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.

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 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.

cited from the abstract of the paper: " Real-time PCR (qPCR) is the most accurate method of quantifying gene expression, provided that suitable endogenous controls are used to normalize the data"

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, experimental or environmental conditions, and experiments (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, Oshlack 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, Orlando, Sigova, Lin, Rahl, Burge, Levens, Lee, and Young, 2012, Wu, Hu, Tong, Williams, Smyth, and Gantier, 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 Fernandes, Mommens, Hagen, Babiak, and Solberg (2008) and Hruz, Wyss, Docquier, Pfaffl, Masanetz, Borghi, Verbrugghe, Kalaydjieva, Bleuler, Laule et al. (2011), universally stably expressed genes may not exist. Hruz et al. (2011) showed that a subset of stably expressed genes from a specific biological context may have more variability than other genes if examined across a broader range of samples and conditions. Many studies have shown that stably expressed genes are subject to change from one experiment to another due to different experimental protocols, different tissue types, or other varying conditions (Reid, Olsson, Schlosser, Peng, and Lund, 2006, Hong, Bahn, Lyu, Jung, and Ahn, 2010). The top 100 stably expressed genes in Arabidopsis developmental series of Czechowski, Stitt, Altmann, Udvardi, and Scheible (2005) shared only 3 genes with top 50 stably expressed genes identified from Arabidopsis seed samples by Dekkers, Willems, Bassel, van Bolderen-Veldkamp, Ligterink, Hilhorst, and Bentsink (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.

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, Jensen, and Ørntoft, 2004), (REF Fernandes et al. 2008). 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, Dheda, Bustin, and Zumla (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, Ngai, Speed, and Dudoit (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, Apperson, Walker, Tian, Xu, Adamczy, Zhan, Liu, Ander, Liao et al. (2009), Dekkers et al. (2012), Gur-Dedeoglu, Konu, Bozkurt, Ergul, Seckin, and Yulug (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. Andersen et al. (2004) uses a linear mixed model to estimate the between-group and within-group variances from expression profiles of microarray data, and then combines the two variance components by a Bayesian formulation. For a future experiment, the stability value for a given gene is defined as the absolute value of posterior mean plus one standard prediction deviation of that gene. 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 housekeeping 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, McCulloch and Neuhaus 2001) 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

Overview of the RNA-Seq data sets

We examined RNA-Seq data from 49 Arabidopsis experiments stored on the NCBI GEO repository (see more details below). After screening, we retained data from 209 biological samples in 23 experiments. For illustrating our methods for finding stably expressed genes, we divided the experiments into three groups: *the seedling group* contains 70 Arabidopsis seedling samples from 10 experiments; *the leaf group* contains 60 Arabidopsis leaf samples from 5 experiments; the *multi-tissue group* contains 79 samples from 8 experiments on multiple tissue types (shoot apical, root tip, primary root, inflorescences and siliques, hypocotyl, flower, carpels, aerial tissue). Table 1 summarizes the basic information about the three groups.

For finding stably expressed genes in each group, we processed the raw sequencing data and summarized the results as count matrices of mapped RNA-Seq short reads (see details below). We removed genes with low mean numbers (less than 3) of mapped read counts. Such genes tend to be more prone to sequencing

Table 1: Summary statistics for three groups.

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

noise, less interesting to biologists, and also cause convergence issues when fitting statistical models. Many other researchers (such as Anders, McCarthy, Chen, Okoniewski, Smyth, Huber, and Robinson 2013) recommend removing such genes before analyzing RNA-Seq data. The number of remaining genes in each group is also summarized in Table 1.

Details of the data processing steps

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 (Sequence Read Archive 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 Sequence Read Archive (SRA) files to FASTQ files using the NCBI SRA Toolkit (Leinonen, Sugawara, and Shumway (2010), version 2.3.5-2).
2. Align short reads in FASTQ files, using the `align()` function from Subread aligner (RSubread, version 1.16.2, Liao, Smyth, and Shi 2013) in the software R (R Core Team, 2015), to the Arabidopsis reference genome

`Arabidopsis_thaliana.TAIR10.22.dna.toplevel.fa`
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
`Arabidopsis_thaliana.TAIR10.22.gtf`
is downloaded from Ensembl plat FTP server. The multi-mapping or multi-overlapping (???) reads were not counted.

