RESEARCH ARTICLE

A reanalysis of mouse ENCODE comparative gene expression data [v1; ref status: indexed, http://f1000r.es/5ez]

Yoav Gilad, Orna Mizrahi-Man

Department of Human Genetics, University of Chicago, Chicago, IL, 60637, USA

v1 **First published:** 19 May 2015, **4**:121 (doi: 10.12688/f1000research.6536.1)
Latest published: 19 May 2015, **4**:121 (doi: 10.12688/f1000research.6536.1)

Abstract

Recently, the Mouse ENCODE Consortium reported that comparative gene expression data from human and mouse tend to cluster more by species rather than by tissue. This observation was surprising, as it contradicted much of the comparative gene regulatory data collected previously, as well as the common notion that major developmental pathways are highly conserved across a wide range of species, in particular across mammals. Here we show that the Mouse ENCODE gene expression data were collected using a flawed study design, which confounded sequencing batch (namely, the assignment of samples to sequencing flowcells and lanes) with species. When we account for the batch effect, the corrected comparative gene expression data from human and mouse tend to cluster by tissue, not by species.

Open Peer Review

Referee Status: ? ✓ ✓ ✓

	Invited Referees 1	2	3	4
version 1 published 19 May 2015	? report	✓ report	✓ report	✓ report

1 **Rafael Irizarry**, Harvard School of Public Health USA

2 **Michael Eisen**, University of California, Berkeley USA

3 **Mick Watson**, University of Edinburgh UK

4 **Lior Pachter**, University of California, Berkeley USA

Discuss this article

Comments (3)

Corresponding author: Yoav Gilad (gilad@uchicago.edu)

How to cite this article: Gilad Y and Mizrahi-Man O. **A reanalysis of mouse ENCODE comparative gene expression data [v1; ref status: indexed, http://f1000r.es/5ez]** *F1000Research* 2015, **4**:121 (doi: 10.12688/f1000research.6536.1)

Copyright: © 2015 Gilad Y and Mizrahi-Man O. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Grant information: This work was supported by NIH grant MH084703.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have no conflicts of interest or competing interests to disclose.

First published: 19 May 2015, **4**:121 (doi: 10.12688/f1000research.6536.1)
First indexed: 25 Jun 2015, **4**:121 (doi: 10.12688/f1000research.6536.1)

Introduction

The mouse ENCODE Consortium has collected multiple types of genomic and functional data in order to better understand the potential utility of the mouse as a model system for biomedical research. To study gene expression levels, the Consortium collected RNA sequencing data from multiple tissues from human and mouse. Their comparative analysis revealed that gene expression patterns tend to support clustering of the data by species, rather than by tissue (Figure 2a in reference 1).

This pattern was confirmed and discussed in greater detail in a companion paper by Lin *et al.*[2], which also acknowledged that this observation is somewhat unexpected. Indeed, previous comparative studies reported that gene expression data from human and mouse (and across other species more generally) tend to cluster by tissues, not by species. Lin *et al.* proposed that previous studies might have been biased in their focus on a few 'specialized' tissues that tend to express the largest number of 'tissue-specific genes', while the overall pattern supports less tissue specificity.

The implications of the observation that human and mouse gene expression data may be clustering by species more than by tissues can be profound. To a large degree, modern biology is built upon the empirical observation that homologous gene regulatory networks establish the identities of homologous cell-types, tissues, and organs across species – the results of Lin *et al.,* if true, challenge these observations and the biological basis of homology. From a more practical perspective, the mouse is arguably the most important animal model for biomedical research. If gene regulation in any mouse tissue is markedly more representative of a general mouse regulatory network than the regulatory network of a corresponding human tissue, this would call into question the utility of the mouse, and perhaps any other non-human animal, as a useful model system for biomedical research.

Here, we present a reanalysis of the mouse ENCODE Consortium comparative RNA sequencing data. We argue that a flaw in their study design raises doubt regarding their conclusions.

