Journal of
Data Mining in Genomics & Proteomics

Yin et al., J Data Mining Genomics Proteomics 2013, 4:4
DOI: 10.4172/2153-0602.1000139

Research Article Open Access

Visual Mining Methods for RNA-Seq Data: Data Structure, Dispersion Estimation and Significance Testing

Tengfei Yin[1]*, Mahbubul Majumder[2], Niladri Roy Chowdhury[2], Dianne Cook[2], Randy Shoemaker[3] and Michelle Graham[3]

[1]*Department of GDCB, Virtual Reality Applications Center, Iowa State University, USA*
[2]*Department of Statistics, Iowa State University, USA*
[3]*USDA-Agricultural Research Service, Corn Insects and Crop Genetics Research Unit and Department of Agronomy, Iowa State University, USA*

Abstract

In an analysis of RNA-Seq data from soybeans, initial significance testing using one software package produced very different gene lists from those yielded by another. How can this happen? This paper demonstrates how the disparities between the results were investigated, and can be explained. This type of contradiction can occur more generally in high-throughput analyses. To explore the model fitting and hypothesis testing, we implemented an interactive graphic that allows the exploration of the effect of dispersion estimation on the overall estimation of variance and differential expression tests. In addition, we propose a new procedure to test for the presence of any structure in biological data.

Keywords: RNA-seq; Differential expression; Dispersion estimation; Visual mining; Lineup inference

Introduction

RNA-seq is a high-throughput sequencing technology used to measure differential expression (DE) of genes. RNA or cDNA samples are broken into small fragments from which short nucleotide sequences are generated.

These fragments are aligned with their corresponding genic sequences and the number of fragments (counts) aligning to a gene provides a relative estimate of the level of expression of that gene under specified experimental conditions. Comparing these counts across several treatments gives insights into how the genome is responding to different treatments.

Many commonly used and freely available software tools have been developed for DE analysis of RNA-seq data. Most work directly on the count data, such as Cufflinks [1], Bioconductor packages edgeR [2], DESeq [3] and baySeq [4]. Some analysts are transforming count data into approximately normally shaped data and using software developed for microarray data, like the limma package with function voom [5]. Recent studies conducting comprehensive comparisons across software packages report that there is no single optimal approach [6,7]. Cu di does not perform as well probably due to its normalization procedure accounting for isoform-specific information. Interestingly, methods based on a variance-stabilizing transformation combined with limma (e.g voom+limma and vst+limma) perform well in general and are not sensitive to effects caused by outliers [6]. In addition, compare to Cu inks, the Bioconductor packages edgeR and DESeq support more complex multi-factor experiments.

There are many factors that may directly impact DE analysis such as normalization, counting, alignment, sequencing depth and sample size. This paper focuses on the effect of dispersion estimation. The packages edgeR and DESeq are designed to adjust for overdispersion, which occurs when variance cross biological replicates is larger than mean expression [8-10]. They use a negative binomial model which defines the relationship between variance and mean as=+2, where is the dispersion factor. In the real world, it is more likely that the individual gene has its own dispersion factor. In order to conduct hypothesis tests the dispersion must be estimated for each gene, which requires sharing information across genes due to the massive multiple testing and few replicates common to these types of projects. Different approaches for estimating dispersion are available. edgeR moderates gene-specific dispersion towards common/trended dispersion effect modeled by mean-variance relationship, while DESeq takes the maximum of individual dispersion and the trended dispersion, that making DESeq more conservative and edgeR more sensitive to outliers [11]. It is reported that various parameter settings in edgeR or DESeq could vary the results of DE analysis a lot, in terms of false discovery rate, type I error control and truly DE genes detection, probably due to inaccuracies in the estimation of the mean and dispersion parameters [6]. Cu di model the single-isoform gene variance similarly to DESeq approach, and uses a mixture model of negative binomial with the beta distribution parameters as mixture-weights for multi-isoforms genes [7].

In edgeR, common dispersion assumes that all genes share the same dispersion-too simplistic but useful as a baseline quantity. Trended dispersion is estimated by the fitted value from a smooth performed on a plot of binned common dispersion versus average abundance. This yields dispersion estimates for genes with similar average count, effectively averaging dispersion of nearest neighbor genes. The dispersion estimate is finalized by a weighted likelihood empirical Bayes approach [8], that shrinks the tagwise dispersion towards the common dispersion or trended dispersion. The prior degrees of freedom (parameter prior.df) indicates the weight given to the prior, and the larger the prior.df, the more the tagwise dispersion are squeezed towards the common/trended dispersion. The choice of final dispersion value used for each gene can affect significance, since the dispersion is the ruler upon which treatment mean differences are measured. Once the dispersion has been calculated, a modified generalized linear model (GLM), developed for multifactor RNA-seq

***Corresponding author:** Tengfei Yin, Department of GDCB, Virtual Reality Applications Center, Iowa State University, 1620 Howe Hall 2274, Ames, IA 50011-2274, USA, E-mail: tengfei@iastate.edu

Received May 16, 2013; **Accepted** August 20, 2013; **Published** August 28, 2013

Citation: Yin T, Majumder M, Chowdhury NR, Cook D, Shoemaker R, et al. (2013) Visual Mining Methods for RNA-Seq Data: Data Structure, Dispersion Estimation and Significance Testing. J Data Mining Genomics Proteomics 4: 139. doi:10.4172/2153-0602.1000139

experiments, where a complex design can be specified [12], can be used to identify genes differentially expressed between treatments. The significance of any co-efficient, or a contrast of treatments in the linear model can be performed using likelihood ratio statistics or quasi-likelihood statistics. Benchmark data demonstrates that shrink tagwise estimates towards trended estimates gave better results than shrinking towards common dispersion estimates, and GLM method usually find more significant genes than exact test [6].