Subread aligner is a recently developed sequence mapping tool that adopts a seed-and-vote paradigm to map the RNA-Seq short reads to the genome. It breaks each short read into a series of overlapping segments called subreads and uses the subreads to vote on the optimal genome location of the original read. The subreads are shorter and can be mapped to the genome much faster. Compared to other alighers such as Bowite 2 (Langmead and Salzberg, 2012) or BWA (Li and Durbin, 2009), Subread aligner is both faster and more accurate (Liao et al., 2013). We compared results from the above pipeline to results from a pipeline described in Anders et al. (2013) over several RNA-Seq experiment data, and Rsubread was more than three times faster and successfully mapped more reads to the reference genome. For people familiar with R, it also has the advantage that it is completely implemented in R.

We divided the experiments into three groups as summarized in Table 1. As an additional data quality control measure, within each group, we computed an initial set of normalization factors from all samples combined using method described in Section 2.2. An experiment is retained only when the normalization factors of all samples in the experiment are between 0.70 and 1.30. If the initial estimated normalization factor is too different from 1 for a sample, it often indicates that the read counts distribution in the corresponding sample is markedly different from the distributions of the rest of the samples. Such samples demand additional attention before being incorporated in studies that we intend to do.

2.2 Count normalization

As explained in the introduction, for analyzing RNA-Seq data, count normalization is needed for 1) adjusting differences in sequencing depths or library sizes; 2) for adjusting the apparent changes in relative read frequencies of non-DE genes simply to accommodate changes in relative read frequencies of truly DE genes.

For the second type of adjustment, we follow Anders and Huber's method (Anders and Huber, 2010) for estimating normalization factors. Let y_{ij} denote the

read count for ith gene of jth sample (m genes and n samples in total). We first create a pseudo-reference sample where each gene's expression value is this gene's geometric mean expression over all real samples,

$$y_{i,0} = (\prod_{j=1}^{n} y_{i,j})^{1/n}, i = 1, \ldots, m. \tag{1}$$

Next we calculate the median fold-change in relative frequency between each sample j and the pseudo-reference sample,

$$R'_j = \text{median}\left(\frac{y_{1,j}/N_j}{y_{1,0}/N_0}, \ldots, \frac{y_{m,j}/N_j}{y_{m,0}/N_0}\right), \tag{2}$$

where N_j is the library size for sample j (the sum of RNA-Seq counts mapped to all genes retained in each sample). Last, The *normalization factor* R_j for sample j is then calculated as

$$R_j = \frac{R'_j}{(\prod_{j=1}^{n} R'_j)^{1/n}}. \tag{3}$$

With the estimated normalization factors, the relative frequencies will be computed as $y_{ij}/N_j R_j$. The assumption made here is that the median fold change between relative frequencies in two samples should be 1. In other words, this normalization method assumes that the majority of genes are not DE. The NBPSeq package (Di, Schafer, and Di, 2014) has an inbuilt function for this procedure and it will be used for count normalization in this paper.

The TMM method of Robinson et al. (2010) is based on a similar principle: assuming the majority of the genes are not DE. The TMM method can be applied to a subset of genes selected based on an initial screening of mean expression level and fold changes.

We want to apply Anders and Huber's method to a subset of stably expressed genes based on our stability measure (i.e., taking the median from a subset of genes in equation (2)). But to identify the stably expressed genes, we first need a set of initially estimated normalization factors. So tackle this circular dependence, we use a one-step iteration method to estimate the normalization factors:

1. First, we use all the genes to calculate the initial normalization factors;
2. Then, we fit a GLMM to each gene and estimate the total variance measure, incorporating the initial normalization factors as an offset term (see Section 2.3);
3. Next, we select the top 1000 stably expressed genes based on the total variance measure estimated from step 2 above, and use them as reference genes to recalculate the normalization factors.

In practice, this one-step method seems to be adequate and further iterations will only slightly change the set of 1000 stably expressed genes. For example, for multi-tissue group of experiments, if we were to run one more iteration of steps 2 and 3, there would be 956 overlapping genes between the top 1000 genes from the first iteration and those from the second iteration.

2.3 Poisson log-linear mixed-effects regression model and the total variance measure of expression stability

We fit a Poisson log-linear mixed-effects regression model to the RNA-Seq counts mapped to each gene and measure gene expression stability using a total variance measure.

Let Y_{ijkl} be the number of RNA-Seq reads mapped to gene i in sample j from treatment group k in experiment l. We will fit regression models to each gene separately and suppress subscript i from the model equations. For each gene, we fit a Poisson log-linear mixed-effects regression model

$$Y_{jkl} \sim \text{Poisson}(\mu_{jkl}), \tag{4}$$

$$\log(\mu_{jkl}) = \log(R_{jkl}N_{jkl}) + \xi + \alpha_l + \beta_{k(l)} + \varepsilon_{jkl}, \tag{5}$$

