2012

Genomic landscape of the soybean (Glycine max) genome

Jenna Lynn Woody

Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd

Part of the Genetics Commons, and the Statistics and Probability Commons

Recommended Citation

Woody, Jenna Lynn, "Genomic landscape of the soybean (Glycine max) genome" (2012). Graduate Theses and Dissertations. 12520.
http://lib.dr.iastate.edu/etd/12520

This Dissertation is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Genomic landscape of the soybean (Glycine max) genome

by

Jenna Lynn Woody

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Randy C. Shoemaker, Co-Major Professor
William Beavis, Co-Major Professor
Michelle Graham
Steven Cannon
Carolyn Lawrence
Alicia Carriquiry

Iowa State University
Ames, Iowa
2012

Copyright © Jenna Lynn Woody, 2012. All rights reserved.
# TABLE OF CONTENTS

## CHAPTER 1. GENERAL INTRODUCTION
- Rationale
- Dissertation organization
- Literature review
- Literature cited

## CHAPTER 2. GENE EXPRESSION PATTERNS ARE CORRELATED WITH GENOMIC AND GENIC STRUCTURE IN SOYBEAN
- Abstract
- Introduction
- Methods
  - RNA extraction, isolation and purification
  - Structural parameters of genes
  - Gene expression
  - GC Content
  - Physical Clustering
- Results
  - Genic and intergenic physical parameters
  - GC content
  - Physical clustering
- Discussion
- Literature cited
- Figures
- Tables

## CHAPTER 3. GENE EXPRESSION: SIZING IT ALL UP
- Abstract
- Introduction
- The models
- The “controversy”
- A novel dichotomy in highly expressed genes compared to lowly expressed genes
- Selection on the individual gene or an entire region?
- Literature cited

## CHAPTER 4. HOMOGENEOUS REGIONS (ISOCHORES) IN SOYBEAN (GLYCINE MAX (L.) MERR.)
- Abstract
CHAPTER 5. GENERAL CONCLUSIONS

Conclusions

ACKNOWLEDGEMENTS

SUPPLEMENTAL FIGURES
Chapter 1. General introduction

Rationale

The research objectives for this dissertation were to: 1) Analyze the relationship between the physical parameters of the gene and the transcriptional demands; 2) Evaluate the positioning of the genes along the genome in respect to their transcriptional demands to look for potential clustering; 3) Identify regions of the genome that are significantly homogeneous in their GC content and examine the correlation between the homogeneous regions and other genic properties.

Dissertation Organization

This dissertation contains five chapters. The first chapter consists of the research objectives, dissertation organization and a literature review. The second, third and fourth chapters are each presented as complete manuscripts. Chapter one has been published in Genome, chapter two has been published in Frontiers in Plant Genetics and Genomics, and chapter three has been submitted to Frontiers in Plant Genetics and Genomics. Chapter 5 is a comprehensive conclusion of the research presented and my recommendations for further research.

Chapter two, “Gene expression patterns are correlated with genomic and genic structure in soybean”, was published in 2011 in Genome 54(1);10-8 (Woody et al. 2011). Next generation sequencing and the whole genome
sequence of soybean were used to analyze the relationship between expression characteristics and the physical parameters of the genome. The collaboration in this piece was instrumental to it’s existence. J. L. Woody and R. C. Shoemaker conceived the analysis and wrote the manuscript. RNAseq analysis was taken primarily from the research projects done by Severin et al. (2010) in which Y. T. Bolon, B. Joseph, B. W. Diers, and C. P. Vance were involved in the RNA extraction, isolation, and purification. A. D. Farmer, G. J Muehlbauer, G. D. May, R. C. Shoemaker, N. Weeks, S. B. Cannon, A. J. Severin and J. L. Woody collaboratively sequenced, filtered, mapped and normalized the RNASeq dataset. The assistance from D. Grant, R. T. Nelson, M. A. Graham, S. B. Cannon, N. Weeks, and R. C. Shoemaker was critical in identifying the physical parameters of the genome.

Chapter three, “Gene expression: sizing it all up”, was published in Frontiers in Plant Genetics and Genomics in 2011; 2:70 (Woody and Shoemaker, 2011). An overview of the relationship between gene expression and other genomic parameters is presented in this manuscript. Several hypotheses explaining the evolutionary significance between transcript level and genic length are discussed. The potential relationship between genomic neighborhoods, chromatin and transcription is also discussed. This publication was derived from a collaboration of ideas between R. C. Shoemaker and J. L. Woody.

Chapter four, “Homogeneous regions (iscochores) in soybean (Glycine max (L.) Merr.) has been accepted with moderate revisions om the journal Frontiers in
Plant Genetics and Genomics. In this manuscript long homogeneic genomic regions (LHGRs) in the soybean genome were identified and analyzed. The regions were identified by J. L. Woody using segmentation points (regions where the genome shifts from being GC-rich to AT-rich), the z’ score along the regions and calculating the homogeneity of the regions compared to the chromosome. Genic properties, transposable element density and gene function were all analyzed in relationship to these regions. R. C. Shoemaker and W. Beavis provided guidance in the methods and editorial revisions for the manuscript.

Literature review

The genomic landscape of plants, while slowly being charted, is still composed primarily of unknown territory. The landscape can be related to chromatin domains, transposable element neighborhoods, gene organization, epigenetic modifications of the genome and more. Certain patterns of expression, tissue specific versus constitutive, or high expression versus low expression, etc., are often associated with physical attributes of the gene and genome. Gene expression is not controlled solely by the promoter, it is also modulated by transcription factors, small RNAs, parachromatin, as well as by all of the components that make up epigenetics (Jorgensen, 2011). Characterizing and identifying the internal cues that regulate transcription and translation within the genome can help us decipher the form, function and evolution of living organisms. Recently, with advances in technology, a correlation between the transcriptional profile of the gene and the physical size of the gene has been
observed. The focus of my research project has been to better understand unexplored internal genomic regulations. Coupling next-generation transcriptome data with the recently published soybean genome has allowed us to get a fuller understanding of the relationship between the structural parameters of the gene, transcriptional demands and genomic neighborhoods.

**Regulation of the gene**

At the very basic level, DNA transcription without any modifications is very similar in eukaryotes and prokaryotes (Hahn, 2004). The binding of activators upstream of the core promoter typically engages the RNA polymerase II transcription cycle. Once the binding is complete, adaptor complexes are brought in which then leads to the binding of general transcription factors and this forms the transcription initiation complex (Thomas and Chiang, 2006). RNA polymerase II is then placed on the core promoter with the aid of several transcription factors and this forms the closed preinitiation complex. The RNA polymerase clears the promoter and the single-strand template is placed in the RNA polymerase II cleft which initiates RNA synthesis (Buratowski, 2003).

Regulatory elements such as enhancers often modify this basic level of transcription. In many cases, the regulatory elements act on the nearest gene in cis and then the linear proximity of a gene to a regulatory site can be used as a predictor of the target gene (de Villiers et al., 1983). However, regulatory elements do not always act on the nearest gene and can also influence multiple genes at once. The number of genes that an element can control depends on the
chromosomal context. For example, the β-globin locus control region (LCR) activator was analyzed in the natural location and at a new location with no globin genes present. In the natural location, the β-globin LCR activated two or three β-globin genes at different developmental time points. When placed in a new location with no globin genes present, the β-globin LCR activated 6-7 genes (Noordermeer et al., 2008). This implies that the organization and nearby genomic content influences the regulators of genes. In differentiated cells it has been estimated that about 60% of the DNA is transcribed in a given cell and condition and up to 95% of genomic DNA is possibly transcribed within the lifespan of an organism. Considering less than 2% of the DNA in higher eukaryotes is actually translated into proteins, the highest level of regulation is possibly the genomic DNA itself, positioning itself spatially into chromosome territories and genomic domains (Scherrer, 2011). It is apparent that gene expression is an intricate multistep process that is influenced through multiple levels of regulation; through transcriptional machinery that is locally recruited to the genes, chromatin remodeling, and the movement of the DNA within the nucleus.

Gene organization

The organization of genes in eukaryotic genomes is non-random and has been suggested to play an important role in regulating gene expression and evolution. Some regions in the genome have a conserved gene order and within these regions, the genes may be expressed together throughout specific
developmental time points or in specific tissues. The concept that the transcriptional profile of a gene is influenced by the region of the genome in which it resides is supported by the study of transgenes. It has been observed that the expression characteristics of a transgene changes depending on the region into which it is inserted (Milot et al., 1996; Gierman et al., 2007; Csink et al., 2002). More recently, whole-genome wide transcriptional studies have shown non-random clustering of co-expressed genes (Guttikonda et al., 2010). These co-expressed gene clusters are also conserved across species, as shown in a comparison between *S. cerevisiae* and *C. albicans* where the co-expressed genes are conserved more than expected by chance (Hurst et al., 2002). Chromosomal breakpoints also avoid co-expressed clusters. In Drosophila, human, mouse and chicken clusters of co-expressed genes are under-represented in the number of breakpoints compared to the rest of the genome (Ranz et al., 2007; Singer et al., 2005; Semon et al., 2006).

Tandem gene arrays, for example the *Hox* and *globin* clusters, are one example of co-expressed clusters in the eukaryotic genomes. Tandem gene arrays consist of individual genes that form in a collinear series, often formed during duplication events. These linear arrays can contain tissue-specific genes or ubiquitously expressed genes but they are all regulated together (share similar transcription profiles). The linear placement of these genes along the chromosome is critical to the transcription of the individual genes. For example, at the *β-globin* locus those genes closest to the locus control region (LCR) are
upregulated first. In humans, the *rDNA* tandem gene arrays have been studied with fluorescence in situ hybridization (FISH) and found that this tandem gene array's position in the nucleus during cellular differentiation is specific and a change in position correlates with a change in the transcription of the genes (Kosak et al., 2002). This suggests that the co-expressed clusters may form as a function of nuclear positioning.

Beyond tandem gene arrays, eukaryotic genomes appear to be organized into clusters of genes with common transcriptional profiles. In humans, genome-wide mRNA expression studies have shown that the genome is organized into compartments of high and low levels of gene activity (Caron et al., 2001). In *S. cerevisiae*, gene pairs involved in cell cycle phases, sporulation and the pheromone response were found clustered together across the genome (Cohen et al., 2000). An expression analysis within the Drosophila genome under 80 different experimental conditions uncovered groups of 10 to 30 genes covering 20 to 200 kilobases (kb) that were co-expressed (Spellman and Rubin, 2002). It is evident that many co-expressed genes cluster into neighborhoods across eukaryotic genomes but the mechanism that controls the clustering, the functional significance, and the consequences of the clustering is still unclear. One likely explanation is that gene clusters share a common regulatory element as was shown in several gene clusters in *S. cerevisiae* (Cohen et al., 2000).

However, it is unlikely that this explains all gene clusters described above. The proximity of the genes could ensure that the concentration of shared proteins in
certain locations is sufficient. The hypothesis that there are localized transcriptional centers or "expression hubs" is supported when nuclear subcompartments were studied and certain regions showed increased localized protein concentrations (Thompson et al., 2003). This idea suggests that higher order chromatin and nucleus interactions are one of the mechanism by which gene clusters function.

**Chromatin**

Chromatin structure and histone modifications within the chromatin also play a key role in transcriptional regulation and the formation of genomic neighborhoods. The chromatin structure is comprised of approximately 200 bp of DNA and a repeating unit of histones (Kornberg and Thomas, 1974) and plays a vital role in gene regulation. Chromatin also functions in a variety of other necessary cell processes of which the main functions are to package DNA, strengthen DNA for mitosis and meiosis, and protect against DNA damage and DNA replication. The structure of chromatin determines its function: the more tightly wrapped chromatin reduces or eliminates transcription while loose chromatin increases transcription. In general, we know that chromatin has three main levels of organization: euchromatin in which DNA is wrapped around histones, heterochromatin which consists of several histones wrapped even tighter into a 30 nm fiber, and finally a very high level of packaging which occurs during mitosis and meiosis. Specific mechanisms regulate the condensation state
of the chromatin and include covalent modification of histone tails and the repositioning of nucleosomes along the DNA fiber (Cosma, 2002).

Chromatin-based gene regulation can be accomplished with covalent post-translational modifications of histones, ATP-dependent chromatin remodeling and replacement of canonical histones with specialized variants. Post-translational modifications include methylation of arginine residues, methylation, acetylation, ubiquitination, ADP-ribosylation, sumoylation of lysines, and phosphorylation of serines and threonines (Li et al., 2007a). Certain modifications are associated with active genes such as the acetylation of histone 3 (H3) and histone 4 (H4) while others associate with inactive genes such as methylation at H3 lysine 9. To be effective, the location of the modification is crucial (Landry et al., 2003). ATP-dependent chromatin remodeling utilizes ATP hydrolysis to change the histone-DNA complex and usually results in de-compaction of the end DNA from histone octamers, formation of the DNA loop, or sliding the nucleosomes along the chromosome (Flaus and Owen-Hughes, 2004). The third main mechanism of chromatin-based gene regulation is the incorporation of histone variants (Li et al., 2007). The main feature that distinguishes variant histones from the canonical core histones is that they are expressed outside of S phase and are incorporated into the chromatin independently of DNA replication. The functional significance of histone variants is broad and not yet completely understood. The histone variant H2AX in animals is found at sites of DNA damage and plays a role in the DNA repair pathway (van
Attikum and Gasser, 2009). The histone variant H2A.Z is located in nucleosomes flanking the transcriptional start site and plays a role in transcriptional regulation by preventing DNA methylation (Zilberman et al., 2008; Conerly et al., 2010).

Histone variant CenH3 is found in all eukaryotes. This variant is incorporated at centromeres and plays a vital role in chromosome segregation (Malik and Henikoff, 2003). It is apparent that the incorporation of a histone variant can take on a variety of roles and is integral to the regulation and maintenance of the genome.

Communication between cis-regulatory DNA elements that can be separated by hundreds of kilobases frequently regulate the transcription of genes in higher eukaryotes. The mechanism behind this communication lies in long-range chromatin interactions, or chromatin loops (Kadauke and Blobel, 2009). Chromatin loops can occur between a variety of elements but is most commonly studied between an enhancer and promoter (O'Sullivan et al., 2004).

