Our main points include:

1. using multiple (a larger number of) public-available existing/past data sets,
    (a) using multiple data sets
        i. as compared to using a single data set: a numerical measure of expression stability (typically, measures of certain aspects of RNA-Seq count variation) can be more reliably estimated by using more data sets (???)
        ii. (leaf) learn variance components (see GLMM)
        iii. (leaf) using top 1000 identified genes for normalization
        iv. (discussion) use as prior information in Bayesian analysis or use the set for quality control/vanity check purpose
    (b) (caveat) genes that are stable under a range of conditions might not be the most stable under a particular condition (under a single experiment).
        i. (leaf) identify different reference sets for different tissue types
    (c) Subtle points on interpretability and comparability (??? do we need some toy examples)
        i. using an explicit reference set: improve interpretability
        ii. using a common reference set (when comparing two or more studies): improve comparability
2. using a numerical measure of stability
    (a) (leaf) stability of house-keeping genes (HKGs)
    (b) (leaf) different numerical measures (geNorm, normFinder)
    (c) (leaf) different data sources (microarray and RNA-Seq)
    (d) (discussion) numerical measure of stability vs biological stability
3. Validate the stably expressed genes that we find? (no leaf yet)
    (a) biological function (online database)
    (b) (GO analysis)
4. Future, use our methods and leaf (stable set, rankings, variance components) in real studies.

Identification of stably expressed genes from Arabidopsis RNA-Seq data

Abstract

We examined RNA-Seq data on 209 biological samples from 23 different experiments carried out by different labs and identified genes that are stably expressed across biological samples, experiment conditions, and labs. We fit a random-effect model to the read counts for each gene and decompose the total variance to into between-sample, between-treatment and between-experiment variance components. Identifying stably expressed genes is useful for count normalization and differential expression analysis. The variance component analysis is a first step towards understanding the sources and nature of the RNA-Seq count variation.

1 Introduction

(overview) RNA sequencing (RNA-Seq) has become the technology of choice for transcriptome profiling over the last few years. The exponential growth in RNA-Seq study has accumulated a large amount of *Arabidopsis thaliana* (Arabidopsis) data under a variety of experimental/enviromental conditions. It is only natural to begin exploring how the large amount of existing data sets can help the analysis of future data. In this paper, we discuss identifying stably expressed genes from multiple existing RNA-Seq data sets based on a numerical measure of stability. We envision that such identified stably expressed genes can be used as a reference set or prior information for count normalization and differential expression (DE) analysis of future RNA-Seq data sets obtained from similar or comparable experiments. We also fit a random-effect model to the read counts for each gene and decompose the total variance to into between-sample, between-treatment and between-experiment variance components. The variance component analysis is a first step towards understanding the sources and nature of the RNA-Seq count variation. To illustrate our methods, we examined RNA-Seq data on 209 Arabidopsis samples from 23 different experiments carried out by different labs and identified genes that are stably expressed across biological samples, experiment conditions, and labs.

A reference set of stably-expressed genes will be useful for count normalization. A key task of RNA-Seq analysis is to detect DE genes under various experimental or environmental conditions. Count normalization is needed for adjusting differences in sequencing depths or library sizes (total numbers of mapped reads for each biological sample) due to chance variation in sample preparation. In DE analysis, gene expression levels are often estimated from relative read frequencies. For this reason, normalization is also needed to account for the apparent reduction or increase in relative read frequencies of non-differentially expressing genes simply to accommodate the increased or decreased relative read frequencies of truly differentially expressing genes. Many existing normalization methods, such as the trimmed mean of M-values normalization method (TMM) (Robinson et al., 2010) and Anders and Huber's normalization (Anders and Huber, 2010), will assume that the majority of the genes are not DE within an experiment and examine the sample distribution of the fold changes between samples. If the experiment condition can affect expression levels of more than half of the genes, many of the existing normalization methods may be unreliable (Lovén et al. (2012), Wu et al. (2013)). This difficulty can be alleviated if one could identify a set of stably expressed genes whose expression levels are known or expected to not vary much under different experimental conditions. Our idea is to identify such a reference set based on a large number of existing data sets.

Our basic intuition is that a numerical quantification of expression stability—which typically measures certain aspects of RNA-Seq count variation—can be more reliably estimated by using more data sets. There is, however, a caveat to this idea: as pointed out by Hruz et al. (2011), universally stably expressed genes may not exist and a subset of stably expressed genes from a specific biological context may have less variability than those identified across varying tissues and conditions. Many studies have shown that stably expressed genes are subject to change from one experiment to another, either due to varying experimental protocols, or due to different organs of a given species (Reid et al. (2006), Hong et al. (2010)). The top 100 stably expressed genes in Arabidopsis developmental series of Czechowski et al. (2005) shared only 3 genes with top 50 stably expressed genes identified from Arabidopsis seed samples by Dekkers et al. (2012). In this study, we try to balance the generality and specificity by identifying different reference gene sets for different tissue types of Arabidopsis.

We can also think that when a normalization method is applied to a single data set, it effectively specifies an implicit reference set of stably expressed genes (those genes that have the least variation after normalization). We can think this as using an internal reference set. In contrast, what we are proposing is that one can also identify an external reference set by looking at past data sets. The internal and external reference sets will provide different contexts for the DE analysis: in other words, one can choose to answer different scientific questions by using different reference sets. In any case, we advocate making the reference set explicit during a DE analysis and using a common reference set when analyzing multiple datasets. We will further discuss these points in the discussion Section 4.

In this paper, we identify stably expressed genes from RNA-Seq data sets based on a numerical measure—the sum of three variance components estimated from a mixed-effect model. We want to clarify that there is a distinction between numerical stability and biological stability—often times, we may not understand the biological functions of genes with numerically stable expression measures. From an operational point of view, however, numerical stability is more tractable. In pre-genomic era, the so-called "house-keeping genes" are often considered as candidates of reference genes for normalization (Bustin (2002), Andersen et al. (2004)). House-keeping genes are typically constitutive genes that maintain basic cellular function, and therefore are expected to express at relatively constant levels in non-pathological situations. However, many studies have shown that house-keeping genes are not necessarily stably expressed according to numerical measures (a review can be found in Huggett et al. (2005) and reference therein). For example, in the microarray analysis of Arabidopsis, Czechowski et al. (2005) showed that traditional house-keeping genes such as ACT2, TUB6, EF-1α are not stably expressed, and thus not good reference genes for normalization. Spike-in genes have also been considered as reference genes for normalization, but Risso et al. (2014) showed that spike-in genes are not necessarily stably expressed according numerical measures either. For microarray data, there are many efforts to numerically find stably expressed genes by quantifying the variation of measured expression levels across a large number of microarray data sets. For example, Czechowski et al. (2005) measured the expression stability of each gene using the coefficient of variation (CV). Genes with lower CVs are considered as more stably expressed. By investigating 721 arrays under 323 conditions throughout development, Czechowski et al. (2005) suggested stably expressed (reference) genes under different experimental conditions for Arabidopsis. Stamova et al. (2009), Dekkers et al. (2012), Gur-Dedeoglu et al. (2009), and Frericks and Esser (2008) screened a large number of microarray data sets to identify stably expressed genes in human blood, Arabidopsis seed, breast tumor tissues, and mice respectively. Validation experiments (Czechowski et al. (2005), Dekkers et al. (2012), Huggett et al. (2005), Stamova et al. (2009)) showed that these genes are more stably expressed than traditional house-keeping genes.

The rest of the paper is organized as follows: in Section 2, we describe the data preparation steps and our method of identifying stably expressed genes; in Section 3, we discuss the stably expressed genes and factors that might affact stability ranking; we also discuss results from variance component analysis and how to use the identified stably expressed genes for count normalization. (AND DISCUSSION...)

2 Methods

(Overview) In this section, we discuss the data preparation and our method of identifying stably expressed genes. In Section 2.1, we describe the steps for collecting and processing data. In Section 2.2, we introduce Anders and Huber's method that will be used for count normalization in our model. In Section 2.3 we describe a generalized linear mixed model (GLMM) (McCullagh and Nelder, 1989) and estimate three variance components for each gene: the *between-sample*, *between-treatment* and *between-experiment* variances. The *total variance* is defined to be the expression stability measure associated with that gene. Genes with smaller total variance are considered to be more stably expressed.

2.1 RNA-Seq data collection and processing

For evaluating the gene expression stability, we prepared RNA-Seq data for 209 Arabidopsis samples in 23 experiments.

