Using Mixtures of Biological Samples as Genome-Scale Process Controls

Jerod Parsons[1], Sarah Munro[1], P. Scott Pine[1], Jennifer McDaniel[3], Michele Mehaffey[2], Marc Salit[1]

Institutions

1- Biochemical Sciences Division, National Institute of Standards & Technology, 100 Bureau Drive, Gaithersburg, MD, 20899

2 – *Leidos Biomedical Research Inc. P.O. Box B Bldg 428, Frederick, MD, 21702*

Corresponding Author: Jerod Parsons (jerod.parsons@nist.gov)

Abstract:

Background:

Genome-scale "-omics" measurements are challenging to benchmark due to the enormous variety of unique biological molecules involved. Mixtures of previously-characterized samples can be used to benchmark repeatability and reproducibility using component proportions as truth for the measurement. We describe and evaluate experiments characterizing the performance of RNA-sequencing (RNA-Seq) measurements.

Results:

The parameters of a model fit to a measured -omic profile can be evaluated to assess bias and variability of the genome-scale measurement of a mixture. A linear model describes the behavior of expression measures of mixtures and provides a context for performance benchmarking. Residuals from fitting the model to experimental data can be used as a metric for evaluating the effect an individual step in an experimental process has on the linear response function and precision of the underlying measurement while identifying signals affected by interference from other sources. Effective benchmarking requires well-defined mixtures, which for RNA-Seq requires knowledge of the messenger RNA (mRNA) content of the individual components. We demonstrate and evaluate an experimental method suitable for use in genome-scale process control and lay out a method utilizing spike-in controls to determine mRNA content.

Conclusions:

Genome-scale process controls can be derived from mixtures. These controls relate prior knowledge of individual components to a complex mixture, allowing assessment of measurement performance. The mRNA fraction accounts for differential enrichment of mRNA from varying total RNA samples. Spike-in controls can be utilized to measure this relationship between mRNA content and input total RNA. Analysis of mixtures can also be employed to determine the composition and proportions of an unknown sample, even when component-specific markers are not previously known, so long as pure components can be measured alongside the mixture.

Keywords:

RNA sequencing, Gene expression, mixture deconvolution, expression deconvolution, process control, spike-in control, ERCC

Background:

Measurement assurance for genome-scale measurements is challenged by the impracticality of creating a sample containing known quantities of tens of thousands of components, such as the RNA transcripts measured in an RNA-seq experiment. Deep sequencing of cellular RNA can generate vast quantities of gene expression information, yet measurement biases have been identified at nearly every step of the library preparation process [1-4].

As RNA-sequencing expression data expands from discovery into clinical applications, the sources and magnitudes of bias and variability must be carefully understood and quantified. Even the basic units of expression in sequencing, such as transcripts per million (TPM) or fragments per kilobase per million reads (FPKM), are undergoing revision [5,6]. Even when using comparable units, it is rarely possible to directly compare gene expression values reported by different labs, on different instruments, or frequently just on different days [6-8], unless special care is taken to use uniform samples and protocols. Identifying the presence and variation of biases in a measurement process over time requires a standard to be used for process control.

Ideally, a measurement process is linear and possesses a known precision. A linear measurement process shows an increase in signal proportional to an increase in the object being measured. It is also helpful if measured signal arises only from the single source, and not from interference from non-targets. Precision consists of repeatability and reproducibility, defined as the degree of closeness in multiple measurements made by a single user and the closeness between multiple labs, respectively. We show that mixtures can demonstrate that a measurement's response function is linear and of high specificity (free of interference) while measuring its variability and precision. Properly constructed mixture samples can be used to correct for systematic measurement errors, provide ongoing monitoring of performance, serve as a tool for interlaboratory comparison, and create a context for evaluating batch effects.

Two approaches to creating useful genome-scale standards include the creation of a limited number of external spike-in controls, such as the External RNA Control Consortium (ERCC) controls, which were created for microarrays and have been applied to next-gen sequencing [9-11]. A second approach utilizes mixtures of previously characterized samples as well as prior knowledge of the ratio of the mixtures, and has also been applied to microarrays [12-14] but has not been utilized in other genome-scale measurements. Using standards in these measurements provides confidence in the ability of a test to detect both positive and negative results, including the limits of that detection.

Mixtures can serve as a test that applies to each of the tens of thousands of transcripts in a mixture's -omics profile. Linearity of the measurement response can be demonstrated using prior knowledge of the ratio of previously characterized mixtures, based on the fundamental understanding that a mixture is a linear combination of its components. Previous work with mixtures in microarrays[12-14] utilized an arbitrary 10-fold "selectivity" cutoff to evaluate the linear dynamic range of microarray measurements and understand the variability of these measurements. The arbitrary selectivity cutoff in previous work prevents the identification of interference, as any genes affected by interference would be filtered by the stringent selectivity cutoff.

Using known mixture compositions, predicted values can be calculated based on the assumption that the measurement response is linear. Deviation of the observed values from the model-predicted value is an indication of bias in the measurement. Systematic biases could be introduced by sample preparation, signal processing, interference from other genes, or sampling variation. Signal arising from off-target molecules, such as a closely related transcript, can cause false positive results and result in a lowered specificity. Mixture samples can provide information about the measurement sensitivity, specificity, repeatability, reproducibility, dynamic range, and limit of detection.

Determining the relative contributions to gene expression of individual components within mixtures of biological states has received some attention in the clinical realm, where biopsies and other patient samples are often mixtures containing desired and undesired components. The process of resolving gene expression signals introduced by each individual component of a mixture [13-23] has been used to account for tumor heterogeneity and to separate whole blood samples into individual cell types. These procedures often separate mixture components based on a subset of "signature" genes that vary uniquely between components. These 'deconvolution' methods have been used [27-30] to develop high-resolution tumor expression signatures from imperfect biological samples [31,32] and differentiate between cell-type-frequency changes and per-cell gene expression changes [33,17]. Many of these methods can determine mixture component types by using a linear model where mixture expression is treated as a combination of expression signatures.

One parameter notably absent from these methods is RNA content. Different cell types express different total amounts of RNA per cell, confounding estimates of cell type proportion made based on the quantification of total RNA [24]. Others have introduced the concept of a biological scaling factor [25,26] to compensate for variation in the RNA content of cells, including the use of spike-in controls to calculate this factor. The filtering of mRNA out from total RNA adds a bias to the experiment due to the different abundance of mRNA between cell types.