Methods

RNA-Seq data, genome and gene annotation files

In December 2014 we asked and were kindly provided by the authors of Lin *et al.*[2] the names of the sequence files used in their comparative analysis. Based on this information we obtained sequence files in FASTQ format (Supplementary Table 1) from the ENCODE project[1] site (https://www.encodeproject.org/; some of the files were only available from early January 2015).

For our analysis, we used the same genome build and gene annotation files as in Lin *et al.*[2]. The ENSEMBL[3] genome build *Mus musculus* GRCm38.68 was downloaded from ftp://ftp.ensembl.org/pub/release-68/fasta/mus_musculus/dna/Mus_musculus.GRCm38.68.dna_sm.toplevel.fa.gz; the corresponding transcript annotation file was downloaded from ftp://ftp.ensembl.org/pub/release-68/gtf/mus_musculus/Mus_musculus.GRCm38.68.gtf.gz. The *Homo sapiens* genome build provided by ENSEMBL[3] contains haplotypic regions that are not part of the primary assembly. To avoid these regions, genome build *Homo sapiens* GRCh37 was downloaded from the Illumina iGenomes page: (http://support.illumina.com/sequencing/sequencing_software/igenome.html). The GENCODE[4] Release 14 transcript annotation file for human was downloaded from ftp://ftp.sanger.ac.uk/pub/gencode/release_14/gencode.v14.annotation.gtf.gz. The chromosome names in the GENCODE gtf file did not match those in the genome sequence file, and were thus modified.

Sequencing study design

Based on the sequence identifiers found in the FASTQ files, we reconstructed the sequencing study design used to collect the gene expression data in Lin *et al.*[2]. The sequence identifier line in a FASTQ file generated from an Illumina sequencing run can take two formats, depending on the version of the Consensus Assessment of Sequence and Variation (CASAVA) pipeline used to generate it. Prior to version 1.8 of this pipeline the sequence identifier line was of the following format (CASAVA v1.7 user guide p.88; downloaded from: http://support.illumina.com/downloads/casava_software_version_17_user_guide_(15011196_a).html

@<machine_id>:<lane>:<tile>:<x_coord>:<y_coord>#<index>/<read_#>

Starting from version 1.8 the sequence identifier line is of the format http://support.illumina.com/help/SequencingAnalysisWorkflow/Content/Vault/Informatics/Sequencing_Analysis/CASAVA/swSEQ_mCA_FASTQFiles.htm

@<machine_id>:<runnumber>:<flowcellID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is filtered>:<control number>:<index sequence>

Below is a sequence identifier line from the mouse pancreas read1 FASTQ file (sequence identifier lines from the remaining FASTQ files were of similar format):

@D4LHBFN1:276:C2HKJACXX:4:1101:3448:12374 1:N:0:AGTTCC

Based on this information we inferred that the FASTQ files were generated by CASAVA version 1.8 or higher. Thus, we could extract from the sequence identifiers the following details that pertain to the sequencing study design: machine identifier, run number, flowcell identifier, and flowcell lane number. We found that the sequencing was performed in five batches, each consisting of a multiplexed single run on a single lane on one of four sequencers (Figure 1; note that two of the batches, composed of human samples only, differed only in their lane number). The design was such that only one batch contained samples from both species. The remaining four batches could be divided into pairs where each of the two batches had a nearly identical tissue composition, but a different species.