Selecting experiments

The *Gene Expression Omnibus* (GEO) repository at *National Center for Biotechnology Information* (NCBI, `http://www.ncbi.nlm.nih.gov/`) stores raw sequencing data from a large number of RNA-Seq experiments. For this study, we restrict our attention to Arabidopsis experiments satisfying the following conditions: 1. Ecotype = "Columbia" (we kept only the Columbia samples from experiments that compare Columbia samples to other ecotypes); 2. Library strategy= "RNA-Seq"; 3. Library source = "transcriptomic"; 4. Library selection= "cDNA"; 5. Library layout = "Single End"; 6. There are at least 2 biological replicates for each treatment. We screened all the Arabidopsis experiments available from the NCBI GEO repository up to May 31, 2015 and downloaded raw RNA-Seq data (SRA files) from 49 experiments.

We assembled our own in-house pipeline to process all the raw RNA-Seq data: align the raw RNA-Seq reads to the reference genome and summarize the read counts at the gene level. In the GEO repository, the mapped read counts are unavailable for some experiments and the available ones are from different processing pipelines. Our pipeline, implemented using the software R (R Core Team, 2015), is summarized as follows:

1. Convert the SRA files to FASTQ files using the NCBI SRA Toolkit (version 2.3.5-2, `http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software`).

2. Align short reads in FASTQ files, using the Subread aligner (RSubread, version 1.16.2, Liao et al. 2013) in the software R (R Core Team, 2015), to the Arabidopsis reference genome downloaded from the *Ensembl plants FTP server* (`http://plants.ensembl.org/info/data/ftp/index.html`).

3. Summarize the read counts at the gene level using the `featureCounts()` function from the Subread aligner and store the read counts as data matrix. The annotation file is set to be `Arabidopsis_thaliana.TAIR10.22.gtf` in `featureCounts()`, which is desired for matching the short reads to Arabidopsis genes.

Subread aligner is a recently developed sequence mapping tool that adopt a seed-and-vote paradigm to map the short reads to the genome location. It uses a relatively large number of short seeds (called subreads) extracted from each read and allows all the seeds to vote on the optimal location. Compared to other alighers such as Bowite 2 (Langmead and Salzberg, 2012) or BWA (Li and Durbin, 2009), Subread is both faster and more accurate (Liao et al., 2013). For people familiar with R, it also has the advantage that it is completely implemented in R.

(**Three groups of experiments**) Czechowski et al. (2005), Hruz et al. (2011), and Dekkers et al. (2012) show that gene expression vary across different tissue types and development stages. For this reason,

We divided the experiments into three groups: the *seedling group* where only Arabidopsis seedling are included, the *leaf group* where only Arabidopsis leaves are included, and the *multi-tissue group*

where different tissue types are included. Within each group, we pre-screened experiments by data quality: we chose experiments with initial normalization factors (discussed in Section 2.2) ranging from 0.70 to 1.30; that is, the experiment is retained only when the minimum of normalization factors is greater than 0.70 and the maximum of normalization factors is less than 1.30. If the initial estimated normalizations is too different from 1, it indicates the read counts distribution in the corresponding sample is markedly different from the distributions of the rest of the samples. Such samples may demand additional attention before being incorporated in studies that we intend to do ....

Eventually, we obtained 23 data sets (with 209 biological samples): 10 experiments (70 samples) for the seedling group, 5 experiments (60 samples) for the leaf group and 8 experiments (79 samples) for the multi-tissue group (that is: shoot apical, root tip, primary root, inflorescences and siliques, hypocotyl, flower, carpels, aerial tissue). Samples from different experiments were merged by their unique gene IDs, and then stored as read count matrices. A brief description of the data sets can be found in Table (REF) of the online supplementary material. More details for each experiment are also accessible at NCBI via the unique accession number provided in Table (REF).

(Selecting genes) Prior to analysis, we removed genes with low overall counts as suggested by Anders et al. (2013). Filtering lowly-expressed genes is helpful not only because such genes provide little information about expression level, but also because they will cause convergence failure to our model. In practice, convergence issue can be avoided by removing genes with overall mean count less than 3. Table 1 summarizes the three groups after lowly-expressed genes are removed.

Table 1: data set summary

Group	# experiments	# treatments	# samples	# genes
Seedling	10	31	70	24379
Leaf	5	28	60	20967
Multi-tissue	8	35	79	23666

2.2 Count normalization

Many existing normalization methods, such as the trimmed mean of M-values normalization method (TMM) (Robinson et al., 2010) and Anders and Huber's normalization (Anders and Huber, 2010), assume that the majority of the genes are not DE within an experiment. Effectively, these methods use genes with relatively small observed fold changes under a single experiment as a reference gene set in normalization. In this paper, we choose Anders and Huber's method (Anders and Huber, 2010) to estimate normalization factors. Briefly, let y_{ij} denote the read count for ith gene of jth sample, then a pseudo-reference sample is created, with gene ith expression value defined as the geometric mean of the same gene over all real samples,

$$z_i = (\prod_{j=1}^{n} y_{ij})^{1/n}, i = 1, \ldots, m$$

The *normalization factor* for sample j is then calculated as the median of the fold-changes between sample j and the pseudo-reference sample over all genes

$$R_j = \text{median}(\frac{y_{1j}}{z_1}, \ldots, \frac{y_{nj}}{z_n}) \quad (1)$$

In this paper, we use the implementation in NBPSeq package (Di et al., 2014) for count normalization.

We apply an iterative procedure to estimate the normalization factors: initially, we use all the genes to calculate the normalization factors, which serve as a part of the offset term in the GLMM, and rank all the genes by our stability measure (see Section 2.3); next, at each iteration, we select the top 1000 stably expressed genes — ranking based on previous iteration — as reference genes to recalculate the normalization factors. In practice, over 950 genes are overlapped between top 1000 genes from the first iteration and those from the second iteration according to our stability measure. We therefore recommend one iteration to be sufficient.

2.3 Poisson log-linear mixed-effects regression model

Let Y_{ijkl} denote the read count for ith gene in jth observational unit of kth treatment group in lth experiment. The index i is suppressed herein since only one gene is evaluated at a time by GLMM. We assume that each read count Y_{jkl} follows a Poisson distribution with mean μ_{jkl} modeled as

$$\log(\mu_{jkl}) = \xi + \log(R_{jkl}N_{jkl}) + \alpha_l + \beta_{k(l)} + \epsilon_{jkl}, \tag{2}$$

where αs are experiment effects, βs are treatment effects nested in experiments, and ϵs are the effects for biological samples. N_{jkl} is the library size (column sum), and R_{jkl} is the normalization factor discussed in Section (2.2). The product term $R_{jkl}N_{jkl}$ is called *normalized library size* in DE analysis.

The between-experiment effect α in equation (2) is treated as random for two considerations. First, we view the collected data sets as a random sample from the pool of all Arabidopsis RNA-Seq experiments. Second, we expect the results from this study to be generalizable to future experiments when it comes to stably expressed genes and variance component analysis. We also consider the treatment effects β as random in the sense that treatment in future experiments may be different from this study.

We assume that αs, βs and ϵs are mutually independent, and

$$\alpha_l \sim N(0, \sigma^2_{\text{experiment}}), \quad \beta_{k(l)} \sim N(0, \sigma^2_{\text{treatment}}), \quad \epsilon_{jkl} \sim N(0, \sigma^2_{\text{sample}}).$$

where $\sigma^2_{\text{experiment}}$, $\sigma^2_{\text{treatment}}$ and σ^2_{sample} are called *variance-components*. They capture the *between-experiment*, *between-treatment* and *between-sample* variation correspondingly. The between-sample variance plays a similar role as the over-dispersion, which represents extra-Poisson variation in read counts under a negative binomial model (Anders and Huber (2010), Di et al. (2011)). The between-treatment term accounts for variation in treatment condition (e.g., genotype, growth medium). The between-experiment variation includes all other possible variations that are difficult to be separated statistically, for example, lab personnel and conditions, day light hours, age of the plants, temperature, sequencing platform, etc.

The stability measure of a gene is defined by the total variance

$$\hat{\sigma}^2 = \hat{\sigma}^2_{\text{sample}} + \hat{\sigma}^2_{\text{treatment}} + \hat{\sigma}^2_{\text{experiment}} \tag{3}$$

We rank all the genes according to their $\hat{\sigma}^2$s, and consider highly ranked (top 1000) genes to be stably expressed.