In this paper, we discuss the effect of dispersion estimation on RNA-seq differential expression analysis. The data used in this paper comes from a study on the iron efficiency response in two genotypes of soybean. Primary interest is on the iron response of one genotype (RPA), where Replication Protein A Subunit 3 has been silenced by a virus. The other genotype (EV) is the control which was infected by virus having an empty vector [13]. Several new visual tools are developed to help this investigation. Interactive plots enable the large data sets to be explored using traditional statistical plots. We can create plots of the dispersion under different scenarios and display the resulting p-values. Mousing over these plots brings up a classic model diagnostic plot for the gene in focus, in a relatively seamless manner, so that many of these individual gene charts can be viewed quickly. In addition, a new visual hypothesis test is conducted by comparing significant genes, with the most significant findings from data where structure has been removed.

The results section describes the experiment that brought the issue to our attention, and an explanation for the contradiction in results we found from using different dispersion estimation methods. It illustrates the interactive graphics capabilities that helped to examine the results, and that will help digest results in larger, complex RNA-seq studies. The section also describes the results of a visual hypothesis test to determine presence or absence of any structure in RNA-seq data. For plant data, like the soybean, where results can be complicated by genome duplications and environmental effects, this can be helpful guide for troubleshooting downstream analyses. It should be noted that a key feature of the visual analysis is the importance of plotting the raw data, in conjunction with estimated elements. Section 2 on materials and methods describes the dispersion estimation methods used, the methods behind the interactive graphics and the visual hypothesis test, and software used.

Materials and Methods

The RNA libraries were prepared and then sequenced by Illumina (http://www.illumina.com/) equipment [13]. The 12 raw fastq les were then aligned by bowtie2 [14], and output to 12 bam format les. The bam les are available at: National Center for Biotechnology Short Read Archive (NCBI SRA Bioproject accession PRJNA190191 1). As documented in Atwood et al. [13], one sample was removed suring pre-processing for quality reasons, resulting in 11 samples. To use Cuffdiff, we split the data into two groups based on treatment to test iron condition effect, renamed the bam les to reflect the experimental design, and then perform two independent analyses for each subset. For the analysis using edgeR, the package Rsamtools was used to import the bam les and the package rtracklayer was used to import the gene features gff file. The package GenomicRanges was used to count reads for genes and output a matrix containing gene counts for each sample. All samples were analyzed together using a negative binomial generalized linear model. False discovery rate methods gave the final gene lists for each analysis.

To determine the effect of dispersion of DE analysis, we used edgeR in two different ways. In method 1, the approach mirrors the Cuffdiff analysis, and the samples corresponding to treatments EV and RPA are separated. Tests for differential expression are run on each subset. In method 2, both treatment and condition are analyzed together. Method 1 treats the analysis like two separate single factor experiments, while method 2 treats the experiment as a 2X2 factorial design, followed by contrasts for checking specific effects. A major difference between the two approaches is the way dispersion is estimated. With method 1 dispersion is estimated separately for each treatment, but for method 2 it is estimated using both treatments, all samples. edgeR is used with negative bionomial GLM method, tagwise dispersion is squeezed towards trended dispersion instead of common dispersion.

The interactive graphics are programmed in R is using package cranvas [15], which is back-ended by packages qtpaint [16] and qtbase [17], handling the graphical elements using Qt libraries. Linking is controlled by the package plumbr [18], which registers the signal generated from the scatterplot. A function is attached to this signal which retrieves the data for the selected gene and generates the static interaction and model estimates plots using the ggplot2 [19] package. The response rate is fast, and performance with this size of data is quite reasonable to get a good overview of the data quickly. The packages cranvas, qtpaint, qtbase are available for most linux distributions and the Mac operating system, but not Windows yet. All of the other packages are available across all platforms. The R script for producing interactive diagnostic graphics is available at bitbucket (https://bitbucket.org/yintengfei/paper_jdmgp_soybean/src/).

The inference test for structure involved recruiting independent observers using Amazon's Mechanical Turk [20]. Amazon's Mechanical Turk is used for tasks that humans can do better than computers, which is the case for reading statistical plots. The web page for this project is http://www.public.iastate.edu/~mahbub/feedback_turk9/homepage.html. To test for the signficance of a treatment or interaction effect, in the RNA-Seq data, multiple lineups are generated and posted on this site. Observers saw three lineups: a very easy one generated from simulated data, and two containing data from the RNA-Seq experiment, with the most significantly expressed genes on treatment and interaction. The simulated data is used as a filter; responses from subjects who correctly pick the most structured plot in this lineup are kept for the two real lineups. For each of the treatment and interaction lineups, multiple versions were available, and chosen randomly to show a subject. Versions were made with the observed data plot placed in different locations on the page, and amongst different null plots. By doing this, we ensure that if the observed data is detectable, then it does not depend on where it was placed in the lineup, or what comparisons were used. Buja et al. [21] introduced the idea, Majumder et al. [22] validated the approach in controlled conditions, and Chowdhury et al. [23] used the approach to examine HDLSS data in controlled conditions.

Results

Data, experiment, contradictions

As can be seen from Table 1, the experimental design is a 2X2 factorial design. There are two treatments (RPA, Empty Vector), two

	Iron Condition	
Treatment	I	S
RPA	1,2,3	1,2,3
EV	1,2,3	1,2,3

Table 1: A 2X2 factorial experimental design with two treatments and two iron conditions, three biological replicates in each.

conditions (iron insufficient and sufficient), and three replicates in all but EV, iron insufficient which had two. In the two treatments, virus induced gene silencing was used to silence the expression of a DNA replication gene RPA (GmRPA3c). An empty vector (EV) control was used to control for the effect of the virus on plant growth and development [13]. The goal of the project was to identify the genes significantly affected by RPA (treatment), and/or iron availability (condition) among the greater than 40,000 genes in the soybean genome.

In other words, the question of interest is whether the two treatments responded differently to the two different conditions, and if so, which genes were responsible. That is, which genes have different expression patterns for the two treatments, particularly in regard to iron conditions.