Transcriptionally silent chromosome territories are typically located at the nuclear periphery while the transcriptionally active domains reside in the interior. The transcriptionally silent chromosome territories are generally gene-poor while the active chromosome territories are generally gene-rich. The location of chromatin loops can be static and studies have shown that as a chromatin loop changes to a different chromatin territory the transcriptional profile of the loop often changes as well. In a study focusing on the interaction between the $\beta$-globin gene and the locus control region (LCR) Palstra et al. (2003) were able to show that changes in
the physical interactions between the β-globin locus and the LCR changed in
correlation with a change in gene expression. Another example was shown in
mice, Meshorer and Misteli (2006) cultured pluripotent mouse embryonic stem
cells and showed that dispersed chromatin was not very compact. Once
differentiation occurred, large regions of the genome underwent chromatin
compaction (Meshorer and Misteli, 2006). Certain biological conditions such as
different developmental stages and changes in environmental conditions also
result in large-scale de-condensation in which the whole genome is de-
compacted (Barlow, 1976; Mathieu et al., 2003; Tessadori et al., 2007). The
observations above have led to the suggestion that the arrangement of the
chromosomes in the nucleus may be cell type-specific and key to understanding
the regulation of transcription. Chromatin domains comprising gene clusters may
influence the way the nucleus is configured. Given the large amount of
transcription occurring at any time within the cell it may be expedient for
transcription if genes are co-localized into domains. This could give rise to
transcription centers or pools in which a group of genes share the same
concentration of proteins.

**Expression breadth/level and genic size**

The previous sections described the transcriptional profile of genes as
being correlated with larger regions of the genome. At an individual gene level,
exon size, intron size, intron density, and length of the intergenic region have
been found to correlate with the transcriptional profile. A correlation between
transcriptional demands and these structures that comprise the gene have been identified. Differences in the structural properties of genes with a “narrow expression breadth” in comparison with genes with a “broad expression breadth” or genes with a high level of expression in comparison with a low level of expression, have recently been discovered. In humans and *Caenorhabditis elegans*, the introns in highly expressed genes are substantially shorter than in those genes that are expressed at low levels (Castillo-Davis et al., 2002). Subsequent research revealed that in humans, high expression level also correlated with smaller protein products, less intronic DNA, and greater codon and amino acid biases (Urrutia and Hurst, 2003).

Genes with high transcriptional demands (those that are highly expressed or broadly expressed) appear to be located in genomic regions that are gene dense, GC rich, with shorter coding regions and introns than those that are lowly expressed or tissue-specific (Versteeg et al., 2003). The size of the protein and the intron sizes vary considerably across eukaryotic organism but a negative relationship between the expression level and the gene size was found across a diverse selection of species including Drosophila, *C. elegans*, humans and *A. thaliana* (Coghlan et al., 2000; Jansen et al., 2000; Castillo-Davis et al., 2002; Urrutia et al., 2003; Seoighe et al., 2005). A negative relationship between expression level and protein length was also found in *S. cerevisiae* (Coghlan and Wolfe, 2000).
A negative relationship between expression breadth (the number of tissues a gene is expressed in) and the physical size has also been observed. In humans, a negative relationship between expression breadth and protein length was identified (Vinogradov, 2004). Similarly in mouse and rat, as expression breadth increases the intron size and intergenic regions decrease (Pozzoli et al., 2007). Thus, it is not just the coding regions that are correlated with the expression breadth but also the region around the gene. Genes and intergenic sequences located in the GC rich regions are shorter than those in the GC poor regions in warm-blooded vertebrates (Bernardi, 2000). The location of GC rich genes is typically in the open chromatin while the GC poor genes are located in the more compact, peripheral chromatin (Saccone et al., 2002).

Three models have been suggested as explanation for the relationship between transcriptional demands and genic size. The “selection for economy” proponents base their argument on the fact that transcription and translation are both time-consuming and costly (Urrutia and Hurst, 2003; Seoighe et al., 2005). To transcribe one nucleotide, two adenine triphosphate molecules and roughly 0.05 seconds are required (Carmel and Koonin, 2009) therefore it would be advantageous to the organism to reduce the cost of those genes ubiquitously and highly transcribed and translated. Within the “selection for economy” model there are two sub-arguments; the energetic cost hypothesis and the time cost hypothesis. The energetic cost hypothesis states that selection is influenced by a drive to minimize the energetic cost of transcription. Alternatively, in the time cost
hypothesis, a selection for shorter genes occurs when limited time periods are
required to transcribe large amounts of mRNA (Rao et al., 2010). The common
thread is that the decrease in genic size is a result of selected mutations with the
purpose to decrease the demands of highly transcribed genes. Carmel et al.
(2009) analyzed the genomic architecture of humans, C. elegans, D.
melanogaster, and A. thaliana and observed that as the expression level
increased, the gene became more compact. They also found that correlation was
significantly weaker between expression breadth and the gene size. However,
the “selection for economy” hypothesis implicitly assumes neutral accumulation
of non-coding DNA and cannot explain the correlation between expression
characteristics and the length of the intergenic region.

Similar to the “selection for economy” model is the “mutational bias”
argument which proposes that the intron length of highly expressed genes
decrease as the transcriptional demands decrease. However, instead of the
individual gene being selected upon, the mutational bias hypothesis focuses on
the entire region of the genome in which the gene resides and is based upon
transcription-associated non-adaptive deletion bias (Urrutia and Hurst, 2003;
Comerón, 2004). Genomic regions containing many genes tend to be GC rich
(Urrutia and Hurst, 2003) and are also regions of high recombination rates
(Fullerton et al., 2001). A simple mutational bias could be considered if
recombination induces mutations. Genes that have a higher transcriptional
demand would also be more prone to reverse transcription and retrotransposition
(Mourier and Jeffares, 2003). The mutational bias model suggests that the ‘neighborhood’ of the gene is the cause for selection rather than the transcriptional demands of the individual gene.

The “genome complexity” model postulates that longer introns and intergenic regions are the result of an increase in regulatory elements (Vinogradov, 2006; Eisenberg and Levanon, 2003). Vinogradov (2004) suggested that broadly expressed genes require simple regulation and therefore less regulatory elements. Conversely, tissue specific genes contain more functional domains and are associated with more complex protein architecture resulting in larger gene “spaces.” In support of this model, intron length negatively correlates with both protein divergence and polymorphism in Drosophila. This trend towards conservation suggests that the longer introns contain more functional units (Haddrill et al., 2005, Petit et al., 2007). This was extended to rodents and they found that an increase in intron length correlated with an increase in regulatory elements (Pozzoli et al., 2003). Studying the intergenic regions is critical to this hypothesis. Enhancers, silencers and insulators of tissue-specific genes are located upstream and downstream of genes and can work over distances of 100 kb (Levine and Tijan, 2003). Recently, intergenic sequences have been shown to participate in chromatin-mediated repression that involves blocks of genes rather than just one gene. A greater amount of noncoding sequence surrounding a gene or a group of genes could help create a 'gene nest' which would help facilitate the suppression of gene
activity in most tissues (Zuckerkandl, 2002). Interestingly, in broadly expressed genes, functional domains cover a higher percent of coding region than in tissue-specific genes. However, this difference is not enough to explain the increase in the average length of the coding sequence in tissue-specific genes (Vinogradov, 2004). Rather than the expression profile impacting the gene size, the “genome complexity” model suggests that functional properties of a gene determine the length of the physical genic properties (Eisenberg and Levanon, 2003; Vinogradov, 2006).

**Plant controversy**

In plants, a controversy has emerged regarding expression and the structure of plant genomes. Ren et al. (2006) studied both *Oryza sativa* and *A. thaliana* and found that highly expressed genes contained more and longer introns and produced a larger primary transcript than genes expressed at a low level. However, this situation is complicated as these contradictions are not found across all plant species. In the moss plant *Physcomitrella patens*, highly expressed genes contained shorter introns (Stenoien, 2007). To make this situation more complex, the housekeeping genes of *A. thaliana* are less compact than the tissue-specific genes when the expression level is controlled (Li et al., 2007). These results are in contradiction to the previous studies done in animals. Another study in *A. thaliana* also found that non-coding regions got larger as expression breadth increased (Yang, 2009). However, in the same study the coding regions got smaller as expression breadth increased. In a separate study
in *A. thaliana* both the non-coding and coding regions of the genes decreased as the expression level increased (Camiolo et al., 2009).

Interestingly, it is not just in plants that opposing correlations have been discovered. In several yeasts and other unicellular organisms, genes expressed at a high level have longer introns than genes expressed at a low level (Vinogradov, 2001). In *Ostreococcus lucimarinus*, a unicellular green algae, the number of introns and the intron density increase as the expression level increases (Lanier et al., 2008). A different study in the unicellular organism *S. cerevisiae* also observed an increase in intron size as the expression level increased but the protein length showed an opposite trend and decreased as the expression level increased (Coghlan and Wolfe, 2000). Another example is in animals. In chicken, genes expressed in every tissue are larger than genes expressed in one tissue (Rao et al., 2010). In the same study, the gene size, coding sequence length, first intron length, average intron length and total intron length all negatively correlate with expression level. The relationship between genic size and both expression level and expression breadth vary across species.

In many of these reports, either expression level or expression breadth was studied independently. Throughout my research project I have attempted to resolve these inconsistencies and determine which of the contrasting patterns are evident in the soybean genome.
**Genic neighborhoods**

If weakly expressed genes evolve under the umbrella of alternative splicing demands, it would appear evident that selection would be at an individual level. However, if the selection was for an economical purpose, it is reasonable to question whether entire neighborhoods are under specific selection. Clustering of highly expressed genes has been established and several physical genomic properties have been associated with these regions. In a study that combines transposable elements, gene length, and gene expression Jjingo et al. (2011) found that all three of those factors are closely related. Combined together, transposable elements and gene length account for 78% of the variation in expression level, 76% of the variation in expression breadth, and 66% of the variation in tissue specificity. The authors proposed that selection for economy plays a role in the evolution of gene size but suggested that the removal of transposable elements may be a stronger mechanism of selection than reduction of gene length. In a study done in rice (Tian et al., 2009) retrotransposons, genetic recombination, and gene density were all correlated and the authors suggested this correlation helped shape the makeup of the rice genome.

In rice, transposable element families were differentially distributed across the genomes in areas of varying methylation patterns (Takata et al., 2007). Kim et al. (2004) found that the expression breadth of a gene is highly correlated with Alu elements (repetitive elements) and expression level is highly correlated with
L1 densities in human. Confirmed by Eller et al. (2007), highly and broadly expressed genes are enriched with Alu elements and depleted in L1 elements. This suggests that rather than gene expression or transposable element insertion accounting for a variation in genic level, epigenetics may be influencing the entire genetic region. Isochores, large regions within the genome that are homogeneous in their GC content have been characterized and analyzed since 1976 (Macaya et al., 1976). Gene density, gene expression, insertion of transposable elements and density of transposable elements are only a few of the basic biological properties associated with isochores (Bernardi, 2004). It is possible that these properties act as a unit and isochores are the homes for these interactions.

**LHGRs**

Isochores, or more recently termed LHGRs (long homogeneous genome regions) are regions of the genome that are homogeneous in their GC content that have been characterized and analyzed since 1976 (Macaya et al. 1976). It has been suggested that most eukaryotic genomes contain LHGR regions, or mosaics of homogeneous GC content that abruptly change from one neighboring LHGR to the next. LHGRs were first observed using ultracentrifugation in CsCl density gradients and were called isochores (Macaya et al. 1976). DNA fractionation by ultracentrifugation, cytogenetic analyses, and recently, analyses of genes and genome sequences, have been utilized to identify these regions.
With advances in technology and the availability of whole genome sequences, LHGRs can now be identified with more precision using computational tools.

LHGR regions have been defined as segments of DNA, typically above 300 kb, that are relatively homogeneous in terms of AT and GC content, and that can have sharp boundaries from the neighboring stretches of DNA. LHGRs are broken into families that are characterized by their GC levels. The LHGR family designations in vertebrates and invertebrates are based on the frequency distributions of LHGRs across GC percentage (Macaya et al., 1976). Multiple peaks in the distribution of LHGRs across GC content are evident in most species although the demarcation between one LHGR family and the adjacent family is often poorly defined. The mid point between two peaks in the frequency of the LHGRs by GC content is often considered the breakpoint between where one family ends and the next family begins. A different technique was applied in Arabidopsis (Zhang and Zhang, 2004) where the LHGRs were identified as GC-rich (GC content of the LHGR was above the average GC content for the chromosome in which it resided) and AT-rich (GC content of the LGHR was below the average GC content for the chromosome in which is resided) because no statistically distinguishable peaks were discovered.

Fundamental biological properties have been found to be tightly correlated with certain LHGR families, defined by the frequency distributions of LHGRs across GC percent, such as repeat sequence distribution, gene density,
replication timing, CpG distribution, genic size and transcription abundance (Bernardi 2007). In humans, the patterns of chromosome banding during prophase are consistent with the number and distribution of LHGRs (Constantini et al. 2006). LHGR families are strongly conserved between invertebrate species (Cammarano et al. 2009) as well as between vertebrate species (Constantini et al. 2009). Even the dinucleotide patterns, LHGR sizes and the relative amount of LHGRs are mainly conserved (Constantini et al. 2009).

Specific dinucleotide frequencies are correlated with each of the LHGR families. The dinucleotide frequencies specific to a family are conserved among species (Constantini and Bernardi, 2008b). In all vertebrate genomes, the GC levels, dinucleotide frequencies, and LHGR sizes are conserved across all vertebrates studied (Constantini et al. 2009). These differences are likely to influence the RNA and protein interactions as well as the chromatin structure. It is possible that there are only a discrete number of chromatin structures and the LHGR families represent these structures. This idea supports the conservative mode of evolution hypothesis which states that evolution (Constantini et al. 2009) is constrained by "negative selection acting at a regional (isochore) level to eliminate any strong deviation from the presumably functionally optimal composition of isochores" (Bernardi et al. 1988).

