2016-03-01

Improving Cotton Agronomics with Diverse Genomic Technologies

Aaron Robert Sharp
Brigham Young University - Provo

Improving Cotton Agronomics with

Diverse Genomic Technologies

Aaron Robert Sharp

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

Master of Science

Joshua A. Udall, Chair
Peter J. Maughan
Stephen R. Piccolo

Department of Plant and Wildlife Sciences

Brigham Young University

March 2016

ABSTRACT

Improving Cotton Agronomics with Diverse Genomic Technologies

Aaron Robert Sharp
Department of Plant and Wildlife Sciences, BYU
Master of Science

Agronomic outcomes are the product of a plant's genotype and its environment. Genomic technologies allow farmers and researchers new avenues to explore the genetic component of agriculture. These technologies can also enhance understanding of environmental effects. With a growing world population, a wide variety of tools will be necessary to increase the agronomic productivity.

Here I use massively parallel, deep sequencing of RNA (RNA-Seq) to measure changes in cotton gene expression levels in response to a change in the plant's surroundings caused by conservation tillage. Conservation tillage is an environmentally friendly, agricultural practice characterized by little or no inversion of the soil prior to planting. In addition to changes in cotton gene expression and biological pathway activity, I assay the transcriptional activity of microbial symbiotes living in and around the cotton roots. I found a large degree of similarity between cotton individuals in different treatments. However, under conventional disk tillage I did find significantly greater activity of cotton phosphatase and sulfate transport genes, as well as greater abundance of the microbes *Candidatus Burkholderia brachynathoides* and *Arthrobacter* species L77.

This study also includes the use of high-throughput physical mapping of DNA to examine the genomic structure of a wild cotton species, *Gossypium raimondii*, which is closely related to the economically significant crop species *Gossypium hirsutum*. This technology characterizes genomic regions by assembling large input DNA molecules labeled at restriction enzyme recognition sites. I created an efficient algorithm and generated 812 whole-genome assemblies from two datasets. The best of these assemblies allowed us to detect 3,806 potential misassemblies in the current release of the *G. raimondii* genome sequence assembly.

Keywords: cotton, agriculture, conservation tillage, RNA-Seq, subgenome expression bias, BioNano physical mapping, genome sequence assembly

ACKNOWLEDGEMENTS

I thank Dr. Joshua Udall for his work as my advisor, as well as Dr. Jeff Maughan and Dr. Stephen Piccolo for advice and feedback as committee members. I thank members of the Udall lab, particularly Chris Hanson, Carrie Evans, Alex Freeman, Erik Winterholler, Susan Dudley and David Amorim for tireless efforts in data collection, insightful conversations about data analysis, and for beta testing my software. I thank Dr. Todd Campbell for his collaboration and feedback on my RNASeq work. I thank the employees of BioNano Genomics, particularly Weiping Wang and Palak Sheth, for answering technical questions regarding DNA extraction and the use of their software. I thank Sara Greenfield and Zach Liechty for maintaining cotton plants in the greenhouse and collecting tissue for DNA extraction. I thank Lindsay Chaney for her insight into effective data analysis and visualization. I thank several anonymous reviewers for helpful insight into the portions of this work that have been submitted for publication. This work was funded by the NSF and Cotton Inc. I thank my wife, Heather Sharp, for all of her support and patience.

Table of Contents

List of Tables

List of Figures

Chapter 1: Sequencing, conservation tillage, and the rhizosphere

Introduction

Massively parallel sequencing of RNA

In the last decade, the cost of DNA sequencing has fallen dramatically (Wetterstrand, 2015), accelerating genomic insights into organismal biology and evolution. Massively parallel sequencing technology (sequencing or MPS) can provide several types of biological insight. It can be used to elucidate the coding sequences and regulatory elements of genes (Anderson et al., 1981; Fleischmann et al., 1995; Myers et al., 2000). It can be used to discover genomic polymorphisms, which in turn can help researchers predict changes in molecular pathways. In association and linkage studies, causal variants for genetic disorders and phenotypically significant genomic regions for desirable agronomic traits can be discovered. With some additional bench work, sequencing technology can be used to discover changes in methylation patterns (Lister & Ecker, 2009), locations of DNA binding elements (Mardis, 2007), and even the relative localization of chromosomes in the cell (Lieberman-Aiden et al., 2009). Another specialized application of sequencing called RNA-Seq uses MPS to count and assign RNA fragments to known genes. From these counts, one can infer gene expression levels and their changes in response to environmental stimuli (Garber, Grabherr, Guttman, & Trapnell, 2011).

Conservation tillage

Increasing demands for agricultural productivity have prompted a closer look at the impact of traditional agricultural practices on cropland yields and sustainability. Conventional disk tillage (DT) plays an important role in modern agriculture, but it comes with certain environmental consequences, such as high fuel usage, and increased soil erosion. An alternative to DT is conservation tillage (CT), either low- or no-tillage practices. Conservation tillage reduces fossil fuel consumption by cultivation machinery, and decreases erosion and runoff through improved soil structure and conservation of crop residues on the soil surface (Soane et al., 2012). The use of CT has increased in recent years because of these benefits, as well as its potential to improve soil load bearing capacity (Soane et al., 2012), increase soil organic matter content and decrease soil CO_2 emissions (Novak, Bauer, & Hunt, 2007). It should be noted, however, that a potential negative environmental consequence of CT is the increased need for herbicide usage.

Conservation tillage has generally been shown to have variable effects in different soils, crops, and environments. For example, flat fields derive less benefit from soil conservation than hilly fields do, where erosion and runoff are more significant problems (Sojka, Karlen, & Busscher, 1991). The focus of our study is mostly flat, loamy fine sands typical of the lower Coastal Plain region in the United States. The crop of interest for our study is *Gossypium hirsutum L.*, upland cotton. Out conservation tillage site was established over thirty years prior to data collection (Hunt, Matheny, & Wollum, 1985; Sojka et al., 1991).

Preservation of the native soil structure can improve cropland productivity and sustainability, since non-compacted soil facilitates water infiltration, root penetration, and nutrient retention (Lachnicht, Hendrix, Potter, Coleman, & Crossley, 2004). For example, thirty years of data collected from CT and DT soybean fields indicated that soils under CT management had more stable soil macroporosity and higher levels of organic carbon than soils managed with DT (So, Grabski, & Desborough, 2009). Additionally, cambisol soil in CT fields has been found to contain higher concentrations of available phosphorous and organic material than DT fields (Horacek, Kolar, Cechova, & Hrebeckova, 2008).

Recent work in CT acknowledges the advantages of reduced need for fossil fuel consumption, manual labor, machine maintenance, and irrigation. However, it also notes some flaws in CT. Crops grown under CT may require more nitrogen fertilizer (Yin & Main, 2015). Although CT reduces surface runoff, the improved penetrance of soil under CT may lead to increased subsurface runoff (Potter, Bosch, & Strickland, 2015).

The rhizosphere

The rhizosphere, is a region of cropland soil characterized by root secretions and soil biota. Soil biota activity can play an important role in plant health and crop yield through symbiosis and nutrient sequestration. Analysis of biota communities in both DT and CT fields has shown that microbial and mycorrhizal activity is higher in CT systems, for example, in increased soluble carbon accumulation in the soil (S. Zhang, Li, Lu, Zhang, & Liang, 2013). This study suggests that the rhizosphere is significantly influenced by tillage management practice.

Both soil characteristics and the rhizosphere can have an impact on plant health and yield. For example, gene transcription level changes were observed in *Gossypium hirsutum* in response to soil structure (loose or compact) (Klueva et al., 2000) and hydration level (Bowman et al., 2013). Other studies have shown that soil quality has a strong influence on crop yield (Lachnicht et al., 2004). As for the rhizosphere, *Zea mays* roots showed increased expression of enzymes involved in the recruitment and infection process of a beneficial fungus following soil inoculation of the fungal symbiont (Fries, Pacovsky, & Safir, 1996). It has also been suggested that due to microbe activity, crops in CT fields are expected to resist environmental and nutrient stresses better than crops grown under DT (Carpenter-Boggs, Stahl, Lindstrom, & Schumacher, 2003). Therefore the impact of tillage practice on crop plants is likely very substantial, both through soil characteristics, and through microbial activity in the rhizosphere. Recent work particularly focusing on the rhizosphere under CT found an increase in actinomycetes, mycorrhizae fungi, and total organic carbon under CT (Mbuthia et al., 2015).

In this study, we examined the effect of conservation tillage on gene expression levels in cotton and its associated microbial community using RNA-Seq. To do so, we exposed several individuals of a single cotton genotype to fields where either DT or CT have been practiced continuously for 30 years.

Methods

Field trial and sample collection

Our research site at the Pee Dee Research and Education Center near Florence, South Carolina is located on a Norfolk loamy sand soil (an acrisol or fine-loamy, siliceous, thermic paleudult). Eight plots make up a long term research site where surface-disked (conventional tillage, DT) and non-disked (conservation tillage, CT) treatments were first established in 1978 (Novak et al., 2007). The plots were seeded with a single genotype of *Gossypium hirsutum*, cv. Siokra-L23, on May 10, 2013. On July 9, 2013, sixteen individuals were selected randomly, two from a single row in each of the eight plots. They were excavated, and a single lateral root from each plant was flash frozen in liquid nitrogen and placed on dry ice. All samples were collected within one hour. RNA was extracted from washed, homogenized root tissue using the Spectrum^{TM} Plant Total RNA Kit (SIGMA-ALDRICH, USA), according to manufacturer's instructions, and prepared as single-end libraries using a TruSeq Kit (Illumina, USA). Sequencing was performed on an Illumina HiSeq 2000 (USA) at Oregon State University's Center for Genome Research and BioComputing. The data have been uploaded to the NCBI short read archive. Reads for CT-treated plants can be found using accession numbers SRR3225337, and SRR3225340-

SRR3225346. Reads for DT-treated plants can be found using accessions numbers SRR3225348-SRR3225355.

