© 2011 KRANTHI KIRAN VARALA

GENOME COMPOSITION OF *GLYCINE MAX* AND SEQUENCE DIVERSITY AMONG CULTIVATED AND EXOTIC ACCESSIONS

BY

KRANTHI KIRAN VARALA

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Crop Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Associate Professor Matthew E Hudson, Chair
Professor Brian W Diers
Assistant Professor Angela D Kent
Assistant Professor Mark A Mikel
Associate Professor Stephen P Moose

ABSTRACT

Soybean is an economically important crop in large portions of the world. Incorporation of soybean in to the food system in many direct and indirect ways has vastly increased the nutritional quality of low cost and plant-based diets. Therefore an enormous amount of effort has gone into increasing the yield and nutritional quality of soybeans through plant breeding over hundreds of years. Despite this economic and nutritional importance the soybean genome was largely uncharacterized until 2004. Research described in here deals with the application of novel sequencing technologies to elucidate the soybean genome composition as an initial step to understanding the organization of the genome.

Three, partially independent, studies were performed to study soybean genome content and diversity. The first study applied 454 pyrosequencing to obtain a low coverage survey that identified repeat composition of the genome. The second study compiled data from numerous small RNA sequence datasets to follow the small RNA level regulation of soybean genes and the maintenance of genomic stability by siRNA mediated heterochromatization. The third study applied a reduced represenatation sampling strategy to identify SNP markers in the non-repetitive regions of the genome that can distinguish between soybean accessions. The method developed in this study should be generally applicable to other lines of soybean or even in other crop plants that have a fully sequenced genome. These studies, along with others reported simultaneously, and those that will be conducted in the near future, together enhance our understanding of soybean and increase our ability to manipulate this important species to our advantage.

ACKNOWLEDGMENTS

This work was made possible by numerous contributions and kindnesses from many people. I would like to acknowledge the constructive role played by my family in my education and training. For my thesis research, I have fortuitously landed in an excellent lab and found a very able and generous advisor in Matt and, intellligent and supportive lab members. I would like to thank all these people for their help, support and tolerance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 An organism and its genome

The genome of an organism is both a catalog of its potential and a partial log of the path it has taken to acquire this potential. Unfortunately the log does not come with time stamps to determine the order of events. Insights into the evolution of a genome are often confounded by the bustle of rearrangement events in each generation of the organism. Nonetheless, by studying the genome, broad perspectives can be gained and utilized to increase our understanding of the behavior, potential and plasticity of an organism.

A vast majority of higher life forms have developed elaborate methods of gametic reproduction to increase the frequency of genetic combinations. In such species, the genome of each individual is unique at the maximum resolution i.e., to say no two individuals have completely identical genomes. Our current understanding of molecular biology indicates that the single base pair resolution of a genome is not usually the determining functional unit. Some base pairs are more important than others and, seemingly, large tracts of the genome play no role in the functioning of an individual. It has proved difficult to identify a singular resolution at which a genome can be dissected to make sense of the parts. If we define a certain granularity of the genome as containing the individual functional unit, then both the numerous subsets that a unit can be divided into and super sets that the unit is a part of seem to interact in complicated overlapping patterns to determine the scope and severity of the function.

Traditionally, single genes have served as the functional units of a genome that are assigned functions and tracked through a population of the species to explain the natural variation in form and function extant in the species. A gene though has subsequently and in ever increasing resolution been divided into multiple alleles, coding and non-coding regions, regulatory regions, and other subsets. Similarly the genomic context of the gene, a facet that differs substantially between individuals for some genes, plays a role in the regulation of the gene. Therefore

the classic definition of a gene as a functional unit responsible for one function is a gross abstraction of a much larger and refined network of influences. Despite these challenges and inconsistencies the complete sequence of bases composing a genome is a vast body of knowledge which through detailed examination and interpretation can by and large tell the story of life. Admittedly we lack the depth of understanding and perhaps even the tools to glean all the information from this extremely large and complicated data, but acquiring the data is the first step down the path of increasing our knowledge and through it our utilization of life forms to our advantage. Advances in algorithms, tools and methodologies over that past few decades has allowed us to gain tremendous insights into the inner workings of a genome and with the rapid development currently taking place in genome studies we are beginning to understand the genome at finer scales.

1.2 Genomic sequencing: methods and improvements

Since the advent of DNA sequencing in the late 70's [1] the genomes of increasingly complexity have been unraveled. Starting with a small DNA virus with a total size of 5,368 base pairs we now have sequenced genomes that are many billions of bases. Up until 2005, however, the method used to sequence DNA remained largely unchanged. Dideoxy termination sequencing, originally described by Sanger et al. [2], remained the method of choice for sequencing DNA, although significant improvements in the reagents and equipment lead to a gradual but significant increase in the number of sequential bases output in one sequencing reaction. Accompanying the increasing read lengths was the increased quality of sequence defined as the confidence in each base call made by the sequencing machine. Despite its advantages the main drawbacks of the Sanger method for genome sequencing were the need to clone and amplify a segment of DNA in a vector and the need for complex robotics and capillary/electrophoresis machines to perform the sequencing in a high-throughput manner. In addition, the cloning

step introduced a bias against sequences recalcitrant to growth in the bacterial host of choice. Therefore, the cost of Sanger sequencing although steadily decreasing, remained high at a per base level, and the throughput remained a bottleneck in all major sequencing efforts.

Vast improvements in the throughput of DNA sequencing and per base cost started to appear with the introduction of so called next-generation sequencing technologies such as pyrosequencing and sequence-by-synthesis. Both these methods used novel chemistry and amplification of immobilized DNA to increase signal. This method allowed sequencing from nano to picogram quantities of DNA. They also had the added bonus of not including a cloning step that removed the significant bias introduced therein. The single biggest advantage of these technologies, however is the tremendous increase in the number of bases sequenced in a single experiment.

Despite the tremendous advances in throughput and the exponential reduction in per-base cost offered by these methods the most important drawback of the method remains the number of sequential bases that can be confidently called (the read length). The number of bases that can be determined to occur contiguously in the original DNA molecule has very important implications on our ability to construct the whole molecular sequence from the reads.

1.2.1 Sampling of genome

Ideally we would like to determine the entire sequence of a DNA molecule such as a whole chromosome from one end to the other in one reaction without any breaks. While some highly experimental techniques like nanopore sequencing claim potential to be able to achieve this goal eventually, there is currently no reliable method of doing so. Hence, the first step in all genomic DNA sequencing efforts is to mechanically, chemically or enzymatically shear the DNA into smaller fragments. These fragments are then sequenced to generate sequence "reads",

and overlapping reads are stitched together in an assembly step to regenerate the sequence of bases in the original DNA molecule.

The initial genome sequencing efforts, including the human genome project in the initial stages, sheared genomic DNA into approximately 120 Kb fragments and cloned them into Bacterial Artificial Chromosomes (BACs) followed by a careful selection of BACs to sequence based on a golden path approach. The golden path approach attempts to direct the selection of clones for sequencing such that the genomic sequence is read more or less sequentially wherever possible. While this approach vastly simplifies the assembly stage of the genome, it tends to be very labor and time intensive.

A major improvement in this strategy was introduced by Fleishmann et al. [3] who suggested abandoning the time-consuming golden path approach and randomly fragmenting and sequencing the BACs until enough coverage was obtained to programmatically assemble local regions of the genome. This strategy was labeled shotgun sequencing for obvious reasons and was made viable by advances in sequencing throughput and computing capabilities. Shotgun sequencing abandons the rigor of sequential selection to take advantage of massive sequencing capacity, but at the cost of clarity in the assembly stage. This approach essentially translates to taking random samples of sequence from the genome. The population, i.e.. total number of such sequences in the genome, is finite. Therefore, a mathematical model can be used to describe how the number of times a genomic region is sequenced increases as the sampling frequency increases.

1.2.2 Lander-Waterman model and genome complexity

Lander and Waterman [4] first described the relation between the number of sequences sampled from a genome of a given size, the chance of repeated sampling from a region for a given read length and the required length of overlap for detecting such a repetition. Assuming that genomic regions are sampled following a

Poisson distribution, they provide an estimate of the number of genomic regions expected to have a given number of overlapping reads. These overlapping reads can then be assembled programmatically to form a contiguous sequence (contig). The underlying assumption here is that two reads would share the same sequence of bases for a significant length (determined by the overlap parameter L) only when these reads are derived from the same genomic region. The number of contigs expected to contain j reads is defined by,

$$Ne^{-2c\sigma}\,(1-e^{c\sigma})^{j-1}$$

$$c=\frac{LN}{G},\sigma=1-\frac{T}{L}$$

where N is the number of reads, L is the read length, G the haploid genome size in base pairs, and T the minimum base pair overlap required for contig formation. The number c, called coverage, is a measure of how many equivalents of the genome length have been captured in the sequencing effort. This model provides an excellent way of estimating the minimum coverage required to assemble most of the genome into large contigs, for a given genome and read length, assuming that most genomic regions are unique over the length of the overlap parameter. If all the genome sequence is unique at this level, the coverage required to assemble the genome increases exponentially as the read length decreases. As the length of the genome increases, the likelihood of regions that are insufficiently covered also increases following a Poisson distribution.

The genomes of most eukaryotes, especially the species with fairly large genomes are filled with many repeats that are long and share a very high degree of sequence identity. This means that the fundamental assumption that two reads with significant overlap must have been sampled from the same region is violated. Therefore, conserved repeats constitute the biggest hindrance to an unambiguous assembly. In certain regions, a series of tandem repeats or a group of highly similar repeats are present with little interspersed unique DNA. Sequencing reads derived from

such regions do not span across enough unique regions to differentiate between copies of the repeats. Such reads can confuse the assembly program by showing sufficient overlapping bases for assembly leading to the joining of genomic regions that are not adjacent to each other in the genome. Such mis-joins are highly deleterious to the quality of the assembly and are usually avoided by removing or in other ways discarding all reads that have spuriously high degrees of overlap with other reads.

The average length of sequence that can be read in a single reaction is therefore a very important factor that determines both the number of reads required to achieve enough coverage and the ability to unambiguously assemble reads into contigs. Ideally, we would like to read the entire sequence of bases in a single continuous run to determine the genome without any additional analysis needed. Improvements in equipment and reagents used in Sanger dideoxy termination sequencing pushed the maximum read lengths to 1000 bp and above and allow the reading of "paired ends" from both sides of a molecule of defined length. While the increased read lengths and paired reads improved the ease of assembly substantially, genome sequencing of higher eukaryotes remains a substantial challenge due to their very large size and the complexity caused by numerous repeats.

1.3 Next generation sequencing: throughput and read lengths

Starting in 2005, a series of novel sequencing platforms have emerged. These platforms introduced massive changes in the sequencing methodologies and outputs. The three major technologies to emerge in this time are the pyrosequencing platform from 454/Roche, Genome Analyzer by Solexa/Illumina and SOLiD by Applied Biosciences (AB). Together these platforms have been referred to as next-generation sequencing (NGS) platforms. NGS platforms were developed in response to the increasing demand to lower the per base cost of sequencing to al-

low the rapid resequencing of genomes and to allow creation of reference genomes for a wider range of organisms. Hence the focus for the newer methods was always on increasing throughput massively so as to reduce the time and labor costs involved in traditional Sanger sequencing.

The common solution arrived at in all three platforms was to immobilize a strand of DNA on a solid platform, amplify it *in situ* and perform the sequencing on this immobilized cluster of DNA molecules that originate from a single DNA fragment. Pyrosequencing and Illumina rely on the incorporation of nucleotides due to the action of DNA polymerase and the coupled release of a signal. AB SOLiD relies on competitive ligation of short oligos to the template to read the sequence. Despite the differences in chemistry these platforms share a set of strengths and weaknesses. The clear advantages of NGS platforms over Sanger based sequencing is the reduced bias in representation of genomic regions and massive throughput while the drawbacks are the much shorter read lengths, an overall lower quality and significant informatics challenges in processing the data.

454 sequencing by far produces the longest reads of the three with the latest versions reaching over 500 bp average read length, while Illumina's latest offering 100-150 bp and SOLiD reaching 75bp. In terms of total throughput per run, Illumina and SOLiD generate about the same number of bases and 454 produces significantly less. Further improvements in technology are expected to increase both read lengths and total throughput across the board in the near future. These developments suddenly made practical a number of approaches to study the role of genetic material in a wide range of phenomena in ecology [5], evolution [6, 7, 8], and genetics [9, 10, 11, 12]. Novel approaches could now be devised to better solve long-standing problems in agriculture [13, 14].

1.4 Applications in crop species

Most agronomically important plant species carry a very large genome that is further complicated by the presence of copious numbers of repeat elements. The process of domestication might itself encourage the proliferation of certain repeat elements [Chapter 4]. Therefore, the determination of the complete genomic sequence of a crop species is extremely challenging. The most agronomically important crops (rice, maize, soybean) already have a reference genome available, but the amount of diversity within each of these species is tremendous and is not represented in the reference genomic sequence. Generating a genome for multiple lines from the same species through clone-based sequencing is still cost-prohibitive and will likely remain so in the foreseeable future. Also, the vast majority of crop species do not yet have a fully sequenced representative genome. It is here that the NGS platforms have a vital role to play. Applications of NGS technologies in various novel and creative designs are rapidly expanding our knowledge of crop genomes which in turn will prove extremely useful in the improvement of these crops.

NGS technologies have serious limitations in *de novo* sequencing of a new genome due to in read length. Recent advances in read length and improvements in assembly algorithms have overcome this limitation somewhat [15]. Nonetheless *de novo* sequencing of a crop species purelyby NGS sequencing is unlikely to yield a sufficiently scaffolded genome assembly in the near future. Therefore, over the past few years NGS platforms have been used primarily to obtain preliminary genome (or RNA) sequence information in species with little to no prior information [9, 16] or to sequence the coding regions in distant relatives of a known genome [6, 16]. In light of the availability of NGS platforms, and the paucity of information available for soybean at the time this work began, this thesis research has focused on the application of these technologies to soybean to determine the composition, regulatory content and extant variation of its genome.

Soybean genetics traditionally has focused on the improvement of crop varieties through incorporation of beneficial alleles and traits through breeding. Marker assisted breeding in the last few decades has brought together the realms of breeding and molecular biology culminating in the development of a high-resolution genetic map [17] and a Department of energy (DOE) led genomic sequencing effort [18]. Prior to the availability of the soybean genome, the largest set of information available for soybean was the extensive sampling of coding regions achieved through the soybean expressed sequence tag (EST) project [19]. Gene space sampling is the logical first step in characterizing a complex genome. EST sequencing reveals the important coding regions of the genome but critically lacks information on the genomic linkage, context and regulatory information associated with those genes. Also lacking is the evolutionarily and structurally important repeat composition of the genome.

Repeat elements play a crucial role in the rearrangement of plant genomes that allows the plant to respond to environmental changes [20]. Knowledge of repeat composition also informs decisions in sequencing efforts to guide clone choice in large sequencing efforts and primer design for more focused studies. Identification of centromeric repeats provides a useful cytogenetic tool to individually identify chromosomes for karyotyping. To determine the repeat composition and distribution of soybean, an early 454 pyrosequencing run was performed using complete soybean genomic DNA [Chapter 2]. This study revealed the repeat composition of the soybean genome in terms of the total repeat content, various families of repeats and copy numbers of different repeat elements. In addition the data obtained allowed the identification of the most abundant repeat families and the tandem repeated units within it that led to the flourescence in situ hybridization (FISH) based karyotyping of the soybean genome[10].

With the availability of the soybean genome, the characterization of the coding regions and their immediately adjacent regulatory regions was completed. This high-confidence gene set allows the study of coding regions and their tran-

scriptional regulation through promoter regions. Advances in the study of gene regulation over the past two decades have revealed the vital role of non-coding RNAs in the epigenetic and post-transcriptional levels of gene regulation. Small RNAs (siRNAs and miRNAs) constitute the most important class of non-coding RNAs in plants but identification of these functional molecules from the genomic or transcriptomic information is non-trivial. The very short read lengths, but exceptionally high throughput provided by Illumina sequencing is ideal for profiling small RNA content in plants. Therefore, multiple studies have been designed to elucidate the role of small RNAs in various biological phenomena in soybean. By leveraging the soybean genome information the combined sequence output of these studies would readily reveal the global profile of small RNA producing loci in the soybean genome. Chapter 3 deals with the identification of siRNA producing/target loci and putative miRNA producing loci in the soybean genome. This study helped characterize the active transposon population in the soybean genome and identified many novel putative miRNA in soybean, some of which are conserved in other plants.

The most powerful application of NGS platforms is the genomic sequencing of close relatives of a species with a well characterized high-quality genome. Especially the Illumina and AB SOLiD platforms with their short, relatively high quality reads and higher throughput are particularly suited to this task. This technique is especially powerful in differentiating the genetically close accessions of the same species. Soybean has a very rich domestication and breeding history stretching back thousands of years in East Asia. The elite i.e., high yielding cultivars of soybean are constantly improved by introgression of desirable traits from lower yielding exotic accessions and this process is greatly aided by molecular markers. Therefore, reliable identification of markers that distinguish between two accessions of interest is extremely useful to soybean breeders. Chapter 4 deals with the development of a protocol to rapidly identify SNP markers that are polymorphic between any two lines of soybean. Reduced representation sampling of the soy-

bean genome, achieved by anchoring read starts to a carefully chosen restriction enzyme site, allowed the generation of thousands of high-confidence SNP markers between the chosen accessions. This study also revealed a surprising amount of heterogeneity within the reference soybean accession (Williams 82) chosen for genome sequencing.

Through the application of NGS technologies, both before and after the availability of the soybean genome, significant knowledge of the genomic content and its interpretation was gained in these studies. Similar studies in soybean and other important crop plants have been attempted in recent years. The continued usage of NGS technologies in studying crops should generate vast datasets that improve our understanding of crop genomes, their spatio-temporal regulation in the organ and developmental dimensions of the plant. Beyond the already sequenced genomes the applications of these methods in closely related species will reveal gene space differences among them. Such information would help illuminate the molecular mechanisms involved in domestication and development of resistance to various biotic and abiotic stresses.

Eventually, with enough sampling of genome composition across the diverse set of life forms, fundamental life processes will be better understood at a molecular level. Specific adaptations that allow life to fit ecological niches and the various paths available to a genome to trans-locate from one niche to another could also be elucidated by studying the molecular events underlying such changes. Population level studies of the genomic variation within a species would reveal the many-to-many relationships linking underlying genotypes and the phenotypes generated by their complex interactions. Sequencing technologies involving single molecule sequencing with no amplification biases are currently being developed and would hopefully offer more unbiased sampling of genomes. These methods will be particularly useful in studying metagenomes such as the rhizosphere. Metagenomes would help explain the entire gene space that is not necessarily intrinsic to the plant but in some cases could offer a better explanation for the observed phe-

notype. The next few decades will offer exciting new perspectives and means for observing organisms that are essential for the sustainable existence of human population.

CHAPTER 2

REPEAT COMPOSITION AND LANDSCAPE OF THE *GLYCINE MAX* GENOME

2.1 Global repeat discovery and estimation of genomic copy number in large, complex genome using a high-throughput 454 sequence survey [1]

2.1.1 Introduction

Genome sequencing has historically been accomplished by fragmenting genomic DNA, amplifying the fragments clonally using bacteria, and sequencing the amplified clones [21]. Although this method has improved to the extent that much larger genomes can be sequenced, and some of the intermediate cloning steps can be circumvented [3, 22], practically all genome sequence until very recently has been generated by the Sanger method. Given the costs of Sanger-based genome sequencing and surveys, significant amounts of genomic information for most of the 129,293 eukaryotic species listed in the NCBI taxonomy database [23] are unlikely to be available for some time. Soybean [Glycine max (L.) Mer], which is the subject of this study, has an existing but incomplete genome project. However, many crop plants, plant pathogens, endangered species and species of evolutionary interest have little or no available genome data. Recently developed microbead technologies capable of sequencing hundreds of thousands of DNA molecules in parallel provide a way to obtain genomic information from these species for reasonable cost, and without any bacterial cloning step. The method used here, 454 pyrosequencing, uses pyrophosphate release as a method for detection of base in-

[1] This section of Chapter 2 was previously published in BMC Genomics 2007, 8:132. The article is reproduced verbatim except for the addition of Figures 2.3 to 2.5 and the their descriptions. The article was published under the Creative Commons Attribution License with the express statement that the copyright for the work is retained by the authors. Authors contributions KS performed analysis of Bioinformatics data and comparison of databases and laboratory experiments to validate predicted repeats, created the data displays and helped draft the manuscripts. KV developed and implemented assembly and database bioinformatics methods, implemented and performed parallel analysis and annotation of sequences and repeats, and developed the web interface for the database and alignment viewer. MEH conceived the study, design, co-ordination and manuscript, developed and implemented the DNA extraction procedure, assembly and repeat detection, analysis of copy number and the remaining bioinformatics and scripting. Specifically Figures 2.1 and 2.6 and Tables 2.1 and 2.2 were generated by other authors.

corporation [24, 25, 26]. Pyrosequencing has been used before to genotype SNPs in a polyploid plant, potato. However, the technology used [27] relied on known primer sequences, greatly limiting the utility of the method for de novo sequencing. The 454 pyrosequencing method uses randomly sheared DNA, has no requirement for known primer sequences (making it suitable for de novo sequence surveys), and makes sequence data faster and cheaper to obtain than Sanger-based methods. However, the accuracy and read length of the method as used here is generally inferior to Sanger-based sequencing of small clones [28].

The first step in characterizing large genomes has frequently been a genome survey, often using end sequences of Bacterial Artificial Chromosome (BAC) vectors [29, 30]. Such a survey gives important information about common repeat sequences, allows the generation of some genetic markers and helps determine the feasibility of building a BAC tiling path. Such surveys are limited, however, as representation of cloned sequences is likely to be somewhat skewed towards those that can be successfully propagated in bacterial vectors [31]. Here we describe a method for performing high coverage, inexpensive and detailed genome surveys without the necessity of cloning, bacteria or vector libraries. The 454 pyrosequencing method described by Margulies et al. [28] allows access to randomly placed, short sequences in large numbers, without the generation of bacterial vectors or a cloning step. Since 454 pyrosequencing produces relatively short reads, without paired end information, it is currently unsuitable for de novo sequencing of eukaryotic whole genomes. However, a high-coverage genome survey using this method can potentially deliver invaluable data about the makeup of a genome, quickly and at relatively low cost. In particular, the identification of sequences present in many copies per genome (essential in order to generate a unique tiling path for a structured sequencing approach) is straightforward.

The soybean genome is relatively well-characterized, and significant progress has been made towards its completion. A survey of BAC clone ends has been performed at relatively low coverage on the soybean genome [29], and extensive

sequencing of soybean ESTs has been performed [19]. However, a complete physical map is not yet available, and the amount of soybean genomic sequence in the public domain is still somewhat limited, although now growing rapidly. The survey described here provides further information about the makeup of the genome of this crop of great commercial importance.

2.1.2 Methods for DNA isolation and sequencing

Soybean nuclear genomic DNA isolation

8 g of young trifoliate leaves were taken from soybean cv. Williams, grown under controlled greenhouse conditions in sterile soil. Leaves were ground to coarse powder in N2(l), transferred to 20 ml NIB (Modified from Zhang et al [32]; 10 mM Tris, 10 mM EDTA, 100 mM KCl, 500 mM sucrose, 4 mM spermidine and 0.1% -mercaptoethanol), and placed on ice for 10' with swirling every 1'. The suspension was filtered through 2 layers of Miracloth and 2 layers of cheesecloth, and 1 ml 10% Triton X-100 in NIB added. The suspension was incubated on ice for a further 10' with swirling every 1', then centrifuged at 2,000 g for 15' at 4C. The supernatant was discarded and the pellet resuspended in 20 ml NIB. After centrifugation for 2' at 100 g, the supernatant was transferred to a fresh tube and the pellet discarded. After centrifugation for 15' at 2,000 g, the supernatant was discarded and the pellet inspected for any green coloration. The centrifugation and resuspension steps were repeated until the pellet was pure white in color. Once free of visible chloroplast contamination, the pellet was resuspended in 10 ml TE (10 mM Tris pH 7.5, 1 mM EDTA). 1 ml of 10% sodium lauryl sulphate was added and 50 mg protease K powder. The resulting suspension was incubated for 48 h at 37C with slow orbital shaking. 1 ml 3 M sodium acetate pH 5.3 and 10 ml phenol/chloroform/IAA were added, the solution was gently emulsified and centrifuged for 5' at 10,000 g. The aqueous phase was removed and the extraction

repeated. 25 ml ethanol was added, the contents mixed and incubated at 20C for 14 h, followed by centrifugation at 13,000 g for 15'. The pellet was washed twice in 70% aqueous ethanol, resuspended in 100 l TE, reprecipitated by the addition of 100 l isopropanol, centrifuged at 13,000 g for 15' and resuspended in 100 l TE.

DNA sequencing

DNA sequencing was performed on sheared soybean genomic DNA isolated as above by 454 Life Sciences, Branford, CT at the 454 Sequencing Center using the GS-20 instrument [28]. Two 1.6 million-picowell plates were sequenced, and reads were filtered and trimmed to 5% or fewer marginal calls as described [28]. Further trimming was then performed based on the phred-equivalent quality score for each base (-10 log P(e), [33]). The reads were further trimmed of leading and trailing bases where $Q < 10$, in order to ensure comparable data to BAC-based surveys [29]. The mean Q value was 26 across the sequences after filtering and trimming, and the longest read was 410 bases and the shortest 35 bases, with a standard deviation of the mean length of 18 base pairs. In the filtered, trimmed sequences, 95% of bases were Q10 or higher, 83% were Q20 or higher. While these quality scores are relatively low by comparison to automated Sanger sequencing of small clones, they are comparable to the levels of quality obtained in whole-genome sample sequencing of the soybean genome using BAC end sequences [29]. Possible contaminants resembling organellar sequences were counted, but not removed, since reads with sequence identity to organelle sequences may be derived from organellar DNA or be genuine genomic sequence with high similarity to the organellar genome. A total of 6,819 reads (0.9%) showed significant (1E-6 BLAST (blastn) hit) identity to a collection of available chloroplast sequences, 958 reads (0.1%) showed a similar level of identity to a collection of mitochondrial sequences.

For organellar contaminant estimation, fully assembled soybean chloroplast and mitochondrial sequences were not available; chloroplast and mitochondrial genome sequence from plants including all available soybean data were assembled into a BLAST database in-house. Of the remaining reads, the overall GC content of the sequence was 33%. The full sequence dataset of the soybean 454 genome survey has been deposited at the NCBI Trace Archive, TI range 1732557604-1733276192. Assemblies are available from the authors on request or at their web site [34].

2.1.3 Results

Genomic DNA was extracted from purified nuclei isolated from leaves of soybean cv. Williams. The DNA was randomly sheared, and sequenced using the 454 pyrosequencing method [28]. This resulted in 717,383 successful sequence reads, together with phred-equivalent quality (Q) values [33]. Mean read length of these filtered, trimmed reads was 109.5 base pairs (bp), with a total of 78,535,105 bp of sequence generated. The soybean haploid genome size has been estimated at 1,115 million base pairs (Mb) [35], therefore the filtered, trimmed reads used in this sequence survey represent an estimated 7% coverage of the soybean genome.

The 103-Kb region surrounding the CHS locus of soybean has been extensively characterized [36]. We utilized the sequence of this region to probe the genomic distribution and accuracy of the genomic survey sequences. Using BLAT [37], 102 reads with 95% or higher identity across 98% or more of the read to this validated sequence were identified. These reads represent 10,542 base pairs of sequence with an overall 97.7% match to the validated sequence, hence there is a minimum estimated error rate of 2.3%. The presence of slightly more than the expected number of matching reads within the pyrosequencing dataset provides evidence that the estimated genome size of soybean [35] is approximately correct.

Analysis of high-abundance sequences in soybean

Repetitive sequences can confound both common methods for de novo genome sequencing: conventional, tiling-path based assembly strategies and shotgun genome sequencing approaches. Consequently, we aimed to develop accurate repeat detection methods and comprehensive cataloging of repetitive sequences.

Using the annotated TIGR databases [38] from multiple species, we are able to estimate the genomic copy number of all of the repeat classes in the TIGR collection. These repeats may be detected either through similarity to Glycine or to repeats known from other plant genomes, including the completed genomes of Arabidopsis and Oryza.

The TIGR plant repeat database is composed largely of transposable element sequences and noncoding RNA genes, and as with any database using incompletely sequenced genome data, it is incomplete. Satellite sequences such as those detected in the assembly of our own soybean repeat database are under-represented in the TIGR repeat database, despite their presence in GenBank, and the types of repeat and organisms of origin of the sequences vary.

For each of the 717,383 reads, we searched for a significant (e ¡ 1E-6) BLAST (blastn) sequence match to the TIGR plant repeat database, which is organized both by species of origin and class of element. Figure 2.1A shows the percentage of reads with top hits that matched each species represented in the TIGR database. The most abundant matches are those to repetitive elements already known to exist in Glycine max. Since the most abundant sequences in soybean are also the most likely to be well-characterized in this organism, this was an expected result. However, the database contains other legume repeat sequences: 64 sequences from Lotus species, 128 from Medicago species, as well as 130 from Glycine species. We were surprised that the Lotus and Medicago matches were not more abundant. We speculate that this may be because the Lotus, Medicago and soybean sequences are mostly related, and hence the reads with a match to legume repeats generally

have their best (lowest blastn expect value) match to the Glycine sequences. Note that most of the de novo detected repeats from our survey, including the SB92 and STR120 satellites (present in the GenBank nucleotide (nt) database) and many retrotransposons described in Table 2.2 (many of which are not present in nt), were not present in the TIGR database.

Figures

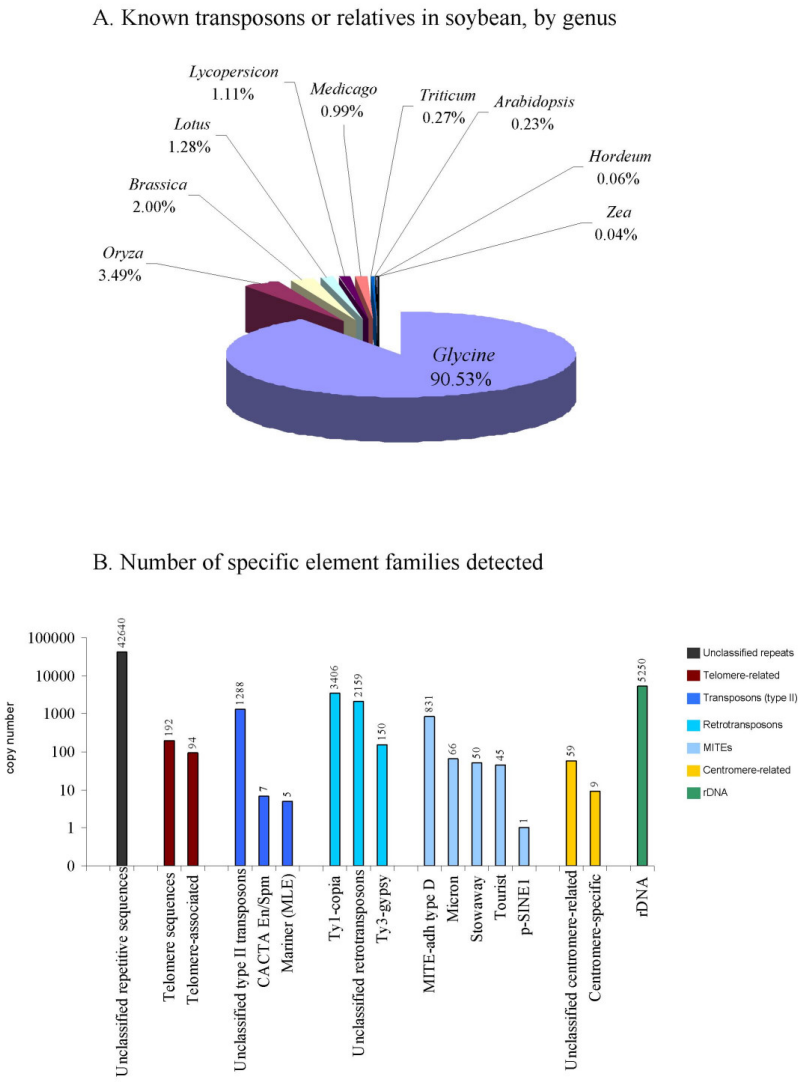

Figure 2.1: **Comparison of sequence survey data with soybean and other plant repeat databases.** A) Distribution of hits to plant repeat databases, by genus. Raw reads were matched using BLAST (blastn) to the TIGR plant repeat databases and the top significant (1E-6) hit recorded. Percentages represent the percentage of reads with hits to sequences from a particular organism with respect to all reads with hits to the TIGR repeats. B) Distribution of hits to plant repeat databases, by class of repetitive element Raw reads from the genomic sequence survey were matched to the combined plant repeat databases as for (A), and the class of repetitive element for the top hit was used to show the relative abundance of different classes of repetitive elements. This gives an estimate of the relative frequency of these families in the soybean genome. Retrotransposons and rDNA are the most common classes of repeat. See Additional File 1 for common repeat sequences not included in the TIGR database.

