INFERENTIAL CONSIDERATIONS FOR LOW-COUNT RNA-SEQ TRANSCRIPTS: A CASE STUDY ON AN EDAPHIC SUBSPECIES OF DOMINANT PRAIRIE GRASS *ANDROPOGON GERARDII*

by

SETH RAITHEL

B.S., Truman State University, 2013

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2015

Approved by:
Major Professor
Dr. Nora Bello

Abstract

Big bluestem (*Andropogon gerardii*) is a wide-ranging dominant prairie grass of ecological and agricultural importance to the US Midwest while edaphic subspecies sand bluestem (*A. gerardii ssp. Hallii*) grows exclusively on sand dunes. Sand bluestem exhibits phenotypic divergence related to epicuticular properties and enhanced drought tolerance relative to big bluestem. Understanding the mechanisms underlying differential drought tolerance is relevant in the face of climate change. For bluestem subspecies, presence or absence of these phenotypes may be associated with RNA transcripts characterized by low number of read counts. So called low-count transcripts pose particular inferential challenges and are thus usually filtered out at early steps of data management protocols and ignored for analyses. In this study, we use a plasmode-based approach to assess the relative performance of alternative inferential strategies on RNA-seq transcripts, with special emphasis on low-count transcripts as motivated by differential bluestem phenotypes. Our dataset consists of RNA-seq read counts for 25,582 transcripts (60% of which are classified as low-count) collected from leaf tissue of 4 individual plants of big bluestem and 4 of sand bluestem. We also compare alternative ad-hoc data filtering techniques commonly used in RNA-seq pipelines and assess the performance of recently developed statistical methods for differential expression (DE) analysis, namely DESeq2 and edgeR robust. These methods attempt to overcome the inherently noisy behavior of low-count transcripts by either shrinkage or differential weighting of observations, respectively.

Our results indicate that proper specification of DE methods can remove the need for ad-hoc data filtering at arbitrary expression threshold, thus allowing for inference on low-count transcripts. Practical recommendations for inference are provided when low-count RNA-seq transcripts are of interest, as is the case in the comparison of subspecies of bluestem grasses.

Insights from this study may also be relevant to other applications also focused on transcripts of low expression levels.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my appreciation and gratitude to all those who helped and supported me with this report. To my advisor, Dr. Bello, who's support and guidance has been an integral part of this report and my development as a statistician. To my committee members, Dr. Gadbury and Dr. Johnson, who have dedicated time and effort to my report. I would also like to thank my coauthors including Dr. Johnson, Matt Galliart, Jennifer Shelton, and Susan Brown, without their research my research would not have been possible. I would also like to thank the Statistics department for their academic and financial support during my time at Kansas State University. In addition this project was supported by a grant funded by the United States Department of Agriculture, Abiotic Stress Program (2008-35001-04545).

Preface

The following report was prepared in the publication format for BMC Genomics.

Chapter 1 - Inferential considerations for low-count RNA-seq transcripts: a case study on an edaphic subspecies of dominant prairie grass *Andropogon gerardii*

Introduction

RNA sequencing (RNA-seq) technology has rapidly become the preferred choice for gene expression analysis as it allows for high throughput over a wide range of expression levels [1]. Yet, some features of RNA-seq data still pose considerable challenges for differential expression (DE) analysis, in particular related to transcripts characterized by low number of read counts [2, 3]. So called low-count transcripts often show large variability of logarithmic fold change (LFC) estimates and thus exhibit inherently noisier inferential behavior [2]. Thus, it is not surprising that low-count transcripts have received little attention. In fact, standard protocols for management of RNA-seq data call for removal of transcripts with read counts below predetermined, though arbitrary, expression thresholds [4]; this practice is often referred to as data filtering. As a consequence of data filtering, low-count transcripts are often excluded from DE analyses and ignored for the purpose of inference. This is problematic as filtering of RNA-seq data can cause transcription factors of low expression levels to be overlooked, despite their key role as master regulators of downstream gene expression [5].

Transcripts with low expression levels are often filtered out from data prior to DE analyses in an attempt to control noise and reduce the impact of multiple testing adjustments on power for DE detection by removal of uninformative or weakly expressed transcripts [4, 6]. Recent advances in statistical methods available for DE analyses of RNA-seq data may provide alternative approaches to deal with weakly expressed transcripts without the need for data

filtering at arbitrary expression thresholds. More specifically, DESeq2 [2], edgeR [7] and edgeR robust [8] recently proposed alternative approaches to handle extreme observations, which may unwittingly also facilitate inference on low-count transcripts. That is, rather than filtering out low-count transcripts at arbitrary threshold and excluding them from analysis, these statistical methods could be used to account for the increased uncertainty associated with low-count transcripts. As a common denominator, both DESeq2 and edgeR implement a generalized linear mixed models approach of the negative binomial family that efficiently borrows information across transcripts to moderate transcript-specific dispersion estimates [2, 8]. As an additional advantage, DESeq2 shrinks LFC estimates towards a common mean in a manner inversely proportional to the amount of information available for a transcript [2]. Little information, due to either a low level of expression or a high level of dispersion, cause transcript-specific estimates to shrink towards zero. In turn, the latest release of edgeR, namely edgeR robust, works by down weighting observations that deviate from the model fit [8], thereby dampening the effect that observations with very high or very low expression levels have on transcript-specific estimates of mean expression and dispersion. As a trade-off, edgeR robust requires explicit user specification of a shrinkage parameter, lets it default to a set predetermined value [8] that may be appropriate for some, but not all, data applications. No such user specification is required by DESeq2; rather, all necessary parameters are estimated from the data. Given these recent developments in statistical methodology, it is timely that guidelines for data filtering of low expression transcripts in protocols for RNA-seq data [4] be revisited, as the impact of this practice on DE inference remains unclear.

Our specific interest in low-count RNA-seq transcripts stems from our ongoing work with the wide-ranging dominant prairie grass big bluestem (*Andropogon gerardii*) and its

edaphic subspecies sand bluestem (*A. gerardii ssp. Hallii*). Big bluestem (BB) is a widely-distributed dominant grass of North American grasslands [9] and constitutes the main native forage grass for cattle [10]. In contrast, sand bluestem's (SB) habitat consists primarily of the Sand Hills in Nebraska [11]. Our preliminary studies [12] provided evidence for morphological and physiological differences between bluestem subspecies that are consistent with enhanced drought tolerance of SB. For instance, we observed a greater quantity of epicuticular wax (ECW) on the leaf surface of SB plants relative to that of BB [12]. Further, analysis of ECW components showed presence of approximately ~20% β-diketones on SB leaves, whereas β-diketones were absent in ECW of BB leaves [12]. Differential quantity and quality of ECW on leaf surfaces could affect heat reflectance and transmittance, thus providing differential relative advantages to heat tolerance in dry conditions. Further, ECW decreased light absorbance in sand bluestem [12], thus potentially lowering internal leaf temperature and protecting against heat stress. Taken together, our preliminary studies indicate that adaptation of SB to water-limited conditions may involve adaptation of leaf cuticle chemistry, morphology, and function. Sand bluestem's enhanced tolerance to dry conditions relative to BB [12] is of interest due to the expected increase in extreme droughts throughout Midwest grasslands [13].

In this study, we further characterize differences between bluestem subspecies at the transcriptome level. Following from the qualitative phenotype differences observed between SB and BB, we initially focused on RNA transcripts that were expressed in only one of the bluestem subspecies, while expression was absent in the other. More specifically, SB-only transcripts where expressed in sand bluestem samples but were absent (i.e. read counts = 0 for all samples) in big bluestem samples. The reverse was true for BB-only transcripts. We further noticed that these transcripts were characterized by few read counts, indicating overall low levels of

expression. Thus, for this study, we purposely defined so called low-count transcripts following the descriptive approach proposed by Bullard [14]. In total we have 25,582 transcripts. Here, low-count transcripts are transcripts below the 60^{th} percentile of least relative abundance and account for approximately 3% of total read counts (Figure 1.1). For contrast, we also defined so-called high-count transcripts, corresponding to transcripts with the top 3% relative abundance, which accounted for 60 % of total read counts (Figure 1.1). So defined, low-count transcripts and high-count transcripts were transcripts with a total read count below 462 or above 12,893, respectively, across all samples in the dataset. Transcripts expressed in one, but not the other, bluestem subspecies were all identified as low-count transcripts (Table 1.1), thus providing specific motivation to study transcripts with low expression levels.