2.4 Other stability measures

There are many other gene expression stability measures for microarray data, among which the M-value in *geNorm* (Vandesompele et al., 2002) and the ρ value in *NormFinder* (Andersen et al., 2004) are popular. In geNorm, the relative variation of genes i to gene i_1 is calculated as the standard deviation of the log fold changes between the two genes, SD_{i,i_1} = standard deviation$(\log_2 \frac{y_{i,1}}{y_{i_1,1}}, \ldots, \log_2 \frac{y_{i,m_0}}{y_{i_1,m_0}})$ where m_0 is the total number of genes in the reference set; then the stability value M_i for gene i is the mean of pairwise variation between gene i and the other $m_0 - 1$ genes,

$$M_i = \sum_{i_1 \neq i} \text{SD}_{i,i_1}/m_0 \tag{4}$$

Stably expressed genes are expected to have low M-values. NormFinder uses a linear mixed model to estimate the between-group and within-group variations from expression values of microarray data, and then combines the two variations by a Bayesian formulation. For a future experiment, the stability of a given gene is defined as the its absolute value of posterior mean plus its standard prediction deviation. Alternatively, Czechowski et al. (2005) and Dekkers et al. (2012) use CV (standard deviation/mean) to measure expression stability, with smaller CV corresponding to more stably expressed genes.

Stability measures differ in summarizing variation of gene expression profiles. The total variance $\hat{\sigma}^2$ considers genes to be stably expressed when they have low $\hat{\sigma}^2$ in the estimated log relative mean

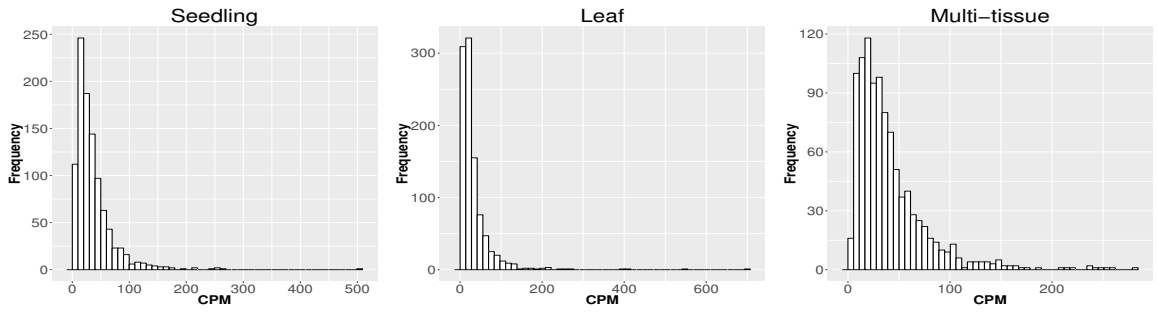


Figure 1: mean CPM (see equation (5)) for the top 1000 most stably expressed genes identified from the seedling (left), leaf (middle) and multi-tissue (right) groups by the total variance $\hat{\sigma}^2$. The mean CPM is computed over all samples for each gene.

counts; the M-value in geNorm identifies as stably expressed those genes that are most similar to each other; and the CV measure (Czechowski et al., 2005) tends to select highly expressed genes as stable since those genes tend to have smaller dispersion (Hruz et al., 2011). Similar to GLMM for RNA-Seq data, the NormFinder uses a linear mixed model to capture the variation in microarray data, which are usually assumed to be normally distributed. In addition, NormFinder requires a minimun of eight samples per treatment (Andersen et al., 2004). Therefore NormFinder will not be discussed in this paper.

3 Results

In Section 3.1, we summarize stably expressed genes identified from three different experiment groups and one emphasis is that stability is context dependent. In Section 3.2, we show that traditional house-keeping genes are not necessarily stably expressed according to our numerical measure, and that microarray data and RNA-Seq data will give different sets of stably expressed genes. In Section 3.3, we further demonstrate that when using a numerical measure to measure gene expression stability, the outcome will depend on the specific numeric measure used. These points should be intuitive, but they are not often emphasized in practice. In Section 3.4, we discuss results from variance component analysis. In Section 3.5, we discussed how to use the identified stably expressed genes for count normalization.

3.1 Stably Expressed Genes

Using the total variance, $\hat{\sigma}^2$, from the GLMM (see equation (2) in Section 2.3) as a stability measure, we identified stably expressed genes in three groups of experiments described in Section 2.1: the group of seedling experiments, the group of leaf experiments, and the group of experiments on different tissue types (see Table 1 for a summary). (The seedling group consists of 70 samples from 10 experiments on Arabidopsis seedlings; the leaf group consists 60 samples from 5 experiments on Arabidopsis leaf; and the multi-tissue-type group consists 79 samples from different 8 experiments on different tissues types.) As we mentioned in Introduction, absolutely stably expressed genes may not exist. Choosing different sample sets as reference allows us to identify the stably expressed genes for different biological contexts.

In Tables 1-3 (REF) in the online supplementary materials, we summarize the top 1000 most stably expressed genes in each group. In Figure 1, we summarize the histograms of the mean Count Per Million (CPM) for the 1000 most stably expressed genes identified in each group. For each gene, the CPM is computed as

$$\frac{\text{count} \times 10^6}{\text{normalized library size}} \tag{5}$$

in each sample and the mean is computed over all samples.

The lists of top 1000 genes in the three groups share 106 genes in common (see supplement material for detail). These genes are stably expressed under a wide range of experimental conditions and in different tissue types, and thus may be worth further study. This list of 106 genes has significant overlap with the top 100 stably expressed genes identified by Czechowski et al. (2005) from a developmental series of microarray samples: 10 out of these 106 genes (see Table REF in the supplement material for details),

AT5G46630, AT4G24550, AT1G13320, AT5G26760, AT1G10430,
AT4G27120, AT3G01150, AT3G10330, AT4G32560, AT2G20790.

appeared in the list of top 100 stably expressed genes out of 14000 genes he examined (the probability is 3.68×10^{-10} for a list of 106 genes random selected from a set of 14000 genes to have an overlap of size 10 or more with a pre-selected list of 100 genes). In particular, one gene, AT1G13320, is in all ten but one list of top 500 stably expressed genes identified by Czechowski et al. (2005) for different experimental and experimental conditions (the only exception is the set of diurnal series, ??? Jeff), and is also identified by Hong et al. (2010) as a stably expressed gene under all six but one experimental conditions he examined. This gene is ranked 446 (top 1.8%), 112 (top 0.5%), 687 (top 2.9%) according to our stably measure in the three groups we examined. (??? Is there anything special about this gene? Jeff) This gene is a subunite of protein phosphatase type 2A complex and involves in regulation of phosphorylation and regulation of protein phosphatase type 2A activity. It has been used as a reference gene for normalization in many papers (e.g., Bournier et al. (2013), Baron et al. (2012); these two papers cited Czechowski et al. (2005) as reference).

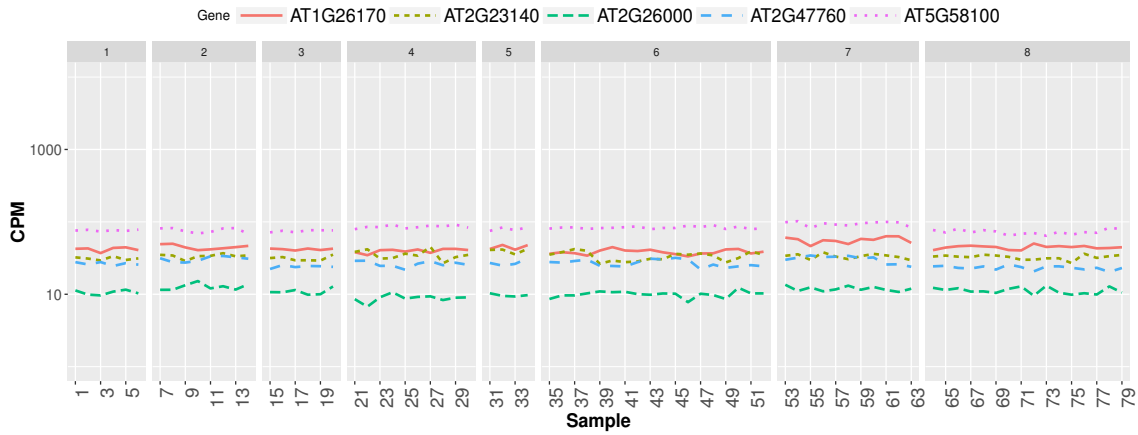
3.2 Comparison to house-keeping genes and stably expressed genes identified from microarray data