Relatively few reads matched the repeat database for Arabidopsis. Most of the reads similar to repeat sequences from other plants (i.e. elements that were previously unknown in Glycine) had their most significant similarity to sequences from Oryza or from Brassica. Brassica is more closely related to soybean than Oryza, but has been the subject of very limited genomic sequencing, while Oryza has been completely sequenced but is much more distantly related to soybean than Brassica or Arabidopsis. The result that there are more similarities between Glycine repeats and Oryza repeats, and between Glycine and the known repeats from Brassica, than to Arabidopsis was therefore unexpected.

This analysis also allows description in broad terms of the abundance of transposable element sequence families in Glycine, given the presence of related sequences in the database used for comparison. Regardless of species of origin, a family was assigned to each soybean sequence read with a significant ($<$ 1E-6) BLAST (blastn) hit to the TIGR plant repeat database. Figure 2.1B gives an overview of the repeat composition of the soybean genome and an expected minimum genome copy number for each element type found in the database used. Again, we cannot expect the reference database to be in any way complete, so no conclusions regarding absent sequences can be made. We estimate that the soybean genome contains a minimum of just over 8,000 transposable elements of types named and present in the TIGR repeat database; many more "unclassified repetitive" sequences that have similarity to sequences in this dataset (at least 42,000) are present. One result that arises from this analysis is that while retrotransposons are common in the soybean genome, Type II transposons are likely to be relatively rare (several examples are present in the database, but few match our soybean survey). More noteworthy was that no hits were observed to MULE (MUtator-Like Element) transposons in the TIGR collection. It is likely therefore that soybean MULEs are sufficiently divergent in sequence from any MULEs in the TIGR repeat database that they are not detected by a BLAST (blastn) search. Conversely, while MITEs (Miniature Inverted-repeat Transpos-

able Elements) were not previously present in the TIGR soybean repeat database, sequences with hits to MITE elements from other organisms in the TIGR plant repeats are present in many different classes, indicating the presence of several MITE families in the soybean genome.

De novo detection of abundant sequences in the soybean genome

Identification of repetitive elements using high throughput survey sequencing is not limited to sequence homology searches to known repeats from other genomes. Repeats can also be identified based on their over-representation in the data set. By clustering non-cognate, overlapping DNA sequence fragments using phrap [39] we were able to identify a comprehensive set of sequences present in many copies in the soybean genome.

The expected number of cognate contigs obtained by sequencing 7% of a non-repetitive genome was calculated according to Lander and Waterman [4]. Since 7% of the genome was sequenced, assembly into non-cognate contigs allows detection of sequences present in 14 copies or more per genome. The observed excess of overlapping sequences from phrap assembly was used to estimate the relative amount of repetitive DNA present in 14 or more copies in the soybean genome. These calculations are summarized in Table 1. Note that most (81%) of the predicted repetitive sequence is in contigs that contain more than seven reads, all of which are likely non-cognate since none are expected to be generated by chance from non-repetitive DNA. In total, approximately 41% of the total reads in the survey (293,889 out of 717,383) were found to form contigs only expected to assemble if the underlying sequence is present in multiple copies [Table 1]. We thus estimate that 41% of the soybean genome is present in more than 14 copies per haploid set. Most of these repeats (comprising an estimated 33% of the soybean genome) are present in 100 or more copies.

This estimate is in strong agreement with past DNA re-association kinetics

(Cot measurements), which predict that 30-45% of the soybean genome consists of highly repetitive DNA, with the total repeat content in the range of 40-60% [40, 41]. However, unlike Cot measurements, this method gives access to the underlying sequence of the detected repeats.

Tables

Table 2.1: **Lander-Waterman model predictions for expected contig depth.** Repetitive sequences in the soybean genome quantified using the difference between the contigs produced by an assembly algorithm with conservative parameters, and the predictions of the Lander-Waterman model for sampling a completely non-repetitive genome

Number of reads in contig	Predicted by model	Observed number of contigs	Repetitive reads (Observed-predicted)
2	41,126	42,221	2,189
3	2,511	9,742	21,693
4	153	3,498	13,379
5	9	1,646	8,183
6	1	937	5,619
7	0	634	4,438
>7	0	4,213	238,389
			total 293,890

Our assembly yielded 20,670 predicted repetitive contigs (contigs assembled with three or more reads per contig). The Missouri repeat database [42] contains 348 sequences, the soybase.org collection [43] 5,010 repeats, and the TIGR Glycine repeat database [38] 130 sequences. Using BLAST with an e value cutoff of 1E-6, we determined that our repeat database contains 19,274 repeats with no similar sequences in the Missouri collection, 16,261 repeats with no similar sequences in the soymap.org collection, and 20,124 with no similar sequences present in the TIGR Glycine repeat database (although more reads from our survey show significant similarity to TIGR repeats from other organisms, as discussed above).

The most abundant repetitive sequences which assembled into higher-order sequence structures were the 92 bp repeat family (GI:402616); these are present in multiple distinct contigs of higher-order repeats (Table2.2). In total, 26,714 reads, or 3.7% of the soybean genome sequence, are contained in SB92-like higher-order repeats. However, the published SB92 repeat sequence, which is found in centromeres as well as other genomic locations in the annual soybeans [44] matches only 4,567 reads by BLAST (blastn with e < 1E-6). This indicates the variability

of the repeat units within the higher-order contigs, many of which are not close enough to the published, canonical SB92 sequence to match it in our BLAST search. This is consistent with observations [44] that the SB92 repeat has a high level of sequence diversity. A total of 51 contigs contain SB92-like sequences (the most abundant are shown in Table 2.2), but these sequences do not assemble into a single unit. This indicates that distinct subtypes and higher-order structures of this satellite sequence are present in the soybean genome.

Table 2.2: Forty most abundant higher-order repeat sequences in soybean.

ContigID	Length	% of genome	Best Genbank hit	Repeat family
80377	13386	0.36	emb|Z26334.1|GMP3X1SAT G.max satellite DNA	SB92 repeat
80376	13092	0.33	emb|Z26334.1|GMP3X1SAT G.max satellite DNA	SB92 repeat
80375	9911	0.26	emb|Z26334.1|GMP3X1SAT G.max satellite DNA	SB92 repeat
80374	8916	0.25	gb|U26701.1|GMU26701 Glycine max satellite STR120-B.1	STR120 satellite
80373	6678	0.23	gb|AF186186.1|AF186186 Glycine max retrovirus-like element Calypso5-1	STR120 satellite and a retroelement
80372	6743	0.23	emb|Z26334.1|GMP3X1SAT G.max satellite DNA	SB92 repeat
80371	9930	0.21	emb|Z26334.1|GMP3X1SAT G.max satellite DNA	SB92 repeat
80370	8197	0.19	emb|Z26334.1|GMP3X1SAT G.max satellite DNA	SB92 repeat
80369	8269	0.16	gb|U26698.1|GMU26698 Glycine max satellite STR120-A.2	STR120 satellite
80368	9309	0.16	gb|AF297983.1|AF297983 Glycine max TRS1 tandem repeat region	SB92 repeat
80367	6325	0.15	previously undescribed retroelement	SIRE
80366	5613	0.14	gb|AF297985.1| Glycine max TRS3 tandem repeat region	SB92 repeat
80365	7401	0.13	gb|AF297985.1| Glycine max TRS3 tandem repeat region	SB92 repeat
80364	3789	0.12	gb|AF297983.1|AF297983 Glycine max TRS1 repetitive repeat region	SB92 repeat
80363	5168	0.12	gb|U26699.1|GMU26699 Glycine max satellite STR120-A.3	SB92 repeat

continued on next page

Table 2.2 continued

ContigID	Length	% of genome	Best Genbank hit	Repeat family
80362	4505	0.12	previously undescribed retroelement	calypso / diaspora
80361	6307	0.12	gb|AF297983.1|AF297983 Glycine max TRS1 tandem repeat region	SB92 repeat
80360	5757	0.12	gb|AF297985.1| Glycine max TRS3 tandem repeat region	SB92 repeat
80359	6040	0.11	unknown rpt sequence	found in soy ESTs
80358	5454	0.11	previously undescribed retroelement	calypso / diaspora
80357	5620	0.11	gb|AF297985.1| Glycine max TRS3 tandem repeat region	SB92 repeat
80356	4775	0.11	emb|Z26334.1|GMP3X1SAT G.max satellite DNA	SB92 repeat
80355	2945	0.11	previously undescribed retroelement	calypso / diaspora
80354	4673	0.11	previously undescribed retroelement	SIRE
80353	2688	0.1	18S ribosomal RNA	rRNA
80352	4832	0.1	previously undescribed retroelement	SIRE
80351	5601	0.1	previously undescribed retroelement	calypso / diaspora
80350	3773	0.09	previously undescribed retroelement	SIRE
80349	5318	0.09	gb|AF297983.1|AF297983 Glycine max TRS1 tandem repeat region	SB92 repeat
80348	3781	0.09	emb|Z26334.1|GMP3X1SAT G.max satellite DNA	SB92 repeat
80347	4451	0.09	previously undescribed retroelement	Calypso
80346	4068	0.09	previously undescribed retroelement	Diaspora
80345	3227	0.09	Previously unknown repeat sequence	
80344	6201	0.09	previously undescribed retroelement	calypso / diaspora
80343	4527	0.09	previously undescribed retroelement	SIRE
80342	4795	0.09	previously undescribed retroelement	calypso / diaspora
80341	3733	0.08	previously undescribed retroelement	calypso / diaspora
80340	3261	0.08	previously undescribed retroelement	calypso / diaspora
80339	5164	0.08	previously undescribed retroelement	calypso / diaspora
80338	4818	0.08	gb|AF297983.1|AF297983 Glycine max	SB92 repeat

continued on next page

Table 2.2 continued

ContigID	Length	% of genome	Best Genbank hit	Repeat family
			TRS1 tandem repeat region	
80337	3649	0.08	previously undescribed retroelement	calypso / diaspora

In addition to a large number of satellite repeats, we detected novel transposable elements (not detected by BLAST (blastn) comparison to the TIGR repeat database above, presumably because no similar elements are present in that collection). These elements correspond to 25 different classes, including both Type I and Type II transposons. In support of the hypothesis that MULEs do in fact exist in the soybean genome, we detected two MULEs in our de novo soybean repeat assembly. These MULE elements have contig IDs 39304 (estimated approx. 25 copies/genome) and 66822 (estimated approx. 40 copies/genome).

The 40 most abundant sequences detected by assembly of the survey data, the number of reads encoding each, and the percentage of the genome that each is predicted to represent, are summarized in Table 2.2. Note that the list is dominated by SB92 repeats, STR120 satellites and calypso/diaspora type retrotransposons. The full list of assembled repeats is available online [34]. An estimated genomic copy number is given, based on the size of the contig and the number of reads it contains (see Methods section). However, we are unable to determine from our survey whether these sequences actually occur in the stated copy number as contiguous units, or whether fragments of these sequences may occur in separate locations. The copy number is our best estimate of the relative abundance of these high-copy-number sequences.

We have compiled and curated the multiple copy sequences discovered using the above sequencing approach and phrap assembly into a soybean repeat database, available from the authors' web site [34].

Methods for Detection and quantitation of repetitive sequences

Phrap [39], compiled with the manyreads option on a dual Xeon 2.4 Ghz server with 4 GB DDR2 RAM, with the -ace output option, was used for high throughput assembly of the short read sequences. Parameters were tested to optimize assembly of higher-order repeats. In most cases the default parameters for scores,

pentalties, trimming (-trim_qual = 13 -trim_score = 20) were found to be optimal. The assembly of the short reads was found to be very sensitive to the -minmatch parameter. Minmatch values above 14 led to higher-order repeats validated by PCR not being assembled, while values of 12 or less caused the program to crash. Ultimately, 14, the default value, was the value used for the assembly described here. The resultant contigs were either 1) sequences which overlapped by chance, or 2) sequences present in multiple copies per genome. We modeled the probability of generating contigs from sequences which overlap by chance using an implementation of the Poisson distribution developed by Lander and Waterman [4]. The number of contigs expected containing a number of reads j is given by equation 1.

$$Ne^{-2c\sigma}\,(1-e^{c\sigma})^{j-1}$$

$$c = \frac{LN}{G}\ ,\ \sigma = 1 - \frac{T}{L}$$

Where N is the number of reads, L is the read length, G the haploid genome size in base pairs, and T the base pair overlap required for contig formation (in this case equivalent to the phrap minmatch parameter, 14). The 'expected' number of reads (from a perfectly non-repetitive genome) was subtracted from the observed number of reads in order to determine the repeated sequences. No contigs containing more than five reads were expected to occur by chance given the depth of coverage of our survey (Table 2.1).

Using survey data for genomic copy number analysis

Assuming that sequences in our genomic DNA survey are sampled without bias for particular sequence types, the genomic dataset provides a method of estimating the copy number of any genomic sequence. Since the reads are shorter than Sanger sequencing reads, the same amount of sequence provides a higher sampling rate

throughout the genome. A 7% coverage survey with 109.5 bp reads provides 6.25 reads per 10 kb of single copy sequence. By comparison, a Sanger-based survey with 700 bp reads, and with no read pairing, would have a sampling rate of 1 fragment/10 kb at 7% coverage. Since most Sanger sequencing is done using read pairs, this would further reduce the effective sampling rate to one read pair (1,400 bp) per 20 kb of genomic sequence. Hence, the 454 pyrosequencing survey data can be used to estimate the copy number of any 10 kb window of genomic sequence with relative accuracy, as well as detect high-copy-number sequences accurately across much shorter windows.

We utilized the sequence of the CHS region, used earlier to probe the accuracy of the genomic survey sequences, to demonstrate the utility of this approach to detect repeats. The CHS sequence is extensively annotated at the gene level but not previously annotated for noncoding repetitive regions, since no databases of repeats were available to the authors of that study [36]. The survey reads with substantial identity to this region were identified with BLAT, then assembled to the genomic sequence backbone, and further inconsistent matches were excluded using a blastz [45] alignment (using default options for gap penalties, MSP and gap thresholds, chaining and word size). The resulting alignment consists of closely related, but not necessarily directly cognate sequences, since repetitive sequences from other genomic regions are intended to assemble with the repetitive regions in the query sequence, allowing them to be visualized.

Since approximately 7% of the genome was sequenced, approximately 7% coverage is expected for single-copy sequences, and higher coverage indicates repeated regions. Expected copy number can thus be calculated from the coverage of each sequence window across the alignment. Many regions would be expected to be present in two or more copies as a result of the history of the soybean genome, which involves relatively recent duplication [46]. Using the laj viewer [47] and scripts written in-house [34] (source code available on request from the authors), we created graphical views of the alignment. The resulting graphic [Figure 2.2A]

shows the superimposition of the microbead reads matching the BAC sequence containing the 103-Kb region surrounding the CHS locus. This clearly defines regions of the BAC that are present in multiple copies per genome, and shows estimated copy number of these regions.

We repeated this analysis with two more BACs available from the soymap.org site [48]. Neither BAC had any associated annotation at the time of writing. The BAC clone GM_WBb0078A23 is derived from a pericentromeric region, whereas GM_WBb0098N11 is from a euchromatic region of the genome [S. Jackson, Purdue University, personal communication]. The two euchromatic BACs [Figure 2.2A and 2.2B] have a similar appearance low or single copy regions form most of the sequence, and they are interspersed by sequences that are found in tens, hundreds or thousands of copies, such as stretches of satellite repeats or transposable elements. In contrast, the pericentromeric sequence is composed to large extent of sequences that are present in hundreds or thousands of copies [Figure 2.2C]. Note that some regions of the pericentromeric BAC are estimated to be present in few copies, possibly as few as one copy, per genome. This approach is thus potentially useful for detecting unique, possibly genic regions within sequences that are largely repetitive.

Copy number estimation

DNA fragments were matched to the fragment for which copy number is to be determined using BLAT [37]. The number of base pairs matching in BLAT hits with 100% sequence identity was used to provide a minimum copy number, since duplicated genes may have highly similar sequences. Estimated copy number, C, within any sequence window was calculated by equation 2:

$$C = \frac{o}{e}$$

$$e = \frac{cw}{L} \ , \ c = 1 - \frac{LN}{G}$$

Where o represents the observed number of reads matching the sequence window, e represents the expected number of reads matching a single-copy sequence window of size w, c represents coverage, w represents window size in base pairs, N the total number of reads in the survey, L the average read length of the survey in base pairs, and G the haploid genome size in base pairs. In this survey, c = 0.07 and L = 109.5. For the purposes of this study, any region of a clone with an estimated copy number less than one was assigned an estimated copy number of one.

Assembly of sequences to exemplar BAC sequences using BLAT and BLASTZ

For estimation of quality using assembled reads to the 103 kb exemplar sequence [36] BLAT [37] was used with a 95% identity cutoff (otherwise with default nucleotide options) to identify strongly matching reads. All matching reads were then excluded where the matching block did not extend across 98% or more of the entire read, thereby removing reads that did not match at this identity level across their entire length. Estimated probability of any base being correct was then calculated by dividing the number of matching bases by the number of mismatched bases, plus any bases at the end of the read not included in the matching block, plus the number of matching bases. Percentage of correctly matched bases was given by the correct-base probability multiplied by 100.

For copy number estimation, survey reads were identified as being contiguous with the BAC sequence using BLAT with default parameters, a tile-size of 11 and a minimum score of 30 (this results in a "significant" match criterion of a minimum exact match of two eleven-base tiles with an intervening gap of two or fewer bases, and a minimum percentage match of 90% across the entire block

generally in our experience this is roughly equivalent to a blastn search with e-value cutoff of 1E -20). In the copy number estimation [Fig 2.2], an alignment was performed using blastz [45] with default options for gap penalties, MSP and gap thresholds, chaining and word size. Reads not producing an alignment matching these criteria were excluded. The Laj applet [47] was then used for visualization for Figure 2.2 and for the web site alignment tool. A modified version of the laj viewer was used to generate a visualization showing the percentage identity of the 454 read to the the BAC sequence. In this view the difference in sequence-level conservation of the repeat units is displayed. Euchromatic clone BAC I clearly shows three longish repeat elements that show a high degree of conservation in the genome [Fig 2.3] while Euchromatic clone BAC II shows a more dispered signature of repeat elements with a significantly lower degree of conservation [Fig 2.4]. The pericentromeric BAC clone [Fig 2.5] shows very high repeat content and the majority of these repeats are highly conserved in the genome.

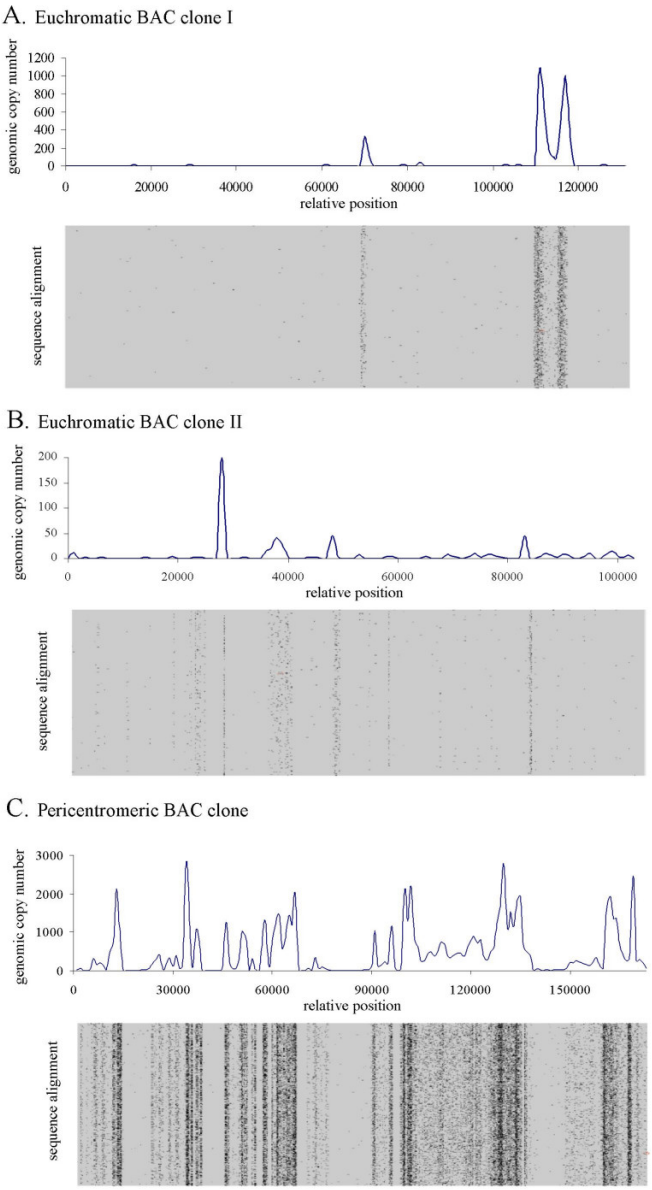

Figure 2.2: **Alignment of sequence survey reads to BAC clones.** The figure shows a graphic of the alignment of survey reads using BLASTZ to three genomic Bacterial Artificial Chromosome (BAC) sequences of soybean DNA, and estimation of copy number. Copy number was estimated according to the number of sequence survey reads aligning to each 1 kb window of the BACs. The alignment represents the superposition of identical or closely related sequences on the BAC sequence, in order to visualize the individual reads showing regions present in many copies per genome. The BAC sequences were: A) The euchromatic BAC described by Clough et al.(20); B) the euchromatic BAC GM_WBb0098N11; C) the BAC GM_WBb0078A23 from a heterochromatic region.

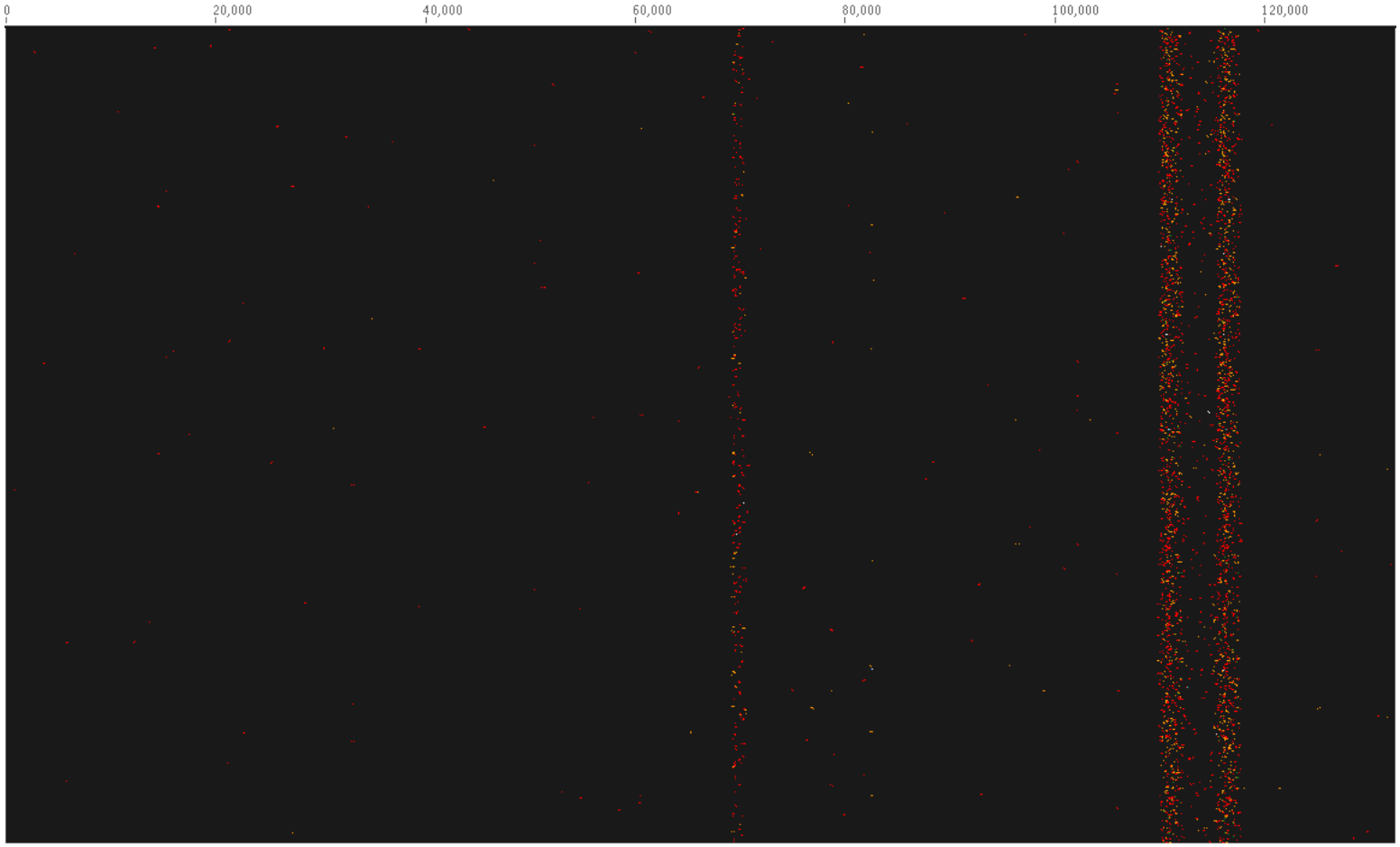

Figure 2.3: **Visualization of the alignments of 454 reads to Gm_Wb0098N11.** Alignments of individual reads to the BAC sequence are shown with the % identity of match indicated by the color. Red indicates very high identity between 454 read and the BAC sequence while White indicates lower % identity . This BAC is mostly low-copy, but carries three highly conserved repeat elements as highlighted by the red vertical clusters of reads.

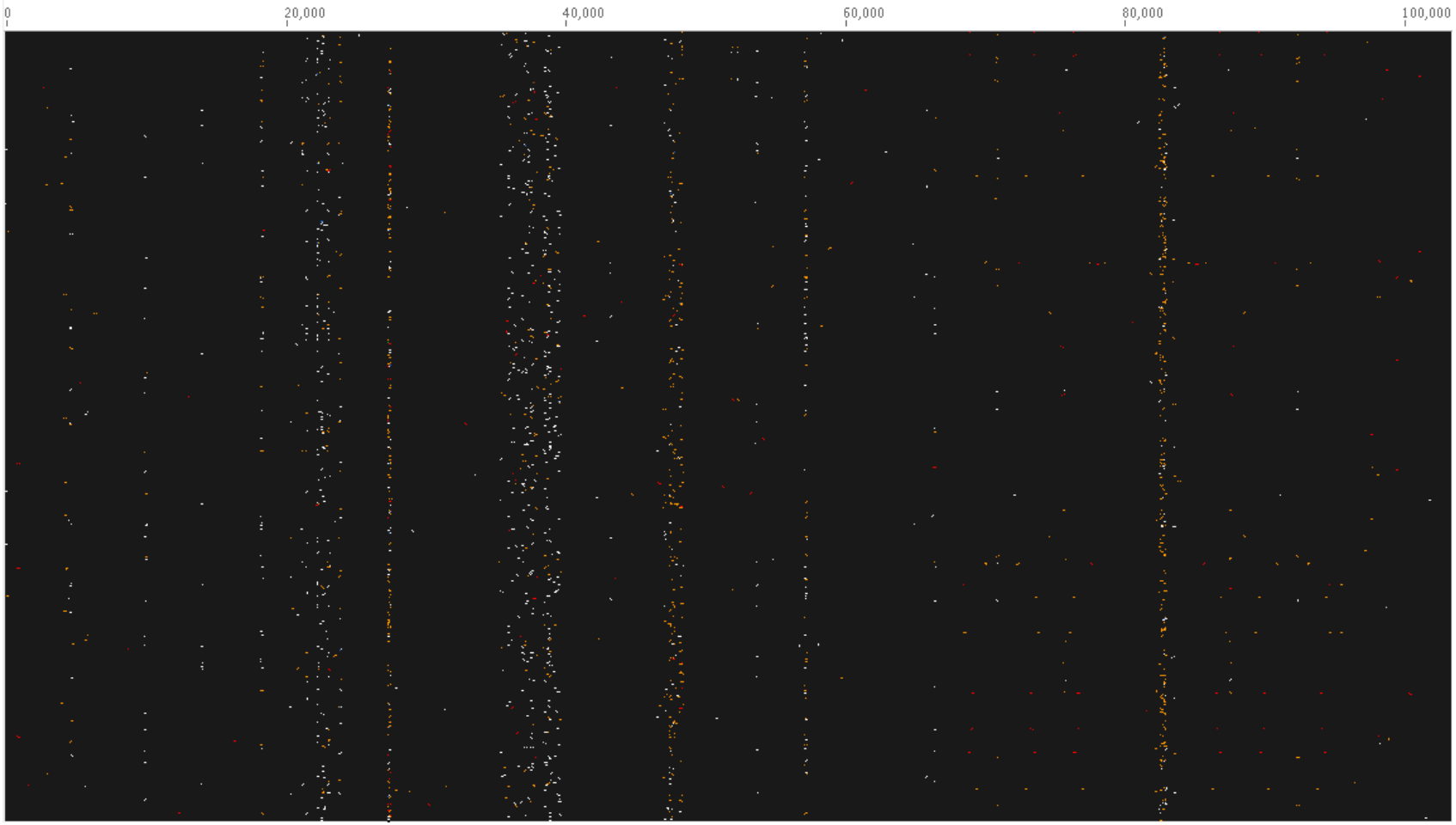

Figure 2.4: **Visualization of the alignments of 454 reads to the CHS containing BAC.** Alignments are shown with the % identity of match indicated by the color. This BAC is largely low-copy with an occasional repeat element of moderate conservation.

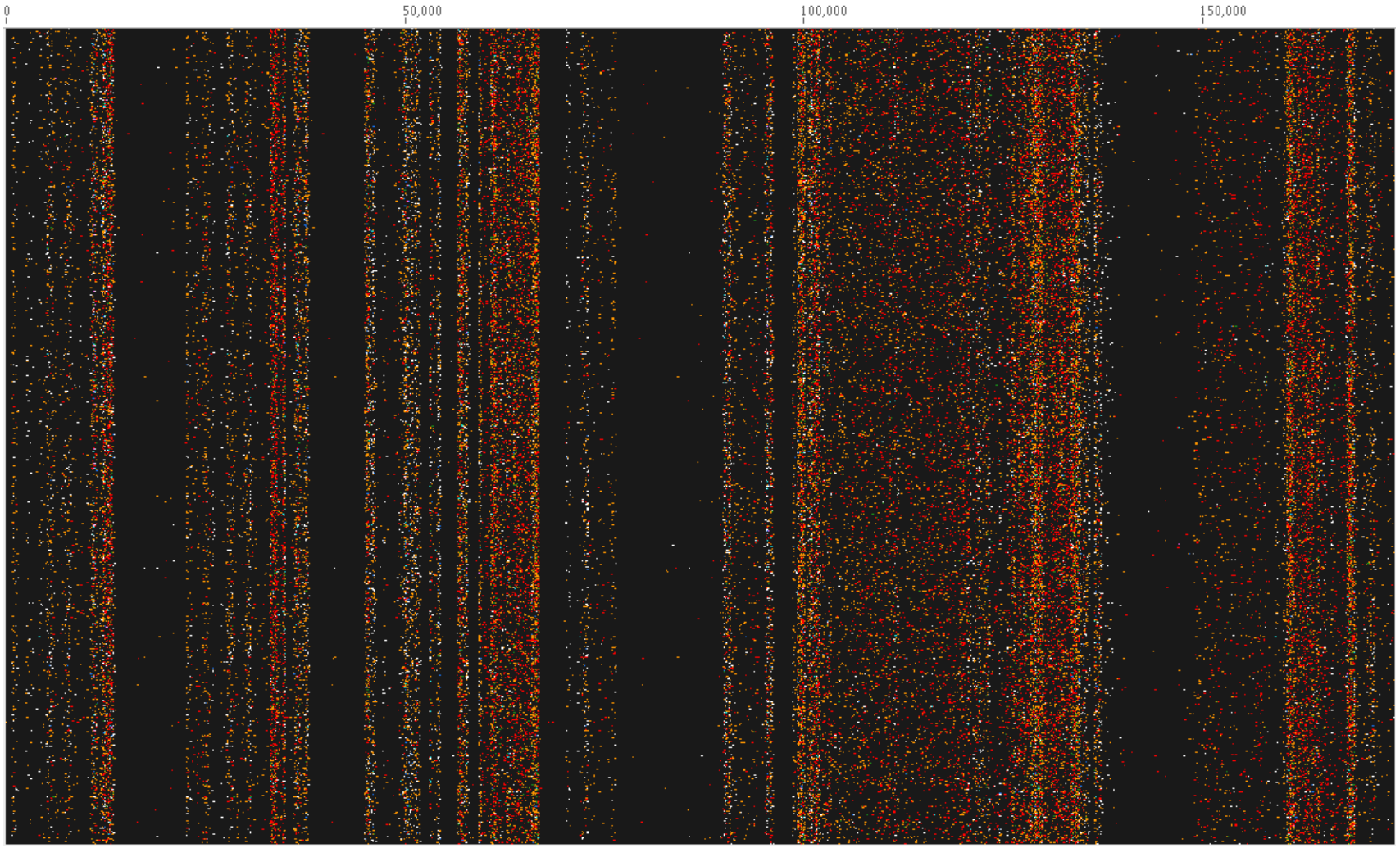

Figure 2.5: **Visualization of the alignments of 454 reads to Gm_Wb0078A23.** Alignments are shown with the % identity of match indicated by the color. Pericentromeric BAC containing a very large number of highly conserved repeat elements.

In addition to developing a database of repetitive sequences, we have developed a graphical tool for alignment of any sequence to the raw read data, to allow the detection of repetitive regions. The whole-genome copy number of sequence fragments from BAC or other genomic clones can be assessed using the search and alignment viewer, which is available at the authors' web site [34].

