A COMPARISON OF THE EFFICIENCY IN FINDING GENES BETWEEN SEQUENCES ENRICHED FOR HYPO-METHYLATED REGIONS AND WHOLE-GENOME SHOTGUN SEQUENCE IN BREAD WHEAT

by

Joshua Lee Watson

Honors Thesis

Appalachian State University

Submitted to The Honors College
in partial fulfillment of the requirements for the degree of

Bachelor of Science

August 2015

Approved by:

Matt C. Estep, Ph.D., Thesis Director

Jill E. Thomley, Ph.D., Second Reader

Leslie Sargent Jones, Ph.D., Director, The Honors College

ABSTRACT

Bread wheat (*Triticum aestivum*) has a roughly 17Gbp hexaploid genome, resulting from a hybridization event between tetraploid emmer wheat (*Triticum dicoccoides*) and diploid goat grass (*Aegilops tauschii*). This large plant genome is composed of at least 80% transposable elements (TE's), making the transcriptionally active regions (genes) difficult to locate. Epigenetic methylation of DNA is a common indicator of low transcriptional activity and is used to silence TE's within a genome. Using restriction enzymes that cannot cut methylated DNA (*HpaII* and *HpyCH4IV*) Illumina sequencing libraries were constructed that are enriched for hypomethylated regions of the wheat cultivar "Chinese Spring". The resulting sequence data (roughly 4.5 Gb) was assembled into contigs with AbySS using *k*-values of 36, 50, and 64. Resulting contigs were then annotated for gene content using Blastx and Blast2GO. Our findings were then compared to un-enriched sequences from a whole genome shotgun sequence to determine the gene enrichment potential of our selection strategy. When contigs were assembled with a *k*-value of 64 for the libraries made with *HpyCH4IV* and *k*-values of 64 and 50 for the *HpaII* libraries, a higher proportion of genes were identified than in the control whole genome shotgun sequence.

INTRODUCTION

Many common cereal crops have extremely large genomes, which makes them difficult to sequence (Nelson et al. 2008). Their nuclear genomes are also filled with repetitive, non-coding DNA, making it very problematic to find and analyze functional genes (Nelson et al. 2008). Plant genome size is affected by several factors, such as polyploidy via hybrid speciation and the amplification of transposable elements (Bennetzen 2002). Bread wheat (*Triticum aestivum*) in particular is problematic to genetically manipulate, as it has a 17Gbp hexaploid genome, (with component genomes known as A, B, and D) which is the result of multiple independent hybridization events (Figure 1) (Brenchley et al. 2012). The A and B genomes (which gave rise to *Triticum urartu* and a now-extinct relative of *Aegilops speltoides*, respectively) diverged roughly 6.5 million years ago and hybridized to create the D genome, *Aegilops tauschii*, roughly 5.5 million years ago (Marcussen et al. 2014). Less than 1 million years ago, the A and B genomes would again hybridize to make a new species, tetraploid emmer wheat (*Triticum turgidum*) (Marcussen et al. 2014). The final hybridization event, which created modern bread wheat, was between tetraploid emmer wheat (*Triticum turgidum*) and diploid goat grass (*Aegilops tauschii*) roughly 10000-8000 years ago, which correlates with the beginnings of human agriculture (Brenchley et al. 2012).