Differentially expressed genes

Raw data were trimmed using trimmomatic v0.35 (Bolger, Lohse, & Usadel, 2014). Illumina TruSeq2-SE adapters sequences were removed. Any leading and trailing bases with phred quality scores below five, and any six-bp regions with average scores below thirty, were trimmed from the ends of reads. Reads shorter than forty-bp after trimming were discarded. The trimmed reads were aligned to the *G. hirsutum* reference genome v1.1 (T. Zhang et al., 2015) using Tophat2 v2.0.7 (Johns Hopkins University, USA) with default parameters, except that the option --no-coverage-search was used in order to skip estimation of transcript isoform abundance. Samtools (Li et al., 2009) allowed us to count the number of reads per replicate per annotated gene. Read counts per gene were normalized by replicate as proportions of total trimmed reads using R version 3.1.0 (R Core Team, 2015). R was also used to calculate Pearson correlation coefficients between replicates, and perform complete-linkage clustering of replicates based on Euclidean distances. We used the R package EdgeR v3.4.2, which creates a generalized linear model to perform a principle component analysis using the 500 most informative genes, and to detect differentially expressed genes that were statistically significant (Robinson, McCarthy, & Smyth, 2010; Robinson & Oshlack, 2010). We used false discovery rate (Benjamini & Hochberg, 1995) to measure statistical significance, with a threshold designed to detect less than one false positive out of all the genes assayed. We also excluded genes with fold-changes between treatments that were less than two. Significant genes were BLASTed against the NCBI non-redundant (nr) protein database using BLASTX (Camacho et al., 2009) with default parameters, including a maximum of 100 hits. This version of BLAST searches for similarity between a nucleotide sequence and known proteins by first translating the query in all six possible reading frames.

Reads that did not align to the *G. hirsutum* genome were pooled from all replicates and assembled into a putative microtranscriptome using Trinity release 2014-07-17 (Grabherr et al., 2011). All assembled transcripts were mapped back to the *G. hirsutum* genome using Tophat2 as before. All transcripts with significant matches to the *G. hirsutum* genome and all transcripts shorter than 500 bp were removed. To quantify the abundance of transcripts per replicate, trimmed reads were mapped back to the assembled trasncriptome using Tophat2 as before, and counted using Samtools. Correlation between replicates at the microtranscriptome level and detection of significant differentially abundant taxa were calculated as above, except that read counts per microbial transcript were normalized by the total non-hirsutum reads per replicate. Significantly different transcripts were BLASTed against representative genomes of NCBI's Microbial database (Chen, Ye, Zhang, & Xu, 2015; Morgulis et al., 2008). Transcripts with no hits were not analyzed further, and transcripts with 100 hits were discarded as ambiguous or chimeric assemblies.

Subgenome bias

We also detected *G. hirsutum* genes with subgenome specific expression biases. To do this, we aligned trimmed reads to the *G. raimondii* reference genome (Paterson et al., 2012) using the SNP tolerant short read aligner, GSNAP v. 2015-07-23 (Wu & Nacu, 2010). We categorized aligned reads to either the A_t or D_t subgenome using PolyCat v. 1.8 (Page, Gingle, & Udall, 2013). Categorized reads were assigned to annotated *G. raimondii* genes under the assumption that head *G. raimondii* gene, and counts were normalized as proportions of total trimmed reads per replicate. EdgeR was used again to detect genes with significantly more A_t biased reads than

D_t biased reads, or vice versa. The significance threshold for this comparison was set at an FDR of 0.5.

Results

A total of 19,578.2 Megabase-pairs (Mbp) were sequenced. After trimming, 14,6178.0 Mbp were left in the dataset. Of these, 13,137.3 Mb aligned to the *G. hirsutum* genome (Figure 1). There was not a significant difference between replicates in the proportion of reads that were of cotton origin (Figure 1). The similarity between treatments in these metrics is expected, as substantial deviations would most likely only be caused by technical error during RNA extraction, amplification, or sequencing.

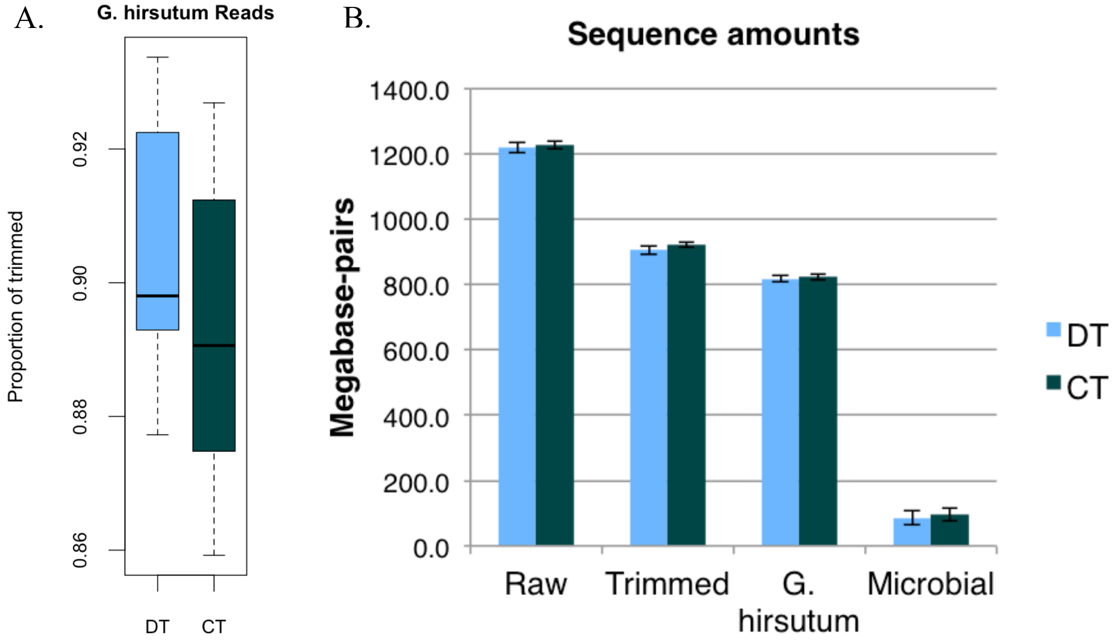

Figure 1: Distributions of RNA-Seq data under conventional disk tillage (DT) and conservation tillage (CT): (A) proportion of trimmed reads aligning to the *Gossypium hirsutum* reference genome in eight replicates per treatment (n=16). (B) Average amount of RNA per replicate sequenced, retained after trimming, successfully aligned to the *G. hirstutum* reference genome, and successfully aligned to the microtranscriptome assembly. Error bars represent standard deviation across replicates.

Cotton transcriptome

At the global level, the treatment showed little effect on the cotton transcriptome. Pearson correlation coefficients between replicates of DT ranged from 0.5 to almost 1.0. Correlation was much stronger within CT; however, correlations between replicates of opposite treatments were also high (Figure 2A). Clustering did not separate replicates by treatment (Figure 2B), and a principle component analysis shows similarity between replicates even between treatments (Figure 2C).

Despite the high degree of similarity, we detected seven genes with a fold change between treatments greater than two and an FDR value lower than 1.48e-5, which was calculated as one divided by the 70,478, the number of annotated *G. hirsutum* genes. All of these genes were more abundant under DT. Of these seven significant genes, three were on chromosome D5 (Gh_D05G0219, Gh_D05G0357, and Gh_D05G1276), and there was one each on chromosomes D11 (Gh_D11G0018), A2 (Gh_A02G1207), A5 (Gh_A05G0155), and A11 (Gh_A11G0020). All seven genes had 100 significant BLAST hits, which were manually examined to determine the function of the significant genes (Table 1).

Table 1: Significant differentially-expressed cotton genes with functions

Genes	**Protein function**
Gh_A11G0020[1], Gh_D11G0018[1]	Purple acid phosphatase
Gh_D05G1276	Inorganic pyrophosphatase
Gh_A05G0155[2], Gh_D05G0219[2]	Sulfate transporter
Gh_A01207, Gh_D05G0357	Unknown

[1] – Likely homoeolog pair
[2] – Likely homoeolog pair

Our subgenome bias analysis indicated that 12,772 genes showed statistically significant subgenome expression bias (FDR<0.5), of which about half, 6,378, showed greater abundance of A-subgenome transcripts.