We aimed to demonstrate the assessment of linear response, specificity, and accuracy of genome-scale measurements using mixtures. In the process, we demonstrate that linear models can be used to separate these mixtures into the proper component signatures. We were mindful that while our mixtures were of total RNA, the measurement filters for mRNA, and that the relationship between these two values is an important factor when interpreting results. We anticipate that a mixture-based approach to measurement assurance is highly generalizable to many types of mixtures and can be extended to the wide variety of genome-scale measurements, including but not limited to proteomic, metabolomic, and transcriptomic systems.

Results:

To assess measurement parameters of genome-scale transcriptome data, we analyzed two RNA-seq experiments measuring synthetic mixtures of commercially available human total RNA samples (Figure 1)[13,14,34]. One experiment included a mixture of two reference total RNA samples, sequenced by 9 labs as a part of the Sequencing Quality Control Consortium (SEQC) [34-35]. This study sequenced “Universal Human Reference RNA” (SEQC-A), “Human Brain Reference RNA” (SEQC-B) and two mixtures of the above (SEQC-C and SEQC-D) with mixture compositions C=3A+1B and D=1A+3B. These four samples were sequenced by 9 labs using either Illumina or Life Technologies sequencing instruments.

The second sample, called BLM, contains two mixtures (BLM-1 and BLM-2) composed of total RNA isolated from human brain (the same RNA as SEQC-B), liver, and muscle tissue. These two mixtures were made with component proportions of 1B:1L:2M and 1B:2L:1M. The total RNA of each individual tissue was also sequenced “neat” to provide an expression signature for each tissue.

These mixtures were designed to have a defined expression signal ratio between them. If the measurement response were linear and unbiased, the signal in the SEQC-C sample would be exactly 1/4 the signal of SEQC-B plus 3/4 the signal from SEQC-A due to the design of the mixture. However, these total RNA mixtures went through RNA-seq library preparation, which purposely filters ribosomal RNA out of the pool. The resulting sequence data reflects this filtration, which can be different between samples. A correction for this differential filtration and an upper-quartile normalization [36] must be applied to accurately reflect the experimental process and allow the model to return the designed ratios of expression between mixtures (Supp.Figure 1).

ERCC spike-in control RNAs were added to the components of these mixtures. ERCC controls [12] were designed as exogenous RNA sequences to function as control RNA. Two spike-in control pools were designed with ratiometric differences in the concentration of individual ERCC spike-ins. As expected based on the mixture designs, ERCCs spiked-in equally yielded equal expression signal, while signal from ERCCs spiked differentially into multiple subpools was at ratios corresponding to the designed fold changes. Poisson sampling at the lower expression levels results in increased dispersion about the expected ratio [49].

Linear model-based analysis of genome-scale gene expression

We posit that mixture expression is a linear combination of the component samples and the mixture proportions of each component. Equation 1 describes the relationship between signal in the mixtures and signal in the constituent samples. A mixture *M* – 2 per dataset in this study - is composed of a number of named components “*C*” (“B”,”L”, and ”M” in the Brain/Liver/Muscle mixture or “A” and “B” in the SEQC dataset), with each component comprising a proportion of the mixture Φ_{C}. $\chi_{i,M}$ is the expression signal arising from a particular gene/transcript *i* in mixture *M*.

Equation 1: $\chi_{i,M}=\sum_{C=1}^{N} \chi_{i,C}\times\Phi_{C,M}$

This study uses four mixtures of the same general form:

$$\chi_{i,BLM1}=\chi_{i,B}\times\Phi_{B,1}+\chi_{i,L}\times\Phi_{L,1}+\chi_{i,M}\times\Phi_{M,1}$$

$$\chi_{i,BLM2}=\chi_{i,B}\times\Phi_{B,2}+\chi_{i,L}\times\Phi_{L,2}+\chi_{i,M}\times\Phi_{M,2}$$

$$\chi_{i,SEQC\text{-}C}=\chi_{i,SEQC\text{-}A}\times\Phi_{A,C}+\chi_{i,SEQC\text{-}B}\times\Phi_{B,C}$$

$$\chi_{i,SEQC\text{-}D}=\chi_{i,SEQC\text{-}A}\times\Phi_{A,D}+\chi_{i,SEQC\text{-}B}\times\Phi_{B,D}$$

These mixtures were made from total RNA, while the expression signal (sequencing reads) arises only from the mRNA. As the fraction of the total RNA mass that is mRNA varies between cells, the filtering of total RNA into mRNA introduces a bias. Supplemental Figure 1 shows the offset from the expected ratios of tissue-specific and ERCC RNA caused by this bias. We correct the specific equations for the mRNA fraction by multiplying each component by a factor *ρ*. This factor corresponds to the measured mRNA compared to the mass of total RNA in each mixture. ρ_C is defined as the amount of measured RNA per unit total RNA in component *C*.

After adding this factor, the BLM1 mixture equation becomes

$$\chi_{i,BLM1}=\chi_{i,B}\times\Phi_{B,1}\times\rho_{B}+\chi_{i,L}\times\Phi_{L,1}\times\rho_{L}+\chi_{i,M}\times\Phi_{M,1}\times\rho_{M}$$

There are a few approaches that have been described to measure *ρ*. One study directly measured the mRNA content between SEQC-A and SEQC-B samples [36] using qRT-PCR. Another described the use of trimmed mean of log expression ratios (TMM)[25] to measure mRNA content from RNA-seq data. TMM-derived factors have been shown to be an appropriate measure in cases where there is no global expression level change (such as the SEQC mixtures), but introduce bias if there are global expression changes (such as in the BLM mixtures)[26]. Supp.Figure 2 demonstrates this.

The ρ factor can be determined using spiked-in RNA[26] as sample reads per microgram of total RNA divided by spike-in reads per microgram of spike-in RNA. This calculation emphasizes that the mRNA fraction is a correction for the differential enrichment between polyadenylated spike-in RNA and total RNA, which is only partly composed of mRNA.

The mRNA fraction ρ is a property of an individual RNA sample and is affected by any RNA manipulation - particularly ribosome elimination. For replicates within a single polyA-selected SEQC experimental run, the ρ of a mix varies slightly, likely due to fluctuations in efficiency of mRNA enrichment. (S.Table 1) It is also important to note that FPKM units should not be used to calculate mRNA fraction (Supp.Figure 3), as the FPKM derivation [6] includes a term which couples sample abundance to spike abundance.