In order to study transcripts with low expression levels, we consider new statistical methods that may account for the additional uncertainty with low-count transcripts. Both edgeR robust and DESeq2 have shown promising results in simulation studies and in selected real datasets [2, 8]. However, relative performance of these methods has often been shown to be data dependent and results may differ [8]. Thus, it is unclear how one might decide between these statistical methods for DE analysis of a specific dataset. Plasmodes have been proposed as a strategy to validate statistical methods or even assess relative performance of competing methods on a given -omics dataset [15]. Thus far, plasmodes have been successfully applied to microarray [16] and qPCR data [17], and have most recently been adapted for RNA-Seq data [18]. Probably one of the main advantages of plasmodes is that some characteristics of experimental data can be preserved, including the overall distribution of the data as well as any potential gene-to-gene correlation structure [18]. In particular, an RNA-Seq dataset can be used to generate a null plasmode dataset by randomly partitioning samples of the same treatment condition into two new

arbitrary groups. Reshuffling of samples creates a null dataset where no differential expression between groups is to be expected beyond sample-to-sample variation [18]. DE analysis of null plasmodes can then be implemented to compare Type I error between methods, since any transcripts identified as DE on a null plasmode would be considered a false positive. In addition, plasmodes allow for introduction of some known truth, as it happens in data simulation. As such , selected transcripts within a null plasmode can be "spiked" with fold changes of known magnitude to create DE transcripts, and thus evaluate statistical power and other performance metrics under the alternative hypothesis [18].

In this study, we use a plasmode-based approach to assess inferential performance of statistical methods for DE analysis of RNA-seq data with a special focus on low-count transcripts, as motivated by our case study on subspecies of bluestem prairie grass. We further evaluate the impact of alternative data filtering strategies that are commonly reported in the literature and discuss their implications for inference on low-count transcripts.

Results

Plasmodes

All plasmodes were generated using data from big bluestem samples only, given its benchmark status as a widely distributed dominant prairie grass. To evaluate inferential performance of statistical methods under the null hypothesis, big bluestem samples were randomly partitioned into 2 groups of 2 samples each, yielding a total of 3 null plasmodes. In turn, performance under the alternative hypothesis was evaluated using DE plasmodes, that is modified null plasmodes for which one of the groups had a known proportion of transcripts spiked with estimates of effect sizes of transcripts called DE from a preliminary analysis [18]. A total of 15 DE plasmodes were generated.

On each plasmode, we conducted DE analyses using DESeq2, edgeR classic, and edgeR robust. All of the methods model read counts assuming a negative binomial distribution and apply shrinkage to moderate the estimation of dispersion parameters. For edgeR robust, we specified degrees of freedom (DF) to be 4, 10 and 50, to reflect increasing levels of arbitrarily specified shrinkage. We note that DF = 10 is the default DF specification in edgeR robust. We also evaluated the performance of edgeR robust with DF specified using an estimate obtained from the classic edgeR software (i.e. DF $=\widehat{DF}$). We note that a quantile-adjusted conditional maximum likelihood approach for estimation of the DF parameter is available in the classic edgeR software for simple, completely randomized, design structures such as that in our motivating problem on bluestem subspecies [19]. Estimates of DF ranged from approximately 3.21 to 3.30 across null plasmodes. To compare performance of the various DE analyses methods, we computed false positive rate (FPR), true positive rate (TPR) or power, positive predictive value (PPV) or precision, negative predictive value (NPV), and accuracy, as defined in Table 1.2.

We first assessed Type I error of DE methods in null plasmodes using false positive rate (FPR). Since both groups pertain to the same subspecies, we do not expect any difference in expression levels beyond sampling variability. Table 1.3 contains estimated FPR for DE methods, after adjustment to a false discovery rate (FDR) of 0.05. Overall, all methods seemed to adequately control FPR below a 0.05 FDR nominal value for both all transcripts as well as low-count transcripts. Nevertheless, DESeq2 had the lowest FPR and was thus the most conservative of the methods evaluated, followed closely by edgeR classic and then by edgeR robust. Within edgeR robust, FPR increased with more degrees of freedom, thus indicating more liberal

inference with greater DF specifications. These patterns in FDR performance between DE methods were apparent when either all transcripts or only low-count transcripts were considered.

Next, we used DE plasmodes to compare inferential performance of statistical methods under the alternative hypothesis to detect true differences in expression levels of transcripts. Estimated TPR or power, PPV or precision, NPV and accuracy based on DE plasmodes are displayed in Table 1.4. Estimated power across methods ranged from approximately 0.54 to 0.65 for all transcripts, and from 0.17 to 0.39 for low-count transcripts. In both cases, DESeq2 showed the lowest power of all methods evaluated, followed by a modest power increase with edgeR classic and more substantial power boost with edgeR robust. Within specifications of edgeR robust, there was no evidence for differences in power when DF were specified to be 10 or less, but a DF=50 specification caused a significant inflation of power for both all transcripts and low-count transcripts. Not surprisingly, results on power mirrored those obtained on FPR based on the null plasmodes; that is, methods with the lowest FPR were also methods with the highest number of false negatives and thus, the lowest power. Precision, or PPV, was maximum using DESeq2 and was estimated at 0.66 and 0.39 for all transcripts and low-count transcripts, respectively (Table 1.4). In both cases, a significant drop in precision of at least 2 to 3 percentage points was apparent with edgeR classic relative to DESeq2, whereas the estimated drop in power was of 10 percentage points or more with edgeR robust. As the specification of DF on edgeR robust increased from 4 to 50, precision decreased further and was nearly halved by DF=50 relative to DESeq2. Noteworthy, for both all transcripts and low-count transcripts, inferential precision using edgeR robust was greater when DF were estimated as opposed to specified by default (i.e. DF=10; Table 1.4). In turn, estimated NPV for all DE methods was high in magnitude and ranged from 0.989 to 0.992 for all transcripts as well as low-count transcripts

(Table 1.4). Overall inferential accuracy of DE analyses ranged from 0.972 to 0.986 for all transcripts and from 0.968 to 0.985 for low-count transcripts. In both cases, maximum accuracy was observed using DESeq2, followed in decreasing order by edgeR classic and then by edgeR robust, with decreasing accuracy as DF increased (Table 1.4). Overall accuracy of DE calling using edgeR robust was greater when DF were estimated as opposed to specified by default (i.e. DF=10), though the absolute magnitude of the difference was small (approximately half a percentage point). All methods appear to control FPR rate in DE plasmodes below the nominal value (Table 1.4), though DESeq2 was more conservative than any of the edgeR methods, particularly for low-count transcripts.

Case study: Comparison of bluestem subspecies

Next, we conducted DE analyses to explore the transcriptomic basis for differences between subspecies of bluestem prairie grass (see Materials and Methods). Our dataset consisted of 4 samples of big bluestem and 4 of sand bluestem, for which read counts on a total of 25,582 transcripts were obtained. Differential expression analyses between subspecies was conducted using DESeq2 and edgeR robust. The specification of degrees of freedom for edgeR-robust was based on quantile-adjusted conditional maximum likelihood estimates using edgeR classic [19], such that $\widehat{DF} = 3.02$. Figure 1.2 A and D contains MA plots of estimated logarithmic fold changes in the complete dataset (i.e. no filtering) using DESeq2 and edgeR robust, respectively. Overall, edgeR robust declared 12.4% of transcripts as DE (Table 1.5) whereas DESeq2 declared only 9.0% of transcripts as DE (Table 1.6). This is consistent with the more conservative Type I error performance of DESeq2 relative to edgeR robust, coupled with greater power of the latter, as previously observed using a plasmode approach. We note that the difference in DE calling between statistical methods may be partially attributed to inference on low-count transcripts,

whereby 14.6% of low-count transcripts were declared DE by edgeR robust but only 9.1% by DESeq2 (Tables 1.5 and 1.6). Instead, DE calling amongst high-count transcripts was 4.8 and 4.6% for edgeR robust and DESeq2, respectively (Tables 1.5 and 1.6).