which is a specific type of generalized linear mixed model (GLMM, McCulloch and Neuhaus (2001)). In equation (5), N_{jkl} and R_{jkl} are the library size and normalization factor discussed in Section 2.2. We will call $R_{jkl}N_{jkl}$ the *normalized library size*. ξ is a fixed-effect term for the baseline log mean of the *relative counts* (counts divided by the normalized library sizes). α, β, and ε represent the experiment effect, the treatment effect (nested within each experiment), and the sample effect respectively. We view α, β and ε as random effects and assume that they are independent and follow normal distributions:

$$\alpha_l \sim N(0, \sigma^2_{\text{experiment}}), \;\; \beta_{k(l)} \sim N(0, \sigma^2_{\text{treatment}}), \;\; \varepsilon_{jkl} \sim N(0, \sigma^2_{\text{sample}}), \tag{6}$$

where $\sigma^2_{\text{experiment}}$, $\sigma^2_{\text{treatment}}$ and σ^2_{sample} are called *variance-components*—they quantify the overall variances of the corresponding random effect terms.

(Sarah: Should effect be singular or plural?) The sample effect ε represents the extra-Poisson variation in read counts among samples in the same treatment group and σ^2_{sample} plays a similar role as the *over-dispersion* parameter in a negative binomial model (Anders and Huber 2010, Di, Schafer, Cumbie, and Chang 2011). The experiment effect, α, accounts for all sources of variation at the experiment level, including differences in lab personnel and conditions, day light hours,

age of the plants, temperature, sequencing platform, and other unidentified sources. The contributions from these different experiment-level sources are often difficult to separate statistically. We treat the experiment effect α as a random effect because we view the collected experiments as a random sample from the pool of all Arabidopsis RNA-Seq experiments. We also treat the treatment effect β as a random effect. In a DE test, β is usually considered as a fixed-effect term. Here for evaluation of expression stability, we are not interested in the specific levels of the individual β's and focus more on the overall variation of β under a range of experimental conditions.

We define the stability measure as the estimated *total variance*,

$$\hat{\sigma}^2 = \hat{\sigma}^2_{\text{sample}} + \hat{\sigma}^2_{\text{treatment}} + \hat{\sigma}^2_{\text{experiment}}. \tag{7}$$

The parameters $(\xi, \sigma^2_{\text{experiment}}, \sigma^2_{\text{treatment}}, \sigma^2_{\text{sample}})$ are estimated from `glmer()` function of R package `lme4` (Bates, Maechler, and Bolker (2012), version 1.1.7), which uses a Gaussian-Hermite quadrature to approximate the likelihood function. We rank all the genes according to their $\hat{\sigma}^2$s, and consider highly ranked (e.g., top 1000) genes to be stably expressed.

Normal models (equation (6)) are commonly assumed for the random effects in the GLMM settings. The normality assumption is likely a simplification of reality, yet it is a good starting point and should be adequate for finding genes with low total variation—the stably expressed ones.

2.4 Other stability measures

The assessment of gene expression stability depends on the specific stability measure used. Czechowski et al. (2005) and Dekkers et al. (2012) used the coefficient of variation (CV) measure, computed as *standard deviation/mean*, for finding stably expressed genes from micorarray data.

The *M-value* in geNorm (Vandesompele, De Preter, Pattyn, Poppe, Van Roy, De Paepe, and Speleman, 2002) is a well-cited measure. For a set of m_0 genes, the M-value measure works as follows: first, it calculates the *relative variation* of gene i_1 to gene i_2 by the standard deviation of their log fold changes across all the n samples;

$$V_{i_1,i_2} = st.dev\left\{\log\left(\frac{y_{1,i_1}}{y_{1,i_2}}\right), \ldots, \log\left(\frac{y_{n,i_1}}{y_{n,i_2}}\right)\right\}$$

next, the M-value for gene i_1 is obtained by taking the average of $m_0 - 1$ relative variations, one for each pair of gene i_1 and the remaining $m_0 - 1$ genes;

$$M_{i_1} = \frac{\sum_{k \neq i_1} V_{i_1,k}}{m_0 - 1}$$

In this way, each of the m_0 genes is assigned an M value to represent their corresponding expression stability.

In the Results section, we compared the M-value to the total variance measure on RNA-Seq data from the multi-tissue group experiments, and compared the stably expressed genes identified from these two measures to those identified from microarray data using the CV measure.

