Supplemental Information

Sequencing technology does not eliminate biological variability
Kasper D. Hansen, Zhijin Wu, Rafael A. Irizarry, Jeffrey T. Leek

Methods

Preprocessing of sequence datasets

Unmapped single-end sequencing reads from the Pickrell *et al*. study[1] were obtained from http://eqtl.uchicago.edu/ in FASTQ format. A number of subjects were assayed twice, we used the first replicate for those (but see below).

Unmapped paired-end sequencing reads from the Montgomery et al. study[2] were obtained from the European Nucleotide Archive, http://www.ebi.ac.uk/ena/ in FASTQ format. For our analysis, we used the first read of the mate-pairs.

All reads were trimmed from the 3' end to have a length of 35bp. They were aligned to hg19 using Bowtie[3] with mapping parameters "-m 1 –v 2 –y".

Annotation for the human genome was obtained from ENSEMBL (version 61) and union gene models were constructed[4]. A union gene model for a given gene, as defined in the *Genominator* R package, is the union of all bases belonging to any isoform of the gene[5]. Bases belonging to multiple genes are removed. Overlap between sequencing reads and union gene models were determined by identifying each read with its center position and declaring an overlap if this center position belongs to the gene model.

For each dataset, log2-transformed RPKMs[6] were formed as:

$$\log_2\left(\frac{\mathrm{G}+1}{\mathrm{L}\times \mathrm{S}/10^9}\right)$$

with G being the number of reads belonging to the gene model for a fixed gene, S being the total number of reads belonging to all the gene models for the sample, and L being the total length of the gene model for the gene in question in bp.

Preprocessing of microarray datasets

Celfiles from the Choy *et al.* study[7] were obtained from GEO (GSE11582). We selected all celfiles corresponding to unrelated Yorubian individuals and normalized them together using RMA[8]. After normalization, we furthermore selected the first technical replicate for each sample (only a few had multiple technical replicates) and only kept samples that were processed at the Broad. Probesets were mapped to ENSEMBL gene identifiers using the hgu133a.db package from Bioconductor, which encodes mapping results from Affymetrix. Probesets mapping to multiple ENSEMBL gene identifiers were discarded. Expression measures from RMA are on the log2-scale.

Normalized (within population) data from Stranger *et al.* study[9] were obtained from GEO (GSE6536). Probes were mapped to ENSEMBL gene identifiers using mappings provided by ENSEMBL version 61. Probes mapping to multiple ENSEMBL gene identifiers were discarded. Expression measures were calculated on the log2-scale.

Matching of microarray and sequencing datasets

After preprocessing, the datasets were matched based on ENSEMBL gene identifiers. For each comparison, we only retained samples that were assayed in both datasets being compared. Genes with zero sequencing reads in all samples in the comparison were discarded. After sample matching, the expression measures for sequencing (log2 transformed RPKMs) were quantile normalized.

For Figure 1 we only retained genes that had an RPKM (calculated as above) greater than 5. For the array data that was assayed on an Affymetrix platform we furthermore required that all samples had an RMA-expression value greater than 5. No such filtering was performed for the Illumina data. The same values were used to make scatterplots of coefficients of variations (**Supplementary Fig. 1**).

Probe-local expression measures

The union gene models described above imply that the expression measures for RNA-Seq are obtained over large genomic intervals, while microarrays either use a single probe (Illumina) or a set of probes near the 3' end of the transcript (Affymetrix). We wanted to investigate whether our results would change if we use a gene model defined locally around the probe(s) on the microarrays. The idea is to define a probe-local gene model as the union of all exons overlapping the probe or probeset.

We obtained mappings of the microarray probes to the hg19 transcriptome from the MySQL interface to ENSEMBL version 61. For each probe we obtained all the ENSEMBL exons overlapping either the 3' or 5' end of the probe (this was done in order to deal with probes overlapping exon-exon junctions). For the Illumina array we defined a probe-local gene model to be the union of all exons overlapping the probe, discarding probes overlapping exons from multiple genes. For the Affymetrix array we likewise defined a probe-local gene model to be the set of exons overlapping any probes in a given probeset, discarding probesets which overlapped multiple genes or where 5 or less (out of 11) probes did not overlap an exon.

RPKMs were calculated as above for these local gene models, including requiring probe-local gene models to have RPKMs of 5 or greater. Defining and filtering probe-local gene models in this way allowed us to match more microarray probes or probesets to the RNA-Seq data.