A considerable amount of overlap in DE calling was apparent between methods. In particular, approximately 91.2% of all transcripts declared DE by DESeq2 were also declared DE using edgeR robust (Figure 1.4.A). For low-count transcripts, edgeR robust declared DE approximately 96.8% those also declared DE by DESeq2 (Figure 1.4.D).

We further considered SB-only and BB-only transcripts, which were expressed in only one bluestem subspecies and absent in the other. Recall that all such transcripts were classified as low-count transcripts due to low expression levels. EdgeR robust identified 80.4% of such transcripts as DE (Table 1.5), whereas DESeq2 called DE only 39.8% (Table 1.6). Yet, approximately 99% of transcripts expressed in only one bluestem subspecies and declared DE based on DESeq2 were also declared DE by edgeR robust, again indicating a substantial amount of overlap between the methods.

Filtering Strategies

We further assessed inferential implications of two commonly used filtering approaches. For this purpose, our bluestem data set was subjected to filtering defined in terms of mapped reads present (RP) [20] and of read counts per million (CPM) [18]. The RP filtering approach removes transcripts if the overall number of samples with mapped reads present (i.e. samples with read counts greater than zero for a given transcript) is smaller than the number of samples per treatment group [20]. In turn, CPM-based filtering removes transcripts if a pre-selected number of samples have counts per million (CPM) smaller than a certain value [18], which for

our dataset was specified at 1 CPM. Table 1.1 shows a breakdown of transcripts available for DE analyses after RP-based and CPM-based filtering. Most notably, RP-based filtering excluded only 129 transcripts (i.e. approximately 0.5%) from the unfiltered dataset, none of which were low-count transcripts or transcripts present in only one of the subspecies. In contrast, when CPM-based filtering was implemented, a total of 10,734 transcripts (i.e. almost 42% of the total) were excluded. More specifically, CPM filtering removed from the data 10,280 low-count transcripts, amongst which were all 455 transcripts present in only one of the bluestem subspecies (Table 1.1). As such, only approximately 29% low-count transcripts, and none of the transcripts present in only one of the bluestem subspecies, were available for DE analyses following CPM-based filtering.

Filtered datasets were subjected to DE analysis using edgeR robust and DESeq2, as described in the previous section. Tables 1.5 and 1.6 show the breakdown of transcripts declared DE on the filtered datasets based by each of the statistical methods. Based on either DESeq2 or edgeR robust, transcripts declared DE in RP-filtered data were essentially the same transcripts as those declared DE in the unfiltered data (i.e. over 99% overlap). Exceptions included 4 (edgeR robust) or 7 (DESeq2) additional transcripts declared DE in the RP-filtered data, but not in unfiltered data. Instead, CPM filtering reduced the number of transcripts declared DE based on edgeR robust and DESeq2 by 68.4% ((3173-1002)/3173, Table 1.5) and 58.4% ((2290-952)/2290, Table 1.6), respectively, relative to unfiltered data. The impact of CPM filtering on DE calling was primarily driven by low-count transcripts, for which DE calling was reduced by 88.8% ((2135-239)/2135, Table 1.5) and 84.6% ((1325-204)/1325, Table 1.6) based on edgeR robust and DESeq2, respectively. Most notably, all 455 transcripts present in only one of the

bluestem subspecies were lost to DE inference as CPM-based filtering excluded them from the data prior to analyses.

Figure 1.2 shows MA-plots obtained from fitting DESeq2 or edgeR robust to RNA-seq data subjected to no filtering, RP-based filtering or CPM-based filtering. Within each DE method, the overall shape of the MA-plots on RP-filtered data resembled that of the unfiltered data. This is not surprising as RP filtering removed only a small number of transcripts from the dataset. In contrast, MA-plots on the CPM-filtered dataset showed a drastically modified pattern relative to unfiltered data, particularly on the left size of each plot, due to exclusion of low-count transcripts, which were also transcripts with more extreme fold-changes.

Approximate tests for DE inference based on DESeq2

The most recent release of the DESeq package, namely DESeq2 [2], implemented a Wald test approach as the default strategy for DE testing on individual transcripts. This approach differs from that of previous versions of DESeq, which specified a likelihood ratio test (LRT) as the default instead [21]. The rationale behind using a Wald test approach as a default relies on its flexibility for testing individual coefficients or functions thereof, without the need to fit a reduced model [2]. Both LRT and Wald are approximate tests that rely on asymptotic chi-square and normal distributions, respectively, under the null hypothesis [22].

Motivated by our interest in low-count transcripts, we further compared the relative performance of LRT and Wald tests for DE inference. Figure 1.3 shows scatterplots of unadjusted p-values for DE inference obtained from Wald tests and LRT for both high-count transcripts and low-count transcripts in unfiltered or filtered datasets. Regardless of data filtering, LRT and Wald tests showed considerable inferential agreement for DE calling on high-count transcripts, as indicated by most points falling along the identity line (Figure 1.3 D, E, F). In

contrast, for low-count transcripts, the Wald test seemed to underestimate P-values for DE inference relative to the LRT approach, particularly in the 0 to 0.20 range (Figure 1.3 A, B, C). This difference between Wald test-based and LRT-based P-values on low-count transcripts was particularly noticeably in unfiltered or RP-filtered data but it was even apparent, though only slightly, in CPM-filtered data, for which most low-count transcripts had already been excluded.

Discussion

In this study, we used plasmodes generated from our RNA-seq data on bluestem subspecies to compare inferential performance on differential expression between alternative statistical methods. This study is one of few to use a plasmode based approach to compare statistical methods in RNA-Seq data [18]. It is important to note that with plasmodes the null distribution can be sensitive to the random division of the original control group into the two new plasmode groups when there are systematic differences such as variation due to technology or operator, especially if the number of samples in the control group is small [16]. However this is more of an attribute rather than a limitation since the goal of the plasmode is to reflect the actual structure of a real dataset including any systematic effects [16]. We were particularly interested in transcripts of low expression levels, as motivated by observed differences in phenotypes between subspecies of bluestem prairie grass. We also considered data filtering strategies that, while often implemented in RNA-seq data pipelines, impose arbitrary criteria for data exclusion with unknown impact on DE inference.

Our plasmode approach indicated adequate control of Type I error within nominal levels using either DESeq2 or edgeR robust, regardless of specification of DF, both for all transcripts and for low-count transcripts. Still, false positive rates increased with greater DF specifications under edgeR robust, indicating the need for careful consideration of this specification. In turn,

edgeR robust showed greater power than DESeq2, which is consistent with previous simulation studies [8] and is hereby shown to also apply to transcripts of low expression levels. Interestingly, within specifications of edgeR robust, there was no evidence for any changes in power when DF were specified to be at default (i.e. DF=10) or at a much smaller value estimated from the data (i.e. DF = [3.21, 3.30], though power was increased at DF=50. Nevertheless, the observed increase in power with increasing DF in edgeR robust came at the expense of an increase in false positives. Not unexpectedly, power for DE calling of low-count transcripts was decreased relative to that of all transcripts, regardless of method chosen for DE inference. This is to be expected as low levels of expression indicate little information available for inference, as shown with previous simulations studies [8]. Furthermore, DESeq2 showed the greatest precision and accuracy of all methods evaluated not only for all transcripts, as already shown by other simulation studies [2], but especially for low-count transcripts.

Our results from the plasmode approach to assess inferential performance suggest that the specification of DF for edgeR robust can impact DE inference of RNA-seq data, particularly that of low-count transcripts and thus, should be considered carefully. Most relevant to our dataset, the default DF specification (i.e. DF=10) was not optimal and lead to a decrease in inferential precision and accuracy relative to using an estimated DF value. The default value of the shrinkage parameter for edgeR robust (i.e. DF = 10) seems to be based on an array of simulation studies [7]. However, it is unclear whether such arbitrarily specified DF value is justified for any particular RNA-seq dataset, for which the amount of dispersion, the correlation structure between transcripts and the sample size may not be aligned with those of simulated conditions [18, 20]. This is further supported by our plasmode approach and suggests that the specification of DF on edgeR robust should be informed carefully and estimated from the data whenever

possible. Alternatively, if complexity of the experimental design prevented proper estimation of the DF parameter, a researcher might consider relying more heavily on inference from DESeq2, for which no arbitrary specification of DF is needed. Regardless, for any given dataset, DE inference based on multiple analysis methods seems to be a standard recommendation [1, 20]. For instance, researchers may consider declaring DE only those transcripts that show low FDR-adjusted P-values by both edgeR robust and DESeq2. This recommendation is further supported by the high level of overlap in DE calling observed between the methods, when properly specified.