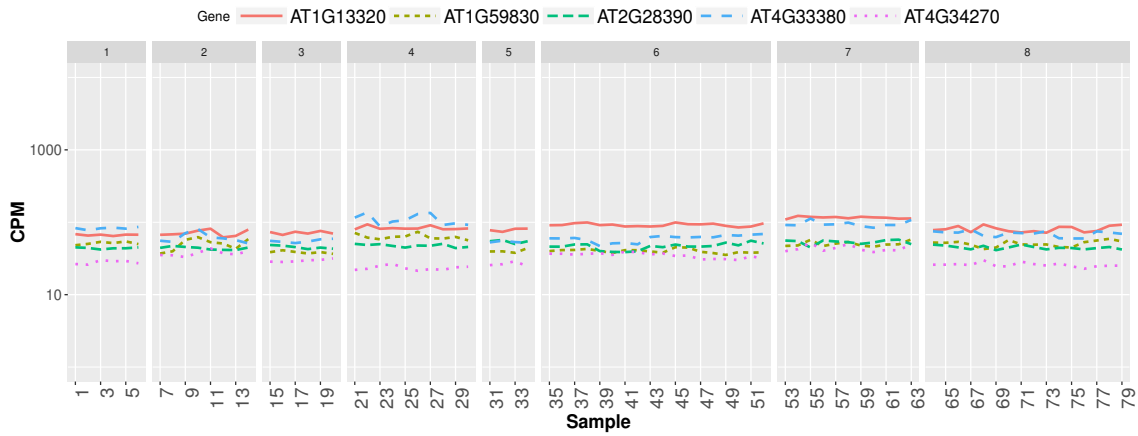
Czechowski et al. (2005) discussed the expression stability of house-keeping genes and showed that the house-keeping genes are not stably expressed according to their numerical measure. In particular, they compared the expression profiles of five traditional house-keeping genes (AT1G13440, AT3G18780, AT4G05320, AT5G12250, AT5G60390) and five genes (AT1G13320, AT5G59830, AT2G28390, AT4G33380 and AT4G34270) that they identified as stably expressed according to the CV measure from a developmental series of microarray experiments (see Figure 1 of that paper). In Figure 2, we compare the expression profiles of these 10 genes from Czechowski et al. (2005) to the expression profiles of five genes (AT1G26170, AT2G23140, AT2G26000, AT2G47760, AT5G58100) that we randomly selected from the top 100 most stably expressed genes identified from the multi-tissue group RNA-Seq data according the total variance $\hat{\sigma}^2$. For each of the 15 genes, Figure 2 shows the expression levels measured in CPM over 79 samples in the eight experiments in the multi-tissue group, and Table 2 summarizes the variance components estimated from the GLMM in 2.3.

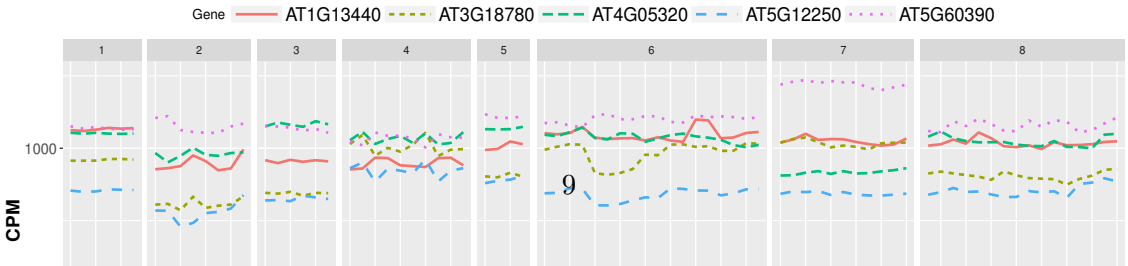
The five house-keeping genes show large total variation with all three variance-components relatively large as compared to the other 10 genes. This is consistent with Czechowski's observation that house-keeping genes are not necessarily stably expressed according to a numerical measure. Three of the five stably-expressed genes identified by Czechowski are among the top 1000 stably-expressed genes according to our stably measure the total variance $\hat{\sigma}^2$. Czechowski et al. identified those five genes from microarray data and different experiments. It is not too surprising those genes might not be the most stable in RNA-Seq experiments: the two technologies differ in many aspects including coverage and sensitivity.

Table 2: Variance components estimated from the multi-tissue group for 15 genes. Column 1 is the source: five stably expressed genes identified by the total variance $\hat{\sigma}^2$ (GLMM), five stably expressed identified by Czechowski according to the CV measure from a developmental series of microarray experiments (Czechowski), and five traditional house-keeping genes (HKG). Columns 3–5 are the values of each variance components. Column 6 specifies each gene's rank according to the total variance $\hat{\sigma}^2$ in the multi-tissue group.

Source	Gene	betweeen-sample	between-treatment	between-experiment	Rank
GLMM	AT5G58100	0.0018	0.0008	0.0042	9
	AT2G23140	0.0049	0.0058	0.0000	49
	AT2G26000	0.0028	0.0002	0.0079	53
	AT2G47760	0.0037	0.0031	0.0042	58
	AT1G26170	0.0025	0.0025	0.0069	77
Czechowski	AT2G28390	0.0032	0.0000	0.0042	13
	AT1G13320	0.0029	0.0008	0.0230	687
	AT1G59830	0.0043	0.0039	0.0199	782
	AT4G34270	0.0062	0.0000	0.0328	1466
	AT4G33380	0.0072	0.0033	0.0534	3136
HKG	AT5G12250	0.0163	0.0192	0.1337	7832
	AT1G13440	0.0189	0.0089	0.1624	8420
	AT5G60390	0.0082	0.0169	0.2150	9573
	AT4G05320	0.0092	0.0089	0.2299	9749
	AT3G18780	0.0360	0.0107	0.4168	12623







3.3 Factors affecting stability ranking

The previous two subsections demonstrate that when using a numerical measure to quantify gene expression stability, the outcome is dependent on 1) the biological context reflected in the reference sample set used and 2) the technology used for measuring gene expression. It should also be intuitive, and we will further clarify in the second half of this subsection, that 3) the stability ranking is also dependent on the specific numerical measure used. In this section, we will first compare the lists of stably-expressed genes identified under different scenarios where one or more of the above three factors differ. We then further discuss the subtle roles played by the specific stability measure and the reference gene set by comparing the total variance $\hat{\sigma}^2$ measure from the GLMM (see equation (2)) to the M-value measure used in the geNorm method (Vandesompele et al., 2002). Last, we discuss the effect of an iterative elimination procedure used by geNorm.

We look at an additional five lists of stably expressed genes identified under different scenarios and examine how each of these five lists overlaps with the the top stably-expressed genes identified from the multi-tissue group of RNA-Seq experiments according to the total variance measure $\hat{\sigma}^2$ (see Section 2.3). The five lists are:

L_1: 100 top stably expressed genes from the multi-tissue group according to the M-value in geNorm (applied to $\log(\text{count} + 1)$) of Vandesompele et al. (2002) ;

L_2: 100 top stably expressed genes from the seedling group according to the total variance $\hat{\sigma}^2$ from the GLMM;

L_3: 100 top stably expressed genes from the leaf group according to the total variance $\hat{\sigma}^2$ from the GLMM;

L_4: 100 stably expressed genes identified from a developmental series of microarray experiments by Czechowski et al. (2005) using the CV measure (REF);

L_5: 50 stably expressed genes identified by Dekkers et al. (2012) from microarray seed experiments using the CV measure.

In Figure 3, we plot the *recall* percentage for each list above against the number of top stably-expressed genes we selected as reference from the multi-tissue-type group. The recall percentage for L_i is defined as

$$\frac{\#\{L_i \cap \text{reference set}\}}{\#\{L_i\}} \times 100,$$

where $\#\{\}$ denotes the number of elements in the list. We have the following observations:

1. The list L_1 is identified from the same set of RNA-Seq experiments as the reference sets, but using a different stability measure (M-value in geNorm). This list has significant overlap with the top stably-expressed genes identified using the total variance measure: 35 and 99 out of the 100 genes from the list L_1 are among the top 100 and 1000 most stably-expressed genes, respectively, from the multi-tissue group identified using the total variance measure.

2. The lists L_2 and L_3 are identified from different sets of RNA-Seq experiments (leaf and seedling experiments) using the same stability measure as used for the reference sets. The lists L_4 and L_5 are identified from microarray experiments (a developmental series and a seed group) and using the CV measure. The overlapping (recall) percentages are still statistically significant, but much less than in the case of L_1. This shows that differences in tissue type and in measuring technology both influence the expression stability ranking, and to comparable degrees. The lists L_3 and L_5 have the least overlapping percentages with the reference sets. These lists are identified from a leaf group and a seed group respectively. Our understanding is that the leaf group and the seed group are more biologically homogeneous than the multi-tissue group and thus provide very different biological contexts for evaluating expression stability.