Ortholog annotation

Following Lin *et al.*[2], we used the protein-coding ortholog list generated by the modENCODE and mouse ENCODE consortia[5]. A file containing all orthologs from human, mouse, fly and worm was downloaded from http://compbio.mit.edu/modencode/orthologs/modencode.common.orth.txt.gz. From this list we extracted 14,744 human-mouse one-to-one ortholog pairs, for which both members were included in the transcript annotation files we used. We note that this number is lower than the ~15,106 ortholog pairs reported in Lin *et al.* We are not certain of the meaning of the '~' in the report

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

Figure 1. Study design. Sequencing batches as inferred based on the sequence identifiers of the RNA-Seq reads.

of the number of ortholog pairs analyzed by Lin *et al.* Nevertheless, we believe that a possible explanation for this disparity is a parsing error. The last two columns of the 'modENCODE ortholog file' represent the number of genes from each species in the ortholog group. One of the steps required to obtain the subset of ortholog groups for analysis is to select those records where the two last columns have a value of 1 (i.e. one-to-one ortholog pairs). We found that if this selection is done through a command line search that does not require that the value in the last column be exactly "1", but rather just begins with "1", then the result is 15,104 putative human-mouse ortholog pairs.

Quality assessment of RNA-Seq data

We used the FastQC software v0.10.0 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess the quality of the individual FASTQ files (Supplementary Table 2–Supplementary Table 6). We were concerned by evidence for GC content bias and overrepresented sequences. To examine the latter in greater detail, we mapped the sequences overrepresented in at least one sample to the genome of the respective species, using BLAT searches[6] against the hg19 (human) and mm10 (mouse) assemblies at the UCSC genome browser site (http://genome.ucsc.edu/)[6]. We found that in both species many of the overrepresented sequences mapped perfectly to the mitochondrial genome (Supplementary Table 3–Supplementary Table 6). For the mouse pancreas sample only, we also found many overrepresented sequences mapped to regions with rRNA repeats from the SSU-rRNA_Hsa and LSU-rRNA_Hsa families.

Mapping RNA-Seq reads to genome sequences

We mapped the RNA-Seq reads to their respective genomes using Tophat v2.0.11[7] with the following options: "--mate-inner-dist 200" (i.e. inner mate distance is 200nt, based on paired-end reads with length 100nt each and an insert size of 350-450nt); "--bowtie-n" (i.e. the "-n" option will be used in Bowtie[8] in the initial read mapping stage); "-g 1" (i.e. multi-mapping reads will be excluded from alignment); "-m 1" (i.e. one mismatch is allowed in the anchor region of a spliced alignment); "--library-type fr-firststrand" (the libraries had been constructed using the Illumina TruSeq Stranded mRNA LT Sample Prep Kit[2]). An exception was the mouse pancreas sample, for which the mapping process stalled consistently at the same stage. For this sample we used Tophat v1.4.1[8] with the same options as above. Tophat requires a Bowtie[8] index. For human we used the Bowtie index that was packaged with the genome sequence in the file downloaded from the Illumina iGenomes page (http://support.illumina.com/sequencing/sequencing_software/igenome.html). For mouse we built an index using the bowtie-build utility from Bowtie v2.2.1 (v 0.12.7 for the index used with Tophat v1.4.1).

Calculating gene GC content

For each of the two species we used the appropriate GTF file to generate a table, which contains for each gene its ENSEMBL gene identifier its common name, and the GC content of the sequence covered by the union of the gene's transcripts. To this end, we first generated a GTF file where overlapping exons from different transcripts of the same gene were merged into a single "exon" with the same sequence coverage, retaining the association with the gene identifier. Next, we computed the nucleotide content of the exons in this new GTF file using the 'nuc' utility from bedtools v2.17.0[9]. Finally, we computed the GC content for each gene identifier by summing the number of 'G' and 'C' nucleotides in its merged exons and dividing by the sum of counts of unambiguous nucleotides in these exons.

Per-gene FPKM values

We used Cufflinks v2.2.1[10] to compute fragments per kilo base of transcript per million (FPKM) values and aggregate them per gene. The only option used was "--library-type fr-firststrand". For the required transcript annotation file ("-G" parameter) we used the GTF file for the respective species described in the "Genome and gene annotation files" section. We then generated a matrix of 14,744 by 26 FPKM values for each gene (in the ortholog table) and sample. While generating this table we noticed that some of the common gene names were associated with more than one ENSEMBL gene identifier. In some cases we determined that this was due to gene identifiers that have been retired from the ENSEMBL database[3] but were retained in the GTF file (27 and 64 retired identifiers for human and mouse, respectively). These retired identifiers were ignored when constructing the FPKM matrix. For the remaining such cases we incorporated the value from the first appearance of the common name.