Mixture analysis models recapitulate known mixture proportions

To demonstrate the accuracy of this analytical framework of mixture sequencing, the mixture proportions Φ_{BLM} were recalculated for the BLM mixtures BLM-1 and BLM-2. The ρ values and the sequencing expression data were used to solve for the mixture proportions Φ_{BLM} by linear regression to the mixture equation**.** Figure 2 shows the Φ_{BLM} values at which residuals were minimized for the two mixtures for each replicate sample in each laboratory. Estimates of the three component proportions in the two mixtures are consistent with the designed 25:25:50 and 25:50:25 proportions in the two BLM mixtures. Supp. Figure 4 shows that the designed proportions of SEQC mixtures can also be calculated by this equation, returning the 75:25 and 25:75 proportions for mixes C and D, with some variability between labs. Equation 1, which lacks correction for mRNA fraction, does not return the designed ratios (Supp. Figure 5).

Linear model-predicted mixture counts are equivalent to replicate measures

In studies by the SEQC [34], differential expression between replicate samples was utilized to evaluate measurement performance based on the hypothesis that the control samples used in the study had no true differences between replicates. We created pseudo-replicate predicted count values from the 'neat' samples for use in benchmarking. These simulated mixtures were built based on the measured mixture expression and the true mixture proportions**.**

Figure 3 shows a dendogram of the distance between actual mixture expression and simulated expression counts of SEQC samples. The four base samples A, B, C and D are most different from one another, reflecting the biological differences between the samples. A and C are more closely related, as C consists of 75% A and 25% B. Modeled pseudo-replicate samples 'Cm' and 'Dm' across each of the six SEQC sites are no more different than cross-lab replicates of the C and D data, indicating that building the model for mixture C from components A and B does not introduce significant variability. This supports the treatment of modeled mixtures as replicate measurements expected to have no true differential expression from the mixture samples.

Discussion:

If the response function of a measurement is linear, mixtures of biological samples can be useful as genome-scale process controls for that measurement. When this condition is met, a mixture can be modeled simply as linear combination of its components. Two experimental datasets with known mixture parameters were used to verify these assertions. In the case of RNA-seq, the mRNA fraction of the total RNA mixture components must be accounted for in order to reflect the true values, when mixtures of RNA are calculated based on mass fractions of total RNA and the sequencing experiments measures only mRNA.

Mixtures with either known or unknown proportions can be analyzed. If mixture proportion information is known *a priori*, genome-scale data can be used as a process control to test the repeatability and sensitivity of measurements by comparing observed and expected measures. Alternatively, if the mixture proportions are an unknown and desired parameter, expression measures from the mixture in combination with the neat components can be used to experimentally determine the mixture proportions. This application can be valuable to un-mixing biological mixtures, including clinical mixtures, cell cultures, and xenografts[27-32]. While the mRNA fraction correction is applicable only to RNA-sequencing measurements, the general mixture model is theoretically applicable to any measurement with a linear response function.

Mixtures can provide measurement process assurance to a sequencing experiment. Using mixture samples alongside pure samples, one can demonstrate the reproducibility and sensitivity of genome-scale RNA, protein, as well as metabolite measurements. The main goal of this type of mixture analysis is to create a known ratio value by which the measurement characteristics of an experiment can be assessed. While an experiment's measurement of this known ratio is not sufficient to prove the validity of the measurement, it is a necessary condition, and any deviations are indicative of bias.

While we demonstrate mixture analysis with two specific samples, the analysis is fully generalizable to any number or type of mixture components. Any mixture split into known individual components can be measured in this way. For example, a clinical researcher may have three samples of interest from healthy, chronically diseased and acutely diseased sources. A mixture of these three cell types would provide confidence in the measurements made on the three samples individually by verifying the repeatability of that measurement. It can also provide a benchmark sample to assess comparability over space and time. These mixtures can detect biases introduced by batch effects, operator effects, sample mislabeling, and technical artifacts while evaluating the variability of the measurement. Mixture samples with known proportions can help determine experimental reproducibility and discover technical artifacts introduced by the measurement process by comparison of the expected to observed proportions.

With this analytical model, end users and core facilities can use known mixtures as a process control to track changes in measurement quality whenever changes to the experimental process are made. By including a predefined mixture, cross-sample comparisons can be made to demonstrate the internal consistency of measurements made using any new experimental technique, kit, or downstream analysis tool. In this way, there is some assurance that changes in experimental protocol have not affected measurement reproducibility. Residuals from modeled counts can be used as a metric to evaluate the magnitude of effect an experimental process has on the linearity and precision of sequence measurements.

In addition to gaining an understanding of the measurement process using the benchmarking workflow, unknown samples can be collected and studied to determine the relative proportion of known components. Proportions of components can be determined even in the absence of any type-specific markers, given measurable differences in expression between the cell types.

Resolving the composition of mixtures has proven useful in determining the purity of cell lines or proportions of heterogeneous cells, in identifying interesting cellular contaminants such as partially differentiated cells, and understanding clinical samples containing mixed cell types. In contrast to approaches using transgene expression [41], the mixture model described here can evaluate tissue sample purity without focusing on a handful of “tissue-specific”, marker, or transgenes. We expect mixed-sample RNA to be useful in regulatory applications, where a demonstration that a therapeutic stem-cell mixture has a specific composition may be key to ensuring safety [48].

Spike-in controls correct for mRNA fraction-caused biases in linearity

In addition to providing limit of detection and cross-experiment comparison characterizations of a dataset, spike-in controls can be used in mixture samples to determine the mRNA fraction of cells. mRNA fraction is a critical parameter for comparing samples that do not have identical total RNA content. This is most relevant to cells with variable global expression [24], including comparisons across and within cell cycle, tissues, and developmental states [40]. mRNA fraction is also critical in single cell gene expression studies, where lysis efficiency and total RNA content can vary greatly from cell to cell.

There are many methods used to determine component gene expression profiles from mixture samples. At present, only the one we describe here explicitly accounts for mRNA fraction. In RNA-seq experiments, mRNA fraction can be calculated with information obtained via spike-in controls. When comparing samples with variable mRNA content, bias arises when that variability is not accounted for. We describe a straightforward method for measuring the enrichment of mRNA in RNA-seq samples using spike-in RNA. We show that mRNA-corrected unmixing of two mixture datasets returns the known mixture proportions (Figure 2, Supp Figure 4), demonstrating suitability for solving unknown mixtures of known components.

