark Cronn, R., Clark, R., and Cronn, R. C. “Polyploidy and the Evolutionary History of Cotton Recommended Citation POLYPLOIDY AND THE EVOLUTIONARY HISTORY OF COTTON” Available at http://lib.dr.iastate.edu/bot_pubs

2. Fang, L., Gong, H., Hu, Y., Liu, C., Zhou, B., Huang, T., Wang, Y., Chen, S., Fang, D. D., Du, X., Chen, H., Chen, J., Wang, S., Wang, Q., Wan, Q., Liu, B., Pan, M., Chang, L., Wu, H., Mei, G., Xiang, D., Li, X., Cai, C., Zhu, X., Chen, Z. J., Han, B., Chen, X., Guo, W., Zhang, T., and Huang, X. “Genomic Insights into Divergence and Dual Domestication of Cultivated Allotetraploid Cottons” *Genome Biology* 18, no. 1 (2017): 33. doi:10.1186/s13059-017-1167-5, Available at http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1167-5

3. Renny-Byfield, S., Page, J. T., Udall, J. A., Sanders, W. S., Peterson, D. G., Arick, M. A., Grover, C. E., and Wendel, J. F. “Independent Domestication of Two Old World Cotton Species” *Genome Biology and Evolution* 8, no. 6 (2016): 1940–1947. doi:10.1093/gbe/evw129, Available at https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evw129

4. Zhang, H.-B., Li, Y., Wang, B., and Chee, P. W. “Recent Advances in Cotton Genomics.” *International journal of plant genomics* 2008, (2008): 742304. doi:10.1155/2008/742304, Available at http://www.ncbi.nlm.nih.gov/pubmed/18288253

5. Zhu, Y.-X. and Li, F.-G. “The *Gossypium Raimondii* Genome, a Huge Leap Forward in Cotton Genomics” *Journal of Integrative Plant Biology* 55, no. 7 (2013): 570–571. doi:10.1111/jipb.12076, Available at http://doi.wiley.com/10.1111/jipb.12076

6. Lin, L., Pierce, G. J., Bowers, J. E., Estill, J. C., Compton, R. O., Rainville, L. K., Kim, C., Lemke, C., Rong, J., Tang, H., Wang, X., Braidotti, M., Chen, A. H., Chicola, K., Collura, K., Epps, E., Golser, W., Grover, C., Ingles, J., Karunakaran, S., Kudrna, D., Olive, J., Tabassum, N., Um, E., Wissotski, M., Yu, Y., Zuccolo, A., ur Rahman, M., Peterson, D. G., Wing, R. A., Wendel, J. F., and Paterson, A. H. “A Draft Physical Map of a D-Genome Cotton Species (Gossypium Raimondii)” *BMC Genomics* 11, no. 1 (2010): 395. doi:10.1186/1471-2164-11-395, Available at http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-395

7. Clouse, J. W., Adhikary, D., Page, J. T., Ramaraj, T., Deyholos, M. K., Udall, J. A., Fairbanks, D. J., Jellen, E. N., and Maughan, P. J. “The Amaranth Genome: Genome, Transcriptome, and Physical Map Assembly” *The Plant Genome* 9, no. 1 (2016): 0. doi:10.3835/plantgenome2015.07.0062, Available at https://dl.sciencesocieties.org/publications/tpg/abstracts/9/1/plantgenome2015.07.0062

8. Jiao, W.-B., Garcia Accinelli, G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., Willing, E.-M., Piednoel, M., Woetzel, S., Madrid-Herrero, E., Huettel, B., Hümann, U., Reinhard, R., Koch, M. A., Swan, D., Clavijo, B., Coupland, G., and Schneeberger, K. “Improving and Correcting the Contiguity of Long-Read Genome Assemblies of Three Plant Species Using Optical Mapping and Chromosome Conformation Capture Data.”