BLAST searches

Where not otherwise stated, BLAST [49] programs were used with an e value cutoff of 1E-6, and repetitive sequence filtering on except when matching to repeat databases, where the filter was off. The number of significant hits and alignments (-v and -b options) was limited to 20. Otherwise the parameters were used at default settings.

Higher-order structure of repeats within satellite sequence

We were able to assemble some of the smaller, tandem satellite repeats detected in our survey (for example, the previously known STR120 repeat) into non-cognate but deeply sequenced higher-order units using the data from our high-coverage survey. Other sequences, such as retrotransposons, were assembled into a single unit.

In order to validate the assembly of selected assembled abundant sequences, both single unit and higher-order satellite, we used PCR amplification to determine the presence of a block of the predicted size in the genome, and used conventional sequencing to confirm the identity of the fragment. Three such amplicons, two higher-order satellite sequences and one putative retroelement, were amplified from genomic DNA to provide validation of the non-cognate assembly data. The fragments from Contig 80285 (gag-pol) and Contig 80369 (STR120 repeat) were cloned and the fragment ends sequenced from vector primers. The

fragment from Contig 80374 (another STR120 higher order repeat) was refractory to cloning, and was sequenced in part directly from the gel-purified PCR product using the amplification primers. All sequences matched the contig assembled from the 454 sequence survey, with some base mismatches. Fragments 1 and 2 matched their predicted contigs with $> 95\%$ sequence identity across the sequenced length in a global pairwise alignment. No sequence was 100% identical to the predicted contig, probably due to the degeneracy between similar repeats expected in vivo. Fragment 3 was more divergent to our predicted sequence, with a BLAST match at $> 95\%$ identity but an overall identity of 87% to the predicted contig in a global pairwise alignment. We attribute this to a higher level of degeneracy within this higher-order repeat family in vivo, with the cloned fragment being divergent from the most common sequence predicted by the genome survey.

Amplification and sequencing of repetitive DNA sequences

DNA was amplified using the PCR in an MJ research DNA Engine thermal cycler (Bio-Rad, Hercules, CA), and reaction conditions were modified to favor amplification of repeats. Reactions were performed in a total volume of 50 ul containing 30 ng/ul Soybean genomic DNA, 1.2 mM MgCl, 0.1 ug/ul BSA, 0.15 mM dNTPs, 0.025 units/ul of Extaq (Takara Mirus Bio, Madison, WI), 0.6 Ex taq buffer, and 0.05 uM of each oligonucleotide primer. Initial denaturing was at 94 C for 2 min. This was followed by 30 cycles of a 30s denaturing step at 94C, a 40s annealing step at 58C and a 3m extension at 72C. This was followed by a 30 m final extension at 72C.

The primer pairs used were:

MH103 (CATCCATGTTGGTAAGCACCAG) and

MH104 (GGGCATAATAAGGCTTTACACGT),

MH123 (GGTGCAGTTATGGTTTGGGA) and

MH124 (TCTAGAGGTATCATCACTCAAG),

MH155 (TAAAGATGTATTGTCGGAAGATGGGGGC) and

MH156 (TCGAGTTTGGTGCTGTGTTAAATGATTGC).

These primers were designed to amplify segments of Contigs 80285, 80374 and 80369 respectively. The primers were designed completely from sequence derived from assembly of non-cognate small 454 sequence reads. The base quality levels from the 454 sequence assembly had Q values of 40 or greater for all bases underlying the primer.

Plasmid cloning of PCR products was performed using T/A overhang cloning into pGEM-T easy (Promega, Madison, WI). The clones were end-sequenced using BigDye terminator premix (ABI, Foster City, CA) and the vector primers SP6 and T7. Products that failed to clone were end sequenced with the primers used to amplify the product.

Analysis of conceptual translations from genomic reads

The average read size of our survey was 109.5 bp, giving a maximum average open reading frame size of 37 amino acids. Consequently, reads that are derived completely from exonic sequence are a potential source of partial protein sequence. The GMGI database v. 12.0 was used to estimate our survey's coverage of coding regions of the genome [50]. This contains 63,676 sequences with an average length of 594 base pairs. A BLAST (blastn) search was performed with each GMGI sequence as a query and the survey reads as a database, with an expect value cutoff of 1E-6. Figure 1.3A shows the number of soybean ESTs with 95% or higher nucleotide level sequence matches to the raw reads, 23,389 of 63,676, or 37%. Since seven percent of the genome was covered with average 109.5 base pair reads, we expect approximately 37% of the ESTs known from soybean to have hits to the genomic reads. This concordance provides further evidence of the unbiased

random sampling of genomic sequence by our sequencing method, and further evidence that the genome size estimate of 1,115 megabases of Arumuganathan and Earle [35] is approximately correct.

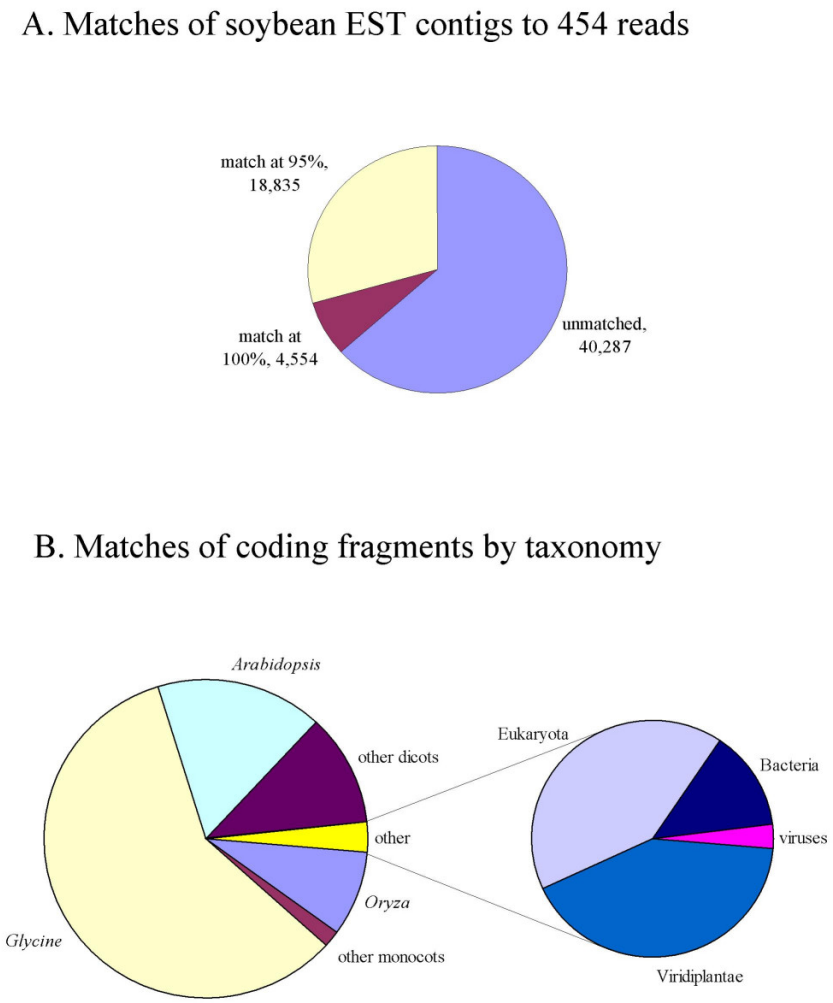

Figure 2.6: **Annotation of protein ORFs with hits to public database.** A) Proportion of EST clones from the Glycine Max Gene Index (GMGI) matched by 454 reads at 95% and 100% sequence identity (using BLAST with e < 1E-6). The total number of sequences matching at 95% or higher identity is 37% of total EST clones. Note that few sequences match at 100% identity due to the error rate of the 454 pyrosequencing used for this study.B) Coding fragments discovered within the short reads (with e values to the GenBank protein (nr) database < 1E-6), and their closest protein-level sequence hit by taxonomy of the source organism of the database sequence.

In addition to sequences that have hits to the GMGI EST collection, a number of reads contained open reading frames with BLAST hits to known proteins from other organisms, but no hits to soybean ESTs or other soybean sequences. Figure 2.3B shows the distribution of coding fragments with an open reading frame giving a 1E-6 or lower BLAST (blastp) e value to the nr database, and the taxonomy of the organism from which the closest sequence in that database was generated. This demonstrates the coverage of the existing EST collection, with over 50% of protein sequence derived from survey reads matching Glycine proteins that are already known. In total, 10,464 of the survey reads were identified as derived from likely conserved protein coding regions (using e < 1E-6 BLAST (blastx) hits to the nr database); 41% of the identified protein fragments have no detectable similarity to known soybean protein sequences, giving over 4,000 potential novel soybean protein fragments with similar, conserved protein sequences known from other organisms.

Protein coding sequence detection and annotation

Sequence reads were translated in all six reading frames and resulting putative peptides were matched to the GenBank nr database. Reading frame translations with BLAST (blastp) hits of 1E-6 or lower were considered to be coding sequence fragments. Percentage identity across the matched region, as given by the BLAST output, was then further used to divide the matches into the groups shown in Figure 1.3.

2.1.4 Discussion

Comparison of 454 survey sequences to previously sequenced BAC clones can reveal regions of multiple-copy sequence and allow approximate quantitation of copy number. Since no bacterial cloning is necessary, a significant advantage of

this approach is that repetitive sequences which are refractory to cloning in E. coli [31] can be characterized without a cloning step.

It is possible to use the survey sequences to reconstruct a representative dataset of soybean highly-repetitive sequences in silico on a whole-genome scale, because sequences which assemble with 7% genome coverage will almost all be present in multiple copies. Using this method 20,670 multi-copy sequences were found, of which 4,213 are estimated to be present in 100 or more copies per genome. These sequences include transposons, satellites, putative centromeric and telomeric repeats (often in higher-order repeat units) and multi-copy genes such as those for ribosomal RNA. We have collated, curated and annotated these repeat sequences and developed an on-line database where these sequences can be accessed and searched, and we believe they have utility and biological interest in addition to the detection of repeats for genome assembly. For example, since MULEs can be domesticated to perform conserved developmental tasks [51] it is possible the MULEs detected using this survey in soybean will be of broader biological interest.

Exclusion of these multi-copy sequences and low-complexity simple repetitive DNA gives a dataset of "low or single-copy" DNA sequences that can be potentially used to derive genetic markers in subsequent experiments. Agreement with previous Cot measurements [40, 41] provides evidence of a lack of bias in genomic sampling using the 454 sequencing procedure, thus it is possible that high-coverage surveys will be able to detect single-copy regions with greater accuracy than current methods.

Of the 20,670 repeats discovered in our survey, an interesting class are the higher-order repeats composed of slightly divergent repeat units of between 30 and 220 nucleotides. This class represents many of the most abundant repeats in soybean [Table 2]. Eukaryotic centromeres are typically composed of satellite sequences with a repeat frequency of between 150 and 210 nucleotides, or approximately the amount of DNA required to fold around a nucleosome [52]. Two 92 bp

repeats (based on our analysis, the most abundant sequence family in the soybean genome) form a repeat unit of 184 bp, making these sequences a candidate for a centromeric or pericentromeric satellite. Such satellite sequences, while conserved in size, are highly variable in sequence even within a plant species [53] and show more rapid evolutionary change than euchromatic sequences [52, 54] consequently it is expected that soybean repeats show little sequence similarity to those known in Arabidopsis and its relatives [Table 2.2].

In humans [55] and in Arabidopsis [56] centromeric repeats have been shown to consist of higher order arrays, composed of closely related yet divergent nucleosomal repeat monomers. Our short-read sequencing data allows global analysis of such higher-order structures within abundant satellite DNA, and several sequences in Table 2.2 and the repeat database [34] represent such higher order repeat families, producing contigs between 2,500 and 14,000 base pairs in length. Speculatively, therefore, some of these sequences may represent novel centromeric repeats. These relatively large, high-copy-number satellite repeats are difficult to access by other means, and are often not included even in "completed" genomes such as Arabidopsis [53] because of difficulties in obtaining or assembling BAC clones. A detailed catalog of these higher-order repeats is an important product of the survey approach we describe. Knowledge of these higher order sequences provides both a screen for clones containing such problem sequences, and potentially a method to generate more detailed knowledge of tandem repetitive regions such as centromeres or telomeres.

In a genome such as soybean, where substantial EST sequencing has been performed, but the genome itself is not completely sequenced, the genomic survey data can also provide estimates of the copy number of any genes characterized at the molecular level. Copy number of genes is known to affect agronomically relevant traits in soybean, such as allergenicity [57]. In addition to gene copy number estimation, detailed knowledge of repeat sequences, and the ability to screen these sequences from any shotgun genome sequencing dataset, are of significant value

to any sequencing and assembly project. While our survey was aimed primarily at investigating repetitive sequences, we also generated some data on partial protein-coding sequences. Most of the sequences we discover with hits to known proteins, but not to known soybean proteins, are likely to represent the regions of incomplete coverage within transcripts partially covered by known ESTs. It is also possible that some of our short sequences are not of sufficient length to generate significant hits. However, some hits from non-plant eukaryotes, bacteria and viruses are seen. These sequences may indicate the presence of a small number of coding sequences in soybean without homology in the completely sequenced plant genomes. We cannot exclude the possibility that our sequences are too short to generate significant scoring alignments with some orthologous plant proteins. It is also possible that these sequences result from microbial DNA contamination, or that homologous proteins exist within, for example Arabidopsis or rice but that these proteins have not been annotated and placed in the nr database. The utility of such a coding region fragment discovery project includes the potential to design microarray probes to coding sequences that may not be present even in detailed EST sequence sets.

2.1.5 Conclusion

We have developed and validated a method for genomic survey sequencing; a high-coverage, short-read genome survey using 454 pyrosequencing. This method provides no de novo assembled sequence, and is not a replacement for conventional shotgun genomic sequencing, or for EST sequencing. However, rapid sequencing of many short genomic fragments gives a clear picture of overall genome composition. Given the much lower cost of the method when compared to Sanger-based whole-genome sequencing or EST sequencing, it can provide a substantial amount of information as a preliminary step to characterize large, unsequenced genomes.

Even much higher coverage sequencing of soybean, using random short reads of the size described here, would be unlikely to allow the assembly of a complete genome sequence. Short sequence fragments, together with the extensive repeats we describe, would cause insoluble difficulties in whole-genome assembly. However, a 454 pyrosequencing genome survey allows the derivation of many types of valuable information, including repeat composition, genome size and genomic copy number. Higher coverage would further increase the value of this type of survey, in particular the coverage of single-copy protein-coding sequences. Ultimately, advances in read length (up to 500 bp or more), and the availability of paired reads, could make possible true whole-genome shotgun sequencing of soybean and other crop plants at greatly reduced cost.

2.2 Identification of repeating units and use in karyotyping *Glycine max* chromosomes

2.2.1 Introduction

Despite the generation of large volumes of sequence information in genome surveys and the subsequent whole genome sequencing effort the unambigous assignment of the sequences to individual soybean chromosomes is still challenging. Numerous efforts by the soybean research community, applying various marker generation methods over the past few decades, helped develop fairly detailed genetic linkage maps for soybean. More recently, twenty molecular linkage groups had been characterized, using SSR markers, to correspond to the twenty soybean chromosomes.

Despite the availability of multiple genetic maps and the development and application of numerous technologies to generate a genetic linkage map, substantial difficulty existed in the production of a cytogenetic map for soybean. The confounding issue was primarily the size and shape of the soybean chromosomes themselves. *Glycine max*, and its close relative *Glycine soja* have a 2n=40 set of chromosomes in vegetative cells. The 20 chromosomes in each set are remarkably identical in size and show almost no distinguishing cytological features. The secondary problem is that the relatively recent whole genome duplication in soybean means that most hybridization based methods are unable to distinguish between the highly similar chromosomal stretches. On the other hand centromeric repeats are known to diverge rapidly [52] and are stably inherited loci within a species, hence are extremely useful as karyotyping probes. The challenge, then, is to identify centromeric repeats and characterize the variation within them, with the purpose of identifying regions of conservation and regions of divergence.

The ability to differentiate different chromosomes of the plant is important, particularly in crosses with other species. Such crosses can be highly beneficial

to crop species to introgress agronomically important traits. Soybean especially has been used numerous times to generate such wide crosses [58]. Additionally a clear cytogenetic map allows tracking of chromosomal translocations and other chromosome scale events within different lines. Chromosomal evolution and centromeric studies would also benefit greatly from an unambigous karyotype that can be linked to a genetic linkage map. Nonetheless it is virtually impossible to karyotype the individual chromosomes of the soybean plant through microscopy owing to their very similar shapes and sizes. Soybean chromosomes were initially karyotyped using the typical methods of chromatin content and chromosomal arm lengths [59]. While useful, the routine usage of this method was both challenging and error-prone due to the extensive experience required to discern these minor differences.

Previous attempts at characterizing the size and content of soybean centromeres had identifed the 92bp short tandem repeat (named as SB92) as strongly linked to the centromeres of the soybean chromosomes [60]. Two more short tandem repeats with repeat unit lengths of 120bp(STR120) [61] and 102bp(STR102) [62]were identified to be enriched in the pericentromeric regions along with numerous longer transposable elements. The gypsy and copia families of transposable elements are highly abundant in the soybean genome [9]. All of these sequences were, in turn, proposed as candidates for centromeric repeats and utilised as in situ probes to determine copy number and localization in the soybean genome. The survey sequencing effort described previously proved instrumental in categorizing all the sub-species of these repeats, especially the SB92 repeat family. Centromeric sequences have been reported to diverge between each chromosome of an organism while still maintaining an overall similarity [52]. Among all the soybean repeats discovered the SB92 family fits this behavior of the centromeric repeats best. Therefore we hypothesized that the variants of the SB92 family might be differentially associated with the centromeres of different soybean chromosomes.

Vahedian et al [44], and independently, Kolchinsy and Gresshoff [63] first described a short tandem repeat by hybridizing labelled total soybean DNA to a soybean genomic DNA library and by studying the low molecular weight bands generated by a restriction enzyme with infrequent sites. Vahedian et al, in their study, showed that the soybean genomic DNA hybridizes heavily to a particular clone and that restriction digests of genomic DNA show a periodicity of about 90bp. It was also shown that digestion by a methylation sensitive enzyme failed to reproduce this periodicity, thereby implying that this sequence is heavily methylated in the soybean genome. The generation of a ladder with units in increment of 90 bp also suggests that the sequence is present in large tandemly repeated blocks but the repeating units themselves show sufficient sequence diversity to alter the presence of the restriction site within each unit. Similar results were obtained by Kolchinsky and Gresshoff using a different restriction enzyme and they were able to isolate and directly clone the smallest repeat unit and the assumed trimer of the repeat.

Also, similar ladders with periodicity of 92 bases were obtained from multiple enzymes which identify different recognition sequences, implying that the sequence variation along the length of this short repeat is fairly large. Kolchinsy and Gresshoff identified 10 differing repeats from genomic DNA and estimated the sequence conservation across them to be about 92%. A fairly crude characterization of the copy number for this sequence in the soybean genome estimated that the sequence is present at approximately 100,000 copies in the genome. Thus the tandem repeat identified seemed to possess all the salient features of a centromeric repeat, namely a high copy number, a short repeating unit arranged in large tandem batteries and a high amount of sequence diversity.

On sequencing the clone Vahedian et al. identified a 190bp sequence fragment that consisted two short repeating units. The first unit was 92 bp long and labelled

the SB92 repeat while the second was 91bp long and labeled the SB91 repeat. Further sequencing using specific primers and amplification from genomic DNA identified a small amount of variation in the identified sequences. Kolchinsy and Gresshoff on the other hand reported that while the repeats show a large amount of sequence diversity the length of the repeating unit is fixed at 92bp. Nonetheless, both papers concluded that the SB92 repeat and its variants are centromeric in nature and that the inherent variation within the sequence should be helpful to karyotype the soybean chromosomes. Subsequent studies [62] attempted to follow the localization, of the two variants orginally reported, with the specific goal of differentiating the two chromosome sets from the most recent genome duplication.

More recently the SB92 family has been classified into two families of repeats, named CentGm-1 and CentGm-2, that show significant divergence from each other. The CentGm-1 family of repeats show a higher degree of similarity to the canonical SB92 sequence generated by Vahedian et al and represented by Contig80377 in non-cognate assembly of survey reads [9]. On the other hand, the CentGm-2 family of repeats is closer to the repeat earlier called SB91 and is best represented by Contig80371 in the non-cognate assembly.

Allopolyploidy in soybean

Soybean and other diploid Glycine species have 20 pairs of chromosomes and are believed to have undergone a whole genome duplication compared to the other crop species from the legume family. Numerous studies based on restriction fragment length polymorphisms (RFLP) [64], sequence analysis of dupilcate genes, and BAC based FISH provide evidence for the presence of two highly similar homeologous regions in the soybean genome. Further evidence was provided by integration of the high density genetic and physical maps of soybean [65]. While the evidence for polyploidy, both recent and ancient, was and is fairly substantial

in the published literature there was little consensus on whether the more recent polyploidization event occured by whole genome duplication of a single ancestor with 2n=20 or through the cross of 2 distinct 2n=20 progenitors. Sequence analysis of duplicated regions did not provide evidence of sequence similarity to two distinct extant species that could have served as, or been derived from, the anticipated progenitors. This result can be interpreted as either evidence for autopolyploidy invovlving a single progenitor or the loss of the both progenitor lines after the hybridization event. One line of evidence for the allopolyploidy hypothesis was the early description of two variants within the SB92 family [44] allowing the possibility of two distinct centromeric sequences, each of which could have been derived from a distinct progenitor. The great degree of sequence similarity between these variants could then be explained as either convergent evolution of centromeric repeats to allow viable meiotic pairing and stable spindle association within the nucleus. Alternately the two progenitor cross to produce the hybrid could have occurred very shortly after the speciation event(s) leading to their separation and the limited timeframe between the speciation and recombination resulted in relatively less divergent centromeric repeats.

2.2.2 Methods for tandem repeat identification

In the survey sequencing study described previously [9] we identified the most abundant repeat units in the soybean genome and computed their approximate copy number, using a 454 sequence survey of randomly sheared genomic DNA from the Williams 82 line of soybean. The non-cognate assembly produced 80377 "contigs" that separate individual elements of a repeat family and join them into their, presumably, higher order tandem occurrences in the genome. Thus, Contig80377 (a sequence that contains multiple slight variations of the CentGm1 repeat in tandem; i.e., a higher order repeat of CentGm1) is the single most abundant sequence in the soybean genome. Contig80377 and numerous other

repeat contigs captured the local variants in the SB92 repeat family. Since the assembly was created from randomly generated survey DNA sequences, the depth of the contig is expected to correlate with the copy number of that repeat element in the genome. Because of this, the higher the number assigned to the contig (the highest being 80377), the higher the copy number of that repeat in the soybean genome.

The low coverage survey and noncognate assembly thus successfully separated the slight variants of the small tandem repeats (such as SB92), as well as generated assemblies of larger repeats such as transposons. This very same property of the tandem repeats confounds whole genome assemblies as the one attempted for the soybean genome [18] and tandemly repeated regions were therefore not included in the genome assembly. While the variant separation in the larger assembled noncognate contigs produced by the survey assembly allows identification of sequence diversity within a family, the individual contigs still need to be deconvoluted to identify the monomer repeating unit constituting each contig. For this purpose, mreps software [66] was used to identify contigs that consisted of tandemly repeated, slightly varied monomers.

The noncognate contigs were sorted by read depth and the 500 contigs with greatest depth were analyzed with mreps. The program was asked to identify all tandemly repeated elements with a 10 bp or longer monomer with a small level of "fuzziness" allowed by sepcifying a value of 5 to the resolution parameter. With these settings the program is expected to identify short tandem repeats that are almost identical while filtering out very small repeats and ignoring simple repeats such as homopolymeric stretches. Since the goal for this analysis was to identify sequences in large batteries showing enough identity to hybridize to the same probe, the stringency level chosen was high. The resulting output file was parsed using a Perl script.

2.2.3 Results

We identified 3 large tandem repeat families, all of which had an estimated copy number of over 1,000 copies in the genome: the SB92 family subdivided into the CentGm-1 and CentGm-2, the STR120 family, and a novel repeat with a monomer of 86 bp that had not been described previously. Following covention we labeled the novel repeat SB86. The SB92 and STR120 repeat families were represented by multiple contigs and thus numerous monomers were identified from each of the contigs. The identified monomers from each contig were aligned to each other to aid in the identification of regions of high and low sequence conservation.

SB92 family

The SB92 family was represented by numerous contigs in the assembly. Often each of the contigs yielded slightly different monomers in different positions along the contig. Given the noncognate nature of the assembly these variants likely represent repeat elements derived from different regions of the centromeric and/or pericentromeric regions. This is especially likely in the contigs that show a greater diversity between the monomers since previous studies in centromeric repeat evolution [52] indicate that centromeric repeats display a strong tendency to undergo convergent evolution locally.

A well conserved 92 bp monomer was identified from 15 different contigs, of which 3 showed multiple monomers within the same contig. Two contigs, Contig80377 and Contig80371, contained many kilobases of contiguous sequence composed locally of highly similar monomers. The monomers from each contig showed enough sequence similarity to be identified by a BLAST alignment against the canonical SB92 sequence and yet had a few regions of sequence diversity. Contig80377, contig with the greatest depth, in particular had multiple monomers each differing slightly from each other and the canonical SB92 sequence.

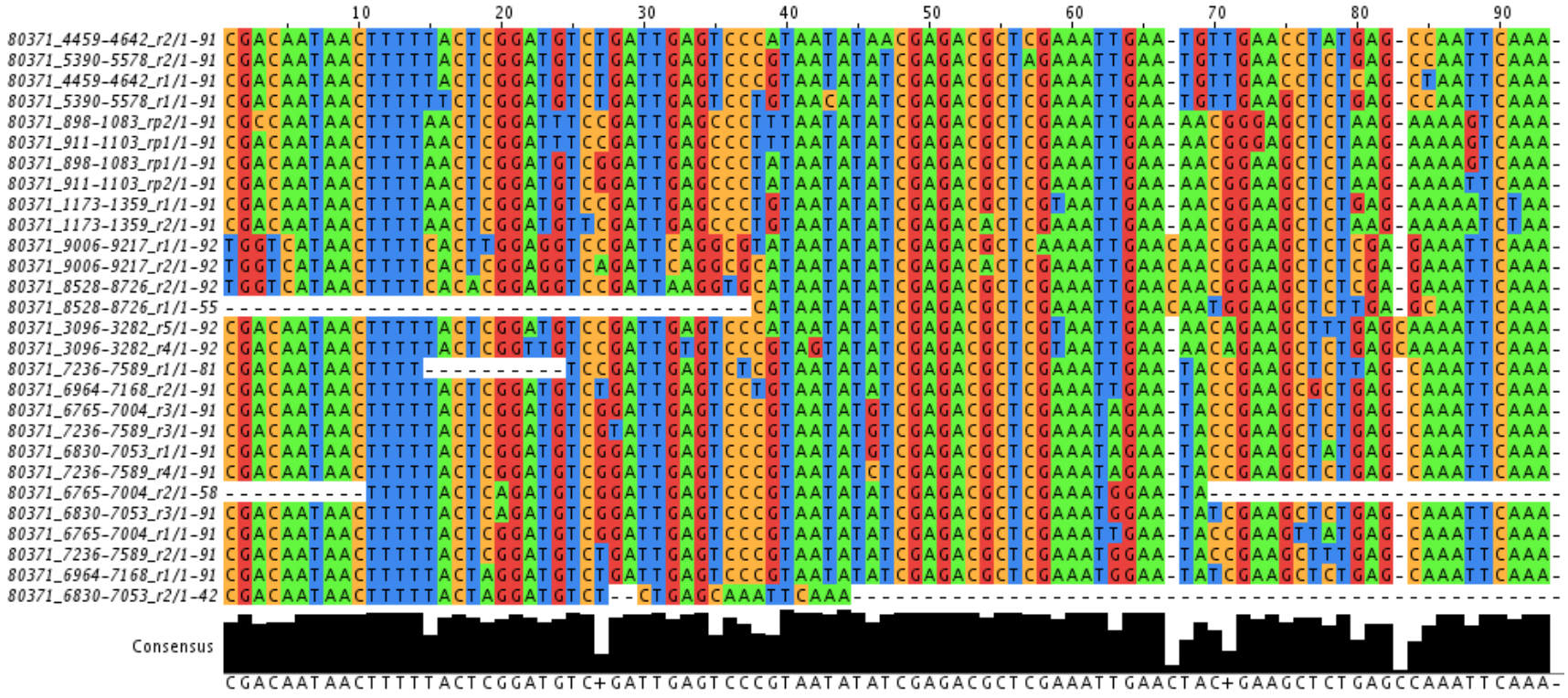

Figure 2.7: **Alignment of monomers from Contig80371.** Multiple sequence alignment of the short tandem repeats identified from Contig80371. These units show greater % identity to the CentGm-2 sequence.

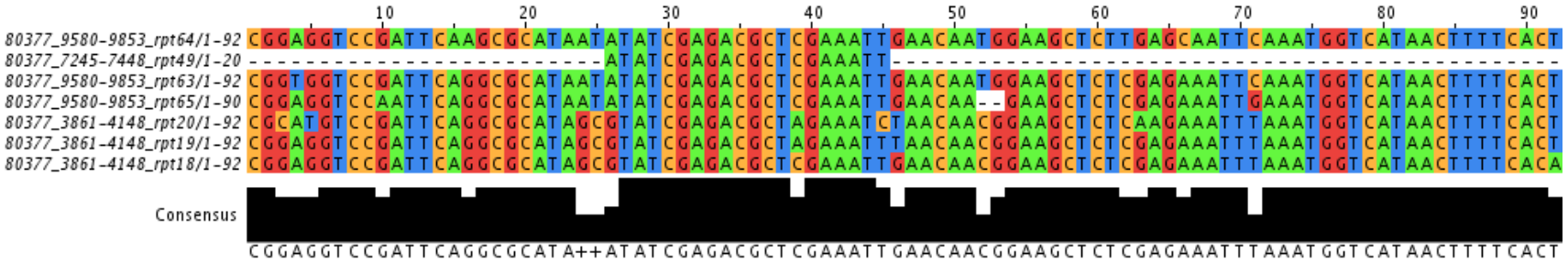

Figure 2.8: **Alignment of few monomers from Contig80377.** Multiple sequence alignment of the short tandem repeats identified from one region of Contig80377. These units show >95% identity to the canonical SB92 sequence and CentGm-1.

SB86 repeat

Contig80186 is a 1703 bp long higher order repeat structure and was assembled from 243 reads. mreps reported that 1300bp of this contig is composed of a repeating monomer of 86 bp. This 86 bp repeat bore no resemblance to any of the previously identified tandem repeats in soybean. Using the equation for copy number estimation [9], it can be estimated that the 86 bp repeat is present in 19000 copies in the genome. Additionally, the absence of much interspersing sequence between the 86bp repeat units in the structure of this contig suggests that this repeat is most likely present in a single or few loci in the genome as opposed to being interspersed with other repeats like STR120, for example. Therefore this sequence was identified as an ideal target for designing a FISH probe.

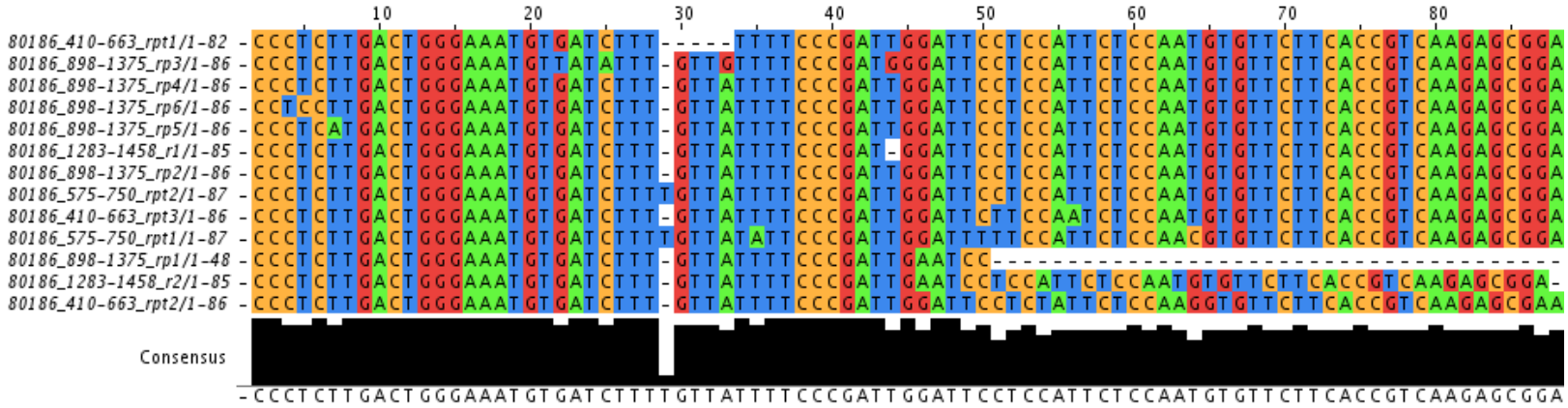

Figure 2.9: **Alignment of 86 bp monomers from Contig80186.** Multiple sequence alignment of the short tandem repeats identified from Contig80186.