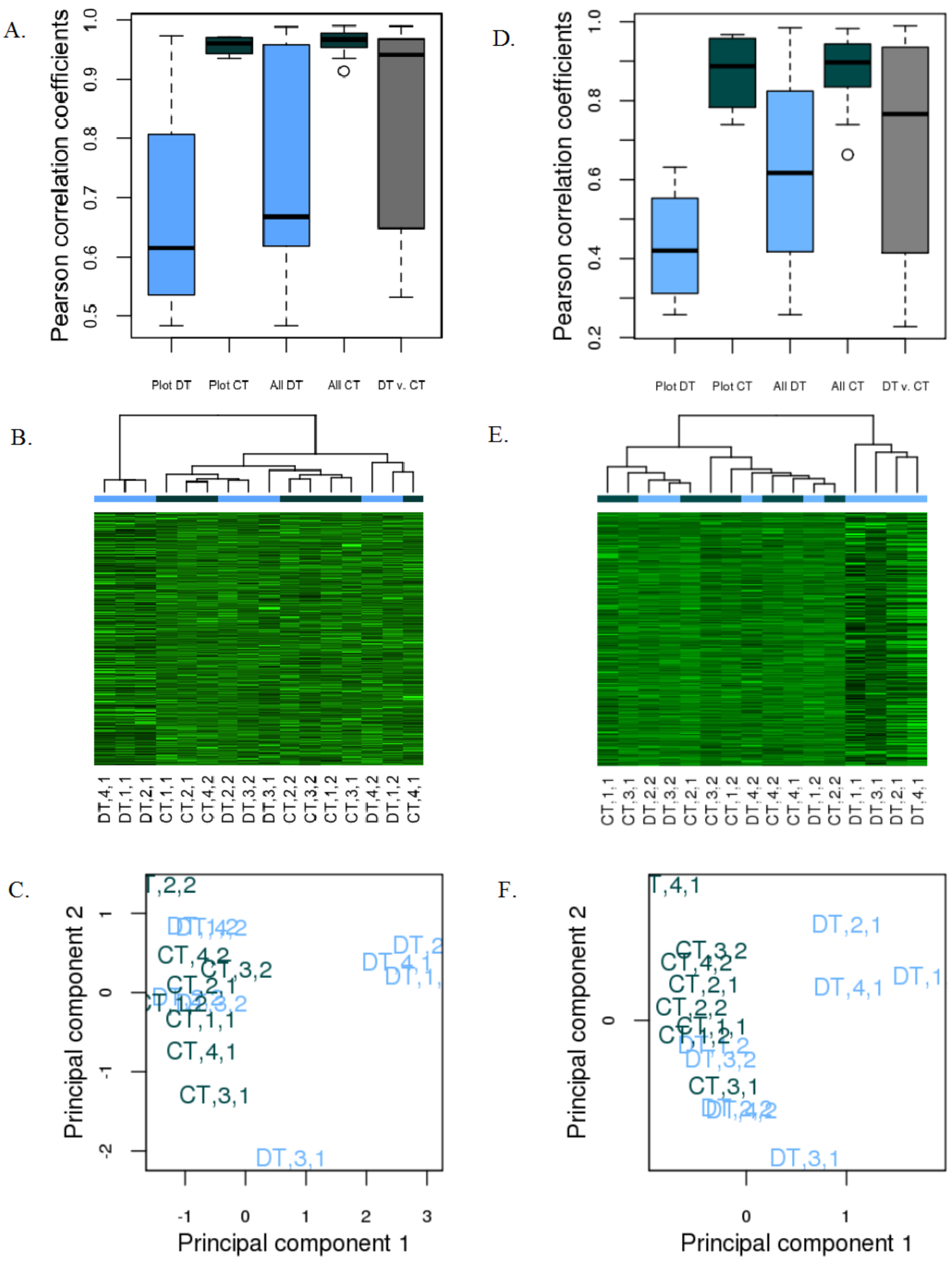

Figure 2: Considerably similarity between replicates in different tillage treatments, conventional disk (DT) or conservation tillage (CT), at the cotton transcriptome (A-C) and microtranscriptome (D-F) level. (A, D) Distribution of Pearson correlation coefficients between pairs of replicates in the same plot (columns 1 and 2, 4 comparisons per column), in the same treatment (columns 3 and 4, 28 comparisons per column), and in different treatments (column 5, 64 comparisons). (B, E) Replicate similarity clusters and gene expression levels. The bar designates replicates from DT (blue) or CT (green). Darker green in the heatmap denotes higher expression levels. The numbers following the treatment label describe the plot and sample number of each replicate. (C, F) Principal component analysis of replicates based on the 500 most informative genes.

Microtranscriptome

The microtranscriptome was assembled from 1,480.5 Mbp into 8,284 microbial transcripts (total length=2.5 Mbp; max length=3,551 bp, N50=293 bp). Only 133 transcripts aligned back to the *G. hirsutum* reference genome, and these ranged from 201 to 484 bp long. After filtering these and other transcripts shorter than 500 bp, we were left with 620 microbial transcripts. As we saw in the cotton transcriptome, there was considerable similarity between replicates of opposing treatments (Figure 2D-F).

Twenty-two transcripts were significantly more abundant under DT. Eleven (50.0%) of these had no significant BLAST hits and seven (31.8%) had 100 hits, and so they were excluded. The remaining four (18.2%) transcripts had between 8 and 22 hits, and for these we filtered out all but the most significant hit per transcript. Two matched the taxon *Candidatur Burkholderia brachynathoides*, and two matched *Arthrobacter* species (sp.) L77.

Discussion

Cotton transcriptome

Overall, there only a few significant differences between cotton individuals grown under DT and CT. The variation we did see between replicates did not appear to be driven exclusively by tillage. This agrees with a recent meta-analysis of CT studies, which found that many crops experience yield loss during the first few years of CT, but cotton was one of the few crops that did not (Pittelkow et al., 2015). Our results also suggest that gene expression variability was greater in individuals under DT than under CT. This may imply that CT reduces the impact of spatial variation.

The few differentially expressed cotton genes for which we were able to determine function were related to phosphatase activity and sulfate transport. All of these were more active under DT.

Phosphate is a major nutrient for plants, and modern agriculture relies on extensive use of external phosphate, generally mined as Rock Pi (Smit, Bindraban, Schröder, Conijn, & Van der Meer, 2009). Although it is not clear from the current analysis whether in the increased activity in purple acid phosphatase and inorganic pyrophosphatase is in response to a higher concentration of plant-available phosphate in DT soils, or phosphate starvation, the observation that conservation tillage has some impact on cotton's ability to utilize phosphate is worth additional consideration. Sulfate is another plant nutrient. Although uptake and metabolism of sulfate is affected by phytohormones (Koprivova & Kopriva, 2016), the observation that no other phytohomrone-controlled genes show significant differential expression implies that tillage has some impact on the availability of sulfate in the soil. Again, whether CT increases or decreases plant-available sulfate is unclear from this study. However, the relationship, once elucidated, would be agriculturally significant.

It is interesting to note that the pair of genes corresponding to purple acid phosphatases are likely homoelogs, since they are located in approximately the same position of homoeologous chromosomes and share a common BLAST hit with a single *G. raimondii* gene. The pair of genes assigned the function of sulfate transporter are also likely homoeologs. This implies that, at least in some cases, homoeologous genes retain similar control elements and patterns of expression.

Others have also explored the preferential expression of one homoeoallele in certain cotton genes (Rambani, Page, & Udall, 2014). At the same significance level they used, our study found a substantially higher number of biased genes (12,772 in our study vs. 2,686 or 3,146 in theirs). However, both studies observed a roughly equal proportions of A_t and D_t biased genes (6,378 and 6,394 in our study). One possible reason for this difference is that we did not first filter out genes that were not expressed at high levels in all replicates. Another possible reason is that they used cotton flower petals, rather than root tissue. Finally, it is possible that we detected more genes because we had a larger number of replicates (n=16 in our study, n=3 in theirs), and therefore greater power to detect subtle but consistent differences.

Microtranscriptome

The most common microorganisms in our dataset were *Candidatus Burkholderia brachynathoides* and *Arthrobacter* sp. L77. Both of these showed greater abundance, or at least greater transcriptional activity, under DT.

Burkholderia brachynathoides is a known leaf endosymbiont that colonizes members of the *Rubiaceae* group (Lemaire, Lachenaud, Persson, Smets, & Dessein, 2012). Cotton is not part of that clade, so it seems unlikely that that particular microbe would colonize cotton roots in this case. It is possible that the transcript BLAST matched with *brachyanthoides* is actually derived from a similar but distinct cotton root endosymbiont. Another possible explanation is that leaves and other residue from a previous crop, which were incorporated into the soil by DT, have a lasting impact on microbial soil communities.

Another member of the *Arthrobacter* genus, *Arthrobacter* sp. Strain AK-YN10, has been reported in Indian agricultural soils, where it degrades the herbicide atrazine (Sagarkar et al., 2014). Given the very uncertain relationship between these Strain AK-YN10 and the microbe we detected, sp. L77, it is only speculation to state that increased abundance or activity of atrazine-utilizing bacteria indicate a greater degree of residual herbicide in DT fields.

Perhaps the clearest insight from our microtranscriptome analysis, one that is already well known, is that short RNA reads are ineffective for unambiguous *de novo* transcript reconstruction. The presence of several transcripts with a large number of BLAST hits may be indicative of chimeric assemblies, a problem that might be exacerbated by widely conserved regions in microbial genomes.

Chapter 2: Physical DNA mapping

Introduction

High-throughput physical mapping

High-throughput physical mapping on the recently developed Irys® platform produced by BioNano Genomics (USA) not sequencing. Rather, this technology labels sparse sequence landmarks, namely restriction endonuclease recognition sites, to characterize much longer input molecules (for a size comparison, see Figure 3). Molecules characterized this way can be assembled into representations of contiguous genomic regions (contigs), and used to scaffold sequence contigs generated with MPS, or for direct comparison and structural variant (SV) detection (Hongzhi Cao et al., 2014; Hastie et al., 2013). This approach is reminiscent of the FingerPrint Contigs approach to bacterial artificial chromosome characterization (Soderlund, Longden, & Mott, 1997) and resembles optical mapping developed by David Schwartz (1993).