Genome research (2017): gr.213652.116. doi:10.1101/gr.213652.116, Available at http://www.ncbi.nlm.nih.gov/pubmed/28159771

9. Chain, F. J. J. and Feulner, P. G. D. “Ecological and Evolutionary Implications of Genomic Structural Variations.” *Frontiers in genetics* 5, (2014): 326. doi:10.3389/fgene.2014.00326, Available at http://www.ncbi.nlm.nih.gov/pubmed/25278961

10. Fan, S. and Meyer, A. “Evolution of Genomic Structural Variation and Genomic Architecture in the Adaptive Radiations of African Cichlid Fishes.” *Frontiers in genetics* 5, (2014): 163. doi:10.3389/fgene.2014.00163, Available at http://www.ncbi.nlm.nih.gov/pubmed/24917883

11. Chakraborty, M., VanKuren, N. W., Zhao, R., Zhang, X., Kalsow, S., and Emerson, J. J. “Hidden Genetic Variation Shapes the Structure of Functional Elements in Drosophila” *Nature Genetics* 50, no. 1 (2018): 20–25. doi:10.1038/s41588-017-0010-y, Available at http://www.nature.com/articles/s41588-017-0010-y

12. Iskow, R. C., Gokcumen, O., and Lee, C. “Exploring the Role of Copy Number Variants in Human Adaptation.” *Trends in genetics : TIG* 28, no. 6 (2012): 245–57. doi:10.1016/j.tig.2012.03.002, Available at http://www.ncbi.nlm.nih.gov/pubmed/22483647

13. “Gossypium Raimondii (D5) Genome JGI Assembly v2.0 (Annot v2.1) | Cottongen” Available at https://www.cottongen.org/analysis/50

14. “Bionano Prep™ Plant Tissue DNA Isolation Base Protocol” Available at https://bionanogenomics.com/wp-content/uploads/2018/02/30068-Rev-D-Bionano-Prep-Plant-Tissue-DNA-Isolation-Protocol.pdf

15. “Bionano Prep ™ Labeling -NLRS Protocol” Available at https://bionanogenomics.com/wp-content/uploads/2017/07/30024-Rev-J-Bionano-Prep-Labeling-NLRS-Protocol.pdf

16. “Irys® User Guide” Available at https://bionanogenomics.com/wp-content/uploads/2017/01/30047-Rev-B-Irys-User-Guide.pdf

17. Dellaporta, S. L., Wood, J., and Hicks, J. B. “A Plant DNA Minipreparation: Version II” *Plant Molecular Biology Reporter* 1, no. 4 (1983): 19–21. doi:10.1007/BF02712670, Available at http://link.springer.com/10.1007/BF02712670

18. “Procedure & Checklist >20 Kb Template Preparation Using BluePippin Size-Selection System (15 - 20 Kb Cutoff) for Sequel Systems - PacBio” Available at https://www.pacb.com/documentation/procedure-checklist-20-kb-template-preparation-using-bluepippin-size-selection-system-15-20-kb-cutoff-sequel-systems/

19. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., and Phillippy, A. M. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation” *bioRxiv* (2016): 071282. doi:10.1101/071282, Available at https://www.biorxiv.org/content/early/2016/08/24/071282

20. Quinlan, A. R. and Hall, I. M. “BEDTools: A Flexible Suite of Utilities for Comparing

Genomic Features.” *Bioinformatics (Oxford, England)* 26, no. 6 (2010): 841–2. doi:10.1093/bioinformatics/btq033, Available at http://www.ncbi.nlm.nih.gov/pubmed/20110278

21. “R: The R Project for Statistical Computing” Available at https://www.r-project.org/

22. Schliep, K. P. “Phangorn: Phylogenetic Analysis in R” *Bioinformatics* 27, no. 4 (2011): 592–593. doi:10.1093/bioinformatics/btq706, Available at https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq706

23. 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, Available at https://www.sciencedirect.com/science/article/pii/S1055790315001669

24. Korlach, J. “Perspective - Understanding Accuracy in SMRT Sequencing” Available at www.pacb.com

25. Shelton, J. M., Coleman, M. C., Herndon, N., Lu, N., Lam, E. T., Anantharaman, T., Sheth, P., and Brown, S. J. “Tools and Pipelines for BioNano Data: Molecule Assembly Pipeline and FASTA Super Scaffolding Tool” *BMC Genomics* 16, no. 1 (2015): 734. doi:10.1186/s12864-015-1911-8, Available at http://www.biomedcentral.com/1471-2164/16/734