Probe Design

Probes were designed for each of the repeating monomers using a simple set of rules, namely, each probe must be at least 50% GC, 25 bp long, and specifically identify one repeat subfamily. Each probe was made as an oligonucleotide labeled with a fluorochrome for screening as karyokaryotyping agents. Due to the multitude of monomers that are part of the SB92 family and hence the lack of a true consensus sequence for the SB92 family of repeats, probes were designed manually from sequence alignments. Some of the probes for the SB92 family were designed to hybridize to the most conserved regions of the repeat while others were designed to the variant regions. Each probe was then used as a karyotyping agent and scored for the ability to produce the widest possible range of signals among the chromosomes. The expectation was that the differences in hybridization efficiency among the variants in the family and the copy number differences of variants would together produce a specific combination of hybridization signals that allowed the distinction of chromosomes. Probes that showed useful, discriminative patterns for karyotyping were retained.

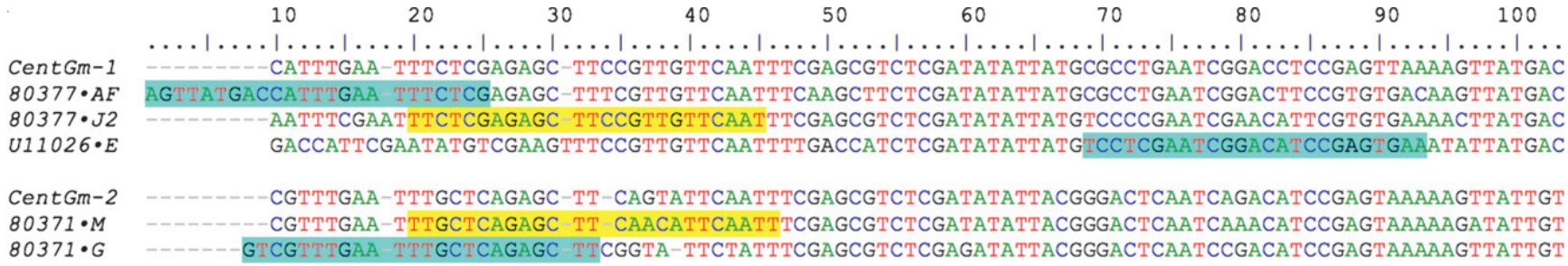

Figure 2.10: **Probes designed to consensus sequences from Contig80371 and Contig80377.** Multiple sequence alignment of the short tandem repeats identified from Contig80377 and Contig80371 and the CentGm-1 and CentGm-2 repeats.

Final cocktail of FISH probes for Soybean karyotyping

None of the probes were individually sufficient to discern all the chromosomes. A methodology to karyotype chromosomes using a mixture of short DNA probes was applied successfully to *Zea mays* [67]. A similar approach was attempted in soybean with many combinations of probes. The STR120 and STR102 probes were rejected due to a high background signal. The cocktail of oligonucleotide probes that provided the best resolution was as follows:

SB92 family probes:

1. TTGCTCAGAGCTTCAACATTCAATT
2. AAGCTCTGAGCAAATTCAAACGAC
3. CGAGAAATTCAAATGGTCATAACT
4. TTCACTCGGATGTCCGATTCGAGGA
5. TTCTCGAGAGCTTCCGTTGTTCAAT

SB86 probe:

ATGTGATCTTTGTTATTTTCCCG

This cocktail of probes was sufficient to label all the chromosome pairs in soybean but lacked the ability to discriminate all the pairs. Therefore, an additional probe to the 18s rDNA and 10 BAC based probes were employed in the final cocktail to completely and unambigously label each of the 20 chromosome pairs.

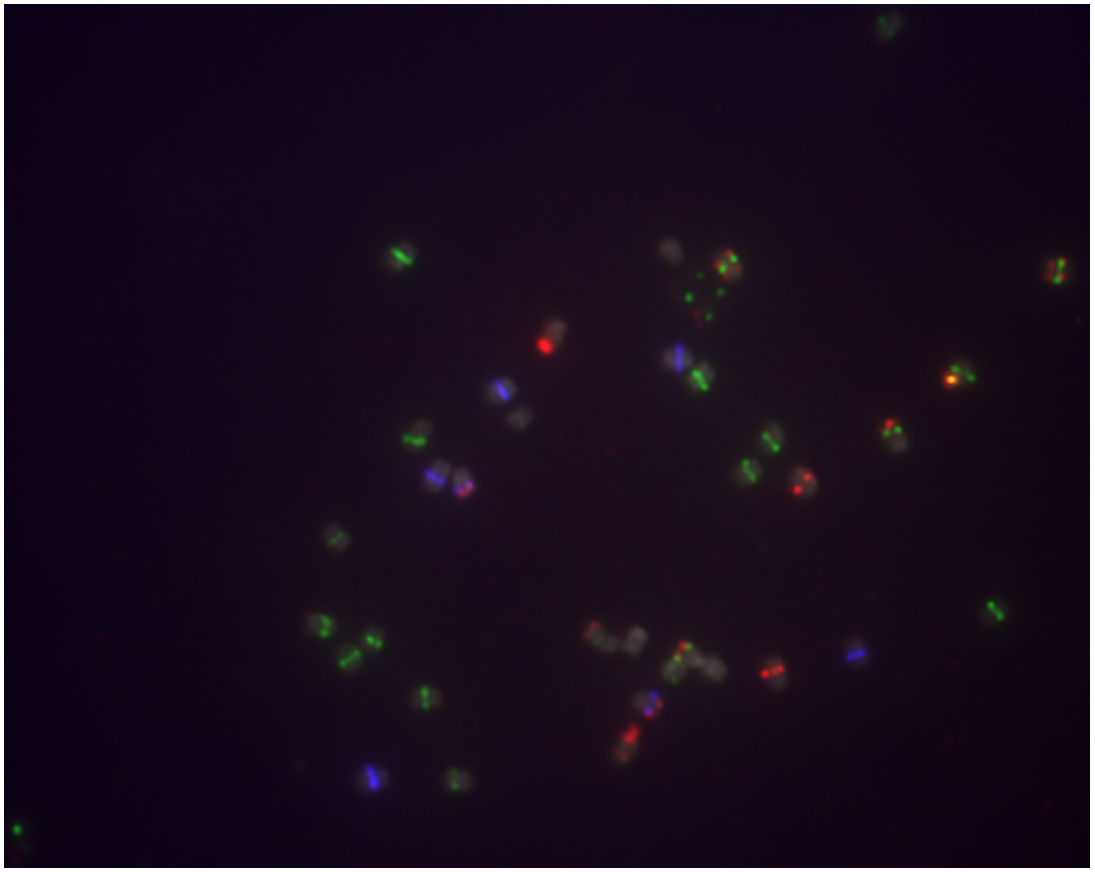

Figure 2.11: **Metaphase spread of soybean chromosomes labeled with a cocktail of probes designed from various units of SB92.** Four SB92 probes designed from the SB92 repeat units, and a SB86 probe were tagged with a range of flourophores and hybridized to metaphase spreads of soybean chromosomes. Shown here are false colored signals from a chromosome spread indicating differential hybridization affinity of the probes to different chromosomes. The combination of probe signals and their affinity allows differentiation of a majority of the soybean chromosomes.

2.2.4 Discussion

Centromeric repeats are unique among tandem repeats in a genome since they seemingly diverge quickly after speciation events and to a lesser degree between the individual chromosomes in the genome. In addition these repeats undergo rapid and uneven expansion through tandem duplications. This process of rapid evolution generates a unique signal of sequence and copy number variation between the chromosomes that in recent years has been exploited to visually differentiate between the chromosomes, often in a mitotic spread. Somewhat countering the rapid evolution is the demonstrated covergent evolution of tandem repeats in close proximity.

The soybean genome, like many plant genomes, is rich in repeats. The repeat content, specifically the high-copy-number repeat content, of the soybean genome was estimated to be as high as 40% by numerous studies. Unlike monocot crop species (e.g. maize and rice), soybean displayed distinct organization of heterochromatic and euchromatic regions [62]. The centromeric and pericentromeric regions were shown to be very high in repeat content and relatively gene poor. Previous studies had identified the SB92 and STR120-STR102 repeats as being strongly associated with the centromeric and pericentromeric regions. Our survey sequencing had identified both these repeat families as being present in many thousands of copies in the genome, with the SB92 repeat being especially dominant. In addition, the SB86 repeat was estimated to be a abundant repeat albeit, with far fewer occurrences than the previous two mentioned. The STR120 family, despite its abundance, proved to be a poor candidate for generating a discernible FISH signal implying that this repeat likely occurs in a more dispersed form in the genome and is most likely not centromeric in nature.

The SB86 repeat is very interesting since it shows a clear hybridization signal on a single chromosome pair. This signal does not correspond to the primary constriction of that chromsome pair, which itself shows a fairly strong SB92 sig-

nal. Therefore, the SB86 repeat is present in a large battery of tandem repeats in a single locus in the soybean genome not associated with the apparent centromere. This sequence could therefore be the location of the incipient evolution of a neocentromere in soybean. Alternately this locus and the short repeat might be the last remnant of a different centromeric repeat that is in the process of being eliminated and replaced by the SB92 family.

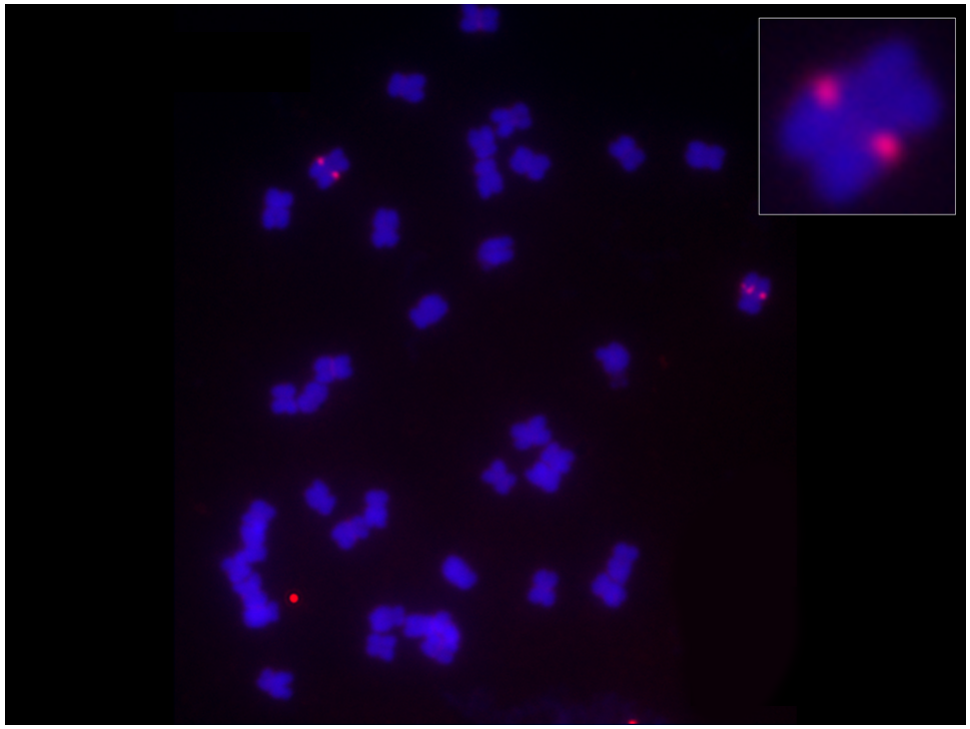

Figure 2.12: **Metaphase spread of soybean chromosomes labeled with a probe designed from SB86.** A SB86 probe was labeled with a flourophore and hybridized to metaphase spreads of soybean chromosomes. Shown here is the signal from a chromosome spread. The SB86 probe hybridizes to a single chromosome pair.

A recent study of soybean centromeric sequences [68] using ChIP pulldowns with a specific antibody characterized the sequences associated with Centromere-specific Histone3 (CENH3). In their study Tek et al. identified the CentGm-1 and a family of previsouly described non-autonomous retrotransposons [69] as being bound to the GmCENH3 protein. The family of nonautonomous retroelements named Family 6 was discovered by Wawrzynski et al. by comparing repeats from multiple legume species and they identified Contig80367 from our noncognate assembly as its best match in the survey sequencing. Tek et al. also identified a new 411 bp repeat element they christened CentGm-4. While this novel repeat sequence fits most classic parameters of centromeric repeats, the size of its repeating unit is much larger than anticipated for a centromeric repeat. Subsequent sequence comparisons showed that the 411 bp repeat was assembled into seven separate contigs namely Contig79817,Contig80235,Contig80240,Contig80260, Contig80262 and Contig270, in the asseembly. The large periodicity (number of base-pairs before the repeat unit starts over again) of the repeat in its tandem batteries and slightly higher variation between the adjoining repeat units seen in this repeat precluded this unit from identification using the repeat unit identification method used in our study. Therefore, even though the unit was present in multiple contigs and the resulting copy number was in between the SB92 and SB86 repeats, our method fails to identify the larger repeat unit. Adjusting the parameters to the mreps program to allow more divergence between the monomers of a tandem repeat would most likely rectify this problem, but beyond a certain level of stringency relaxation would increase the false positive rate. There would also be an accompanying reduction in the ability to discriminate between divergent members of a repeat family that was critical in our development of probes for the SB92 family. Therefore any future applications of this method in other species would likely benefit from repeating the tandem repeat identification by decreasing the stringency level recursively and collecting the newly identified units from each stage.

Soybean researchers had over many years developed a very fine linkage map for the 20 chromosomes using a wide variety of molecular markers. With the development of a karyoptyping method and the co-localization of BACs containing known markers it became possible to assign individual linkage groups to physical molecules for the first time. In addition to providing a molecular framework for understanding inheritance of traits, this development allows the tracking of large chromosomal translocations and rearrangements among the soybean chromosomes.

CHAPTER 3

SMALL RNA CONTENT AND EXPRESSION

3.1 Introduction

Small RNAs are a relatively modern discovery in the field of genetics that has vastly increased our understanding of numerous biological phenomena, which could not be easily explained by single or multi gene inheritance. Small RNAs are now understood to play an important role in regulating and modulating gene expression by down regulating a single gene or a group of genes in a wide range of tissues and organs. Most small RNAs act locally within the cell where they are produced. Additionally, the small size and abundant copies of the small RNA molecules are believed to allow some of them to dissipate quickly through tissues and act as signalling molecules [70].

In the early 1990's, transformation methods began to be extensively applied to many plant species to produce transgenics with desirable traits. While the methods have been extremely useful in agriculture, horticulture and related fields in generating numerous useful transgenics, there have been a few setbacks in the process. Some of the early researchers noticed that in some transgenic individuals the expression of the transgene would be shutdown shortly or within a few generations of the transformation event. In two studies in particular, researchers attempted to darken the flower color in Petunia by increasing the copy number of the Chalcone synthase (CHS) gene that is critical to anthocyanin synthesis [71, 72]. Both sets of researchers found that the mRNA level of CHS was lower than the wild type in a portion of the transformed plants. This observation of transcriptional repression induced by the incorporation of extra copies of an endogenous gene remained unexplained at the time. At the same time, other groups working with virus resistance in plants discovered that expression of viral transgenes in a susceptible plant rendered high levels of resistance against that and other similar viruses. This resistance was later determined to be caused by a molecular mechanism involving a sequence-specific degradation of viral RNA [73].

The presence of small RNA molecules responsible for the repression of an endogenous gene, was first reported in *C.elegans* in a breakthrough study by Lee et al. [74]. Specifically the role of double stranded RNA in this suppression mechanism was discovered by Fire et al. [75]. Simultaneous discoveries in virus resistance studies in plants [76] confirmed the role of small RNA in the two major clades of multicellular organisms. Further research has since shown the existence of small RNA mediated post-transcriptional regulation in the algal and fungal kingdoms, but in a less extensive role.

Numerous genomic regions outside the traditional protein-coding gene space have been shown to be transcribed. A generic term used to describe such transcribed regions that do not seem to code for any protein is noncoding RNAs. Small RNAs, as a class of the broader non-coding RNAs, are primarily identified by the size of the functional molecules. RNA molecules in the size range 21 to 25 nucleotides are termed "small RNA" and within this range different size classes show distinct yet overlapping functions and origins. The two major classes of small RNA in plants are small interfering RNA (siRNA) and microRNA (miRNA). They play diverse roles in the maintenance, management and expression of the plant genome.

3.1.1 siRNA function and biogenesis

Early experiments trying to incorporate resistance to viruses discovered that expressing transgenic viral RNA in plants made them resistant to the virus and to other viruses that shared significant sequence similarity [77]. Further studies proved that the plants achieved resistance by silencing the expression of viral RNA through a species of small RNA molecules [76] later labeled small interfering RNA (siRNA). This class of molecules has since been shown to function in a wide range of plants and animals. Its main characteristics are a size range of 21-24 nucleotides and they are produced from a double stranded RNA precursor.

siRNA biogenesis is a multi-step pathway involving a few key components that are well conserved between the plant and animal kingdoms. They are generated from long double stranded RNA molecules that are formed by the co-existence of a sense strand RNA and an anti-sense RNA, long fold-back RNA structure or through the action of an RNA-dependent RNA polymerase (RDR2) [78]. While the origin of the double stranded RNA molecule is not always evident, the mechanism for its processing is fairly well elucidated. The double-stranded RNA (dsRNA) is cleaved by Dicer-like3 (DCL3) to produce 24nt siRNA molecules which are then methylated by Hua-Enhancer1 (HEN1) [70].

The major described roles for siRNAs in plant cells are virus defense and genome integrity maintenance via transcriptional inactivation and post-transcriptional gene silencing (PTGS). Transcriptional inactivation is achieved by directing a chromatin modification apparatus to specific regions of the genome. In short, the HEN1 methylated siRNAs bind to a specific ARGONAUTE family member (AGO4), which then recruits other DNA binding and methylation enzymes, leading to the heterochromatinization of the target region. This mechanism of chromatin modification is believed to be largely responsible for the extensive methylation of transposable elements and other repeat regions of the plant genome. Thus the siRNA pathway is very important in the maintenance of genome integrity by silencing the large transposon load carried by plants.

PTGS is a powerful method of regulating the mRNA levels of target genes to fine tune their expression. Overexpression and/or ectopic expression of genes is likely to trigger the siRNA mediated silencing pathway possibly as a broad scale defense mechanism against viral infections. RNA dependent RNA polymerase 2 (RDR2) or other RNA dependent polymerases, use the target mRNA as template to generate a complementary strand. These long dsRNA molecules trigger the generation of siRNAs, which in turn direct the RNA induced silencing complex (RISC) to cleave the original RNA, as seen in the silencing of viral genes and transgenes.

A large majority of the siRNA population of plant cells is heterochromatic i.e., it directs the methylation of its target region of the genome. These siRNAs are typically processed from a dsRNA generated either from the target locus or a trans-site with a high sequence similarity. Based on origin, two more classes of siRNAs have been reported in plants. In plant genomes certain genes exist in close proximity in a head-to-head orientation such that they share the same sequence (complemented) in their UTRs. In certain tissues or developmental phases the coexpression of these genes leads to the occurence of natural anti-sense RNAs which complement perfectly in their shared UTRs. These transcripts then hybridize and initiate the production of siRNAs called natural anti-sense siRNA (nat-siRNA). These siRNAs have been shown to silence the expression of these two genes and any other genes sharing significant sequence similarity [11]. The second class of siRNAs differing significantly from the heterechromatic siRNA are tasi-RNAs. tasi-RNAs are generated by transcription of a tasi locus, which is then cleaved by a specific microRNA (miRNA) in a phasing manner. The miRNA binds to the transcript at a specific position of complementarity and cleaves it. One of the cleaved portions, either 5' or 3', is bound by SGS3 and RDR6 and complemented to produce a dsRNA molecule that is then processed by DCL4 and HEN1. This process generates 21nt long siRNAs in a phased manner starting from the miRNA binding site [79]. This special class of siRNA was postulated to have evolved to allow the transmission of an miRNA signal across cell boundaries [70].

3.1.2 miRNA function and biogenesis

The earliest discovery of small RNA form and function involved the transcriptional regulation of lin-14 by lin-4 [74, 80]. Lin-4 encodes an mRNA that can form a fold-back structure, leading to the generation of a small RNA whose sequence complements partially to the mRNA of to lin-14. A specific small RNA generated

from the fold-back structure of a gene coded transcript and targeting other mRNA for cleavage is called miRNA.

The second miRNA discovered (let-7) was shown to be conserved across the animal kingdom, implying that these molecules have evolved early in the development of eukaryotic organisms. The basic pathway of miRNA biogenesis and function shares a high degree of functional similarity between the plant and animal kingdoms. However, none of the individual components in the pathway are very well conserved in protein sequence. Besides, none of the known miRNAs are conserved across the plant and animal kingdoms. Therefore, the miRNA pathway in plants is thought to have evolved independently of the counterpart in animals. Evolution of the miRNA pathway added to the complexity of spatio-temporal gene regulation which enables the spatio-temporal development of the highly differentiated organs in the higher organisms. Partial evidence for this hypothesis is provided by the rough correlation between the number of miRNAs in a genome and the complexity of the organism [81].

miRNA genes show a wide range of conservation across organisms. No known miRNA is conserved amongst plants and animals, but many miRNAs are known to be well conserved within each kingdom, suggesting that they evolved early in evolutionary history. In addition to these highly conserved miRNAs, a large number of miRNAs that are specific to one organism or a small clade are being discovered [82]. Multiple degrees of conservation in this class suggest that miRNA genes undergo rapid evolution with a few exceptions. A convenient model to describe this observation is that novel miRNA genes arise fairly commonly and along with their targets are subjected to selection. All interactions that confer a beneficial regulation model to the plant are fixed and passed on to the progeny [83]. Since the miRNA pathway is a secondary regulation method with possible redundant effectors in the primary transcriptional regulation and/or translational regulation, the organism might be more robust to the higher gain and loss rate of the miRNA genes compared to protein coding genes.

miRNA generation in the plant cells is a multi-step process involving transcription of the non-protein coding sequence, excision of the hairpin from the longer molecule, separation of the strands, methylation of the single stranded RNA molecule and finally loading of the single stranded miRNA into the RNA-induced silencing complex (RISC). miRNA genes are transcribed by RNA polymerase II (Pol II) [84] and undergo similar modifications to protein coding mRNAs, namely they are polyadenlated at the 3' end and have a 5' cap to protect the molecule from degradation. This long noncoding RNA molecule, called pri-miRNA, typically does not have an open reading frame of significant length and is not bound by the translation apparatus. Instead the molecule forms a fold-back structure, attracts a DCL family protein called DCL1 [85] which excises the stem-loop precursor of miRNA, called pre-miRNA. DCL1 further processes the pre-miRNA to excise the dsRNA comprised of the mature miRNA molecule and its complemetary sequence (miRNA*). Both strands of this dsRNA are methylated at the 3' end by HEN1 and the miRNA strand is then loaded into a RISC complex primarily by the action of AGO1. Single stranded RNA degrading enzymes act in parallel to degrade the free miRNA and miRNA* molecules, to remove the miRNA* sequence from the cell and to regulate miRNA levels in steady state.

Two studies aimed at following the evolution of miRNA genes focused on the sequences immediately flanking the novel mature miRNA in Arabidopsis [86, 87]. By focusing on distinct subsets of these sequences they arrived at two different models for the genesis of novel miRNA genes. Allen et al. [86] studied the subset consisting of novel miRNA genes that show a high degree of conservation between the flanking sequences of the mature miRNA in the known miRNA gene and the predicted targets for this gene. By comparing the sequence similarity and positions of the matching fragments they postulated that novel miRNA might evolve by the generation of tandem inverted copies of a particular gene. This duplication event generates a double stranded RNA initially and accumulates mutations over time, to generate the classic mismatched pairing required for DCL1 processing and thus

a new miRNA is born. de Felippes et al [87] instead studied the subset that show no similarity to their targets or to other regions of the genome and concluded that miRNA genes arise from natural inverted repeats occurring randomly in the genome. The novel miRNAs thus generated are subjected to selection, and any beneficial molecules are incorporated into the regulation mechanism. Appearance of novel miRNA genes might be a result of either or both of these mechanisms or other yet undescribed processes. In all cases the requisite conditions for the development of an miRNA gene is the development of a hairpin with enough mismatches in the stem to allow recognition and processing by DCL1.

An intriguing aspect of miRNAs is that most conserved miRNAs exist not as a single locus in the genome but a multitude of loci spreading over the genome with little conservation in the flanking sequence. For example, miR156 is a highly conserved miRNA in plants and is believed to target SBP transcription factors. The soybean genome has 12 loci that encode this miRNA (Figure 3.1) and the alignment of these loci shows clearly that there were 6 loci for the gene before the last genome duplication event in this organism's history. Additionally many miRNAs exist as families of nearly identical sequences with one or two differing bases, or showing length polymorphisms. Therefore, the amount of variation within an active miRNA molecule, when normalized to the length of the molecule, is substantially greater than a protein coding gene without affecting the function of the small RNA molecule significantly. It is unclear what the importance of these polymorphisms is in the function of the miRNA pathway. Interestingly some of the length polymorphisms are expressed differentially among tissue types [88]. It is possible that the length and sequence polymorphisms allow an miRNA molecule to regulate a wider range of targets that might themselves show tissue-specific expression.

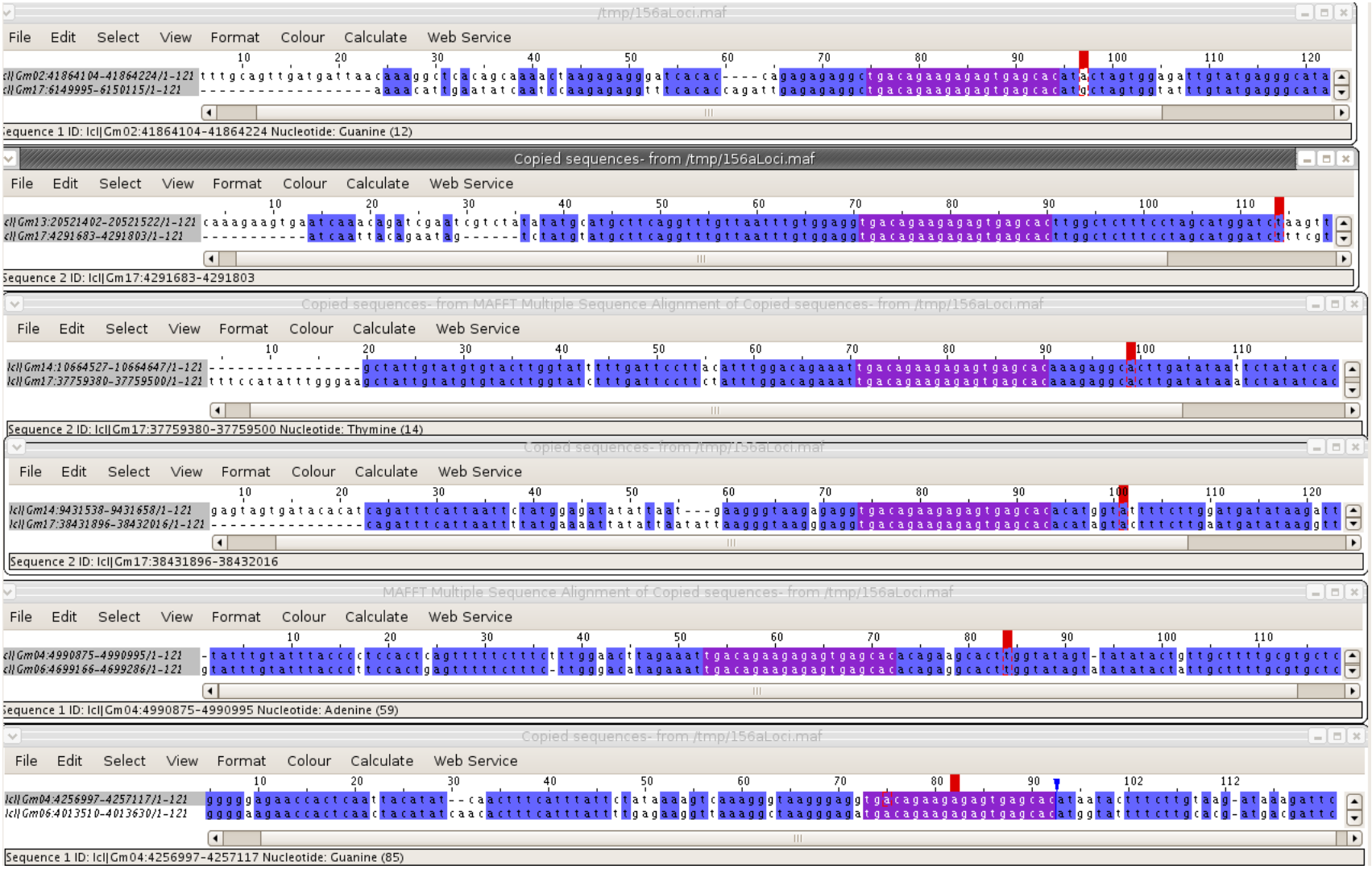

Figure 3.1: **Genomic loci encoding miR156.** mir156 is a crucial miRNA in the development of plants. There are twelve genomic loci in soybean that encode this single member of the mir156 family.

3.1.3 miRNA prediction

Numerous high-throughput expression studies in a wide range of organisms subjected to various treatments have examined the expressed small RNA populations[11, 89, 90]. In plants, the overall population of small RNAs tends to be dominated by the endogenous siRNA class. Apart from these abundant species, a large number of small RNA molecules have been discovered to exhibit many of the salient characteristics [91] of miRNAs. While high-throughput small RNA sequencing (RNAseq) results provide evidence of expression, the verification of existence of hairpin precursors for each of these small RNAs is not possible with the experiments and would take enormous time and effort to confirm experimentally. Instead, computational methods and genome sequence data can be used to identify those small RNAs that are processed from a stem-loop hairpin structure.

The main method of screening the huge number of potential miRNAs from an RNAseq experiment is to identify the presence and quantify the stability of a genomic stem-loop hairpin structure involving the small RNA under investigation. A computer algorithm that folds any given stretch of nucleotides into all possible secondary structures and estimates their thermodynamic stability is often employed in this process. mFold/UNAFold [92] has emerged as a popular program for this purpose. UNAfold allows a DNA or RNA molecule to fold into every possible secondary structure through unconstrained base pairing and calculates the thermodynamic stability of such structures. Stability of the structure is measured as the free energy (dG) of the molecule in solution. Using a gross simplification, the dG scales with the number of bases paired. The number of possible stable secondary structures for a given molecule is often large, even above a user-specified threshold, necessitating the use of other heuristics to reduce the list of possible miRNAs.

3.1.4 Small RNA content of the soybean genome

With the availability of the completed soybean genome [18] and numerous small RNAseq datasets from various groups, it has become possible to increase our understanding of the structure and maintenance of the genome through small RNA. The genome sequencing effort has broadly defined the repeat composition and distribution in the soybean genome. Sequence analysis of the final genome assembly and stages preceding that has allowed the documentation of transposable elements families into databases such as soyTEdb [93]. Approximately 60% genome of soybean was estimated to be composed of high copy number repeat elements, a majority of which are transposable elements [9, 18]. Some highly expressed transposable elements were inadvertently captured in cDNA libraries and EST sequencing efforts and provided insight into the most active transposable elements [19].

The presence of a wide variety of high copy number repeat families implies a great expansion of transposable elements, either gradually or sporadically, in the evolutionary history of soybean. Most of these elements are believed to be dormant because they lack functional copies of the genes required to excise and transpose. However, these elements might yet be passively transposed by the action of other elements or through meiotic recombination. Given the large copy number of repeats, a mass activation of transposons would be highly deleterious to the genome, as it would affect its fundamental stability and disrupt the function of numerous genes. Therefore, the plant needs to constantly suppress the activity of transposable elements in a broad manner and specifically react to the sudden activation of one or more kinds of repeats, herein referred to as maintenance of genomic stability. This role has been shown to be fulfilled by hererochromatic siRNA in other plants. Therefore, it is highly likely that some or all of these repeat regions of the genome serve as origins or targets for the siRNA mediated methylation pathway. Due to the exact complementarity of siRNAs to

their targets, drawing a disctinction between origin and target is likely to prove challenging.

The second type of information not readily revealed by the genome sequencing is the miRNA content of the genome. Even the identification of well conserved miRNAs is not trivial, since the number of loci matching a known miRNA vary between species and its not obvious whether all the loci produce a functional miRNA. Studies in many species of plants have revealed miRNA genes that are not well conserved in other fully sequenced genomes. Therefore with each new plant species sequenced, the number of miRNAs discovered will likely increase, until all major branches of the phylogenetic tree have a representative fully sequenced and analyzed. The prediction of novel miRNAs based on expression evidence and stability of secondary structures is therefore most likely to dramatically increase our knowledge of the miRNA network in plants. The increasing knowledge of novel miRNAs and their targets will greatly enhance our understanding of the basic regulation mechanisms involved in cellular processes and responses to stimuli. However, current methods for miRNA target prediction though lack the desired specificity and sensitivity and will likely need more improvements in their algorithms. Nonetheless, the identification of likely miRNAs and more tentatively their targets should advance our understanding of cellular and developmental biology.

3.2 Methods and results

To discover the noncoding RNA content of the soybean genome we collated small RNAseq data from multiple sources. Data from Tuteja et al. [11], and two as yet unpublished studies focusing on pathogen response and photomorphogenesis were analyzed in this study. While the experimental design for each of these studies is diverse and not comparable across them, they do provide a large amount of small RNA sequence and expression information from a wide range of soybean

tissues. Each of the individual experiments involved the extraction of the small RNA component of the target tissues followed by massively parallel sequencing on a Illumina genome analyzer platform. The tissue types represented across these experiments include seeds (various sub-structures and development stages), developing embryo, seedlings, stem and leaf. While most plants sequenced were from the Williams 82 genotype also used in the soybean genome sequencing effort, some of the datasets were derived from other accessions of soybean.