The LHGR structure of plants was initially characterized by looking only at the sequence of several genes at a time and at the surrounding genomic
sequence (Matassi et al. 1989, Montero et al. 1990, Salinas et al. 1988). Monocots and dicots differ in their compositional pattern of LHGRs. Monocots have a much wider range of GC content than do dicots (Salinas et al. 1988). Interestingly, compositional similarities exist between monocots and warm-blooded vertebrates. Both groups have higher GC content relative to dicots and cold-blooded vertebrates and even higher GC content in the coding regions. Similar to warm-blooded vertebrates, monocots also show a similar distribution of housekeeping genes compared to tissue-specific genes with the housekeeping genes having a much higher GC content (Salinas et al. 1988). Warm-blooded vertebrates and monocots also show a bimodal distribution of GC3, or the wobble codon, while cold-blooded animals and other plants show a unimodal distribution (Macaya et al. 1976). Since this bimodality is present in monocots and warm-blooded vertebrates it is suggested that these organisms have developed this evolutionary feature independently. It is also possible that the bimodality is a result of a larger genomic bias, such as the LHGR structure (Tatarinova et al. 2010).

There are also two competing processes that may influence this separation between warm-blooded vertebrates and monocots compared to cold-blooded vertebrates and dicots. The action of GC based mismatch repair is pronounced in highly recombinating organisms such as grasses and warm-blooded vertebrates (Birdsell 2002). In self-pollinating and asexually reproducing species
the effect of recombination is counteracted by AT biased mutational pressure (Birdsell 2002). Cytosine deamination and oxidative damage to C and G bases are the cause of AT biased mutational pressure (Coulondre et al. 1978; Newcomb and Loeb 1998). Comparison between the coding regions of Arabidopsis and three of its closest relatives (Raphanus sativus, Brassica rapa, and Brassica napus) showed that the GC3 values of R. sativus, B. rapa, and B. napus genes are on average 0.05 higher than Arabidopsis (Villagomez et al. 2009). This supports the idea that selfing individuals are under different mutational pressures as B. rapa and R. sativus are self-incompatible while Arabidopsis is self-pollinating.

Arabidopsis is the only plant, until now, in which LHGRs have been identified and characterized at a whole genome level. In previous studies, only specific genes and small regions of the genome have been analyzed. Several distinctions were found in Arabidopsis compared to humans. The range of GC content was much smaller in Arabidopsis, the average GC content for GC rich LHGRs was 0.37 and AT rich LHGRs was 0.34 compared to the average for humans, the AT rich LHGRs had an average of 0.38 while GC rich LHGRs had an average of 0.47 (Zhang and Zhang 2004). Another distinction was in the discovery of centromeric LHGRs. These LHGRs were GC rich yet had a low gene density and different T-DNA insertion sites than the other GC rich LHGRs. Each of the centromeric regions were contained within centromeric LHGRs and all of
the predicted centromere sequences were part of an LHGR region (Zhang and Zhang 2004).

Plants are unique in their ability to make remarkable changes in genome size and composition. The question remains as to what structural constraints genomes face. Understanding the role of genome size and composition could help decipher the “rules” of genome structure and function in plants. Through my research I use the recently published soybean genome sequence (Schmutz et al. 2010) to identify and characterize the LHGRs in the soybean genome.

**Literature cited**


Jorgensen RA. 2011. We’re all computation biologists now… Next stop, the global brain? Front Genet 2:68.


Kim TM, Yung YC, and Rhyu MG. 2004. Alu and L1 retroelements are correlated with the tissue extent and peak rate of gene expression, respectively. J. Korean Med Sci 19:783–792.


CHAPTER 2. GENE EXPRESSION PATTERNS ARE CORRELATED WITH GENOMIC AND GENIC STRUCTURE IN SOYBEAN

A paper published in Genome.

Jenna L. Woody¹, Andrew Severin¹, Yung-Tsi Bolon², Bindu Joseph¹, Brian W. Diers³, Andrew T. Farmer⁴, Nathan Weeks⁵, Gary J. Muehlbauer⁶, Rex Nelson⁵, David Grant⁵, James E. Specht⁷, Michelle A. Graham⁵, Steven B. Cannon⁵, Gregory D. May⁵, Carroll P. Vance²,⁶ Randy C. Shoemaker⁵.

Abstract

Studies have indicated that exon and intron size and intergenic distance are correlated with gene expression levels and expression breadth. Previous reports on these correlations in plants and animals have been conflicting. In this study, next-generation sequence data, which has been shown to be more sensitive than previous expression profiling technologies, were generated and analyzed from 14 tissues. Our results revealed a novel dichotomy. At the low expression level, an increase in expression breadth correlated with an increase in transcript size because of an increase in the number of exons and introns. No significant changes in intron or exon sizes were noted. Conversely, genes

¹Department of Agronomy, Iowa State University, Ames, IA 50011, ²United States Department of Agriculture-Agricultural Research Service, Plant Research Unit, St. Paul, MN 55108, ³Department of Crop Sciences, University of Illinois, 1101 West Peabody Dr., Urbana, IL 61801, ⁴National Center for Genome Resources, Santa Fe, NM 87505, ⁵United States Department of Agriculture-Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, ⁶Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, and ⁷Department of Agronomy, University of Nebraska, Lincoln, NE 68583.
expressed at the intermediate to high expression levels displayed a decrease in transcript size as their expression breadth increased. This was due to smaller exons, with no significant change in the number of exons. Taking advantage of the known gene space of soybean, we evaluated the positioning of genes and found significant clustering of similarly expressed genes. Identifying the correlations between the physical parameters of individual genes could lead to uncovering the role of regulation owing to nucleotide composition, which might have potential impacts in discerning the role of the noncoding regions.

Introduction

The uniqueness of living organisms is dependent not only upon the suite of genic material available to them but also upon the subtleties of gene expression patterns. Some genes have a narrow breadth of expression and are expressed in only a single tissue, while other genes are more broadly expressed in many tissues. Some genes produce very little transcript while others are veritable gushers of message. The underlying mechanisms controlling these differences in expression patterns are often attributed to regulatory factors at the genic level. However, the control of gene expression may also have components related to gene structure or genomic context.

Previous studies have revealed striking differences in the structural properties of genes with a narrow expression breadth vis-a`-vis in comparison with genes with a broad expression breadth. Differences in gene structure were
also observed between genes with high and low levels of expression. In humans, high expression level correlates with smaller protein product, less intronic DNA, and greater codon and amino acid biases (Urrutia and Hurst 2003). A negative relationship between expression level and protein length was also found in *Saccharomyces cerevisiae* (Coghlan and Wolfe 2000). In addition to increased levels of expression correlating with a decreased protein size, a negative relationship between expression breadth and protein and intron length also was identified in humans (Vinogradov 2004). In mouse and rat, as expression breadth increases, the intron size and intergenic region lengths also decrease (Pozzoli et al. 2007). Three hypotheses have been proposed to explain these relationships: natural selection for economy, mutational bias, and genome complexity (Vinogradov 2006).

The “natural selection for economy” argument suggests that highly expressed genes are smaller as a result of selected deletions. The implication of this hypothesis is that genes highly expressed will be less “costly” if minimized in size (Eisenberg and Levanon 2003). The “mutational bias” hypothesis also contends that intron length of highly expressed genes decrease as expression levels increase. It is proposed that this negative relationship is caused by regional transcription-associated deletion. In *Drosophila*, an increase in intron length was correlated with a decrease in GC content but was not correlated with a change in the number of functional elements of the coding regions. This evidence supports mutational bias rather than natural selection for economy.
The “genome complexity” hypothesis maintains that the size variation of genic and intergenic regions is based upon the intricacy of the gene where those with more complex structures require more regulatory elements for nucleosome formation. In mouse and rat, for example, more chromatin-regulated suppression (large intergenic regions) is required for genes expressed in just one tissue, whereas the more ubiquitous genes are smaller and are located in more compact regions (Chen et al. 2005). As a consequence, as a gene’s complexity increases, it’s intron length, coding, and intergenic regions would be predicted to increase.

When these hypotheses were tested in plants, contradictory results were found for both Arabidopsis and rice. Genes expressed at a high level were found to have more and longer introns and a longer primary transcript than those genes expressed at a lower level (Ren et al. 2006). In a different study, genes expressed in the male gametophyte in Arabidopsis at high levels were found to have less intron density than genes expressed exclusively in the sporophyte at lower levels. This observation indicates support for the natural selection for economy hypothesis of gene evolution (Seoighe et al. 2005). In another study, Arabidopsis genes with a broad expression breadth were found to have larger transcripts than those with a smaller breadth of expression, while genes with a high expression level were found to have smaller transcripts than those with a lower expression level (Camiolo et al. 2009).

In many of these reports, either expression level or expression breadth
was studied independently. In an attempt to resolve these inconsistencies and to determine which of the contrasting patterns are evident in the soybean genome, we combined both factors and looked at expression intensity in conjunction with expression breadth. Our results revealed an intriguing pattern of genic and genomic physical parameters associated with expression intensity and expression breadth.

**Materials and methods**

**RNA extraction, isolation, and purification**

Soybean seeds (experimental line LD0–15146) were grown in conditions that mimicked Illinois field growing conditions as specified in Severin et al. (2010).

Briefly, tissue samples were taken from a minimum of three plants for each extraction and harvested at approximately 1400 h. Tissue samples were collected from 14 tissues: root, nodule, flower, young leaf, one cm pod, seven stages of seed development, and two stages of pod-shell development. The stages in seed development include 10 days after flowering (DAF), 14 DAF, 21 DAF, 25 DAF, 28 DAF, and 42 DAF. The stages of pod-shell development include 10 DAF and 14 DAF. Root tissue was harvested after 12 days and nodules were harvested 20–25 days after inoculation. Standard RNA isolation and purification was performed according to Severin et al. (2010). Samples were ground by mortar and pestle with liquid nitrogen and the RNA isolation was done using a modified TRizol (Invitrogen, Carlsbad, California, USA) protocol. A digest
with on-column RNase-free DNase (QIAGEN, Valencia, California, USA) was used to remove the DNA. The RNA was purified and concentrated using a RNeasy column (QIAGEN).

Next generation Solexa sequencing of the RNA samples was performed at the National Center for Genome Resources (Santa Fe, New Mexico, USA). The Solexa software was used to analyze the 36 bp short-reads for image analysis, base-calling, quality filtering, and per base confidence scores. Sequences were aligned with GS Nap (Wu and Nacu 2010) using the default settings against the soybean genome sequence assembly. The short read alignments were filtered for alignment with no more than two mismatches or one indel. The uniquely mapped reads that passed our filtering criteria were normalized using the reads per kilobase per million method (Mortazavi et al. 2008; Nagalakshmi et al. 2008) where the length corresponded to length of the cDNA for the longest splice variant for a particular gene model. A summary of the short read alignments can be found in Table S1.2

Structural parameters of genes

To obtain our genomic data, we used the version of the soybean genome available from Soybase.org [accessed 25 August 2009]. The splice variants and gene models were as reported by Schmutz et al. (2010). For gene models with alternative splicing variants, we used the variant with the longest mRNA. To determine the intergenic region, we counted the distance between the end of the 3' region to the beginning of the next 5' region. Given two overlapping genes, we
counted the intergenic region to the left for gene A. If gene B overlapped and extended beyond gene A then the intergenic region to the right would be counted and annotated to gene B but not to gene A. If gene B was encompassed within gene A, we counted the intergenic region around gene A and annotated the distance to gene A and no intergenic region was counted for gene B.

To ensure that a tissue bias did not skew our results, we also ran the same analyses with data from Libault et al. (2010), a study independent from ours. We used their normalized data found at http://digbio.missouri.edu/soybean_atlas. All correlation analyses described in our study were also conducted on the Libault et al. dataset. Their tissues included root hairs isolated 12, 24, and 48 h after inoculation and mock inoculation and a sample of stripped roots from 3-day old seedlings. Other tissues included apical meristem, flower, green pods, leaves, nodules, root, and root tip.

**Gene expression**

After normalization, we removed genes that did not have a total transcript count of at least two. For our analysis, 43 353 gene models out of the 66 210 total gene models were considered expressed. The relationship between the various structural parameters and expression breadth and expression intensity were quantified using a Pearson’s parametric correlation (r).

Expression level was categorized as low (transcript count of at least two and no more than nine in a specific tissue), intermediate (transcript count
between ten and forty-nine), or high (transcript count of fifty or greater). The number of genes in each expression group at each level of expression breadth is included for both the Severin et al. (2010) analysis (Table S2) and the Libault et al. (2010) analysis (Table S3). Also included is the percent of pericentromeric genes in each group. Pericentromeric and euchromatic regions were determined by the coordinates found at SoyBase.org. If a coordinate was found in the middle of the gene, the gene was counted as a pericentromeric gene. All correlation analyses done specifically on pericentromeric or euchromatic genes were done as described for the Severin et al. dataset.

**GC content**

The data for the GC content and exon lengths were acquired from the Glyma1.gff file on SoyBase (Soybase.org) documented as the “gene models” track (Glyma 1.01 genome assembly). We analyzed the first, internal, and last exons separately. If exon length is different because of Atbiased mutations leading to a gain of stop codons and thus a shortening of the transcript, then a significant difference in GC content of those longer transcripts should be primarily found in the last exon (Xia et al. 2003). The average percent GC content was used for genes with multiple internal exons. To ensure that genes with no introns were not significantly biasing our data, we analyzed intronless genes separately. Taking out the intronless genes did not significantly affect the results. However, genes without introns were quite different than genes with introns and are included in the results. The Mann–Whitney U two-sample rank-sum test (Mann
and Whitney 1947) was performed on individual categories. We used the exact procedure because the asymptotic variant increases the probability of a type I error (Neuhaus 2005). The expression categories had various sizes with varying distributions, but the nonparametric Mann–Whitney U test will account for both factors and will generate a robust measurement provided there are sufficient sample sizes. Our sample sizes were consistent with the recommended size as noted by Siegel and Castellan (1988).

**Physical clustering**

The recently completed whole-genome sequence of soybean (Schmutz et al. 2010) allowed us to develop statistical parameters by modeling a random distribution of the number of genes in each of our categories by taking into account the actual distribution of the gene models in the genome. We then determined the actual position of each of the genes in our categories.