Figure 1 illustrates the results. For method 1, treatment EV was reported to have 304 differentially expressed genes on iron condition, and treatment RPA had only 75 differentially expressed genes. An argument could be made that EV was responding more feverishly, either activating or de-activating genes, to the two conditions. But repeating the analysis using edger, flipped these results. With this method, EV was reported to have 90 differentially expressed genes and RPA, many more at 133 differentially expressed genes. Using only method 2, we might argue that RPA was responding more vigorously to the two conditions. The two methods report paradoxical conclusions.

Proposition: Dispersion for RPA is substantially greater than dispersion of EV. This would account for the contradictory numbers of significant genes based on analyzing data together or separately.

All tests for differences between treatments were constructed by measuring the difference between means using a ruler calibrated by the variance. The mean calculations are straightforward, and typically always the same. It is the variance estimation which can differ. When there is one experiment, the variances are calculated on the replicates of each treatment, and these are used to gauge the difference between the means. In RNA-Seq data, there are many genes being tested simultaneously, thought of as many simultaneous experiments. Therefore, the dispersion calculation takes into account the shared dispersion of all the genes, along with that of the treatment replicates for the individual gene. Figure 2 illustrates some of the ways that variance estimation difference can affect the interpretation of the difference between means.

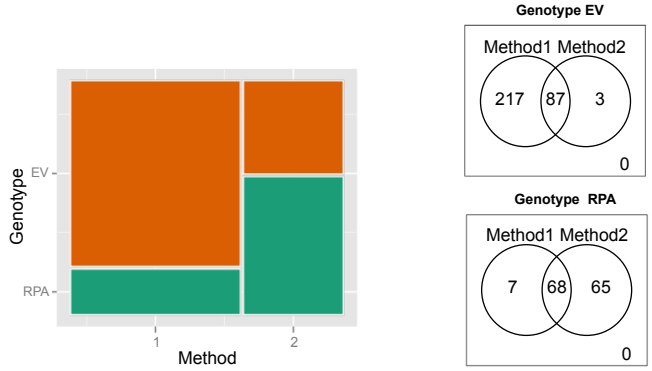

Figure 1: Representation of counts of genes found to be signi_cant for iron condition for treatments EV and PRA, based on two di_erent approaches: (left) mosaic plot, and (right) conventional Venn diagrams of signi_cant genes.

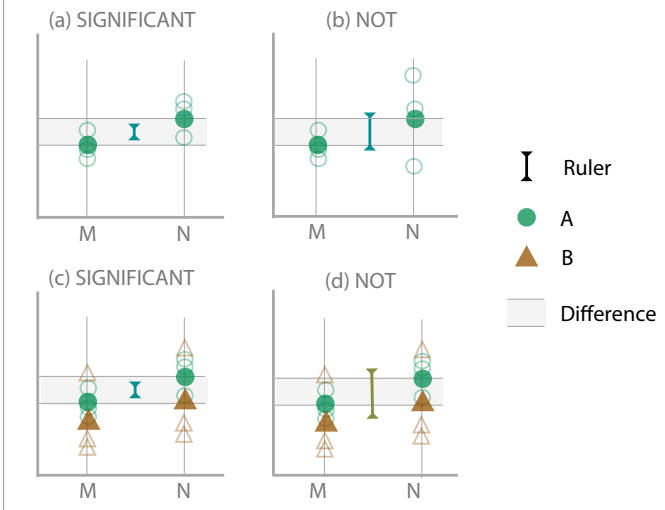

Figure 2: Diagram illustrating how estimation of the variance can a_ect the interpretation of the mean di_erence, under di_erent scenarios. (a) Conventional situation, where each condition (M, N) has the same small variance, mean di_erence is viewed as signi_cant, because it is big relative to the small variance. (b) One condition (N) has higher variance than the other, ignoring the di_erence, and pooling the variances would increase the length of the ruler upon which the di_erence is gauged, to the extent that the di_erence is not considered to be signi_cant. (c) Addition of an extra treatment level (B), that has larger variance, but this has no e_ect because each level of the treatments (A, B) is handled as separate samples { di_erence is still seen as signi_cant. (d) Variances across treatments (A, B) are pooled, which increases the length of the ruler upon which the di_erence is measured, resulting in no signi_cant di_erence.

We estimated an appropriate measure of dispersion for the soybean data, common biological coefficient of variation (BCV) [12], using the two different methods, one for all 11 samples and one for split data for different treatments (6 samples in RPA and 5 samples in EV). In addition, we estimated BCV for each factor combination to identify which factor contributed more to the common dispersion. The results are shown in Table 2. In general, treatment RPA has higher common dispersion, especially under insufficient iron conditions.

If the data is split and dispersion is estimated separately for each treatment, the ruler used to assess the difference between iron condition means for EV is smaller (0.076) than that for RPA (0.151), and thus more genes will be detected as different (304 vs 75). On the other hand, if dispersion is estimated across both treatments the ruler will be medium sized (0.095), which changes the interpretation of how big a difference there is between the two iron condition means. This is why the count of significant genes flips from 90 to 133 for the two treatments, EV and RPA, respectively when using different programs.

Interactive diagnostic graphics

To explain the diagnostic procedures, we will backup a few steps, to the start of the analysis. A commonly used diagnostic plot for DE analysis is to examine the mean-variance relationship and plot the biological coefficient of variation (BCV) against the mean abundance, on a $\log_2$ scale. BCV is the square root of the tagwise estimated dispersion [12]. This is a scatterplot, where one point represents a single gene. When created using the R package, cranvas, this plot is interactive: mousing over the plot, or clicking, will create an event which can be used to induce changes in other charts. An interaction plot is linked to the scatter plot for this experiment, because it is ideal

for examining the results of the 2X2 factor experiment. For a 2X2 experiment, the interaction plot contains points for each gene for each treatment and replicate. There are three replicates for each treatment, except for one treatment which only has two. This gives 11 data points for each gene. The horizontal axis is the iron condition, and the vertical axis displays $\log_2$ count per million. Color and symbol indicate VIGS treatment. A line connects the means for the treatments. Examining this line indicates important aspects of the results, with the most interesting being whether the treatments are responding differently to the iron condition, as would be indicated by different slopes of these lines.