Raw sequencing data from a total of 19 libraries of soybean small RNA was compiled. The raw reads from each sequencing run were appropriately trimmed and filtered based on the particular qualities of each run. Specifically, each run was processed to trim the 3' adapter sequence used in the sequencing technology and all reads with a resultant length of less than 16 base pairs were removed. Then low quality reads, defined as any read with one or more ambiguous base calls, were removed. Some of the datasets contained a large proportion of reads (40-60% of total reads) derived from rDNA and tRNA while others had less than 3% of total reads derived from such sequences. The datasets with a large number of rDNA and tRNA reads were filtered to remove all reads matching known rDNA and tRNA sequences. No restrictions based on minimum number of reads were applied to any of the datasets. After all data preparation steps were completed a total of approximately 124 million reads remained. Many of these reads represented the same small RNA sequence and their frequencies were unaltered.

3.2.1 Mapping to genome and block analysis

All sequencing reads were then aligned to the Glyma1 version of the soybean genome assembly (JGI) using the short read alignment program Novoalign (v 2.04). Despite the various filtering steps about 23% of the reads were filtered out by the alignment program for failing various quality thresholds. Of the total set approximately 58% i.e., 72 million reads were sucessfully aligned to the soybean

genome when no mismatches were allowed. The high level of stringency was enforced to allow the unambiguous determination of origin for the small RNA sequence. Relatively low percentage of reads mapping to the genome is likely caused by one or a combination of the following factors:

1. Poorly sequenced regions of the soybean genome do not allow mapping of reads. This problem is especially relevant in the repetitive pericentromeric regions which are likely the targets of heterochromatic siRNAs.

2. Reads derived from a different genetic background are more likely to carry single nucleotide polymorphisms (SNPs), especially in the transposable elements which show rapid divergence between lines.

3. Residual nucleotide variation has been reported in the Williams 82 line [94] which could mean the lines sequenced in this study differ from the line used for genome sequencing. [Also see Chapter 4.]

4. Erroneous adapter trimming and sequencing errors in the small RNAseq experiments create superfluous bases and SNPs disallowing perfect matches.

Heterochromatic and other endogenous siRNA are formed by the cleavage of long transcripts and act in collaboration to the recruit histone methylation apparatus to their targets. Therefore, individual siRNA sequences are not the functional unit of regulation. Instead the set of such siRNAs targeting a single genomic locus is the functional unit. Therefore, grouping or clustering siRNAs based on their origin on the genome gives a more appropriate perspective to look at the siRNA population, rather than analyzing individual molecules. Such clustering also has the added advantage of reducing the computational load for handling the rather large datasets produced by high-throughput sequencing. Most of the soybean genome outside the telomeric, ribosomal, centromeric and pericentromeric regions is very well sequenced. However, there remain small unsequenced patches which prevents any small RNA reads produced from that region from mapping. Additionally, due to the random nature of sampling inhererent to the RNAseq method, regions that produce a small proportion of the total small RNA will not

be evenly covered by reads even if those regions are producing small RNA actively, but at a low level. The number of small RNA reads mapping to any given position of the genome is hereby called the"coverage" of that region. Therefore any method attempting to cluster small RNA reads based on genomic origin should be robust to the small gaps in coverage caused by these deficiencies. The maximum length of a gap allowed is fairly arbitrary and any value chosen would wrongly cluster or fail to cluster genomic loci correctly. In this study we used a fairly arbitrary cutoff of 500 base pairs as the maximum allowed gap in coverage before a cluster is terminated and a new one started, partly because this value of the parameter produces a manageable number of clusters. Lowering the maximum allowed gap size drastically increases the number of clusters while raising this threshold too far results in adjoining genes or transposable elements being grouped into a cluster with very varied coverage.

3.2.2 siRNA blocks and maintenance of genomic stability

The soybean genome is very rich in repetitive elements, with estimates of upto 40% of the genome being consisting of high copy number repeats [9]. Apart from the large batteries of tandem repeats characteristic to the soybean centromeres and pericentromeres, a large proportion of the repeat content is composed of transposable elements. Therefore we expect to see a large number of siRNAs directed towards the control of these elements through the heterochromatinization pathway. Two approaches were taken to identify the small RNA associated regions of the genome. These regions are either the origin, or target, or both, of any given cluster of siRNA.

The first approach was aimed at identifying all regions of the genome that produced or were targeted by siRNA in the tissues and conditions sampled. To this end the small RNA mapping and clustering methodology, described earlier, was performed. Of the 955,054,837 bases in the Glyma1 assembly 149,350,689

bases were covered by at least one small RNA read i.e., approximately 15.6% of the soybean genome is involved to some extent in the small RNA pathways. The vast majority of these loci are likely to be transposable elements that are targeted for methylation. The clustering approach identified 357,889 active blocks of siRNA in the soybean genome. Blocks that were very "sparse" i.e., had less than a total of 100 reads (all libraries pooled) were excluded from analysis to identify regions of relatively high activity. This filter also removed false clusters formed by spurious mapping of siRNAs to short regions of similarity. As a result 23,226 blocks were left that had at least 100 or more reads mapping to them. Higher cutoffs for depth reduces the number of blocks substantially. While this approach allows the identification of small RNA producing/targeted loci interpretation of the results is not trivial and depends heavily on the availability of reliable annotation for these regions. Nonetheless studies aimed at observing the small RNA regulation of specific loci would greatly benefit from this approach [11].

The second approach attempted to directly estimate the number of siRNA directed at the various families of transposable elements known to exist in the soybean genome. The collection of known transposable elements in soyTEdb [93] was used as the reference sequences to map the pooled small RNA from all libraries. The program maq [95] was used for this mapping since it allows more mismatches and filters out fewer reads. The sequences in this database span a total of 169,491,207 bases in the soybean genome. Of these only 61,627,752 bases (36.4%) had any small RNA mapping to them. This smaller portion of soyTEdb could represent the active portion of the transposable elements in the genome. The distribution of the covered regions among the various elements does not show the complete absence of small RNA mapping to any single element. Instead some of the transposable elements show a very low frequency of reads mapping to them which can indicate that these elements are no longer active and therefore do not need to be supressed or they undergo methylation through a different mechanism that does not involve siRNA.

To allow the comparison of relative levels of small RNA among the various elements, the reads mapping to mulitple loci were normalized to reduce their contribution to any one locus linearly with the increasing number of loci the read maps to. This normalization allows the computation of weighted counts of small RNA reads mapping to any one transposable element. The raw counts were further normalized to the length of the element to compute the reads per Kb (RPK) of the element. These counts can then be fairly compared among the elements to measure the amount of siRNA within the cell that is directed towards that element. Elements that are actively transcribed would initiate the siRNA pathway and therefore have a larger number of small RNA that exactly match its sequence.

Table 3.1: **Twenty most abundant siRNA mapping transposable elements.** Small RNA was mapped to the soyTEdb database of transposable elements and weighted count of reads mapping to each element was normalized to length of the element (RPK). Shown here are the 20 elements that show the greatest number of siRNA generation. The superfamily refer to the kind of transposable element while the family represents the subtype whose elements share a great deal of sequence similarity.

ID	Superfamily	Family	RPK	Length
RLG_Gmr448_Gm2-1	Gypsy	Gmr448	28151	3969
RLG_Gmr330_Gm7-1	Gypsy	Gmr330	29251	4912
RLG_Gmr169_Gm15-10	Gypsy	Gmr169	29341	10808
DTM_uuu_Gm20-93	Mutator	Ukn	31321	1602
RLG_Gmr21_Gm1-13	Gypsy	Gmr21	31386	19508
DTM_uuu_Gm8-81	Mutator	Ukn	32208	7908
DTM_uuu_Gm10-128	Mutator	Ukn	34913	391
RLC_Gmr73_Gm17-1	Copia	Gmr73	36491	1067
RLG_Gmr4_Gm18-58	Gypsy	Gmr4	38766	10866
RLG_Gmr213_Gm5-1	Gypsy	Gmr213	39605	5269
RLG_Gmr213_Gm3-1	Gypsy	Gmr213	39691	5253
DTM_uuu_Gm17-110	Mutator	Ukn	40017	726
RLG_Gmr213_Gm13-1	Gypsy	Gmr213	40728	5131
RLG_Gmr9_Gm2-145	Gypsy	Gmr9	41482	18283
RLG_Gmr213_Gm19-1	Gypsy	Gmr213	46886	4453
RLC_Gmr5_Gm1-177	Copia	Gmr5	50144	6770
DTM_uuu_Gm15-118	Mutator	Ukn	57237	9485
RLG_Gmr17_Gm4-19	Gypsy	Gmr17	63053	944
RLG_Gmr539_Gm11-1	Gypsy	Gmr539	212932	5190
RLG_Gmr539_Gm2-1	Gypsy	Gmr539	1787865	5060

The most common kind of transposable element in the soybean genome is the Gypsy element followed by Copia and Mutator. Unsurprisingly the siRNA population mirrors this distribution with the Gypsy superfamily garnering 4.65 million RPK while the Copia superfamily gains 2.28 milliion Reads per Kilobase (RPK) and the Mutator superfamily gains 2.03 million RPK. A look at the top 20 most abundant siRNA targeted loci (Table 3.1) shows the subtype of elements that are most actively suppressed. The Gmr539 and Gmr213 families of Gypsy elements are responsible for the largest number of siRNA in the soybean cells. These families most likely represent the subtypes of Gypsy elements that are currently active in the soybean genome. Similarly the Gmr5 and Gmr73 families of Copia element are responsible for a large number of siRNAs and are also actively transposing in soybean. The mutator class of elements lack the family level classification in soyTEdb. The most abundant mutator elements show a large variation in size ranging from 391 bases to 9485 indicating that these individual elements represent a range of Mutator elements that are all actively transposing in the soybean genome.

The presence of a large number of gypsy, copia and mutator elements and their activity implies that there is a tremendous potential for genomic rearrangements in soybean. Particularly the gypsy family of elements is very large and there is some evidence of its expansion in the high yielding cultivars of soybean (see Chapter 4). Normally the activity of these elements is heavily suppressed by the activity of siRNA, as shown here, thus conferring a fair degree of genomic stability. But under conditions of stress or other activation events these elements are activated and may cause large scale genomic changes that can lead to the development of desirable traits [20].

3.2.3 Learning miRNA folding parameters from known miRNA

miRNA folding algorithms are typically optimized for human miRNA genes in the sense that the default settings of the programs are optimized to find stable folds for pre-miRNA sequences from the human genome. The primary measure of goodness of a fold used by the different folding algorithms is the free energy of the folded structure represented by dG. The dG associated with a molecule is inversely related to the thermodynamic stability of the folded structure, which in turn scales linearly with the number of hydrogen bonds between nucleotides in that structure. A perfect base pairing similar to the DNA double helix though would violate the unique requirement of DCL1 action in miRNA processing, which is the presence of a few mismatches in the stem-loop pre-miRNA structure. Therefore, optimizing purely for the minimization of free energy is not advisable for detecting novel miRNA loci.

A group of plant small RNA researchers established a set of rules to facilitate the standardized annotation and nomenclature of novel miRNA from the rapidly growing deep sequencing data [96]. These broad rules encompass the criteria used by most previous studies in this field. While the fold characteristics of a miRNA locus can be somewhat defined, the set of parameters assigned to the folding software are not readily defined and are likely to change for each species. For example, the length of the pri-miRNA molecules is distributed over a much wider range in plants than animals and the required number of paired bases is also lower. To determine the correct program parameters for maximizing the sensitivity of the prediction process, known miRNA generating loci were folded with a range of parameter values to ascertain the set conferring maximum sensitivity.

To accomplish this task all known *Glycine max* miRNAs and their precursors were downloaded from miRBase (as of Feb 2010). This set included 75 miRNAs with experimentally or computationally defined pre-miRNA sequences and their

predicted secondary structures. These pre-miRNA sequences fit the aforementioned rules for miRNA loci to various degrees with some of the well-defined and widely conserved miRNA, for example miR160, fitting the rules perfectly, while some of the newer, predicted folds, for example miR1521, fitting less well. A few of the predicted loci, for example miR1536, seemed to completely lack the stem-loop characteristics associated with miRNA and were most likely mischaracterized by the submitting authors.

The various parameters allowed to change in the training set are:

1. Folding temperature i.e., the assumed ambient temperature at which the thermodynamic stability of the molecule is calculated.

2. Threshold dG i.e., the maximum free energy cutoff above which a fold is rejected.

3. Window i.e., the amount of variation allowed in the secondary structures explored to find the best fold. As the size of the window increases the diversity of structures considered increases.

4. Length of pri-miRNA sequence i.e., number of adjoining bases to the mature miRNA that are submitted to the folding algorithm.

The set of parameters showing the highest specificity was determined by visually comparing the predicted folds from the program to the submitted folds in miRBase and scoring them individually for similarity. Following values for each parameter were found to be optimal for finding the "correct" secondary structure of the pri-miRNA: Folding temperature = 25C, Threshold dG = -40, Window = default, Length =170 bp. With these parameters all the conventional miRNA folds were readily detected (Figure 3.2 and 3.3). While a reasonably stable secondary structure was identified for some of the "unconventional" miRNA (sequences not showing good pairing in miRBase), it proved difficult to determine the correctness of this fold (Figure 3.4). Hence the latter set was removed from further analysis.

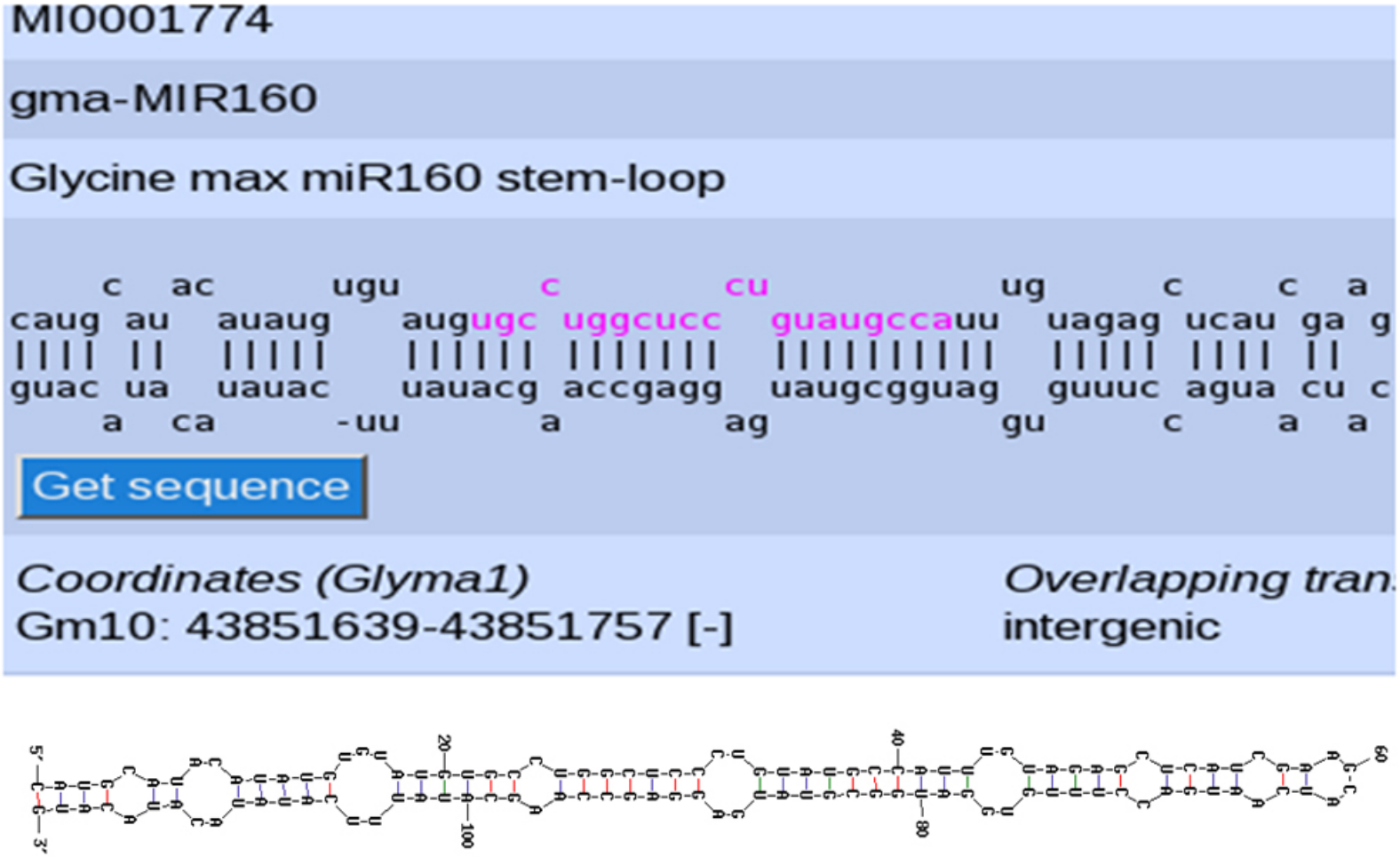

Figure 3.2: **Canonical secondary structure of pri-mRNA.** Glycine max miR160 locus showing a traditional pre-miRNA fold in miRBase is depicted at the top. The corresponding predicted fold from the folding pipeline used in this study is shown at the bottom to compare. The two folds are almost identical when the chosen parameters were used for UNAfold.

gma-MIR1521

Glycine max miR1521 stem-loop

```
----------------------------------------------------uacuca             aa          c  a          c         u  -aa   u
                                                  uugacuguuaaugg  aauguugacug ca augucauaauc uauuggaugu aa   uug a
                                                  ||||||||||||||  |||||||||||| || |||||||||||| |||||||||| ||   ||| 
                                                  aacugacaauuacc  uugcaauugac gu uacaguauuag auaaccugca uu   agc u
cuuguaauuaacaauaacuuauuacaacugacaauacacaguguuaaaauaaccugc             ca          a  a          u         c  aaa   a
```

Get sequence

Coordinates (Glyma1)	*Overlapping transcripts*
Gm11: 18237729-18237918 [-]	intergenic

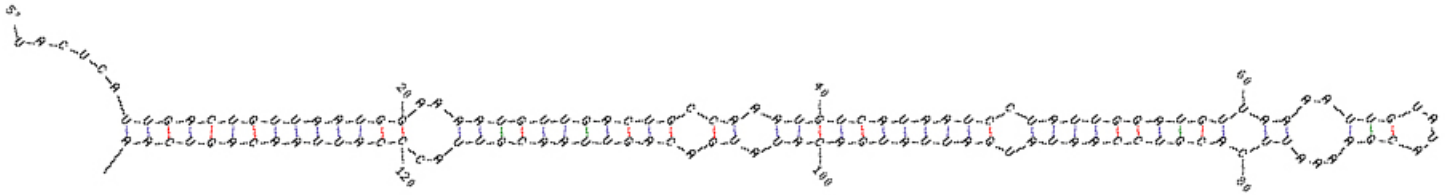

Figure 3.3: **Extended stem in a predicted pri-miRNA.** Glyince max miR1521 showing an extened 3' stem as deposited in miRBase is shown in the top half and the corresponding predicted fold from the pipeline is shown below. The structure predicted with UNAfold with the chosen parameters is almost identical and differs only in minor positions at the ends.

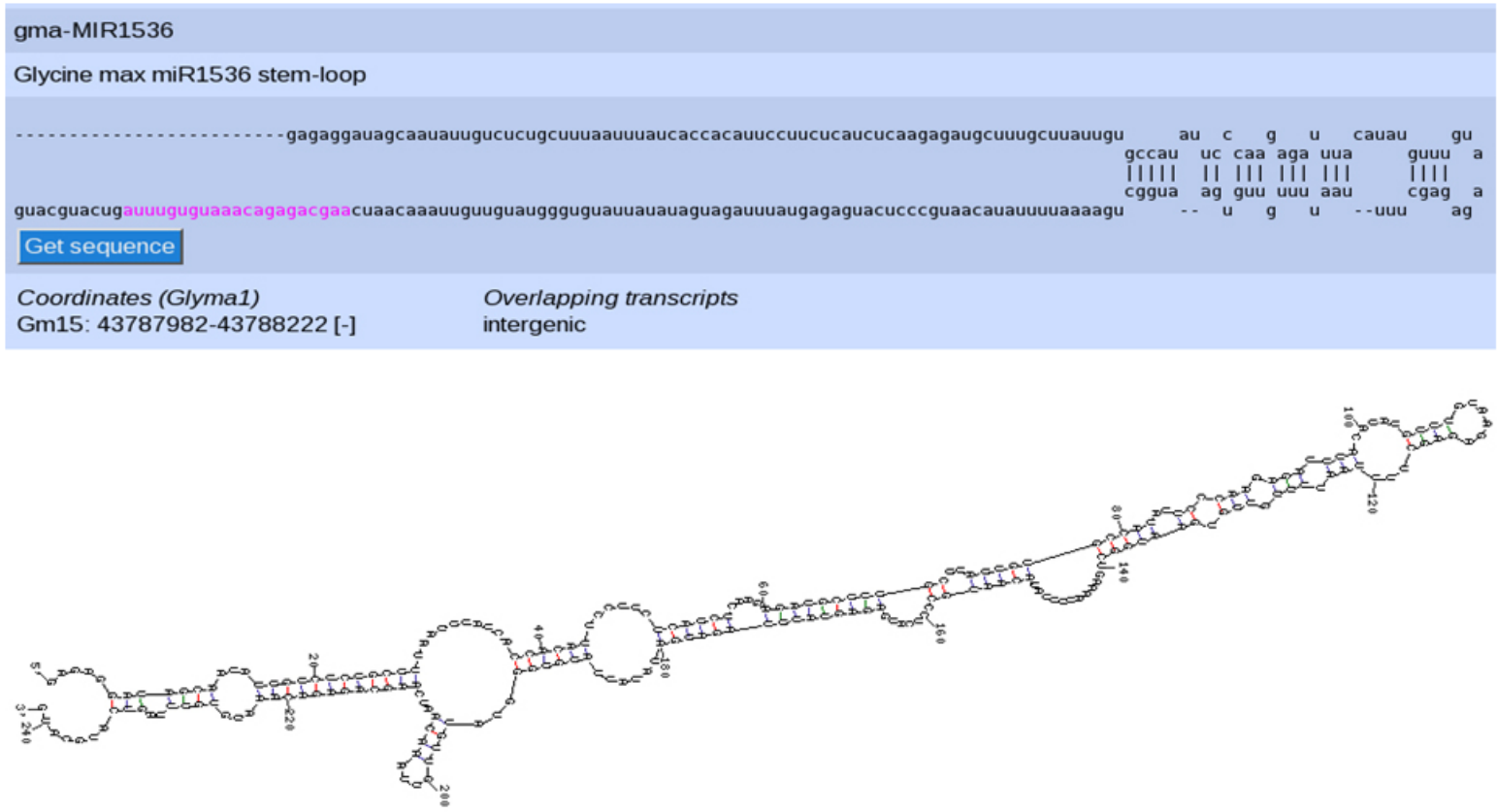

Figure 3.4: **Computationally predicted miRNA with improper pairing.** A computationally predicted miRNA locus shows insufficient pairing of miRNA and miRNA* sequences. A stable secondary structure was identified at this locus, but its specificity cannot be determined due to lack of a known structure.

In addition to detecting the right parameters of the fold, quantifiable criteria for the goodness of a predicted fold are needed to process the large number of small RNA producing loci in the soybean genome. Various properties of the output stable secondary structures were studied to identify a set of filters that can be readily applied to screen the large number of genomic loci expected to produce stable secondary structures. The primary metric used by the folding algorithm is the dG which represents the overall stability of the molecule but does not pay special attention to the miRNA or miRNA* sequences. Genomic regions with short tandem repeats or large nucleotide biases can stochastically produce stable folds around the small RNA. The likelihood of a stable fold occurring stochastically increases with the length of sequence submitted for folding. Therefore the selection of putative miRNA based purely on dG causes too many false positives. Nonetheless dG offers a convenient first filter to identify regions that form any form of a stable fold. Empirical observations based on the training set indicate that for a length of 170 bases a maximum dG of -40 is a good threshold for filtering chance occurrences. Additional properties of the folds that were highly predictive of a canonical loop were that over 75% of the bases in the small RNA need to be paired and the length of the complementary sequence (predicted miRNA*) should not be more than 1.5 times the length of the small RNA, and that no bases in the small RNA or within 10 bases of its end can be in the complementary strand of the stem-loop i.e., the distance between an miRNA and its complement should be at least 20 bases.

Applying this knowledge small RNAs from the deep sequencing projects that mapped to less than or equal to 20 genomic loci were submitted to the prediction pipeline. 75 bases on either side of the small RNA was extracted and the approximately 170 bp sequence was submitted to UNAfold to find stable structures at 25C with a maximum free energy of -40 kCal/mole. The small RNA producing loci that exhibited stable structures were checked against the other criteria to screen out spurious folds. The next filter applied was to screen these loci against

known repeat sequences, as indicated by the soybean genome project, to remove the small number of repeat sequences that pass all these filters. Finally the set of putative miRNAs were aligned to known miRNA sequences from miRBase to filter out the known miRNA sequences.

3.2.4 Identification of novel miRNA

Deep sequencing data from 11 small RNAseq experiments was first pooled into a single set. These experiments were designed to look for small RNA regulation in various phenomena of interest. Data from Tuteja et al. [11] was combined with small RNAseq from germinating cotyledons, young embryo, and normal and pathogen challenged stems and leaves. A total of 26,695 small RNA sequences that mapped to less than 20 loci in the genome were used. A total of 86,972 loci were thus submitted to UNAfold. Of these approximately 800 small RNA sequences are predicted to form a stable fold with a dG lower than -40 and fulfilling the other structural criteria mentioned from at least one of the loci they map to.

While plants display a very diverse population of small RNAs, thousands of miRNA genes in a single genome may be an overestimation of the miRNA gene space. Further filters were therefore applied to reduce the number of false positives from this set. These loci were filtered to remove any that lie within a known repeat or a gene or match a known Glycine max miRNA sequence. Additionally only the small RNA in the size range of 21-22 nucleotides were retained. These filters reduced the number of putative, previously undescribed miRNA to 155 sequences. The 20 most abundant predicted miRNAs are shown in Table 3.2

Table 3.2: **Twenty most abundant novel miRNA from Stem, Leaf, Seed and Root tissues.** Sequences of novel putative miRNA is shown along with the number of genomic loci encoding it. The total number of reads matching this sequence perfectly, across all libraries, is shown. The fourth column shows the tissue types if any that most reads are derived from. None in this column implies that none of the tissues showed a substantially higher count than others for this sequence.

Sequence	No. of mapping loci	Total reads	Abundant tissues	Predicted EST targets
AGGGATAGGTAAAACAACTAC	1	383676	Seed, Stem,Leaf	12
AGGGATAGGTAAAACAATGAC	1	63362	Seed coat, Cotyledon	10
GGAATGGGCTGATTGGGAAGC	1	57093	Seed coat, Stem, Leaf	4
GGAGATGGGAGGGTCGGTAAAG	2	18826	None	3
AGAGGTGTATGGAGTGAGAGA	1	11376	Seed coat, Stem, Leaf	16
AGAGATGTATGGAGTGAGAGA	1	10230	None	13
TTGAAGCTGCCAGCATGATCTT	2	7771	Cotyledon	3
TTGACAGAAGAAAGGGAGCAC	2	7723	Seed	22
GGAATGGGCTGATTGGGAAGT	1	7259	Seed	4
CTGAGACCAAATGAGCAGCTGA	1	5635	Leaf	4
TATGGGGGGATTGGGAAGGAA	2	5397	Seed	12
TAAGACGGAACTTACAAAGATT	2	4353	Stem, Seed coat	7
GCGTATGAGGAGCCAAGCATA	2	2200	Root	0
AGAGGTGTTTGGGATGAGAGA	1	1429	Leaf	11
TGACAGAAGAGAGTGAGCACTT	2	1119	None	17
GGAGGCGTAGATACTCACACC	1	1049	None	0
CTGACAGAAGAGAGTGAGCAC	2	984	None	22
TAGCCAAGAATGACTTGCCGG	1	972	Root	9
TGACAGAAGAGAGTGAGCACAA	2	795	None	19
TCGGATTAGCAGGTTTGAGGAA	1	744	Leaf	2

Table 3.3: **Twenty most abundant novel miRNA from seedlings.** Sequences of novel putative miRNA is shown along with the number of genomic loci encoding it. The total number of reads matching this sequence perfectly, across all libraries, is shown. The fourth column shows the tissue types if any that most reads are derived from. None in this column implies that none of the tissues showed a substantially higher count than others for this sequence.

Sequence	No. of mapping loci	Total reads	Abundant tissues	Predicted EST targets
GGAGGCGTAGATACTCACACC	1	565765	Hypocotyl, Hook	0
TGTTGCGGGTATCTTTGCCTC	1	79355	Hypocotyl, Hook	6
TGAACTATACAAAGACGGTTA	3	40862	Hypocotyl, Hook	2
AGAGATGTATGGAGTGAGAGA	1	35423	Hypocotyl, Hook	13
TTGTTCGATAAAACTGTTGTG	1	27976	None	15
GAGTGGATCTGAGAACACAAGG	1	17305	Hypocotyl, Hook	6
AGAGGTGTATGGAGTGAGAGA	1	16393	Hook	16
GTTGAATGGTATTGTAGTACT	1	14081	Hypocotyl, Hook	11
AATCATCGTCAGATTTTGAGGC	5	12498	Hook	11
TTTTGGTATATCGTTAGACGAC	7	11392	Hook	2
TGGAGACCTGAACTGAAGAGG	6	10734	Hypocotyl, Hook	7
AGTTGAATGGTATTGTAGTACT	1	10403	Hypocotyl, Hook	9
GTGGTATCAGGTCCTGCTTCA	1	10328	Hypocotyl, Hook	6
TTTGGTATATCGTTAGACGACG	7	9331	Hypocotyl, Hook	5
TGGAGACCTGAACTGAAGAGGC	6	8306	Hypocotyl, Hook	6
TGAGAATTTGGCCTCTGTCCA	2	7327	None	5
ACACTGATATGTTTGCAAGCGA	5	7143	Hypocotyl, Hook	4
GGAGATGGGAGGGTCGGTAAAG	2	6810	None	3
CGATGATTAGTTAGTTGTAGTA	6	6476	Hypocotyl, Hook	7
ATGAATTTGATTGTAGATGGC	2	5560	None	14

The same approach was applied to a small RNAseq dataset involving germinating seeds. Small RNA from the cotyledon, hook and hypocotyl of dark grown and light treated seedlings was extracted and deep sequenced. The sequence data from this set was also processed in the same way as the previous set. Approximately 200,000 loci were tested within this set and after all the filtering steps a total of 158 putative miRNA sequences were identified. The 20 most abundant predicted miRNAs are shown in Table 3.2.

There is a very small overlap between the two sets of putative miRNA predicted yielding a net result of 312 putative novel miRNA sequences. A large number of these molecules, especially in the first dataset with fewer total reads, are represented by very few reads in the underlying libraries. This could imply either that they are very low abundance transcripts in the tissues examined or that some of them are false positives that were not adequately filtered. The putative miRNA represented by a large number of reads are more likely to be novel functional miRNA. This assertion is partially confirmed by the sequence conservation of some of these highly expressed putative miRNA sequences in other legume genomes and some even in Arabidopsis. For example, 20 of the putative miRNA from the seedling dataset are present at 100% identity in the Arabidopsis genome. While individual characterization of each of these putative miRNAs is a time and labor intensive task, an estimation of the role these miRNAs might play in soybean transcriptional control was attempted by computationally identifying the targets of these miRNA genes.

3.2.5 Predicted targets of novel miRNA

Putative miRNA identification based on expression and structural properties is a potent tool for the discovery of novel miRNA. However, it faces the difficult challenge of controlling the false positive rate. An extra line of evidence that can support the prediction of a novel miRNA is the presence of complementary regions

in the transcribed regions of protein coding genes. Plant miRNA target identification is considerably easier than animal miRNAs since the dominant form of regulation discovered is post-transcriptional suppression involving the stable hybridization of the miRNA to the target mRNA. This process requires a sufficient degree of sequence similarity to be maintained between the miRNA and its target, allowing the identification of such targets through simple sequence alignment algorithms. The program psRNATarget [97] is designed to identify miRNA targets in EST databases based on target binding rules previously specified [98]. This tool was used to identify the potential targets of the novel miRNAs predicted.

The 20 most abundant miRNAs from each dataset were uploaded to psRNAtarget and processed against GMGI v15. With the exception of two miRNA sequences all novel miRNAs were predicted to target multiple genes in the soybean genome. A total of 235 individual ESTs from GMGI were identified as potential targets with a large majority predicted to be targeted for cleavage, as opposed to translational regulation. These ESTs in turn map to 292 gene models in the soybean genome. To summarize the results and detect patterns in the potential targets GO terms associated with these gene models were looked up and overrepresented Gene Ontology (GO) terms identified throught the AgriGO server [99]. The GO terms falling under the biological process branch were tested for this query set against the background of all genes in the genome. The significantly over-represented terms as determined by a T-test, assuming a hypergeometric distribution, and an FDR correction at *alpha* of 0.05 are shown in Figure 3.5.