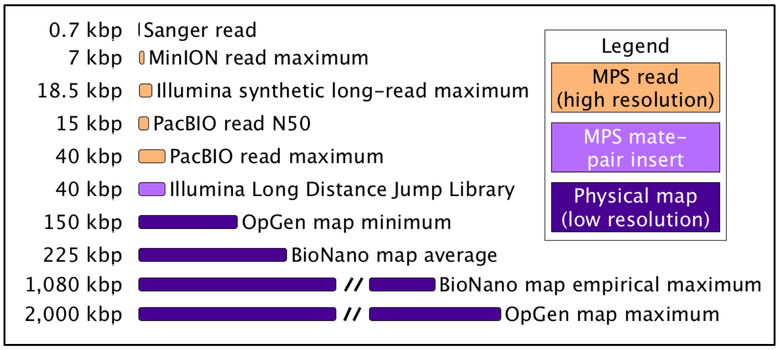

Figure 3: Comparison of DNA fragment lengths characterized by different technologies. Illumina, PacBIO, OpGen, and BioNano are company names. MinION is a sequencing platform produced by Oxford NANOPORE Technologies. MPS, massively parallel sequencing; kbp, kilobase-pairs.

Data collection

The process for data collection in the Irys system begins with a high molecular weight (HMW) DNA extraction, typically facilitated by embedding unlysed nuclei in agarose gel to protect DNA from shearing (M. Zhang et al., 2012). The purified DNA is subjected to enzymatic single-strand nicking at restriction endonuclease recognition sites after which a modified ligation-repair process is used to incorporate fluorescent nucleotide analogs near the site of the nick. Single molecules are then loaded using an electric current into nanoscopic channels (Han Cao et al., 2002), which confine them in a linear conformation while they are imaged using a high-powered microscope (Lam et al., 2012).

As with any single molecule technology, there is considerable noise in the raw data. For example, distances measured between fluorescent labels might not accurately reflect distributions of restriction endonuclease recognition motifs in the genome, because to non-uniform behavior of fluorophore, because of stretching of the DNA duplex, and because of camera resolution limits. Some recognition motifs might not be labeled, and some labels may occur at locations other than restriction motifs because of enzyme inefficiency and single-strand nicks existing in damaged DNA (Valouev, Schwartz, Zhou, & Waterman, 2006). Obtaining high quality data from plants is particularly difficult because of natural contaminants such as polyphenols, polysaccharides, and proteinase inhibitors (Varma, Padh, & Shrivastava, 2007). Lab procedures should aim to minimize these contaminants; however, protocols for this new technology are limited. Part of my work was to explore best practices in the lab for improved data quality.

Physical map assembly

Input mapping data are assembled into genome map, which is a set of consensus contigs, each representing a unique genomic interval. That pattern of distances between observed labels gives each contig a unique fingerprint by which it can be identified.

In order to recreate accurate genomic contigs, algorithms that assemble molecule data must compensate for noise inherent in physical mapping data. In order to detect true overlaps, assembly algorithms use inexact length matching and model probabilities of both missed and erroneous labels. In order to prevent spurious overlaps, assembly algorithms use a significance threshold or p-value, requiring label-pattern matches between molecules to be so similar that they are unlikely results of chance. Depending on genome size and complexity, restriction endonuclease recognition site patterns may be similar at multiple loci. An algorithm's ability to mitigate input noise relies in large part on user-provided input parameters that describe the error profile of the input dataset (Valouev, Li, et al., 2006; Valouev, Schwartz, et al., 2006). Therefore, accurate assembly requires that a user select reasonable input parameters.

There are methods for empirical estimation of reasonable input parameters. However, most of them rely on significant existing genomic resources. For example, BNG provides software that maps a random subset of input data to a reference genome sequence assembly, and selects input parameters that maximize the number of molecules that align, as well as the goodness of fit for those alignments. When a reference assembly is not available, one potential alternative is to select input parameters by trial and error. Using a variety of input parameter combinations yields a variety of assemblies, from which an optimal solution might be chosen. Part of my work was to develop software that would make this approach computationally feasible.

Methods

Data collection

We selected the species *Gossypium raimondii* because it is the closest living relatives to one of the subgenome progenitors of the agriculturally significant allopolyploid cotton, *G. hirsutum* (Brubaker, Paterson, & Wendel, 1999). It also has a high quality reference genome sequence assembly that was created using MPS, as well as a genetic maps and a traditional, BAC-based physical map (Paterson et al., 2012).

We collected two separate datasets. For the first (dataset1), ~10g young leaf tissue from several *G. raimondii* plants was flash frozen in liquid nitrogen and shipped on dry ice to Kansas State University, a Certified Service Provider for physical mapping with BNG technology. They performed HMW DNA extraction according to a proprietary protocol that includes physical disruption of the cell wall with a mortar and pestle, polyphenol isolation with PVP (SIGMA-ALDRITCH, USA) and Percoll (GE Healthcare Life Sciences, USA), plastid contaminant removal using Triton X-100 (SIGMA-ALDRITCH, USA), protein digestion with proteinase K (NEB, USA), and embedding of unlysed nuclei in agarose gel to prevent DNA shearing (M. Zhang et al., 2012). Purified DNA molecules were subjected to single-strand nicking at sites recognized by two modified restriction endonucleases, Nt.BspQ1 (6 ul) and Nt.BbvCl (4ul) (NEB, USA), simultaneously. The second dataset (dataset2) was collected in our own lab. Approximately 2g young leaf tissue was harvested from a single, mature *G. raimondii* individual for DNA extraction. Mapping data were collected as above with a few modifications. Unfrozen tissue was fixed in formaldehyde and immediately homogenized using a Qiagen TissueRuptor (Qiagen, Belguim). Multiple wash steps with Triton X-100 removed mitochondria and plastids until centrifuged pellets were not visibly green. Additional SDS was used during protein digestion with proteinase K. High molecular weight DNA was digested with 8ul of Nt.BspQ1 only. Additional differences between datasets outlined in Table 2. For both datasets, restriction endonuclease recognition sites were labeled with fluorescent nucleotide analogs provided by BNG, which were incorporated by Taq polymerase (NEB, USA). The DNA backbone was stained with a non-specific, intercalating dye, provided by BNG. Labeled, stained DNA molecules were linearized by physical constriction in nanoscopic channels on an Irys Chip v2.0

(BioNano Genomics, USA), immobilized with an electric current, and imaged automatically with a high-powered microscope and high-resolution camera using the Irys system.

Table 2: Improved lab methods

Dataset1	Target application	Dataset2
Liquid nitrogen	Fix cells to protect DNA and halt nucleases	Formaldehyde
Mortar and pestle	Rupture cell walls	Qiagen TissueRuptor
Centrifugation	Remove cellular debris	Micrometer filters
Single Triton X-100 wash	Remove plastid contaminants	Multiple Triton X-100 washes
Mix with paintbrush	Homogenize nuclei suspension	Mix with paintbrush and with non-stick pipett tip
Float in two separate Percoll gradients	Isolate nuclei	Float in a single Percoll gradient
Proteinase K alone	Digest proteins	Proteinase K with more SDS
RNAse	Digest RNA	RNAse
Embed unlysed nuclei in agarose plugs	Maintain long DNA molecules	Embed unlysed nuclei in agarose plugs
Cut with two restriction enzymes simultaneously	Nick recognition sites	Cut with a single restriction enzyme
Incorporate fluorescent analogues with Taq polymerase	Label recognition sites	Incorporate fluorescent analogues with Taq polymerase
Run in nanochannels on Irys machine	Image molecules	Run in nanochannels on Irys machine

The Irys Chip v2.0 contains two arrays of channels divided into flow cells. The first dataset consisted of multiple, individually labeled DNA aliquots and 19 total flow cell runs over 5 chips, each at about 20-cycles per flow cell run. The second dataset required only a single labeled aliquot and was run on both flow cells of a single Irys chip for four, 30-cycle runs. Software provided by BNG converted raw images into digital molecule representations. Data were filtered to remove labels with low ratios of label to background intensities. The threshold was determined dynamically by IrysView®, based on the distribution of background intensities in that flow cell run. All data were filtered to remove molecules shorter than 150 kbp.

Physical map assembly

Because assembly is highly sensitive to input parameters, we attempted multiple assemblies with different input parameter combinations. In order to empirically estimate parameters using the reference genome, we ran a molecule quality report (dataset1), and AutoNoise (dataset2), both of which are software packages provided by BNG. We also used OMWare (Sharp & Udall, 2016) on both datasets to efficiently run a large number of assemblies with a wide variety of input parameter values.

The user interface provided by BNG allows the user to specify a number of input parameters that are known to affect map assembly algorithms (see Mendelowitz & Pop, 2014; Valouev, Schwartz, et al., 2006). A *significance threshold* for accepting pairwise molecule alignments is an assumption about genome complexity, which frequently, but not necessarily, scales with genome size. It is an indication of how probable a match between two molecules is expected to occur because of random chance instead of a common genomic locus. The *false positive label rate* explains the frequency of observed labels found at locations other than the expected restriction endonuclease recognition sites. The *false negative rate* describes the proportion of restriction sites that do not have observed labels, due to enzyme inefficiency. It is an assumption of the BNG assembly algorithm that false positive labels and false negatives are distributed randomly throughout the genome. *Minimum molecule length* and *minimum labels per molecule* are not assumptions about the data error profile, or the genomic complexity. Rather, they represent a compromise between the number of molecules included and the reliability of each molecules, where longer, more label-dense molecules are more reliable. Although OMWare does not test their effect, the BNG user interface also includes multiple parameters to describe variance in observed distances between labels compared to actual restriction endonuclease recognition site distributions, as well as options relevant to the assembly refinement processes (see Valouev, Zhang, Schwartz, & Waterman, 2006). Although all of these parameters do not apply uniformly to all of the steps in the assembly process, the user interface only allows a single designation for each.

