FUNCTIONAL DIVERGENCE OF DUPLICATED GENES IN THE SOYBEAN GENOME

Paul L. Auer

R. W. Doerge

Follow this and additional works at: https://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

Recommended Citation

FUNCTIONAL DIVERGENCE OF DUPLICATED GENES IN THE SOYBEAN GENOME

Paul L. Auer [1] and R. W. Doerge

Department of Statistics, Purdue University, West Lafayette, IN 47907-2066

Abstract

The soybean genome has undergone many different evolutionary changes that are observable with modern technologies. Of particular interest to scientists and plant breeders is the fact that the soybean genome exhibits features of genome duplication from millions of years ago. Genes that were copied during the duplication event have since diverged functionally. Identifying functionally divergent duplicate genes may provide insight into the evolution of soybean. To investigate functional divergence, transcripts from seven different tissue samples of pooled soybean messenger RNA were sequenced using the Solexa next-generation sequencer and analyzed for gene expression. We tested differential expression of duplicated genes within tissue by employing an integrated normalization and statistical testing methodology. Blocks of duplicate genes (i.e., gene sets) were tested for unanimity of over- or under-expression. These same genes were also analyzed for differential expression across tissues. We identified thousands of duplicate genes that displayed differential expression patterns within each tissue. In some cases these genes were over-represented in duplicate blocks, suggestive of functional divergence of a large genomic region.

Keywords: next-generation sequencing, RNA-Seq, differential expression, soybean

1 Introduction

Soybean (*Glycine max*) is a prominent nutritional resource for both animal feed and cooking oil and is one of the most important agricultural crops worldwide. The sequence of the soybean genome was recently released (Schmutz et al., 2010), making it the only legume species with a currently available reference sequence. As such, it serves as a reference for over 20,000 other species (Schmutz et al., 2010). Soybean is an ancient polyploid (i.e., a "paleopolyploid"), meaning that its genome was duplicated millions of years ago. The current soybean genome contains 20 pairs of chromosomes, making it a diploid species. It is the product of a diploid ancestor with 11 pairs of chromosomes, which lost a chromosome pair, and underwent polyploidization (i.e., its genome duplicated; each pair of chromosomes doubled making four copies of each chromosome). Over time the duplicated regions were segmented and shuffled throughout the genome, which then underwent diploidization (i.e., it lost two of the four copies) (Shultz et al., 2006). Because these duplicated regions mixed in with the rest of the genome, the current soybean genome reveals traces of the duplication events (Schmutz et al., 2010). Many of these duplicated regions contain genes. Duplicate genes that behave differently are a major feature of polyploid evolution in plants (Blanc and Wolfe, 2004).

---

[1] Corresponding author: Department of Statistics, Purdue University, 150 N. University St., West Lafayette, IN 47907. E-mail: plivermo@purdue.edu

Genes are often considered functional regions of the genome because they encode for proteins, which are the primary determinants of biological form and function (Griffiths et al., 2008). Genes encode for proteins by transcribing into messenger ribonucleic acid (mRNA), which is translated into protein (Crick, 1970). To understand the behavior or function of a gene, it is important to know both the protein it encodes for and its level of activation. One approach to measuring the level of activation of a gene (i.e., "gene expression") is by measuring mRNA levels. The presence of mRNA in a cell indicates that a gene has been expressed and will be translated into protein. Thus, a measure of mRNA abundance provides a reasonable indication of gene expression.

To study the functional divergence of duplicated genes in soybean, mRNA abundance levels were investigated by the Jackson Laboratory at Purdue University. Using the Illumina Genome Analyzer (i.e., "Solexa," Illumina, 2010), a next-generation sequencing (NGS) technology, mRNA (i.e., "transcripts") were directly measured through a new methodology called RNA sequencing (RNA-Seq). mRNA was extracted from the cells of seven different tissues (Apical Meristem, Flower, Green Pods, Leaves, Nodule, Root, and Root Tip) from multiple soybean plants. The mRNA was then pooled according to tissue type. These seven pools of transcripts were then randomly fragmented by sonication, and reverse transcribed into complementary DNA (cDNA). Only those cDNA fragments meeting a certain size specification (roughly between 250-500 bases) were retained. Small adapters approximately 20 bases long were ligated to the ends of the size-selected cDNA fragments. The resulting cDNA fragments with attached adapters were then amplified by several rounds of Polymerase Chain Reaction (PCR; Saiki et al., 1988). Every round of amplification doubles the amount of cDNA from the previous round. Each of the seven pools of transcripts was then input to the Solexa NGS device (Figure 1).

The Illumina sequencing platform consists of a flow cell, a cluster station, and a sequencing machine. The flow cell is a small glass slide onto which eight lanes have been channeled. The flow cell was loaded into the cluster station where the seven pools of transcripts were input to seven different lanes of the flow cell. Once loaded with the samples, the flow cell was placed inside the sequencing machine. Each cDNA fragment in the lanes of the flow cell is comprised of a combination or "sequence" of four nucleotides [adenine (A), thymine (T), guanine (G), and cytosine (C)]. The Solexa sequencing machine works by reading off the combination or "sequence" of nucleotides comprising the cDNA fragments in the lanes of the flow cell. When the machine is finished sequencing, it outputs approximately five million 36 base "sequencing reads" in each of the seven lanes. These sequencing reads represent the cDNA fragments that were loaded into the flow cell. In order to interpret these reads, they were aligned to the soybean reference genome using the Genomic Short-read Nucleotide Alignment Program (GSNAP; Wu and Watanabe, 2005). Alignment entails searching the reference genome for regions that match the sequencing reads. Allowing the reads to match a position with a two base mismatch tolerance resulted in 56% of the reads mapping to a unique genomic location. Gene expression was quantified by adding the number of reads that mapped uniquely to each gene.