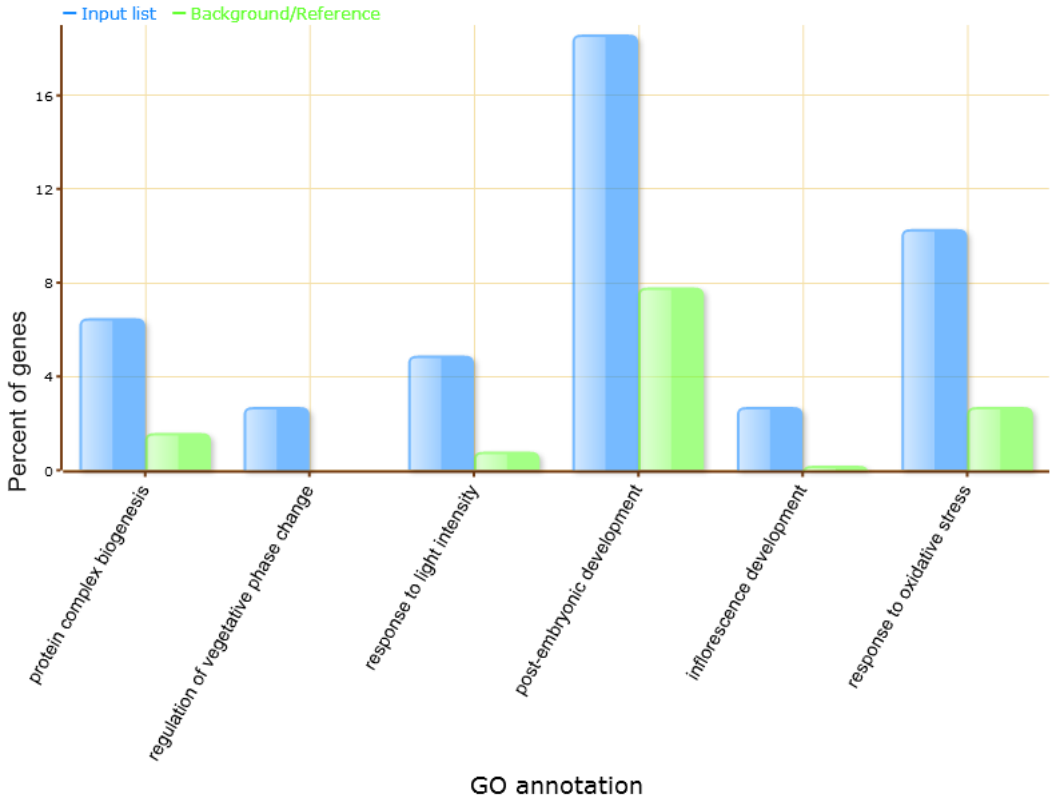

Figure 3.5: **Statistically significant GO terms among miRNA target mRNAs.** GO terms associated with the predicted targets were compared to the background reference to identify significantly over-represented terms using a hypergeometric test with a p-value cutoff of 1e-6. Significant GO terms from the biological process branch are shown.

3.3 Discussion

The small RNA population in soybean tissues is dominated by the heterochromatic siRNAs generated to suppress the activity of certain genomic regions. As was observed here, the most abundant elements in the genome are most likely the elements that have been actively transposing and increasing their copy number in the genome since older elements would likely have been fractured/deleted through recombination. Therefore they tend to be more actively suppressed by the siRNA mediated methylation pathway. Under conditions of stress such as biotic or abiotic challenges these elements might be activated and generate a great deal of genomic diversity. Therefore the knowledge of actively transposing elements as revealed by their copy number and the amount of siRNA suppression will enhance our understanding of the genomic regions that have undergone recent rearrangements or hold the potential for such changes.

Considering the diverse set of tissues and treatments from which the small RNA libraries were derived, the miRNAs predicted in this study are expected to regulate a wide range of processes in the plant. Fortuitously, the developmental stages and tissue types concerned in these experiment share certain similarities. A common thread across many of the libraries sequenced was the development of a young seedling and especially the development of meristem and generation of early plant organs. miRNAs are known to be heavily involved in plant development and transitions of phase [70]. Therefore, the observation of development associated GO terms in the predicted targets (Figure 3.5) provides a line of evidence that the predicted miRNAs are likely real and are involved in the regulation of developmental genes. The second dataset utilized in this study was studying light responses in seedlings and therefore the presence of "response to light intensity" and "response to red and far red light" GO terms further down the list is likely the result of these processes also involving one or more miRNA regulators. Finally the oxidative stress group of targets might represent genes involved in

the reaction to the start of photosynthesis in germinating seedlings or to light stress.

With the increasing amount of small RNA data available in the public domain, the role of small RNA in the maintenance of the genome and the regulation of transcription should become increasingly apparent. The efficient management of the retrotransposon load on the plant genome provides the plant a much higher genomic stability and appropriation of fewer resources to constantly suppress transposon activity. The implications of the saved energy in improving the yield of important crop plants like soybean are tremendous. Recent studies have focused on the role of small RNAs in the phenomenon of heterosis and concluded that global changes in 22-24nt small RNA activity leading to changes in methylation pattern are observed in hybrid plants showing increased phenotypic traits [100]. While the link between the two phenomena is currently tentative, the potential influence of the cost of genomic stability maintenance on the yield potential of a crop opens new avenues to explore in crop improvement. Discovery of novel miRNAs regulating the responses of plants to environmental stimuli will greatly increase our understanding of the processes underlying disease resistance and stress tolerances. Adjusting these responses, especially in certain target tissues, can potentially improve crop yields by allowing crop varieties to better resist environmental changes. Further, adjustments via small RNA mediation to nutritional quality and control of allergens or other harmful chemicals in the consumed parts of plants may offer exciting possibilities in the improvement of food quality and safety.

CHAPTER 4

RAPID GENOTYPING OF SOYBEAN CULTIVARS USING HIGH THROUGHPUT SEQUENCING

4.1 Introduction [1]

Soybean (*Glycine max*) lines grown in the US were originally introduced from East Asia. A wide range of cultivars are grown in east Asia and have been selected over centuries for yield and suitability to local environment. Earlier domestication of wild soybean involved selection for larger seed and improved nutritional quality. After introduction into the US, commercially grown cultivars were selected for improved yield and biotic/abiotic stress tolerance traits. Studies on diversity of soybean germplasm in the United States have suggested that the introduction and multitude of selection steps may have served as a genetic bottleneck and reduced the genetic diversity within the elite germplasm in the US [101]. An elite US cultivar called Williams 82 [102] was chosen for whole genome sequencing [18].

Breeding practices often involve introgression of desirable traits from a non-elite or wild variety into an elite line. The progeny from such crosses are generally backcrossed to the elite line to recover a near-isogenic line (NIL) with similar yield properties to the elite line and the added trait from the target locus. Molecular markers allow a breeder to rapidly screen a large number of lines for markers associated with the trait, allowing the selection of the molecular marker and thus specific introgression of a single genomic locus. Fine mapping the locus with molecular markers allows the amount of target DNA that will be integrated into the NIL to be reduced. This reduction in linkage drag can also allow reduction in the yield drag often associated with introgression. Therefore the availability of a large number of markers, spread more or less evenly over the genome of a specific exotic line targeted for introgression of a trait, is very valuable.

[1] Portions of this chapter have been submitted to PLoS One for publication. The authors retain the copyrights to the document. Author Contributions: Kranthi Varala performed all the bioinformatics analysis of data before and after the molecular experimental stage. Kankshita Swaminathan performed the extraction of DNA and sample preparation. Ying Li confirmed the predicted SNPs by amplification of genomic DNA fragments followed by sanger sequencing. Matthew E. Hudson conceived the study, design, co-ordination and manuscript. All authors contributed to the development of the manuscript.

Although fine mapping a locus with tightly linked markers is cost and labor intensive, it might need be done only once per allele of interest. Such an association will, potentially, be applicable in crosses between a different set of parents. The first generation molecular markers were restriction fragment length polymorphism (RFLP) [103], random amplified polymorphic DNA (RAPD) [104], amplified fragment length polymorphism [105] markers or microsatellite DNA markers [106]. Later, Simple Sequence Repeat (SSR) markers provided finer resolution and greater power for cultivar identification [107, 108, 109, 110]. More recently SNP markers [111, 112, 113] have grown in stature as an important tool in soybean breeding. Many genetic linkage maps using these marker sets individually or in combination have been constructed to assist breeding [114, 115, 116, 117, 118]. While SSR markers have a higher distinguishing power between lines, the distribution of SSRs is sparse in the genome and hence may limit the resolution offered in fine mapping. SNP markers, on the other hand have not only been shown to be highly effective in distinguishing lines in the soybean germplasm [113] but are also vastly more common in genomes, especially between two distantly related lines and are thus the predominant markers used in fine mapping. SNP density is expected to be particularly high in the non-coding regions of the genome [119]. One disadvantage of SNP markers is that they are usually available in the form of an array developed for specific genotypes, often not the genotype from which introgression is necessary in a given breeding project. Thus, fewer of the markers on the array are informative in the case of some introgression experiments. We explore the possibility of rapidly and cheaply developing SNP markers for any accession of soybean as referenced against the Williams 82 genome using major recent advances in short read sequencing technologies. These methods potentially allow rapid, low cost genotyping without the high initial costs of developing an array.

In this study we chose four cultivars of soybean (based on the presence of useful resistance traits) to explore a method for exploiting the stated advantage of short

read sequencing in generating SNP markers. Dowling is a low-yielding southern US accession of soybean that has proven useful as a source for the Rag1 allele, which provides soybean aphid resistance [119, 120]. Dwight is an elite soybean cultivar often used as the high yielding recurrent parent in breeding [121]. The Komata and PI594538A accessions carry the Rpp1 and Rpp1-b alleles that confer resistance to soybean rust [122] and are being used in attempts to integrate this trait into commercial lines. Williams 82 was also included to provide a base line for interpreting the experimental results, since it was the source of the DNA used for the reference genome sequence.

Sequencing technologies such as Illumina genome analyzer and ABI SOLiD offer the ideal combination of depth of coverage and frequency of sampling to generate a large set of SNP markers. The approach used in this work is to generate a large number of short reads from genomic DNA of these cultivars and align them to the reference genome. Numerous methods of SNP calling from sequence read data have been developed, each differing in the details of implementation and confidence measures used for calling SNPs. SNPs detected by high throughput sequencing of genomic DNA and called by these programs have the potential to provide very fine resolution of the genomic differences between the cultivar sequenced and the reference assembly, here the Glyma1 assembly [18] of the Williams 82 genome. Owing to reduced selection pressure outside protein-coding regions, a large number of variant SNPs can be expected between the genomes of the cultivars of interest if intergenic sequences are included in the analysis.

Assuming a minimum requirement of three reads covering the base in question to call a SNP and given the size of the *G. max* genome, estimated to be n = 1.1 gigabases (Gb) [35], a random whole genome shotgun sequencing effort will have to be produce at least 3.3 Gb of raw sequence per cultivar to provide sufficient confidence in most SNP calls to produce a high density map. While the cost of producing such large amounts of sequence data has steadily decreased over time, it is nonetheless a substantial investment. Furthermore, sequencing a randomly

sheared genomic DNA library in a complicated eukaryotic genome such as *G. max* is estimated to produce a large proportion of reads from the repetitive fraction of the genome. The proportion of reads sampled from a repeated region is directly proportional to the fraction of the genome representing repeats. Up to 60% of the *G. max* genome is estimated to be composed of moderately to highly repetitive elements [18, 9]. The correct alignment of reads sampled from a repeat region is ambiguous by definition due to the presence of many repeating units from a single repeat family. Therefore for the purpose of identifying reliable SNP markers it is desirable to reduce the representation of repeat elements in the sequencing library. It is possible to exclude many repeats by sequencing mRNA in the form of Expressed Sequence Tags and to mine these for SNPs, and this has been done previously in soybean [116]. However, a higher rate of mutation in non-genic sequences, (including both repetitive and non-repetitive elements) is expected compared to the protein-coding regions of the genome, which are more functionally conserved. Thus, the ideal SNP discovery method would exclude repeats while preferentially targeting non-protein-coding DNA. In an earlier survey sequencing effort we characterized the repeat content and composition of the *G. max* genome (Williams 82) [9]. The study also identified the SB92 repeat family as being a predominant repeat that represents close to 3% of the soybean genome. We hypothesized that a method devised to target non-repetitive sequences on the basis of this information would reduce the representation of the repeat content in the sequencing library. During the genome fragmentation stage of library preparation, directed cleavage of DNA by Type II restriction enzymes, as opposed to random shearing, offers an effective way to anchor the start of a short read preferentially to certain sites. Such complexity reduction methods have been successfully applied to alter the genome sampling frequency in multiple organisms [123]. More recent improvements in sequence yield and multiplexing protocols allow a vastly more intricate design to develop high density linkage maps at a population level [124]. We thus deployed a method that targets deep sequencing at Type II restriction enzyme recognition sites, a procedure that has recently been used by others in soybean [125]. In the

study presented here, the enzyme choice was determined on the basis of numerical analysis of a prior repeat survey [9] in order to reduce the likelihood of cleavage within repeats and maximize the number of useful sequence reads. We attempt to develop a general strategy to sequence a reduced-representation library (RRL) from soybean genomic DNA and use this method to identify SNPs polymorphic between the reference genome and any soybean line of interest.

4.2 Results

4.2.1 Restriction enzyme choice

Choosing an enzyme that does not cleave in any highly repetitive region (as identified previously [9]) for library preparation causes a larger proportion of reads to begin in low-copy regions. While it is highly unlikely to find an enzyme that does not cleave in any repeated sequence in any given genome, knowledge of repeat composition allows selection of an enzyme that substantially increases the representation of non-repetitive regions in the library.

The choice of restriction enzyme to digest genomic DNA was made based on the following criteria: 1. The enzyme should cut often enough in the genome to sample sites at a fairly small physical interval. 2. The recognition site should be present more often in the non-repetitive regions of the genome than the repetitive and 3. The enzyme recognition site should be absent in the extremely abundant 92 bp peri-centromeric repeats CentGm-1 and CentGm-2 [126]. To satisfy the first condition we limited the candidate enzymes to those whose recognition sequence is four to six base pairs. To identify the relative frequency of recognition sites in highly repetitive regions compared to less repetitive regions, it is imperative to identify the high ly repetitive fraction of the soybean genome. A whole genome survey sequencing [9], done at 0.7X coverage, of the 1.1 Gb soybean genome

was used to identify the highly repetitive component. The Lander-Waterman model[4], originally developed to describe a non-repetitive genome, when applied with these parameters of genome size and coverage predicts that it is unlikely to sample any region repeatedly, and this likelihood decreases exponentially with the depth of coverage. Therefore, any overlapping reads seen in such a survey are expected to come from repetitive regions. Based on the model predictions, it was estimated that any sequence covered by three reads or more is expected with high confidence to occur in multiple locations in the genome. Therefore all contigs from the non-cognate assembly [9] containing three reads or more were classified as repeats. Any read with no detectable overlap with another is assumed to have been derived from unique or low copy number regions of the genome. Restriction enzyme site frequencies were computed independently in the repetitive and low copy number sets of sequences. Each enzyme was then scored on the relative frequency of its recognition site in the non-repetitive set compared to the repetitive one. The enzyme *Mse*I was selected, as it showed the highest bias towards low copy regions while still matching the other two criteria mentioned above. *Mse*I is a type II restriction enzyme with a four base recognition site TTAA and cleaves after the first base, leaving a 5 TAA overhang. As such, the sites for this enzyme are extremely common in the genome, (around every 100 bp on average). In order to reduce the number of total sites sequenced, nuclear DNA from each accession was digested with *Mse*I and the 100-150 bp fraction from each digestion was sequenced on a single lane of the Illumina genome analyzer. This size-fractionation step causes only those *Mse*I sites located within 100-150 bp of another *Mse*I site to be targeted for sequencing, around 10% of the total number of sites. *Mse*I sites within 100-150 bp of each other occur on average every 1032 bp in the Williams 82 reference genome sequence.

The number of bases covered by at least one read in each of the lines is as follows: Dowling: 55730683, Dwight: 57862248, Komata: 93217361, PI594538A: 96473008 and Williams 82: 57661682. The GC composition of reads mapping to

Glyma1 ranged from 33 to 39%, which is comparable to the 36.8% GC composition of Glyma1. This implies that the restriction strategy did not introduce a bias towards or against the GC richer or poorer regions of the genome.

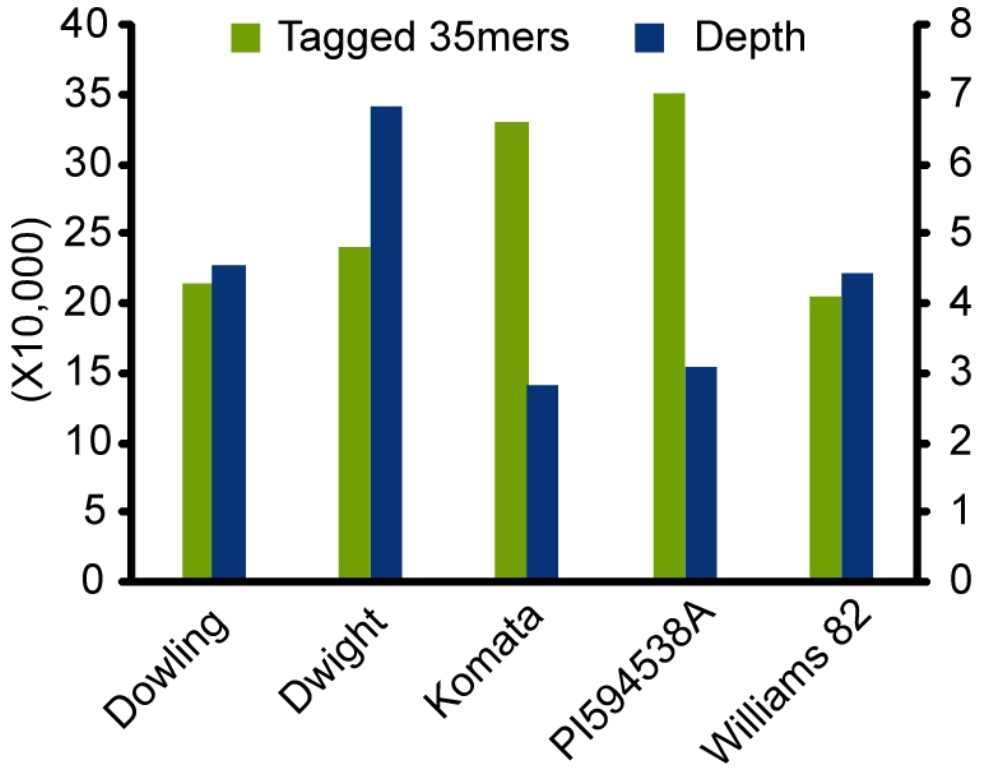

Figure 4.1: **Sequence coverage at tagged sites across varieties subjected to genotyping by sequencing.** Average depth of coverage across all bases covered by resequencing reads is shown by blue columns with the value indicated by the vertical axis on the right. The variation in depth observed between libraries is a combination of variation in the amount of reads obtained in a sequencing run and the number of loci tagged by a read in that library. The total number of 35mers from the genome that were tagged by at least one read is shown by the green columns. The vertical axis on the left depicts the number of sites tagged (in tens of thousands).

Table 4.1: **Efficiency of sampling strategy.** Reads were mapped to the Glyma1 reference assembly at at least 90% identity. Tagged 35mers represents the number of unique 35 bp frames in the genome that are covered by at least 1 read. Low copy reads are those that map to less than 5 loci in the genome.

Variety	Dowling	Dwight	Komata	PI594538A	Williams 82
Total Reads	7625036	11764203	8452272	9309921	7583486
% of reads mapped	95	96	88	91	97
Tagged 35mers	2148006	2410953	3310171	3506523	2053175
low copy %	57	43	69	72	55

4.2.2 SNP discovery

Sequencing yielded 35 base reads from each of the libraries. The number of reads sequenced from each library varied appreciably, as expected, and hence the amount of sequence coverage is unequal across the libraries (Table 4.1). Efficiency of restriction site anchoring was tested by counting the percentage of reads that begin with the expected TAA overhang from the restriction digestion (Table 4.1). Frequency of the trinucleotide TAA in the soybean genome assembly is 53,859,048 i.e., approximately 16.92% of all trinucleotides in the genome. Given this background, if the genome were to be sheared randomly and sequenced, about 17% of the reads are expected to start with the bases TAA. 81-94% of the reads in each of our libraries began with the trinucleotide TAA. Hence a very significant over representation of reads starting with TAA was obtained, thus confirming that the DNA library built from the restriction digested DNA was heavily biased towards true *Mse*I fragments. All reads were aligned to the Glyma1 assembly using the m.a.q. alignment program [95]. The percentage of reads, from each library, which align successfully to the Glyma1 assembly either uniquely and/or in multiple locations are listed in Table 4.1. The anchoring strategy restricted the sampling sufficiently to increase depth from approximately 0.25X coverage expected from a random shotgun sampling to approximately 4X, thus allowing greater confidence in calling SNPs. The number of SNPs identified from each accession is listed in

Table 4.2. The list of high confidence SNPs described here was generated from a larger set of SNP calls generated by m.a.q. by applying a high stringency filter to increase confidence in the SNP call and reduce false positives. SNPs were filtered to only include calls that were very high confidence: covered by at least 3 reads, minimum consensus quality of 20 for the polymorphic base and two bases on either side, no indels within 6 bp and no more than 2 SNPs in a 10 bp window. This set of parameters ensures that no SNP calls are based exclusively on repetitively mapped or poorly aligned reads. The number of SNPs discovered correlates with the expected genetic distances of these accessions, since Dwight has a known higher coefficient of parentage with Williams 82, and Williams 82 is the line from which the reference sequence was generated. Dwight and Williams 82 share a common parent [102, 121]. Dwight is expected to share at least a quarter of its genome with Williams (Fig. 4.2). This common parentage explains the lower diversity between Williams 82 and Dwight.

Table 4.2: **High confidence SNPs and SNP density.** The SNP density between each line and the reference assembly is measured as the total number of good quality bases resequenced in that line divided by the number of high confidence SNPs. This measure is called Mean Distance Between SNPs (MDBS). Mean Depth At SNP (MDAS) assesses confidence measured as number of reads aligned at the SNP position.

Variety	Dowling	Dwight	Komata	PI594538A	Williams 82
Mean Coverage	4.56	6.85	2.81	3.07	4.45
Filtered SNPs	6019	4294	12727	14550	1122
MDBS	626	904	609	649	4168
MDAS	17.1	17.33	6.59	8.26	12.89

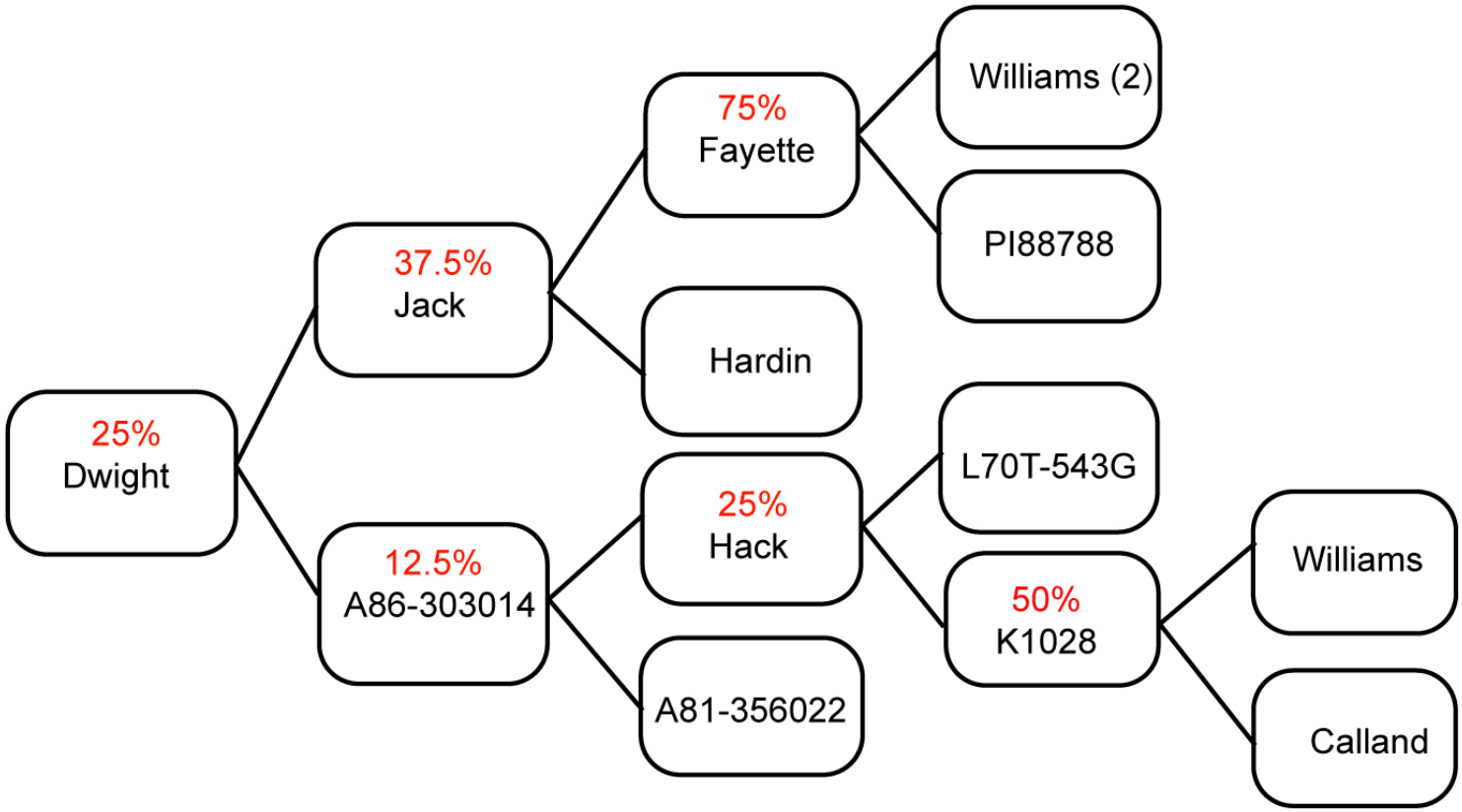

Figure 4.2: **Pedigree of Dwight.** The Dwight variety of soybean was produced by crossing Jack and an experimental line. Following the parentage back 3 generations reveals that the Williams line served as an ancestor on both sides of the cross and was used as recurrent parent to varying degrees. Numbers in parenthesis indicate the number of times a line was used as a recurrent parent. Based on the parentage, the proportion of the genome that is expected to be from Williams is indicated in Red. Lineage is depicted with the progenitors to the right.

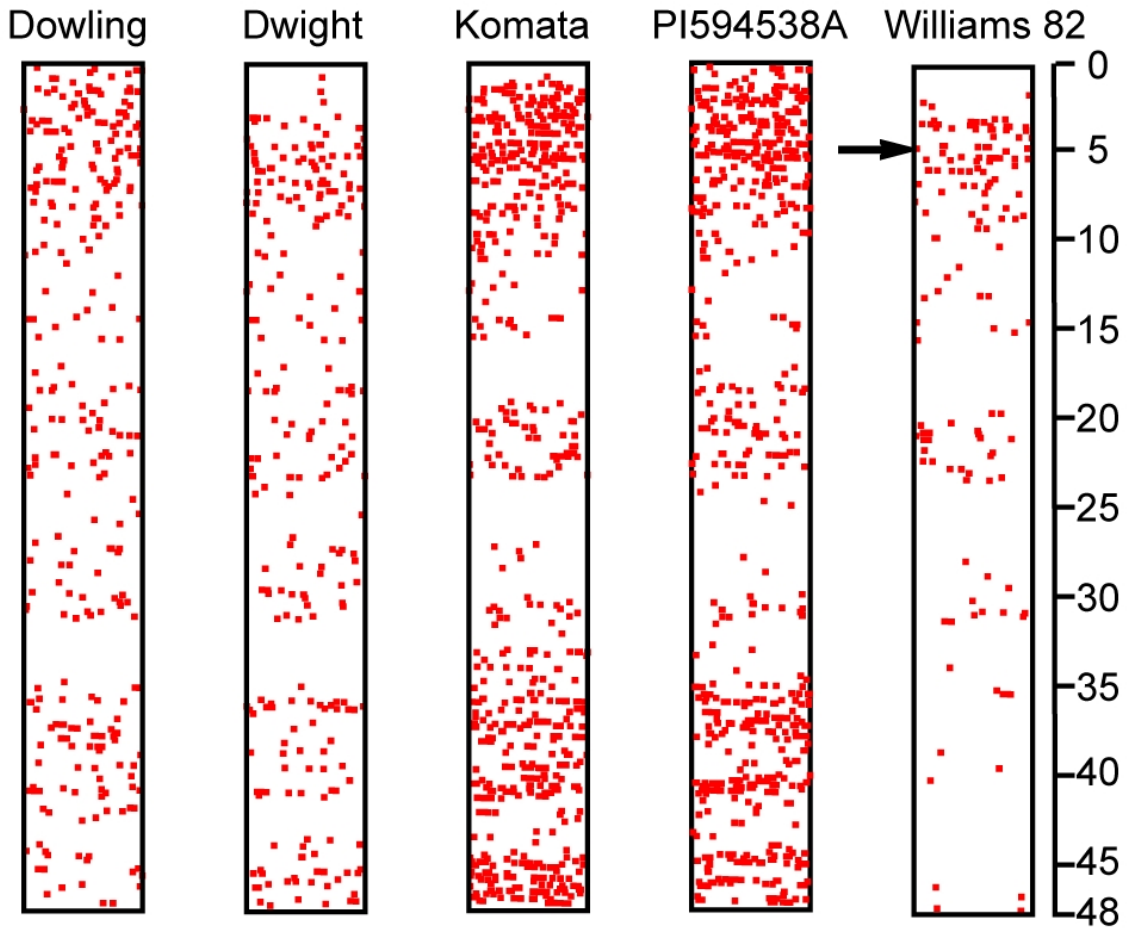

Figure 4.3: **SNPs polymorphic between each variety sequenced and the Glyma1 assembly of chromosome 3.** High confidence SNPs on Gm03 (Linkage group N) are shown. SNPs occurring in a one million basepair (MB) bin are dithered, left-to-right, along the X axis based on the position of SNP within that bin. Y-axis represents the length of the chromosome. The presence of a large number of SNPs and their non-random distribution on Gm03 for the Williams 82 data suggests that the Williams 82 lines carry significant portions of the non-recurrent parent Kingwa. The distribution of SNPs in other lines shows a density proportional to sampling frequency and shared parentage, while the Williams 82 line shows higher diversity around the 5MB mark of Gm03 (arrow). Note that repetitive sequences (predominant in the centromere) prevent unique mapping of sequence reads and thus show substantially reduced SNP density. Scale is in megabases (Mb) of physical distance.

To express the degree of diversity between lines we used Mean Distance Between SNPs (MDBS), the average distance between polymorphisms within regions sequenced in this experiment. This was calculated by computing the number of bases covered by at least three reads and dividing that by the number of SNPs, called at high confidence relative to the reference genome, without accounting for the base quality. Since the average quality for bases is comparable between the 5 libraries (data not shown) we estimate that any local biases in mapping quality will even out over the genome. After normalization the MDBS is almost even between the Dowling, Komata and PI594538A lines but shows a marked increase in Dwight and is very large in Williams 82, as expected (Table 4.2). Even though the raw count of SNPs is higher in the two East-Asian accessions (Komata and PI594538A as compared to Dowling) the number of sites sampled in them is significantly higher (Figure 4.1) thus providing an important correction for the SNP density estimation.

To test the predicted high-confidence SNPs, we independently sequenced the SNP loci using traditional Sanger sequencing. Twenty SNP loci, with good quality flanking sequence, were identified between the Komata and Williams 82 accessions and chosen for confirmation. Sequence around this region was extracted from the Glyma1 assembly, and used for primer design. Of these twenty regions, two failed to amplify with the designed primers. The eighteen other primer sets amplified a single region, as evidenced by a single band on a gel. Using Sanger sequencing, we confirmed the predicted SNP in these eighteen loci. In two cases additional SNPs in the vicinity that had not passed the high confidence SNP filter were also confirmed, implying that the SNP density estimate we arrived at is likely to be a conservative estimate of the true variation between these lines, and that our false SNP discovery rate is less than 5%.

4.2.3 Heterogeneity in Williams 82

Williams 82 was created by crossing Williams and Kingwa accessions of soybean [102]. The expected proportion of Kingwa genomic DNA in the Williams 82 genome is 1.75% based on the pedigree information, with most of it expected to be centered around the Rps1 locus on Gm03 [127, 17]. We sequenced Williams 82 to serve as a negative control for the experiment in anticipation of having to adjust SNP calling parameters for the inherent biases/errors in Illumina short read sequencing. Ideally no SNPs should exist between the Williams 82 sequencing run and the reference genome at the correct level of stringency required to remove false positives, since our genomic DNA was derived directly from the reference allele acquired from the USDA Soybean Germplasm collection. Despite repeatedly increasing the stringency level we continued to observe SNPs between the resequenced Williams 82 library and the reference Glyma1 assembly (Figure 4.4). Additionally we observed a 100% confirmation rate by Sanger sequencing for the SNPs identified between the Komata reads and the reference assembly at the default stringency level. Thus the SNPs detected between our Williams 82 reads and the Glyma1 reference assembly are highly unlikely to be the result of errors either in our sequence or in the Glyma1 sequence.