If the number of genes actually located in our designated bins exceeded the simulated data by three standard deviations, we considered the genes to be clustered. With this method, only gene models were used in the simulation. Clustering along the soybean genome was determined by an in-house script (Table S4) that was written in the programming language R (R Development Core Team 2005). The approximated pericentromeric and centromeric regions were based on positions found at Soybase.org. This clustering algorithm, by using gene models, accounted for variations in gene density.
Results

To capture genomic structural parameter differences associated with patterns of gene expression, we first allocated expressed genes into low, intermediate, or high expression intensity categories. The genes were further categorized based upon expression breadth, where the latter was the number of tissues in which they were considered expressed, from 1 to 14.

Genic and intergenic physical parameters

The total exon length (sum of the length of all exons in the gene) and the total intron length (the sum of the length of all introns in the gene) significantly increased across breadth of expression for genes in the low expression group (nine transcripts or less) \((r = 0.95, P < 0.01; \text{Fig. 1A} \text{ and } r = 0.93, P < 0.01; \text{Fig. 1B, respectively})\). Inversely, total exon and total intron length \((r = -0.88, P < 0.01; \text{Fig. 1A} \text{ and } r = -0.78, P < 0.01; \text{Fig. 1B, respectively})\) decreased across expression breadth in genes that were highly expressed (fifty transcripts and over). Those genes in the intermediate expression group (ten to forty-nine transcripts) showed no correlation between total exon length or total intron length and expression breadth \((r = -0.41, P < 0.05; \text{Fig. 1A} \text{ and } r = -0.36, P < 0.05; \text{Fig. 1B, respectively})\).

A previous study reported that highly expressed genes experience selective pressure for smaller transcripts (selection for economy hypothesis) (Li et al. 2007). Our data only partially supports that conclusion. We found that, for highly expressed genes, the total exon length and average exon length
significantly decreased as the number of tissues in which expression was detected became greater \((r = -0.88, P < 0.01; \text{Fig. 1A and } r = -0.71, P < 0.01; \text{Fig. 1C, respectively})\). Thus, our data only partially support the hypothesis as it was true only when the genes were expressed in multiple tissues, but the hypothesis did not hold true for tissue-specific genes. Interestingly, genes with a low expression level showed a significant increase in the total exon length and the average number of exons per gene \((r = 0.95, P < 0.05; \text{Fig. 1A and } r = 0.95, P < 0.05; \text{Fig. 1D, respectively})\) as they were expressed in more tissues. Genes expressed at the intermediate expression level showed no significant change in their total exon length, average exon length, or average number of exons across expression breadth (Figs. 1A, 1C, and 1D).

These results present several interesting observations. The increase in total exon length and total intron length across expression breadth for genes with low expression levels appears to simply be due to an increase in the number of exons and introns per gene, respectively (Figs. 1A, 1B, and 1D). On the other hand, because no change in the average number of exons is noted for genes expressed at high levels, the decrease in total exon length across expression breadth appears to result from an actual decrease in the size of the individual exons (Fig. 1C). Thus, highly expressed genes have smaller transcripts only when they are constitutively or near constitutively expressed, whereas genes with low levels of expression have smaller transcripts when they are tissue-specific or near-tissue-specific, but have vastly larger transcripts as they are constitutively
or nearly constitutively expressed.

The mean distance between intermediately and highly expressed genes decreased significantly as those genes became more broadly expressed in all tissues ($r = -0.81$, $P < 0.01$ and $r = -0.71$, $P < 0.01$, respectively, Fig. 1E). But genes with low levels of expression showed no significant change in intergenic length across expression breadth (Fig. 1E). Intergenic distances were not detectably different among expression categories when the genes were tissue or near tissue-specific (data not shown). Thus, distance between genes is only significantly different for intermediate and highly expressed genes, and only when they are broadly expressed.

To avoid possible tissue bias, we repeated the same analysis with next-generation sequencing data generated by Libault et al. (2010) from 14 separate samples taken from flower, pods, leaves, nodules, root tip, root, root hair, and apical meristem. Although significance values varied slightly, we found consistent results with the Libault et al. (2010) data with variation only in the intermediate expression range. For genes expressed at the low level, an increase in expression breadth correlated with an increase in total exon length ($r = 0.73$, $P < 0.01$; Figure S1A) and total intron length ($r = 0.96$, $P < 0.001$; Figure S1B). Highly expressed genes decreased in total exon length ($r = -0.84$, $P < 0.001$; Figure S1A) and total intron length ($r = -0.78$, $P < 0.001$; Figure S1B) as the expression breadth increased. At the low expression level, the number of exons positively correlated with expression breadth ($r = 0.89$, $P < 0.001$; Figure S1C).
while at the high expression level the number of exons negatively correlated with expression breadth \( r = -0.90, P < 0.001 \); Figure S1C) similar to that found in our analysis. For all expression levels, the average exon size significantly decreased in the Libault et al. samples: low expression \( r = -0.81, P < 0.001 \); Figure S1D), intermediate expression \( r = -0.57, P < 0.05 \); Figure S1D), and high expression \( r = -0.57, P < 0.05 \); Figure S1D). A significant decrease in intergenic length was only apparent at the highest expression level \( r = 0.56, P < 0.05 \); Figure S1E).

It is known that the pericentromeric regions of chromosomes have reduced recombination rates. Soybean is unusual in that it has extremely large pericentromeric regions (Schmutz et al. 2010) and to eliminate bias from this region, we performed the same analyses using genes from only the euchromatic region and then from only the pericentromeric region. Interestingly, when the genes in the euchromatic region were analyzed alone, the correlations between expression characteristics and genic parameters in the two data sets were remarkably similar (Table 1) and were consistent with our previous results. Fewer significant correlations were found in genes within the pericentromeric regions (Table S5).

**GC content**

To determine if differences in GC content were associated with patterns of gene expression, we divided all expressed genes into a set of 2 × 2 factorial categories. The four categories included (i) genes that had low expression levels
(one to nine transcripts) and also were narrowly expressed (one tissue only), (ii) genes that had low expression levels and also were broadly expressed (expressed in all 14 tissues), (iii) genes that had high expression levels (fifty or more transcripts) and also were narrowly expressed, and (iv) genes that had high expression levels and also were broadly expressed. We made comparisons in all directions. When considering GC content we examined the first exon in a gene, the average GC content of all internal exons, and the GC content of the last exon in a gene. We also considered the average GC content of all genes in each category with no introns independently (Figs. 2A, 2B, 2C, and 2D).

With a single exception, broadly expressed genes had a significantly higher GC content in all exons (start, middle, and end) than did narrowly expressed genes (Figs. 2A, 2B, and 2C). The one exception, genes with low expression (nine or fewer transcripts), for which the average GC content of all middle exons was significantly greater for narrowly expressed genes than for broadly expressed genes. In all comparisons GC content of exons of highly expressed genes were greater than that of genes with low expression. This was observed regardless of whether genes were broadly expressed or narrowly expressed. The same pattern was seen in GC content of genes with no introns; broadly expressed genes had a higher GC content than narrowly expressed genes, and highly expressed genes had a higher GC content than genes with low expression (Fig. 2D).
Physical clustering

To examine whether genes exhibiting unique patterns of expression were physically clustered along the genome, we again considered the four categories described above. We established our statistical parameters by a modeling process that randomly placed the number of genes in each of our categories into annotated gene models as they were distributed across the genome. This enabled us to take into account differences in gene density within the genome. We conducted 1000 iterations of this process and then evaluated the results from genomic bins of 100 Kb. We then determined the actual positions of each of the genes in our categories. If the number of genes located in our designated bins exceeded the simulated data by three standard deviations, we considered the genes to be clustered.

More clusters of tissue-specific narrowly expressed genes were found than constitutively broadly expressed genes regardless of expression level (high or low) (Table S4). Approximately two-times the percentage of the total number of tissue-specific, broadly expressed genes expressed at high levels were found in clusters than in any other category (13.4% vs. 5.8% – 6.4%). On the other hand, broadly expressed genes expressed at low levels had more genes per cluster in the pericentromeric region than any other category (4.5 vs. 2 – 2.9) (Table 2). Approximately a quarter of the clustered genes in each category were found in pericentromeric regions (24% – 29%).

Very few (eight) genes that were expressed in all 14 tissues and also
expressed at high levels were clustered. However, the total number of genes in this category was low and the number of clustered genes was proportional to the total number for the other categories. The physical distribution of clusters of genes across all 20 chromosomes in 100 kb bins is depicted in Figure S2. A numerical representation of this distribution is shown in Table S4.

**Discussion**

The soybean genome has been largely unexamined with regard to relationships of gene expression profiles to gene structural parameters. Our goal was to identify possible associations between transcriptional expression of genes and physical parameters of the genes and their environs. We used a whole-genome RNA-Seq analysis of soybean tissues sampled during a progression of seed developmental time points and various tissues. We found unique patterns relative to changes in physical parameters of genes with regard to expression breadth and expression intensity. At high levels of expression, introns and exons were smaller as genes were expressed across a larger number of tissues. However, this relationship was reversed in genes expressed at a low level. At low levels of expression, introns and exons were larger as genes were expressed in a larger number of tissues. In the intermediate to high expression intensity ranges, intergenic regions were smaller when expression breadth was examined on a narrow basis.

Our results provide evidence that could support previous seemingly contradictory findings. In an Arabidopsis and rice study, sequence expansion
was correlated with an increase in expression intensity (Ren et al. 2006) whereas conversely, in humans an increase in expression intensity was correlated with sequence contraction (Urrutia and Hurst 2003). In another Arabidopsis study, sequence contraction was correlated with an increase in expression intensity, whereas sequence expansion was correlated with an increase in expression breadth (Camiolo et al. 2009). The question then is how does one interconnect these contradictory findings in a meaningful manner? The answer seems to lie in a joint consideration of both intensity and breadth of gene expression. Our findings show that the total length of the exons or introns within the gene “increases” as expression breadth increases at low expression intensities, but at high intensities of expression, the relationship flips, such that total intron and exon length of the gene “decreases” as expression breadth increases. This “flip” is also evident in the average number of exons observed across expression breadth.

Alternative splicing may explain the different characteristics of genes at low levels of expression. Camiolo et al. (2009) proposed that exon–exon junction complexes, when placed on mRNA during splicing, imposed a post transcriptional effect which then promoted an increase in the size of the transcript and the efficiency of translation (Le Hir and Seraphin 2008). Alternative splicing can yield isoforms that are broken down by non-sense mediated decay or other such mechanisms, which subsequently decrease the transcript count (Hillman et al. 2004). In our study, genes that fit into the lowest expression groups were
found to have several other common features of polytypic genes. In humans, polytypic genes were found to have more exons and larger transcripts when expressed at a lower level (Wegmann et al. 2008). In agreement with that study, our study showed that genes expressed at a low level in more tissues were found to have more exons and larger transcripts than any of the other expression categories. It is possible that those genes expressed at a low level are less inclined to be pressured by a selection for economy and are instead influenced by the demands of being polytypic.

Alternatively, at high levels of expression, it may be that codon usage bias is affecting the physical properties of the genes. GC content and codon bias in Physcomitrella patens were found to be highly correlated (Stenøien 2007). In another study, 13 species had a correlation between GC content and many of the physical properties of the genes (Zhu et al. 2009). In our study, genes with a higher expression intensity and a larger expression breadth had the largest GC content (Figs. 2A, 2B, 2C, and 2D). Seoighe et al. (2005) found that Arabidopsis genes with high expression levels also had a higher GC content, which is consistent with our results. In rice, there was a stronger selective constraint on codon usage in highly expressed housekeeping genes compared to highly expressed tissue-specific genes determined by a lower synonymous substitution rate (Mukhopadhyay et al. 2008). In the same study, weakly expressed genes did not show a significant difference in synonymous substitution rate across expression breadth suggesting that a codon usage bias is not a likely cause of
structural differences in the genes expressed at a low level. In our study, GC content increased across breadth at all expression levels, but the rate of increase was much higher in those genes expressed at a high level. The increase in GC content across expression breadth at a high expression level was approximately three times greater than the increase in GC content across expression breadth in those genes expressed at the low level. The increase in the percentage of GC content across expression level in broadly expressed genes was more than two times greater than the increase across expression level in narrowly expressed genes. At high intensities of expression, a decrease in the length of exons and introns rather than in the number of exons and introns was apparent, also suggesting a codon usage bias, which was similar to that found in Gramineae genes (Guo et al. 2007). As the hypothesis for selection for economy suggests, genes with large transcriptional demands are prone to selection for miniaturization (Li et al. 2007) and codon usage bias might be the means by which the genome achieves this. Meanwhile, genes expressed at a low level are not as transcriptionally demanding, even when expressed in all tissues, making selection for the miniaturization in these genes economically irrelevant. Thus, as high percentage of GC content is positively correlated with high levels of expression in many tissues it could be suggested that a relationship occurs between expression characteristics and codon usage bias.

The bendability of the DNA and B-Z transitions, both of which promote open chromatin and active transcription, increases in GC-rich areas of the
genome. Therefore, those genes with a higher GC content are more likely to be heavily transcribed (Vinogradov 2003). The GC content in conjunction with the intronic and intergenic sequences could be manipulated by epigenetic factors in an effort to perfect chromatin-mediated suppression of tissue-specific genes and regulation of expression level (Vinogradov 2004). If the physical properties of genes are based upon the genomic complexity hypothesis highly and broadly expressed genes should be located in regions of open chromatin at a higher frequency than those not so expressed (Vinogradov 2003). Conversely, genes that are less intensively and narrowly expressed would be expected to be found in the condensed chromatin regions. In this study, we found that genes expressed at the low intensities with a narrow expression breadth had a larger number of genes in clusters in the approximated pericentromeric regions: regions with more compact chromatin than genes expressed at high levels with a narrow expression breadth. Also, genes expressed at the low intensities with a broad expression breadth had more genes in clusters in the approximated pericentromeric regions than genes expressed at high intensities with a broad expression breadth. We found that genes expressed at the highest expression level in only one tissue had twice the percentage of genes in clusters than any other expression group. Identifying the correlations between the physical parameters of the individual genes, the possible role of regulation owing to nucleotide composition, and the regulatory effects of chromatin structure could have potential impacts in identifying the role of the noncoding regions. In this
study, we found significant clustering and regions with multiple clusters of the same expression category that indicates possible clustering domains.