Duplicate genes were defined as genes located within larger duplicate regions called "homologous blocks." These homologous blocks represent regions of the soybean genome that were duplicated and then shuffled throughout the genome over evolutionary time. Various computational approaches have been developed for identifying homologous blocks (de Peer, 2004); in this work,
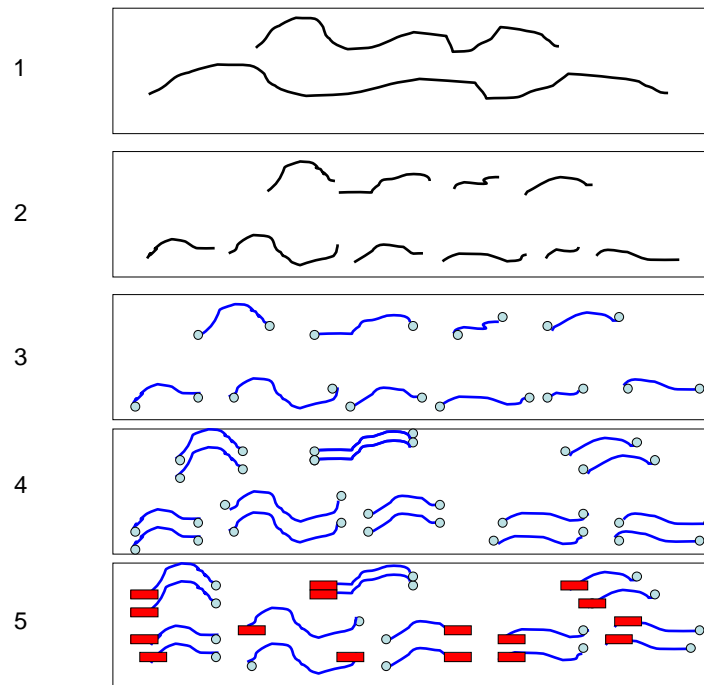
Figure 1: Description of the RNA-Seq experimental process: mRNA (represented in black) is isolated from cells (1), randomly fragmented (2), and copied into cDNA (3, cDNA is represented in blue). Adapters (open circles) are ligated to the ends of the cDNA strands, and the cDNA is size selected (4). The size selected cDNA is then amplified using PCR (4). Finally, the ends of the amplified cDNA are sequenced (5, the sequenced portion is illustrated by a red rectangle). Image taken from Auer (2010).

the recently developed i-ADHoRe software (Simillion et al., 2008) was used to identify homologous blocks in the soybean genome (Schmutz et al., 2010). Although pairs of duplicate genes (i.e., "paralogues") have similar sequences (this is a criterion for identifying duplicate regions with the i-ADHoRe software), often they contain enough dissimilarity so that 36 base RNA-Seq reads can be mapped unambiguously to one copy of the duplicate pair. Given RNA-Seq read counts for a pair of duplicate genes, we investigated functional divergence by statistically testing several different hypotheses. We tested for differential expression, within each tissue, of pairs of duplicate genes. With results from this analysis, we tested for overall differential expression of the homologous blocks. Demonstrating the flexibility of these data to answer several different biological questions, we also tested for differential expression of each gene across tissues as well as the overall differential expression, across tissues, of the homologous blocks.

This analysis confronts many unresolved statistical issues in the analysis of RNA-Seq data. Although several methods have been proposed for testing differential gene expression across ex-

perimental conditions or tissues types (see Auer and Doerge, 2010, for a review), there has been limited research on testing differential expression between genes. The fact that different genes have different lengths (even though they are duplicates), suggests that gene length should taken into account in the statistical tests (Blekhman et al., 2010). Currently, there are no gold-standard methods that address this issue. Differences in gene length also complicate analyses of gene sets (Oshlack and Wakefield, 2009). Although newly developed methods effectively handle differences in gene lengths for gene set analyses (Young et al., 2010), gene length remains a problem when testing for unanimity of over- or under-expression in a gene set. The methods presented here provide a statistical framework that acknowledges these well documented, unresolved issues in the analysis of RNA-Seq data.

2   Modeling Differential Expression of Duplicate Genes

Although duplicate genes exhibit high sequence similarity, they often produce transcripts of different length. Recall the experimental process that generates RNA-Seq data (Figure 1): longer transcripts will produce more random fragments and thus more sequencing reads. Let Y_{ig} be the number of reads mapping to gene g in the ith tissue type ($i = 1, ..., 7$). The expected value of Y_{ig} is proportional to the total abundance of gene g in the ith tissue (λ_{ig}) times the length of gene g (L_g): $E(Y_{ig}) = \lambda_{ig}L_g$. For inference comparing the transcription levels, within the ith tissue, of duplicate genes g and g', we compared the read counts Y_{ig} and $Y_{ig'}$ *normalized* by the gene lengths L_g and $L_{g'}$, respectively.