Per-gene raw fragment counts

To compute per gene raw counts from the alignment files produced by Tophat[7], we used the program featureCounts v1.4.4[11] with the respective species' GTF file specified in the "Genome and gene annotation files" section. For all runs we used the following options: "-p" - indicates that fragments rather than reads should be counted; "-C" - indicates that chimeric fragments will not be included in the summarization process; and "-s 2" - indicates that the paired-ends are reversely stranded. We next generated a matrix of 14,744 by 26 raw counts for each gene (in the ortholog table) and sample. Since the output from featureCounts identifies genes by their gene identifier (the ENSEMBL identifier in our case), whereas the ortholog table uses the gene's common name to identify it, we used the GC content table, which contains both these identifiers to match counts to the correct row in the ortholog table. As we did when generating the FPKM matrix, we ignored the values from retired ENSEMBL identifiers, and if there were still multiple identifiers for the same common name, we used the value from the identifier that appeared first.

Results

In this reanalysis effort, we focused solely on the RNA sequencing data that can be mapped to coding regions. Lin *et al.*[2] reported additional results, related to data on the expression of non-coding transcripts and histone marks. We did not reanalyze these additional data types.

Lin *et al.*[2] analyzed both previously published and newly collected human and mouse gene expression data. The previously published data consist of RNA sequencing from ENCODE, the Illumina Human BodyMap 2.0, and the Roadmap Epigenomics Mapping Consortium. In these previously collected data sets, human and mouse samples were analyzed by different labs at different times, such that there is a clear batch effect that is confounded with species. Lin *et al.*[2] clearly explains this limitation of the previously published data. They state that in order to address this issue they focus on the analysis of only the newly collected data – RNA sequencing data of samples from 13 human and mouse tissues that were collected by the same lab, using the same sample processing protocol. We focus our reanalysis study on the same newly collected data set (see Methods).

Replication of sample clustering by species

As a first step of our study we set out to replicate the analysis of Lin *et al.*[2]. To do so, we started with the matrix of FPKM values (computed, using Cufflinks[10], based on the read alignments to the genome). This analysis was done within R environment v 3.1.3 GUI 1.65 Snow Leopard build (6912)[12]. See Supplementary Text 1 for detailed commands, and a supplement zip file for the R input (available in Zenodo: http://dx.doi.org/10.5281/zenodo.17606).

We $\log_2$-transformed the FPKM matrix (after adding 1 to avoid undefined values). To visualize the data, we used an approach that is similar in principle to that used by the ENCODE mouse consortium and Lin *et al.* Specifically, we used the function 'prcomp' (with the 'scale' and 'center' options set to TRUE) to perform principal component analysis (PCA) of the transposed FPKM matrix (so that samples were now in rows and genes in columns), after removal of invariant columns (genes). Scatter plots of the PCA results were generated using the ggplot2 package[13]. In agreement with the findings of Lin *et al.*[2] the samples cluster mostly by species (Figures 2a, Figure S1 and Figure S2). We also plotted the heatmap of the matrix of Pearson correlations between the 26 samples, using the pheatmap function from the pheatmap package v1.0.2[14] with default settings (i.e. complete linkage hierarchical clustering using the Euclidean distances). Again, samples from the same species tend to cluster together (Figure 2b).