Figure 3 illustrates the scatter plot linked with an interaction plot of the raw data, and a plot of the estimated means generated by the GLM. To begin, the pattern in the BCV *vs* abundance plot should be examined. It is expected that as abundance increases dispersion decreases. For genes with small mean abundance, dispersion varies a lot. There is a curious string of points with high abundance, which cluster apart from the others as having unusually high dispersion (explored using linked brushing in Figure 4). From later investigation, this cluster contains genes that come from repeated elements of the genome, and in regions not of interest, so they were removed in later analysis.

In Figure 3, in the BCV *vs* abundance scatterplots, yellow indicates the gene that is the focus of the user's interaction-the user has actively selected this point to investigate. When the point is selected, the interaction plot for the respective gene is shown. In the top row, a gene (call this A) that has a high overall mean abundance, but relatively low dispersion is highlighted, and in the bottom row a gene (call this B) with similar mean, but relatively higher dispersion is highlighted. The gene with relatively low dispersion would be considered the more interesting gene, and should emerge from the hypothesis testing as more significantly expressed than the other. The difference in dispersion is visible in the interaction plots. Gene B on RPA treatment has one replicate with an unusually high value on the iron sufficient condition, which likely resulted in the high dispersion value. The plot of the model estimates from GLM in edgeR on the right shows what the model sees when expression is fitted to the treatment levels. The model estimates don't seem to match the raw data. The estimated mean for RPA (green) is pulled towards the replicate with the extremely high value, which suggests something strange is happening with the model. In Figure 4, two genes from the strange cluster are highlighted, and from the interaction plots, we can see the pattern of each is almost identical, supporting the conclusion that they are from repeated elements of the genome where counting of reads would be problematic.

Figures 5-7 illustrate the effect of the shrinkage parameter (prior.df) on the dispersion calculation, and hence, the p-value and the significance of genes. Two different shrinkage parameters (1 and 10) are chosen. Here, we have chosen a very small value (0), and the default value (10). The top two plots on the left in each figure provide a comparison of the two p-values that would result from the two different shrinkage values, on full scale and zoomed in to small values. If the shrinkage parameter did not matter, the points would lie very close to an X=Y line. We can see that there is a positive linear association, a low p-value for initial shrinkage corresponds to a low p-value for another, but the spread is much larger than we might expect. This says that the p-values are changing substantially in relation to shrinkage, and popping in and out of significance. Three different genes are highlighted. In Figure 5a, gene that has low initial p-value that increases for the converged shrinkage value, so that it would initially be a candidate for an interesting gene, but is dropped from the list in the final analysis. The next two plots shows the BCV for the two different shrinkage values, and it can be seen that the BCV initially is close to 0, and is increased to 0.27 with the change in shrinkage. The last two plots show the raw data interaction plot (left), and the model estimates with dispersion shown as bars. Shrinkage with prior.df=10 pulls the tagwise dispersion towards the trended dispersion value, effectively reducing significance of this gene. Figure 6 shows a gene with the opposite pattern, one that initially has large dispersion, but this is reduced substantially to increase the significance from 0.06 to about 0.005. Figure 7 shows a gene for which there is little effect of the shrinkage. It is a gene with a very small p-value, which doesn't change much.

	Iron Condition		
Treatment	I	S	Pooled
EA	0.114	0.042	0.076
RPA	0.176	0.108	0.151
Pooled	0.174	0.087	0.095

Table 2: Dispersion for all possible combinations of factor levels.Generally, larger dispersion is observed with treatment RPA, treatment RPA under condition iron insufficient has largest dispersion.

The interactive graphics demonstrated here enable the analyst to investigate the effects of the choices that they make in the data pipeline. Particularly, the effect on significance needs to be understood to ensure reliable results in these large high-throughput studies.

Lineup inference

RNA-Seq data is an example of high-dimension low sample size (HDLSS) data. In this type of data, the high-dimensionality can dwarf any signal in the data. Chowdhury et al. [23] examined whether differences between groups are visible in the presence of many noise variables. This is equivalent to the multiple testing involved in identifying gene expression differences in RNA-seq data analysis. Chowdhury et al. [23] used the lineup protocol described in Buja et al. [21]. The lineup protocol is a very new approach to test discoveries made using visualization while data mining.

Figure 8 shows a lineup constructed on the 2X2 experiment described in this paper. There are 20 interaction plots laid out in a grid. One of the plots displays actual data and the others show what might be seen when data is randomized (null plots). The actual data plot shows the most significantly expressed gene from a test of whether RPA (green) silencing affects expression of the gene depending on iron condition, but EV (orange) does not. This corresponds to a pattern where the slope of the green line is steep, and the spread of the green points is small. The null plots are generated by permuting the experimental design (Table 3). The full analysis is conducted on this permuted data, and the most significant gene is recorded and plotted. The process is repeated 19 times to give the 19 null plots. These represent the most extreme patterns we might see if there is no treatment effect. An independent judge is employed to examine the lineup and choose the plot with the most structure. If the observed data is selected by the judge, this is equivalent to rejecting a null hypothesis. The null hypothesis for this lineup is "RPA silencing does not affect gene expression", that is, there are NO genes on the RPA treated plants that respond differently depending on iron condition. If the null hypothesis is rejected, it is also saying that there is no structure in the data.