Analysis of technical replicates

In the Pickrell *et al.* study 11 subjects were assayed twice, with separate RNA extractions from the same cell line and library preparations. In the Choy *et al.* study 14 subjects were assayed twice, again with separate RNA extractions from the same cell line and library preparations. These two sets of subjects were analyzed as above (without requiring the two sets of samples to cover the same individuals). Instead of filtering as described above, we retained exactly the genes retained in the analysis of the full datasets described above.

For each gene we fitted a mixed effect model,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where Y_{ij} is the log2 RPKM for individual i, replicate j, μ is a fixed effect describing the population level gene expression, α_i is a random effect with variance σ^2_{bio} describing the biological (individual to individual) variation and ε_{ij} is an error term with variance σ^2_{tech} describing technical variation. This model was fitted using the lmer function from the R package lme4, and we computed the ratio:

$$\frac{\sigma^2_{bio}}{\sigma^2_{bio} + \sigma^2_{tech}}$$

as a measure of the amount of biological variation compared to total variation. These values were used as basis for **Supplementary Figure 2a**.

Increased Variability in Sequencing May Be Due to Larger Dynamic Range

In **Figure 1** and **Supplementary Figure 1,** on average gene expression is more variable as measured with sequencing compared to microarrays. An analysis of spike-in experiments suggests that sequencing experiments accurately measure expression levels {Mortazavi, 2008 #56}. So the increased estimates of variability may reflect true biological variability and mean that sequencing measurements simply have a greater dynamic range. A larger scale spike-in experiment with biological replicates would be needed to confirm this result.

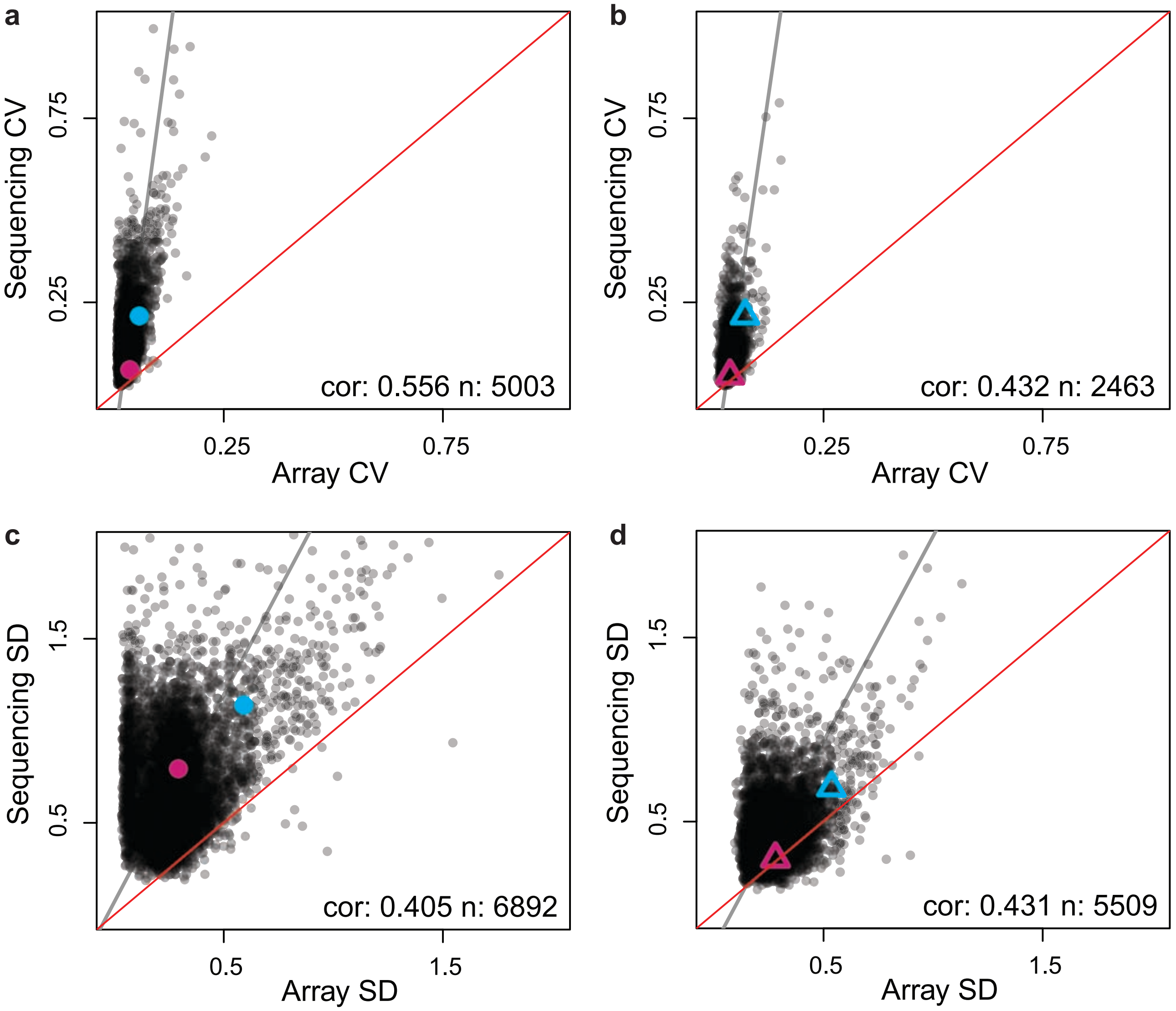
References:

1. Pickrell, J.K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772 (2010).
2. Montgomery, S.B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777 (2010).
3. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
4. Langmead, B., Hansen, K.D. & Leek, J.T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol* **11**, R83 (2010).
5. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
6. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).
7. Choy, E. et al. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet* **4**, e1000287 (2008).
8. Irizarry, R.A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003).
9. Stranger, B.E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-853 (2007).
10. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349 (2008).
11. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848 (2008).
12. Sultan, M. et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956-960 (2008).
13. Cloonan, N. et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**, 613-619 (2008).
14. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476 (2008).
15. Yoder-Himes, D.R. et al. Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing. *Proc Natl Acad Sci U S A* **106**, 3976-3981 (2009).
16. Lipson, D. et al. Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* **27**, 652-658 (2009).
17. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382 (2009).
18. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010).

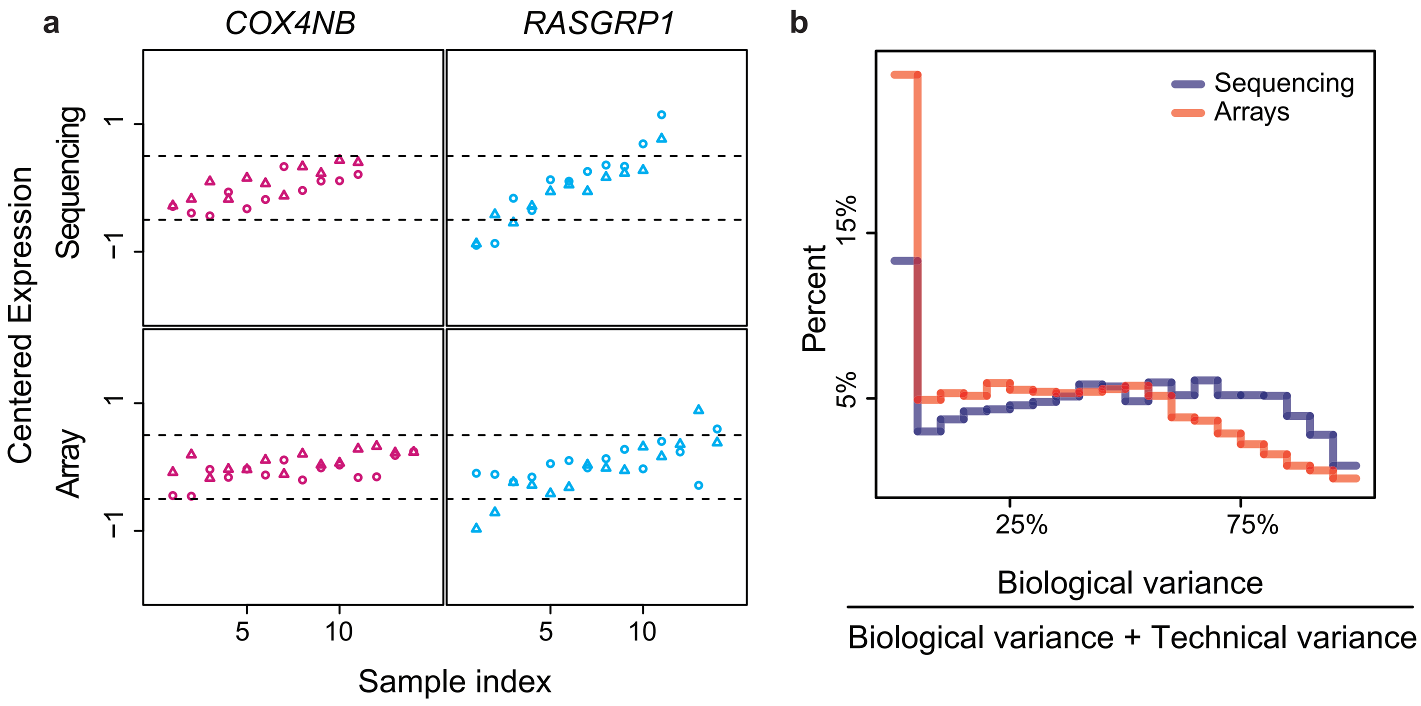
19. Bruno, V.M. et al. Comprehensive annotation of the transcriptome of the human fungal pathogen Candida albicans using RNA-seq. *Genome Res* **20**, 1451-1458 (2010).
20. Wilhelm, B.T. et al. RNA-seq analysis of 2 closely related leukemia clones that differ in their self-renewal capacity. *Blood* **117**, e27-38 (2010).
21. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009-1015 (2010).
22. Hammer, P. et al. mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome Res* **20**, 847-860 (2010).
23. Yang, F., Babak, T., Shendure, J. & Disteche, C.M. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res* **20**, 614-622 (2010).