The differences in the physical parameters of genes within euchromatic regions compared with genes within pericentromeric regions, as related to expression variables suggest that the lack of recombination or chromatin structure has a strong effect on the mechanisms giving rise to the characteristics we observed in this study. As results of analyses of both datasets were consistent in the euchromatic regions, the implication is that a common mechanism is functioning in these regions. It is possible that genes in the pericentromeric regions are under different evolutionary constraints.

It is apparent that the structural parameters of plant genes are determined by more interacting forces than any single hypothesis so far proposed. The effect of expression breadth on genic size is dependent on the effect of expression intensity. This study has provided evidence in support of both the selection for economy and the genomic organization hypotheses (Vinogradov 2004). Further analysis into the effects of splicing events and codon usage bias could provide more insight into additional fine-tuning of gene regulatory networks.

**Literature cited**


Figure 1A  Correlation between expression breadth and total exon length. The mean is shown for each level of tissue specificity in each expression category.
Figure 1B. Correlation between expression breadth and total intron length. The mean is shown for each level of tissue specificity in each expression category.
Figure 1C. Correlation between expression breadth and average exon length. The mean is shown for each level of tissue specificity in each expression category.
Figure 1D. Correlation between expression breadth and average number of exons. The mean is shown for each level of tissue specificity in each expression category.
Figure 1E. Correlation between expression breadth and intergenic length. The mean is shown for each level of tissue specificity in each expression category.
Figure 2A. Change in percent gc content across expression level and breadth categories for the first exon. The significance levels are shown between comparisons.
Figure 2B. Change in percent gc content across expression level and breadth categories for the internal exons. The significance levels are shown between comparisons.
Figure 2C. Change in percent gc content across expression level and breadth categories for the last exon. The significance levels are shown between comparisons.
Figure 2D. Change in percent gc content across expression level and breadth categories for genes with no introns. The significance levels are shown between comparisons.
Table 1. Correlations between expression breadth and physical parameters for genes in the euchromatic regions for both Severin et al. (2010) and Libault et al. (2010).

<table>
<thead>
<tr>
<th></th>
<th>Low expression&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Intermediate expression&lt;sup&gt;b&lt;/sup&gt;</th>
<th>High expression&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total exon length</td>
<td>0.74**&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.98***</td>
<td>-0.86***</td>
</tr>
<tr>
<td></td>
<td>0.74**&lt;sup&gt;e&lt;/sup&gt;</td>
<td>0.98***</td>
<td>-0.80***</td>
</tr>
<tr>
<td>Total intron length</td>
<td>0.95***</td>
<td>0.94***</td>
<td>-0.79***</td>
</tr>
<tr>
<td></td>
<td>0.96***</td>
<td>0.94***</td>
<td>-0.90***</td>
</tr>
<tr>
<td>Number of exons</td>
<td>0.91***</td>
<td>0.96***</td>
<td>-0.64*</td>
</tr>
<tr>
<td></td>
<td>0.91***</td>
<td>0.96***</td>
<td>-0.84***</td>
</tr>
<tr>
<td>Average exon length</td>
<td>-0.80***</td>
<td>-0.54*</td>
<td>-0.83***</td>
</tr>
<tr>
<td></td>
<td>-0.80***</td>
<td>-0.54*</td>
<td>-0.53*</td>
</tr>
<tr>
<td>Length of intergenic region</td>
<td>-0.85***</td>
<td>-0.71**</td>
<td>-0.77***</td>
</tr>
<tr>
<td></td>
<td>-0.85***</td>
<td>-0.71**</td>
<td>-0.60*</td>
</tr>
</tbody>
</table>

<sup>a</sup>Genes expressed with a transcript count of nine or under.  
<sup>b</sup>Genes expressed with a transcript count of ten to forty-nine.  
<sup>c</sup>Genes expressed with a transcript count of fifty or over.  
<sup>d</sup>Correlations for Severin et al. (2010).  
<sup>e</sup>Correlations for Libault et al. (2010).
Table 2. Physical clustering of genes in four expression categories.

<table>
<thead>
<tr>
<th></th>
<th>Narrow/Lo&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Broad/Lo&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Narrow/Hi&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Broad/Hi&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes clustered</td>
<td>345</td>
<td>367</td>
<td>262</td>
<td>8</td>
</tr>
<tr>
<td>Number of genes clustered/peri centromeric&lt;sup&gt;e&lt;/sup&gt;</td>
<td>87</td>
<td>59</td>
<td>59</td>
<td>2</td>
</tr>
<tr>
<td>Percent of genes in approximated pericentromeric</td>
<td>29</td>
<td>26</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>Average number of genes in each cluster</td>
<td>2.4</td>
<td>2.5</td>
<td>2.2</td>
<td>2</td>
</tr>
<tr>
<td>Average number of genes in each cluster/peri centromeric&lt;sup&gt;f&lt;/sup&gt;</td>
<td>2.9</td>
<td>4.5</td>
<td>2.4</td>
<td>2</td>
</tr>
<tr>
<td>Percentage of total genes that are in a cluster</td>
<td>5.8</td>
<td>6.0</td>
<td>13.4</td>
<td>6.4</td>
</tr>
</tbody>
</table>

<sup>a</sup> Number of genes expressed in one tissue with a transcript count of nine or under.<br>
<sup>b</sup> Number of genes expressed in fourteen tissues with a transcript count of nine or under.<br>
<sup>c</sup> Number of genes expressed in one tissue with a transcript count of fifty or over.<br>
<sup>d</sup> Number of genes expressed in fourteen tissues with a transcript count of fifty or over.<br>
<sup>e</sup> The number of genes clustered in the approximated pericentromeric region.<br>
<sup>f</sup> The average number of genes in a cluster in the pericentromeric region.
CHAPTER 3. GENE EXPRESSION: SIZING IT ALL UP


Jenna L. Woody\(^1\) and Randy C. Shoemaker\(^2\)

Abstract

Genomic architecture appears to be a largely unexplored component of gene expression. That architecture can be related to chromatin domains, transposable element neighborhoods, epigenetic modifications of the genome, and more. Although surely not the end of the story, we are learning that when it comes to gene expression, size is also important. We have been surprised to find that certain patterns of expression, tissue specific versus constitutive, or high expression versus low expression, are often associated with physical attributes of the gene and genome. Multiple studies have shown an inverse relationship between gene expression patterns and various physical parameters of the genome such as intron size, exon size, intron number, and size of intergenic regions. An increase in expression level and breadth often correlates with a decrease in the size of physical attributes of the gene. Three models have been proposed to explain these relationships. Contradictory results were found in several organisms when expression level and expression breadth were analyzed.

\(^1\) Department of Agronomy, Iowa State University, Ames, IA 50011

\(^2\) United States Department of Agriculture-Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011
independently. However, when both factors were combined in a single study a novel relationship was revealed. At low levels of expression, an increase in expression breadth correlated with an increase in genic, intergenic, and intragenic sizes. Contrastingly, at high levels of expression, an increase in expression breadth inversely correlated with the size of the gene. In this article we explore the several hypotheses regarding genome physical parameters and gene expression.

**Introduction**

Ever since Beadle and Tatum conducted simple but elegant experiments that led to a basic understanding that “genes act by regulating definite chemical events” (Beadle and Taum, 1941) we have known that mutations can influence the fate of an organism. This profound finding led to their receiving the Nobel Prize in Physiology or Medicine in 1958. We now know that the regulation of expression of genes is more complex. Expression is no longer thought to be controlled solely by the “strength” of a promoter, but is modulated by transcription factors, small RNAs, parachromatin, as well as by all of the components that make up epigenetics (Jorgensen, 2011).

Identifying the internal cues that regulate gene expression can help in deciphering the form and function of living organisms. With the surge in whole-genome sequencing, exploring the uncharted territories and complex evolutionary constraints is now possible. Until recently, genic properties such as exon size and intron size have been assumed to evolve under stochastic
processes. In the last ten years, a correlation between transcriptional demands and genic properties has been identified. Each gene has an individual profile varying in the level of transcription and the number of tissues in which it is expressed. As the transcriptional demands of a gene increase, the genic size tends to decrease. Proposals for the explanation of this relationship have focused on selection for economy, a regional mutational bias or genomic organization. While this relationship is seemingly constant in animals, in plants many contradicting results have been found (Ren et al. 2006, Camiolo et al. 2010, Yang et al. 2010, Woody et al. 2011). It is apparent that different selective forces are acting on the plant genomes than what has been previously thought.

**The models**

“Selection for economy” proponents base their argument on the fact that transcription and translation are both time-consuming and costly (Urrutia and Hurst 2003; Seoighe et al 2005). To transcribe one nucleotide, two adenine triphosphate molecules and roughly 0.05 seconds are required (Carmel and Koonin 2009) thus it would be advantageous to the organism to reduce the cost of those genes ubiquitously and highly transcribed and translated. As might be apparent, within the selection for economy argument, there are two sub-arguments; the energetic cost hypothesis and the time cost hypothesis. The energetic cost hypothesis states that selection is influenced by a drive to minimize the energetic cost of transcription. Alternatively, in the time cost hypothesis, shorter introns and shorter exons are selected when limited time
periods are required to transcribe large amounts of mRNA (Rao et al. 2010). The common thread is that the decrease in genic size is a result of selected mutations with the purpose to decrease the demands of highly transcribed genes.

If indeed, selected mutations occur that result in decreased gene sizes and increased transcription one has to wonder when and how does this take place. Selection for gene reduction based on economical reasons can occur at two stages, transcription and translation. An equal decrease in intron and exon size would suggest selection is occurring at the transcription stage while a decrease solely in the exon size would point to selection at the point of translation. To make this even more complex, selection could be occurring at both stages. For this reason, it appears that there are two facets to the argument for selection for economy, is it occurring and if so, is it in transcription or translation?

While the selection for economy hypothesis is reasonable, it does not explain the shortening of non-coding regions in genes that are highly and/or broadly transcribed. Vinogradov (2004) suggested that broadly expressed genes required simple regulation and therefore less regulatory elements. Conversely, tissue-specific genes contain more functional domains and are associated with more complex protein architecture (Vinogradov 2004) resulting in larger gene “spaces”. The genome complexity model postulates that the functional properties of a gene determine the length of the physical genic properties (Eisenberg and Levanon, 2003; Vinogradov 2006). Intron and intergenic regions are
hypothesized to be involved in chromatin-mediated suppression and higher order regulation thus introns and intergenic regions are increased when genes are transcribed at a low level or in a tissue-specific manner.

The mutational model focuses on transcription-associated non-adaptive deletion bias, the idea that highly expressed regions are in chromosomal regions with high deletion rates resulting in the bias (Urrutia and Hurst, 2003; Comeron, 2004). The selection for economy model and the mutational bias model share a lot of overlap but the underlying concept is different for the two. The selection for economy model refers to the strain an individual gene’s transcription and translation puts on the cell. At a larger level, the mutational bias model suggests that the “neighborhood” of the gene is the cause for selection. Highly expressed genes tend to cluster in the chromosomes (Caron et al., 2001) and it is hypothesized that this clustering might result in local mutational bias.

Eukaryotic genomes are composed with a myriad of distinct regions of varying GC content. Genomic regions containing many genes tend to be GC rich (Urrutia and Hurst, 2003) and thus are also regions of high recombination rates (Fullerton et al., 2001). It is possible that the increase in recombination imposes a mutational bias on these highly expressed genes (Seoighe et al., 2005). However, the mutational bias model has also been suggested at the individual gene level. As a gene is transcribed more it is more disposed to retroposition and reverse transcription (Mourrier and Jeffares, 2003).
In chicken, (Rao et al., 2010) gene size, CDS length, first intron length, average intron length and total intron length are negatively correlated with expression level and expression breadth. In humans, (Eisenberg and Levanon, 2003) 575 constitutively expressed genes were analyzed and were found to have shorter introns, untranslated regions and coding sequences than tissue specific genes. These studies add support to the selection for economy model as the regions that are transcribed are decreasing in size as expression increased. They also found that the difference in genic size between tissue specific and housekeeping genes was larger for the introns than for the exons and proposed that the coding sequences and UTRs would be less susceptible to change based on selection. Another study in humans and Caenorhabditis elegans identified a significant decrease in the intron size of highly expressed genes and this decrease was much larger than the decrease in coding region size suggesting that the reduction is not functional but a result of natural selection (Castillo-Davis et al., 2002).

It is readily apparent that the models allow for conceptual overlap. A reduction in intron size could also support the genome complexity model. An increase in expression correlates with a decrease in regulatory elements and thus a decrease in intron and intergenic size according to the model. However, Li et al. (2007) analyzed genes with high functional/regulatory complexity in M. musculus, human and Arabidopsis thaliana and found that these genes did not have longer introns or longer proteins. In addition, they did not find that house-
keeping genes were more compact than tissue specific genes expressed at similar expression levels. And so, the controversy grows.

**The “controversy”**

A controversy has emerged regarding expression and the structure of plant genomes. In a contradiction to the models, Ren et al. (2006) studied both *Oryza sativa* and *A. thaliana* and found that highly expressed genes contained more and longer introns and a produced a larger primary transcript than genes expressed at a low level. The genic parameters also increased as the expression breadth increased which is different than what had been found in animals. However, in a subsequent study in *A. thaliana* both the noncoding and coding regions of the genes decreased as the expression level increased (Camiolo et al., 2009).

In accordance with the previous study, another study in *Arabidopsis* found that expression breadth positively correlated with the non-coding structural parameters (Yang, 2009), e.g. non-coding regions got larger as expression breadth increased. However, in the same study expression breadth was negatively correlated with the coding regions, e.g. coding regions got smaller. It is possible that plant genomes are under a different selection pressure than animals and that different methods are needed to decipher the evolutionary process.