Formally, this was accomplished by assuming that the gene counts Y_{ig} are Poisson($\lambda_{ig}L_g$) random variables. The Poisson distribution is often used to model count data (Agresti, 2002), and RNA-Seq gene counts in particular (Marioni et al., 2008). Differential expression within the ith tissue between duplicate genes g and g' was tested with the following hypotheses

$$H_0 : \lambda_{ig} = \lambda_{ig'} \qquad \text{vs.} \qquad H_A : \lambda_{ig} \neq \lambda_{ig'}. \tag{1}$$

Conditioning on the total number of reads for the duplicate pair ($k = Y_{ig} + Y_{ig'}$), the distribution of Y_{ig} is

$$
\begin{aligned}
Y_{ig}|k &\sim \quad \text{Binomial}(\pi, k), \\
\pi &= \quad \frac{L_g\lambda_{ig}}{L_g\lambda_{ig} + L_{g'}\lambda_{ig'}}.
\end{aligned}
\tag{2}
$$

Under the null hypothesis of no differential expression, the Binomial proportion is $\pi_0 = \frac{L_g}{L_g+L_{g'}}$. This gives the form of the Exact Conditional Test (ECT; Przyborowski and Wilenski, 1940) for testing the equality of two Poisson rates (e.g., Equation 1). Given observed counts y_{ig} and $y_{ig'}$, the two sided P-value from the ECT is

$$2 \times \min\left\{ \sum_{j=y_{ig}}^{k} P(Y_{ig} = j|k, \pi_0), \sum_{j=0}^{y_{ig}} P(Y_{ig} = j|k, \pi_0) \right\}. \tag{3}$$

Table 1: Number and percent of differentially expressed (DE) paralogues within each tissue. Differential expression was tested with the ECT, and the FDR was controlled at the 0.05 level using the BH method.

Tissue	DE paralogues	Percent DE
Apical Meristem	8838	53.59
Flower	8600	51.61
Green Pods	5356	33.50
Leaves	7730	48.49
Nodule	7856	50.38
Root	8764	52.79
Root Tip	7703	48.75

P-values for testing differential expression, within each of the i tissues, of duplicate genes g and g' were obtained in this manner for all 17,538 pre-defined paralogues. To account for multiple testing, we adjusted the P-values using the Benjamini-Hochberg procedure (BH; Benjamini and Hochberg, 1995), controlling the False Discovery Rate (FDR) within the ith tissue at $q = 0.05$. Only gene pairs for which $y_{ig} \neq 0$ or $y_{ig'} \neq 0$ were included in this analysis. Table 1 summarizes the distribution of differentially expressed duplicate genes within each tissue.

3 Expression Profiles within Duplicated Blocks of Genes

Recall that the duplicate genes are located within larger duplicated regions called homologous blocks. For each tissue, we visualized the expression profiles of the duplicate genes within pairs of pre-defined homologous blocks (Figure 2). For this pair of duplicate blocks, notice that the orientation of the genes on chromosome 13 is flipped on chromosome 15. Also note that within this particular block of genes, there appear to be a large number of blue vertical bars representing over-expression of the gene on chromosome 13 compared to its duplicate on chromosome 15. To determine whether this observation is in keeping with the distribution of differential expression across the genome as a whole, we interpreted the differential expression results summarized in Figure 2 as a hypergeometric sampling problem.