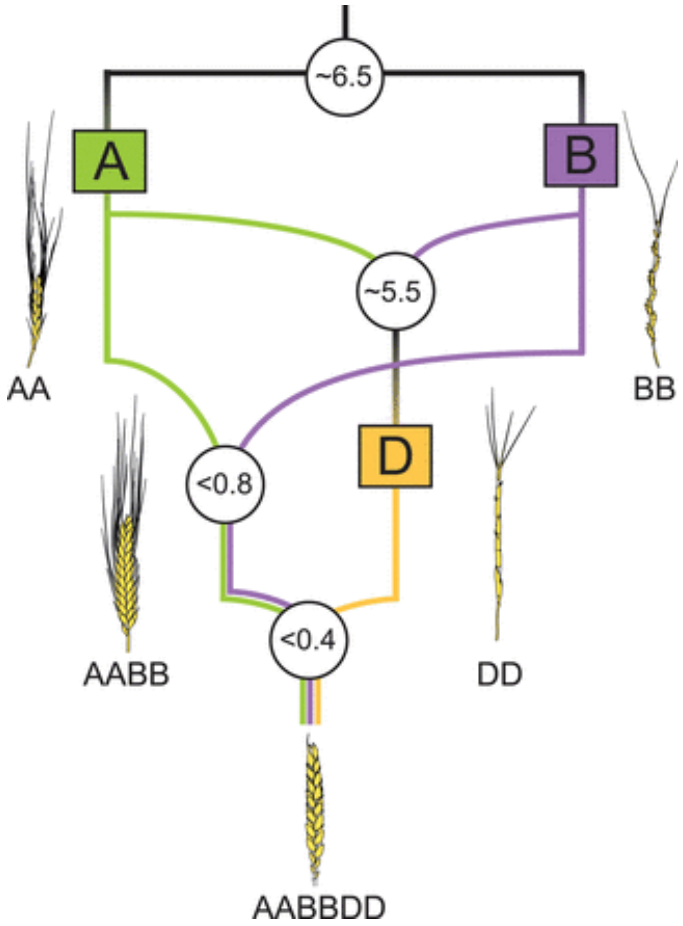

Figure 1. Model of the Phylogenetic History of Bread Wheat (*Triticum aestivum*; AABBDD). Units indicate millions of years ago (Marcussen et al. 2014).

Compounding the difficulty of sequence analysis in bread wheat is the high percentage of highly repetitive non-gene structures in its nuclear genome (Brenchley et al. 2012). The most prevalent non-gene structures are the transposable elements (TEs), which can be classified as either DNA transposons or retrotransposons (Charles et al. 2008). Of these two classes, retrotransposons have a greater impact on the size of the genome in plants, composing a significant portion (80%) of the maize (*Zea mays*) genome (Nelson et al. 2008). Retrotransposons are mobile genetic elements that can create copies of their DNA sequence and then move their copies to other sections of the genome at will, by creating an RNA transcript of the enzymes reverse transcriptase,

RNaseH, and integrase as well as the repetitive DNA (Kumar et al. 1999). The transcripts for these enzymes are then translated into their respective proteins, which use the RNA as a template to reverse transcribe the DNA sequence and insert it at a new location (Kumar et al. 1999). These mobile genetic elements cause mutations when they insert themselves into or near genes in a host genome, and the replicating mechanism they use produces mutations that are often stable, as they do not alter the DNA sequence at the point of replication (Kumar et al. 1999).

Retrotransposons can be further broken down into the subgroups of long terminal repeat (LTR) retrotransposons, long interspersed repetitive elements (LINEs), and short interspersed repetitive elements (SINEs) (Kumar et al. 1999). Of these three, LTR retrotransposons are the most common and the most complex, being found in most plant species and taking up a significant portion of the genome, such as over 60% of the maize genome and slightly less than 60% of the wheat genome (Emberton et al. 2005; Charles et al. 2008; Kumar et al. 1999). LTR retrotransposons contain protein-coding genes specifically for their own replication and insertion, but they do not generally contain genes relevant to other cell processes (Kumar et al. 1999). LINEs are similar to LTR retrotransposons, but they lack a gene to insert themselves back into the genome, and SINEs (non-autonomous elements), the smallest retrotransposons, lack any kind of genes to move themselves, relying on those encoded by LINEs or LTR retrotransposons (Kumar et al. 1999).

DNA Transposons, the other large category of transposable element, differ from retrotransposons in that their mechanism of transposition is a DNA intermediate instead of an RNA intermediate (Wicker et al. 2007). There are two subclasses of DNA

transposons: subclass 1 and subclass 2 (Wicker et al. 2007). Subclass 1 DNA transposons are the more researched of the two, and were thought to be the only kind of DNA transposon until recently (Wicker et al. 2007). These transposable elements are generally much smaller than retrotransposons, containing a short terminal inverted repeat (TIR) of at most 200 bp and a gene that encodes the enzyme transposase (Wessler 2006). Transposase binds to the TIRs and cuts both strands of DNA, excising the transposon, and inserts it into another region of the genome (Wessler 2006). Subclass 2 DNA transposons are not well understood, but they do replicate a strand of DNA and then move a single strand to another area, much like retrotransposons but without an RNA intermediate (Wicker et al. 2007). DNA transposons can be present in the genome in numbers comparable to retrotransposons, but due to their smaller size, they are responsible for a smaller percentage of the genome, such as 8.6% in maize or 10-14% in wheat (Charles et al. 2008; Schnable et al. 2009).

The wheat genome has undergone multiple hybridization events and transposable element amplifications, creating a genome with low gene density, which, based on sequence analysis, varies widely, ranging from 1 gene per 87 kb to 1 gene per 184 kb (Choulet et al. 2010). When genes do occur, they are found in groups described as “gene islands”, as previous work has determined that 94% of wheat genes are found in gene dense regions that cover 29% of the genome (Choulet et al. 2010). Maize (*Zea mays*) has a similarly low gene density, even with a vastly different evolutionary history (Bennetzen et al. 2005). Previous studies have shown that the gene density in maize can range from 0.5-10.7 genes per 100kb (Haberer et al. 2005). Genes in maize are less

likely to be grouped together, with 64% of genes being found in islands containing only one gene (Bennetzen et al. 2005).