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 quantify 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 (5) 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). 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

$$\frac{\text{count} \times 10^6}{\text{normalized library size}} \tag{8}$$

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

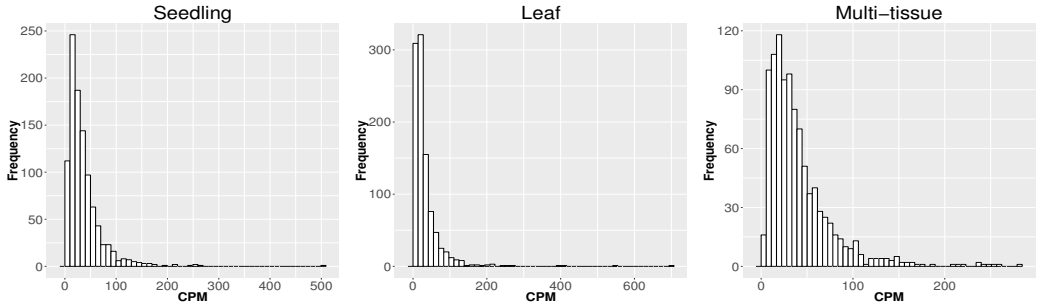

Figure 1: mean CPM (see equation (8)) for the top 1000 most stably expressed genes identified by the total variance $\hat{\sigma}^2$ from the seedling (left), leaf (middle) and multi-tissue (right) groups . The mean CPM is computed over all samples for each gene.

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 stability 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, Tissot, Mari, Boucherez, Lacombe, Briat, and Gaymard (2013), Baron, Schroeder, and Stasolla (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 estimated variance components. Column 6 specifies the ranking 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