26. 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

27. Hurles, M. E., Dermitzakis, E. T., and Tyler-Smith, C. “The Functional Impact of Structural Variation in Humans.” *Trends in genetics : TIG* 24, no. 5 (2008): 238–45. doi:10.1016/j.tig.2008.03.001, Available at http://www.ncbi.nlm.nih.gov/pubmed/18378036

28. Chain, F. J. J. and Feulner, P. G. D. “Ecological and Evolutionary Implications of Genomic Structural Variations.” *Frontiers in genetics* 5, (2014): 326. doi:10.3389/fgene.2014.00326, Available at http://www.ncbi.nlm.nih.gov/pubmed/25278961

29. Stelly, D. M., Kautz, K. C., and Rooney, W. L. “Pollen Fertility of Some Simple and Compound Translocations of Cotton” *Crop Science* 30, no. 4 (1990): 952. doi:10.2135/cropsci1990.0011183X003000040041x, Available at https://www.crops.org/publications/cs/abstracts/30/4/CS0300040952

FIGURES

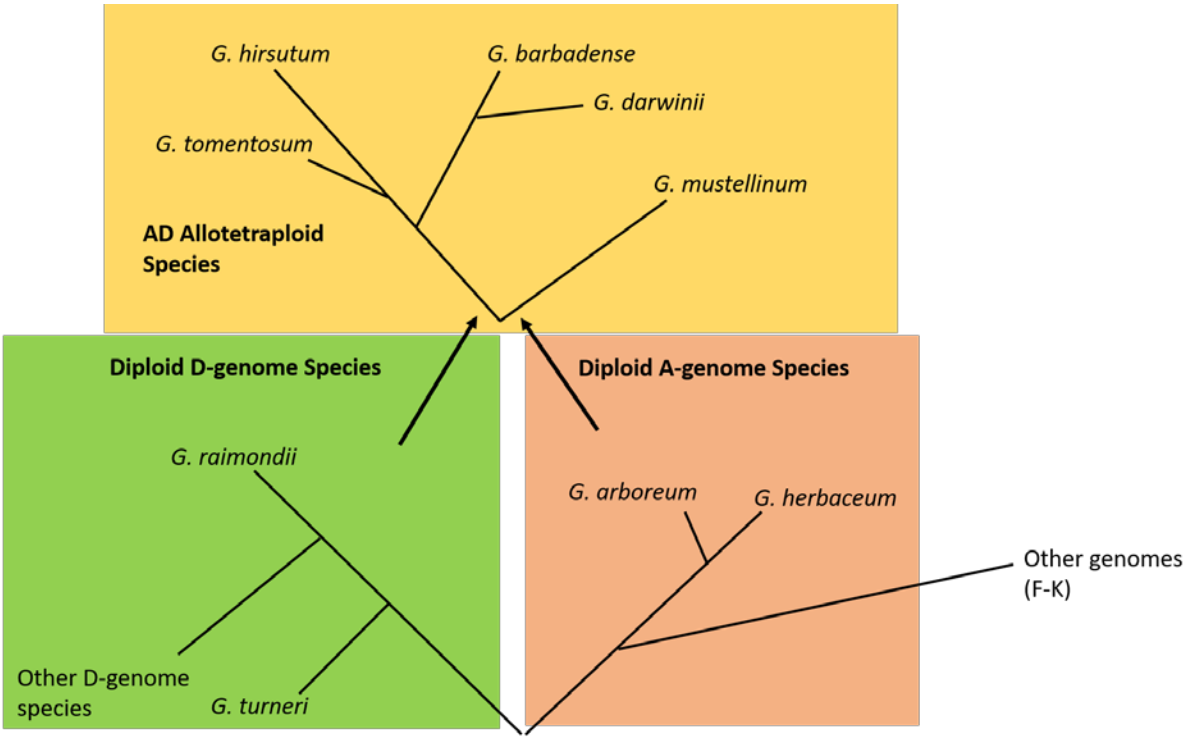

Figure 1. *A phylogeny of Gossypium created using genomic sequence data. The green box represents the D-genome diploid side of the phylogeny and the red box represents the A-genome diploid side. The yellow box represents everything post polyploidization (circa 1.2 MYA)[1].*