Specifically, $B_1, ...B_H$ represent the pre-defined duplicate blocks, where H is the total number of pre-defined duplicate blocks. We assigned every gene pair within each duplicate block a label indicating its status as significantly over-expressed (SOE), significantly under-expressed (SUE) or equally expressed (EE) as classified in the differential expression analysis (Section 4.1). Let (b_{1h}, b_{2h}, b_{3h}) represent the total number of SOEs, SUEs, and EEs in duplicate block B_h. For duplicate block h we tested whether (b_{1h}, b_{2h}, b_{3h}) represents a random sample of $n_h = \sum_{i}^{3} b_{ih}$ draws without replacement from the overall list of SOEs, SUEs, and EEs throughout the genome. To do so, for each duplicate block h we arranged the results of the differential expression analysis in a 2×3 table (Table 2).
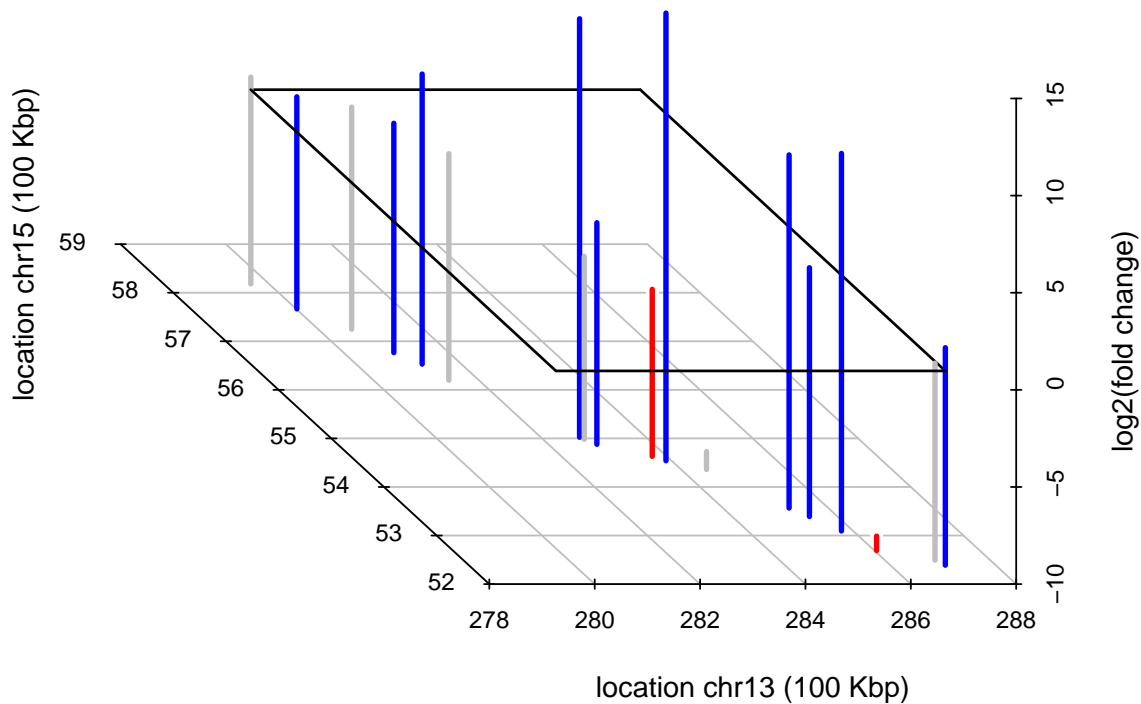
Figure 2: The expression profiles of the paralogues in the duplicate blocks located on chromosomes 13 and 15, for the Nodule tissue sample. The values on the vertical axis represent the $\log_2$(fold change) of each gene pair, after normalizing for gene length [e.g., for gene pair g and g' in tissue i, the fold change is $(Y_{ig}L_{ig'})/(Y_{ig'}L_{ig})$]. The horizontal axis represents location of each gene on chromosome 13 in 100 kilo-basepairs, and the axis representing depth shows location of each gene on chromosome 15 in 100 kilo-basepairs. Blue and red vertical bars represent duplicate genes that are significantly differentially expressed (blue when the gene on chromosome 13 is over-expressed, red when the gene on chromosome 15 is over-expressed). The black box provides a visual aid for identifying zero $\log_2$(fold change). Grey vertical lines represent gene pairs that were not found to be differentially expressed. Image taken from Auer (2010).

The probability of observing b_{1h} SOEs, b_{2h} SUEs, and b_{3h} EEs under a simple random sample is given in McCullagh and Nelder (1989) as

$$P(b_{1h}, b_{2h}, b_{3h}) = \frac{\binom{b_{1.}}{b_{1h}}\binom{b_{2.}}{b_{2h}}\binom{b_{3.}}{b_{3h}}}{\binom{n_.}{n_h}}. \tag{4}$$

By fixing the marginals of the 2×3 table, P-values can be obtained by summing the probabilities of all 2×3 tables that are at least as unlikely to occur. Unfortunately, there is no standard software for efficiently calculating these P-values, and since the marginal totals are so large, implementation is excessively time-consuming, computationally. Instead we took a Monte Carlo approach to finding the P-values. For each block h, we took n_h draws without replacement from a collection of $b_{1.}, b_{2.},$

Table 2: A 2×3 table of the results from the differential expression analysis for duplicate block h. The counts b_{1h}, b_{2h}, and b_{3h} represent the number of significantly over-expressed (SOEs), significantly under-expressed (SUEs), and equally expressed (EEs) gene pairs, respectively, in block h. The column totals ($b_{1.}$, $b_{2.}$, and $b_{3.}$) represent the total number of SOEs, SUEs, and EEs, respectively, across the whole genome. The marginal row total n_h represents the number of gene pairs in block h. These variables define the 2×3 table.

	SOE	SUE	EE	Row total
Block h	b_{1h}	b_{2h}	b_{3h}	n_h
Remaining blocks	$b_{1(-h)}$	$b_{2(-h)}$	$b_{3(-h)}$	$n_{(-h)}$
Column Total	$b_{1.}$	$b_{2.}$	$b_{3.}$	$n_.$

Table 3: Number and percent of homologous blocks that showed non-random patterns of differential expression (DE) of their constituent duplicate genes. The FDR was controlled at $q = 0.1$.

Tissue	DE Blocks	Percent DE
Apical Meristem	46	6.40
Flower	33	4.59
Green Pods	55	7.65
Leaves	31	4.31
Nodule	27	3.76
Root	31	4.31
Root Tip	25	3.48

and $b_{3.}$ SOEs, SUEs, and EEs respectively. After aggregating the results of the n_h draws into b_{1h}, b_{2h}, and b_{3h}, we then calculated the probability (Equation 4). After 20,000 iterations, our Monte Carlo P-value was calculated as the proportion of times a draw produced a probability (Equation 4) at least as unlikely as the one we observed in block h. Since we performed this testing procedure for each of the 719 pre-defined pairs of homologous blocks, a multiple testing correction was necessary. We adjusted the P-values using the BH method, controlling the FDR within the ith tissue at $q = 0.1$.