Why is this important? In RNA-Seq analysis, a substantial number of the genes will appear to be significant simply due to the massive multiple testing. Even with false discovery rate adjustments, some genes may still have p-values that would consider being small. With small p-values, it is so tempting to believe that the genes are

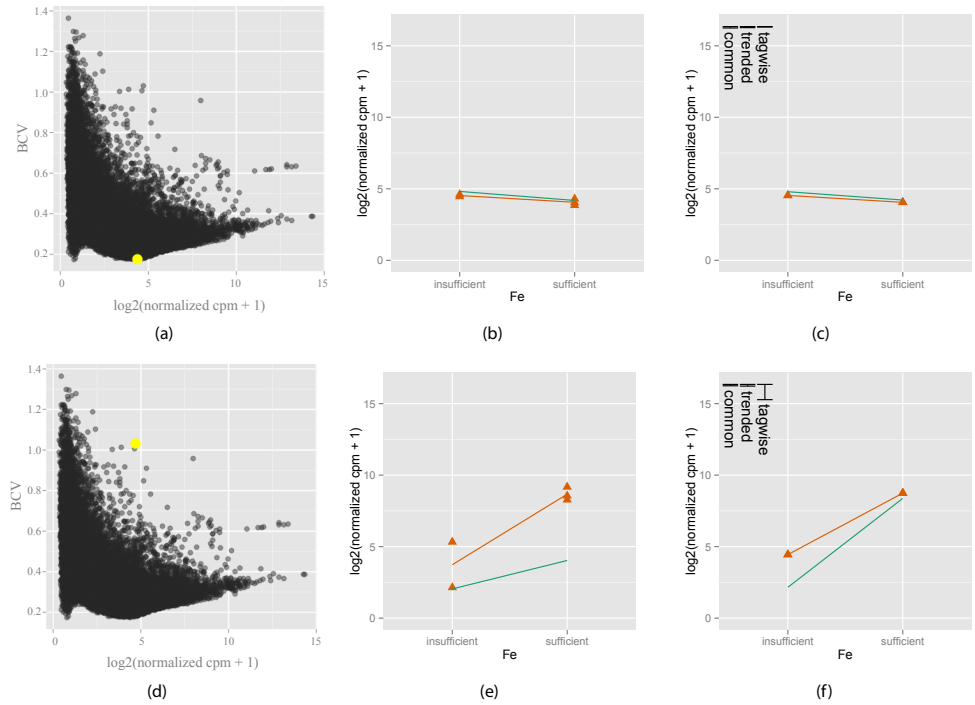

Figure 3: Illustration of linking plots to examine mean-variance relationships and gene expression. The left plots (a, d) show the mean-variance (BCV vs log2 of count per million (CPM)). One point represents one gene. The yellow point indicates highlighting of a point by identi_cation using mouse action, which immediately displays this gene in the two other plots on the right. The middle column shows interaction plots of the raw data (b, e) and the right column shows the _tted value from the model (c, f). The two genes that are highlighted have similar overall mean abundance but di_erent dispersion, one low (top) and one high (bottom). In the interaction plots, the green color represents RPA and the orange color represents EV.

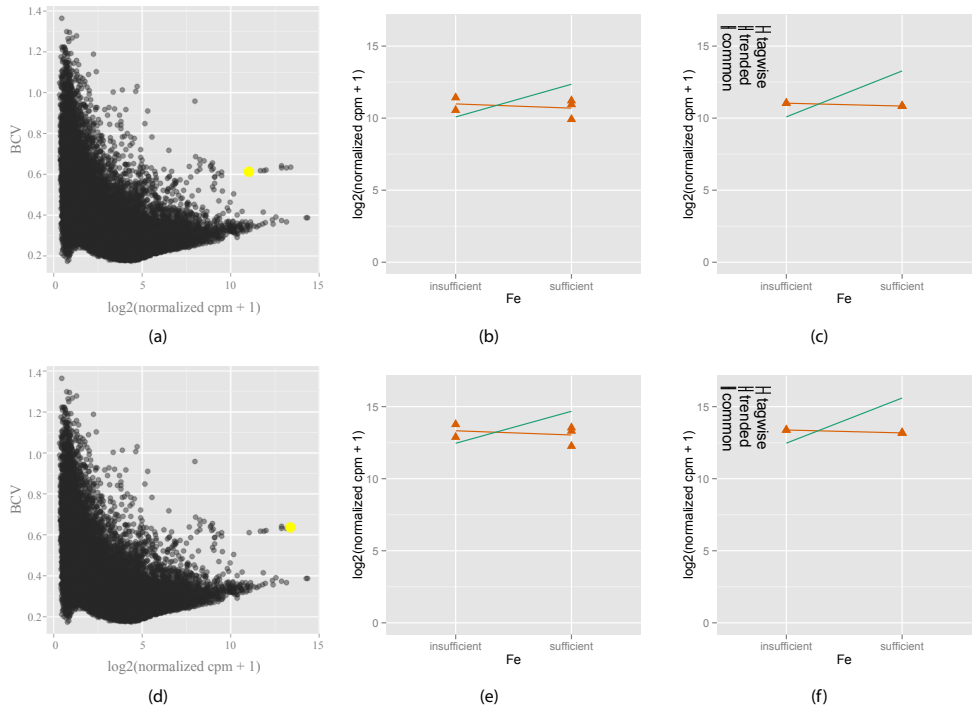

Figure 4: Illustration of linking plots to examine mean-variance relationships and gene expression for genes in a strange cluster. Two points are identi_ed in a group of outliers. Clearly they all have similar expression patterns, so do other points in that group, where one replicate has an unusually low expression value. Later investigation revealed that these genes are repetitive elements of the soybean genome, and this may have caused some counting problems. In the interaction plots, the green color represents RPA and the orange color represents EV.