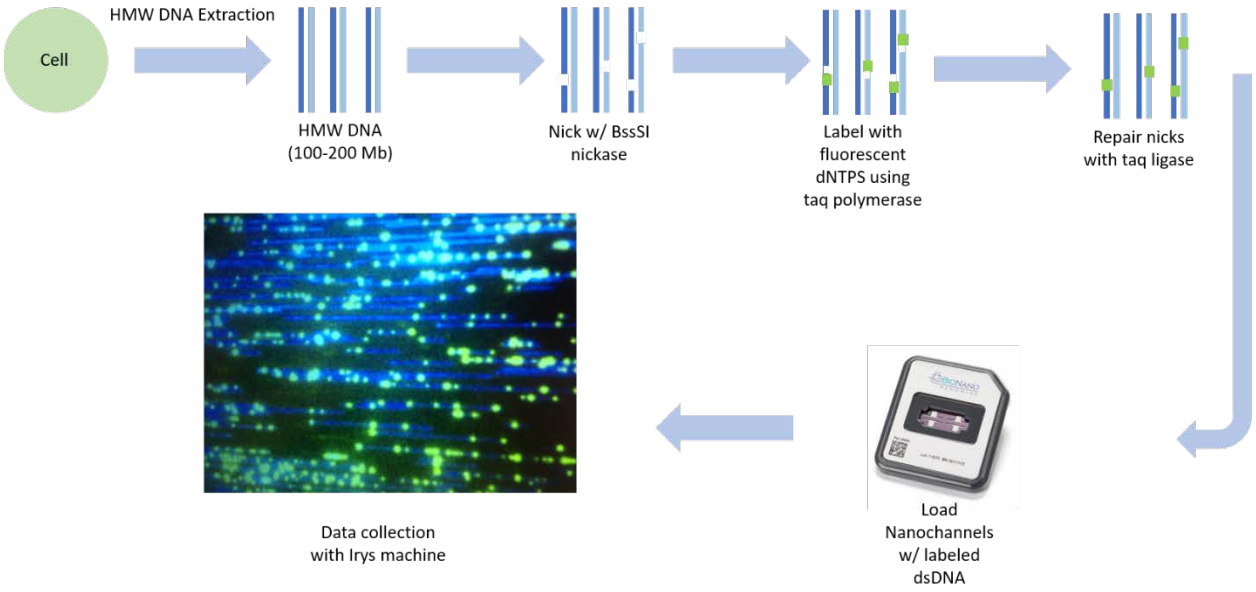

Figure 2. *A workflow depicting the process of data collection used for physical map creation. HMW DNA was extracted from nuclei from extracted from leaf tissue. The DNA was nicked with an enzyme and fluorescent labels are inserted in the nick sites. The nick sites are repaired with ligase. The DNA is then loaded on the Irys chip and onto the Irys machine. The machine then images the molecules and detects label positions and molecule length.*