Data filtering is a common processing step in the RNA-seq pipeline [4], though its implications have not been thoroughly explored. Filtering was original proposed with the goal of reducing the impact of multiple testing adjustment on power for DE detection [4, 6]. However, our results indicate only a small difference in DE calling following RP-based filtering compared to no filtering, with 99% overlap between the two, regardless of DE method. This suggests that both edgeR robust and DESeq2 retained similar number of transcripts declared as DE regardless of whether the data has been RP-filtered or not, thus questioning the need to impose arbitrary filtering rules on the data.

In turn, more extreme filtering rules such as those based on a CPM criterion can cause a drastic reduction in the number of transcripts available for DE analyses. In our case, CPM filtering excluded almost 42% of the original transcripts, most of which were low-count transcripts. Filtering by CPM criterion was originally designed to remove transcripts considered challenging for inference due to shortage of available information [4]. However, we showed that CPM-based filtering also excluded from the data all transcripts expressed in only one of the bluestem subspecies and absent in the other, which were of particular interest to researchers in

this case. Removal of these transcripts may impair understanding of the transcriptomic basis for phenotypic differences between bluestem subspecies and misinform further exploration of candidate genes. Moreover, CPM-based filtering also reduced both the total number and the proportion of transcripts called DE relative to no filtering, whereas little gain was obtained in uniquely identified DE-declared transcripts (i.e. approximately 0.3% and 0.1% gain with DESeq2 and edgeR robust, respectively). On a more general note, data exclusion based on CPM filtering may have even more serious implications for inference on transcription factors, which have low expression levels despite their key role as master switches that regulate gene expression [5].

Overall, the rationale for arbitrary filtering RNA-seq data based on either an RP or CPM criteria, seems unclear, particularly given the availability of powerful state-of-the-art statistical methodology that can deal with most of the challenges in RNA-seq data. Instead, researchers may consider using the complete unfiltered RNA-seq data for DE analyses, ensuring use of modern statistical methods to properly borrow information across transcripts and moderate (i.e. shrink or weigh) DE inference based on expression levels. In particular, DESeq2 and edgeR robust have shown promising inferential performance in handling low-count transcripts with minimal effect on the DE analysis for the remaining transcripts. Forgoing the use of data filtering at arbitrary thresholds in favor of more elegant approaches to deal with the inherent challenges of RNA-seq data may be particularly relevant for research questions focused on transcripts of low expression levels.

Finally, when implementing DESeq2, differential expression assessments for low-count transcripts based on the default Wald-test may be rather liberal relative to those based on likelihood ratio tests. For small sample sizes, the performance of these approximations is known

to deteriorate rapidly, particularly for Wald tests [23]. In addition, both tests assume certain regularity conditions hold, but are often not verified in practice [22]. Both tests implemented by DESeq2 constitute approximations that may require careful attention to and consideration of the assumptions made on a case-by-case basis, in order to ensure sound inference.

Conclusions

We implemented a recently adapted plasmode approach to compare inferential performance of modern statistical methods, namely DESeq2 and edgeR robust, on RNA-seq data, with a special focus on low-count transcripts as motivated by bluestem grass species. Implications of these results may be relevant to biological applications beyond our study involving transcripts of low expression levels, such as transcription factors. Both DESeq2 and edgeR robust seemed to properly control family-wise type 1 error on all transcripts as well as on low-count transcripts. For low-count transcripts, edgeR robust showed greater power whereas DESeq2 showed greater precision and accuracy. Overall, both methods showed promising inferential performance on low-count transcripts and yielded a substantial amount of overlap in DE calling. Still, a note of caution is in order regarding the approximate nature of DE tests, particularly when applied to low-count transcripts, in particular those of DESeq2.

The specification of DF under edgeR robust was non-trivial as it impacted precision and accuracy of DE inference. This finding questions the use of a default DF value that may not be appropriate for all datasets and that was certainly not optimal in our case study. Whenever possible, the DF should be estimated from the data.

Filtering of RNA-seq data can have serious implications for inference as mostly low-count transcripts are removed from the data and excluded from DE analyses. Researchers may reconsider standard RNA-Seq data pipelines that call for filtering at arbitrary thresholds. Instead,

researchers may implement modern statistical methodologies specifically developed to deal with the inherent challenges of RNA-seq data.

Materials and Methods

Data collection

RNA was extracted from leaf tissue of 4 individual plants of each of two phenotypically divergent bluestem subspecies, namely big bluestem (*Andropogon gerardii,* Saline population) and sand bluestem (*A. gerardii ssp. Hallii,* Arapahoe population). All plants were grown under greenhouse conditions. Samples were sequenced using the Illumina HighSeq 2000 and Roche 454 sequencers on a single flow cell using multiplexed sequencing. Reads were mapped to a *de-novo* reference transcriptome assembly (described next) [24] and number of aligned reads were counted on putative transcripts.

Transcriptome Assembly

All reads were stringently cleaned to remove tags, ambiguous bases, duplicates, and low quality bases. The resultant Illumina and 454 assemblies were merged with miraEST v3.4.11 [25] to produce the final merged transcriptome. Assemblies were evaluated on the basis of N25, N50, N75, cumulative length of contigs, and number of contigs. Ortholog Hit Ratio (OHR) [26] was calculated. N-values and OHR values suggest that the merged assembly was more contiguous and complete than either the Illumina or the 454 assemblies individually. A larger proportion of BLASTX [27] hits was identified in the merged assembly than in any of the single-*k*-mer Illumina assemblies indicating that the contigs in the merged assembly may be more complete in addition to having higher N-values. Cleaned Illumina reads were mapped to the final emerged assembly using Bowtie2 v.2.1.0 [28] in the best mapping mode. Two reference

transcriptomes were used for alignment. The final transcriptome used for analysis contained transcripts with greater than or equal to 400 base pairs.

RNA-seq data

Prior to analysis, any transcripts with zero reads present on all samples were removed. The final dataset used for analysis consisted of a total of 25,582 transcripts. We first defined transcripts with expression levels present in sand bluestem and absent (i.e. read counts = 0 for all samples) in big bluestem as SB-only transcripts. In turn, BB-only transcripts were defined as transcripts with expression levels present in big bluestem and absent (i.e. read counts = 0 for all samples) in sand bluestem.

We then organized and partitioned data based on relative abundance of transcripts. In brief, transcripts were ranked from largest to smallest number of total mapped reads across all samples. We adapted the approach proposed by Bullard [14] and defined high-count transcripts as the top 3 % transcripts with the highest relative abundance, which accounted for 60 % of total read counts (Figure 1.1). We also defined low-count transcripts as transcripts within the 60^{th} percentile of least relative abundance and accounting for approximately 3% of total read counts (Figure 1.1). So defined, high-count transcripts and low-count transcripts were transcripts with at least 12,893 read counts or at most 462 read counts, respectively, across all samples in the dataset. We note that the proposed definitions of high-count and low-count transcripts are specific to our motivating problem and the corresponding structure of our data. Table 1.1 shows the breakdown of transcripts into high-count and low-count categories in the filtered and unfiltered data.

Construction of plasmode datasets

All plasmodes were generated using data from big bluestem samples only, given its benchmark status as a widely distributed dominant prairie grass.

Null plasmode datasets were constructed as previously described [18]. Briefly, for each null plasmode, samples of big bluestem were randomly partitioned into two arbitrary groups. A total of 3 unique null plasmodes were created, reflecting the 3 possible unique combinations of 4 samples in groups of 2. So defined, no differential expression is to be expected between groups other than sample-to-sample variation. Thus, null plasmodes allow for evaluation of analysis models under the null hypothesis [18].