A molecular technique used successfully for gene enrichment in maize is the construction of hypomethylated partial restriction (HMPR) libraries using methylation-sensitive *HpaII* (cut site 5'-CCGG-3') and *HpyCH4IV* (cut site 5'-ACGT-3') restriction endonucleases (Emberton et al. 2005). Methylation in DNA involves the addition of a methyl group to a 5' cytosine (converting it to 5-methyl cytosine) in the sequences 5′-CG-3′ and 5′-CNG-3′(Emberton et al. 2005). In maize, genes are usually unmethylated, while LTR retrotransposons are mostly methylated (Bennetzen et al. 1994). The same pattern of methylation has also been identified in wheat (Cantu et al. 2010). Many bacterial restriction endonucleases, such as *HpaII* and *HpyCH4IV,* cannot cut methylated DNA, which means that genic regions are much more likely to be extracted (Emberton et al. 2005). Construction of HMPR libraries with the two previously mentioned enzymes has been shown to reduce the amount of LTR retrotransposons in sequence data from 70% to less than 5% (Emberton et al. 2005). Library construction with this technique has been previously performed by Matt Estep at the University of Georgia, and the DNA was sequenced at Purdue University by Rick Westerman. Access to the data was provided by Jeff Bennetzen of the University of Georgia. The focus of this portion of the experiment was to complete a computational analysis on the previously generated data.

This technique has not yet been used to filter wheat DNA, but since wheat and maize have similar gene density and DNA methylation patterns, this technique was tested in order to determine if it could prove useful as a starting point for future studies

on bread wheat and other grasses with similar genetic structures. Library construction is only a part of the process of finding genes in large, LTR-dense genomes, as the large amounts of sequence data (often reaching several gigabytes) render manual analysis techniques impossible (Skuse et al. 2008). An array of bioinformatics tools needs to be utilized to discover genes and their regulatory systems at any useful rate (Skuse et al. 2008). In this experiment, AbySS (Assembly by Short Sequence), Blast+, and Blast2GO were used to create contigs (longer sections of DNA sequence data made from overlapping raw sequences) and annotate the sequence data, respectively. AbySS is a parallel sequence assembly program that creates contigs using a de Bruijn graph, which represents a homogenous overlap between sequences (Simpson et al. 2009). AbySS was developed specifically to address the limitations of previous sequence assembly programs, as using a de Bruijin graph allows the assembly algorithm to be calculated across a network of computers (Simpson et al. 2009). The level of stringency in AbySS is given by the *k-mer* value (Simpson et al. 2009). The *k-mer* value is used to determine all possible subsequences of the given sequence of length *k,* so if a sequence is 100 bp in length and *k* is 64, then *k-mer* refers to all of the sections in the 100 bp sequence of length 64 (Simpson et al. 2009). For the purpose of assembling contigs, sequences must overlap by *k*-1 to be considered a contig (Simpson et al. 2009). Blast+ is the newest version of the BLAST (basic local alignment search tool) search program, which searches a given sequence against a database of sequences with known function, and returns sequences with similar nucleotides (Camacho et al. 2013). Blast is the most commonly used sequence similarity search tool (Camacho et al. 2013). Blast2GO performs a three-step annotation procedure: first, input sequences can be blast

searched for homologous sequences or blast results can be imported; second, each sequence is mapped, which finds Gene Ontology (GO) terms for each known blast hit; and third, annotation, which calculates an annotation score for each GO term and selects the term with the lowest annotation score above the user-given threshold (Conesa et al. 2008).

Using AbySS, several sets of contigs were made from the raw HMPR data with a range of *k-mer* values, as well as a library of contigs from a whole genome shotgun sequence to act as a control. A whole genome shotgun sequence is an unbiased sequencing method in which all of the organism's nuclear DNA is broken up randomly and sequenced. The contigs were then blasted, mapped to GO terms, and annotated with Blast2GO. We were able to demonstrate that our gene enrichment strategy recovered a higher proportion of genes than in the control whole genome shotgun sequence.

METHODS

Library Construction (experimental)

Hypomethylated partial restriction (HMPR) libraries were constructed as previously described, using nuclear DNA extracted from bread wheat (*Triticum aestivum*) cultivar "Chinese spring" (Emberton et al. 2005, Nelson et al. 2008). Digestion was performed with restriction enzymes *HpaII* (used to generate data set 1080) and *HpyCH4IV* (used to generate data set 1081), in order to allow the enzymes to shear the DNA at unmethylated cut sites. The sequences were generated with the Illumina sequencing platform.