Of the 719 pre-defined pairs of homologous blocks, very few duplicate blocks demonstrated non-random expression patterns of their constituent duplicate genes (Table 3). However, there was a homologous block found on chromosome 7 that displayed over-expression of its genes relative to their duplicate pairs in the block on chromosome 15 in every tissue. In fact, the chromosome 7 block is also duplicated on chromosome 19, as well as on chromosome 3. In every case, and in every tissue, the genes on chromosome 7 were over-expressed relative to their duplicates on the other chromosomes.

Table 4: A 2×2 contingency table for testing differential expression between Tissue 1 and Tissue 2 of gene g. The counts Y_{ig} represent the count in Tissue i, $(i = 1, 2)$, for gene g or the remaining genes $(-g)$. The ith marginal column total is denoted $Y_{i.}$ and represents the total number of mapped reads for the sample in Tissue i. The gth marginal row total is denoted $Y_{.g}$ and represents the total gene counts summed across columns. $Y_{..}$ is the sum of the two column totals.

	Tissue 1	Tissue 2	Row Total
Gene g	Y_{1g}	Y_{2g}	$Y_{.g}$
Remaining Genes	$Y_{1(-g)}$	$Y_{2(-g)}$	$Y_{.(-g)}$
Column Total	$Y_{1.}$	$Y_{2.}$	$Y_{..}$

4 Modeling Differential Expression Across Tissues

To investigate how the behavior of duplicate genes change across tissue types, for each duplicate gene we tested differential expression for the root versus nodule and the leaf versus flower tissue comparisons. We proceeded on a gene-by-gene basis by organizing the data for each gene and each comparison in a 2×2 contingency table (Table 4). Because the same gene is being compared across tissues, the length of the gene is unimportant. However, the total number of mapped sequencing reads in each tissue is critical because we are comparing across tissues. Often in RNA-Seq analyses, the gene counts Y_{ig} are assumed to be proportional to the total number of mapped reads in the ith tissue (Marioni et al., 2008; Mortazavi et al., 2008). Thus, for inference comparing the transcription levels, of duplicate gene g, across tissues i and i', we compared the read counts Y_{ig} and $Y_{i'g}$ *normalized* by the total number of reads in tissue i ($Y_{i.}$) and i' ($Y_{i'.}$), respectively.

Due to the fact that for some genes the corresponding 2×2 table contained small cell counts (i.e., $y_{ig} < 5$), for each gene we tested for differential expression with Fisher's Exact Test (Fisher, 1935b). Fisher's Exact Test assumes that the marginal totals of the 2×2 table (i.e., Table 4) are fixed and tests differential expression using the hypotheses

$$H_{0g} : \theta_g = 1 \qquad \text{vs.} \qquad H_{Ag} : \theta_g \neq 1, \qquad (5)$$
$$\theta_g = \frac{\pi_{1g} \pi_{2(-g)}}{\pi_{2g} \pi_{1(-g)}}$$

where π_{ig} is the true level of expression for gene g in the ith tissue. In Table 4 consider there being $Y_{.g}$ white balls and $Y_{.(-g)}$ black balls in an bag. If one were to draw $Y_{1.}$ total balls from the bag, one may ask "What is the probability of observing an outcome at least as unlikely as Y_{1g} white balls?" If this probability (i.e., the P-value from Fisher's Exact Test) is small, then the column classification has affected the draw from the bag. In our application, this implies that gene g is differentially expressed between Tissue 1 and Tissue 2. We calculated two-sided P-values by

summing the probabilities of all 2×2 tables (fixing the marginal totals) with probabilities less than or equal to that of the observed table, where the probability of a 2×2 table (e.g., Table 4) is

$$P = \frac{Y_{.g}! Y_{.(-g)}! Y_{1.}! Y_{2.}!}{Y_{..}! y_{1g}! y_{2g}! y_{1(-g)}! y_{2(-g)}!}. \tag{6}$$

P-values for testing differential expression of gene g across tissues i and i' were obtained in this manner for all 35,076 duplicate genes for both the leaf versus flower and root versus nodule comparisons. To account for multiple testing, we adjusted the P-values using the BH procedure, controlling the FDR for each comparison at $q = 0.05$. Only genes for which $y_{ig} \neq 0$ or $y_{i'g} \neq 0$ were included in this analysis. There were 16,539 genes that were differentially expressed (out of 32,250 that had at least one non-zero value) between the root and nodule tissues. The performance of Fisher's Exact Test can be seen in Figure 3; for genes with larger expression values, differential expression is easier to detect. Note that the plot is also symmetric, suggesting that overall both tissues have approximately the same number of over- and under-expressed genes. The flower versus leaf comparison produced a similar plot, with 10,753 genes (out of 32,388 that had at least one non-zero value) differentially expressed between the two tissues.