A total of 5 DE plasmodes were generated from each null plasmode for a total of 15 DE plasmodes, as previously described [18]. The proportion of differentially expressed transcripts in each DE plasmode was set at $\pi = 0.2$. We used edgeR classic [19] to obtain a list of estimated effect sizes for transcripts declared DE at FDR = 0.05. Estimates of effect sizes were sampled without replacement and added to log-transformed counts of randomly selected transcripts on all samples of one of the arbitrary groups in the null plasmode dataset, then back transformed to the count scale. As such, DE plasmodes combine random reshuffling of data with known effects estimated from real data and added to known transcripts. Thus, DE plasmodes allow for evaluation of analysis models in identifying truly DE as well as non-DE transcripts [18].

Differential expression analyses

DESeq2

The R package DESeq2 [2] for which the read count K_{ij} for transcript i in sample j is described with a generalized linear model of the Negative Binomial family with logarithmic link, such that $K_{ij} \sim NB(mean = \mu_{ij}, dispersion = \alpha_i)$ with mean $\mu_{ij} = s_j q_{ij}$ and link function

$\log(q_{ij}) = x_j\beta_i$. Where s_j is the normalized library size for sample *j* as previously defined [21]. Where α_i is the variability between samples for transcript *i*, $x_j$as the design matrix elements, and β_i as the design matrix coeffcients. Estimation of the dispersion parameter is conducted in 3 steps [2]. First, gene-wise dispersion estimates are obtained using maximization of the Cox-Reid adjusted conditional likelihood of the dispersion. Then, a dispersion trend is estimated using a Gamma-family generalized linear model regression. Last, a maximum a-posteriori (MAP) dispersion estimate is obtained by shrinking the gene-wise dispersion estimates toward the overall dispersion trend using an empirical Bayes approach that enables borrowing of information across transcripts. DESeq2 further incorporates empirical Bayes shrinkage of logarithmic fold changes, thus enabling further borrowing of information and stable estimation for gene expression fold changes to count data, particularly for low-count genes [2]. More specifically, maximum-likelihood estimates of logarithmic fold changes are shrunk towards a zero-centered normal prior distribution to yield the final MAP estimates. The amount of shrinkage is inversely proportional to the amount of information an experiment provides for a given log fold change coefficient, so that transcripts with low estimated mean values μ_{ij} and high dispersion α_i in small datasets are pulled more strongly toward zero. Default DE testing on the shrunken LFCs is based on a Wald test, whereas a likelihood ratio test (LRT) alternative is also available [2].

EdgeR robust

R package edgeR robust [8] for which the read count Y_{ij} for transcript *i* in sample *j* is described with a generalized linear model of the Negative Binomial family with logarithmic link, such that $Y_{ij} \sim NB(mean = \mu_{ij}, dispersion = \phi_i)$ with link function $\log(\mu_{ij}) = X\beta_i +$ $\log(N_j)$. Where *X* is the design matrix containing the covariates, β_i is a vector of regression

parameters, N_j is the library size for sample j, and ϕ_i is the square of the biological coefficient of variation for transcript i. Dispersion parameters are estimated as follows. First, initial gene-wise dispersion is estimated using adjusted penalized likelihood (APL). These estimates are then moderated by shrinkage towards a common dispersion estimate obtained by maximizing a common likelihood function. Shrinkage is determined by a prior degree of freedom parameter afforded to the shared likelihood and specified arbitrarily by the researcher [4]. Unless explicitly specified, the default value for the prior degrees of freedom is equal to 10 [29]. In turn, regression parameters β_i are estimated using maximum likelihood that incorporates working weights attached to each observation. Weights are attached to each observation so observations that deviate strongly from model fit are given a lower weight. Observations weights are defined as functions of a Pearson residual [8] that are iteratively updated during estimation. The dispersion estimation machinery also receives the same observation weights, so that the influence of outliers is dampened on both regression and dispersion estimates. Testing for DE is conducted using a LRT-based approach.

Specification of the shrinkage parameter for edgeR robust

As previously indicated, edgeR robust uses DF = 10 as a default to specify the amount of shrinkage applied to dispersion parameters [7]. While the default value for degrees of freedom is provided in the edgeR robust package as a "rule of thumb", there is little guidance available to accurately inform specification of the DF parameter in a given dataset. Greater values of DF indicate greater shrinkage of tagwise dispersion estimates towards an overall dispersion parameter common to all transcripts.

We compared the performance of edgeR robust at varying DF specifications. In particular, we considered DF = 4, 10 and 50, to indicate a range of shrinkage around the default

specification. Further, we considered using the classical edgeR software [19] to estimate DF using a quantile-adjusted conditional maximum likelihood [19]. Estimation of DF is facilitated by the simple design structure of our bluestem dataset, in which only 2 groups are being compared (i.e. BB vs SB) and no blocking or nesting design structure is apparent. We refer to this scenario as a DF=classic specification under edgeR robust.

Multiple testing adjustments

Following DE analyses based on either DESeq2 or EdgeR robust, transcripts were called DE based on a FDR = 0.05 using the Benjamini-Hochberg procedure [30].

Performance metrics

Table 1.2 defines performance metrics used to compare inferential performance of statistical methods. More specifically, we defined false positive rate (FPR) as the number of false positives over the sum of false positives and true negatives. Power was defined as the number of true positives over the sum of true positives and false negatives. In turn, precision was the number of true positives over the sum of true positives and false positives, and it is also referred to as positive predictive value. Further, negative predictive value (NPV) was the number of true negatives over the sum of the true negatives and false negatives. Finally, accuracy was defined as the sum of true positives and true negatives over the total number of transcripts.

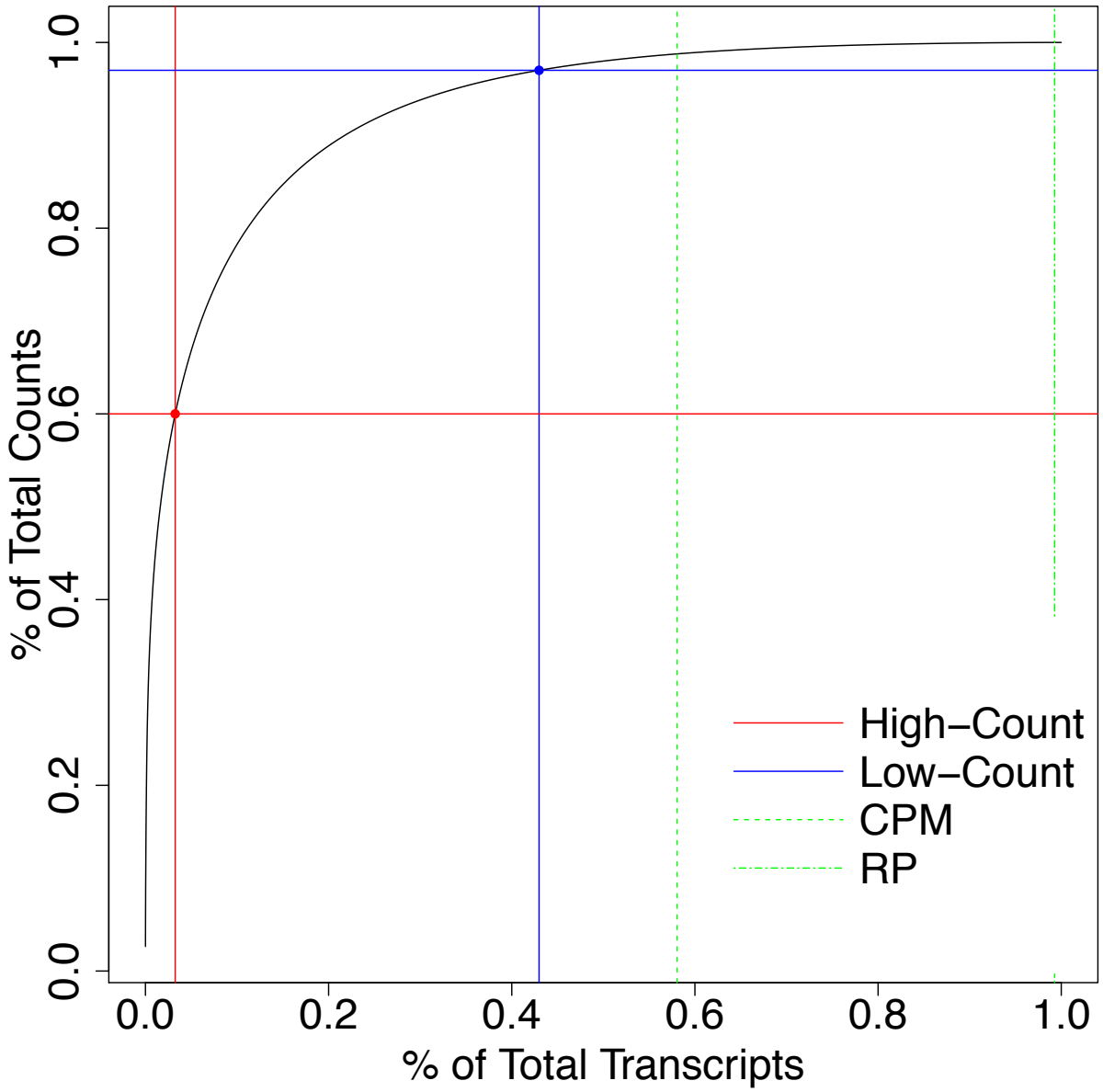
Performance metrics were computed on each plasmode dataset fitted with each statistical method for DE analyses. Each metric was then fitted with a general linear mixed model to compare methods for DE analysis accounting for plasmode dataset as a random blocking factor. Models were fitted using the GLIMMIX procedure of SAS (Version 9.3, SAS Institute Inc., Cary, NC). Residual assumptions were evaluated using studentized residuals. Pairwise