Whole genome shotgun sequences (Control)

DNA sequences were downloaded from EBI study ERP000319 to act as controls for our experiment. The control data (DNA from a whole genome shotgun sequence of Chinese spring) was taken from 5 μg of nebulized (sheared randomly) Chinese spring nuclear DNA, and sequences were selected of lengths between 500-800 bp (Brenchley et al. 2012). The sequences were then generated with 454 pyrosequencing in a similar fashion as the Illumina reads to generate five-fold coverage of the wheat genome, roughly 85 GB (Brenchley et al. 2012). As a control, a random selection of whole genome shotgun sequence data of roughly equal size to one of the HMPR data files ($5 * 10^6$ sequences) was selected with a Python random sampler script. Selected sequences shorter than 100 bp were eliminated, and then a random 100 bp section of each remaining sequence was taken and put into a separate file for further analysis.

Bioinformatics

The raw HMPR data (1080 and 1081) (Purdue Genomics core) was assembled into paired-end contigs using AbySS version 1.5.1 (Simpson et al. 2009) with a range of *k-mer* values (36 (the lowest allowed in AbySS), 50, and 64 (the highest *k* value allowed)) to give a wide coverage for the parameter. Resulting contigs were then compared to the NCBI non-redundant protein database using the blastx function of Blast+ version 2.2.27+, which translated each contig of nucleotides to the six potential reading frames (patterns of dividing the sequence of nucleotides into groups of three, three of which are found on each of the two strands of DNA) and then searches the resulting peptide sequences (Camacho et al. 2013). Blast2GO version 3.0 was then used to further analyze the top ten blastx hits of each contig by mapping each hit with an e-value (expect value, the possibility of finding the given sequence at random in a database the size of the one searched) of $1*10^{-6}$ or lower to Gene Ontology (GO) terms, a value which had been used in the previous study with maize (Emberton et al. 2005). Annotation was then calculated using an annotation score cutoff of 55 (Conesa et al. 2008). The control sequence data was then analyzed with AbySS, blastx, and Blast2GO in the same manner as described previously.

For each of the annotated sets of contigs for the 1080, 1081, and shotgun sequence data, the following statistics were calculated and compared with Blast2GO: the percent of contigs with blastx results, the percent of successfully mapped contigs, and the percent of contigs with annotation. Specifically, the proportions of each level of analysis in the 1080 and 1081 data were compared to the equivalent category in the control.

RESULTS

Two pairs of sequence libraries were generated from partial restriction digestions using both enzymes independently. The paired data sets for both the 1080 and 1081 data had the same number of sequences, and the amount of DNA in each data set for both enzymes was similar (Table 1). Since the DNA was sequenced with Illumina, reads averaged around 100 bp in length. While the range of sequence length varied from 30bp to 101bp, the mean length is between 92 and 96 bp, indicating that the majority of the sequences in all four HMPR sequence libraries were close to the optimum length (Table 1). Due to the sequence selection process, the randomly selected control data had sequences of exactly 100 bp in length (Table 1).

Table 1. HMPR Library Sequence Number and Size

	Sequences	Nucleotides	Length range	Mean length
1080 R2	4,229,132	404,287,056	30-101	96
1080 R1	4,229,132	391,246,103	30-101	93
1081 R1	5,378,707	502,236,596	30-101	92
1081 R2	5,378,707	517,224,183	30-101	95
Control	5,000,000	500,000,000	100	100

Sequences were classified in one of four ways: “no blast hits”, “with blast hits”, “with mapping”, and “with GO annotation”. Each classification represents the highest level of analysis achieved by the sequences in that category, so sequences in the “no blast hits” category did not match with any blastx hits, those in the “with blast hits” group did not match any GO terms, the sequences in the “with mapping” section had no GO terms whose annotation score was above the cutoff, and those in the “with GO annotation” were able to achieve the highest level of analysis (Figure 2).