We designed and wrote Python code that would facilitate automatic assembly using a variety of values for those input parameters. This approach is similar to that used by Kansas State University in their program Irys-scaffolding (Shelton et al., 2015), except that it does not perform assembly refinement steps, and it breaks each assembly into its component parts in order to reduce the computational resources required. We ran OMWare twice to generate a total of 910 unrefined, *de novo* assemblies, 405 for each of our *G. raimondii* datasets, each time with a different combination of the input parameters shown in Table 3.

Table 3: Input parameter values tested with OMWare

Parameter	Overlap significance threshold	False positive labels per 100 kbp	Proportion restriction sites unlabeled	Min. molecule length (kbp)	Min. labels per molecule
Values	1.11E-04	0.5	0.15	100	6
	1.11E-06	1.5	0.3	150	8
	1.11E-08	2.5	0.45	180	10
	1.11E-10				
	1.11E-12				

We assessed the quality of the assemblies based on their contiguity and their internal consistency. We also used the reference genome to assess their accuracy. Assemblies were scored for total length, contig N50 length, and length of longest contig for contiguity. Internal consistency was divided into two metrics, the average number of overlapping molecules in which each label is observed, and the proportion of input molecules excluded from the assembly as singletons. Finally, we measured accuracy by comparing our assemblies to a highly contiguous reference genome sequence, using software provided by BNG. We report the weighted average confidence score, where confidence is the negative, 10-base logarithm of the p-value of an alignment.

Comparison to the reference genome

Once an optimal assembly was chosen, we compared it to the *G. raimondii* reference genome. To do so, we first converted reference sequence information into a physical map format by detecting restriction motifs *in silico* with software called Knickers (v.1.5.3) provided by BNG. Our initial comparison with BNG software allowed only a single, best match for each physical map contig. This helps assess map assembly quality and estimate error parameters for the next comparison. The second comparison allowed for multiple consecutive "match groups," on a single mapping contig. Portions of contigs that fall between significant matching groups are called as structural variants (SVs) or misassembles. We also ran HybridScaffold (v3659), with and without conservative filtering rules, in an attempt to join sequence scaffolds into collinear superscaffolds based on physical map evidence.

To assess the nature of disagreements between the reference genome sequence assembly and our BNG physical map assembly, we first filtered out discrepancies that could be explained with known shortcomings of the technologies that produced them. Partial matches, SVs, and mapping contigs that overlapped were filtered out if they matched near a genetic map join or putative collapsed repeat, or if the mapping contigs had low coverage regions. False positives were filtered if they fell within a gap, or could be explained as a single nucleotide variant. For each false positive, the sequence regions spanning 150 base-pairs on either side of the label was searched for seven consecutive N's, or any seven-base sequence that was one nucleotide off from the motif recognized by Nt.BspQI. False positives and false negatives were also filtered if they fell within 300 bp of another label. Disagreements between the MPS and BNG genome representations remaining after filtering will make good targets for additional follow up.

Results

Data collection

Dataset qualities were assessed using the metrics yield per cycle, proportion of expected label density observed, molecule N50, and fluorescence (signal to noise ratio, SNR) of both labels and molecule backbones.

In dataset1, we collected a total of 217.28 Gigabase-pairs (Gbp) of physical map data over nine, two-flow-cell runs of BNG's Irys machine. This is enough data for ~241x coverage of the similar to 900 Mbp *G. raimondii* genome. The weighted average across datasets of the molecule N50 length was 165.37 kbp. The expected label density using Nt.BspQ1 and Nt.BbvCl simultaneously was 12.6 labels per 100 kbp. Our observed label density was consistently lower than the expected (max 11.3 labels per 100 kbp, weighted average 9.2).

In our second dataset, collecting sufficient coverage required a single BNG chip, and a total of two flow cells. Individuals flow cells were run for 120 cycles. Dataset2 includes 230.49 Gbp of data (~256x coverage) with an N50 of 209.8 kbp. The average observed label density is 6.1 out of the expected 7.5 labels per 100 kbp. The quality improved in dataset2 (Figure 4).

Physical map assembly

Using OMWare, we generated a total of 810 unrefined assemblies, 405 for each dataset. Using IrysView we generated two refined assemblies, one for each dataset.

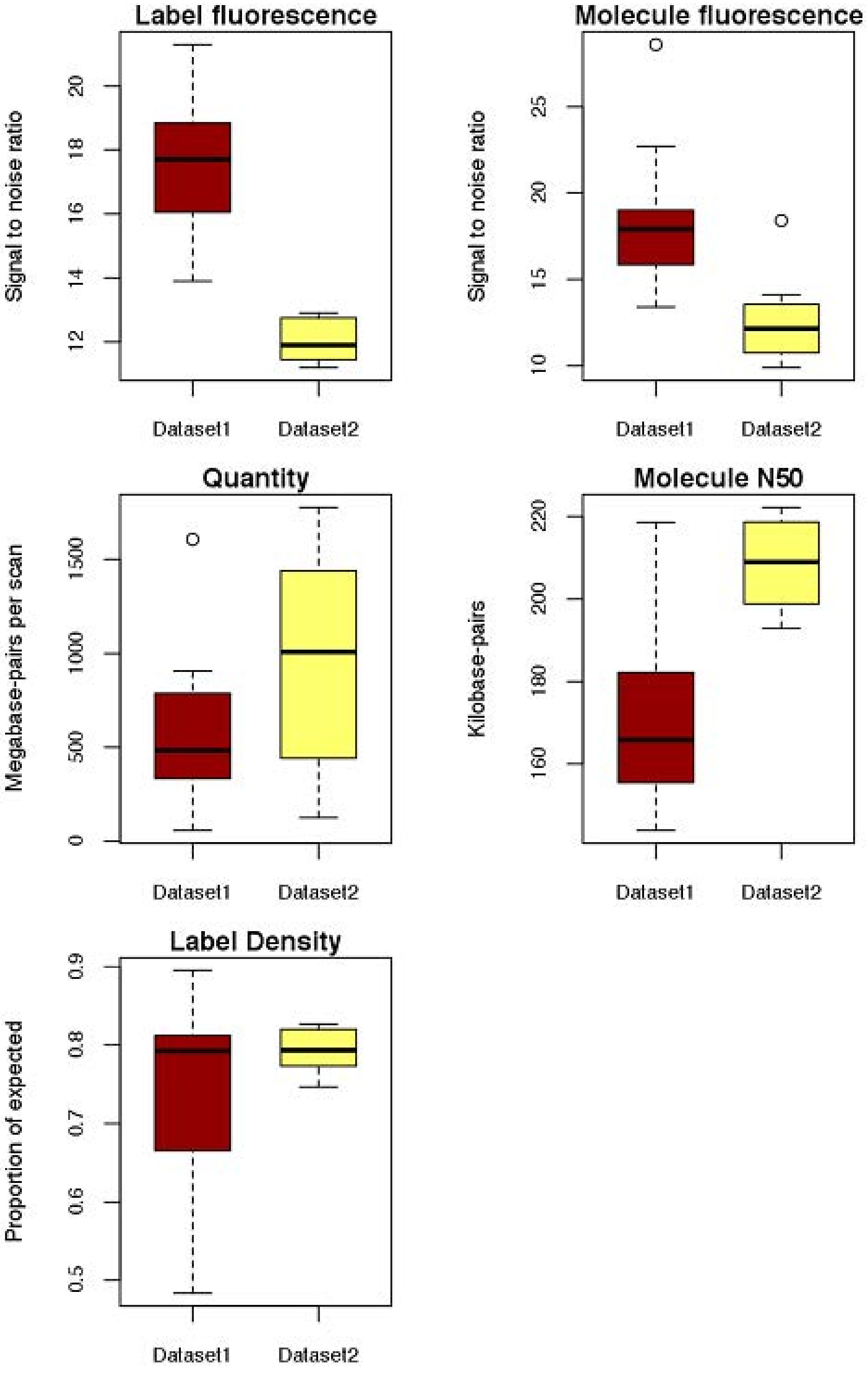

Figure 4: Differences in quality distributions between dataset1 and dataset2. Dataset1 (red) included 18 flow cell runs. Dataset2 (yellow) included 8 flow cell runs. Flow cell runs are divided into 20-30 scans, each of which begins when an electric current pulls a new aliquot of labeled DNA into the nanochannels. The expected label density for dataset1 was 12.6 labels per 100 kilobase-pair (kbp). The expected label density for dataset2 was 7.5 labels per 100 kbp.

Contiguity and internal consistency varied widely between assemblies, and were predominantly controlled by two input parameters, minimum molecule length and significance threshold. The maximum total length of any assembly was about 1.78 Gbp, which is much larger than the expected genome size. However, assembly refinement generally reduces the total assembly length (Table 4). The shortest assembly covered only 78 Mbp. Contig N50 lengths ranged from 252 to 1,821 kbp, and the maximum length of any single contig was 15.24 Mbp. In every assembly generated using dataset1, a large proportion of input molecules, from 0.90 to 0.993, were excluded as singletons. A smaller but still substantial proportion was excluded from dataset2, from 0.65 to 0.90. Across parameter combinations, the average number of molecules in which each label was observed fell between five and fourteen (Figure 5 and Figure 6).