Using a “primitive” plant, Stenoien (2007) studied the possible effect of selection on genome organization in the haploid moss *Physcomitrella patens.*
They found that total intron length, the number of introns, and the total length of genes are negatively correlated with the level of expression. They suggest that if animals and plants have followed separate evolutionary pathways then this difference must have occurred after the split between vascular and nonvascular plants (250 mya, Palmer et al., 2004). One suggested explanation for this difference is that plants tend to have much smaller introns. *Arabidopsis* has an average intron length per gene of 152bp, 387 bp in rice (Ren et al., 2006) compared to 5.5 kbp in humans (Sakharkar et al., 2004). A much larger transcriptional demand on the introns of humans seems plausible. However, *P. patens*’ average intron length is 252 bp, not significantly different from *Arabidopsis* and smaller than rice (Rensing et al., 2005). Subsequent expression studies done in *Arabidopsis* and other plant species revealed different results. Colinas et al. (2008) found that the size of the introns and exons negatively correlated with expression levels. This seemingly nullified the argument that vascular and nonvascular plants are evolving under different constraints.

Interestingly, it is not just in plants that opposing correlations have been discovered. In several yeasts and other unicellular organisms, highly expressed genes have longer introns than genes expressed at a low level (Vinogradov, 2001). In the unicellular green algae *Ostreococcus lucimarinus*, intron number and intron density are positively correlated with expression level (Lanier et al., 2008). Even in animals, as in the mouse example above, controversy has occurred. In chicken, ubiquitously expressed genes were compared with narrowly
expressed genes and they found that ubiquitously expressed genes were larger (Rao et al., 2010). However, they found that gene size, CDS length, first intron length, average intron length and total intron length all negatively correlated with expression level. Throughout the dispute, it is unclear as to whether the source of the contradictions is expression level or expression breadth.

An important consideration when evaluating the contradictions is the quantification and characterization of expression and genic properties both within and across species. Can an ancient polyploid with a large genome such as soybean be compared to a genome such as rice? Both have experienced dramatically different evolutionary trajectories. Can the evolutionary processes of plants be analyzed and compared with animals? Even within a species experiments vary. Expression breadth is relative to the tissue and time points analyzed in the study. This is not to say that we cannot compare across studies but this should be contemplated when making generalizations. A similar conflict occurs when analyzing genic properties. Each individual property (exon length, intron length, intergenic region, individual exon lengths) can tell us a different story to complement the fluid movements of the whole gene. Understanding the evolutionary differences between intron and exon length can give us a wealth of information on what may be occurring during transcription compared to translation.
A novel dichotomy in highly expressed genes compared to lowly expressed genes

A recent study in soybean took a unique approach and partitioned the genes first into categories of expression level (low, mid, high) and then into categories of expression breadth (Woody et al. 2011). A unique division was observed; genes that were expressed at high levels decreased in size as the expression breadth increased while genes that were expressed at low levels increased in size as the expression breadth increased. This lead to the hypothesis that multiple divergent evolutionary paths may be present. Those genes at a low level of expression may be under a different model of selection than those at a high level of expression. In humans, Zhu et al. (2008) looked at 17,288 RefSeq loci across 18 tissues and found that, on average, highly expressed genes are more compact but that genes expressed at a low level show a lot more variation. They suggested that highly expressed genes could be the only genes under an economical selection pressure (selection for economy). In *Arabidopsis* and rice, it was found that housekeeping genes, compared to tissue-specific genes, are under stronger selective constraints and that weakly expressed genes, compared to highly expressed genes also are under stronger selective constraints (Mukhopadhyay et al., 2008). When analyzed further they found that highly expressed housekeeping genes had a lower synonymous substitution rate than lowly expressed housekeeping genes. Berg and Martelius (1995) suggested that a lower synonymous substitution rate was due to a
transcriptional selection for economy. Mukhopadhyay et al. (2008) found that by analyzing preferred codon usage, highly expressed genes that were broadly expressed were under selection for economy through tRNA copy number that was used to optimize the synonymous codon usage. Lowly expressed genes are under a stronger selective pressure than highly expressed genes but highly expressed housekeeping genes are also under a selective pressure and this can be localized to a codon usage bias.

Selection for economy may explain the evolution of highly expressed genes but other selective forces, potentially stronger forces, are acting upon weakly expressed genes. This selection appears to increase as the expression breadth increases. In Woody et al. (2011) it was observed that tissue specific genes did not display a large difference in genic size between low, mid and highly expressed genes, although the physical parameters of highly expressed tissue specific genes were always slightly larger than lowly expressed tissue specific genes. It was postulated that the genes expressed at a low level of expression are selected upon by the demands of being polytypic (genes involved in alternative splicing events). Genes that are lowly expressed, with an increasing breadth of expression share many properties with polytypic genes. Genes expressed at a low level increased in total genic length by increasing the number of exons, not the size of exons and this is dissimilar to highly expressed genes. In humans, an increase in exons and larger transcripts were shown to correlate with polytypic genes expressed at a low level.
What properties of alternative splicing lead to a selection for an increase in exon number? Exon-exon junction complexes are placed on mRNAs during splicing. These complexes result in a post-transcriptional effect in that the size of the transcript and the efficiency of translation are both increased (Camiolo et al. 2009). In a previous study on alternative isoforms in humans, it was found that many gene isoforms of alternative splice genes contained premature termination codons and were subject to non-sense mediated decay and subsequently decreases the transcription level (Hillman et al. 2004). Thus, a selection for economy could be suggested in the highly expressed genes but the lowly expressed genes have a different method of evolutionary selection that possibly rises from the demands of being polytypic.

**Selection on the individual gene or on an entire region?**

If weakly expressed genes evolve under the umbrella of alternative splicing demands, it would appear evident that selection would be at an individual level. However, if nature was selecting for an economical purpose, it is reasonable to question whether entire neighborhoods are under specific selection. Clustering of highly expressed genes has been established and several physical genomic properties have been associated with these regions. In a study that combines transposable elements, gene length, and gene expression Jijingo et al. (2011) found that all three of those factors are closely related. Combined together, transposable elements and gene length account for 78% of the variation in expression level, 76% of the variation in expression breadth, and
66% of the variation in tissue specificity. The authors proposed a role for selection for economy but suggested that the removal of transposable elements may be a stronger mechanism of selection than reduction of gene length. In a study done in rice (Tian et al., 2009) retrotransposons, genetic recombination, and gene density were all correlated and they suggested this relationship helped shape the makeup of the rice genome.

In rice, transposable element families were found to be differentially distributed across the genomes in areas of varying methylation patterns (Takata et al., 2007). Kim et al. (2004) found that the expression breadth of a gene is highly correlated with Alu elements and expression level is highly correlated with L1 densities in human. Confirmed by Eller et al. (2007), highly and broadly expressed genes are enriched with Alu elements and depleted in L1 elements. This suggests that rather than gene expression or transposable element insertion accounting for a variation in genic level, epigenetics may be influencing the entire genetic region. Isochores, large regions within the genome that are homogeneous in their GC content have been characterized and analyzed since 1976 (Macaya et al. 1976). Gene density, gene expression, insertion of transposable elements and density of transposable elements are only a few of the basic biological properties associated with isochores (Bernardi 2004). It is possible that these properties act as a unit and isochores are the homes for these interactions.
If different gene sizes and transposable element densities change across the isochore families and these properties have a large influence on expression, it follows that expression profiles are also influenced by these homogeneous structures. Two questions would arise if this was the case: what is the relationship between these characteristics in the homogeneous regions and do heterogeneous regions have different sets of characteristics with their own distinguishing features. This brings us back to the cost of transcription and translation, the nucleosome formation potential, related to homogeneity and heterogeneity, could influence both the chromatin domain and the size of the gene.

Another variable to consider when studying the evolution of individual components and their relationship with expression level at a whole-genome level is replication timing. Replication timing and expression profiles do not directly influence each other but both seem to be regulated through a mediator (Gilbert, 2002; MacApline and Bell, 2005; Gilbert and Gasser, 2006; Hiratani et al., 2008; Farkash-Amar and Simon, 2009; Schwaiger et al., 2009; Ryba et al., 2010). There are two main stages in replication, early and late. If a replication domain changes timing, the chromatin state usually changes and transcriptional activation or suppression usually follows. Replication timing correlates with isochore structure as well suggesting overarching domains.

Could chromatin domains be the top order of regulation? Chromatin domains have been well studied in many higher eukaryotes although Arabidopsis
is the only plant with extensive research done. Replication domains in *Arabidopsis* are correlated with chromatin conformation and sequence content (Lee et al., 2010). Co-expression can be coordinated by sharing a promoter in neighboring genes. However, co-expressed domains at large distances have also been identified (Chen et al. 2010). It is known that epigenetics helps regulate transcription but it’s effects in whole-genome view are still unclear. Are the replication domains determining the chromatin domains which in turn regulate gene expression? Does the sequence composition, the isochores family, enrich these determinants or are they the determinants for the replication domain?

A circular debate seems inevitable if we try to account for the actions of one biological property such as gene size acting on another property such as presence or absence of transposable elements. It is becoming clear that we need to consider gene expression in a more holistic manner. A complex array of neighborhoods appears to be covering the genome. Jorgenson (2011) described the genome as comprised of two types of chromatin, “orthochromatin” which is the stable, constant function of the chromatin and “parachromatin”, a dynamic and reactive chromatin. Parachromatin could provide a large but dynamic and flexible cloud over the active properties within the genome. Each element, transcriptional demands, transposable element insertion, small RNAs, etc. impact the other but survival is not possible unless the elements are fit to live under the epigenetic cloud.
Literature cited


CHAPTER 4: HOMOGENEOUS REGIONS (ISOCHORES) IN SOYBEAN

(GLYCINE MAX (L.) MERR.)

J. L. Woody¹, W. Beavis², and R. C. Shoemaker³

Accepted with moderate revisions in the journal Frontiers in Plant Genetics and Genomics

Abstract

The landscape of plant genomes, while slowly being characterized and defined, is still composed primarily of regions of undefined function. Many eukaryotic genomes contain isochore regions, mosaics of homogeneous GC content that can abruptly change from one neighboring isochore to the next. Isochores are broken into families which are characterized by their GC levels. We identified 4,339 compositionally distinct domains and 331 of these were identified as Long Homogeneous Genome Regions (LHGRs). We assigned these to four families based on finite mixture models of GC content. We then characterized each family with respect to exon length, gene content, and transposeable elements. The LHGR pattern of soybeans is unique in that while the majority of the genes within LHGRs are found within a single LHGR family with a narrow GC-range (Family B), that family is not the highest in GC content as seen in vertebrates and invertebrates. Instead Family B has a mean GC content of 35%.

The range of GC content for all LHGRs is 16-59% GC which is a larger range

¹ Interdepartmental Genetics Program, Iowa State University, Ames, IA 50011.
² Department of Agronomy, Iowa State University, Ames, IA 50011.
³ United States Department of Agriculture, Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011.
than what is typical of vertebrates. This is the first study in which LHGRs have been identified in soybeans and the functions of the genes within the LHGRs have been analyzed.

Introduction

The genomes of living organisms are often organized into unique patterns, the purposes of which are mostly unknown. It has been reported that at least some eukaryotic genomes contain isochore regions, mosaics of homogeneous GC content that abruptly change from one neighboring isochore to the next. In vertebrates, isochore regions have been defined as segments of DNA, typically above 300 kilobases (kb), that are homogeneous (AT- or GC-rich) with sharp boundaries from the neighboring stretches of DNA (Constantini et al., 2009).

Isochores were first observed using ultracentrifugation in CsCl density gradients (Macaya et al., 1976). DNA fractionation by ultracentrifugation, cytogenetic analyses, and recently, analyses of genes and genome sequences, have been utilized to identify these regions. With advances in technology and the availability of whole genome sequences isochores can now be identified with more precision using computational tools. Initially, sliding-window-based methods were used but these techniques can only determine isochores based on the window size used. Surprisingly, fundamental biological properties have been found to be associated with certain isochore families. Repeat sequence distribution, gene density, replication timing, CpG distribution, genic size and transcript abundance are several of the main features found to associate with
isochores (Bernardi, 2004). In vertebrates, these regions have been mapped and named Long Homogenous Genome Regions (LHGRs) (Zhang et al., 2010). Recently, LHGRs also have been identified in invertebrates (Cammarano et al., 2009). LHGR GC content is strongly conserved among invertebrate species (Cammarano et al., 2009) as well as among vertebrate species (Costantini et al., 2009). However, vertebrates and invertebrates differ in the GC content of their LHGR families. The existence of LHGRs across two kingdoms suggests that all metazoan genomes may contain LHGRs and suggests a biological or evolutionary importance, yet their function remains unknown.

LHGRs are classified into a number of families based on the frequency distributions of LHGRs across GC percentage (Macaya et al., 1995). Often, members of LHGR families will share similar additional biological features such as frequency of transposeable elements (Bernardi, 2000). Multiple peaks in the distribution of content are evident in most species although the demarcations between one LHGR family and the adjacent families have been based on ad hoc decisions. For example, the mid-point between two peaks has been arbitrarily considered a threshold, where one family ends and the next family begins. A different technique was applied in Arabidopsis (Zhang and Zhang, 2004) where the LHGRs were identified as GC-rich (GC content of the LHGR was above the average GC content for the chromosome in which it resided) and AT-rich (GC content of the LHGR was below the average GC content for the chromosome in which it resided) because no distinct peaks within the distribution were observed.
Plants are unique in their wide range of genome size and composition. In the grasses, chromosome size, genome size and GC content have been found to be evolutionarily associated. Small genomes in the genus *Festuca* appear to be AT-rich, are better at adapting to extreme environmental conditions, are more species-rich and are rapidly diverging (Smarda et al., 2008). Isochores of plants were initially characterized by looking only at limited stretches of DNA (Matassi et al., 1989; Montero et al., 1990; Salinas et al., 1988). It was determined that the compositional pattern of isochores was different between monocots and dicots (Salinas et al., 1988). Curiously, compositional similarities were found between monocots and warm-blooded vertebrates. Both groups have higher GC content relative to dicots and cold-blooded vertebrates and even higher GC content in coding regions. Warm-blooded vertebrates and monocots also show a similar distribution of housekeeping genes compared to tissue-specific genes with the housekeeping genes having a much higher GC content (Salinas et al., 1988).