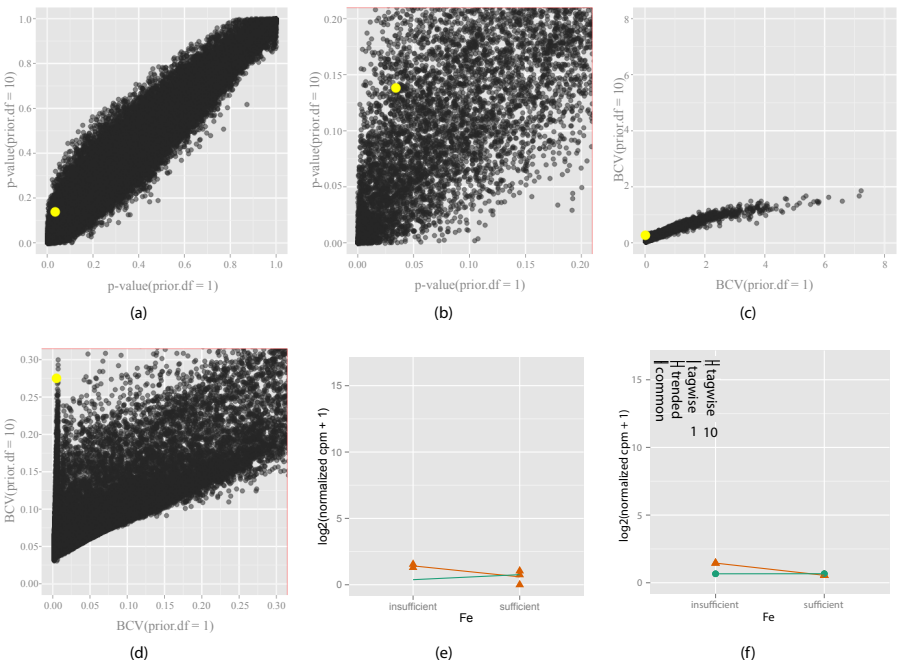

Figure 5: Exploring the e_ect of shrinkage parameter (prior.df) choice on the di_erential expression of EV (orange) when gene signi_cantly di_erentially expressed only with prior.df = 1. Two di_erent shrinkage values (prior.df = 1 and prior.df = 10) are chosen, and compared using the change in p-values (5(a)) and zoomed view (5(b)), BCV (5(c)) and zoomed view (5(d)). Yellow indicates gene under investigation, chosen by mouse action on the plot. The interaction plot (5(e)) for this gene, and the model estimates are shown (5(f)). Bars in the model estimates plot (5(f)) show the changes in di_erent dispersion estimates, for tagwise estimates, number means prior.df. The gene investigated here becomes less signi_cant from the _rst shrinkage value (1) to the second (10), as seen that it has a smaller p-value on the horizontal axis (< 0:05) and higher on the vertical axis (> 0:12) (5(b)), and the increase in tagwise dispersion (5(f)) with prior.df = 10. It makes sense because the variation in RPA (green) is larger than for EV (orange) (5(d) and 5(e)), and trended dispersion is bigger. EV should really be considered more signi_cantly expressed. Shrinkage increases the inuence of RPA dispersion on the EV dispersion estimate.

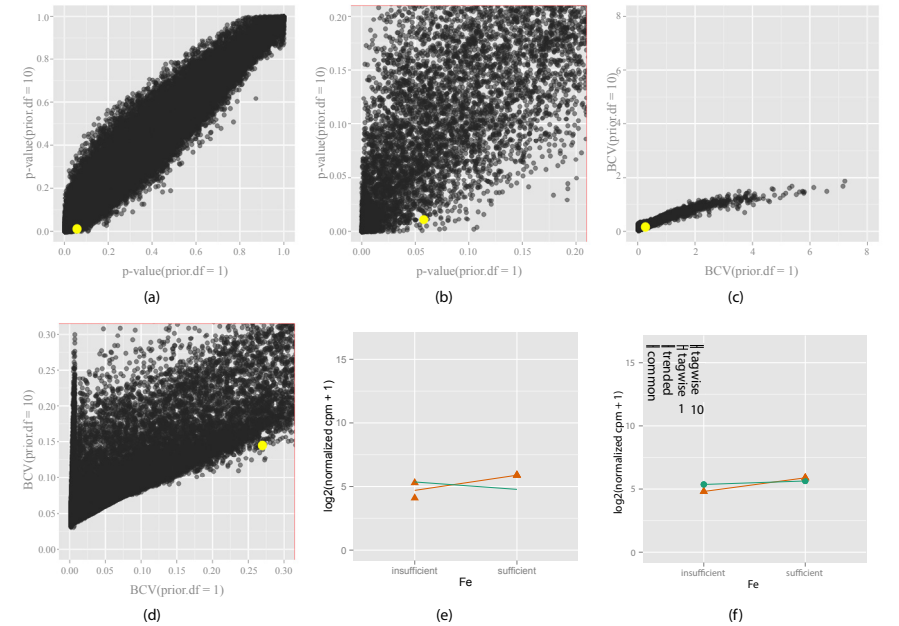

Figure 6: Exploring how the shrinkage parameter (prior.df) choice a_ects the di_erential expression of EV (orange) when gene signi_cantly di_erentially expressed only with prior.df = 10. Two di_erent shrinkage values (prior.df = 1 and prior.df = 10) are chosen, and compared using the change in p-values (6(a)) and zoomed view (6(b)), BCV (6(c)) and zoomed view (6(d)). Yellow indicates gene under investigation, chosen by mouse action on the plot. The interaction plot (6(e)) for this gene, and the model estimates are shown (6(f)). Bars in the model estimates plot (6(f)) show the changes in di_erent dispersion estimates, for tagwise estimates, number means prior. df. The gene investigated here becomes more signi_cant from the _rst shrinkage value (1) to the second (10), as seen that it has a larger p-value on the horizontal axis (> 0:05) and smaller on the vertical axis (< 0:02) (6(b)), and the decrease in tagwise dispersion (6(f)) with prior.df = 10. It makes sense because the smoothed trended dispersion for this gene is smaller (6(f)) than observed raw dispersion (6(e)). Shrinkage increases the inuence of the global smoothed dispersion on the EV dispersion estimate.