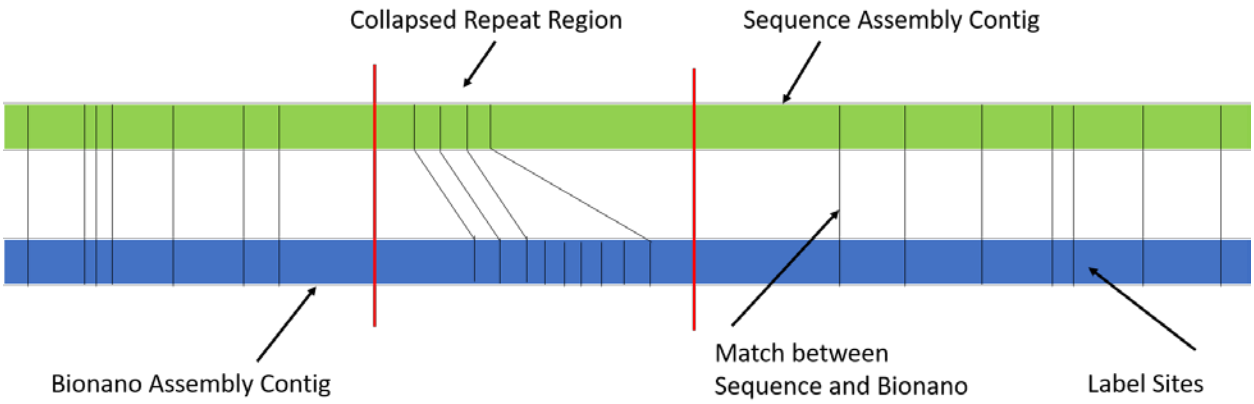

Figure 3. *An example alignment of a sequence assembly with a Bionano physical assembly. The region between the red lines shows a collapsed repeat in the sequence assembly. The green line represents the sequence assembly contig while the blue line represents the Bionano assembly contig. Label sites are where we have used a nicking enzyme to nick the DNA and insert fluorescently labeled nucleotides, this is done in vitro for the Bionano DNA and it's done in silico for the sequence assembly. This diagram shows a region where the Bionano assembly is showing a collapse of a repetitive region of the genome in the sequence assembly.*