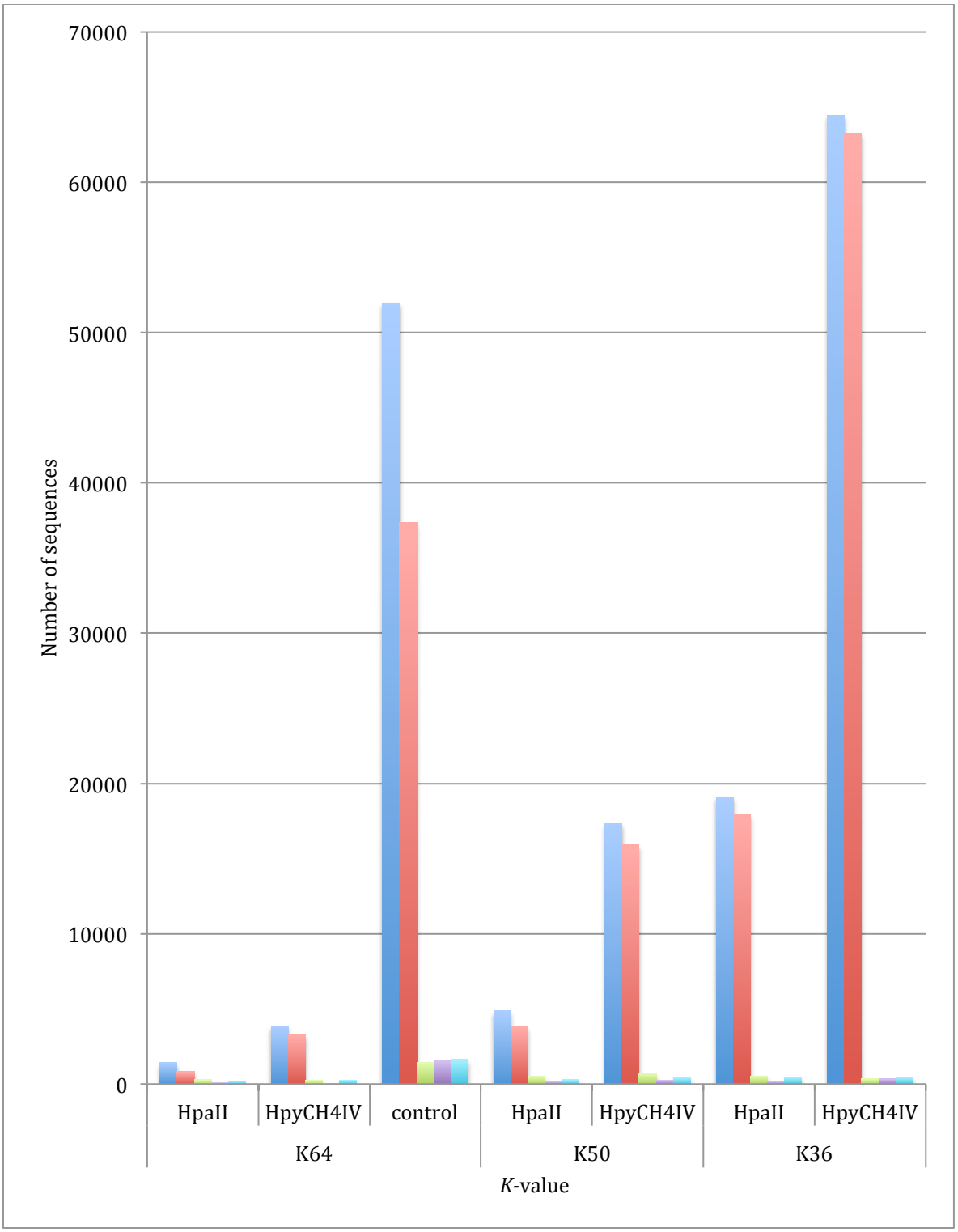

Figure 2. Distribution of Annotation Categories for *HpaII*, *HpyCH4IV,* and Shotgun Sequences (control).

This figure shows the number of sequences in each category of analysis (total sequences (blue), no blastx hits (red), with blastx hits (green), with mapping (purple), and with GO annotation (turquoise)) in the *HpaII*, *HpyCH4IV,* and control libraries.

The control sample of shotgun sequence data with a k-mer of 64 assembled 41974 contigs, 11.0% of which had blastx hits and 3.9% of which had annotation (Figure 3, Table 2). The 1080-*k*64 data had the highest percentages with 41.5% of its 1421 contigs having blastx hits and 15.3% having annotation (Figure 3, Table 2). The 1080 data generated 4860 contigs with a *k* of 50, and had 20.9% of its sequences get blast hits, and 6.4% were annotated (Figure 2, Table 2). With a *k* of 36, 17119 contigs were generated from the 1080 data, 6.2% of which had blast hits and 2.3% of which were annotated (Figure 2, Table 2). The 1081-*k*64 data generated 3852 contigs and had the second-highest percentage of sequences with blastx hits and annotation, with percentages of 14.8% with blastx hits and 7.2% having GO annotation (Figure 3, Table 2). The 1081 data with a *k* of 50 produced 17309 contigs, but had only 8.1% blast hits, and 2.6% with annotation (Figure 2, Table 2). The 1081 data assembled with a *k* of 36 had the lowest percentages of blasted and annotated sequences, as 1.8% out of its 64461 sequences had blastx hits and 0.71% had GO annotation (Figure 2, Table 2).