The accuracy of assembled contigs also varied, and appeared to correspond very little with measures of contiguity or internal consistency. The lowest average confidence score of any assembly was 20.0, and the highest was 39.1. There were no outliers in confidence. The confidence scores are more responsive to changes in false positive label rates, false negative label rates, and minimum labels per molecule than metrics of contiguity appear to be (Figure 5 and Figure 6). For dataset1, the highest accuracy obtained using OMWare was greater than the unrefined accuracy from the molecule quality report. However, for dataset2, AutoNoise generated an assembly with considerably higher quality than OMWare (Table 4).

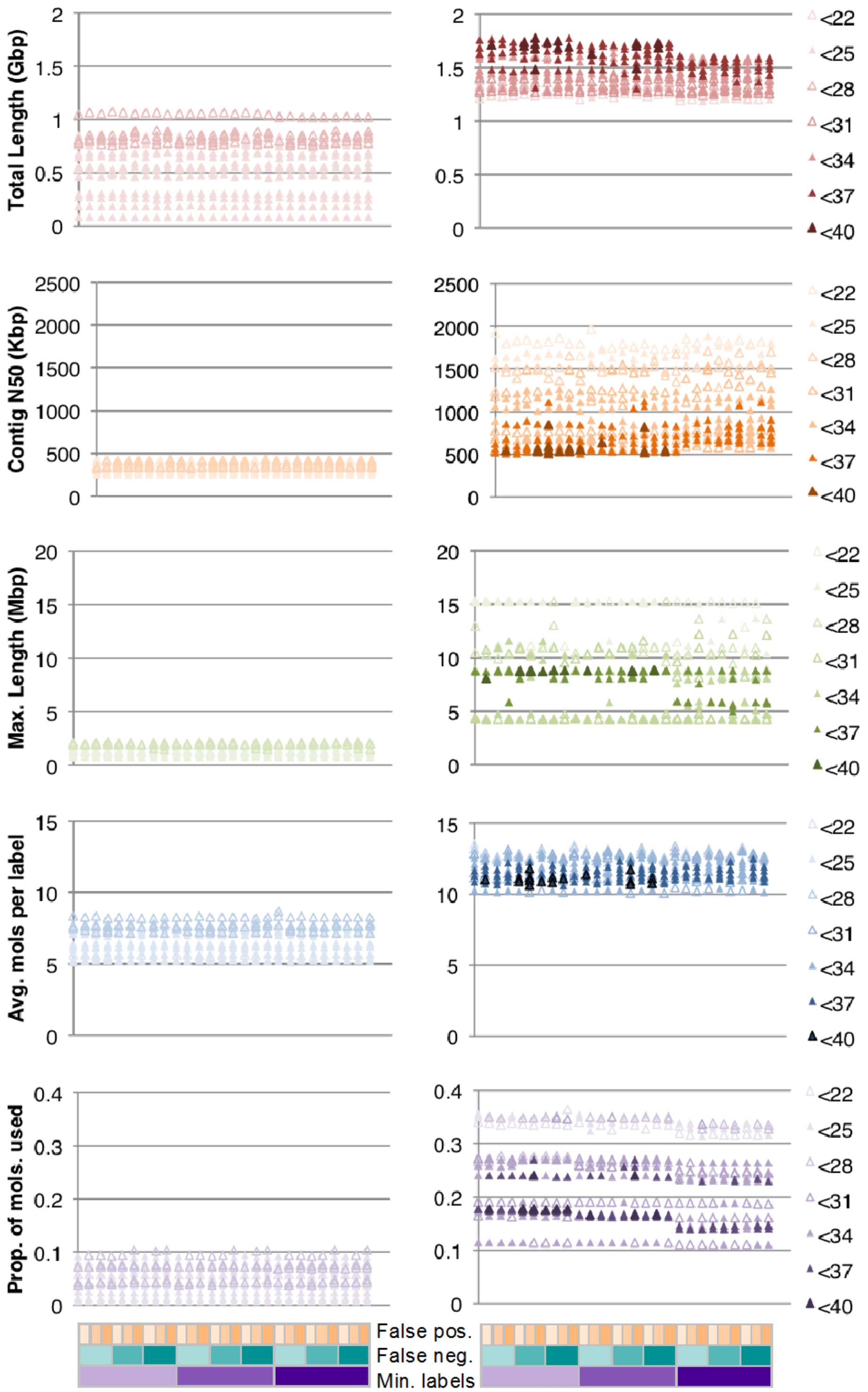

Figure 5: OMWare created 405 assemblies of dataset1 (left) and 405 of dataset2 (right). Contiguity or internal consistency are depicted on the y-axis (Gbp, Gigabase-pairs; kbp, kilobase-pairs; Mbp, Megabase-pairs; Max., maximum; avg., average; prop., propotion; mols., molecules). Accuracy is measured as confidence values ranging from 20 to 40, and is depicted with the color of each data point. The x-axis describes the combination of some of the input parameters used. False positives (False pos.) are one of 0.5 (lightest orange), 1.5, or 2.5 (darkest orange) false labels per 100 kbp. False negatives (False neg.) are one of 15 (lightest green), 30, or 45 (darkest green) percent of restriction motifs unlabeled. Minimum labels per molecule (Min. labels) are one of 6 (lightest purple), 8, or 10 (darkest purple).

Table 4: Refined vs. unrefined assemblies from dataset1 and dataset2

Dataset	Dataset1	Dataset1	Dataset1	Dataset2	Dataset2	Dataset2
Refined?	No	No	Yes	No	No	Yes
Error estimation method	OMWare	MQR	MQR	OMWare	AutoNoise	AutoNoise
Number of contigs	3,217	1,012	779	3,758	2,196	410
Total length (Mbp)	1,015.2	286.4	207.0	1,753.7	1,384.6	800.8
N50 (Kbp)	339.3	284.0	272.4	531.5	1,111.9	2,751.1
Average confidence (-log10(Pval))	27.8	23.2	31.16	39.1	108.5	285.7

Comparison to the reference genome

The best assembly was created using dataset2 and AutoNoise. Initial comparison with the reference genome showed considerable agreement. The weighted average confidence (-log$_{10}$(p-value)) of matches between map contigs and sequence scaffolds was 286. Out of 410 contigs, 402 (98.0%) found at least one significant match. Significant matches do not usually cover the entire contig; however, the median proportion of a contig length that was included within a significant match was 0.9998. The mean proportion was 0.9476, and there were 46 matches (11.4%) where the proportion of contig length within the match was lower than 0.9.

Our comparison also called 782 SVs, of which 752 (96.2%) were insertions or deletions, 14 (1.8%) were translocations, and 16 (2.0%) were inversions. We further categorized these SVs, determining that 61 (7.8%) did not have high enough coverage to be confident, and 26 (3.3%) of them were likely to be part of a collapsed repeat.