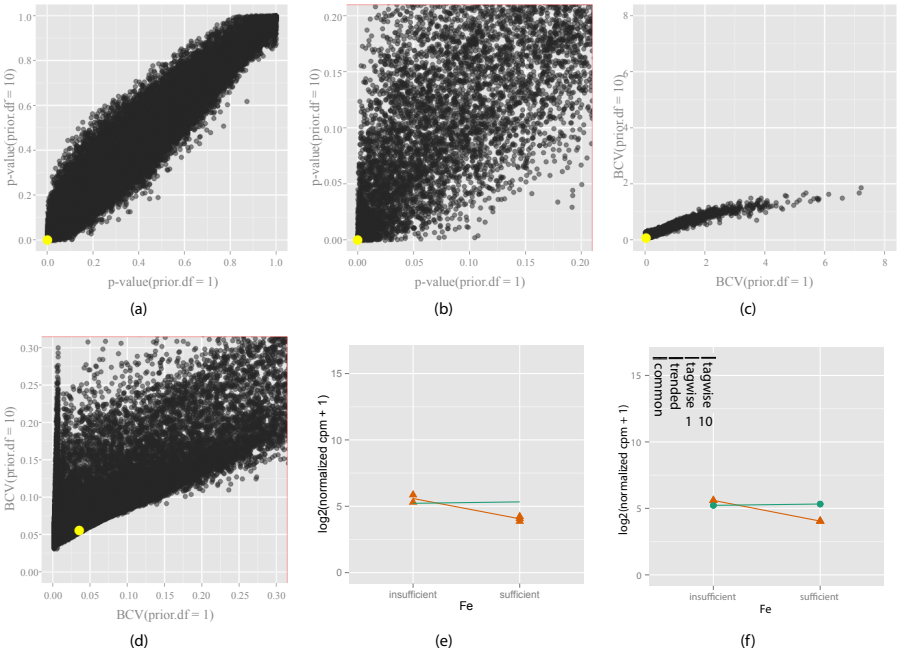

Figure 7: Exploring how the shrinkage parameter (prior.df) choice a_ects the di_erential expression of EV (orange) when gene signi_cantly di_erentially expressed with both prior.df = 1 and 10. Two di_erent shrinkage values(prior.df = 1 and prior.df = 10) are chosen, and compared using the change in p- values (7(a)) and zoomed view (7(b)), BCV (7(c)) and zoomed view (7(d)). Yellow indicates gene under investigation, chosen by mouse action on the plot. The interaction plot (7(e)) for this gene, and the model estimates are shown (7(f)). Bars in the model estimates plot (7(f)) show the changes in di_erent dispersion estimates, for tagwise estimates, number means prior.df. Signi_cance of this gene is e_ectively unchanged by shrinkage, as seen that the p-value is similar in value for both shrinkage values, and observed dispersion is small for both RPA (green) and EV (orange) (7(e)). So this is a gene that is not a_ected much by the choice of shrinkage.

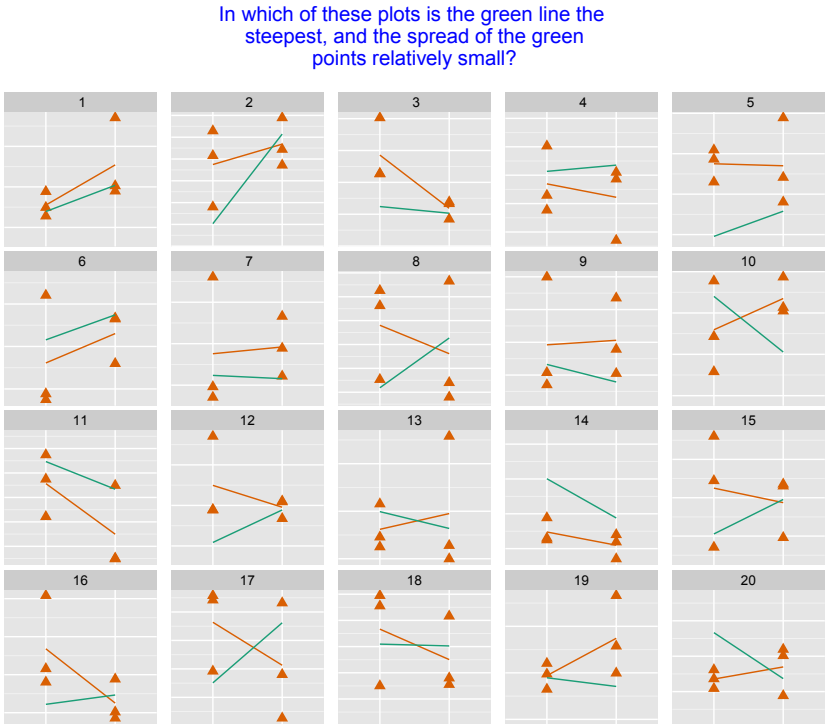

Figure 8: Lineup to examine the signi_cance of the iron e_ect on RPA (green), regardless of the e_ect on EV (orange). The question at the top of the lineup is the request put to the judges. One of the plots uses the real data, and the true experimental design, the rest use permuted treatment levels e_ectively breaking any real association between experimental factors and gene expression. In each case the gene shown is the one that has the most signi_cant di_erence between iron response on RPA.

The position of the observed data plot in the lineup is the solution of this expression $\sqrt{2^4 - 2^2 / 3}$.