5 Modeling Blocks of Differentially Expressed Genes Across Tissues

Using the spatial structure of the soybean data, we summarized the results of the differential expression analysis across tissues by homologous block. For the two tissue comparisons (leaf versus flower and root versus nodule), we visualized the expression levels of genes within pairs of pre-defined homologous blocks (Figure 4). Notice that within this particular pair of homologous blocks, there appear to be a large number of red points representing over-expression in the root tissue compared to the nodule tissue. Just as before with the duplicate gene analysis, we interpreted these results as a hypergeometric sampling problem.

For each tissue comparison (leaf versus flower and root versus nodule) and each of the 719 pairs of homologous blocks, we organized the results of the differential expression analysis exactly as shown in Table 2. We tested for over-abundance of differentially expressed genes (in a particular direction) using the same formulation as before, calculating P-values as the sum of the probabilities (Equation 4) of all 2×3 tables that are at least as unlikely to occur, given fixed marginal totals. Again, we used a Monte Carlo approach since the closed form expression is excessively time-consuming to calculate. P-values were adjusted for each tissue comparison, using the BH procedure for controlling the FDR at $q = 0.1$. We found 22 pairs of homologous blocks that were enriched for genes over-expressed in the root tissue compared to the nodule tissue. No such pairs of blocks were found that were enriched for over-expression in the nodule tissue compared to the root. There was one pair of homologous blocks that was enriched for genes over-expressed in the leaf tissue compared to the flower, and four blocks that were enriched for genes over-expressed in the flower tissue compared to the leaf tissue.
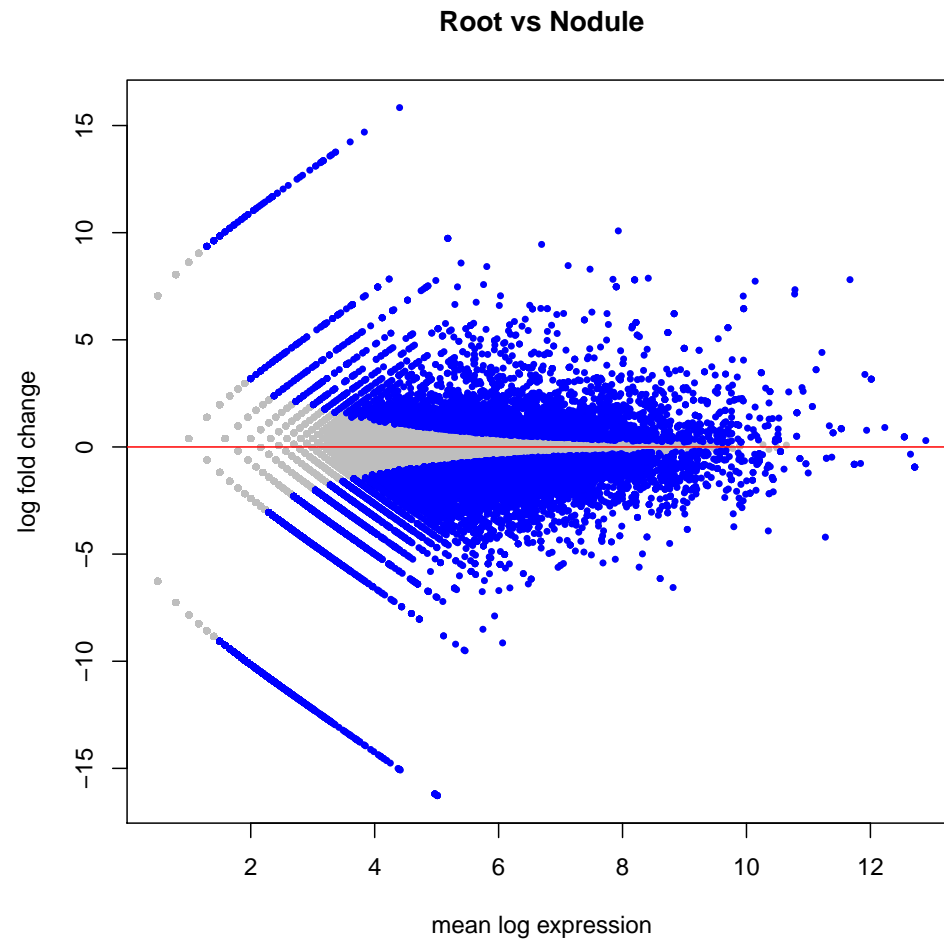
Root vs Nodule

Figure 3: The $\log_2$ fold change, between the root and nodule tissues, is plotted on the y-axis, and the mean $\log_2$ expression is plotted on the x-axis. Gene expression counts were normalized by the column totals of the corresponding 2×2 table. Blue dots represent significantly differentially expressed genes as established by Fishers Exact Test; grey dots represent genes with similar expression. The red horizontal line at zero provides a visual check for symmetry. The plot appears symmetric suggesting that overall both tissues have approximately the same number of over- and under-expressed genes. Image taken from Auer and Doerge (2010).