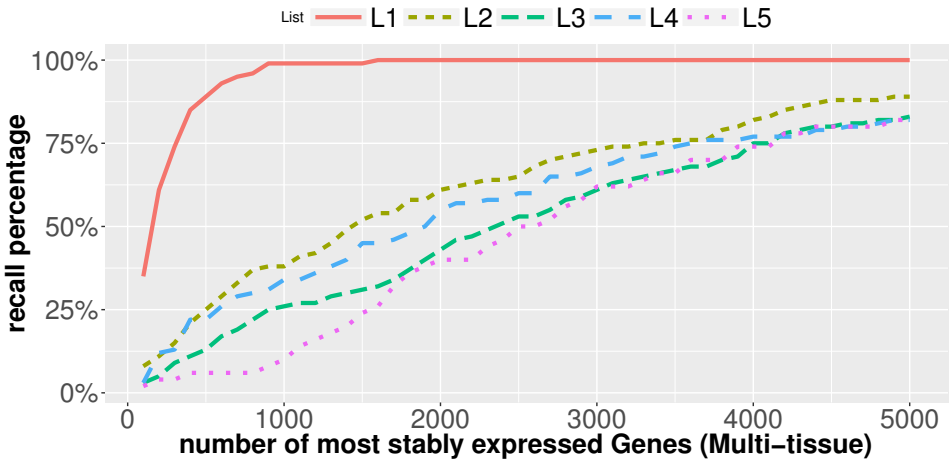


Figure 3: Recall percentage for list L1 — L5 of 3.3. we choose top 100 stably expresed genes, except Dekkers(2012) where only a list of 50 genes is available. x-axis is the number of most stably expressed genes and y-axis shows the recall percentage.

When applied to the same set of samples, the M-value and total variance measure $\hat{\sigma}^2$ give similar expression stability ranking: the rank correlation is 0.97 (see also, observation 1 above). We point out that the reason is because the M-value and normalization step needed for computing our total variance measure have similar fundamental assumptions. The basic principle behind the M-value is that the expression ratio of two stably-expressed genes should be identical in all samples. In formula, it means that the expression values of two stably-expressed genes i_1, i_2 in any two samples j_1, j_2 should satisfy

$$\frac{y_{i_1,j_1}}{y_{i_2,j_1}} = \frac{y_{i_1,j_2}}{y_{i_2,j_2}}. \tag{6}$$

Our total variance measure $\hat{\sigma}^2$ is estimated from normalized data. The basic assumption in the normalization step is that majority of genes are not DE. In formula, it means for any stably-expressed gene i_1, its expression level as measured by the relative frequency should be stable across all samples,

$$\frac{y_{i_1,j_1}}{S_{j_1}} = \frac{y_{i_1,j_2}}{S_{j_2}}, \tag{7}$$

where S_{j_1} to S_{j_2} are the normalized library sizes (i.e., $R_j N_j$ in equation (2). This implies for any two stably-expressed genes i_1 and i_2

$$\frac{y_{i_1,j_1}}{y_{i_1,j_2}} = \frac{y_{i_2,j_1}}{y_{i_2,j_2}} = \frac{S_{j_1}}{S_{j_2}}. \tag{8}$$

The first equation in (8) is equivalent to equation (6). (In practical application of both methods, the stability of any single gene is evaluated by comparing its expression to a set of reference genes. See the Method section for more details.)

In practice, the geNorm program (Vandesompele et al., 2002) is usually used to rank a set of reference genes identified from other methods. An iterative elimination procedure is used along with the M-value to determine the final ranks of the expression stability: after each iteration, the gene receiving the largest M-value will be removed and a new set of M-values will be computed for the remaining genes, and the iteration will go on until there are only two genes left. We did not use such an iterative procedure in the comparisons above (i.e., we only computed one set of M-values for all genes).

This iterative elimination procedure creates an extra layer of complexity that is not well explored in literature. We use a toy example below to illustrate one subtle aspect of the iterative elimination procedure. In this example, we consider the expression values of 7 genes in two samples shown in Table 3. When M-value is used to rank all 7 genes, the initial ranking of expression stability is given in column 4 of the Table: gene 7 is the least stable one and genes 4 and 5 are considered the most

stable ones. Once genes 6 and 7 are eliminated, however, the recalculated M-values will rank genes 1–3 as more stable than genes 4 and 5 (see column 5 of Table 3). The root cause of this reversal of ranking is that when an iterative elimination procedure is used, effectively, the reference gene set is changing after each iteration: in the initial ranking, the expression patterns genes 4 and 5 are close to the "middle of the pack" and thus considered as the most stable, and the expression patterns of genes 1–3 and genes 6 and 7 are considered relatively more extreme; once genes 6 and 7 are removed, however, the "middle of the pack" is shifted towards the expression patterns of genes 1–3, and thus genes 1–3 become the most stably expressed. With this understanding, one could and should make a concious decision on whether such a behavior as described above is desirable or not.

The point we want to emphasize is that the gene stability is a relative concept and the stability ranking depends on which set of genes we use as reference. In an iterative elimination procedure, the reference gene set will change after each iteration. The procedure can thus give surprising results and the adaption of it in practice should not be automatic.

Table 3: A toy example showing the effect of iterative elimination. Columns 2 and 3 are expression levels for two samples, Column 4 is the ranking of genes by M-value without iterative elimination, and Column 5 is the ranking after two iterations.

	Raw Counts		Rank	
Gene	sample 1	sample 2	rank1	rank2
Gene1	1	1	3	1
Gene2	1	1	3	1
Gene3	1	1	3	1
Gene4	1	2	1	4
Gene5	1	2	1	4
Gene6	1	3	6	
Gene7	1	4	7	
Library Size	7	14		
Note:	rank1 = initial ranking by M-value, rank2= ranking after Gene 6 and 7 removed.			

3.4 Sources of variation

For each gene, the GLMM (equation (2) of section 2.3) allows us to decompose total count variance into between-sample, between-treatment and between-experiment variance components. The estimated variance components tell us how much each component contributes to the overall count variation. Table 4 summarizes the percentages—averaged over all genes—of the total variance attributable to each of the three components for three groups of RNA-Seq samples (seedling, leaf and multi-tissue groups in Section 2.1). Figure 4 shows the histograms of the percentages. Figure 5 shows the stacked bar plot of variance components estimated for the multi-tissue group for 20 genes randomly selected from the top 1000 stably expressed genes and 20 genes randomly selected from all of the 23666 genes. As expected, the between-experiment variance component, on average, explains the largest proportion of the total variation. In the group of leaf experiments, the between-treatment variation is markedly greater than the between-sample variation, suggesting the existence of a higher proportion of DE genes.

([Ask Jeff], one implication is that DE is easier to detect in this group of experiments. Our intuition is that leaf samples tend to be more homogeneous and thus the treatment effect is easier to detect between leaf samples.)

Table 4: proportion of estimated variance components for the seedling, the leaf and the multi-tissue groups. Every entry represents the average proportion of a certain variance component, computed by averaging the ratios — the variance component being considered to the total variance $\hat{\sigma}^2$ — over all the genes.

	Seedling	Leaf	Multi-tissue
between-sample	12.3%	16.0%	8.4%
between-treatment	13.4%	28.0%	6.6%
between-experiment	74.3%	56.0%	85.0%

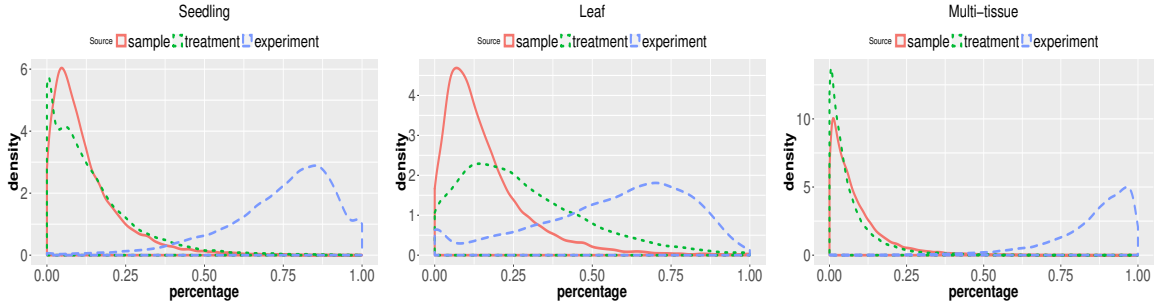


Figure 4: Density plot of percentages of variance components for the seedling, the leaf, and the multi-tissue groups (from left to right).

3.5 Reference gene set for normalization

Once we have ranked the genes according to our numerical stability measure (i.e, the total variance measure, $\hat{\sigma}^2$), one application is to use an explicit set of most stably expressed genes as reference genes for count normalization. The key difference of this approach from existing normalization approach that uses only the one data set under study is that this new approach allows investigators to prescribe a specific biological context for evaluating gene stability by choosing the most relevant reference samples and experiments when computing the stability measure. For example, the most stably expressed genes identified from the multi-tissue group and those identified from the seedling group will provide different biological contexts.