Previous methods used to determine the composition of RNA-seq mixtures make inaccurate estimates of mixture proportion in the BLM sample where the mRNA fractions vary substantially between mixture components. These methods are nearer to true values in the SEQC sample, where the mRNA fraction difference is less significant, but all estimates are improved by incorporating mRNA content (Supp. Figure 5)**.**

Limitations

Technical artifacts identified by mixture modeling are differentially expressed between replicates and should not be confused with the conventional usage of differentially expressed genes, which are compared between samples. Truly differentially expressed genes, such as tissue-specific genes in BLM mixtures, fit well to the model. Technical artifacts would not be identified as differentially expressed by modern differential gene expression methods due to their extreme variance between replicates as a result of crosstalk or nonlinear response. This means the measurement is not sensitive to these genes, and they could be false negatives. If transcripts are identified as artifacts in a process, alternative preparations need to be employed to achieve an unbiased quantification of affected transcripts.

Although mean mixture proportion values returned from a linear combination of mixture components approximate the nominal mixture proportion in both measured samples, the increased variability of the muscle estimate in the BLM mixture (error bars, Figure 2) suggests that there is a lower limit to being able to determine low-abundance mixture components. Due to mRNA fraction, the muscle component of the BLM mix was as low as 10 percent of sequenced RNA in BLM-2. It may be possible to determine lower-proportion mixture components with confidence, but this study did not generate the required data to do so.

Our estimation of mRNA fraction is imperfect; an assumption of the model we build is that the mRNA fraction is constant between replicates of the same sample. Supplemental Table 1 shows that the mRNA fraction varies by as much as 5 percent from library to library. This variability is a source of error in our model. The variability in mRNA fraction is likely due to batch effects in the mRNA enrichment process. This hypothesis is reinforced by the prevalence of non-mRNA transcripts incorrectly called as differentially expressed between mixture replicates.

The sequencing technology and library preparation methods used in these experiments added limitations to the experiments. These are described in supplemental note 1.

Conclusions

We demonstrate the linear response function and specificity of RNA-sequencing measurements using mixtures of biological samples. Such mixtures can be used as benchmarks to characterize the repeatability and reproducibility of experiments or separated to identify the relative proportion of their components. Spike-in controls can be used to calculate the mRNA content of total RNA mixtures, compensating for biases introduced by mRNA enrichment. Our method creates a framework for using mixtures in measurement process control and corrects for biases introduced by ribosomal depletion. Using an mRNA fraction correction improves the accuracy of mixture proportion determination in RNA-seq experiments.

Benchmarking genome-scale measurements using mixed samples will remain useful even after the era of short-read sequencing is over. Answering the biological question of "what types of cells are in the mixture I'm sequencing?" requires more information than even a perfect transcriptome reconstruction could provide. The biological and measurement value added by mixed samples are demonstrated here to be platform-independent. We anticipate that mixtures can provide the same measurement assurance to protein and metabolite measurements. Confidence in the reproducibility of measurement and understanding the components in complex biological samples will always be a staple of quality science.

Methods:

Library Preparation:

For the BLM experiment, Human Brain Reference RNA, Human Liver Total RNA, and Human Skeletal Muscle Total RNA were purchased from Ambion. This purified RNA was quantified by absorbance on a NanoDrop 1000, mixed in the specified proportions, then spiked with ERCC RNA transcribed from NIST SRM 2374. For Illumina sequencing, the Illumina TruSeq protocol was followed. HiSeq runs generated 100+100bp paired-end reads. Solid 5500 sequencing followed the Life Technologies Whole Transcriptome protocol, yielding 75+35 bp paired-end reads. Spike-in composition and amounts are included in the data submission to GEO.

Quantitation and Data Normalization:

BLM gene counts were based on raw count data quantified using HTSeqCounts [40] based on a variety of genome and transcriptome references [42-45] after mapping reads to the genome with Topha t[46]. Raw counts were then normalized using the upper quartile method implemented in EdgeR [36]. Supplemental Figure 3 utilizes RSEM [47]. HTSeq-counts version 0.5.4 was run with options to deal with non-stranded reads in the intersection-nonempty mode. The SEQC data used are available as count tables from GEO GSE47774.

Calculating Unknown Mixture Estimates:

The relative abundance of components in unknown mixtures were calculated by first observing the mean mRNA fraction for the neat components across replicates. The count data in the mixture was set as the response, predicted by the count data from the individual components modified by the mRNA fraction, as based on the mixture equations. An example R script 'generalmixturesolver' is provided at http://github.com/jeparson/mixtureprocesscontrol as a supplemental file to clarify this procedure.

Availability of supporting data:

The SEQC data is available from GEO GSE47774.
[http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47774]
The BLM data is available from the European Nucleotide Archive, PRJEB8231.
[http://www.ebi.ac.uk/ena/data/view/PRJEB8231]
Figure code, count tables, and example scripts available on
https://github.com/usnistgov/mixtureprocesscontrol
List of Abbreviations: ERCC - External RNA Control Consortium, TPM – Transcripts per Million, FPKM – Fragments per Kilobase per million mapped reads, mRNA – messenger RNA

Competing Interests:

Certain equipment and instruments or materials are identified in the paper to adequately specify the experimental details. Such identification does not imply recommendation by NIST, nor does it imply the materials are necessarily the best available for the purpose.

Author Contributions:

Analysis by JP, MS, PP, MM, and SM. Manuscript by JP. Experimental design by PP, SM, JM, MS, and MM. Sample preparation and sequencing by JM+NCI sequencing core.

Author Information

JP MS PP SM - NIST/ABMS. JM NIST MM ?(currently U of W, previously NIH/NCI)

References

1. Van Dijk EL, Jaszczyszyn Y, Thermes C: Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res 2014, 322:12–20.

2. Hansen KD, Brenner SE, Dudoit S: Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 2010, 38:e131.

3. Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W, Fodor SPA: Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proc Natl Acad Sci USA 2014, 111:1891–1896.

4. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, Pizarro A, Kim J, Irizarry R, Thomas RS, Grant GR, Hogenesch JB: IVT-seq reveals extreme bias in RNA sequencing. Genome Biol 2014, 15:R86.

5. Hart T, Komori HK, LaMere S, Podshivalova K, Salomon DR: Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics 2013, 14:778.

6. Wagner GP, Kin K, Lynch VJ: Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 2012, 131:281–285.

7. Raz T, Kapranov P, Lipson D, Letovsky S, Milos PM, Thompson JF: Protocol dependence of sequencing-based gene expression measurements. PLoS ONE 2011, 6:e19287.