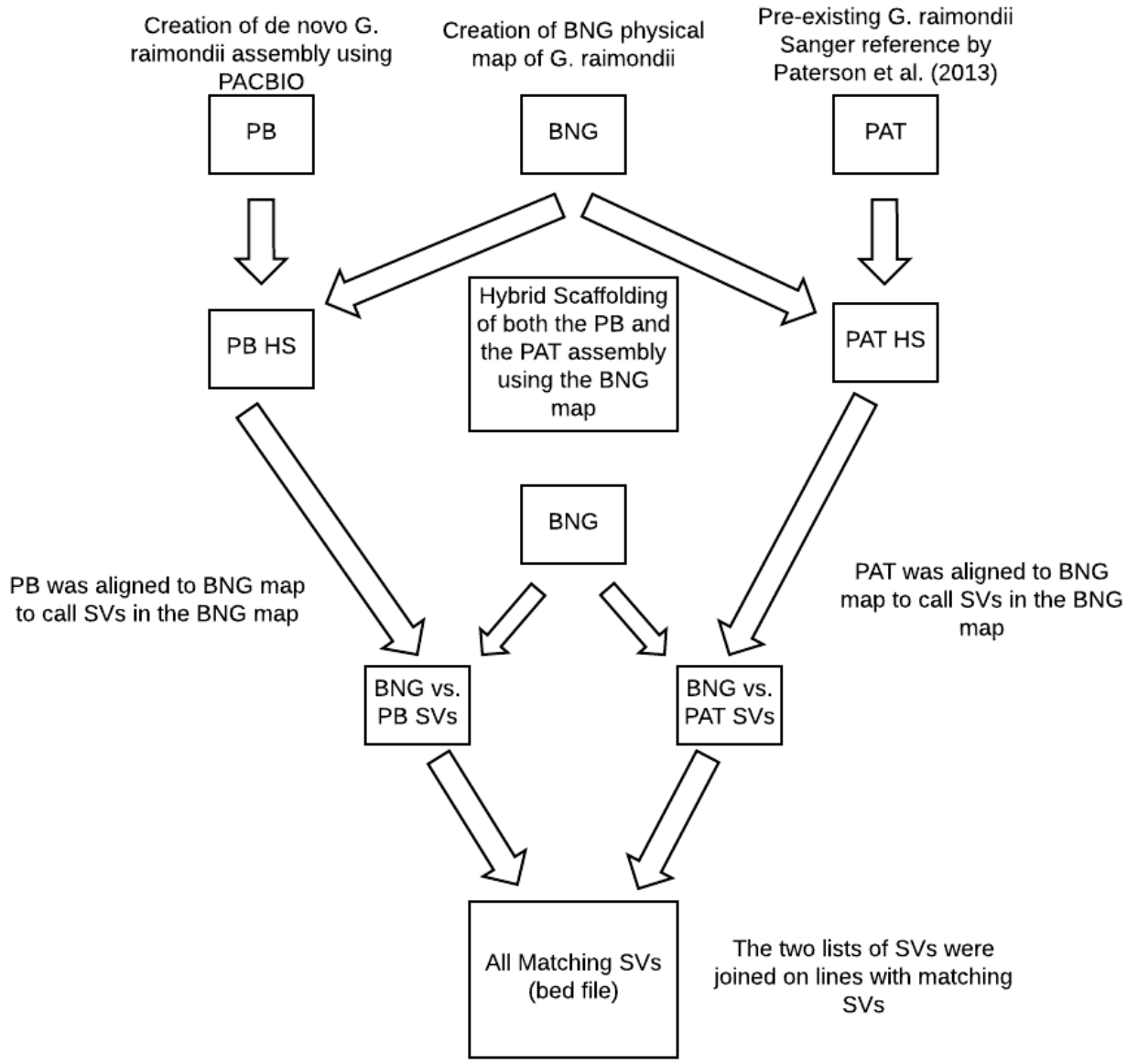

Figure 4. *A flowchart showing the process of creation, improvement and validation of three assemblies. Three assemblies were created. The BNG map was used to scaffold the other two and then SVs called using both sequence assemblies separately were compared to validate the assemblies. If the two lists have a high amount of similar SVs the assemblies are likely to be correct.*

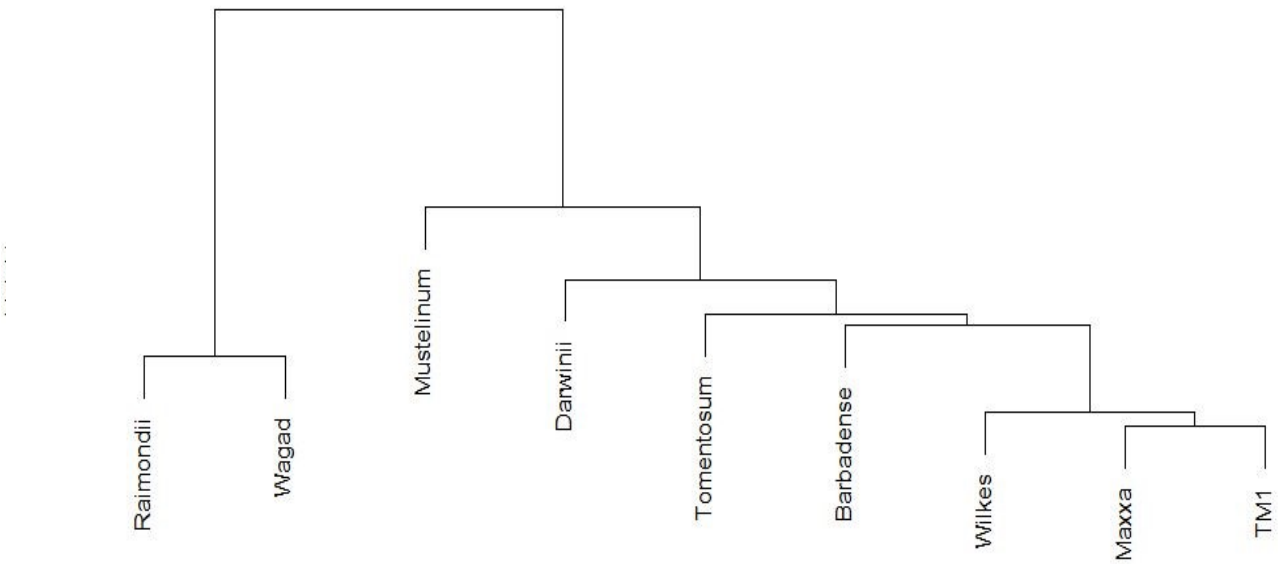

Figure 5. *A phylogeny, newly created using the data generated during this project. The factor used in determining the relationships between species was similar and differing structural variations in the respective genomes.*

TABLES

Table 1. *Assembly Statistics for the Nt.BssSI BNG map of G. raimondii before and after hybrid scaffolding*

	After Initial Map Assembly	Scaffolded Hybrid Assembly
N. Genome Maps	1,020	107
Total Genome Map Len (Mb)	737.645	811.903
Avg. Genome Map Len (Mb)	0.723	7.588
Median Genome Map Len(Mb)	0.523	5.865
Genome Map N50 (Mb)	0.936	10.958
Max Genome Map Len (Mb)	5.6	35.38
Total Ref Len (Mb)	752.729	N/A
Total Genome Map Len / Ref Len	0.980	N/A
N. Genome Maps total align	979 (96%)	N/A
Total Aligned Len (Mb)	634.486	N/A
Total Aligned Len / Ref Len	0.843	N/A
Total Unique Aligned Len (Mb)	611.004	N/A
Total Unique Aligned Len / Ref Len	0.812	N/A