Table 2. Number of Contigs and Percentages of Blasted and Annotated Contigs for all Data Sets

	K value	Total Contigs	% of Contigs with blastx hits	% of Contigs with annotation
Control	64	41974	11.0	3.9
HpaII	64	1421	41.5	15.3
	50	4860	20.9	6.4
	36	17119	6.2	2.3
HpyCH4IV	64	3852	14.8	7.2
	50	17309	8.1	2.6
	36	64461	1.8	0.71

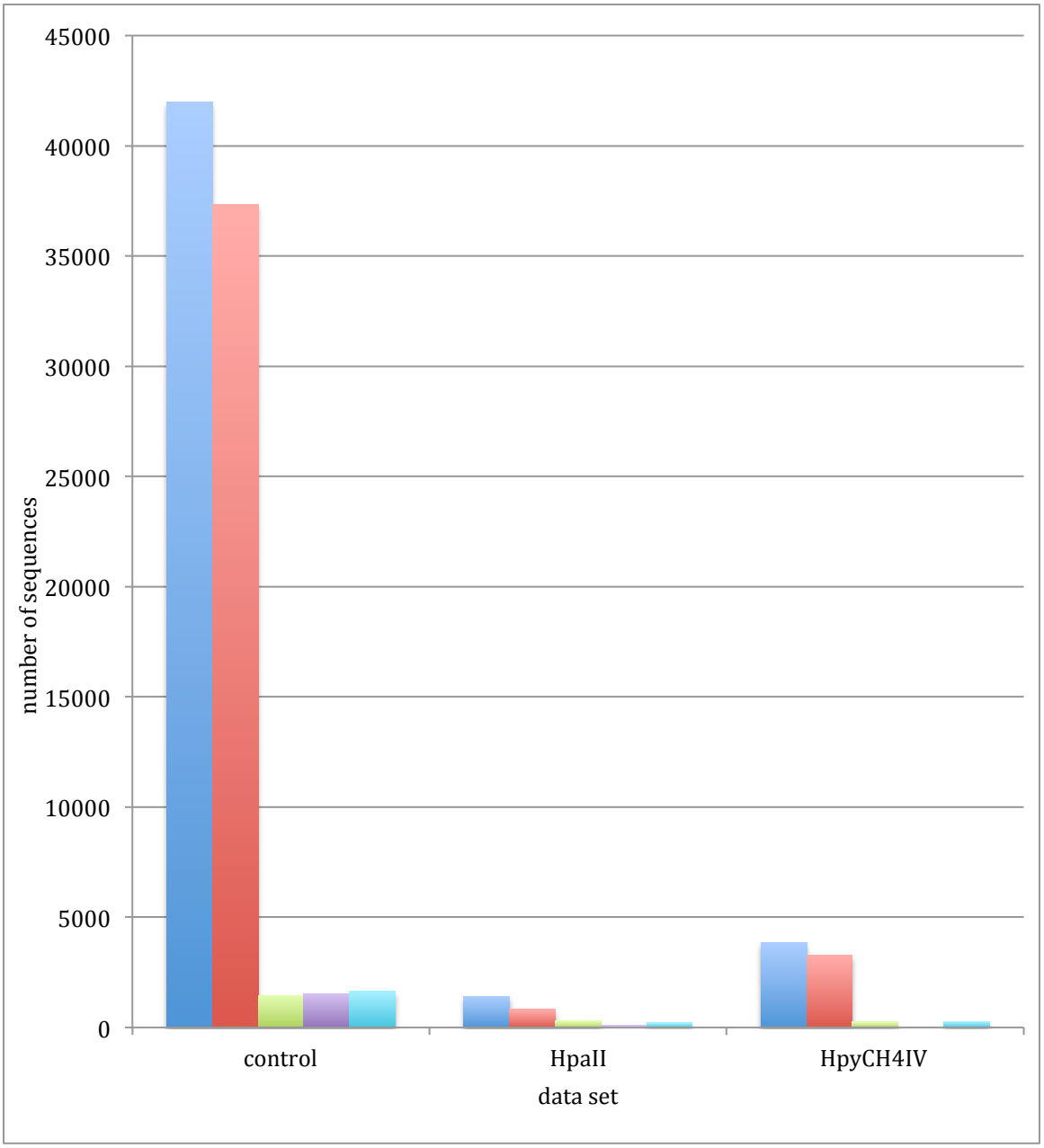

Figure 3. Data Distribution for Control, *HpaII*, and *HpyCH4IV* with K-value of 64.