8. Jue NK, Murphy MB, Kasowitz SD, Qureshi SM, Obergfell CJ, Elsisi S, Foley RJ, O'Neill RJ, O'Neill MJ: Determination of dosage compensation of the mammalian X chromosome by RNA-seq is dependent on analytical approach. BMC Genomics 2013, 14:150.

9. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B: Synthetic spike-in standards for RNA-seq experiments. Genome Res 2011, 21:1543–1551.

10. External RNA Controls Consortium: Proposed methods for testing and selecting the ERCC external RNA controls. BMC Genomics 2005, 6:150.

11. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, Foy C, Fuscoe J, Gao X, Gerhold DL, Gilles P, Goodsaid F, Guo X, Hackett J, Hockett RD, Ikonomi P, Irizarry RA, Kawasaki ES, Kaysser-Kranich T, Kerr K, Kiser G, Koch WH, Lee KY, Liu C, Liu ZL, Lucas A, et al.: The External RNA Controls Consortium: a progress report. Nat Methods 2005, 2:731–734.

12. Pine PS, Rosenzweig BA, Thompson KL: An adaptable method using human mixed tissue ratiometric controls for benchmarking performance on gene expression microarrays in clinical laboratories. BMC Biotechnol 2011, 11:38.

13. Thompson KL, Rosenzweig BA, Pine PS, Retief J, Turpaz Y, Afshari CA, Hamadeh HK, Damore MA, Boedigheimer M, Blomme E, Ciurlionis R, Waring JF, Fuscoe JC, Paules R, Tucker CJ, Fare T, Coffey EM, He Y, Collins PJ, Jarnagin K, Fujimoto S, Ganter B, Kiser G, Kaysser-Kranich T, Sina J, Sistare FD: Use of a mixed tissue RNA design for performance assessments on multiple microarray formats. Nucleic Acids Res 2005, 33:e187.

14. Duewer DL, Jones WD, Reid LH, Salit M: Learning from microarray interlaboratory studies: measures of precision for gene expression. BMC Genomics 2009, 10:153.

15. Li Y, Xie X: A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. BMC Bioinformatics 2013, 14 Suppl 5:S11.

16. Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ: Cell type-specific gene expression differences in complex tissues. Nat Methods 2010, 7:287–289.

17. Gaujoux R, Seoighe C: Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. Infect Genet Evol 2012, 12:913–921.

18. Quon G, Morris Q: ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. Bioinformatics 2009, 25:2882–2889.

19. Gong T, Szustakowski JD: DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics 2013, 29:1083–1085.

20. Gong T, Hartmann N, Kohane IS, Brinkmann V, Staedtler F, Letzkus M, Bongiovanni S, Szustakowski JD: Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. PLoS ONE 2011, 6:e27156.

21. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, Mills GB, Verhaak RGW: Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun 2013, 4:2612.

22. Yadav VK, De S: An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. Brief Bioinformatics 2014.

23. Kuhn A, Thu D, Waldvogel HJ, Faull RLM, Luthi-Carter R: Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. Nat Methods 2011, 8:945–947.

24. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA: Revisiting global gene expression analysis. Cell 2012, 151:476–482.

25. Aanes H, Winata C, Moen LF, Østrup O, Mathavan S, Collas P, Rognes T, Aleström P: Normalization of RNA-sequencing data from samples with varying mRNA levels. PLoS ONE 2014, 9:e89158.

26. Robinson MD, Oshlack A: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010, 11:R25.

27. Quon G, Haider S, Deshwar AG, Cui A, Boutros PC, Morris Q: Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome Med 2013, 5:29.

28. Shen-Orr SS, Gaujoux R: Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol 2013, 25:571–578.

29. Bock C, Lengauer T: Managing drug resistance in cancer: lessons from HIV therapy. Nat Rev Cancer 2012, 12:494–501.

30. Yuan Y, Failmezger H, Rueda OM, Ali HR, Gräf S, Chin S-F, Schwarz RF, Curtis C, Dunning MJ, Bardwell H, Johnson N, Doyle S, Turashvili G, Provenzano E, Aparicio S, Caldas C, Markowetz F: Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Sci Transl Med 2012, 4:157ra143.

31. Zhao Y, Simon R: Gene expression deconvolution in clinical samples. Genome Med 2010, 2:93.

32. Durham AL, Wiegman C, Adcock IM: Epigenetics of asthma. Biochim Biophys Acta 2011, 1810:1103–1109.

33. Liu W, Hou Y, Chen H, Wei H, Lin W, Li J, Zhang M, He F, Jiang Y: Sample preparation method for isolation of single-cell types from mouse liver for proteomic studies. Proteomics 2011, 11:3556–3564.

34. SEQC/MAQC-III Consortium, SEQC/MAQC-III Consortium: A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol 2014, 32:903–914.

35. MAQC Consortium, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD,

Lobenhofer EK, Puri RK, Schrf U, Thierry-Mieg J, Wang C, Wilson M, et al.: The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006, 24:1151–1161.

36. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010, 26:139–140.

37. Shippy R, Fulmer-Smentek S, Jensen RV, Jones WD, Wolber PK, Johnson CD, Pine PS, Boysen C, Guo X, Chudin E, Sun YA, Willey JC, Thierry-Mieg J, Thierry-Mieg D, Setterquist RA, Wilson M, Lucas AB, Novoradovskaya N, Papallo A, Turpaz Y, Baker SC, Warrington JA, Shi L, Herman D: Using RNA sample titrations to assess microarray platform performance and normalization techniques. Nat Biotechnol 2006, 24:1123–1131.

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

39. Qing T, Yu Y, Du T, Shi L: mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. Sci China Life Sci 2013, 56:134–142.

40. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, Heisler MG: Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 2013, 10:1093–1095.

41. Amaral AJ, Brito FF, Chobanyan T, Yoshikawa S, Yokokura T, Van Vactor D, Gama-Carvalho M: Quality assessment and control of tissue specific RNA-seq libraries of Drosophila transgenic RNAi models. Front Genet 2014, 5:43.

42. Thierry-Mieg D, Thierry-Mieg J: AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 2006, 7 Suppl 1:S12.1–14.

43. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, Murphy MR, O'Leary NA, Pujar S, Rajput B, Rangwala SH, Riddick LD, Shkeda A, Sun H, Tamez P, Tully RE, Wallin C, Webb D, Weber J, Wu W, DiCuccio M, Kitts P,

Maglott DR, Murphy TD, Ostell JM: RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 2014, 42(Database issue):D756–763.

44. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ: ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res 2013, 41(Database issue):D56–63.

45. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, et al.: GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012, 22:1760–1774.

46. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 2013, 31:46–53.

47. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011, 12:323.

48. Assessing Human Stem Cell Safety[Stem Cell Information] [http://stemcells.nih.gov/info/scireport/pages/chapter10.aspx]

49. Munro SA, Lund SP, Pine PS, Binder H, Clevert D-A, Conesa A, Dopazo J, Fasold M, Hochreiter S, Hong H, Jafari N, Kreil DP, Łabaj PP, Li S, Liao Y, Lin SM, Meehan J, Mason CE, Santoyo-Lopez J, Setterquist RA, Shi L, Shi W, Smyth GK, Stralis-Pavese N, Su Z, Tong W, Wang C, Wang J, Xu J, Ye Z, et al.: Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nat Commun* 2014, **5**.

Figure Legends:

Figure 1: RNA samples used in this study. RNA isolated from pure tissues is used to generate pairs of mixtures used in two separate experiments. (a): Two SEQC mixtures (SEQC-C and SEQC-D) are built from two components (SEQC-A and SEQC-B). (b): Two BLM mixtures (BLM-1 and BLM-2) are built from three components. The SEQC-B component (HBRR) is from the same source as the Brain BLM component. Per-sample target ratios of tissue proportion between mixtures are shown.

Figure 2: Accuracy of model-derived BLM mix estimates. The grey center point is the nominal 'truth' ratio in which the samples were mixed. Concentric circles with radius 0.025 and 0.05 are added to visually clarify distance from the center point. Colored points depict mixture proportion *(Φ)* estimates generated from measurements of 4 replicate libraries. Black points are the mean of the replicates. Error bars show one standard deviation of the four replicate measures

Figure 3: Clustering of Expression measures in 4 SEQC samples and 2 *in-silico* replicate samples across participating sites: The close agreement between modeled (Cm, Dm) counts and actual counts (A,B,C,D) at sites numbered 1-6 supports the validity of assumptions used to model Cm and Dm counts. Euclidian distance measures between samples show that the various samples are of greater distance from one another, while the *in-silico* modeled samples are most similar to the correct corresponding sample.

Supplemental Figures:

Supplemental Figure 1: Mixture proportion (*Φ*) estimates for samples A in SEQC-C and SEQC-D. The mean (black hollow circle) and standard deviation (error bars) of four individual replicates (colored) of the *Φ* estimate for each sample are shown. The nominal mixture proportions are grey points at the center of the target. Circles centered at that nominal ratio with radii in multiples of .025 are included to more easily identify magnitude of total error. LT and ILM tags indicate the manufacturer of the sequencer used at each lab (Life Technologies and Illumina, respectively). Especially given that the actual creation of the SEQC mixtures was done one time, independent from these laboratories, deviations from the target indicate process variability or errors brought about in these labs. Lab 2 appears to have done something to seriously distort the repeatability of SEQC-C, for example. This could be an indication of inconsistent polyA selection from (for example) inconsistent recovery of RNA off of magnetic beads.

Supplemental Figure 2: Bland-Altman log-ratio(M) - log average(A) plots comparing gene expression in BLM-1 to BLM-2, which were mixed with a designed ratio of 1:1 brain RNA, 2:1 muscle RNA and 1:2 liver RNA. Points representing gene expression values for genes expressed at 5-fold greater levels in a specific tissue are colored based on the sample in which they are selectively expressed. The left panel is Illumina HiSeq expression data, while the right panel is from a SOLiD 5500. ERCC spike-ins in the SOLiD 5500 dataset were in three sub-pools added in the same ratios as the three tissues, while all 96 ERCC controls were spiked at a 1:1 ratio in the HiSeq dataset. Non-tissue selective mRNAs are omitted for clarity.

Supplemental Figure 3:

MA plots of HiSeq counts obtained from BLM-1 vs. BLM-2 are presented here without mRNA fraction correction, using typical normalization methods. Library size normalization scales all libraries to a common total number of counts, while upper quartile normalization scales to the 75th percentile of the counts for each library. Supp. Figure 2 shows the data after correcting for mRNA fraction differences.

	BLM1-a	BLM1-ad	BLM1-au	BLM1-b	BLM1-bd	BLM1-bu	BLM2-a	BLM2-b	BLM2-bd	BLM2-bu
Count Ratio	.0695	.0095	.6698	.0719	.0098	.6342	.0706	.0737	.0098	.6649
Spike Added	.08	.01	.64	.08	.01	.64	.08	.08	.01	.64
message fraction **ρ**	1.152	1.058	.955	1.112	1.017	1.009	1.132	1.085	1.025	.962

Supplemental Table 1: Message RNA fraction (ρ) calculations as a function of spike amount. Spike mass is accounted for in the mRNA fraction calculation. The spike-ins varied by amount (“u” or “d” samples) and content (pools ‘a’ or ‘b’) in both tissue mixtures (1 and. 2). Calculated mRNA fractions vary by +/- 6% across these 10 BLM mixtures, showing that the calculation is robust to spike-in mass and content. mRNA fraction calculations for the ERCC pools must account for the 3-plex nature of the mixes. The shown ratios are for the subset of spike-ins which are present at a 1:1 ratio in each sample.

Supplemental Figure 4:

The effect of using FPKM units. Estimates of mRNA fraction (light points are calculated using count values, dark points using FPKM values) result in a relatively poor solution to the mixture proportion. Both data types are taken from the same RSEM output.

Supplemental Figure 5: Mixture proportions returned by a simple model (Equation 1, blue circles), by an mRNA-corrected model(*ρ-corrected* mixture equations, green triangles) and by the DeconRNASeq package[36] (red diamonds) on SEQC data. Lab # - LT and - ILM indicate the manufacturer of the sequencer used at each participating lab (Life Technologies and Illumina, respectively).DeconRNASeq implements the same general idea, but lacks mRNA fraction correction. In the SEQC data, there is a relatively small mRNA fraction difference between samples, but significant improvements are achieved by correcting for the mRNA fraction. The mean distance from true value across all labs is 0.052(Simple model), 0.033(mRNA-corrected), and 0.048(DeconRNASeq). Error bars represent the SD of four independent libraries from the same RNA source.