In parallel with our study, Haun et al.[94] found that the Williams 82 cultivar contains a significant level of residual genetic variation. We found that many SNPs are detected between the Williams 82 used for this study (acquired from the Soybean Germplasm Collection at the University of Illinois) and the reference genome sequence [18]. This is likely a result of genetic variation between the plants used for sequencing of the reference and the Germplasm Collection line used in this study.

4.2.4 SNP distribution

The SNP positions for each accession were plotted on the twenty soybean chromosomes to visualize their distribution (Figure 4.4). Large gaps in the distribution coincide with the highly repetitive centromeric and pericentromeric regions of the chromosomes [18]. Among the four sequenced accessions (excluding the Williams 82 control) SNP distribution is fairly even in the low copy regions, implying a lack of bias in sampling (Figure 4.4). The observed even distribution of SNPs discovered using this method in conjunction with the consistent mean SNP density of 600 base pairs, within sequenced regions, among the accessions makes this method of SNP discovery an excellent tool for marker development in soybean. The SNP density obtained in this study for mapping purposes, measured as the median distance on the chromosome between any two SNPs discovered using this method, is 46.6 Kb for Dowling, 44.4 Kb for Dwight, 19Kb for Komata and 16 Kb for PI594538A. All of these distances are short enough to be used for extremely fine genetic mapping. In contrast to the data for the other accessions, the SNP distribution plot for Williams 82 SNPs relative to the reference Williams 82 assembly clearly shows regions of high and low diversity (Figure 4.4) between the genotype that was sequenced and the reference. Interestingly the Rps1 gene cluster maps to Gm03 at approximately position 5,000,000 in the Glyma1 assembly [17], a region showing the highest SNP density among all chromosomes between the reference sequence and the genotype sequenced in this study (Figure 4.3). The same region was identified as a location with high variation between individuals of Williams 82 in another study [94]. In addition high SNP density regions are seen in Gm07, Gm14 and Gm15 (Figure 4.4). These regions likely correspond with the portions of the Kingwa genome retained in Williams 82 during back-crossing and subsequent selection.

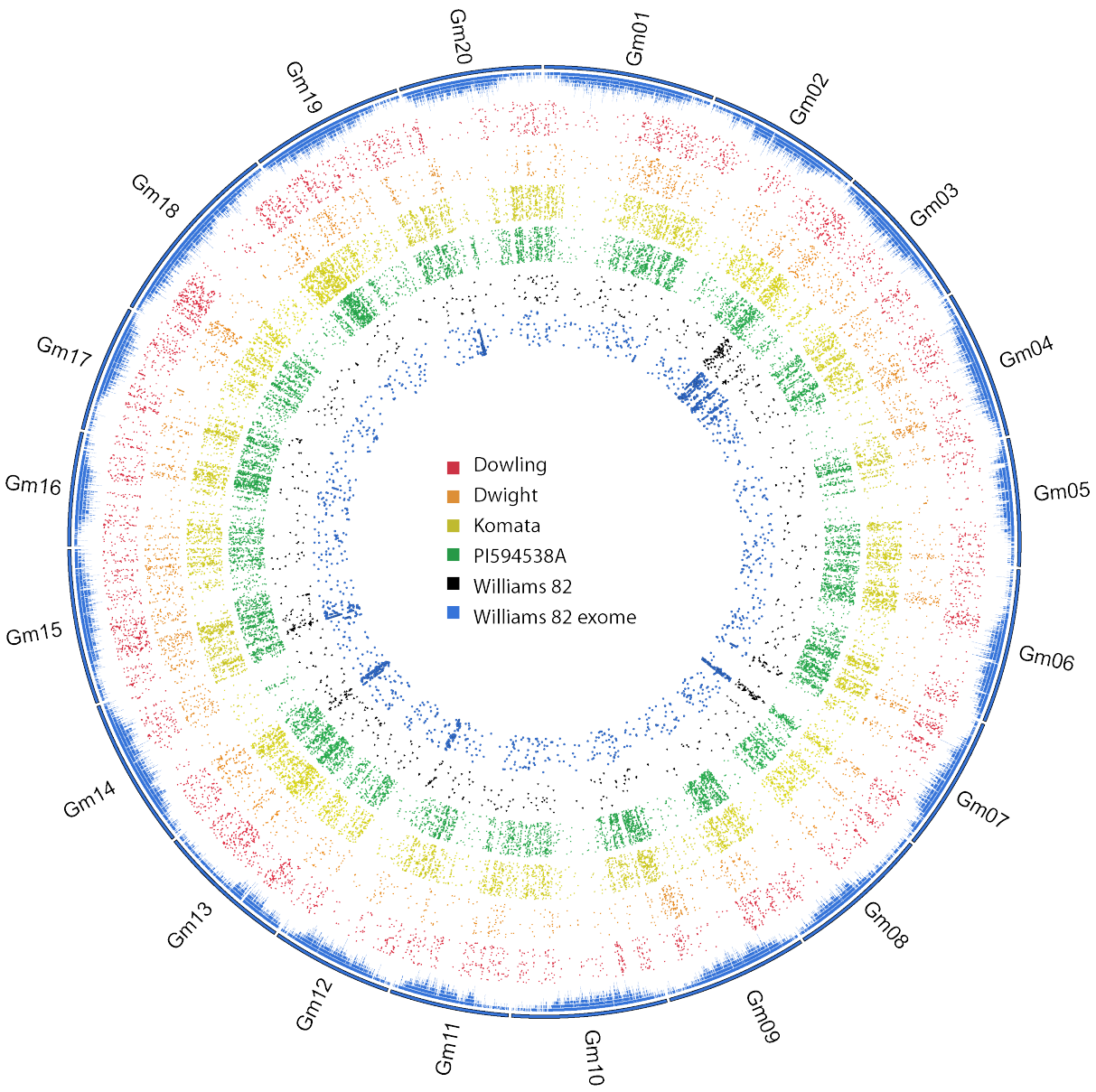

Figure 4.4: **SNP positions on the soybean genome.** High confidence SNPs called by aligning reads against the reference assembly are shown. Outer ring shows the density of repeat elements locally. SNPs from each accession sequenced are plotted in concentric rings in the order: Dowling, Dwight, Komata, PI594538A, Williams82 from outside to the center. The innermost ring depicts SNPs discovered from exome capture sequencing of two Williams 82 individuals [94]. Genomic regions rich in repeats have poor read alignments, hence lacking in SNP predictions. Both Williams 82 datasets show regions of high heterogeneity implying areas of linkage drag from loci selected for in the parental Williams x Kingwa cross.

4.2.5 Transposable element families

A number of reads from each library mapped to known transposable elements (TEs) in soybean. Reads were assigned to TE families by mapping them to the elements listed at soyTEdb [93]. The number of reads matching TE families was normalized to the total number of reads mapping to the reference genome generated from each accession. Since some reads matched to many members of a TE family, the contribution of a read matching multiple TEs was divided by the number of elements the read mapped to. Weighted read counts were summed up for each family of TEs based on the family-level annotation from soyTEdb. Interestingly, the Gypsy family of elements shows higher numbers in the Williams 82 and Dwight genomes relative to the other accessions sequenced (Supp. Fig. 2). The other noticeable expansion is in the Copia and CACTA families in the Dowling genome. These results indicate evidence for variability in TE content between soybean accessions.

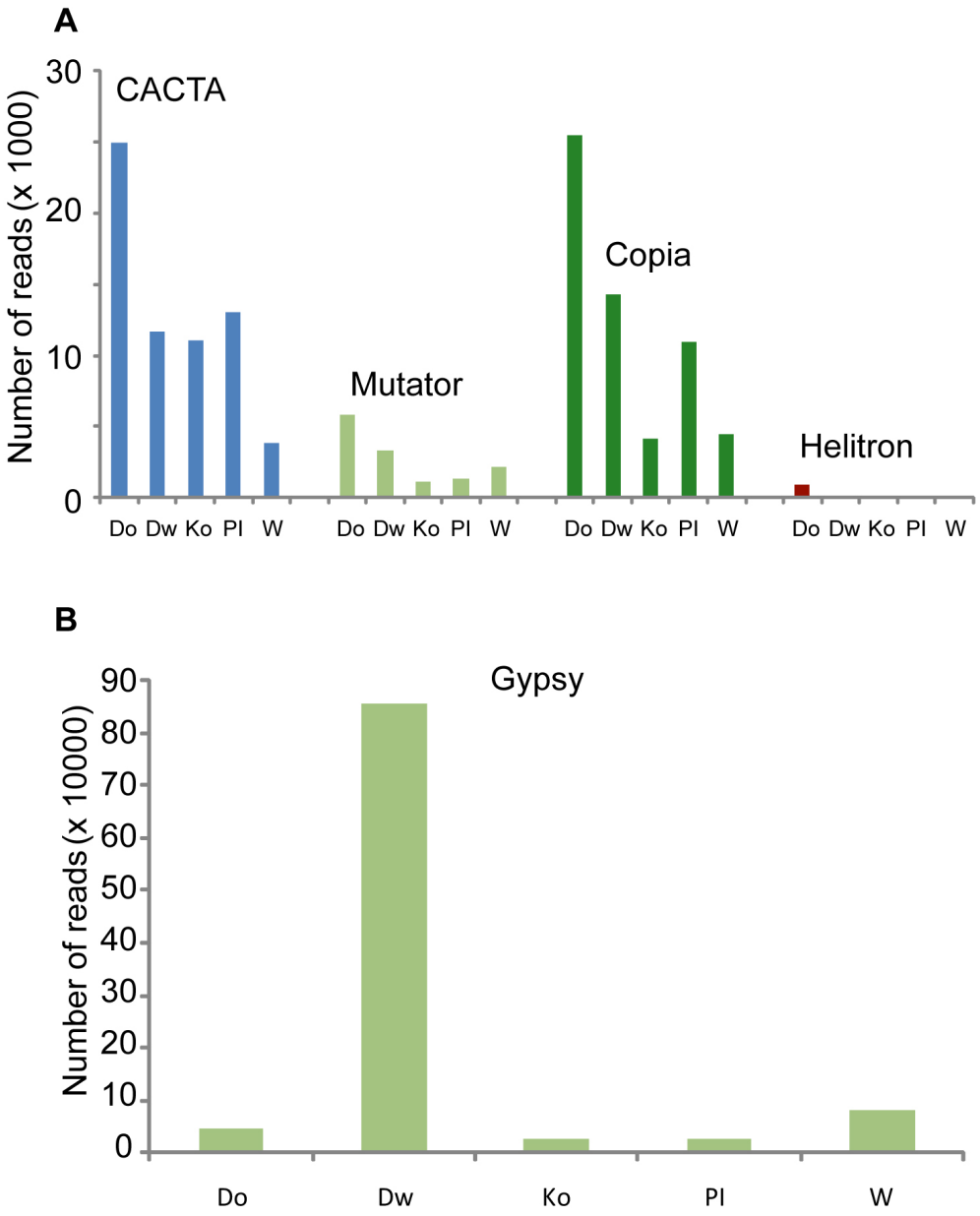

Figure 4.5: **Reads aligning to each transposon family.** All reads were aligned to the soybean transposable element database (soyTEdb) and grouped based on the transposon family they match. The number of reads assigned to each family was normalized to the total number of reads from that library, to allow comparison across lines. Abbreviations for soybean genotypes: Do = Dowling, Dw = Dwight, Ko = Komata, PI = PI594538A, W = Williams 82. A) CACTA and Copia families show a significant expansion in the Dowling accession. B) Elements of the Gypsy family have substantially increased numbers in the Dwight genome and are increased to a lesser extent in the Williams 82 genome.

4.3 Materials and methods

4.3.1 Restriction enzyme selection

Survey sequencing data [9] was used to classify soybean sequences into a low copy set and a highly repetitive set. The Lander-Waterman model [4] for a non-repetitive genome predicts the number of contigs expected to be formed by an overlap of n reads for a given size of the genome and coverage obtained. Fitting this model to the size of soybean genome and coverage obtained in that study predicted that there should be almost no contigs formed by overlap of 5 reads or more by random chance. The non-congate assembly showed contigs far exceeding the number of expected contigs beginning at n = 3. Therefore all contigs formed by an overlap of 3 reads or more were defined as sequences with a high copy number for the purposes of this study. This set was composed of 20,670 contigs and all reads in these contigs were extracted to form the repeat sequence set (n=384,339). Conversely all single reads with no detectable overlap with any other read from the survey sequencing were classified as low copy sequences. Approximately 333,000 reads fell in to the low copy set. The (fortuitously comparable) number of reads in each set removed the need for any normalization of site frequencies. Type II restriction enzymes with a recognition site length of four or six were selected from REBASE [128] and grouped by site. To ease data analysis at later stages all recognition sites with ambiguous bases were removed. In addition all enzymes that do not have a defined sequence at the cut site were removed. To avoid sampling the extremely abundant 92 bp repeat, all enzymes that would cut this repeat were also disqualified. Frequency of each site was then computed in the two sets of sequences. The remaining enzymes were ranked based on the ratio of site frequency in the low copy read set versus repeats. *Mse*I, a type II restriction enzyme with recognition site TTAA, emerged as the best enzyme on this list. The raw site count in the low copy set was 489,740 versus 397,173 in the repetitive

set giving a frequency ratio of 1.23 in favor of low copy sequences. *Mse*I cuts the recognition site TTAA leaving a 5 overhang of TAA.

4.3.2 Plant material

Seed for each soybean line described in the text (Dowling (PI 548663), Dwight (PI 597386), Komata (PI200492), PI594538A and Williams 82 (PI 518671)) was obtained from the USDA soybean germplasm collection. Plants were grown in pots in a temperature- and light-controlled greenhouse in long day (18hr) light conditions for 4-12 weeks. Young leaf and stem tissue, tips of branches with at least two visible leaves, was collected from four to six individuals for each line.

4.3.3 DNA extraction and digestion

10-20 μg of Nuclear DNA, extracted from all five lines according to protocols described in Swaminathan et al. [9] was subjected to complete digestion with *Mse*I. The digest was end-repaired with T4 DNA polymerase run on a 3% low melting point agarose gel. The size fraction from 100-150 bp was electroeluted using Spectrapore dialysis tubing (MW cutoff 3500) and precipitated. 200-500 ng was sequenced by Illumina (Hayward, CA).

4.3.4 DNA sequencing

After library construction for Illumina sequencing, each library was loaded onto one lane of the sequencing flow cell. Sequencing was done on the Illumina GAI genome analyzer system, performed for 35 cycles and bases called. Each library was sequenced twice, except Dwight (sequenced three times), to satisfy quality criteria. Sequence and quality data was obtained in fastq format.

4.3.5 Mapping to Glyma1

Reads were aligned to the Glyma1 version of the soybean genome assembly. Maq (v 0.7.1) was used to align the reads to the genome using the maq.pl script with easyrun option.

4.3.6 SNP calling

SNP calling was done as part of the easyrun option for maq. Additional stringency levels were tested by running maq.pl with the SNPfilter option and varying the cut off parameters for minimum mapping quality at SNP position and in a 6 bp window around it. These changes did not change the number of SNP calls appreciably. Increasing the minimum depth required to call a SNP decreased the number of SNP calls substantially, especially in the Komata and PI594538A lines. The SNPs reported in this report were obtained at the default maq parameters of d=3, -n=20, -q=20, -w=5, and N=2.

4.3.7 SNP verification

Primers for SNP verification were designed using an in-house script. Komata genomic DNA extraction was performed as described earlier. PCR was performed with Ex Taq (Takara, Japan) according to the product menu. The reactions were first heated at 94C for 5 min, followed by 35 cycles using a 30 seconds denaturation step at 94C, a 30 seconds annealing step at 58C, and a 1-min extension step at 72C. An additional 7-min extension step at 72C was added after 35 cycles. PCR products were analyzed on a 1% agarose gel. In all cases that the PCR product showed as a single band in the gel, the Qiagen PCR purification kit (Qiagen, CA, US) was used to purify the PCR product for sequencing. Sanger sequencing reaction of the PCR product was performed with BigDye Terminator v3.1 Cycle

Sequencing Kit (Applied biosystems, CA, USA) according to the manual with the forward primer used in PCR reaction. BigDye reactions were submitted to Keck Center, UIUC for purification and capillary electrophoresis. Sequences were analyzed and compared to Glyma1 genome sequence using Sequencher (Gene codes, MI, US).

4.4 Discussion

Deep sequencing of reduced representation libraries from genomic DNA provides a rapid and relatively inexpensive method of generating markers in lines of agronomic interest. Restriction digestion of genomic DNA offers an excellent way of creating reduced representation libraries. Typically, the restriction enzyme used for digestion is chosen using a general strategy. One such strategy is to use a methylation sensitive enzyme [129]. This approach preferentially targets single-copy sequences, but also targets conserved protein-coding sequences where SNPs are rarer (a disadvantage for less diverse crops such as soybean). It also requires complex procedures to reduce the size of the restriction fragments to a suitable size for Illumina sequencing [13]. An alternative approach is to empirically pick one, or cocktail of few enzymes that give the desired result based on experimental digestion of genomic DNA. Both strategies have been employed with success in plant genomes [130, 131]. In species where the approximate genomic repeat composition is known, it is possible to apply a more rational strategy of choosing a restriction enzyme that provides both an increased depth of sampling and an intentional bias towards the non-repetitive regions of the genome.

In the case of *Glycine max* the enzyme *Mse*I works exceptionally well at reducing the genome complexity sufficiently to allow SNP discovery, while also preferentially targeting intergenic DNA as a result of its lower GC content. A recent study used CviRI digested DNA, based on an in silico analysis of the draft genome and annotated repeat elements, to identify SNPs between Forrest and Williams 82

[125]. Our method of identifying site frequency in the low and high copy regions indicates that CviRI has a frequency ratio of 0.88:1 in low:high copy genomic DNA, which compares unfavorably to the 1.4:1 ratio for *Mse*I. We believe this was the result of our using the mathematically defined repeats by non-cognate assembly of a genome survey [9] rather than using annotated repeats from a genome sequencing project. Mathematically defining repeats is likely to identify more repeats than sequence annotation due to its power to overcome the need to detect sequence similarity across species, and the fact that tandem repeats are often excluded from genome assemblies. A similar strategy can in principle be followed for genotyping multiple accessions of any other crop species with a reference genome and known repeat composition. Even in unsequenced genomes where a survey sequence is available to detect repeat sequences, this method can still be applied, since the length of the restriction site roughly determines the mean distance between such sites in the genome.

With the falling costs of short read sequencing and coupled increases in the number of sequencing reads produced per run, such a deep sequencing strategy is likely to be the most rapid and, perhaps, even the more economical method to generate a large amount of SNP markers for any new accession of interest to plant breeders. A single lane of Illumina sequencing, at the time of this manuscript being written, will likely tag the vast majority of *Mse*I sites in the genome with sufficient depth to allow high confidence SNP detection and provide a very high density SNP map for several accessions using barcoded libraries, yet would still likely be insufficient for full whole-genome resequencing of a single line. In such an experiment the cost per accession is expected to fall further. At the estimated SNP density of 600 bp such genotyping should allow fine mapping a trait of interest down to a very small interval. In specific regions of interest, where a higher density of SNPs is needed, the SNP filter stringency can be lowered accordingly at the cost of increasing the false positive rate (which, as we indicate here, is very low for the procedure as described). Selection based on such markers will facilitate high

throughput genotyping of progeny to select for traits of interest. High resolution mapping will also allow reduction of the yield drag often introduced in such crosses by allowing genotyping to select the progeny with least amount of DNA from the lower yielding parent.

There was a noticeable increase in the number of sites sampled in two libraries out of the five. This difference can most likely be ascribed to small differences in the efficiency of digestion between the different DNA samples or in the fraction of genome obtained during the size selection from gel. While this variation suggests that great care must be taken during those steps, the result still provided sufficient sampling from all libraries to enable SNP calling.

Our survey also detected significant residual variation between different sources of the cultivar Williams 82 (the source used for the reference genome sequencing and the soybean germplasm collection). This is a violation of a perhaps unrealistic assumption made historically by many crop biologists, that varieties of an inbred selfing crop such as soybean should be almost completely homogeneous and homozygous. Our results on variation within Williams 82 confirm those of Haun et al.[94] who found that both SNP and copy number variants exist within this line using different techniques to those used in this study. Our Williams 82 DNA was prepared from pools of several individuals, whereas that of Haun et al. was prepared from individual plants. Therefore while we are not able to determine the extent of variation between the individual plants derived from seed from the germplasm collection, we have determined in the case of many of our polymorphisms that the reference genome contains an allele not found in any of the individuals in our pool.

We conclude that individual plant lines must be closely examined using this or a similar genotyping technique to ensure that within-cultivar variation does not cause errors in experiments involving genetic comparison (particularly those that involve the creation of variation using chemical mutagens, such as TILLinG). For important crop species with large germplasm collections or novel plants being

introduced into agriculture this method provides a rapid, economical and easy protocol to catalog diversity and generate molecular markers. The advent of longer sequence reads makes this technology also potentially applicable to organisms with unsequenced genomes.

REFERENCES

[1] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase," *Journal of Molecular Biology*, vol. 94, no. 3, pp. 441–446, May 1975. [Online]. Available: http://www.sciencedirect.com/science/article/B6WK7-4DN8Y1V-HJ/2/6f9da4af0ad704416b22ae9b90af8194

[2] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5463 –5467, Dec. 1977. [Online]. Available: http://www.pnas.org/content/74/12/5463.abstract

[3] R. Fleishmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J.-F. Tomb, B. Dougherty, J. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. Gocyne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J. Kelley, J. Weidman, C. Phillips, T. Spriggs, E. Hedblom, M. Cotton, T. Utterback, M. Hanna, D. Nguyen, D. Saudek, R. Brandon, L. Fine, J. Fritchman, J. Fuhrmann, N. Geoghagen, C. Gnehm, L. McDonald, K. Small, C. Fraser, H. Smith, and J. Venter, "Whole-genome random sequencing and assembly of haemophilus influenzae rd," *Science*, vol. 269, pp. 496–512, 1995.

[4] E. Lander and M. Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics*, vol. 2, pp. 231–239, 1988.

[5] F. E. Angly, B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer, "The marine viromes of four oceanic regions," *PLoS Biol*, vol. 4, no. 11, p. e368, 11 2006.

[6] A. L. Toth, K. Varala, T. C. Newman, F. E. Miguez, S. K. Hutchison, D. A. Willoughby, J. F. Simons, M. Egholm, J. H. Hunt, M. E. Hudson, and G. E. Robinson, "Wasp gene expression supports an evolutionary link between maternal behavior and eusociality," *Science*, vol. 318, no. 5849, pp. 441 –444, Oct. 2007.

[7] D. Schwarz, H. Robertson, J. Feder, K. Varala, M. Hudson, G. Ragland, D. Hahn, and S. Berlocher, "Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot rhagoletis pomonella," *BMC Genomics*, vol. 10, no. 1, p. 633, 2009.

[8] S. H. Woodard, B. J. Fischmann, A. Venkat, M. E. Hudson, K. Varala, S. A. Cameron, A. G. Clark, and G. E. Robinson, "Genes involved in convergent evolution of eusociality in bees," *Proceedings of the National Academy of Sciences*, 2011.

[9] K. Swaminathan, K. Varala, and M. Hudson, "Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey," *BMC Genomics*, vol. 8, no. 132, 2007. [Online]. Available: http://www.biomedcentral.com/1471-2164/8/132

[10] S. D. Findley, S. Cannon, K. Varala, J. Du, J. Ma, M. E. Hudson, J. A. Birchler, and G. Stacey, "A fluorescence in situ hybridization system for karyotyping soybean," *Genetics*, vol. 185, no. 3, pp. 727–744, July 2010. [Online]. Available: http://www.genetics.org/cgi/content/abstract/185/3/727

[11] J. H. Tuteja, G. Zabala, K. Varala, M. Hudson, and L. O. Vodkin, "Endogenous, tissue-specific short interfering rnas silence the chalcone synthase gene family in glycine max seed coats," *Plant Cell*, vol. 21, no. 10, pp. 3063–3077, 2009.

[12] K. Varala, K. Swaminathan, Y. Li, and M. E. Hudson, "Rapid genotyping of soybean cultivars using high throughput sequencing," *PLoS ONE*, 2011.

[13] M. A. Gore, M. H. Wright, E. S. Ersoz, P. Bouffard, E. S. Szekeres, T. P. Jarvie, B. L. Hurwitz, A. Narechania, T. T. Harkins, G. S. Grills, D. H. Ware, and E. S. Buckler, "Large-scale discovery of gene-enriched SNPs," *Plant Gen.*, vol. 2, no. 2, pp. 121–133, 2009. [Online]. Available: https://www.agronomy.org/publications/tpg/abstracts/2/2/121

[14] A. J. Severin, G. A. Peiffer, W. W. Xu, D. L. Hyten, B. Bucciarelli, J. A. O'Rourke, Y. Bolon, D. Grant, A. D. Farmer, G. D. May, C. P. Vance, R. C. Shoemaker, and R. M. Stupar, "An integrative approach to genomic introgression mapping," *Plant Physiology*, vol. 154, no. 1, pp. 3 –12, 2010.

[15] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang, "De novo assembly of human genomes with massively parallel short read sequencing," *Genome Research*, vol. 20, no. 2, pp. 265 –272, Feb. 2010.

[16] K. Swaminathan, M. Alabady, K. Varala, E. D. Paoli, I. Ho, D. Rokhsar, A. Arumuganathan, R. Ming, P. Green, B. Meyers, S. Moose, and M. Hudson, "Genomic and small RNA sequencing of miscanthus x giganteus shows the utility of sorghum as a reference genome sequence for andropogoneae grasses," *Genome Biology*, vol. 11, no. 2, p. R12, 2010. [Online]. Available: http://genomebiology.com/2010/11/2/R12

[17] Q. Song, L. Marek, R. Shoemaker, K. Lark, V. Concibido, X. Delannay, J. Specht, and P. Cregan, "A new integrated genetic linkage map of the soybean," *TAG Theoretical and Applied Genetics*, vol. 109, no. 1, pp. 122–128, June 2004. [Online]. Available: http://dx.doi.org/10.1007/s00122-004-1602-3

[18] J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. D. May, Y. Yu, T. Sakurai, T. Umezawa, M. K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X. Zhang, K. Shinozaki, H. T. Nguyen, R. A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R. C. Shoemaker, and S. A. Jackson, "Genome sequence of the palaeopolyploid soybean," *Nature*, vol. 463, no. 7278, pp. 178–183, Jan. 2010.

[19] R. Shoemaker, P. Keim, L. Vodkin, E. Retzel, S. Clifton, R. Waterston, D. Smoller, V. Coryell, A. Khanna, J. Erpelding, X. Gai, V. Brendel, C. Raph-Schmidt, E. Shoop, C. Vielweber, M. Schmatz, D. Pape, Y. Bowers, B. Theising, J. Martin, M. Dante, T. Wylie, and C. Granger, "A compilation of soybean ests: generation and analysis," *Genome/Genome*, vol. 45, pp. 329–338, 2002.

[20] B. McClintock, "The significance of responses of the genome to challenge," *Science*, vol. 226, no. 4676, pp. 792 –801, Nov. 1984. [Online]. Available: http://www.sciencemag.org/content/226/4676/792.short

[21] F. Sanger, A. Coulson, G. Hong, D. Hill, and G. Petersen, "Nucleotide sequence of bacteriophage lambda dna," *J Mol Biol*, vol. 162, pp. 729–773, 1982.

[22] J. Venter, H. Smith, and L. Hood, "A new strategy for genome sequencing," *Nature*, vol. 381, pp. 364–366, 1996.

[23] [Online]. Available: http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi

[24] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren, "Real-time dna sequencing using detection of pyrophosphate release," *Anal Biochem*, vol. 242, pp. 84–89, 1996.

[25] M. Ronaghi, M. Uhlen, and P. Nyren, "A sequencing method based on real-time pyrophosphate," *Science*, vol. 281, pp. 363–365, 1998.

[26] M. Ronaghi, "Pyrosequencing sheds light on dna sequencing," *Genome Res*, vol. 11, pp. 3–11, 2001.

[27] A. Rickert, A. Premstaller, C. Gebhardt, and P. Oefner, "Genotyping of snps in a polyploid genome by pyrosequencing," *BioTechniques*, vol. 32, pp. 592–603, 2002.

[28] M. Margulies, M. Egholm, W. Altman, S. Attiya, J. Bader, L. Bemben, J. Berka, M. Braverman, Y.-J. Chen, Z. Chen, S. Dewell, L. Du, J. Fierro, X. Gomes, B. Godwin, W. He, S. Helgesen, C. Ho, G. Irzyk, S. Jando, M. Alenquer, T. Jarvie, K. Jirage, J. Kim, J. Knight, J. Lanza, J. Leamon, S. Lefkowitz, M. Lei, J. Li, K. Lohman, H. Lu, V. Makhijani, K. McDade, M. McKenna, E. Myers, E. Nickerson, J. Nobile, R. Plant, B. Puc, M. Ronan, G. Roth, G. Sarkis, J. Fredrik Simons, J. Simpson, M. Srinivasan, K. Tartaro, A. Tomasz, K. Vogt, G. Volkmer, S. Wang, Y. Wang, M. Weiner, P. Yu, R. Begley, and J. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, pp. 376–380, 2005.

[29] F. Marek, J. Mudge, L. Darnielle, D. Grant, N. Hanson, M. Paz, Y. Huihuang, R. Denny, K. Larson, D. Foster-Hartnett, A. Cooper, D. Danesh, D. Larsen, T. Schmidt, R. Staggs, J. Crow, E. Retzel, N. Young, and R. Shoemaker, "Soybean genomic survey: Bac-end sequences near rflp and ssr markers," *Genome/Genome*, vol. 44, pp. 572–581, 2001.

[30] C. Boysen, M. Simon, and L. Hood, "Analysis of the 1.1 mb human alpha/beta t-cell receptor locus with bacterial artificial chromosome clones," *Genome Research*, vol. 7, pp. 330–338, 1997.

[31] S. Chissoe, M. Marra, L. Hillier, R. Brinkmann, R. Wilson, and R. Waterston, "Representation of cloned genomic sequences in two sequencing vectors: correlation of dna sequence and subclone distribution," *Nucleic Acids Res*, vol. 25, pp. 2960–2966, 1997.

[32] H.-B. Zhang, X. Zhao, X. Ding, A. Paterson, and R. Wing, "Preparation of megabase-size dna from plant nuclei," *Plant J*, vol. 7, pp. 175–184, 1995.

[33] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. ii. error probabilities," *Genome Res*, vol. 8, pp. 186–194, 1998.

[34] [Online]. Available: http://stan.cropsci.uiuc.edu

[35] K. Arumuganathan and E. Earle, "Nuclear DNA content of some important plant species," *Plant Molecular Biology Reporter*, vol. 9, no. 3, pp. 208–218, Aug. 1991. [Online]. Available: http://dx.doi.org/10.1007/BF02672069

[36] S. Clough, J. Tuteja, M. Li, L. Marek, R. Shoemaker, and L. Vodkin, "Features of a 103-kb gene-rich region in soybean include an inverted perfect repeat cluster of chs genes comprising the ilocus," *Genome/Genome*, vol. 47, pp. 819–831, 2004.

[37] W. Kent, "Blat - the blast-like alignment tool," *Genome Res*, vol. 12, pp. 656–664, 2002.

[38] S. Ouyang and C. Buell, "The tigr repeat databases: a collective resource for the identification of repetitive sequences in plants," *Nucleic Acids Res*, vol. 32, pp. D360–D363, 2004.

[39] P. Green, "Phrap, swat, crossmatch," *Available from the author. University of Washington*, 1999.

[40] R. Goldberg, "Dna sequence organization in the soybean plant," *Biochemical Genetics*, vol. 16, pp. 45–68, 1978.

[41] W. Gurley, A. Hepburn, and J. Key, "Sequence organization of the soybean genome," *Biochim Biophys Acta*, vol. 561, pp. 167–183, 1979.

[42] A. Nunberg, J. Bedell, M. Budiman, R. Citek, S. Clifton, L. Fulton, D. Pape, Z. Cai, T. Joshi, H. Nguyen, D. Xu, and G. Stacey, "Survey sequencing of soybean elucidates the genome structure, composition, and identifies novel repeats," *Functional Plant Biol*, vol. 33, pp. 765–773, 2006.

[43] [Online]. Available: http://www.soybase.org/

[44] M. Vahedian, L. Shi, T. Zhu, R. Okimoto, K. Danna, and P. Keim, "Genomic organization and evolution of the soybean sb92 satellite sequence," *Plant Mol Biol*, vol. 29, pp. 857–862, 1995.

[45] S. Schwartz, W. Kent, A. Smit, Z. Zhang, R. Beartsch, R. Hardison, D. Haussler, and W. Miller, "Human-mouse alignments with blastz," *Genome Res*, vol. 13, pp. 103–107, 2003.

[46] R. Shoemaker, J. Schlueter, and J. Doyle, "Polyploidy and gene duplication in soybean and other legumes," *Curr Op Plant Biol*, vol. 9, pp. 104–109, 2006.

[47] M. Wilson, C. Riemer, D. Martindale, P. Schnupf, A. Boright, T. Cheung, D. Hardy, S. Schwartz, S. Scherer, L.-C. Tsui, W. Miller, and B. Koop, "Comparative analysis of the gene dense ache/tfr2 region on human chromosome 7q22 with the orthologous region on mouse chromosome 5," *Nucleic Acids Res*, vol. 29, pp. 1352–1365, 2003.