Despite the potential impact that studies of isochores could have on understanding genome evolution, Arabidopsis is the only plant, until now, in which isochores have been identified and characterized at a whole genome level. In the study of the Arabidopsis genome Zhang and Zhang (2004) identified GC-, AT- and centromeric isochores. The centromeric isochores are GC rich yet a low gene density and different T-DNA insertion sites than the GC-isochores are present. Each of the centromeric regions was contained within a centromeric-
isochore meaning that all of the predicted centromere sequences were part of an isochore region (Zhang and Zhang, 2004).

The soybean genome is paleopolyploid with 2n=40 (Pfeil et al., 2005). Recently it was discovered that 57% of the genome is comprised of repeat-rich heterochromatin. However, in these regions, located near the centromeres, 21.6% of the high-confidence genes were discovered (Schmutz et al., 2010). This genetic composition and the soybean’s genome size more closely resembles that of human than that of Arabidopsis. Understanding the role of genome size and compositional patterns could help uncover some of the basic “rules” of genome structure and function during evolution.

Using the recently published genome sequence (Schmutz et al., 2010) we sought an understanding of the LHGRs in the soybean genome. Our goals were to identify the compositionally distinct domains in the genome, isolate the homogeneic regions and classify the families of LHGRs found in the soybean genome. We used the program GC-Profile (http://tubic.tju.edu.cn/GC-Profile/) (Zhang et al., 2005; Gao and Zhang, 2006) to identify compositionally distinct segments within the genome based on nucleotide organization. GC-Profile utilizes a segmentation algorithm and provides a windowless view of the chromosomes. These domains were then given a homogeneity score using the homogeneity index `h’ (Zhang and Zhang, 2004). Long homogeneic genome regions (LHGRs) were identified based on their homogeneity and were assigned to families using a parametric approach to identify the most likely mixture of
distributions (Mclachlan and Peel, 2000) underlying the overall distribution of GC content.

Results

LHGRs in the soybean genome

We used a segmentation algorithm based on z’ curves to identify soybean LHGRs. The z’ curve separates the entire chromosome into overlapping, compositionally distinct domains based on the nucleotide sequence. There is an inverse relationship between GC content and the z’ curve. Thus, when the slope of the curve is positive it is indicative of a decreasing GC content. The soybean genome is composed of many compositionally segmented regions similar to what has been observed in pig (Zhang et al., 2010). With the z’ curve we were able to determine to base-pair resolution the locations of the non-overlapping regions.

To determine whether a region was a LHGR (significantly homogeneic), we analyzed each of the domains using an ‘h’ value (Zhang and Zhang, 2003). The h value evaluates the homogeneity by dividing the variance in the GC content of the region by the variance in the GC content of the whole chromosome. An h-value < 1 is generally considered to be a region with little variation. We found a total of 4,339 compositionally distinct domains in the 20 chromosomes of the soybean genome. The h-values of our regions ranged from 0.0001 to 4.17. We decided to define our LHGRs as regions with an h-value less than 0.01. Using this criterion, three hundred thirty-one LHGRs were identified.
**LHGR families**

The distribution of the 331 LHGRs appeared to be a mixture of at least four overlapping distributions of GC content (Figure 1). To determine the most number of components within this mixture of distributions we used a parametric approach in which we modeled the distribution as:

\[ f(y; \Psi) = \sum_{i=1}^{g} \Pi_i \phi(y; \mu_i, \sigma_i); \]

where \( i \) represents the number of mixtures, or groups, \( \pi \) are unknown, but estimable mixing proportions and:

\[ \phi(y; \mu_i, \sigma) = \left\{ \frac{(2\pi)^{-1/2}}{\sigma^{-1}_i} \right\} \exp\left\{ -\frac{1}{2} \frac{(y - \mu_i)^2}{\sigma^2_i} \right\} \]

with unknown, but estimable means and variances. We calculated the log(likelihood) for each model beginning with \( g = 1 \) and proceeded to sequentially add groups until no significant improvement of the −log(likelihood) was observed (Table 1).

We found that four groups produced the best fit to the data. Based on this analysis the estimated average GC content was 24%, 35%, 44% and 55% (Figure 2) with 41, 152, 123 and 6 members in each family, respectively. This translated into family-weighted values of 0.13, 0.49, 0.36 and 0.02, respectively. This analysis permitted us to determine how much each of our distributions overlapped, i.e. how much of one family overlapped into another family’s distribution. The amount of Family A (24%) in Family B (35%) was 10%, the amount of B in A was 2.6%. The estimated amount of B in C (45%) was 6% while
the amount of C in B was 5%. The amount of C in D (55%) was 0.01% while the amount of D in C was 1%. With this information we can recognize the amount of bias, for example, that Family C puts on Family B (5%) when looking at the biological properties. This would mean that for each of the LHGRs we classified as a member of Family B LHGR, there is a five percent chance that the LHGR is actually from Family C.

**LHGR size and location**

Table 2 shows the distribution of LHGRs across chromosomes and the respective chromosome length. The number of LHGRs per chromosome does not appear to be associated with the size of the chromosome. For example, chromosomes 3, 4, 7, 9, 13 and 20 have the largest number of LHGRs (34 – 45) while two chromosomes (5 and 11) have no LHGRs. The largest LHGR in the soybean genome is 2 mb in length and is located on chromosome 7. There are only three LHGRs that are longer than 1 mb and two of these are located on chromosome 13 (Figure 3). All of the three longest LHGRs are located in the euchromatic arms of the chromosomes and two of them are located near the telomeric region. Although the top ten largest LHGRs are all part of Family B, on average, LHGRs in Family C are the largest which means that a large number of few small LHGRs in Family B. The average size of LHGRs in Family C is 41.1 kb while Family A has an average of 12.8 kb, Family B has an average of 13.8 kb and Family D has an average of 7.3 kb (Figure 3). This is consistent with results observed in Arabidopsis in which the GC-LHGRs are larger than AT-LHGRs and
is different than mammals such as pig and human where the AT-LHGRs are larger (Zhang et al., 2010). The average size of the soybean LHGRs is 0.82 mb; smaller than the average size for pigs (0.91 mb) and humans (1.20 mb) (Zhang et al., 2010).

**Gene distribution and transposable elements in soybean LHGRs**

The majority of the genes located within soybean LHGRs are found in Family B (Table 3). The coding regions of genes in Family C were much larger than those in Family A and Family B (Table 4). This includes the individual exon lengths, the total exon length and the average exon length as well as the total number of exons. As seen in Table 4, the average total exon length for Family C is approximately twice the average total exon length for either Family A or Family B. Interestingly, the average length of the intergenic region, or regions between genes, is 50,000 bp for Family A and 70,000 bp for Family B and then jumps up to 120,000 bp for Family C showing that not only are the coding regions longer but the noncoding region are also. This is emphasized when you consider that the total intron length and average intron length of genes are the longest for Family C. Family D has been excluded from this comparison as there is only one gene in the LHGRs of Family D.

The number of transposable elements in the LHGRs follows a trend similar to that of gene density. There are more than 2.5 X the number of transposable elements in Family B than any of the other families. However, the difference in the density of transposable elements is not as extreme as the difference in gene
density per LHGR. The number of genes per LHGR for Family B averages 13 while it is only 0.23 for Family C, which means Family B has about 50 times more genes than Family C. Alternatively, the number of transposable elements per LHGR averages 5 for Family B and 2.5 for Family C.

**Gene function**

To explore the predicted function of the genes in the families we used Gene Ontology (GO), KEGG and Panther annotations. Following the protocols of O'Rourke et al. (2008) and Bernardini et al., (2004) we used a Fisher's exact test (Fisher, 1949) and a Bonferroni correction (Bonferroni, 1936) on Family B to determine which gene functions are over- or under-represented in our gene list compared to those gene functions in the whole soybean genome. Only Family B contained enough genes to use this statistic. Family B contains two groups of genes with molecular function gene ontologies that are over-represented, GO:0008683, 2-oxoglutarate decarboxylase and GO:0030976, thiamin pyrophosphate binding. The only gene in Family D is Glyma18g06990 and is predicted to be an ATP-dependent CLP protease. The protein products of these genes are involved in cell regulation and they help to stabilize key metabolic enzymes and also remove damaged polypeptides (Clarke, 1999). Next we pooled all of the genes found in the LHGRs and again performed a Fisher’s exact test (Fisher 1949) with a Bonferroni correction (Bonferroni, 1936). The same two GO molecular functions were over-represented, oxoglutarate decarboxylase and
thiamin pyrophosphate binding. This is not surprising as numerous gene functions are shared between the four families.

Interestingly many of the LHGRs are clustered along the genome. Figure 4 shows the physical locations of the LHGRs across the twenty soybean chromosomes separated by family. Surprisingly, many of the LHGRs from different families cluster together along the chromosome.

**Discussion**

The definition of an LHGR, or isochore, is based on its homoegeneity but the transition between homogeneic to heterogeneic is unclear. There are no regions of the genome that are completely homogeneic and the question is, at what point does a region shift from being homogeneic to heterogeneic. For this reason we chose a conservative cutoff in hopes of eliminating false positives. Heterogeneic regions could have a separate set of biological properties so understanding the differences between the regions could help differentiate between homogeneous and heterogeneous regions. Therefore, one problem in analyzing isochores is the somewhat arbitrary level of acceptable heterogeneity (Chen and Gao, 2007).

The regions that fit our criteria as homogeneic displayed four overlapping mixture models across the GC percent. We fit our families of LHGR count across GC percent with maximum likelihood. Using this statistic allowed for unrestricted variances as our components appear as an asymmetric multimodal density and there was no evidence to restrict the variance. It was apparent that after four
components the log of the likelihood did not significantly change. Using a method to statistically determine the number of components, or families, may be considered in future work however there are many considerations when choosing the statistics. We considered using the LRTS, $-2 \log \lambda$, which adds a penalty for each additional parameter however there is a problem with the parameters being bounded correctly when used on mixture models as discussed in McLachlan and Peel (2000). In this analysis we used a one parameter, the dimension of GC content. Including other parameters such as biological properties could be useful in future research.

LHGRs have been considered a “fundamental level of genome organization” (Eyre-Walker and Hurst, 2001) and have given us insight into the complexity of large regions of the genome. Various important biological properties such as gene expression, gene size and transposable element density have been correlated with LHGRs (Mouchiroud, 1991; Zouback, 1996; Aota, 1986; Jabbari, 1998). To identify LHGR families in soybean we used a novel approach. Instead of defining our families based on the peak in a graph of GC content by frequency, we used an approach to determine the most likely number of distributions in our data and with that, determined the parameters of each of the families. LHGRs in soybeans comprise four families, each of a different size. There are two predominant families, Family B (35% mean GC) and Family C (44% mean GC) and two minor families, Family A (24% mean GC) and Family D (55% mean GC). This is different from what was found in Arabidopsis where no
distinct peaks or families were apparent (Zhang and Zhang, 2004). Soybean
exhibits a very wide range of GC content in LHGRs, the lowest LHGR has a 16%
GC content and the highest LHGR has a 59% GC content, greater than that
found in vertebrates or invertebrates. Vertebrate LHGR families are conserved
and are identified as L1 (>37% GC), L2 (37%-41% GC), H1 (41%-46% GC), H2
(46%-53% GC), H3 (>53% GC) (Zhang et al. 2010). Peaks, or centers of families,
in LHGRs appear approximately in 5% bins while in soybean they appear in
approximately 10% bins. The biological or evolutionary significance of this
remains unknown.

In previous studies of other genera, a majority of the genes identified
resided in one narrow GC range, similar to our observation. However, in most
species the gene density increased as the GC content increased (Constantini et
al., 2009; Cammarano et al., 2009), but in soybeans the gene density is highest
in the family with the second lowest GC content, Family B (35%). This is similar
to what was found in Zebrafish (Constantini et al., 2007b). Family B also contains
most of the transposable elements. As observed in other species, transposable
elements also seem to be enriched in LHGRs at one specific level of GC content
and depleted in others (Mouchiroud, 1991).

The physical properties of the genes in Family B are similar to those found
in Family A and are similar to the average size for a soybean gene. However, the
genes in Family C are much larger both in the coding and the noncoding regions
than the other two families. Family D has the greatest GC content but only
contains one gene. However, it is important to note that Family D consists of only six LHGRs. In previous studies change in gene density correlated with a change in the physical properties of the gene as did the expression level of the gene (Zouback, 1996, Aota, 1986, Woody et al., 2011). It is interesting to note then, that the LHGR families are also correlated with several of these features.

This is the first study done in plants in which a functional analysis has been conducted on the genes in LHGRs. We analyzed the genes in Family B using a Fisher’s exact test with a Bonferroni correction, to identify which functions were under/over-represented. We found that 2-oxoglutarate decarboxylase and thiamin pyrophosphate binding are overrepresented. The two enzymes work closely together as thiamin pyrophosphate is needed to maximize the efficiency of 2-oxoglutarate (Shigeoka and Nakano, 1991) and both are involved in the Krebs cycle (Mitsuda et al., 1975). The similarity in gene function across families is striking. We were unable to perform a statistical analysis on Family A, C or D independently as there were not enough genes but the function of the genes in these families are consistent with the results found in Family B. However, we were able to pool all of the genes in our analysis into one large group and performed the same statistical analysis. Again we found that 2-oxoglutarate decarboxylase and thiamin pyrophosphate binding are significantly overrepresented which is not surprising as the families share many gene functions. The majority, if not all of the genes in the LHGRs are related to metabolic and cell cycle functions such as ATP and AMP proteases and serine-
threonine kinases. In Family D, there is only one gene and even this gene has a metabolic function, ATP-dependent CLP protease. In Arabidopsis, one LHGR and one LHGR-like regions are nucleolar organizer regions while another LHGR is a mitochondrial insertion (Chen et al., 2005).