This figure shows the number of sequences in each analysis category (total sequences (blue), no blastx hits (red), with blastx hits (green), with mapping (purple), and with GO annotation (turquoise)) for the control, *HpaII*, and *HpyCH4IV* assembled with a *k*-value of 64.

DISCUSSION

Bread wheat has one of the most complicated genomes of any common crop, thanks to the multiple hybridization events in its evolutionary history and the massive quantity of retrotransposons and DNA transposons constantly copying, excising, and inserting themselves throughout the genome (Charles et al. 2008; Marcussen et al. 2014). One of the main challenges to analyzing the gene content of wheat is finding the genes when they are surrounded by transposable elements (Brenchley et al. 2012). The construction of libraries with methylation sensitive restriction endonucleases and using multiple enzymes has been shown to reduce transposable element content from over 70% to less than 5% of the sequence in maize (Emberton et al. 2005). Maize and wheat have been previously shown to have very similar gene densities, even though wheat's genome is several times larger (Choulet et al. 2010). Also, both plants have demonstrated a pattern of having unmethylated genes and methylated transposable elements (Bennetzen et al. 1994; Cantu et al. 2010). The initial presumption was that since wheat and maize have a similar genetic structure and methylation patterns, and the construction of HMPR libraries was effective in filtering out transposable elements in maize, then the same technique should also be useful in bread wheat.

Computational methods have become crucial tools in gene identification. After the DNA was sequenced, bioinformatics programs were used for every step of the analysis. From making the contigs to calculating the annotation score for each GO term, it would have been impossible to gather the data in a meaningful amount of time without bioinformatics techniques. The DNA sequence itself may be easy enough to acquire, but to find genes within the sequence (not to mention how they are regulated,

organized, or expressed) in large sequences needs computational tools (Skuse et al. 2008).

There were two major components to this experiment: the contig assembly and the sequence analysis. For the contig assembly, there were multiple mentions in the AbySS literature of finding the optimum *k* value for the assembly, all of which said to run multiple trials and inspect the results. However, the literature is scarce on details of what indicates an optimal *k* value. We decided to use *k* values of 36, 50, and 64 to give a wide coverage of the parameter, representing the minimum, midpoint, and maximum *k* values that AbySS could assemble. The initial idea was that the lower *k* value would give more contigs, but the assemblies with higher *k* values would have proportionally more blasted and fully annotated sequences, which ended up being true. The results from the 1080/1081 data both supported the theory that contigs assembled with a higher *k* value a higher proportion of annotated sequences, so the control data was only analyzed fully with a *k* value of 64. In all data sets, a *k* value of 64 gave the best results proportionally, as both the 1080 and 1081 data had a higher percent of annotated contigs than the control when their contigs were assembled with a *k* value of 64.

There was a general inverse trend between the length of the *k* value and the number of contigs produced, to the point that there are orders of magnitude more sequences in the low *k* assemblies than there are in those of the high *k* (Figure 2). On the same note, as *k* decreased, the number of blasted and annotated sequences increased, but the proportions of valuable data decreased dramatically, as the number of annotated sequences would increase by several dozen, while the number of total contigs without blast hits would increase several times over (Figure 2).

The 1080 contigs produced a higher percent of blasted and annotated sequences than the 1081 dataset. In particular, the 1080 assembly made with a *k-mer* of 64 had the highest percentage of blast results and annotated sequences of all, with over 40% of sequences getting a blastx hit and over 15% getting annotated. The high *k-mer* 1080 data compared especially well to the control, having several times more blasted (3.5x) and annotated (3.9x) sequences, even with the highest possible *k-mer* for the control data. Both experimental data sets, when assembled with high *k-mer* values, had a higher percentage of successfully analyzed sequences than the control. Lower *k-mer* values yield more contigs, but higher *k-mer* contigs are more likely to be the actual DNA sequence, which helps to explain the higher percentages of blasted and annotated results in the high-*k* experimental data sets.

While not quite as effective at removing transcriptionally inactive material in bread wheat as it is in maize, the construction of HMPR libraries in the wheat genome still eliminates a large portion of repetitive sequence data when contigs are assembled with highly selective *k-mer* values. This supports the initial idea that using methylation-sensitive restriction enzymes is effective in enriching for genes in large plant genomes (Emberton et al. 2005). Less selective parameters generate a number of contigs comparable to the control, but fall short of actually producing a higher percentage of genes.

References

Bennetzen, J. (2002). "Mechanisms and rates of genome expansion and contraction in flowering plants." Genetica **115**(1): 29-36.