6   Discussion

The comparison of expression between duplicate genes within tissue enjoys a particular advantage that is due to the study design. Recall that each mRNA sample from each tissue is sequenced in a different lane of the Solexa flow cell. Since the comparisons of duplicate genes and duplicate sets of genes (i.e., homologous blocks) take place within tissue, any differences between tissues are irrelevant to the analysis. This most certainly adds power to the statistical analysis by not
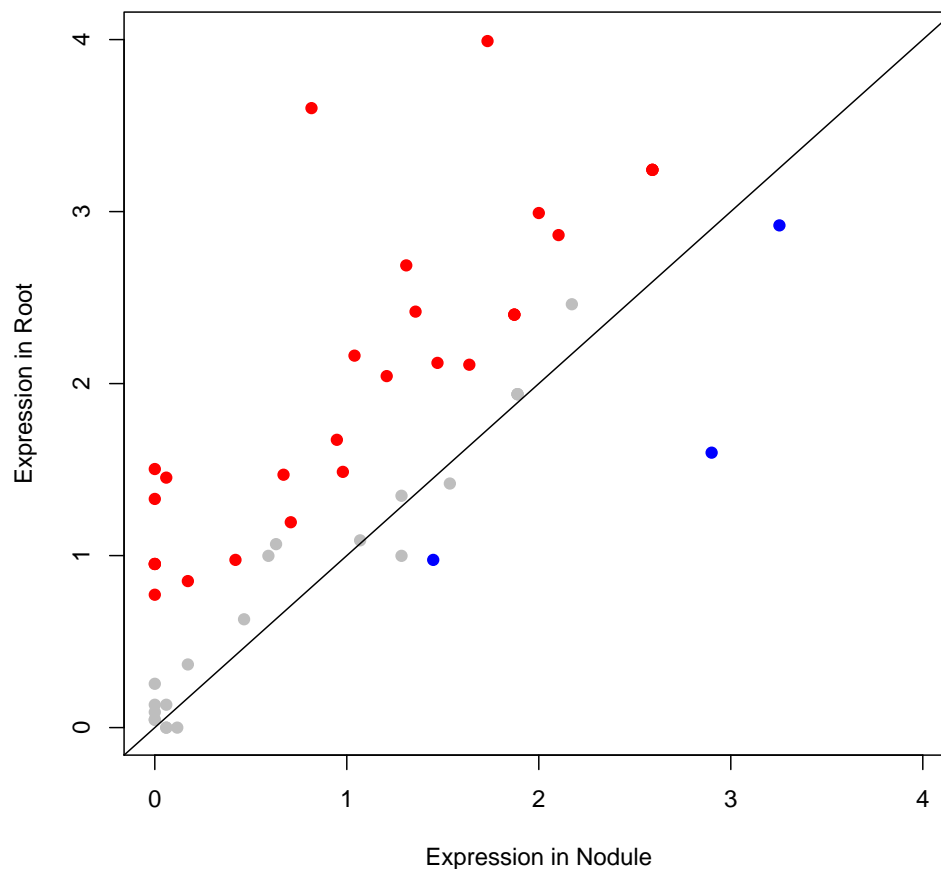
Figure 4: The results of the differential expression analysis of the root versus nodule tissue comparison within a particular pair of homologous blocks. The $\log_2$ normalized gene expression levels for the root tissue are plotted on the y-axis, the corresponding values for the nodule tissue are plotted on the x-axis. Red points represent genes found over-expressed in the root tissue over the nodule tissue, blue points represent genes found over-expressed in the nodule tissue over the root tissue, and grey points represent genes found similarly expressed between the two tissues. The diagonal identity line provides a visual check for symmetry. Notice that there are many more red points than the others, suggesting that overall the genes in this pair of homologous blocks are over-expressed in the root tissue compared to the nodule tissue.

introducing extra variation to the statistical tests for differential expression of duplicate genes. Furthermore, any systematic lane-to-lane variation (i.e., technical variation) is also irrelevant to this analysis, since we only make within-lane comparisons. However, the analysis also lacks key inferential capabilities due to the study design. Since the RNA was captured from multiple plants but then pooled and sequenced as a single sample, there is no way of estimating variability among

different plants. In other words, no meaningful biological replication occurred in this design. The ECT assumes that the within tissue variability is well approximated by the Poisson distribution, but there is no way to verify this assumption. If the within tissue biological variability between plants is more than that expected by the Poisson assumption, then the analysis risks making too many type I errors. In this case, this amounts to calling duplicate pairs of genes differentially expressed, when in fact they are not. A type I error at the gene level would propagate throughout the rest of the analysis onto the duplicate block level, potentially providing misleading results for duplicate blocks as well as duplicate genes.

The comparisons of gene expression between tissues are also severely limited. Since there was no meaningful biological replication, it is impossible to estimate within-tissue variability. This precludes us from making inferences about differences between tissues (Fisher, 1935a). Since sequencing with NGS is very expensive, biological replication is often not possible. In these instances, the results from the analysis should not be generalized, although they may still be useful for bench scientists by providing a list of "interesting genes" worthy of follow-up. Replication notwithstanding, in this design the lanes of the Illumina flow cell are confounded with the tissue types. Since there is no way to separate systematic lane-to-lane variation from the effects of the different tissues, it is not clear that observed differences between tissues can be attributed to a biological effect. However, the confounding of effects is not inevitable with unreplicated data; given this experimental setup other designs are possible that eliminate the possibility of confounding technical artifacts with true biological differences (Auer and Doerge, 2010; Auer, 2010).