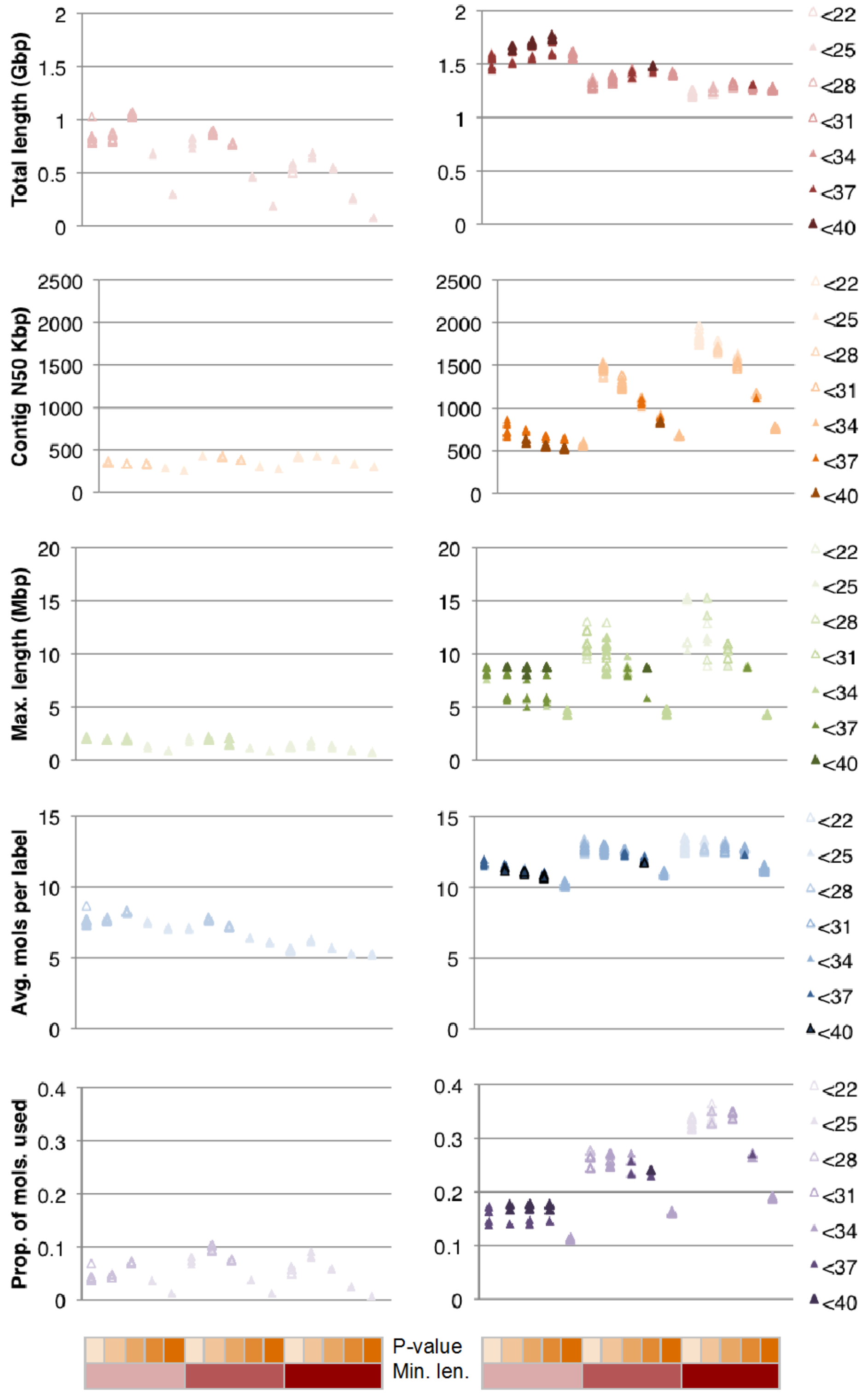

Figure 6: Certain input parameters have a greater impact on assembly contiguity and internal consistency, but not necessarily accuracy. The y-axis and data-point colors have the same meaning as in Figure 5. The x-axis describes some of the assembly input parameter combinations. Overlap significance threshold (P-value) is one of 1.11E-4 (lightest orange), 1.11E-6, 1.11E-8, 1.11E-10, or 1.11E-12 (darkest orange). Minimum molecule length (Min. len.) is one of 100 kbp (lightest red), 150 kbp, or 180 kbp (darkest red).

HybridScaffold suggested no changes to the sequence assembly when it was run with strict filtering parameters. The more lenient run, however, recommended combining sequence scaffolds one and six, both of which are long pseudomolecules, into a single 108 Mbp superscaffold. This merge would likely be inaccurate.

In addition to SVs and suggested joins, the final assembly also had a 2,648 false positive labels (0.12 labels per 100 kbp), and 1,681 false negative labels (1.3% of reference restriction endonuclease recognition motifs). A total of 49,082 labels were found in both genome representations (Table 5).

Table 5: Disagreements between MPS and BNG genome representations

Disagreement type	Partial matches	Overlapping contigs	SVs	False positives	False negatives
Prior to filtering	46	24	782	2,648	1,681
Low coverage regions	13	2	61	-	-
Collapsed repeat	10	2	26	-	-
Genetic map join	23	0	0	-	-
FP in gap	-	-	-	367	-
FP by SNV	-	-	-	536	-
FP/FN too close to another label	-	-	-	153	173
After filtering	0	20	695	1,592	1,508

Discussion

Data collection and physical map assembly

There was a substantial improvement in data quality in dataset2. It is unclear if the lower signal to noise ratios represent an improvement, however, molecule lengths were much higher and label densities were much more uniform across flow cells in dataset2. This hints that consistent label densities may be more advantageous for assemblies than widely dispersed label densities, even if the latter are, on average, closer to the expected. There were some flow cell runs in dataset2 with higher yields than the average in dataset1, but the main factor in decreasing the number of chips used for dataset2 was running each flow cell for 120 rather than 20 cycles. Our experience indicates that that using a single restriction endonuclease rather than a cocktail of enzymes improves data quality considerably. The higher expected label density in dataset1 might also have contributed to shorter molecule N50s.

The data quality improvements in dataset2 are reflected in the assembly quality. OMWare assemblies using dataset2 vastly outperform dataset1 assemblies in metrics of contiguity and internal consistency. Interestingly, however, there some are assemblies of dataset2 data that have lower confidences than the lowest confidence of any dataset1 assembly. It is also interesting to note that the accuracies of the refined assemblies are, in general, a full order of magnitude higher than the highest of the unrefined assemblies. Additionally, while metrics of contiguity and internal consistency seem to respond most to minimum molecule length and overlap significance threshold in both datasets, in dataset2 these measures are more responsive to false positive label rate, false negative label rate, and minimum labels per molecule than dataset1 assemblies. Again, no metrics of internal consistency or contiguity seem to correspond to accuracy, except for perhaps total length.

Comparison to the reference genome

There was a substantial amount of agreement between the published reference genome sequence assembly and our best assembled genome map, dataset2 assembled using AutoNoise. A total of 745 Mbp (98.9%) of the reference were covered by significant map matches, with a weighted average confidence of 285.7 and over 49,000 restriction endonuclease recognition sites were detected in both. Of 5,181 total disagreements between the two, 816 (15.7%) could be satisfactorily explained as low coverage regions (76, 1.5%, map is probably wrong), probable collapsed repeats (38, 0.7%, sequence is probably wrong), genetic map joins (23, 0.4%, sequence is almost certainly wrong), end variants (9, 0.2%, neither is wrong), gaps containing restriction sites (367, 7.1%, sequence is probably wrong), or labels actually within Irys' detectable resolution limit (326, 6.3%, disagreement is likely artificial). Additionally, 536 (10.3%) disagreements are labels detected in the map that would also be present in the sequence genome if just one nucleotide were changed. These disagreements may represent natural variability between individuals. After filtering, 3,806 disagreements remain that merit additional follow up.

Conclusions

Conservation tillage

Conservation tillage is unlikely to have a substantial impact on cotton phenotype, and therefore the benefits of lower greenhouse gas emission, better soil infiltration, and reduced erosion make it a reasonable choice. However, the uncertain but statistically significant effect of conservation tillage on plant-available phosphate and sulfate are worth following up, as are the potential implications of incorporating leaf-derived microbes into the soil with disk tillage.

Physical mapping

The quality of data collected using the Irys platform may improve considerably in response to more thorough, homogenous blending, additional Triton X-100 washes, and digestion with a single restriction endonuclease.

Selecting optimal input parameters with OMWare, at least for this dataset, yielded lower quality assemblies than the BNG software AutoNoise. Additionally, contiguity and internal consistency are unreliable indicators of accuracy in the absence of a reference genome.

There are 3,806 discrepancies between the currently accepted *G. raimondii* reference genome sequence assembly and our best physical map constructed using Irys. These merit follow up and validation, for example, with BAC-end sequencing. Addtionally, 38 regions that are probable collapsed repeats might be corrected in the reference genome using physical map data alone.

Agronomics

Diverse genomic technologies allow researchers to explore crop genetics and molecular responses to the environment. Small but potentially meaningful insights into plant biology may drive increased agricultural productivity that will help to feed a growing world population.

List of Abbreviations

BAC – Bacterial artificial chromosome

BNG – BioNano Genomics

CT – Conservation tillage

DT – Conventional disk tillage

FAO – Food and Agriculture Organization of the United Nations

FDR – False Discovery Rate

G. – Gossypium

Gbp – Gigabase-pairs

HMW – High molecular weight

kbp – Kilobase-pairs

Mbp – Megabase-pairs

MPS – Massively parallel sequencing

MQR – Molecule quality report

NCBI – National Center for Biotechnology Information

SNR – Signal to noise ratio

Sp. – Species

SV – Structural variant

Bibliography

Anderson, S., Bankier, A. T., Barrell, B. G., Debruijn, M. H. L., Coulson, A. R., Drouin, J., . . . Young, I. G. (1981). SEQUENCE AND ORGANIZATION OF THE HUMAN MITOCHONDRIAL GENOME. *Nature, 290*(5806), 457-465. doi:10.1038/290457a0

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Bowman, M. J., Park, W., Bauer, P. J., Udall, J. A., Page, J. T., Raney, J., . . . Campbell, B. T. (2013). RNA-Seq Transcriptome Profiling of Upland Cotton (Gossypium hirsutum L.) Root Tissue under Water-Deficit Stress. *PLoS One, 8*(12), e82634.

Brubaker, C., Paterson, A., & Wendel, J. (1999). Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome, 42*(2), 184-203.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics, 10*(1), 1.

Cao, H., Hastie, A. R., Cao, D., Lam, E. T., Sun, Y., Huang, H., . . . Xu, X. (2014). Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience*, *3*(1), 34. (Vol. 3(1), pp. 1-11): *GigaScience*.

Cao, H., Yu, Z., Wang, J., Tegenfeldt, J. O., Austin, R. H., Chen, E., . . . Chou, S. Y. (2002). Fabrication of 10 nm enclosed nanofluidic channels. *Applied physics letters, 81*(1), 174-176.