In the Introduction, we also argued that using an explicit set of genes as reference for normalization can improve interpretability of DE results, especially when one wants to compare results from two or more experiments, in the sense that DE can always be interpreted as relative to the explicit reference gene set used. We use a toy example to illustrate this point in Table 5 where we examine the mean counts for 5 genes in two two-group comparison experiments. If we use different reference gene set for count normalization, for example, we use genes 1–3 as reference in experiment 1, but use genes 3–5 as reference in experiment 2, we may conclude that gene 3 is not DE in either experiment. If we use

Table 5: My caption

	Exp. 1		Exp. 2	
Gene	Control	Treatment	Control	Treatmetn
1	10	20	10	20
2	10	20	10	20
3	10	20	10	10
4	10	10	10	10
5	10	10	10	10

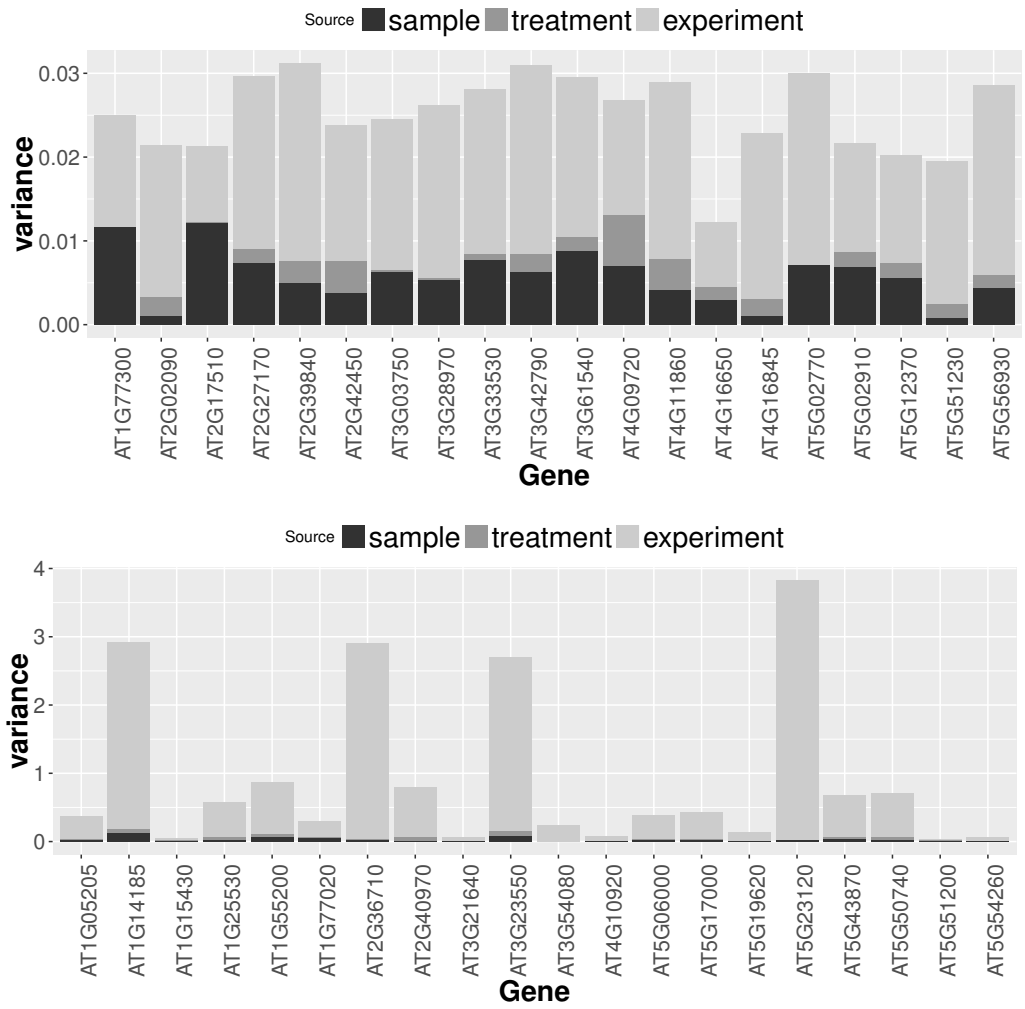


Figure 5: stacked bar plot of the three variance components for the multi-tissue group. Top: 20 genes randomly selected from top 1000 stably expressed genes; Bottom: 20 genes randomly selected from all the genes.

a common reference gene set—either genes 1–3 or genes 3–5—for normalization, however, we will be able to discover, in either case, that the DE behavior of gene 3 is different in the two experiments. Note that the DE conclusion in both experiments will depend on the reference genes used: if genes 1–3 are used as reference, gene 3 is not DE in experiment 1, but will be DE in experiment 2; if genes 3–5 are used as reference, gene 3 will be considered DE in experiment 1, but not DE in experiment 2. The point is, in either case, we will notice that the DE behavior of gene 3 is different between the two experiments. This information will be lost if one uses different reference sets to assess DE in the two experiments.

In practice, we recommend using the top 1000 most stably genes for estimating normalization factors. The key is to avoid using too few (e.g., less than 10) or too many (e.g., using all genes): intuitively, using too few, the estimates will be unstable; using too many, the results will be influenced by less stably expressed genes. Our simple simulations suggest that using between 100 to 10000 genes seems to give stable results. In the first set of three examples, we use Anders and Huber's method (see equation (1)) to estimate normalization factors for samples in each of the seedling, leaf and multi-tissue groups of experiments (see Section 2.1). We use the top 10, 100, 1000, and 10000 stably expressed genes identified earlier (see 3.1 for details) as reference gene set. Figure 6 shows the pairwise scatter plots and correlation coefficient between the normalization factors when different numbers of top stable genes are used as reference. The plots and correlation coefficients suggest using between 100 and 1000 genes tend to give similar normalization factor estimates. We also used the top 10, 100, 1000, and 10000 stably expressed genes identified from the seedling group as reference set for

estimating normalization factors for a set of six seedling samples from a new experiment. The largest Pearson correlation 0.995 is between the normalization factors estimated using the top 1000 and top 10000 stably expressed genes as reference.