comparisons in performance metrics between analyses methods were conducted using a Tukey-Kramer adjustment to prevent the inflation of type 1 error rate.

Filtering strategies

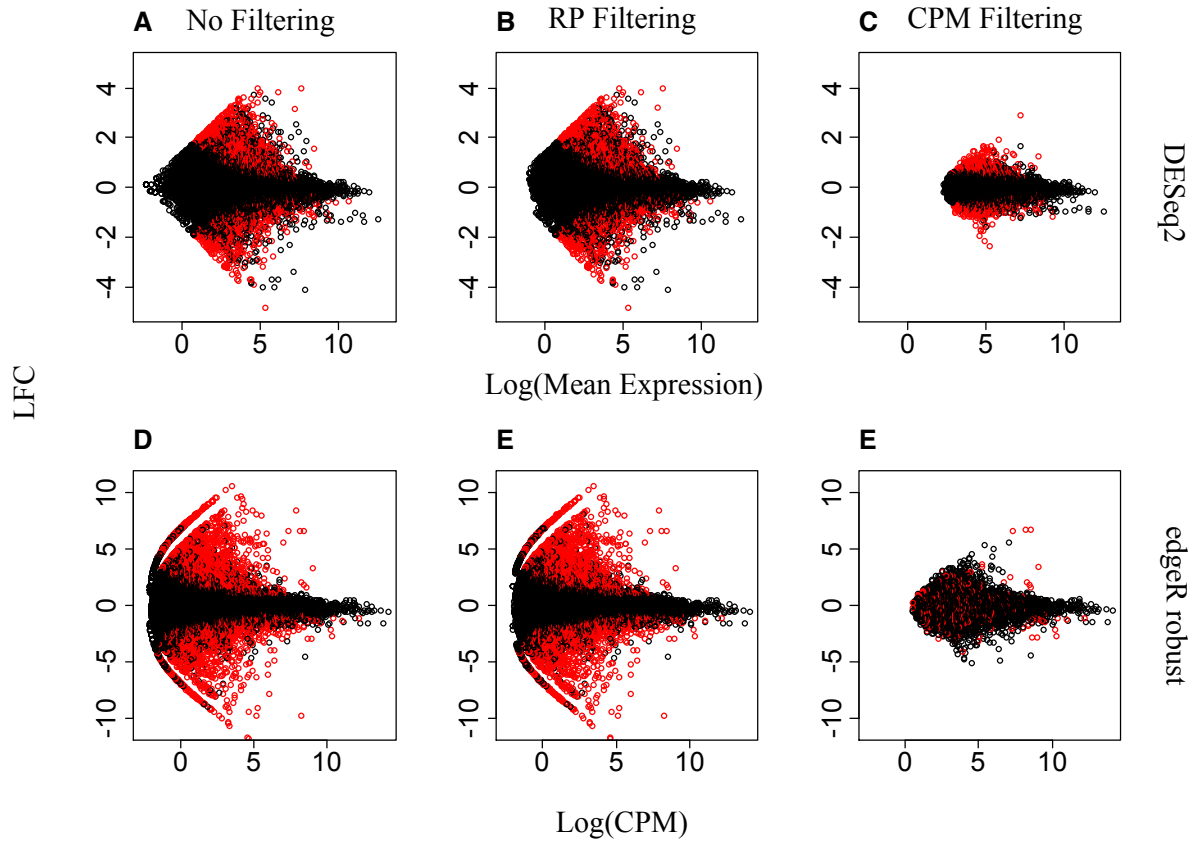
Filtering criteria are often applied to RNA-set datasets prior to DE analyses. A relatively common filtering criterion removes transcripts from a dataset if the number of samples with mapped reads present (RP) is smaller than the number of samples per treatment [20]. For our data, RP filtering removes transcripts with fewer than a total of 4 mapped reads across all samples. Another common filtering alternative strategy removes transcripts if two or more samples have counts per million (CPM) smaller than an arbitrary number [18]. For our data CPM-based filtering removed transcripts if two or more samples have CPM less than 1 CPM. This was analogous to removing any transcripts with fewer than 80 mapped reads across all samples. The number of transcripts remaining in the dataset after applying RP filtering or CPM filtering is shown in Table 1.1, along with the total number of transcripts in the unfiltered dataset, whereby all transcripts with at least one read count in any of the samples is included.

Figure 1.1 Partitioning of high-count transcripts and low-count transcripts



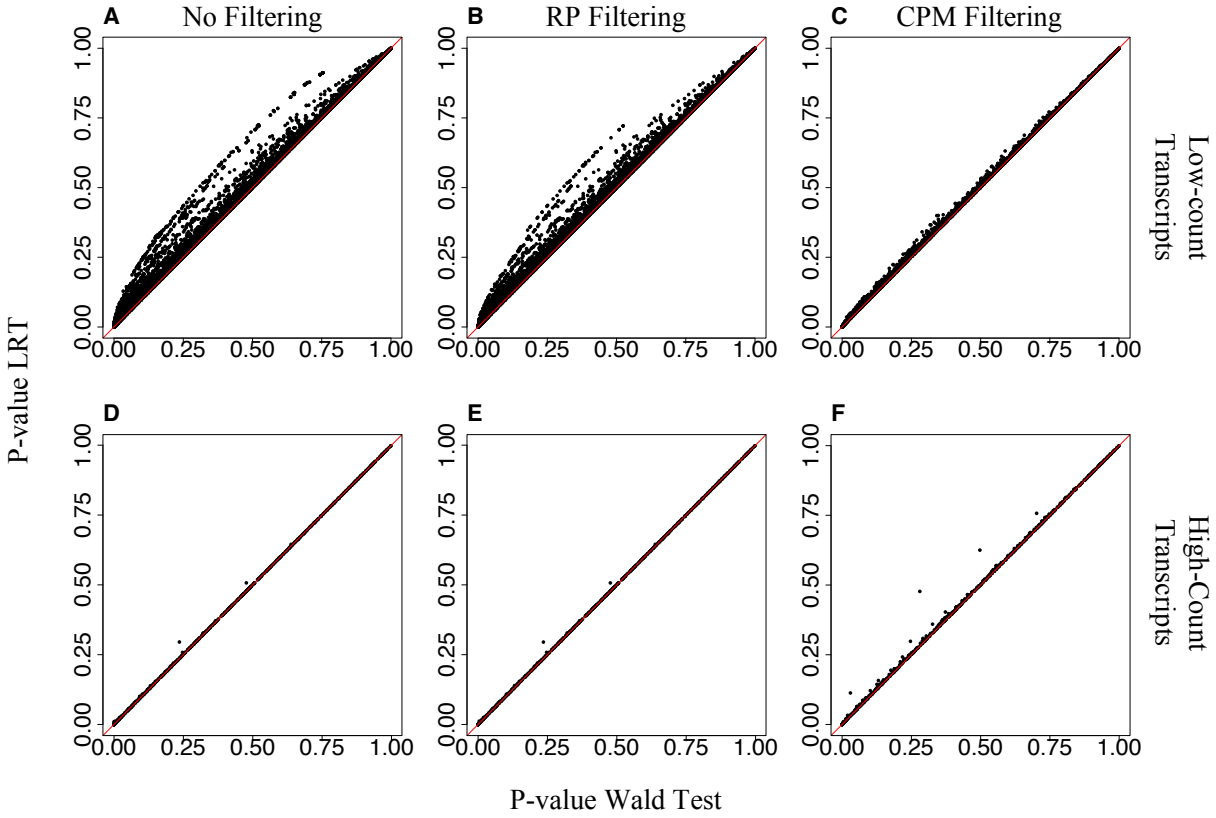
Cumulative percentage of total read counts (*y*-axis) as a function of cumulative percentage of transcripts (*x*-axis), starting on the left with transcripts of highest read counts. Solid colored lines indicate cumulative read counts for the 3rd percentile (red line) most highly expressed transcripts and for the 60th percentile (blue line) least expressed transcripts thereby defining high-count transcripts (to the left of the vertical red line) and low-count transcripts (to the right of the vertical blue line), respectively.