This type of genome-wide analysis of duplicate genes in soybean has only become possible very recently, with the advent of NGS. It requires both the reference sequence of the soybean genome (which was released in January 2010, and was partially obtained using NGS), and a high-throughput sequencing based approach to mapping and quantifying transcripts (i.e., RNA-Seq). First, without the reference sequence, duplicate genes and duplicate blocks of genes cannot be defined. Second, since the majority of 36 base sequencing reads from RNA-Seq can be mapped uniquely to a specific genomic location, RNA-Seq data can differentiate the transcription products of duplicate genes. Other technologies for generating genome-wide data (e.g., microarrays) are not sensitive enough to accurately distinguish transcripts with very similar sequences. This analysis is indicative of the power of NGS (both in providing a reference sequence and generating data) to transform the types of questions that can be investigated on a genome-wide scale. Furthermore, with just one set of data we were able to investigate several different research hypotheses; testing both within tissue comparisons of duplicate gene pairs as well as across tissue comparisons of each gene.

The cost of NGS experiments is rapidly declining, allowing for the potential to design customized, optimal experiments for particular research questions. Although some work in this area was done for microarrays (Kerr and Churchill, 2001b,a), the extension to NGS is not straightforward. Additionally, since NGS platforms are updated frequently, proper experimental designs should be both appropriate to the physical layout of the platform and flexible enough to accomodate changes. As NGS and whole-genome data become more prevalent in the biological and agricultural sciences, the development of statistical methods and bioinformatic tools must keep pace. NGS data are least an order of magnitude larger than microarrays, perhaps limiting the abil-

ity of many statisticians to contribute to this new field. But just as the statistical issues related to microarray data helped to advance statistical and computational methods in high-dimensional data analysis, design, and hypothesis testing, we anticipate the recent influx of NGS data to have a similar effect in the coming years.

7 Summary

Much of the microarray based genomic research in the agricultural sciences is transitioning to NGS applications. In this project, both the reference sequence of the soybean genome and a particular set of RNA-Seq data were used to address several different biological questions. This analysis of gene expression in soybean exemplifies the type of analysis that is required for NGS data. The analysis featured classical statistical methods developed for contingency tables, modern methods for controlling the FDR, and a host of computational tools for processing the raw data and visualizing the results.

8 Acknowledgments

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, Hoboken, New Jersey, 2 edition.

Auer, P. L. (2010). *Statistical Design and Analysis of Next-Generation Sequencing Data*. Ph.D. dissertation, Purdue University, West Lafayette, IN, USA.

Auer, P. L. and Doerge, R. W. (2010). Statistical design and analysis of RNA-Seq data. *Genetics* **185**:405–416.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statisitical Society, Series B* **57**:289–300.

Blanc, G. and Wolfe, K. H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *The Plant Cell* **16**:1679–1691.

Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M., and Gilad, Y. (2010). Sex-specific and lineage-specific alternative splicing in primates. *Genome Research* **20**:180–189.

Crick, F. (1970). Central dogma of molecular biology. *Nature* **227**:561–563.

de Peer, Y. V. (2004). Computational approaches to unveiling ancient genome duplications. *Nature Reviews Genetics* **5**:752–763.

Fisher, R. A. (1935a). *The Design of Experiments*. Oliver and Boyd, Edinburgh, 2 edition.

Fisher, R. A. (1935b). The logic of inductive inference. *Journal of the Royal Statistical Society* **98**:39–82.

Griffiths, A. J. F., Wessler, S. R., Lewontin, R. C., and Carroll, S. B. (2008). *Introduction to Genetic Analysis*. W.H. Freeeman and Company, New York, 9 edition.

Illumina (2010). `http://www.illumina.com`.

Kerr, M. K. and Churchill, G. A. (2001a). Experimental design for gene expression microarrays. *Biostatistics* **2**:183–2001.

Kerr, M. K. and Churchill, G. A. (2001b). Statistical design and the analysis of gene expression microarray data. *Genetics Research* **77**:123–128.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-Seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18**:1509–1517.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, New York, 2 edition.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**:621–628.

Oshlack, A. and Wakefield, M. J. (2009). Transcript length bias in RNA-Seq data confounds systems biology. *Biology Direct* **4**:14.

Przyborowski, J. and Wilenski, H. (1940). Homogeneity of results in testing samples from Poisson series. *Biometrika* **31**:313–323.

Saiki, R., Gelfand, D., Stoffel, S., Scharf, S., Higuchi, R., Horn, G., Mullis, K., and Erlich, H. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**:487 – 491.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* **463**:178–183.

Shultz, J. L., Kurunam, D., Shopinski, K., Iqbal, M. J., Kazi, S., Zobrist, K., Bashir, R., Yaegashi, S., Lavu, N., Afzal, A. J., et al. (2006). The soybean genome database (soygd): a browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of glycine max. *Nucleic Acids Research* **34**:D758–D765.

Simillion, C., Janssens, K., Sterck, L., and de Peer, Y. V. (2008). i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* **24**:127–128.

Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**:1859–1875.

Young, M. D., Wakefield, M. J., Smyth, G. K., and Alici (2010). Gene ontology analysis for RNA-Seq: accounting for selection bias. *Genome Biology* **11**:R14.