Original			Permuted		
Treatment	Condition	log_2 CPM	Treatment	Condition	log_2 CPM
A	M	5.0	A	N	5.0
A	M	10.0	B	M	10.0
A	N	2.5	B	M	2.5
A	N	2.0	A	N	2.0
A	N	3.0	A	M	3.0
B	M	1.5	B	N	1.5
B	M	2.0	A	N	2.0
B	N	7.0	B	M	7.0
B	N	6.5	A	N	6.5

Table 3: Example of how permutation of experimental design is conducted. Experimental treatment labels are permuted, which breaks any association with expression. Any patterns seen in these permuted data sets is consistent with random variability.

significantly responding to the treatment. Plots of the data give additional information that portrays a different aspect of the response, the effect size, how strongly the genes respond to the treatment. Two genes that have very similar p-values may have entirely different expression differences. The lineup enables an assessment of this effect size, and closely approximates fold change, while taking p-value into account. It is important to note that although the lineup results are significant, there is nothing biologically significant about the gene that was tested. It simply happened to be the gene that had the smallest p-value in the original analysis. The lineup results say that there really is some structure in the data, based on the assessment that observers can pick the actual data plot as different from the other plots. There is something in this gene's expression pattern that is more than would be expected by chance. In the original analysis of this data [13], about 2000 genes were found to be significantly expressed, in terms of some factor in the design.

For this experiment, the results of the hypothesis testing are very strong. The p-values for testing of the presence of an interaction effect (as in the lineup shown), and the treatment effect are 0.

Discussion

The way dispersion is estimated substantially affects the significance testing in RNA-Seq data. The effect of heterogeneity between treatment groups can result in radically different, and possibly contradictory, gene lists depending on the way dispersion is estimated. It is historically known as Simpson's paradox [24], and is observed in many other types of data analyses, for example, calculating correlations across groups. It became famous with the Berkeley admissions controversy, when it was alleged that graduate programs were unfairly accepting more male applicants. Combining acceptances across colleges meant that admission rates for women were much lower than for men, but it was dismissed when admission rates in each college showed the reverse pattern.

In differential expression analysis, there are multiple sources of variation that need to be understood in order to arrive at lists of genes to investigate further. In multi-factor studies, a further source of different variance is introduced with the factor levels. Understanding these different types of variation is greatly assisted by making plots of data. It is important to make appropriate plots, which for a 2X2 factor experiment are the classical interaction plots. Using interactive graphics helps to cover the seemingly high hurdle of massive amounts of data, while still incorporating these important plots into the analysis. The interactive graphics demonstrated here enable the analyst to investigate the effects of the choices that they make in the data pipeline. In particular, the effect on significance needs to be understood to ensure reliable results in these large high-throughput studies. Other experimental design may classically use different diagnostic plots, which can be easily substituted for the interaction plots and linked accordingly. Streamlining the interactive graphics, so that they work more smoothly, and for broader types of investigation is planned for the future.

The lineup protocol, available as part of visual inference, helped to reassure us that for the experiment being investigated, that there was structure in the data with both treatments and conditions altering gene expression. It is quite possible with these large experiments that patterns found are purely due to random variation, and this new test enables this to be examined in a rigorous manner. While this is not a substitute for classical inference, visual inference enables the assessment of the effect size in the data, and a sampling of random patterns that one might find in data.

Acknowledgements

The authors are grateful for the financial support from the United Soybean Board, North Central Soybean Research Program, the Iowa Soybean Association, the NSF Plant Genome Research Program (award number 0820642), the USDA-Agricultural Research Service and National Science Foundation grant DMS 1007697.

References

1. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, et al. (2012) Differential analysis of gene regulation at transcript resolution with rna-seq. Nat Biotechnol 31: 46-53.
2. Robinson MD, McCarthy DJ, Smyth GK (2010) Edger: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.
3. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.
4. Hardcastle TJ, Kelly KA (2010) Bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11: 422.
5. Smyth GK (2004) Linear models and empirical bayes methods for assessing di erential expression in microarray experiments. Stat Appl Genet Mol Biol 3.
6. Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of rna-seq data. BMC Bioinformatics 14: 91.
7. Rapaport F, Khanin R, Liang Y, Krek A, Zumbo P, et al. (2013) Comprehensive evaluation of differential expression analysis methods for RNA-seq data.
8. Robinson MD, Smyth GK (2207) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23: 2881-2887.
9. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509-1517.
10. Ugrappa N, Wang Z, Waern K, Shou C, Raha D, et al. (2008) Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344-1349.
11. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, et al. (2013) Count-based di erential expression analysis of RNA sequencing data using r and bioconductor.
12. McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 40: 4288-4297.
13. Atwood SE, O'Rourke JA, Peiffer GA, Yin T, Majumder M, et al. (2013) Replication protein a subunit 3 and the iron efficiency response in soybean. Plant Cell Environ 3: 3.
14. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.
15. Xie Y, Hofmann H, Cook D, Cheng X, Schloerke B, et al. (2013) Cranvas: Interactive statistical graphics based on Qt.

16. Lawrence M, Sarkar D (2013) Qt-based painting infrastructure.

17. Lawrence M, Sarkar D (2012) qtbase: Interface between R and Qt.

18. Lawrence M, Wickham H (2012) Plumbr: Mutable and dynamic data models.

19. Wickham H (2008) ggplot2: An implementation of the grammar of graphics.

20. (2010) Mechanical Turk. Amazon, USA.

21. Buja A, Cook D, Hofmann H, Lawrence M, Lee E, et al. (2009) Statistical inference for exploratory data analysis and model diagnostics. Phil Trans R Soc A 367: 4361-4383.

22. Majumder M, Hofmann H, Cook D (2013) Validation of visual statistical inference, Applied to Linear Models. J Am StatAssoc.

23. Chowdhury NR, Cook D, Hofmann H, Majumder M, Lee EK, et al. (2013) Visual statistical inference for high dimension, small sample size data. Comput Stat.

24. Simpson EH (1951) The interpretation of interaction in contingency tables. J Royal Stat Soc B 13: 238-241.

This article was originally published in a special issue, **Bioinformatics for High-throughput Sequencing** handled by Editor: Dr. Heinz Ulli Weier, Lawrence Berkeley National Laboratory, USA