Figure 1.2 MA-Plots for edgeR robust and DESeq2 with and without filtering



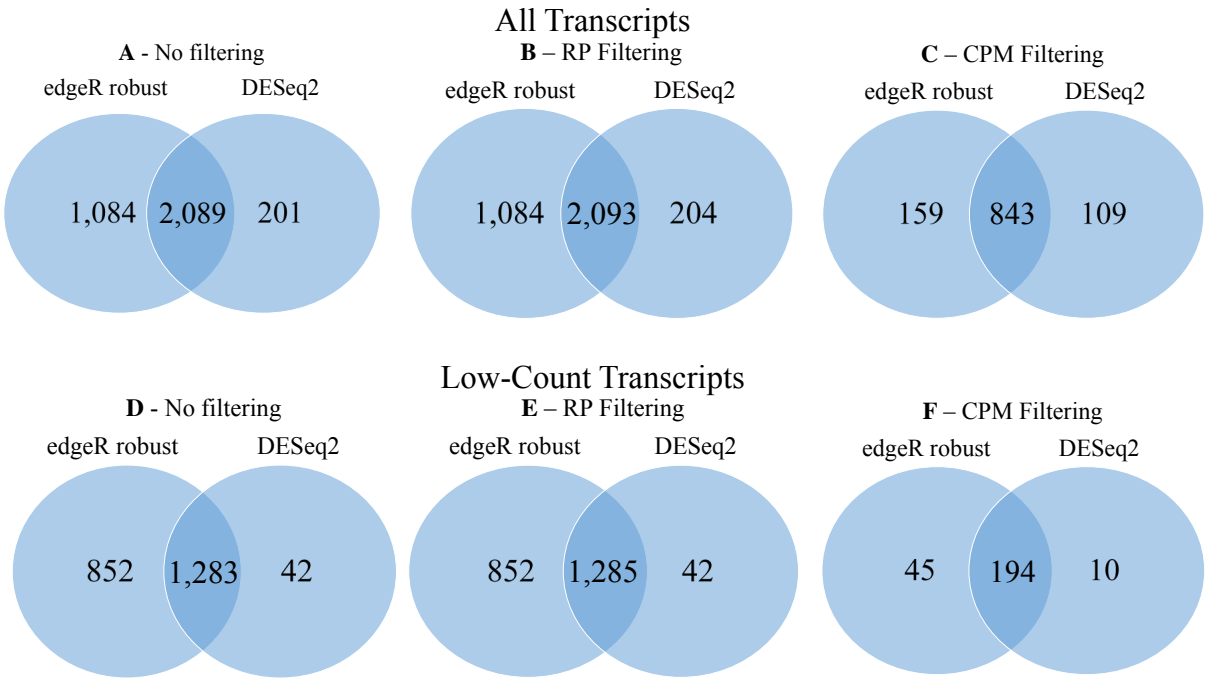
Estimated fold-change in expression of RNA-seq transcripts for SB relative to BB as a function of transcript abundance following differential expression analyses with DESeq2 or edgeR robust (DF = Classic) on data subjected to no filtering or to filtering with CPM or RP methods. For DESeq2, fold-changes are plotted over mean transcript expression on a log scale. For edgeR robust, fold-changes are plotted against counts per million on a log scale. Transcripts declared DE at FDR=0.05 are colored in red.

Figure 1.3 Comparison of P-values for DESeq2 tests on differential expression



Scatterplot of p-values for differential expression obtained using DESeq2's likelihood ratio test (LRT) and Wald test on low-count and high-count transcripts subjected to no filtering or to filtering with CPM or RP methods. Diagonal identity line is indicated in red.

Figure 1.4 Transcripts declared differentially expressed (DE) using edgeR robust and DESeq2



Venn diagrams of sets of all transcripts and of low-count transcripts declared DE using edgeR robust (with degrees of freedom specified based on the corresponding estimate obtained using classical edgeR software) and DESeq2 on unfiltered data or on data filtered on reads-present (RP) or counts per million (CPM) criteria.

Table 1.1 Number of transcripts in the dataset

	All Transcripts			High Count Transcripts			Low Count Transcripts		
	Total	**SB only**	**BB only**	**Total**	**SB only**	**BB only**	**Total**	**SB only**	**BB only**
No Filter	25,582	323	132	831	0	0	14,588	323	132
RP	25,453	323	132	831	0	0	14,588	323	132
CPM	14,848	0	0	828	0	0	4308	0	0

The table contains the number of total-transcripts, high-count transcripts and low-count transcripts available for differential expression analyses following either no data filtering or filtering based on a reads-present (RP) or a counts per million (CPM) criteria. Also listed are number of transcripts with expression levels present in sand bluestem and absent in big bluestem (SB-only transcripts), and transcripts with expression levels present in big bluestem and absent in sand bluestem (BB-only transcripts).

Table 1.2 Classification rules to compute performance metrics

	Transcripts not DE	Transcripts spiked as DE	Total
Transcripts not declared significantly DE	TN	FN	R_0
Transcripts declared significantly DE	FP	TP	R_1
Total	S_0	S_1	G

FP, number of false positives (transcripts in S_0 set declared differentially expressed); TP, number of true positives (transcripts in S_1 set declared differentially expressed);TN, number of true negatives; FN, number of false negatives; FPR, false positive rate = FP/S_0; TPR, true positive rate or power = TP/S_1; PPV, positive predictive value or precision = TP/R_1; NPV, negative predictive value = TN/R_0; accuracy = $(TP+TN)/G$.

Table 1.3 Estimated false positive rates (FPR) on null plasmodes

edgeR robust DF=50	edgeR robust DF=10	edgeR robust DF=4	edgeR robust DF=$\widehat{DF}$	edgeR Classic DF=$\widehat{DF}$	DESEQ2
FPR All Transcripts					
0.0177 a (0.00040)	0.0093 b (0.00040)	0.0063 c (0.00040)	0.0061 c (0.00040)	0.0042 d (0.00040)	0.0031 e (0.00040)
FPR Low-Count Transcripts					
0.0176 a (0.00037)	0.0094 b (0.00037)	0.0064 c (0.00037)	0.0063 c (0.00037)	0.0043 d (0.00037)	0.0032 e (0.00037)

Least square mean estimates (and corresponding SEM, shown in parentheses) of FPR for differential expression at FDR = 0.05 on all transcripts and on low-count transcripts based on DESeq2, EdgeR classic and EdgeR robust, implemented on null plasmodes of RNA-seq data. a,b,c, ,e, indicate differences (Tukey-Kramer adjusted P<0.05) within a row.