Bennetzen, J., R. Liu, J. Ma and A. Pontaroli (2005). "Maize genome structure and rearrangement." Maydica **50**(3/4): 387.

Bennetzen, J. L., K. Schrick, P. S. Springer, W. E. Brown and P. SanMiguel (1994). "Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA." Genome **37**(4): 565-576.

Brenchley, R., M. Spannagl, M. Pfeifer, G. L. A. Barker, R. D/'Amore, A. M. Allen, N. McKenzie, M. Kramer, A. Kerhornou, D. Bolser, S. Kay, D. Waite, M. Trick, I. Bancroft, Y. Gu, N. Huo, M.-C. Luo, S. Sehgal, B. Gill, S. Kianian, O. Anderson, P. Kersey, J. Dvorak, W. R. McCombie, A. Hall, K. F. X. Mayer, K. J. Edwards, M. W. Bevan and N. Hall (2012). "Analysis of the bread wheat genome using whole-genome shotgun sequencing." Nature **491**(7426): 705-710.

Camacho, C., T. Madden, N. Ma, T. Tao, R. Agarwala and A. Morgulis (2013). "BLAST Command Line Applications User Manual."

Cantu, D., L. Vanzetti, A. Sumner, M. Dubcovsky, M. Matvienko, A. Distelfeld, R. Michelmore and J. Dubcovsky (2010). "Small RNAs, DNA methylation and transposable elements in wheat." BMC Genomics **11**(1): 1-15.

Charles, M., H. Belcram, J. Just, C. Huneau, A. Viollet, A. Couloux, B. Segurens, M. Carter, V. Huteau, O. Coriton, R. Appels, S. Samain and B. Chalhoub (2008). "Dynamics and Differential Proliferation of Transposable Elements During the Evolution of the B and A Genomes of Wheat." Genetics **180**(2): 1071-1086.

Choulet, F., T. Wicker, C. Rustenholz, E. Paux, J. Salse, P. Leroy, S. Schlub, M.-C. Le Paslier, G. Magdelenat, C. Gonthier, A. Couloux, H. Budak, J. Breen, M. Pumphrey, S. Liu, X. Kong, J. Jia, M. Gut, D. Brunel, J. A. Anderson, B. S. Gill, R. Appels, B. Keller and C. Feuillet (2010). "Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces." The Plant Cell **22**(6): 1686-1701.

Conesa, A. and S. Götz (2008). "Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics." International Journal of Plant Genomics **2008**: 619832.
Emberton, J., J. Ma, Y. Yuan, P. SanMiguel and J. L. Bennetzen (2005). "Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries." Genome Research **15**(10): 1441-1446.

Haberer, G., S. Young, A. K. Bharti, H. Gundlach, C. Raymond, G. Fuks, E. Butler, R. A. Wing, S. Rounsley, B. Birren, C. Nusbaum, K. F. X. Mayer and J. Messing (2005). "Structure and Architecture of the Maize Genome." Plant Physiology **139**(4): 1612-1624.

Kumar, A. and J. Bennetzen (1999). "Plant retrotransposons." Annual Review of Genetics **33**: 479-532.

Marcussen, T., S. R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, The International Wheat Genome Sequencing Consortium, K. S. Jakobsen, B. B. H. Wulff, B. Steuernagel, K. F. X. Mayer and O.-A. Olsen (2014). "Ancient hybridizations among the ancestral genomes of bread wheat." Science **345**(6194).

Nelson, W., M. Luo, J. Ma, M. Estep, J. Estill, R. He, J. Talag, N. Sisneros, D. Kudrna, H. Kim, J. Ammiraju, K. Collura, A. Bharti, J. Messing, R. Wing, P. SanMiguel, J. Bennetzen and C. Soderlund (2008). "Methylation-sensitive linking libraries enhance gene-enriched sequencing of complex genomes and map DNA methylation domains." BMC Genomics **9**(1): 621.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C.-T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A.-P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J.-M. Chia, J.-M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing and R. K. Wilson (2009). "The B73 Maize Genome: Complexity, Diversity, and Dynamics." Science **326**(5956): 1112-1115.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones and İ. Birol (2009). "ABySS: A parallel assembler for short read sequence data." Genome Research **19**(6): 1117-1123.

Skuse, G. R. and C. Du (2008). "Bioinformatics Tools for Plant Genomics." International Journal of Plant Genomics **2008**: 910474.

Wessler, S. R. (2006). "Transposable elements and the evolution of eukaryotic genomes." Proceedings of the National Academy of Sciences of the United States of America **103**(47): 17600-17601.

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante and O. Panaud (2007). "A unified classification system for eukaryotic transposable elements." Nature Reviews Genetics **8**(12): 973-982.