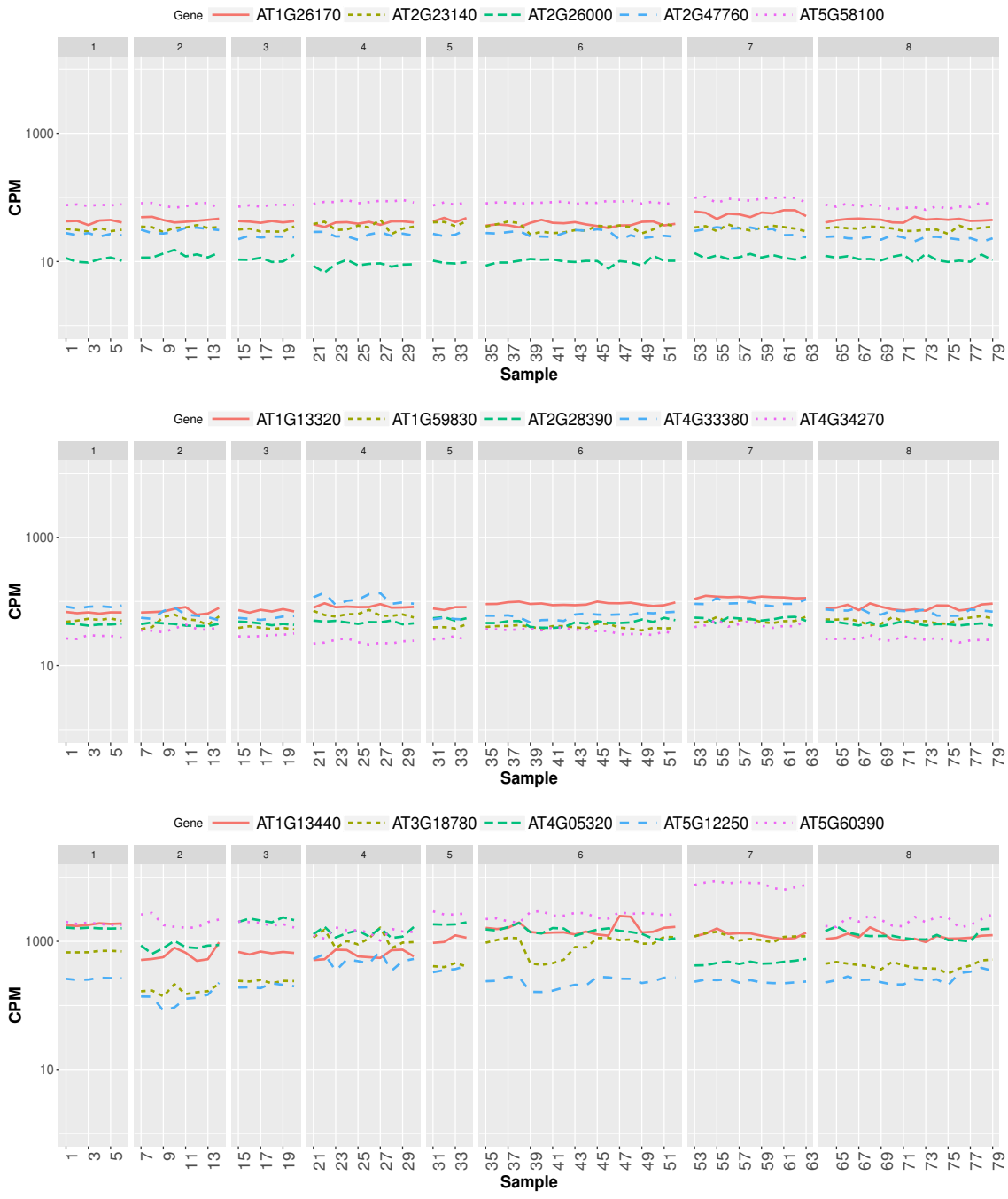

Figure 2: CPM of 15 genes for each sample in the multi-tissue group (described in Section 2.1) across 79 samples: five stably expressed genes identified by the total variance $\hat{\sigma}^2$ (top), five stably expressed identified by Czechowski according to the CV measure from a developmental series of microarray experiments (middle), and five traditional house-keeping genes (bottom).

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 (5)) 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 $(\text{count}+1)$) 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 (see Section 2.4);
- 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 group. The recall percentage for L_i is defined as

$$\frac{\#\{L_i \cap \text{reference set}\}}{\#\{L_i\}} \times 100, \tag{9}$$

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.

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

$$\frac{y_{i_1,j_1}}{y_{i_2,j_1}} = \frac{y_{i_1,j_2}}{y_{i_2,j_2}}. \tag{10}$$

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,

$$\frac{y_{i_1,j_1}}{S_{j_1}} = \frac{y_{i_1,j_2}}{S_{j_2}}, \tag{11}$$

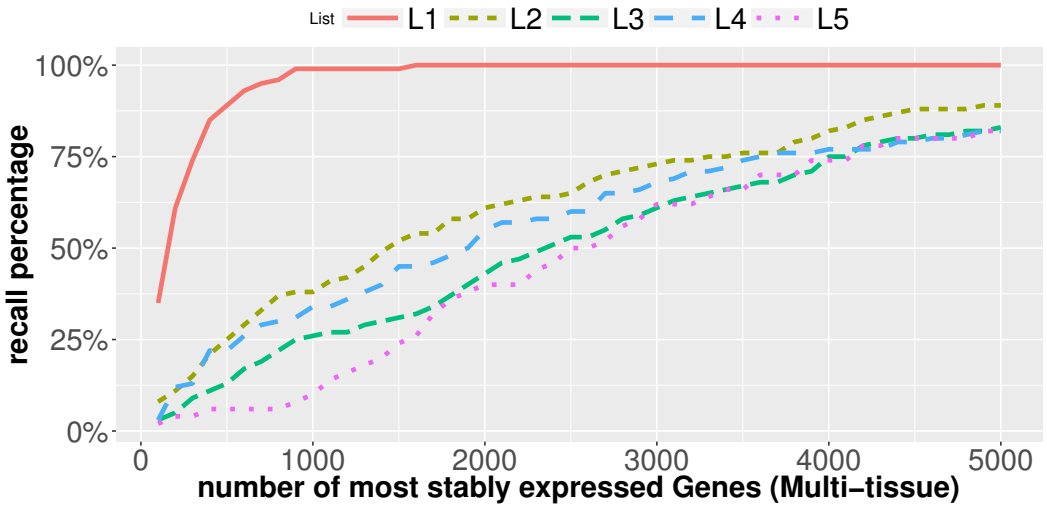

Figure 3: Recall percentage for list L_1 — L_5 (see Section 3.3). We choose top 100 stably expresed genes for L_1–L_4, and top 50 stably expressed genes for L_5. x-axis is the number of most stably expressed genes in multi-tissue group according to the total variance measure, and y-axis shows the recall percentage (see equation (9)) for each of the five lists.

where S_{j_1} to S_{j_2} are the normalized library sizes (i.e., $R_j N_j$ in equation (5)). This implies for any two stably-expressed genes i_1 and i_2

$$\frac{y_{i_1,j_1}}{y_{i_1,j_2}} = \frac{y_{i_2,j_1}}{y_{i_2,j_2}} = \frac{S_{j_1}}{S_{j_2}}. \tag{12}$$

The first equation in (12) is equivalent to equation (10). (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 2 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 geNorm 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 (5) 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 from the multi-tissue group for 20 genes randomly selected from the top 1000 stably expressed genes and 20 genes randomly selected from 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%

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. This new approach allows investigators to prescribe a specific biological context for evaluating