[48] [Online]. Available: http://www.soymap.org/

[49] S. Altschul, T. Madden, A. Schaffer, J. Zhang, W. Miller, and D. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Res*, vol. 25, pp. 3389–3402, 1997.

[50] [Online]. Available: http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=soybean

[51] M. Hudson, D. Lisch, and P. Quail, "The fhy3 and far1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome a signaling pathway," *Plant J*, vol. 34, pp. 453–471, 2003.

[52] S. Henikoff, K. Ahmad, and H. Malik, "The centromere paradox: stable inheritance with rapidly evolving dna," *Science*, vol. 293, pp. 1098–1102, 2001.

[53] S. Hall, G. Kettler, and D. Preuss, "Centromere satellites from arabidopsis populations: maintenance of conserved and variable regions," *Genome Res*, vol. 13, pp. 195–205, 2003.

[54] A. Hall, G. Kettler, and D. Preuss, "Dynamic evolution at pericentromeres," *Genome Res*, vol. 16, pp. 355–364, 2006.

[55] K. Choo, B. Vissel, A. Nagy, E. Earle, and P. Kalitsis, "A survey of the genomic distribution of alpha satellite dna on all the human chromosomes, and derivation of a new consensus sequence," *Nucleic Acids Res*, vol. 19, pp. 1179–1182, 1991.

[56] S. Hall, S. Luo, A. Hall, and D. Preuss, "Differential rates of local and global homogenization in centromere satellites from arabidopsis relatives," *Genetics*, vol. 170, pp. 1913–1927, 2005.

[57] M. Gijzen, K. Kuflu, and P. Moy, "Gene amplification of the hps locus in glycine max," *BMC Plant Biol*, vol. 14, pp. 6–6, 2006.

[58] R. J. Singh, K. P. Kollipara, and T. Hymowitz, "Monosomic alien addition lines derived from glycine max (L.) merr. and g. tomentella hayata: Production, characterization, and breeding behavior," *Crop Sci.*, vol. 38, no. 6, pp. 1483–1489, 1998. [Online]. Available: https://www.crops.org/publications/cs/abstracts/38/6/1483

[59] R. J. Singh and T. Hymowitz, "Identification of Four Primary Trisomics of Soybean by Pachytene Chromosome Analysis," *Journal of Heredity*, vol. 82, no. 1, pp. 75–77, 1991. [Online]. Available: http://jhered.oxfordjournals.org/content/82/1/75.short

[60] N. Gill, S. Findley, J. G. Walling, C. Hans, J. Ma, J. Doyle, G. Stacey, and S. A. Jackson, "Molecular and chromosomal evidence for allopolyploidy in soybean," *PLANT PHYSIOLOGY*, vol. 151, no. 3, pp. 1167–1174, Nov. 2009. [Online]. Available: http://www.plantphysiol.org/cgi/content/abstract/151/3/1167

[61] M. Morgante, I. Jurman, L. Shi, T. Zhu, P. Keim, and J. Rafalski, "The STR120 satellite DNA of soybean: organization, evolution and chromosomal specificity," *Chromosome Research*, vol. 5, no. 6, pp. 363–373, Sep. 1997. [Online]. Available: http://dx.doi.org/10.1023/A:1018492208247

[62] J.-Y. Lin, B. H. Jacobus, P. SanMiguel, J. G. Walling, Y. Yuan, R. C. Shoemaker, N. D. Young, and S. A. Jackson, "Pericentromeric regions of soybean (glycine max l. merr.) chromosomes consist of retroelements and tandemly repeated dna and are structurally and evolutionarily labile," *Genetics*, vol. 170, no. 3, pp. 1221–1230, 2005. [Online]. Available: http://www.genetics.org/cgi/content/abstract/170/3/1221

[63] A. Kolchinsky and P. M. Gresshoff, "A major satellite dna of soybean is a 92-base pairs tandem repeat," *TAG Theoretical and Applied Genetics*, vol. 90, pp. 621–626, 1995, 10.1007/BF00222125. [Online]. Available: http://dx.doi.org/10.1007/BF00222125

[64] R. C. Shoemaker, K. Polzin, J. Labate, J. Specht, E. C. Brummer, T. Olson, N. Young, V. Concibido, J. Wilcox, J. P. Tamulonis, G. Kochert, and H. R. Boerma, "Genome duplication in soybean (glycine subgenus soja)," *Genetics*, vol. 144, no. 1, pp. 329–338, 1996. [Online]. Available: http://www.genetics.org/cgi/content/abstract/144/1/329

[65] J. G. Walling, R. Shoemaker, N. Young, J. Mudge, and S. Jackson, "Chromosome-level homeology in paleopolyploid soybean (glycine max) revealed through integration of genetic and chromosome maps," *Genetics*, vol. 172, no. 3, pp. 1893–1900, 2006. [Online]. Available: http://www.genetics.org/cgi/content/abstract/172/3/1893

[66] R. Kolpakov, G. Bana, and G. Kucherov, "mreps: efficient and flexible detection of tandem repeats in DNA," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3672 –3678, July 2003. [Online]. Available: http://nar.oxfordjournals.org/content/31/13/3672.abstract

[67] A. Kato, J. C. Lamb, and J. A. Birchler, "Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 37, pp. 13 554 –13 559, 2004. [Online]. Available: http://www.pnas.org/content/101/37/13554.abstract

[68] A. Tek, K. Kashihara, M. Murata, and K. Nagaki, "Functional centromeres in soybean include two distinct tandem repeats and a retrotransposon," *Chromosome Research*, vol. 18, pp. 337–347, 2010, 10.1007/s10577-010-9119-x. [Online]. Available: http://dx.doi.org/10.1007/s10577-010-9119-x

[69] A. Wawrzynski, T. Ashfield, N. W. Chen, J. Mammadov, A. Nguyen, R. Podicheti, S. B. Cannon, V. Thareau, C. Ameline-Torregrosa, E. Cannon, B. Chacko, A. Couloux, A. Dalwani, R. Denny, S. Deshpande, A. N. Egan, N. Glover, S. Howell, D. Ilut, H. Lai, S. M. del Campo, M. Metcalf, M. O'Bleness, B. E. Pfeil, M. B. Ratnaparkhe, S. Samain, I. Sanders, B. Segurens, M. Sevignac, S. Sherman-Broyles, D. M. Tucker, J. Yi, J. J. Doyle, V. Geffroy, B. A. Roe, M. S. Maroof, N. D. Young, and R. W. Innes, "Replication of nonautonomous retroelements in soybean appears to be both recent and common," *Plant Physiol.*, vol. 148, no. 4, pp. 1760–1771, 2008. [Online]. Available: http://www.plantphysiol.org/cgi/content/abstract/148/4/1760

[70] X. Chen, "Small rnas and their roles in plant development," *Annual Review of Cell and Developmental Biology*, vol. 25, no. 1, pp. 21–44, 2009. [Online]. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev.cellbio.042308.113417

[71] C. Napoli, C. Lemieux, and R. Jorgensen, "Introduction of a chimeric chalcone synthase gene into petunia results in reversible Co-Suppression of homologous genes in trans," *THE PLANT CELL*, vol. 2, no. 4, pp. 279–289, Apr. 1990. [Online]. Available: http://www.plantcell.org/cgi/content/abstract/2/4/279

[72] A. R. van der Krol, L. A. Mur, M. Beld, J. Mol, and A. R. Stuitje, "Flavonoid genes in petunia: Addition of a limited number of gene copies may lead to a suppression of gene expression," *THE PLANT CELL*, vol. 2, no. 4, pp. 291–299, Apr. 1990. [Online]. Available: http://www.plantcell.org/cgi/content/abstract/2/4/291

[73] D. C. Baulcombe and J. J. English, "Ectopic pairing of homologous dna and post-transcriptional gene silencing in transgenic plants," *Current Opinion in Biotechnology*, vol. 7, no. 2, pp. 173 – 180, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/B6VRV-45765VG-2H/2/1665436420a7123a19e50e739a2f1531

[74] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14," *Cell*, vol. 75, no. 5, pp. 843–854, 1993.

[75] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello, "Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans," *Nature*, vol. 391, no. 6669, pp. 806–811, Feb. 1998. [Online]. Available: http://dx.doi.org/10.1038/35888

[76] A. J. Hamilton and D. C. Baulcombe, "A species of small antisense RNA in posttranscriptional gene silencing in plants," *Science*, vol. 286, no. 5441, pp. 950 –952, Oct. 1999. [Online]. Available: http://www.sciencemag.org/content/286/5441/950.abstract

[77] S. N. Covey, N. S. Al-Kaff, A. Langara, and D. S. Turner, "Plants combat infection by gene silencing," *Nature*, vol. 385, no. 6619, pp. 781–782, Feb. 1997. [Online]. Available: http://dx.doi.org/10.1038/385781a0

[78] Z. Xie, L. K. Johansen, A. M. Gustafson, K. D. Kasschau, A. D. Lellis, D. Zilberman, S. E. Jacobsen, and J. C. Carrington, "Genetic and functional diversification of small rna pathways in plants," *PLoS Biol*, vol. 2, no. 5, p. e104, 02 2004.

[79] E. Allen, Z. Xie, A. M. Gustafson, and J. C. Carrington, "microRNA-Directed phasing during Trans-Acting siRNA biogenesis in plants," *Cell*, vol. 121, no. 2, pp. 207–221, Apr. 2005. [Online]. Available: http://www.sciencedirect.com/science/article/B6WSN-4G0KCXX-9/2/744a1d955fbc43418b5728758156c5b4

[80] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun, "The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans," *Nature*, vol. 403, no. 6772, pp. 901–906, Feb. 2000. [Online]. Available: http://dx.doi.org/10.1038/35002607

[81] A. Kozomara and S. Griffiths-Jones, "mirbase: integrating microrna annotation and deep-sequencing data," *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D152–D157, 2011.

[82] S. Subramanian, Y. Fu, R. Sunkar, W. B. Barbazuk, J. Zhu, and O. Yu, "Novel and nodulation-regulated microRNAs in soybean roots," *BMC Genomics*, vol. 9, no. 1, p. 160, 2008.

[83] N. Fahlgren, M. D. Howell, K. D. Kasschau, E. J. Chapman, C. M. Sullivan, J. S. Cumbie, S. A. Givan, T. F. Law, S. R. Grant, J. L. Dangl, and J. C. Carrington, "High-throughput sequencing of *Arabidopsis* microrna: Evidence for frequent birth and death of *MIRNA* genes," *PLoS ONE*, vol. 2, no. 2, p. e219, 2007. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0000219

[84] Y. Lee, M. Kim, J. Han, K. Yeom, S. Lee, S. H. Baek, and V. N. Kim, "*MicroRNA* genes are transcribed by rna polymerase ii," *EMBO J*, vol. 23, no. 20, pp. 4051–4060, Oct. 2004. [Online]. Available: http://dx.doi.org/10.1038/sj.emboj.7600385

[85] W. Park, J. Li, R. Song, J. Messing, and X. Chen, "CARPEL FACTORY, a dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in arabidopsis thaliana," *Current Biology*, vol. 12, no. 17, pp. 1484–1495, Sep. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/B6VRT-46RVM50-J/2/437ca8de0a695b489e611797b9c93d95

[86] E. Allen, Z. Xie, A. M. Gustafson, G. Sung, J. W. Spatafora, and J. C. Carrington, "Evolution of microRNA genes by inverted duplication of target gene sequences in arabidopsis thaliana," *Nat Genet*, vol. 36, no. 12, pp. 1282–1290, Dec. 2004. [Online]. Available: http://dx.doi.org/10.1038/ng1478

[87] F. Fenselau de Felippes, K. Schneeberger, T. Dezulian, D. H. Huson, and D. Weigel, "Evolution of Arabidopsis thaliana microRNAs from random sequences," *RNA*, vol. 14, no. 12, pp. 2455–2459, 2008.

[88] B. J. Reinhart, E. G. Weinstein, M. W. Rhoades, B. Bartel, and D. P. Bartel, "MicroRNAs in plants," *Genes & Development*, vol. 16, no. 13, pp. 1616 –1626, July 2002. [Online]. Available: http://genesdev.cshlp.org/content/16/13/1616.abstract

[89] R. Rajagopalan, H. Vaucheret, J. Trejo, and D. P. Bartel, "A diverse and evolutionarily fluid set of microRNAs in arabidopsis thaliana," *Genes & Development*, vol. 20, no. 24, pp. 3407 –3425, Dec. 2006. [Online]. Available: http://genesdev.cshlp.org/content/20/24/3407.abstract

[90] D. Moldovan, A. Spriggs, J. Yang, B. J. Pogson, E. S. Dennis, and I. W. Wilson, "Hypoxia-responsive microRNAs and trans-acting small interfering RNAs in arabidopsis," *Journal of Experimental Botany*, vol. 61, no. 1, pp. 165 –177, Jan. 2010. [Online]. Available: http://jxb.oxfordjournals.org/content/61/1/165.abstract

[91] V. AMBROS, B. BARTEL, D. P. BARTEL, C. B. BURGE, J. C. CARRINGTON, X. CHEN, G. DREYFUSS, S. R. EDDY, S. GRIFFITHS-JONES, M. MARSHALL, M. MATZKE, G. RUVKUN, and T. TUSCHL, "A uniform system for microRNA annotation," *RNA*, vol. 9, no. 3, pp. 277 –279, Mar. 2003. [Online]. Available: http://rnajournal.cshlp.org/content/9/3/277.abstract

[92] N. R. Markham and M. Zuker, "UNAFold," in *Bioinformatics*, ser. Methods in Molecular Biology, May 2008, vol. 453, pp. 3 –31.

[93] J. Du, D. Grant, Z. Tian, R. Nelson, L. Zhu, R. Shoemaker, and J. Ma, "SoyTEdb: a comprehensive database of transposable elements in the soybean genome," *BMC Genomics*, vol. 11, no. 113, 2010. [Online]. Available: http://www.biomedcentral.com/1471-2164/11/113

[94] W. J. Haun, D. L. Hyten, W. W. Xu, D. J. Gerhardt, T. J. Albert, T. Richmond, J. A. Jeddeloh, G. Jia, N. M. Springer, C. P. Vance, and R. M. Stupar, "The composition and origins of genomic variation among individuals of the soybean reference cultivar williams 82," *Plant Physiol.*, vol. 155, no. 2, pp. 645–655, 2011. [Online]. Available: http://www.plantphysiol.org/cgi/content/abstract/155/2/645

[95] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Research*, vol. 18, no. 11, pp. 1851 –1858, Nov. 2008. [Online]. Available: http://genome.cshlp.org/content/18/11/1851.abstract

[96] B. C. Meyers, M. J. Axtell, B. Bartel, D. P. Bartel, D. Baulcombe, J. L. Bowman, X. Cao, J. C. Carrington, X. Chen, P. J. Green, S. Griffiths-Jones, S. E. Jacobsen, A. C. Mallory, R. A. Martienssen, R. S. Poethig, Y. Qi, H. Vaucheret, O. Voinnet, Y. Watanabe, D. Weigel, and J. Zhu, "Criteria for annotation of plant MicroRNAs," *The Plant Cell Online*, vol. 20, no. 12, pp. 3186 –3190, Dec. 2008. [Online]. Available: http://www.plantcell.org/content/20/12/3186.abstract

[97] X. Dai and P. X. Zhao, "psrnatarget," unpublished data, Available as a web service. [Online]. Available: http://bioinfo3.noble.org/psRNATarget/

[98] M. W. Rhoades, B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel, and D. P. Bartel, "Prediction of plant MicroRNA targets," *Cell*, vol. 110, no. 4, pp. 513–520, Aug. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/B6WSN-4C5HCWR-D/2/7035a7026f715f5b967d9b4f75897758

[99] Z. Du, X. Zhou, Y. Ling, Z. Zhang, and Z. Su, "agriGO: a GO analysis toolkit for the agricultural community," *Nucleic Acids Research*, vol. 38, no. suppl 2, pp. W64 –W70, July 2010.

[100] M. Groszmann, I. K. Greaves, Z. I. Albertyn, G. N. Scofield, W. J. Peacock, and E. S. Dennis, "Changes in 24-nt siRNA levels in arabidopsis hybrids suggest an epigenetic contribution to hybrid vigor," *Proceedings of the National Academy of Sciences*, Jan. 2011.

[101] D. L. Hyten, Q. Song, Y. Zhu, I. Choi, R. L. Nelson, J. M. Costa, J. E. Specht, R. C. Shoemaker, and P. B. Cregan, "Impacts of genetic bottlenecks on soybean genome diversity," *Proceedings of the National Academy of Sciences*, vol. 103, no. 45, pp. 16 666 –16 671, Nov. 2006. [Online]. Available: http://www.pnas.org/content/103/45/16666.abstract

[102] R. Bernard and C. Cremeens, "Registration of 'Williams 82' soybean," *Crop Science*, vol. 28, no. 6, pp. 1027–1028, Dec. 1988.

[103] P. Keim, R. C. Shoemaker, and R. G. Palmer, "Restriction fragment length polymorphism diversity in soybean," *TAG Theoretical and Applied Genetics*, vol. 77, no. 6, pp. 786–792, June 1989. [Online]. Available: http://dx.doi.org/10.1007/BF00268327

[104] J. G. Williams, A. R. Kubelik, K. J. Livak, J. Rafalski, and S. V. Tingey, "DNA polymorphisms amplified by arbitrary primers are useful as genetic markers," *Nucleic Acids Research*, vol. 18, no. 22, pp. 6531 –6535, Nov. 1990. [Online]. Available: http://nar.oxfordjournals.org/content/18/22/6531.abstract

[105] P. Keim, J. M. Schupp, S. E. Travis, K. Clayton, T. Zhu, L. Shi, A. Ferreira, and D. M. Webb, "A High-Density soybean genetic map based on AFLP markers," *Crop Sci.*, vol. 37, no. 2, pp. 537–543, 1997. [Online]. Available: https://www.crops.org/publications/cs/abstracts/37/2/537

[106] J. Rongwen, M. S. Akkaya, A. A. Bhagwat, U. Lavi, and P. B. Cregan, "The use of microsatellite DNA markers for soybean genotype identification," *TAG Theoretical and Applied Genetics*, vol. 90, no. 1, pp. 43–48, Jan. 1995. [Online]. Available: http://dx.doi.org/10.1007/BF00220994

[107] Y. Fu, G. W. Peterson, and M. J. Morrison, "Genetic diversity of canadian soybean cultivars and exotic germplasm revealed by simple sequence repeat markers," *Crop Sci.*, vol. 47, no. 5, pp. 1947–1954, 2007.

[108] L. Wang, R. Guan, L. Zhangxiong, R. Chang, and L. Qiu, "Genetic diversity of chinese cultivated soybean revealed by SSR markers," *Crop Science*, vol. 46, no. 3, pp. 1032–1038, 2006.

[109] T. Y. Hwang, Y. Nakamoto, I. Kono, H. Enoki, H. Funatsuki, K. Kitamura, and M. Ishimoto, "Genetic diversity of cultivated and wild soybeans including japanese elite cultivars as revealed by length polymorphism of SSR markers," *Breeding Science*, vol. 58, no. 3, pp. 315–323, 2008.

[110] Y. Li, R. Guan, Z. Liu, Y. Ma, L. Wang, L. Li, F. Lin, W. Luan, P. Chen, Z. Yan, Y. Guan, L. Zhu, X. Ning, M. Smulders, W. Li, R. Piao, Y. Cui, Z. Yu, M. Guan, R. Chang, A. Hou, A. Shi, B. Zhang, S. Zhu, and L. Qiu, "Genetic structure and diversity of cultivated soybean (*Glycine max (L.)* merr.) landraces in china," *TAG Theoretical and Applied Genetics*, vol. 117, no. 6, pp. 857–871, Oct. 2008. [Online]. Available: http://dx.doi.org/10.1007/s00122-008-0825-0

[111] Y. L. Zhu, Q. J. Song, D. L. Hyten, C. P. V. Tassell, L. K. Matukumalli, D. R. Grimm, S. M. Hyatt, E. W. Fickus, N. D. Young, and P. B. Cregan, "Single-nucleotide polymorphisms in soybean," *Genetics*, vol. 163, no. 3, pp. 1123–1134, 2003.

[112] K. Van, E. Y. Hwang, M. Y. Kim, H. J. Park, S. H. Lee, and P. B. Cregan, "Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the united states and korea," *Journal of Heredity*, vol. 96, no. 5, pp. 529–535, 2005.

[113] M. S. Yoon, Q. J. Song, I. Y. Choi, J. E. Specht, D. L. Hyten, and P. B. Cregan, "BARCSoySNP23: a panel of 23 selected SNPs for soybean cultivar identification," *TAG Theoretical and Applied Genetics*, vol. 114, no. 5, pp. 885–899, 2007.

[114] M. S. Akkaya, R. C. Shoemaker, J. E. Specht, A. A. Bhagwat, and P. B. Cregan, "Integration of simple sequence repeat DNA markers into a soybean linkage map," *Crop Science*, vol. 35, no. 5, pp. 1439–1445, 1995.

[115] P. B. Cregan, T. Jarvik, A. L. Bush, R. C. Shoemaker, K. G. Lark, A. L. Kahler, N. Kaya, T. T. VanToai, D. G. Lohnes, J. Chung et al., "An integrated genetic linkage map of the soybean genome," *Crop Science*, vol. 39, no. 5, pp. 1464–1490, 1999.

[116] I. Choi, D. L. Hyten, L. K. Matukumalli, Q. Song, J. M. Chaky, C. V. Quigley, K. Chase, K. G. Lark, R. S. Reiter, M. Yoon, E. Hwang, S. Yi, N. D. Young, R. C. Shoemaker, C. P. van Tassell, J. E. Specht, and P. B. Cregan, "A soybean transcript map: Gene distribution, haplotype and single-nucleotide polymorphism analysis," *Genetics*, vol. 176, no. 1, pp. 685–696, May 2007. [Online]. Available: http://www.genetics.org/cgi/content/abstract/176/1/685

[117] P. B. Cregan, "The soybean molecular genetic linkage map," *Genetics and genomics of soybean*, pp. 79–90, 2008.

[118] T. M. Seversike, J. D. Ray, J. L. Shultz, and L. C. Purcell, "Soybean molecular linkage group B1 corresponds to classical linkage group 16 based on map location of the lf 2 gene," *TAG Theoretical and Applied Genetics*, vol. 117, no. 2, pp. 143–147, 2008.

[119] K. A. Kaczorowski, K. S. Kim, B. W. Diers, and M. E. Hudson, "Microarray-based genetic mapping using soybean near-isogenic lines and generation of SNP markers in the Rag1 aphid-resistance interval," *The Plant Genome*, vol. 1, no. 2, pp. 89–98, 2008.

[120] K. Kim, S. Bellendir, K. Hudson, C. Hill, G. Hartman, D. Hyten, M. Hudson, and B. Diers, "Fine mapping the soybean aphid resistance gene *Rag1* in soybean," *TAG Theoretical and Applied Genetics*, vol. 120, no. 5, pp. 1063–1071, Mar. 2010. [Online]. Available: http://dx.doi.org/10.1007/s00122-009-1234-8

[121] C. Nickell, G. Noel, T. Cary, and D. Thomas, "Registration of 'Dwight' soybean," *Crop. Sci Crop Science*, vol. 38, no. 5, p. 1398, 1998.

[122] N. Chakraborty, J. Curley, R. D. Frederick, D. L. Hyten, R. L. Nelson, G. L. Hartman, and B. W. Diers, "Mapping and confirmation of a new allele at from soybean PI 594538A conferring RB LesionType resistance to soybean rust," *Crop Sci.*, vol. 49, no. 3, pp. 783–790, 2009.

[123] N. Baird, P. Etter, T. Atwood, M. Currey, A. Shiver, Z. Lewis, E. Selker, W. Cresko, and E. Johnson, "Rapid SNP discovery and genetic mapping using sequenced RAD markers," *PLoS ONE*, vol. 3, no. 10, 2008.

[124] W. Xie, Q. Feng, H. Yu, X. Huang, Q. Zhao, Y. Xing, S. Yu, B. Han, and Q. Zhang, "Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing," *Proceedings of the National Academy of Sciences*, vol. 107, no. 23, pp. 10 578 –10 583, June 2010. [Online]. Available: http://www.pnas.org/content/107/23/10578.abstract

[125] X. Wu, C. Ren, T. Joshi, T. Vuong, D. Xu, and H. Nguyen, "SNP discovery by high-throughput sequencing in soybean," *BMC Genomics*, vol. 11, no. 469, 2010. [Online]. Available: http://www.biomedcentral.com/1471-2164/11/469

[126] N. Gill, S. Findley, J. G. Walling, C. Hans, J. Ma, J. Doyle, G. Stacey, and S. A. Jackson, "Molecular and chromosomal evidence for allopolyploidy in soybean," *Plant Physiol.*, vol. 151, no. 3, pp. 1167–1174, 2009. [Online]. Available: http://www.plantphysiol.org/cgi/content/abstract/151/3/1167

[127] T. Kasuga, S. S. Salimath, J. Shi, M. Gijzen, R. I. Buzzell, and M. K. Bhattacharyya, "High resolution genetic and physical mapping of molecular markers linked to the phytophthora resistance gene *Rps1-k* in soybean," *Molecular Plant-Microbe Interactions*, vol. 10, no. 9, pp. 1035–1044, Dec. 1997.

[128] R. J. Roberts, T. Vincze, J. Posfai, and D. Macelis, "REBASE a database for dna restriction and modification: enzymes, genes and genomes," *Nucleic Acids Research*, vol. 38, no. suppl 1, pp. D234 –D236, Jan. 2010.

[129] S. Deschamps, M. la Rota, J. P. Ratashak, P. Biddle, D. Thureen, A. Farmer, S. Luck, M. Beatty, N. Nagasawa, L. Michael, V. Llaca, H. Sakai, G. May, J. Lightner, and M. A. Campbell, "Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the illumina genome analyzer," *The Plant Genome Journal*, vol. 3, no. 1, pp. 53–68, 2010. [Online]. Available: https://www.crops.org/publications/tpg/articles/3/1/53

[130] D. Hyten, S. Cannon, Q. Song, N. Weeks, E. Fickus, R. Shoemaker, J. Specht, A. Farmer, G. May, and P. Cregan, "High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence," *BMC Genomics*, vol. 11, no. 38, 2010. [Online]. Available: http://www.biomedcentral.com/1471-2164/11/38

[131] J. Emberton, J. Ma, Y. Yuan, P. SanMiguel, and J. L. Bennetzen, "Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries," *Genome Research*, vol. 15, no. 10, pp. 1441 –1446, Oct. 2005. [Online]. Available: http://genome.cshlp.org/content/15/10/1441.abstract

APPENDIX A. SCRIPT TO PARSE MREPS OUTPUT

```
#!/usr/bin/perl

use strict;

if($ARGV[0] eq "" || $ARGV[1] eq ""){usage();}
open RPOUT, ">$ARGV[1]" or die "Cant open output file\n";

my ($size,$rep_size,$read,$c);
my (%repeats,%rep_reads);
open FILE, $ARGV[0];
my $flag=0;
while(my $line = <FILE>){
      chomp $line;
      if($line =~ /Processing sequence/){
 my @contents = split/\s+/,$line;
 $read = $contents[2];
 $read =~ s/\'//g;
 $flag=0;
 $c=0;
 print "parsing $read\t";
      }
      elsif($line =~ /Processing/){
 $size = (split /\s+/, $line)[5];
 $size =~ s/]//;
 print "of size $size\n";
       }
       if($line =~ /-------/){$flag++;next;}
 if($flag==1){
```

```
      print "\tFound repeat ";
      my @contents = split /\t/, $line;
      my @reps = split /\s+/, $contents[$#contents];
              $contents[2] =~ s/</;
              $contents[2] =~ s/>//;
     print "of length $contents[2] with $#reps\n";
              if($#contents >1 && $contents[2] >25){
  $c++;
  $contents[0] =~ s/\s+//g;
  $contents[0] =~ s/>//;
  if ($contents[4] == 0){
      print RPOUT ">".$read."_$contents[0]rpt$c\n$reps[0]\n";next;
  }
      else{
       foreach my $rep(@reps){
   print RPOUT ">".$read."_$contents[0]rpt$c\n$rep\n";
   $c++;
       }
      }
      }
 }
}

close FILE;
close RPOUT;

sub usage()
{
    print "perl parseMreps.pl <mreps result file> <Output file>\n";
}
```

APPENDIX B. SCRIPT TO DETECT GENOMIC REGIONS THAT FORM STABLE HAIRPINS.

```
#!/usr/bin/perl
#Script expects a novoalign input file. Make sure the
#columns make sense, since they change between versions.

open FL, $ARGV[0];
open OUT, ">Longer.ids";
$c=1;
my %checked;
while(<FL>){
 s/>//g;
 @tmp = split /\t/;
 $name = "putmiR$c"; # id of small rna/Query sequence
 $loc = $tmp[1]; # id of target sequence/chromosome
 $loc =~ s/>//;
 #$loc =~ s/scaffold/s/; # only valid for soybean
 $reg = "$loc.$tmp[2].fas";
 $done =0;
 $dir=$tmp[3];
 $len = length($tmp[0]);
 $end = 70 + int $len;
# for(my $j=$tmp[2]-5;$j<=$tmp[2]+5;$j++){
#  if(defined $checked{"$loc.$j.fas"}){$done=1;}
# }
# if($done ==1){next;}
 if($tmp[0] =~ /N/){next;}
 $c++;
 if($dir eq 'R'){
  $len = length($tmp[0]);
```

```
 $end = $tmp[2]+$end;
 $st = $tmp[2]-(100-$len);
 system "fastacmd -d /scratch/bio/db/Glyma1 -s $loc -S 2 -L $st,$end >$reg";
}
else{
 $end = $tmp[2]+100;
 $st = $tmp[2]-70;
 system "fastacmd -d /scratch/bio/db/Glyma1 -s $loc -L $st,$end >$reg";
}
print "fastacmd -d /scratch/bio/db/Glyma1 -s $loc -L $st,$end >$reg\n";
$fs = -s "$reg";
# if the filesize is bigger than 400 the value of
# $end was greater than the length of the sequence.
#Replacing this by 0 retrieves until the end of sequence.
if($fs > 450 || $fs < 50){
 print "Not enough bases. Skipping.\n";
 system "rm $reg";
 next;
}
#system "fastacmd -d /scratch/bio/db/Glyma1 -s $tmp[8] -L $st,0 >$reg";}
system "UNAFold.pl --run-type=html --label=10 --temp=25 $reg >/dev/null";
@cts = <$reg*_*ct>;
$flag = 0;
$more = 0;
foreach $struc(@cts){
 open ST, $struc;
 $pair=$maxpair=0;
 $minpair=600;
 ($seqlen,$dG,$id)= split /\t/, <ST>;
 $dG =~ s/dG\s=\s//;
 for (my $i=1;$i<=70;$i++){<ST>;}
 for (my $i=1;$i<=$len;$i++){
  @tmp = split /\t/, <ST>;
  if ($tmp[4] != 0){
   $upair=0;
   if($tmp[4] >60 && $tmp[4] <(60+$len)){$pair=0;last;}
```

```
    $pair++;
    if($tmp[4] > $maxpair){$maxpair=$tmp[4];}
    if($minpair > $tmp[4]){$minpair=$tmp[4];}
   }
   else{$upair++;}
   if($pair>0 && $upair>3){$pair=0;last;}#print "$upair unpaired ";}
  }
  close ST;
  #print "$reg\t$len\t$struc\t$pair\t$minpair\t$maxpair\t$flag\n";
  if($minpair <100 && $minpair >62){next;}
  if($maxpair >62 && $maxpair <100){next;}
  if($pair >=($len*0.8) && (abs($maxpair-$minpair) < ($len *1.3))){
   $flag=1;$accSt=$struc;
   last;
  }
  elsif($minpair<5 || $maxpair>165){$more++;}
  #print "$reg\t$len\t$struc\t$pair\t$minpair\t$maxpair\t$flag\n";
}
if($more){
 print OUT "$reg\t".(($more/($#cts+1))*70)."\n";
}
open ANN, ">$reg.ann";
for(my $i=1;$i<=171;$i++){
 print ANN "$i\t0\n";
 if($i==70){
  for(my $j=1;$j<=$len;$j++){
   $i++;print ANN "$i\t5\n";
  }
 }
}
close ANN;
foreach $struc(@cts){
 system "sir_graph -ar -lab 10 -pnum -ab -col sir_graph.col \\
  -o $struc.png -png 800 $struc";
}
if(scalar @cts <0){
```

```
  system "sir_graph -ar -lab 10 -pnum -ab -col sir_graph.col \\
  -o $reg.ct.png -png 800 $reg.ct";
 }
 if($flag){
  push @acc, "$reg\t$accSt";
  system "mkdir folds/$name";
  system "mv $reg* folds/$name/";
 }
 else{push @rej, $reg;
  system "mkdir rejects/$name";
  system "mv $reg* rejects/$name";
 }
 $checked{$reg}++;
}

close FL;
close OUT;

open OUT, ">folds/accepted.txt";
foreach $dG(sort @acc){print OUT "$dG\n";}
close OUT;
open OUT, ">folds/rejected.txt";
foreach $dG(sort @rej){print OUT "$dG\n";}
print OUT "A total of $c locations were tested.\n";
close OUT;
```