Supplemental Note 1:

RNA-seq is capable of making transcript isoform-specific measurements. However, long reads of high depth are required to adequately differentiate between isoforms. Investigations of isoform-level measurements from the BLM dataset, (Table 2) which utilized 75x35bp paired-end reads on the 5500 and 100x100bp paired-end reads on the HiSeq, showed that while the model is extensible towards such measurements, the reduced mean read counts make transcript isoform-level expression measurements less precise due to shorter read length and lower sequencing depth. 92 percent of genes were modeled to within 1 log2 unit of the measured value, while only 85 percent of transcripts were.

	Genes Measured	Genes Modeled (+/- 1 log2)	Percent	Transcripts Measured	Transcripts Modeled (+/- 1 log2)	Percent
BLM	19036	17641	92.6	23182	19772	85.3
SEQC	23947	22820	95.3	40333	38434	95.3

The substantially increased read depth in the SEQC experiment led to 95% of both isoforms and genes being consistently modeled. In the SEQC dataset, 95% of detected isoforms could be consistently modeled to within a factor of 2, and the same percentage of genes could be reasonably predicted. After applying a variance-stabilizing transformation using DEseq[38], every gene and transcript (100%) in the SEQC dataset were correctly modeled by these criteria. The BLM dataset does not contain sufficient replication for variance-stabilizing analysis.

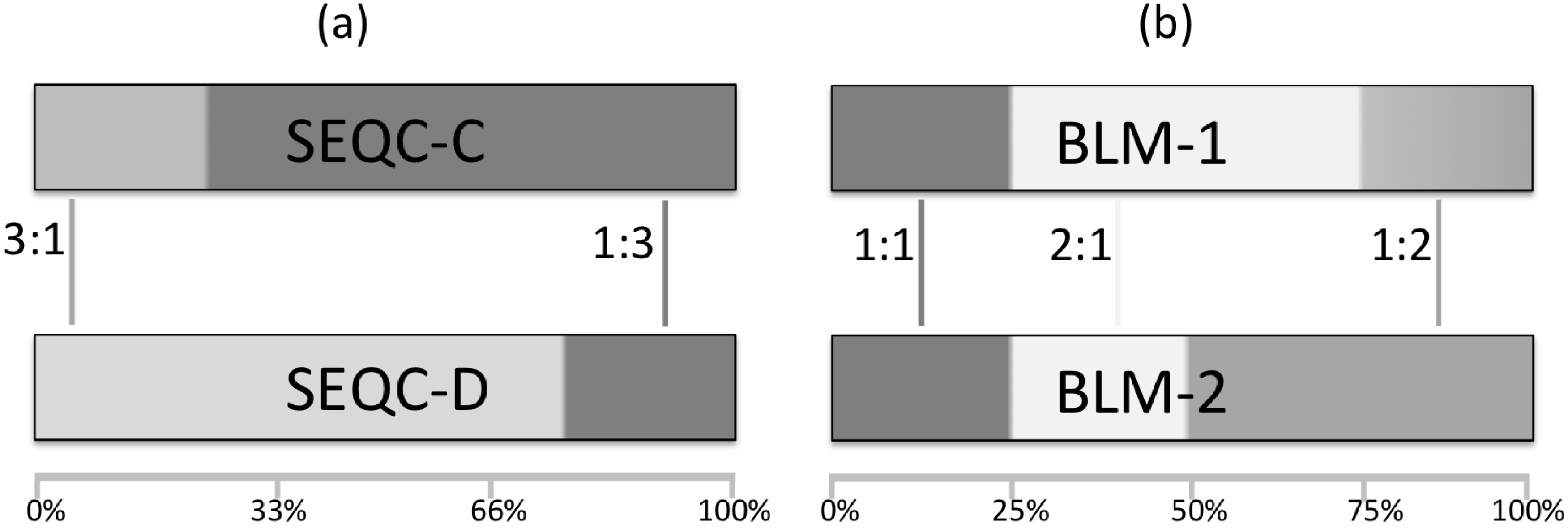

Figure 1

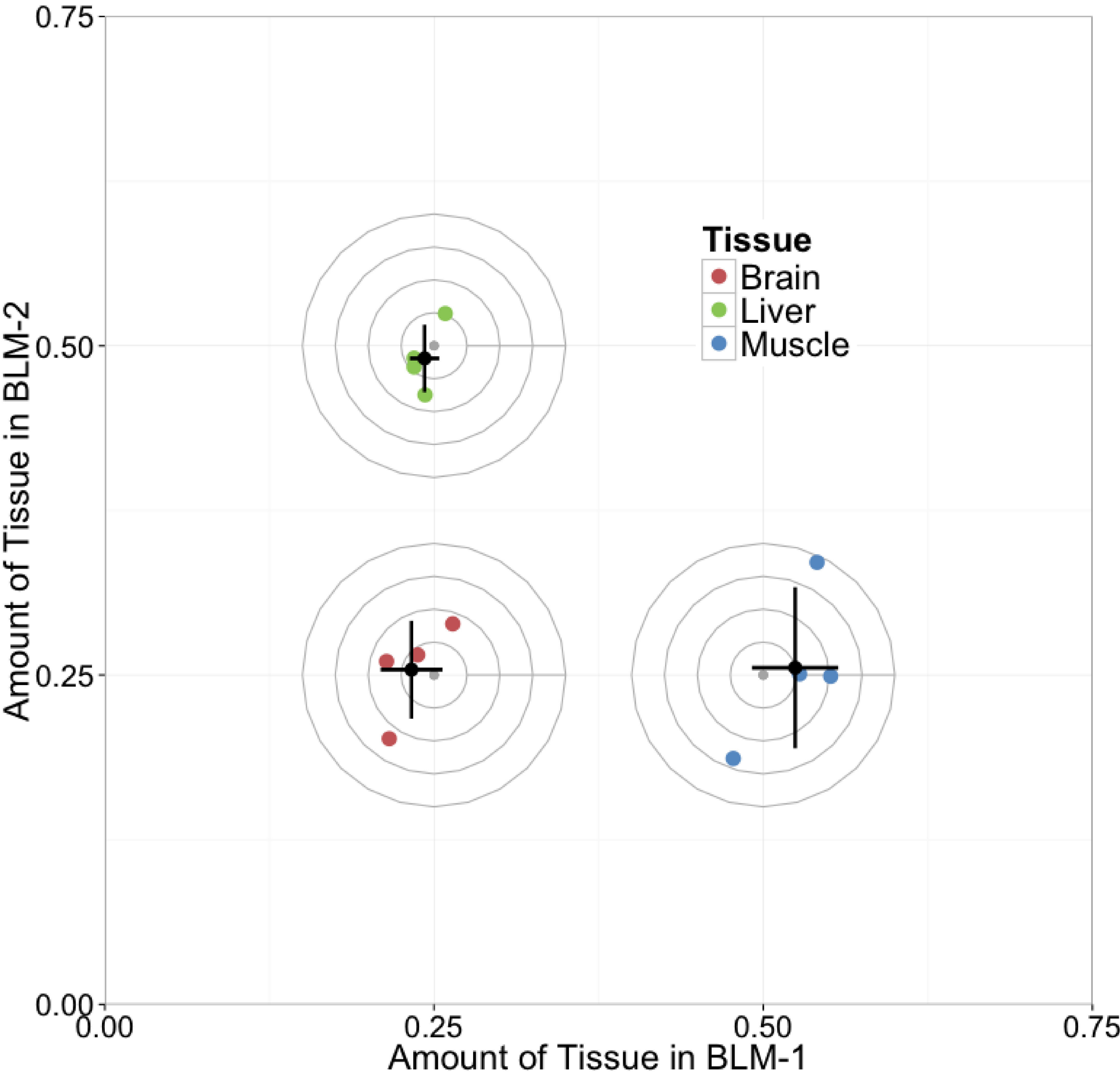
0.75
0.50
0.25
0.00
Amount of Tissue in BLM-2
0.00
0.25
0.50
0.75
Amount of Tissue in BLM-1
Tissue
Brain
Liver
Muscle