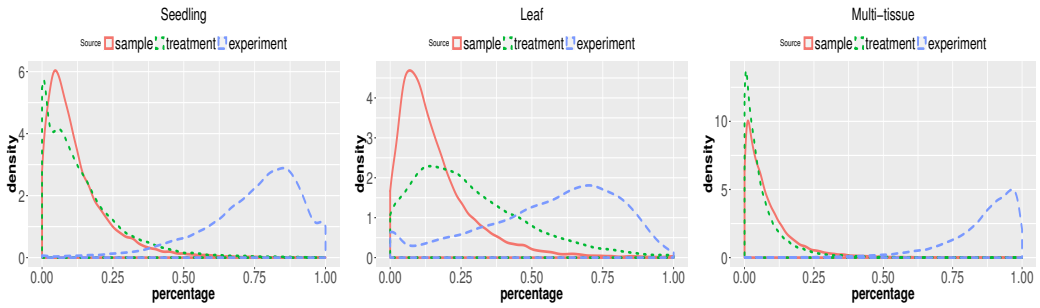

Figure 4: Density plot of percentages of variance components for the seedling, the leaf, and the multi-tissue groups.

Table 5: A toy example—using a common explicit set of reference genes to compare among multiple experiments.

	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

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 contrast, existing normalization approaches are often applied to the single data set under study. A 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. he 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.

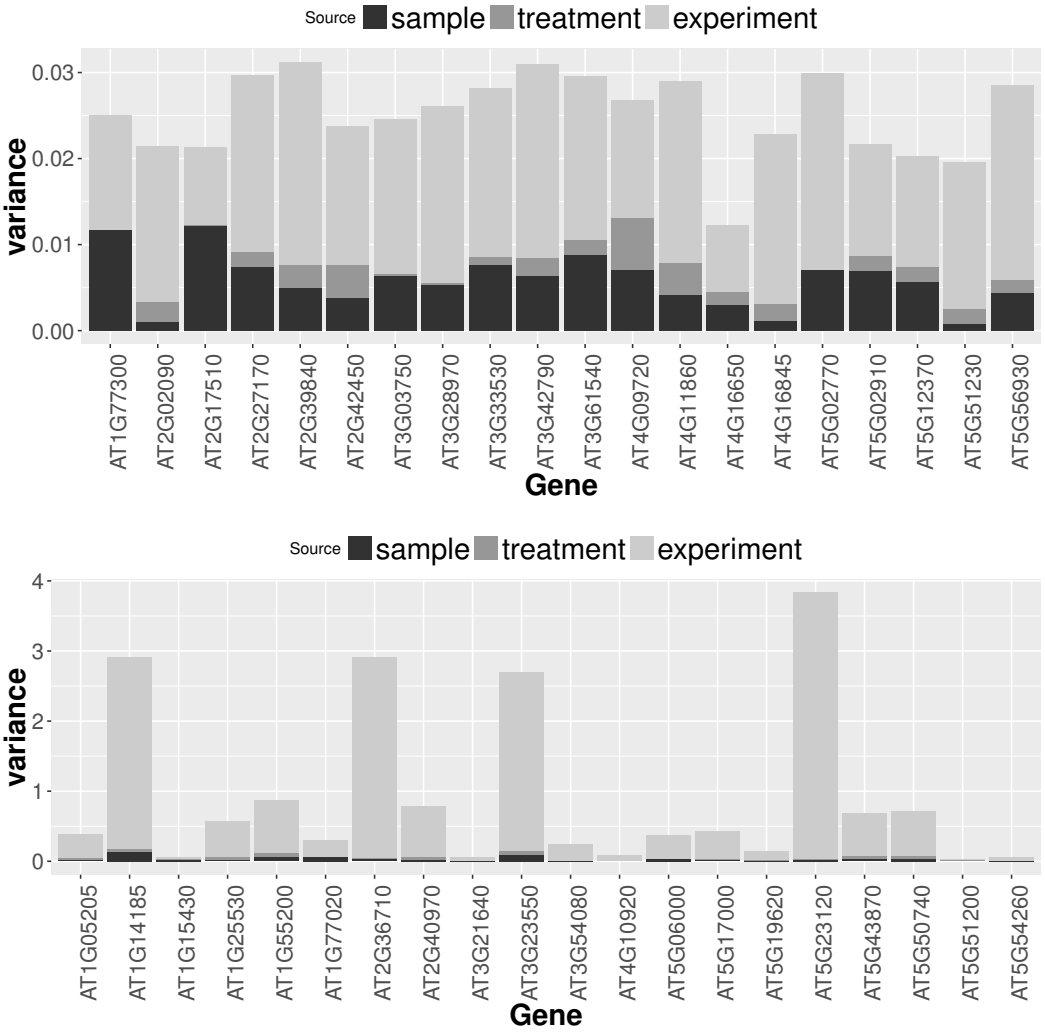

Figure 5: stacked bar plot of the three variance components for the multi-tissue group. Top: 20 genes randomly selected from top 1000 stably expressed genes; Bottom: 20 genes randomly selected from all the genes.