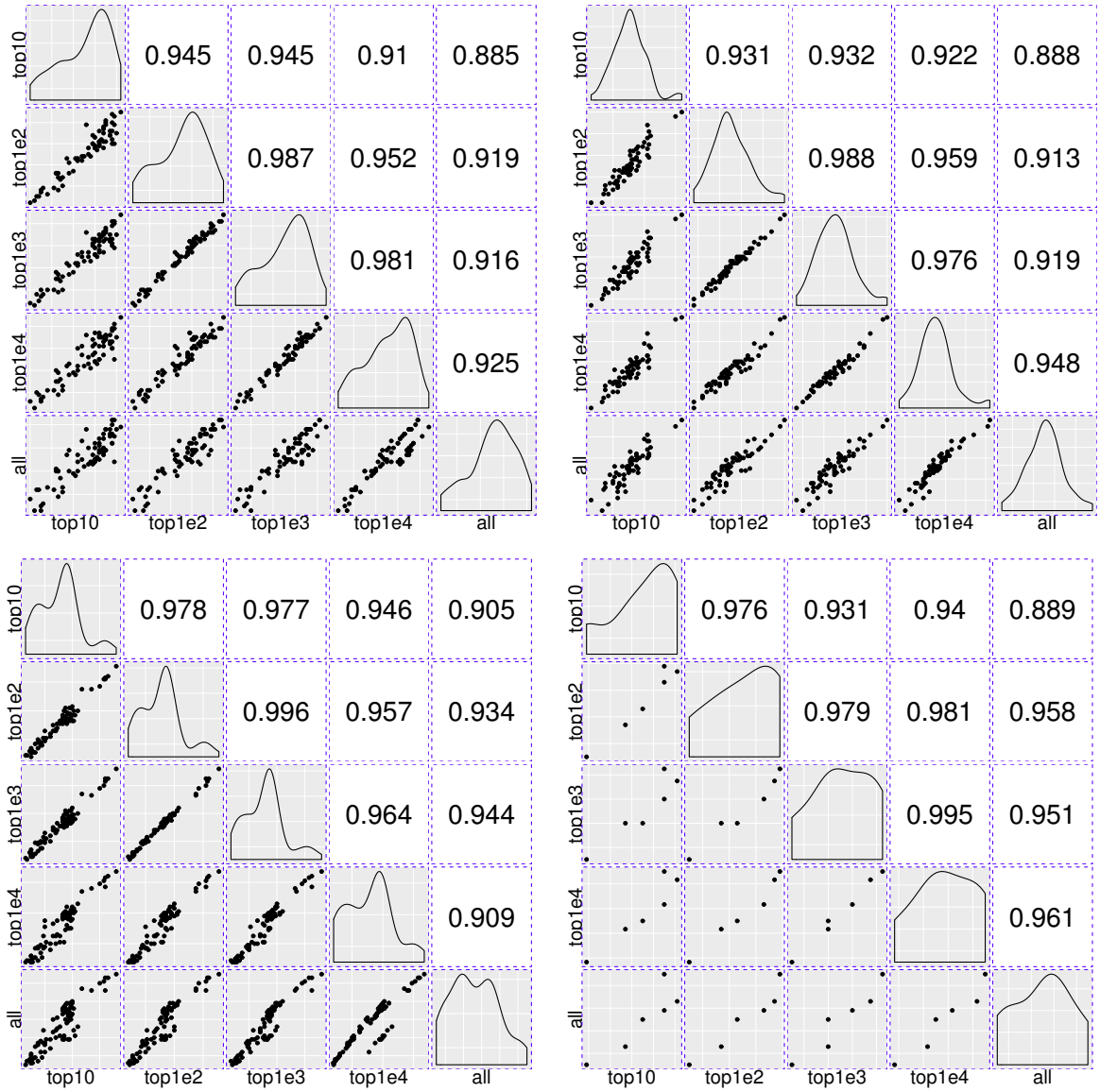


Figure 6: matrix plot of normalization factors by choosing top 10, 100, 1000, and 10000 stably expressed genes for the seedling (top left), leaf (top right), multi-tissue (bottom left) groups. The bottom right plot shows the normalization factors of a new seedling experiment GSE66666 (sample size is 6) when using top 10–10000 stably expressed genes from seedling group as reference genes.

4 Discussion

In this paper, we advocate quantifying gene expression stability by applying a numerical stability measure to a large number of existing data sets. This strategy has also been used by many for finding stably expressed genes using microarray data. When using such a strategy, the outcome is determined by two factors: the data sources used and the specific numerical stability measure. We emphasize that numerical stability is not equivalent to biological stability (??? Jeff). (Ask Jeff about "biological stability": do biologists talk about "biological stability"?) For example, we demonstrated that the expression levels of traditional house-keeping genes are not necessarily stable according to a

numerical measure. Biological stability is a vague term and not easy to quantify. Numerical stability is generally more tractable. (Some have argued that numerically identified stably expressed genes are more reliable. We feel that argument is somewhat circular??? The same numerical measure is used to both identify the stably expressed genes and to verify their stability.) It should be obvious but worth emphasizing that 1) different stability measures will give rise different ranking of gene stability and 2) the stability measure will also depend on the technology used for measuring gene expression: for example, microarray data and RNA-Seq data will give different sets of stably expressed genes.

Major findings of the study Normalization is an essential step for accurate inference in RNA-Seq data analysis. Global normalization methods may lead to erroneous conclusions when they rely on inappropriate reference genes. In this study we identified three sets of stably expressed genes for different organs and tissues of Arabidopsis. We conclued that traditional HKGs are not necessarily stable under not only microarray, but also in RNA-Seq experiment when they are evaluated by numerical stability measures. While Czechowski et al. (2005) identified novel reference genes for different experimental conditions, we demonstrated that they are not among the best candidates for RNA-Seq study. We recommend a set of 1000 reference genes to calculate normalization factors for RNA-Seq data.

4.1 Moved from Introduction part

(To Discussion ???) **Further discussion** Previous studies showed that normalization are needed to account for nuisance effects, including *between-sample* effects, e.g., sequencing depths, flow-cell/library preparation effects (Bullard et al. (2010), Robinson et al. (2010)), as well as *gene-specific* effects, e.g., gene length or GC-contents (Risso et al. (2011), Hansen et al. (2012)). A number of normalization approaches are proposed to address different types of unwanted nuisance effects (Dillies et al. (2013), Risso et al. (2014)). Different from global-scaling normalization, Risso et al. (2014) proposed a regression-based normalization-remove unwanted variation (RUV). In that paper, they regressed the read counts on the known covariates of interest (e.g. treatment effects) and unknown factors of nuisance effects. The factors of nuisance effects are estimated from a subset of data, and are then adjusted for in DE analysis. In RUVg approach, they are estimated through a factor analysis. A main assumption for RUVg is that a set of stably expressed genes can be identified first.

Why is stably expressed gene important/difficult Count normalization would have been easy if one could identify and use as reference a set of stably expressed genes whose expression levels are known or expected to not vary much under different experimental conditions. Effectively, TMM or DESeq method use genes with relatively small observed fold changes under a single experiment as a reference gene set in normalization. There are two obvious issues with this strategy: 1. The available sample size in any single experiment may be too small for us to reliably estimate true fold changes. 2. If the experiment condition can affect expression levels of more than half of the genes (Lovén et al. (2012), Wu et al. (2013)), many of the existing normalization methods (???) may be unreliable.

To identify a set of stably expressed genes, our method still need to estimate an initial set of normalization factors where we need to make assumptions about relative fold changes between samples. This kind of circular dependence seems unavoidable (Vandesompele et al., 2002). Our stretagy is to use an iterative procedure: we rank all the genes based on our stability measure, and use top 1000 stably expressed genes to calculate normalization factors, which are the new offsets in the next iterative GLMM estimating procedure. After five iterations, the top 1000 genes have a large overlap (90.9%) with the top 1000 genes from the first iteration. In practice, we recommend one iteration to be enough.

Stably expressed genes may or may not be the most stably expressed ones for a particular experiment when they are identified by pooling multiple experiments. (Discussion) Two subtle points we want to make: 1) using an explicit reference set improves interpretability of DE test leaf; 2) using a common reference set improve comparability when analyzing two or more data sets. (moot ???) ... Another subtle point we want to make is that a reference set does not have to be absolutely stable to be useful as a reference set: we can slightly change our perspective and interpret all DE leaf as relative to the reference set. ??? For example, a fold change of 2 can be interpreted as the fold change of this gene is 2 more than those genes in the reference set. ??? Any of the simple normalization methods discussed earlier (REFs) are effectively specifying an implicit set of genes as a referent set.

Our proposal is to make the reference set explicit to improve interpretability of the leaf. Furthermore, using an explicit common reference set becomes more useful when the interest is in comparing different experiments. For example, when two RNA-Seq data sets are separately normalized with different reference sets, a fold change of two observed in one experiment may not be directly comparable to a fold change of two observed in the other. This concern can be alleviated by using a common set of reference genes. Different estimated normalization factors effectively specify a different reference set ... (Improve intepretability and comaprability)

(???) Furthermore, new biological insights (REFs) ... Stably expressed genes are likely to be involved in basal metabolic or 'house-keeping' functions, such as kinase activity, nucleotide binding and protein modification processes. Sekhon et al. (2011) and Wang et al. (2010) showed that stably expressed genes are involved in biological processes included cellular processes, transport, protein modification, translation and signal transduction by Gene Ontology enrichment analysis. Besides, in expression study, a high correlation between translational signiture and mRNA level is found in human stably expressed genes(Line et al., 2013). In that paper, a significant increase in mRNA variation prediction was obtained by selecting genes that are stably expressed in more than 1 tissue.

4.2 Alternative method

Another widely adopted approach of fitting GLMM to the data is via negative binomial (NB) regression. In many cases, RNA-Seq data analysis begins with the assumption that Y_{jkl} follows a negative binomial distribution (a.k.a Poisson-Gamma mixture). The NB model introduces a dispersion parameter to capture the extra-Poisson biological variation. In NB regression, we estimate between experiment and between treatment variation. Specifically, for each gene, we assume $Y_{jkl} \sim NB(\mu_{jkl}, \phi)$ with the link function

$$\log(\mu_{jkl}) = \xi + \log(R_{jkl}N_{jkl}) + \alpha_l + \beta_{k(l)}$$

where similarly, α_l is the random effect for experiment, and $\beta_{k(l)}$ is random treatment effect nested in experiments. The only difference is that the dispersion ϕ, rather than variance of biological sample in Poisson regression, is estimated in NB setting. We saw no significant difference in estimating the variance components between these two approaches. The NB regression is run by `glmer.nb()` in `lme4` package(Bates et al., 2012) and `glmmadmb()` in `glmmADMB` package(Bolker et al., 2012). Unfortunately, both implementations of NB regression experienced convergence failure when modeling over 20,000 genes.

A limitation of this study is that the inherent design structures are not taken into account unless when the experiment is a case-control (single factor) study. Our concern is two fold: one, although we collected more than 150 samples, they are far from enough for a complicated design structure because usually there are only 2 or 3 replicates within each treatment; two, efficient algorithm is not available for generalized linear mixed model with more than three random effects (Bolker et al., 2009). However, model (2) is sufficient for the purpose of identifying stably expressed genes in this paper.