Figure 2

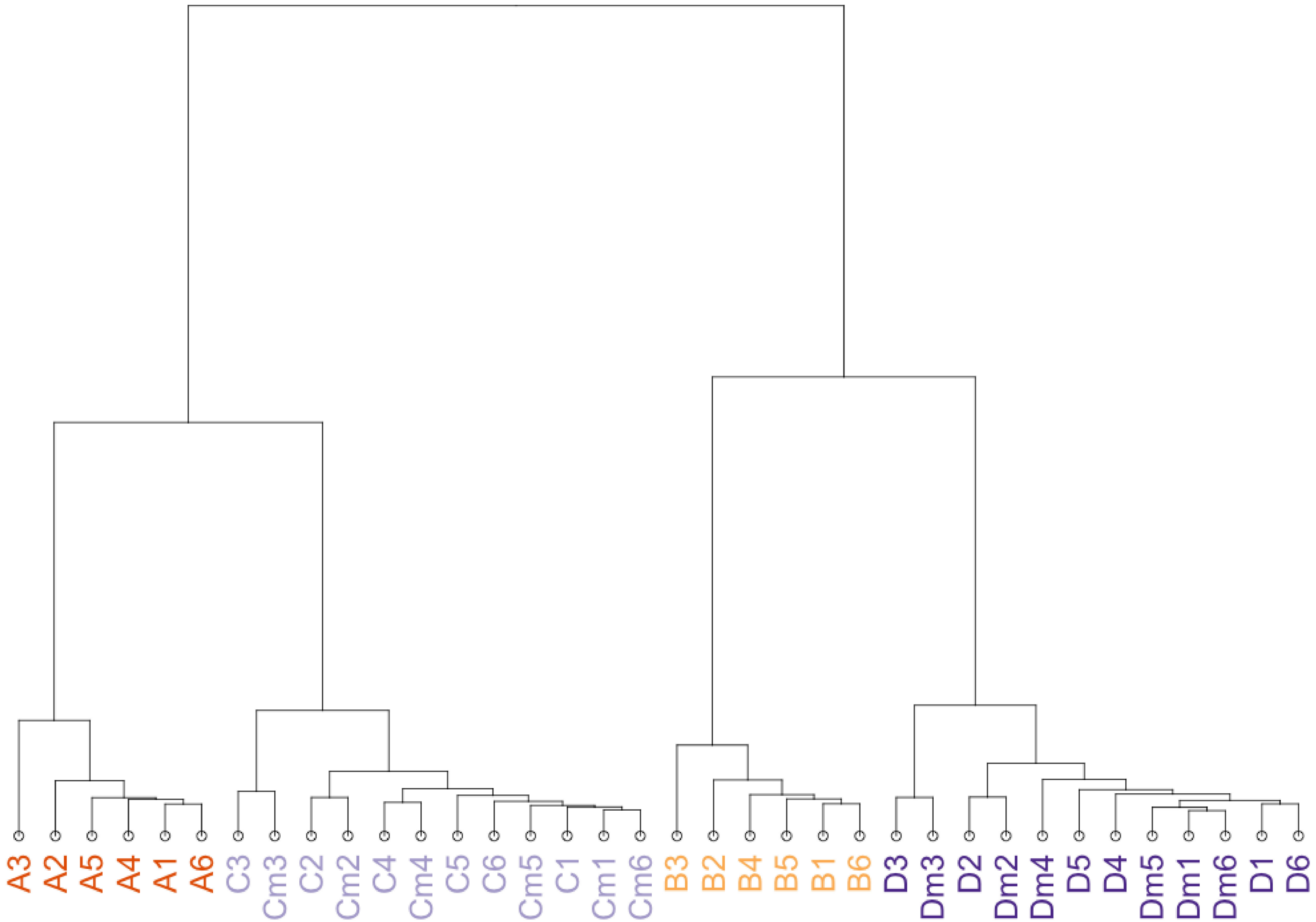
A3
A2
A5
A4
A1
A6
C3
Cm3
C2
Cm2
C4
Cm4
C5
C6
Cm5
C1
Cm1
Cm6
B3
B2
B4
B5
B1
B6
D3
Dm3
D2
Dm2
Dm4
D5
D4
Dm5
Dm1
Dm6
D1
D6

Figure 3

Additional files provided with this submission:

Additional file 1: FSF1.png, 328K
http://www.biomedcentral.com/imedia/1842199884154088/supp1.png
Additional file 2: FSF2.png, 27K
http://www.biomedcentral.com/imedia/1916775531154088/supp2.png
Additional file 3: fsf3.png, 65K
http://www.biomedcentral.com/imedia/9248942721540888/supp3.png
Additional file 4: fsf4.png, 82K
http://www.biomedcentral.com/imedia/3317422451540888/supp4.png
Additional file 5: fsf5.png, 91K
http://www.biomedcentral.com/imedia/1126498645154088/supp5.png