Table 1.4 Performance metrics on differentially expressed (DE) plasmodes

edgeR robust DF=50	edgeR robust DF=10	edgeR robust DF=4	edgeR robust DF=$\widehat{DF}$	edgeR Classic DF=$\widehat{DF}$	DESEQ2
Power - All transcripts					
0.6495 a (0.00610)	0.6275 b (0.00610)	0.6215 b (0.00463)	0.6229 b (0.00463)	0.5704 c (0.00435)	0.5418 d (0.00435)
Power - Low-count Transcripts					
0.3922 a (0.00120)	0.3692 b (0.00120)	0.3657 b (0.00010)	0.3677 b (0.00010)	0.2647 c (0.00008)	0.1785 d (0.00008)
Precision - All Transcripts					
0.3607 a (0.00792)	0.4393 b (0.01153)	0.5218 c (0.01192)	0.5323 c (0.01114)	0.6321 d (0.00589)	0.6586 e (0.00371)
Precision - Low-Count Transcripts					
0.1800 a (0.00811)	0.2289 b (0.01305)	0.2904 c (0.01432)	0.2963 c (0.01378)	0.3605 d (0.01363)	0.3915 e (0.01727)
NPV - All Transcripts					
0.9917 a (0.00016)	0.9915 a (0.00016)	0.9915 a (0.00016)	0.9915 a (0.00016)	0.9902 b (0.00016)	0.9891 c (0.00016)
NPV - Low-Count Transcripts					
0.9917 a (0.00016)	0.9915 a (0.00016)	0.9914 a (0.00016)	0.9915 a (0.00016)	0.9902 b (0.00016)	0.9891 c (0.00016)
Accuracy - All Transcripts					
0.9718 a (0.00080)	0.9778 b (0.00080)	0.9821 b (0.00024)	0.9826 c (0.00024)	0.9858 d (0.00012)	0.9862 e (0.00014)
Accuracy - Low-Count Transcripts					
0.9679 a (0.00122)	0.9744 b (0.00122)	0.9794 c (0.00079)	0.9798 c (0.00079)	0.9841 d (0.00028)	0.9855 e (0.00028)
FPR - All Transcripts					
0.0221 a (0.00086)	0.0155 b (0.00086)	0.011 b (0.00024)	0.0106 c (0.00024)	0.0063 d (0.00001)	0.0053 e (0.00010)
FPR - Low-Count Transcripts					
0.0244 a (0.00128)	0.0175 b (0.00128)	0.0124 c (0.00082)	0.0121 c (0.00082)	0.0063 d (0.00019)	0.0037 e (0.00019)

Least square mean estimates (and corresponding SEM, shown in parentheses) for true positive rate (TPR; i.e. power), positive predictive value (PPV; i.e. precision), negative predictive value (NPV), accuracy and false positive rate (FPR) for differential expression at FDR = 0.05 on all transcripts and on low-count transcripts yielded by DESeq2, EdgeR classic or EdgeR robust, implemented on DE plasmodes of RNA-seq data. a,b,c,d,e, indicate differences within a row (Tukey-Kramer adjusted $P<0.05$).

Table 1.5 Number of transcripts declared differentially expressed (DE) using edgeR robust

	All Transcripts			High-Count Transcripts	Low-Count Transcripts		
	Total	**SB only**	**BB only**	**Total**	**Total**	**SB only**	**BB only**
No Filter	3,173	248	126	40	2,135	245	121
RP	3,177	248	126	40	2,137	245	121
CPM	1,002	0	0	23	239	0	0

The table contains the number of total transcripts, high-count transcripts and low-count transcripts declared DE using edgeR robust (with degrees of freedom specified based on the corresponding estimate obtained using classical edgeR software) on unfiltered data or on data filtered based on reads-present (RP) or counts per million (CPM) criteria. Also listed are transcripts with expression levels present in sand bluestem and absent in big bluestem (SB-only transcripts) and transcripts with expression levels present in big bluestem and absent in sand bluestem (BB-only transcripts).

Table 1.6 Number of transcripts declared differentially expressed (DE) using DESeq2

	All Transcripts			High Count Transcripts	Low Count Transcripts		
	Total	**SB only**	**BB only**	**Total**	**Total**	**SB only**	**BB only**
No Filter	2,290	112	69	38	1,325	112	69
RP	2,297	111	69	38	1,327	111	69
CPM	952	0	0	30	204	0	0

The table contains the number of total transcripts, high-count transcripts and low-count transcripts declared DE using DESeq2 on unfiltered data or on data filtered based on reads-present (RP) or counts per million (CPM) criteria. Also listed are transcripts with expression levels present in sand bluestem and absent in big bluestem (SB-only transcripts) and transcripts with expression levels present in big bluestem and absent in sand bluestem (BB-only transcripts).

References

1. Kvam VM, Liu P, Si Y: **A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data**. *Am J Bot* 2012, **99**:248–256.

2. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2**. *bioRxiv* 2014:1–34.

3. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot NS, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**. *Brief Bioinform* 2013, **14**:671–683.

4. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD: **Count-based differential expression analysis of RNA sequencing data using R and Bioconductor.** *Nat Protoc* 2013, **8**:1765–86.

5. Spitz F, Furlong EEM: **Transcription factors: from enhancer binding to developmental control**. *Nat Rev Genet* 2012, **13**:613–626.

6. Bourgon R, Gentleman R, Huber W: **Independent filtering increases detection power for high-throughput experiments.** *Proc Natl Acad Sci U S A* 2010, **107**:9546–9551.

7. McCarthy DJ, Chen Y, Smyth GK: **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation**. *Nucleic Acids Res* 2012, **40**:4288–4297.

8. Zhou X, Lindsay H, Robinson MD: **Robustly detecting differential expression in RNA sequencing data using observation weights**. *Nucleic Acids Res* 2014, **42**:1–10.

9. Owsley M: **Plant Fact Sheet: Big Bluestem**. *USDA - Nat Resour Conserv Serv* 2011.

10. Knapp A, Carter G: **Variability in leaf optical properties among 26 species from a broad range of habitats**. *Am J Bot* 1998, **85**:940–946.

11. Barnes PW: **Variation in the big bluestem (Andropogon gerardii)- sand bluestem (Andropogon hallii) complex along a local dune meadow gradient in the Nebraska Sandhills.** *American Journal of Botany* 1986:172–184.

12. Shelton J: **Epicuticular wax chemistry, morphology, and physiology in sand bluestem, andropogon gerardii ssp. hallii, and big bluestem, andropogon gerardii spp. gerardii**. Kansas State University; 2012.

13. IPCC: **Summary for Policymakers. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change**. 2013:4–5.

14. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *BMC Bioinformatics* 2010, **11**:94.

15. Mehta T, Tanik M, Allison DB: **Towards sound epistemological foundations of statistical methods for high-dimensional biology.** *Nat Genet* 2004, **36**:943–947.

16. Gadbury GL, Xiang Q, Yang L, Barnes S, Page GP, Allison DB: **Evaluating statistical methods using plasmode data sets in the age of massive public databases: An illustration using false discovery rates**. *PLoS Genet* 2008, **4**.

17. Steibel JP, Poletto R, Coussens PM, Rosa GJM: **A powerful and flexible linear mixed model framework for the analysis of relative quantification RT-PCR data**. *Genomics* 2009, **94**:146–152.

18. Reeb PD, Steibel JP: **Evaluating statistical analysis models for RNA sequencing experiments**. *Front Genet* 2013, **4**(September):1–9.

19. Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to SAGE data**. *Biostatistics* 2008, **9**:321–332.

20. Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data**. *BMC Bioinformatics* 2013, **14**:91.

21. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11**:R106.

22. Hogg R, McKean J, Craig A: *Introduction to Mathematical Statistics*. Pearson; 2013:341–347.

23. Agresti A: *An Introduction to Categorical Data Analysis*. 2nd edition. Wiley; 2007:74 – 90. [*Wiley Series in Probability and Statistics*]

24. Shelton J, Samarakoon T, Song Z, Jeannotte R, Basil N, Welti R, Bello N, Raithel S, Galliart M, Johnson L: **Divergent Epicuticular waxes and transcriptomes of sand and big bluestem**. 2015.

25. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome Res* 2004, **14**:1147–1159.

26. Zeng V, Villanueva KE, Ewen-Campen BS, Alwes F, Browne WE, Extavour CG: **De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean Parhyale hawaiensis**. *BMC Genomics* 2011, **12**:581.

27. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.

28. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**:357–359.

29. Chen Y, Mccarthy D, Robinson M, Smyth GK: **edgeR : differential expression analysis of digital gene expression data User ' s Guide**. 2014(April).

30. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *J R Stat Soc Ser B* 1995:289–300.