5 Appendix

Estimation of Variance Components and Identification of Stably Expressed Genes

The estimation procedure starts from the joint density function of $\boldsymbol{Y} = (Y_{jkl})'$ given $\boldsymbol{\mu} = (\mu_{jkl})'$,

$$f(\boldsymbol{Y}|\boldsymbol{\mu}) = \prod_{j,k,l} f(y_{jkl}|\mu_{jkl}) = \prod_{j,k,l} \frac{[\mu_{jkl}]^{y_{jkl}} \exp(-\mu_{jkl})}{y_{jkl}!}$$

A re-expression of (2) in matrix form gives

$$\log \boldsymbol{\mu} = \boldsymbol{\xi} + \log \boldsymbol{NR} + \boldsymbol{Z_1\alpha} + \boldsymbol{Z_2\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\xi} = \mathbf{1} \cdot \xi$ and $\mathbf{1}$ is a vector of 1s, $\boldsymbol{Z}_1$ is the design matrix for random effects $\boldsymbol{\alpha} = (\alpha_l)$, and $\boldsymbol{Z}_2$ is the design matrix for random effects $\boldsymbol{\beta}$. Therefore $\boldsymbol{\mu} \sim \log N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu}_0 = \boldsymbol{\xi} + \log(\boldsymbol{NR}),$$

$$\boldsymbol{\Sigma} = \sigma_1^2 \boldsymbol{Z_1 Z_1'} + \sigma_2^2 \boldsymbol{Z_2 Z_2'} + \sigma_0^2 \boldsymbol{I}$$

and $\boldsymbol{I}$ is an identity matrix of dimension Q where Q is the total number of biological samples. And

$$f(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \prod_{j,k,l} \mu_{jkl}^{-1} \cdot \frac{1}{\sqrt{(2\pi)^Q |\boldsymbol{\Sigma}|}} \exp[-\frac{1}{2}(\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)]$$

the joint density is then

$$f(\boldsymbol{Y}, \boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^Q |\boldsymbol{\Sigma}|}} \exp[-\mathbf{1}^T \boldsymbol{\mu} - -\frac{1}{2}(\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\mu} - \boldsymbol{\mu}_0)] \prod_{jkl} \frac{[\mu_{jkl}]^{y_{jkl}-1}}{y_{jkl}!}$$

Therefore the likelihood function or the marginal distribution is

$$L(\xi, \sigma_1^2, \sigma_2^2, \sigma_3^2|\boldsymbol{Y}) = f(\boldsymbol{Y}|\boldsymbol{\xi}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\epsilon}} f(\boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\epsilon}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) d\boldsymbol{\alpha} d\boldsymbol{\beta} d\boldsymbol{\epsilon} \tag{9}$$

where the integral in (9) can be approximated by Gaussian-Hermite (GH) quadrature. The estimate of $\boldsymbol{\theta} = (\xi, \sigma_0^2, \sigma_1^2, \sigma_2^2)'$ is obtained by maximizing the log-likelihood after GH approximation. This procedure is done with `glmer()` in `lme4` package (Bates et al. (2012), version 1.1.7) with option "optimizer= 'bobyqa' "

6 Supplementary Material

The details of experimental data is summarized as below

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of rna sequencing data using r and bioconductor. *Nature protocols*, 8(9):1765–1786.

Andersen, C. L., Jensen, J. L., and Ørntoft, T. F. (2004). Normalization of real-time quantitative reverse transcription-pcr data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer research*, 64(15):5245–5250.

Baron, K. N., Schroeder, D. F., and Stasolla, C. (2012). Transcriptional response of abscisic acid (aba) metabolism and transport to cold and heat stress applied at the reproductive stage of development in arabidopsis thaliana. *Plant science*, 188:48–59.

Bates, D., Maechler, M., and Bolker, B. (2012). lme4: Linear mixed-effects models using s4 classes.

Bolker, B., Skaug, H., Magnusson, A., and Nielsen, A. (2012). Getting started with the glmmadmb package. *Available at glmmadmb. r-forge. r-project. org/glmmADMB. pdf*.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135.

Bournier, M., Tissot, N., Mari, S., Boucherez, J., Lacombe, E., Briat, J.-F., and Gaymard, F. (2013). Arabidopsis ferritin 1 (atfer1) gene regulation by the phosphate starvation response 1 (atphr1) transcription factor reveals a direct molecular link between iron and phosphate homeostasis. *Journal of Biological Chemistry*, 288(31):22670–22680.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):94.

Bustin, S. (2002). Quantification of mrna using real-time reverse transcription pcr (rt-pcr): trends and problems. *Journal of molecular endocrinology*, 29(1):23–39.

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K., and Scheible, W.-R. (2005). Genome-wide identification and testing of superior reference genes for transcript normalization in arabidopsis. *Plant physiology*, 139(1):5–17.

Dekkers, B. J., Willems, L., Bassel, G. W., van Bolderen-Veldkamp, R. M., Ligterink, W., Hilhorst, H. W., and Bentsink, L. (2012). Identification of reference genes for rt–qpcr expression analysis in arabidopsis and tomato seeds. *Plant and Cell Physiology*, 53(1):28–37.

Di, Y., Schafer, D. W., Cumbie, J. S., and Chang, J. H. (2011). The nbp negative binomial model for assessing differential gene expression from rna-seq. *Statistical Applications in Genetics and Molecular Biology*, 10(1):1–28.

Di, Y., Schafer, D. W., and Di, M. Y. (2014). Package 'nbpseq'. *Molecular Biology*, 10:1.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683.

Frericks, M. and Esser, C. (2008). A toolbox of novel murine house-keeping genes identified by meta-analysis of large scale gene expression profiles. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1779(12):830–837.

Gur-Dedeoglu, B., Konu, O., Bozkurt, B., Ergul, G., Seckin, S., and Yulug, I. G. (2009). Identification of endogenous reference genes for qrt-pcr analysis in normal matched breast tumor tissues. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 17(8):353–365.

Hansen, K. D., Irizarry, R. A., and Zhijin, W. (2012). Removing technical variability in rna-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216.

Hong, S. M., Bahn, S. C., Lyu, A., Jung, H. S., and Ahn, J. H. (2010). Identification and testing of superior reference genes for a starting pool of transcript normalization in arabidopsis. *Plant and cell physiology*, 51(10):1694–1706.

Hruz, T., Wyss, M., Docquier, M., Pfaffl, M. W., Masanetz, S., Borghi, L., Verbrugghe, P., Kalaydjieva, L., Bleuler, S., Laule, O., et al. (2011). Refgenes: identification of reliable and condition specific reference genes for rt-qpcr data normalization. *BMC genomics*, 12(1):156.

Huggett, J., Dheda, K., Bustin, S., and Zumla, A. (2005). Real-time rt-pcr normalisation; strategies and considerations. *Genes and immunity*, 6(4):279–284.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Liao, Y., Smyth, G. K., and Shi, W. (2013). The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research*, 41(10):e108–e108.

Line, S. R., Liu, X., de Souza, A. P., and Yu, F. (2013). Translational signatures and mrna levels are highly correlated in human stably expressed genes. *BMC genomics*, 14(1):268.

Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., Levens, D. L., Lee, T. I., and Young, R. A. (2012). Revisiting global gene expression analysis. *Cell*, 151(3):476–482.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, K. E., Olsson, N., Schlosser, J., Peng, F., and Lund, S. T. (2006). An optimized grapevine rna isolation procedure and statistical determination of reference genes for real-time rt-pcr during berry development. *BMC plant biology*, 6(1):27.

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nat Biotech*, 32(9):896–902.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). Gc-content normalization for rna-seq data. *BMC bioinformatics*, 12(1):480.

Robinson, M. D., Oshlack, A., et al. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25.

Sekhon, R. S., Lin, H., Childs, K. L., Hansey, C. N., Buell, C. R., de Leon, N., and Kaeppler, S. M. (2011). Genome-wide atlas of transcription during maize development. *The Plant Journal*, 66(4):553–563.

Stamova, B. S., Apperson, M., Walker, W. L., Tian, Y., Xu, H., Adamczy, P., Zhan, X., Liu, D.-Z., Ander, B. P., Liao, I. H., et al. (2009). Identification and validation of suitable endogenous reference genes for gene expression studies in human peripheral blood. *BMC medical genomics*, 2(1):49.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative rt-pcr data by geometric averaging of multiple internal control genes. *Genome biology*, 3(7):research0034.

Wang, L., Xie, W., Chen, Y., Tang, W., Yang, J., Ye, R., Liu, L., Lin, Y., Xu, C., Xiao, J., et al. (2010). A dynamic gene expression atlas covering the entire life cycle of rice. *The Plant Journal*, 61(5):752–766.

Wu, D., Hu, Y., Tong, S., Williams, B. R., Smyth, G. K., and Gantier, M. P. (2013). The use of mirna microarrays for the analysis of cancer samples with global mirna decrease. *RNA*, 19(7):876–888.