In the Introduction, we also argued that using an explicit set of genes as reference for normalization can improve interpretability of DE results, in the sense that DE can always be interpreted as relative to the explicit reference gene set used. This new perspective is especially relevant when one wants to compare results from two or more 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. 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 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 expressed 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) reference genes: intuitively, using too few, the estimates will be unstable; using too many, the results may be subject to influence from highly unstably 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 (2)) 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 Section 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.

4 Conclusion and Discussion

In this paper, we advocate quantifying gene expression stability by applying a numerical stability measure to a large number of existing RNA-Seq data sets. Similar

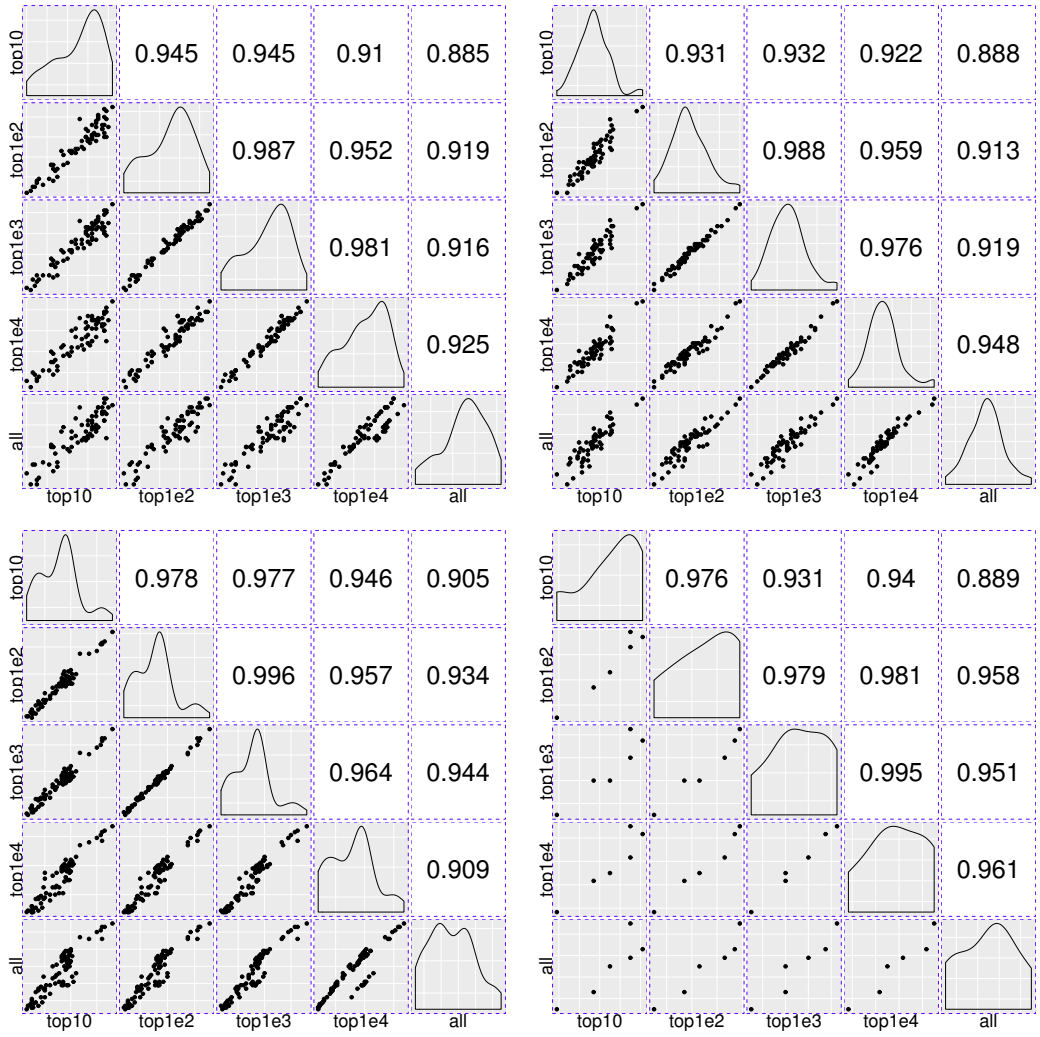

Figure 6: Matrix plot of normalization factors. Top 10, 100, 1000, and 10,000 stably expressed genes are chosen as reference to calculate the normalization factors for each of the three groups: the seedling (upper left), the leaf (upper right), and the multi-tissue (lower left) groups. The lower right plot shows the estimated normalization factors for a new seedling experiment (GSE66666, with sample size 6) when top 10—10,000 stably expressed genes from the seedling group are used as reference. The normalization factors are estimated by Anders and Huber's method.

strategy has also been used by others for finding stably expressed genes from microarray data. Since DE is measured by relative frequencies, we argue that DE is a

relative concept and using an explicit reference gene set can improve interpretability of DE results, and furthermore, when comparing multiple experiments using a common reference gene set can avoid inconsistent conclusions (see section REF).