In conclusion, LHGRs (isochores) are identifiable in soybean. LHGRs in soybean share similarity with animals in that gene density is focused in a narrow GC range. The average sizes exhibited by LHGR families resemble the trends found in human and pig. However, plants are unique and possibly quite variable. The GC mean of the four families in soybean is more diverse and thus the families are spread further along the range of GC content than what has been found in animals. Alternatively, the mean of the two groups found in Arabidopsis is much smaller than that found in animals. A comparative analysis of LHGRs in plants could help increase our knowledge of compositional evolution across species and the relationship between evolutionary adaptation and large, conserved blocks of the genomes.

Materials and methods

Identification of LHGRs

The genome sequence was downloaded from soybase.org (accessed on 02.02.2011). To identify compositionally distinct domains we used a program GC-profile (http://tubic.tju.edu.cn/GC-Profile) (Gao and Zhang, 2006). GC-Profile utilizes a segmentation algorithm that allows for a windowless view of the chromosomes. Each chromosome is considered individually and separated into
non-overlapping domains. We used the parameters as suggested in Zhang et al. (2010) with a minimum size of 3 kb and we ignored gaps shorter than 1% of the chromosome. GC-profile is based off of a Z curve statistic which is a way of viewing the unique compositional pattern of each chromosome. The z' score is calculated based off of the cumulative A, C, G and T occurrence along the specific regions.

The homogeneity within the compositionally distinct domains was measured using a homogeneity index h as described in (Zhang et al., 2004) and is defined by:

\[ h = \frac{d_{\text{LHGR}}}{d_{\text{Chromosome}}} \]

\[ d_{\text{LHGR}} = \sqrt{\frac{1}{M} \sum_{n=1}^{M} (z_n - kn)^2} \]

\[ d_{\text{Chromosome}} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (z_n - kn)^2} \]

where k is the slope of the line through the z' score within the region (chromosome or LHGR) and \( z_n \) is the cumulative z' score across the region. Only absolute homogeneity, a region comprised of only GC or AT, will result in an h value of zero: some heterogeneity has been present in all investigations to date. The level of this heterogeneity is chosen to distinguish groups of LHGRs is arbitrary as a method to finding an absolute cutoff is not known. Future work is needed on measures that can delineate the shift from homogeneous regions to heterogeneous regions.
Transposable elements and gene annotation

The transposable elements were taken from the SoyBase website (www.soybase.org). These were identified from the Soybean Transposable Element Database on 03/08/2011 (Du et al. 2010). Close to 40% of the soybean genome is identified as some type of repeat, mostly active retrotransposons and simple-sequence-repeats. We then identified any transposable elements in our segments. If there was any overlap, the transposable element was counted as part of the segment. We quantified the density of transposable elements by calculating the number of transposable elements per mb. To identify which transposable elements were in LHGR regions we used the fjoin program (Richardson 2006).

SoyBase gene annotations (www.soybase.org) on 07/19/2011 were used for identification of genes. If any exon of a gene was contained within an LHGR, it was considered part of the LHGR. SoyBase was also the source of information on individual exon sizes, intron sizes and intergenic regions. The gene density was calculated by the number of genes within an LHGR per mb. The gene coverage was calculated by the sum of the coding regions of the genes within an LHGR (bp) divided by the total length of the LHGR (bp).

Gene function

To obtain the functional annotation of the genes in our LHGRs we used several methods. For Family B we used the annotations as described in O'Roarke et al. (2008) and Bernardini et al. (2004). The predicted gene
sequences in soybean (Glyma1.01 genome assembly) were compared using TBLASTX (E < 10^{-6}, (Altschul et al. 1997)) to the predicted genes in Arabidopsis. The annotation of the gene model of Arabidopsis that best fit the soybean gene model was used as the basis for the gene ontologies. To determine if any gene functions were over-represented in our LHGRs compared to the entire genome we used gene ontology (GO) annotations. For each GO category, a count was taken for the number of genes connected to it in the LHGRs (specified group) and in the entire genome (population). A Fisher's exact test was done on the each GO category in the LHGRs (O'Rourke et al. 2009) and a Bonferroni adjustment was used (Bonferroni 1936). GO categories with a p-value < 0.05 were considered significant. Families 1, 3 and 4 did not have enough genes to be able to perform an accurate GO analysis. Thus we looked at the annotated functions for these genes in SoyBase.

**Literature cited**


Volume 273, Issue 8, pages 1637–1648.


Figure 1. Distribution of LHGRs (binned into 1% GC intervals) across GC content (in percent).
Figure 2. The mixture model of frequency distributions of the four LHGR families across GC content (in percentage). The different colors represent the different families. Yellow is Family A, green is Family B, blue is Family C and red is Family D. The numbers in white at the base of the density distributions represent the amount of the overlap estimated to come from the adjacent family. For example, the 0.03 value at the base of Family A (yellow) indicates the amount of that family’s distribution that may be attributed to Family B (green).
Figure 3. The size distribution of LHGRs in each family; A) Family A, B) Family B, C) Family C, and D) Family D.
Figure 4. Physical location of LHGRs across chromosomes. The shaded area represents the pericentromeric region (SoyBase.org). The vertical line represents the predicted centromeric region (SoyBase.org). Blue bar = Family A. Red bar = Family B. Orange bar = Family C. Green bar = Family D.
### Table 1. Value of the log of the likelihood for the mixture components.

<table>
<thead>
<tr>
<th>Number of Components</th>
<th>Unrestricted Variances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1135.29</td>
</tr>
<tr>
<td>2</td>
<td>-1134.512</td>
</tr>
<tr>
<td>3</td>
<td>-1119.974</td>
</tr>
<tr>
<td>4</td>
<td>-1113.617</td>
</tr>
<tr>
<td>5</td>
<td>-1113.617</td>
</tr>
<tr>
<td>6</td>
<td>-1112.655</td>
</tr>
</tbody>
</table>

### Table 2. The number of LHGRs in each of the 20 soybean chromosomes and the length of the corresponding chromosome.

<table>
<thead>
<tr>
<th>Chromosome Number</th>
<th>Chromosome Length (bp)</th>
<th>Number of LHGRs/Family</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>1</td>
<td>55915595</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>51656713</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>47781076</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>49243852</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>41936504</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>50722821</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>44683157</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>46995532</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>46843750</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>50969635</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>39172790</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>40113140</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>44408971</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>49711204</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>50939160</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>37397385</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>41906774</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>62308140</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>50589441</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>46773167</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 3. The number of genes, number of LHGRs and the number of transposable elements located in each family.

<table>
<thead>
<tr>
<th>Family</th>
<th>Number of genes</th>
<th>Number of transposable elements</th>
<th>Number of LHGRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>18</td>
<td>20</td>
<td>41</td>
</tr>
<tr>
<td>B</td>
<td>1949</td>
<td>795</td>
<td>152</td>
</tr>
<tr>
<td>C</td>
<td>28</td>
<td>313</td>
<td>123</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 4. The physical parameters of the genes in soybean LHGR Family A, Family B and Family C.

<table>
<thead>
<tr>
<th></th>
<th>Family A</th>
<th>SD</th>
<th>Family B</th>
<th>SD</th>
<th>Family C</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>First exon GC</td>
<td>42</td>
<td>2</td>
<td>43</td>
<td>4</td>
<td>47</td>
<td>4</td>
</tr>
<tr>
<td>Mid Exon GC</td>
<td>43</td>
<td>3</td>
<td>41</td>
<td>3</td>
<td>45</td>
<td>2</td>
</tr>
<tr>
<td>Last Exon GC</td>
<td>42</td>
<td>2</td>
<td>43</td>
<td>5</td>
<td>46</td>
<td>3</td>
</tr>
<tr>
<td>First Exon Length (bp)</td>
<td>240</td>
<td>97</td>
<td>300</td>
<td>232</td>
<td>640</td>
<td>323</td>
</tr>
<tr>
<td>Mid Exon Length (bp)</td>
<td>240</td>
<td>84</td>
<td>200</td>
<td>121</td>
<td>440</td>
<td>147</td>
</tr>
<tr>
<td>Last Exon Length (bp)</td>
<td>200</td>
<td>76</td>
<td>270</td>
<td>314</td>
<td>500</td>
<td>191</td>
</tr>
<tr>
<td>Intergenic Region (bp)</td>
<td>49000</td>
<td>29304</td>
<td>72000</td>
<td>85329</td>
<td>116000</td>
<td>84806</td>
</tr>
<tr>
<td>Total Exon Length (bp)</td>
<td>1060</td>
<td>450</td>
<td>1000</td>
<td>520</td>
<td>2420</td>
<td>1233</td>
</tr>
<tr>
<td>Mean Number of Exons</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Mean Intron Length (bp)</td>
<td>290</td>
<td>72</td>
<td>300</td>
<td>192</td>
<td>400</td>
<td>195</td>
</tr>
<tr>
<td>Total Intron Length (bp)</td>
<td>1810</td>
<td>565</td>
<td>2080</td>
<td>1471</td>
<td>3520</td>
<td>1417</td>
</tr>
</tbody>
</table>
Chapter 5: General Conclusions

The overall goal of this research was to increase the body of knowledge of the genetic landscape in the soybean genome. Genetic architecture has evolved into a large-scale topic as the sequencing technology reveals vast areas of unexplored components in the genome. Genomes from closely related species can vary greatly in terms of gene content and yet large blocks of DNA are conserved over multitudes of species. The soybean genome has been largely unexamined with regard to the relationship of gene expression profiles to gene structural parameters. With the availability of the whole-genome sequence and a whole-genome RNA-Seq analysis of soybean tissues sampled during a progression of seed developmental time points and various tissues we were able to analyze the correlation between expression profiles and genic properties. The focus of this project was to identify possible associations between transcriptional expression of genes, the physical parameters of the genes, and their environs. The identification and characterization of LHGRs helped us better understand the larger genomic blocks and their relationship with basic biological properties such as gene density, repeat sequence distribution and functional gene analysis.

In this body of research, the correlation between transcriptional demands and the physical parameters of the gene was analyzed by partitioning the genes first into categories of expression level (low, mid, high) and then into categories of expression breadth (expressed in one to fourteen tissues). In the genes expressed at a low level, an increase in expression breadth correlated with an
increase in transcript size. The average size of the individual exons did not change but the number of exons and introns significantly increased. Conversely, genes that are expressed at a high level of expression displayed a decrease in transcript size as their expression breadth increase. Highly expressed genes that were also broadly expressed had larger average exon sizes than highly expressed genes that were tissue-specific. Thus, at low levels of expression an increase in the number of exons and introns correlates with an increase in expression breadth while at high levels of expression a decrease in the average exon size correlates with an increase in expression breadth. The GC content of the genes was also found to be highly correlated with the expression characteristics. An increase in transcriptional demands, either in breadth and depth, correlated with an increase in the GC content of the genes.

A cluster analysis was then performed to examine whether genes exhibiting unique patterns of expression were physically clustered along the genome. We found that there were clusters of genes that were expressed at a low intensity with a narrow expression breadth, a low intensity with a broad expression breadth, high intensity with a narrow expression breadth, and high intensity with a broad expression breadth. Genes that were highly and narrowly expressed were twice as likely to be in a cluster than the other three expression categories. Interestingly, genes that were expressed at a low level in many tissues typically had twice as many genes in a cluster in the pericentromeric region than the other three categories.
Larger genomic blocks were also analyzed in this research. Long Homogeneous Genome Regions (LHGRs) were identified and characterized. In soybean, 331 LHGRs were identified and they were categorized by GC content into four different families. The majority of the genes within LHGRs were found within a single LHGR family with a narrow GC-range (Family B). We discovered the LHGR pattern in soybeans is unique compared to the studies done in vertebrates and invertebrates. The majority of genes do not reside within a family with highest GC and instead reside in Family B which has a mean GC content of 35%. Interestingly, the majority of the genes in the LHGRs across all families are related to metabolic and cell cycle functions such as ATP and AMP proteases and serine-threonine kinases. A comparative analysis of LHGRs between plants could further our understanding of long, conserved blocks of the genome. More specifically, a comparative analysis between the LHGRs in monocots and dicots.

It is apparent that the structural parameters of not only plant genes, but large genomic blocks are not stochastically determined. Further analysis into these relationships could provide more insight into additional fine-tuning of gene regulatory networks, understanding the epigenetic landscape and aid in the development of transgenic research. In this research the homogeneous regions of the genome were analyzed but a comparison of the homogeneous regions and the heterogeneous regions could help decipher the underlying evolutionary differences. In addition, an analysis of the homogeneous regions in different soybean lines could uncover the level of conservation within the species. For
future research I would propose a study comparing the LHGR regions of both monocots and dicots. It is possible that epigenetic marks in these regions are involved in the correlation between replication timing, LHGRs and placement within the cell nucleus. A whole-genome epigenetic study, DNA methylation, histone variation, etc. would help uncover whether these regions have unique epigenetic variations between families or unique variations that separate the homogeneous regions from the heterogeneous regions.
Acknowledgements

The authors gratefully acknowledge the financial support of the North Central Soybean Research Program and the U.S.D.A Agricultural Research Service. Names are necessary to report factually on the available data; however, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by the USDA implies no approval of the product to the exclusion of others that may also be suitable.
Supplemental figure 1A. Correlation between expression breadth and total exon length in data from Libault et al. (2010). The mean is shown for each level of tissue specificity in each expression category.
**Supplemental figure 1B.** Correlation between expression breadth and total intron length in data from Libault et al. (2010). The mean is shown for each level of tissue specificity in each expression category.
Supplemental figure 1C. Correlation between expression breadth and average exon length in data from Libault et al. (2010). The mean is shown for each level of tissue specificity in each expression category.
Supplemental figure 1D. Correlation between expression breadth and average number of exons in data from Libault et al. (2010). The mean is shown for each level of tissue specificity in each expression category.
Supplemental figure 1E. Correlation between expression breadth and intergenic length in data from Libault et al. (2010). The mean is shown for each level of tissue specificity in each expression category.
Supplemental figure 2. The physical distribution of clusters of genes across all 20 chromosomes in 100 kb bins. Red box on top = highly and broadly expressed gene clusters; red box on bottom = lowly and broadly expressed gene clusters; blue box on top = highly and tissue-specific gene clusters; blue box on bottom = lowly and tissue-specific gene clusters.