Carpenter-Boggs, L., Stahl, P. D., Lindstrom, M. J., & Schumacher, T. E. (2003). Soil microbial properties under permanent grass, conventional tillage, and no-till management in South Dakota. *Soil & Tillage Research, 71*(1), 15-23. doi:10.1016/s0167-1987(02)00158-7

Chen, Y., Ye, W., Zhang, Y., & Xu, Y. (2015). High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Research, 43*(16), 7762-7768. doi:10.1093/nar/gkv784

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., . . . Venter, J. C. (1995). WHOLE-GENOME RANDOM SEQUENCING AND ASSEMBLY OF HAEMOPHILUS-INFLUENZAE RD. *Science, 269*(5223), 496-512. doi:10.1126/science.7542800

Fries, L. L. M., Pacovsky, R. S., & Safir, G. R. (1996). Expression of isoenzymes altered by both Glomus intraradices colonization and formononetin application in corn (Zea mays L) roots. *Soil Biology & Biochemistry, 28*(8), 981-988. doi:10.1016/0038-0717(96)00115-0

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods, 8*(6), 469-477. doi:10.1038/nmeth.1613

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology, 29*(7), 644-U130. doi:10.1038/nbt.1883

Hastie, A. R., Dong, L. L., Smith, A., Finklestein, J., Lam, E. T., Huo, N. X., . . . Xiao, M. (2013). Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate

De Novo Sequence Assembly of the Complex Aegilops tauschii Genome. *Plos One, 8*(2), 10. doi:10.1371/journal.pone.0055864

Horacek, J., Kolar, L., Cechova, V., & Hrebeckova, J. (2008). Phosphorus and carbon fraction concentrations in a Cambisol soil as affected by tillage. *Communications in Soil Science and Plant Analysis, 39*(13-14), 2032-2045. doi:10.1080/00103620802134867

Hunt, P. G., Matheny, T. A., & Wollum, A. G. (1985). RHIZOBIUM-JAPONICUM NODULAR OCCUPANCY, NITROGEN ACCUMULATION, AND YIELD FOR DETERMINATE SOYBEAN UNDER CONSERVATION AND CONVENTIONAL TILLAGE. *Agronomy Journal, 77*(4), 579-584.

Klueva, N. Y., Joshi, R. C., Joshi, C. P., Wester, D. B., Zartman, R. E., Cantrell, R. G., & Nguyen, H. T. (2000). Genetic variability and molecular responses of root penetration in cotton. *Plant Science, 155*(1), 41-47. doi:10.1016/s0168-9452(00)00205-3

Koprivova, A., & Kopriva, S. (2016). Hormonal control of sulfate uptake and assimilation. *Plant molecular biology*, 1-11.

Lachnicht, S. L., Hendrix, P. F., Potter, R. L., Coleman, D. C., & Crossley, D. A. (2004). Winter decomposition of transgenic cotton residue in conventional-till and no-till systems. *Applied Soil Ecology, 27*(2), 135-142. doi:10.1016/j.aspoil.2004.05.001

Lam, E. T., Hastie, A., Lin, C., Ehrlich, D., Das, S. K., Austin, M. D., . . . Kwok, P. Y. (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology, 30*(8), 771-776. doi:10.1038/nbt.2303

Lemaire, B., Lachenaud, O., Persson, C., Smets, E., & Dessein, S. (2012). Screening for leaf-associated endophytes in the genus Psychotria (Rubiaceae). *FEMS microbiology ecology, 81*(2), 364-372.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data, P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science, 326*(5950), 289-293. doi:10.1126/science.1181369

Lister, R., & Ecker, J. R. (2009). Finding the fifth base: Genome-wide sequencing of cytosine methylation. *Genome Research, 19*(6), 959-966. doi:10.1101/gr.083451.108

Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nature Methods, 4*(8), 613-614. doi:10.1038/nmeth0807-613

Mbuthia, L. W., Acosta-Martinez, V., DeBruyn, J., Schaeffer, S., Tyler, D., Odoi, E., . . . Eash, N. (2015). Long term tillage, cover crop, and fertilization effects on microbial community structure, activity: Implications for soil quality. *Soil Biology & Biochemistry, 89*, 24-34. doi:10.1016/j.soilbio.2015.06.016

Mendelowitz, L., & Pop, M. (2014). Computational methods for optical mapping. *GigaScience, 3*(1), 33.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics, 24*(16), 1757-1764.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., . . . Venter, J. C. (2000). A whole-genome assembly of Drosophila. *Science, 287*(5461), 2196-2204. doi:10.1126/science.287.5461.2196

Novak, J. M., Bauer, P. J., & Hunt, P. G. (2007). Carbon dynamics under long-term conservation and disk tillage management in a Norfolk loamy sand. *Soil Science Society of America Journal, 71*(2), 453-456. doi:10.2136/sssaj2005.0284N

Page, J. T., Gingle, A. R., & Udall, J. A. (2013). PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms. *G3-Genes Genomes Genetics, 3*(3), 517-525. doi:10.1534/g3.112.005298

Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D. C., . . . Schmutz, J. (2012). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature, 492*(7429), 423-+. doi:10.1038/nature11798

Pittelkow, C. M., Linquist, B. A., Lundy, M. E., Liang, X. Q., van Groenigen, K. J., Lee, J., . . . van Kessel, C. (2015). When does no-till yield more? A global meta-analysis. *Field Crops Research, 183*, 156-168. doi:10.1016/j.fcr.2015.07.020

Potter, T. L., Bosch, D. D., & Strickland, T. C. (2015). Tillage impact on herbicide loss by surface runoff and lateral subsurface flow. *Science of the Total Environment, 530*, 357-366. doi:10.1016/j.scitotenv.2015.05.079

R Core Team (2015). R: A language and environment for statistical computing Vienna, Austria: R Foundation for Statistical Computing.

Rambani, A., Page, J. T., & Udall, J. A. (2014). Polyploidy and the petal transcriptome of Gossypium. *BMC Plant Biol, 14*(1), 3. doi:10.1186/1471-2229-14-3

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi:10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology, 11*(3). doi:10.1186/gb-2010-11-3-r25

Sagarkar, S., Bhardwaj, P., Yadav, T. C., Qureshi, A., Khardenavis, A., Purohit, H. J., & Kapley, A. (2014). Draft genome sequence of atrazine-utilizing bacteria isolated from Indian agricultural soil. *Genome announcements, 2*(1), e01149-01113.

Schwartz, D. C., Li, X., Hernandez, L. I., Ramnarain, S. P., Huff, E. J., & Wang, Y.-K. (1993). Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science, 262*(5130), 110-114.

Sharp, A. R., & Udall, J. A. (2016). *OMWare: A tool for efficient assembly of genome-wide physical maps*. Brigham Young University. BMC Bioinformatics.

Shelton, J. M., Coleman, M. C., Herndon, N., Lu, N., Lam, E. T., Anantharaman, T., . . . Brown, S. J. (2015). Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC genomics, 16*(1), 734.

Smit, A. L., Bindraban, P. S., Schröder, J., Conijn, J., & Van der Meer, H. (2009). Phosphorus in agriculture: global resources, trends and developments. *Report to the Steering Committee Technology Assessment of the Ministry of Agriculture, The Neetherlands, Wageningen*.

So, H. B., Grabski, A., & Desborough, P. (2009). The impact of 14 years of conventional and no-till cultivation on the physical properties and crop yields of a loam soil at Grafton

NSW, Australia. *Soil & Tillage Research, 104*(1), 180-184. doi:10.1016/j.still.2008.10.017

Soane, B. D., Ball, B. C., Arvidsson, J., Basch, G., Moreno, F., & Roger-Estrade, J. (2012). No-till in northern, western and south-western Europe: A review of problems and opportunities for crop production and the environment. *Soil & Tillage Research, 118*, 66-87. doi:10.1016/j.still.2011.10.015

Soderlund, C., Longden, I., & Mott, R. (1997). FPC: a system for building contigs from restriction fingerprinted clones. *Computer applications in the biosciences: CABIOS, 13*(5), 523-535.

Sojka, R. E., Karlen, D. L., & Busscher, W. J. (1991). A CONSERVATION TILLAGE RESEARCH UPDATE FROM THE COASTAL-PLAIN SOIL AND WATER CONSERVATION RESEARCH-CENTER OF SOUTH-CAROLINA - A REVIEW OF PREVIOUS RESEARCH. *Soil & Tillage Research, 21*(3-4), 361-376. doi:10.1016/0167-1987(91)90031-r

Valouev, A., Li, L., Liu, Y.-C., Schwartz, D. C., Yang, Y., Zhang, Y., & Waterman, M. S. (2006). Alignment of optical maps. *Journal of Computational Biology, 13*(2), 442-462. doi:10.1089/cmb.2006.13.442

Valouev, A., Schwartz, D. C., Zhou, S., & Waterman, M. S. (2006). An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proceedings of the National Academy of Sciences, 103*(43), 15770-15775.

Valouev, A., Zhang, Y., Schwartz, D. C., & Waterman, M. S. (2006). Refinement of optical map assemblies. *Bioinformatics, 22*(10), 1217-1224.

Varma, A., Padh, H., & Shrivastava, N. (2007). Plant genomic DNA isolation: An art or a science. *Biotechnology Journal, 2*(3), 386-392. doi:10.1002/biot.200600195

Wetterstrand, K. (2015). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP): http://www.genome.gov/sequencingcosts

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics, 26*(7), 873-881.

Yin, X., & Main, C. (2015). Nitrogen fertilization and critical nitrogen concentration for contemporary high yielding cotton under no-tillage. *Nutrient Cycling in Agroecosystems, 103*(3), 359-373. doi:10.1007/s10705-015-9751-0

Zhang, M., Zhang, Y., Scheuring, C. F., Wu, C.-C., Dong, J. J., & Zhang, H.-B. (2012). Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *7*(3), 467-478.

Zhang, S., Li, Q., Lu, Y., Zhang, X., & Liang, W. (2013). Contributions of soil biota to C sequestration varied with aggregate fractions under different tillage systems. *Soil Biology & Biochemistry, 62*, 147-156. doi:10.1016/j.soilbio.2013.03.023

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., . . . Chen, Z. J. (2015). Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. *33*(5), 531-537.