Table 2. *Assembly Statistics for the Nt.BspQI BNG map of G. raimondii before and after hybrid scaffolding*

	Initial Map Assembly Statistics:	Scaffolded Hybrid Assembly Statistics
N Genome Maps	413	89
Total Genome Map Len (Mb)	758	841
Avg. Genome Map Len (Mb)	1.835	9.45
Median Genome Map Len(Mb)	1.359	6.77
Genome Map n50 (Mb)	2.6	13.873
Max Genome Map Len (Mb)	9.8	38.14
Total Ref Len (Mb)	749.575	N/A
Total Genome Map Len / Ref Len	1.011	N/A
N Genome Maps total align	385 (93%)	N/A
Total Aligned Len (Mb)	570.885	N/A
Total Aligned Len / Ref Len	0.762	N/A
Total Unique Aligned Len (Mb)	563.329	N/A
Total Unique Aligned Len / Ref Len	0.752	N/A

Table 3. *PacBio sequencing statistics before and after assembly of the sequenced data*

	PacBio Sequence Data Before Assembly	PacBio Sequence Data After Assembly
# of reads/contigs	N/A	2,379
Read/Contig N50 (Mb)	0.0225	4.9
Mean Read/Contig Length (Mb)	0.014	0.413
Total base pairs in Reads/Contigs (Mb)	18,810	982.72

***Table 4**. Hybrid Scaffold statistics of the PacBio sequence data. The data was scaffolded twice, first with the Nt.BssSI BNG map and then the output of that scaffolding was scaffolded with the Nt.BspQI BNG map.*

	PacBio Sequence Assembly Scaffolded with BNG Nt.BssSI Map (Step 1)	Output Fasta from Step 1 Scaffolded with BNG Nt.BspQI Map (Step 2)
# of scaffolds	106	43
Scaffold N50 (Mb)	13.1	25.8
Mean Scaffold Length (Mb)	7.3	18.3
Max Scaffold Length (Mb)	37.0	57.5
Total Sequence Assembly Length After Scaffolding (Mb)	778.1	785.2

Table 5. *A table containing how many of each type of SV was called in each of two BNG assemblies of G. raimondii.*

	Nt.BssSI BNG assembly	Nt.BspQI BNG assembly
Insertion	449	611
Deletion	46	101
Translocation	126	150
Inversion	0	1
Complex	189	1,007
End	373	143

Table 6. *Statistics of the eight maps that were compared along with G. raimondii to create a phylogeny.*

	# of contigs	Total map length (Gb)	Average contig length (Mb)	Contig N50 (Mb)
G. herbaceum (wagad)	1,838	1.5	0.852	1.2
G. hirsutum (maxxa)	2,087	2.3	1.1	1.6
G. hirsutum (TM1)	2,196	2.2	0.995	1.3
G. barbadense	1,711	2.1	1.2	1.8
G. tomentosum	3,176	1.9	0.618	0.740
G. darwinii	4,939	2.0	0.406	0.473
G. stevensii (wilkes)	3,246	2.1	0.666	0.846
G. mustelinum	4,622	1.8	0.397	0.438

Table 7. *Two bed files, one containing SVs called in the BNG map by the PacBio sequence and one containing SVs called in the BNG map by the Paterson Sanger reference, were analyzed to find all the SVs they had in common and a new bed file was created containing only these shared SVs. The data in this table is a depiction of the contents of this new bed file.*

	Insertions	Deletions
Total Number	288	489
Average SV size (Kb)	24.58	22.22
Max SV size (Kb)	2.08	3.45
Min SV size (Kb)	667.44	160.78

Table 8. *The phylogeny that was created by finding regions of structural homology was created from a similarity matrix. This table summarizes the data found in that similarity matrix.*

	Total regions	Regions with similarity to G. hirsutum (TM1)	Similarity Score
G. raimondii	43,402	2,057	0.0474
G. herbaceum	43,402	7,664	0.1766
G. mustelinum	43,402	21,115	0.4865
G. darwinii	43,402	27,751	0.6394
G. tomentosum	43,402	29,530	0.6804
G. barbadense	43,402	30,272	0.6975
G. stevensii	43,402	34,964	0.8056
G. hirsutum (Maxxa)	43,402	35,411	0.8159