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.

It should be clear but worth emphasizing that when using a numerical measure to identify stably expressed genes, the outcome depends on multiple factors: the reference sample sets used, the technology used for measuring gene expression, and the specific numerical stability measure used. In this study, for illustration of methods, we identified three sets of stably expressed genes from three sets of Arabidopsis experiments. The major point is that stably genes identified from different background will provide different biological contexts for evaluating different expression. In practice, researchers can choose the specific context. A practical challenge in applying such a philosophy is that no two experiments will have identical background, researchers have to decide what experiments can be considered comparable. This is a difficult question, however, we believe it has to be asked from now on: biologists perform comparative experiments with the intension that the conclusions from a single experiment will be generalizable beyond the context of a single lab. If we don't understand comparability between different experiments, such generalization is impossible. Defining and characterizing comparability is a challenging topic we would like to investigate more in future.

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). In this paper, we used a one-step iteration strategy to reduce the dependence on the initially estimated normalization factors. In future, we intend to look at the genes through evolutionary genetics methods: for example, and examine their etc. However, there is no guarantee such methods will work since a gene is well conserved does not necessarily mean it has stable gene expression.

The R codes for reproducing results in this paper are available at Github:

5 Supplementary Material

The details of experimental data is summarized as below

References

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

Anders, S., D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber, and M. D. Robinson (2013): "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor," *Nat. Protoc.*, 8, 1765–1786.

Andersen, C. L., J. L. Jensen, and T. F. Ørntoft (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 Res.*, 64, 5245–5250.

Baron, K. N., D. F. Schroeder, and C. Stasolla (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 Sci.*, 188, 48–59.

Bates, D., M. Maechler, and B. Bolker (2012): "lme4: Linear mixed-effects models using S4 classes," .

Bournier, M., N. Tissot, S. Mari, J. Boucherez, E. Lacombe, J.-F. Briat, and F. Gaymard (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," *J. Biol Chem.*, 288, 22670–22680.

Bustin, S. (2002): "Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems," *J. Mol. Endocrinol.*, 29, 23–39.

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

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

Di, Y., D. W. Schafer, J. S. Cumbie, and J. H. Chang (2011): "The NBP negative binomial model for assessing differential gene expression from RNA-Seq," *Stat. Appl. Genet. Mol. Biol.*, 10, 1–28.

Di, Y., D. W. Schafer, and M. Y. Di (2014): "Package 'NBPSeq'," *Mol. Biol.*, 10, 1.

Fernandes, J. M., M. Mommens, Ø. Hagen, I. Babiak, and C. Solberg (2008): "Selection of suitable reference genes for real-time PCR studies of Atlantic halibut development," *Comp. Biochem. Phys. B*, 150, 23–32.

Frericks, M. and C. Esser (2008): "A toolbox of novel murine house-keeping genes identified by meta-analysis of large scale gene expression profiles," *BBA-Gene Regul. Mech.*, 1779, 830–837.

Gur-Dedeoglu, B., O. Konu, B. Bozkurt, G. Ergul, S. Seckin, and I. G. Yulug (2009): "Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues," *Oncol. Res.*, 17, 353–365.

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

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

Huggett, J., K. Dheda, S. Bustin, and A. Zumla (2005): "Real-time RT-PCR normalisation; strategies and considerations," *Genes Immun.*, 6, 279–284.

Langmead, B. and S. L. Salzberg (2012): "Fast gapped-read alignment with Bowtie 2," *Nat. Methods*, 9, 357–359.

Leinonen, R., H. Sugawara, and M. Shumway (2010): "The sequence read archive," *Nuc. Acids Res.*, gkq1019.

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

Liao, Y., G. K. Smyth, and W. Shi (2013): "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote," *Nuc. Acids Res.*, 41, e108–e108.

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

McCulloch, C. E. and J. M. Neuhaus (2001): *Generalized linear mixed models*, Wiley Online Library.

R Core Team (2015): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`.

Reid, K. E., N. Olsson, J. Schlosser, F. Peng, and S. T. Lund (2006): "An optimized grapevine RNA isolation procedure and statistical determination of reference genes for real-time RT-PCR during berry development," *BMC Plant Biol.*, 6, 27.

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

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

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

peripheral blood," *BMC Med. Genom.*, 2, 49.

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

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