Pubmed ID	Journal (Year)	# of Biological Groups	# of Technical Replicates	# of Biological Replicates
18451266[10]	Science (2008)	1	2	2
19056941[11]	Science (2008)	1	1	2
18516045[6]	Nature Methods (2008)	3	2	1
18599741[12]	Science (2008)	2	2	1
18516046[13]	Nature Methods (2008)	2	3	1
18978772[14]	Nature (2008)	15	1	1 (6 in 1 group)
19234113[15]	PNAS (2009)	2	1	2
19581875[16]	Nature Biotechnology (2009)	1	3	1
19349980[17]	Nature Methods (2009)	4	1	2
20436464[18]	Nature Biotechnology (2010)	4	1	1
20810668[19]	Genome Research (2010)	9	1	1
20980679[20]	Blood (2010)	2	1	2
21057496[21]	Nature Methods (2010)	2	1	1
20452967[22]	Genome Research (2010)	4	1	2
20363980[23]	Genome Research (2010)	1	1	1
20220758[1]	Nature (2010)	1	2	69
20220756[2]	Nature (2010)	1	1	60

Supplementary Table 1 RNA-sequencing studies with n≤3 biological replicates. The first column is the study name, the second column the journal and publication year, the third column is the number of biological groups assayed, the fourth column is the number of technical replicates, and the fifth column is the number of biological replicates. The studies analyzed in the main text are highlighted in blue.



Supplementary Figure 1 Comparing Different Measures of Biological Variability. **(a)** A plot of the coefficient of variation of expression values as measured with microarrays in the Stranger *et al.* study[9] (x-axis) and sequencing in the Montgomery *et al.* study[2] (y-axis). The coefficient of variation estimates from sequencing are larger than the estimates from microarrays, substantiating the result in **Figure 1**. These are the exact same genes as **Figure 1a**. **(b)** As (a), but with expression values as measured with microarrays in the Choy *et al.* study[7] (x-axis) and sequencing in the Pickrell *et al.* study[1] (y-axis). The coefficient of variation estimates from sequencing are again larger than estimates from microarrays. **(c)** A plot of the standard deviation of expression values as measured with microarrays in the Stranger *et al.* study[9] (x-axis) and sequencing in the Montgomery *et al.* study[2] (y-axis), where only reads overlapping exons containing microarray probes are used to measure expression. The estimates of expression variability from sequencing are similar to the estimates from microarrays. **(d)** As c but with expression values as measured with microarrays in the Choy *et al.* study[7] (x-axis) and the Pickrell *et al.* study[1] (y-axis). The estimates of expression variability from sequencing are again almost the same as estimates from microarrays. In all four panels, the two higlighted genes are the same as in **Figure 1**.



Supplementary Figure 2 Technical vs. biological variability in microarray and sequencing experiments. **(a)** A plot of the centered mean expression for technical replicates of the two genes *COX4NB* and *RASGRP1* as measured with sequencing (top row) and microarrays (bottom row). These are the same two genes depicted as in **Figure 1c**, but unlike **Figure 1c** it is not the same samples assayed by microarrays and sequencing. **(b)** A histogram of the estimated proportion of variance attributed to biology for sequencing (blue) and microarrays (red). More genes have a high proportion of variability attributable to biology in the sequencing experiments.