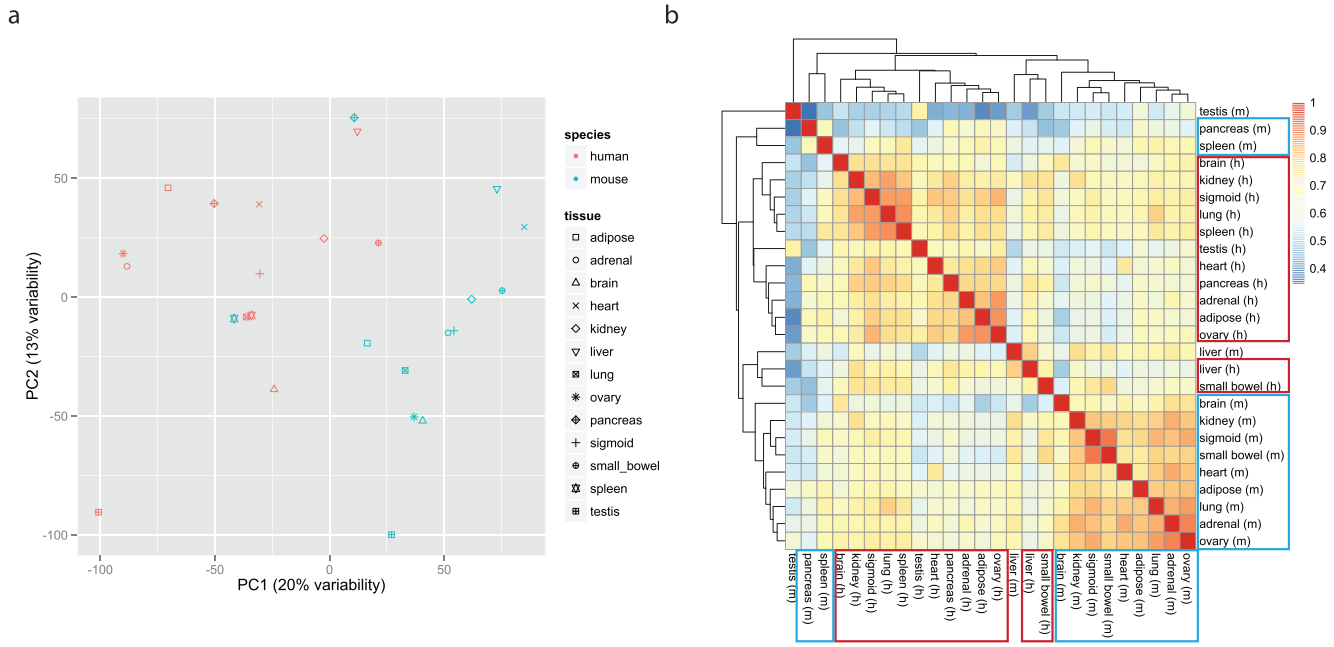

Figure 2. Recapitulating the patterns reported by the mouse ENCODE papers. a. Two-dimensional plots of principal components calculated by performing PCA of the transposed log-transformed FPKM values (from 14,744 orthologous gene pairs) for the 26 samples, after removal of invariant columns (genes). **b**. Heatmap based on pairwise Pearson correlation of expression data used in panel **a**. We used Euclidean distance and complete linkage as distance measure and clustering method, respectively.

Analysis of normalized data after accounting for batch effects yields clustering by tissue

A previous evaluation of normalization methods for RNA-Seq data[15] suggested that FPKM values were not optimal for clustering analysis. Therefore, as a basis for our reanalysis, we used the matrix of per-gene raw fragment counts. The entire analysis was done within R environment v 3.1.3 GUI 1.65 Snow Leopard build (6912)[12]. See Supplementary Text 2 for detailed commands, and a supplement zip file for the R input (available in Zenodo: http://dx.doi.org/10.5281/zenodo.17606).

Following Li *et al.*[16], we removed the 30% of genes with the lowest expression as determined by the sum of fragment counts across all samples. Next, due to the presence of mitochondrial genes among the overrepresented sequences in the data, we also removed reads that map to the 12 mitochondrial genes. This left us with expression data from 10,309 genes for analysis. We note that merely limiting the analysis to this subset of genes does not have a marked effect on the patterns reported by Lin *et al.* (Figure S3; detailed commands in Supplementary Text 3, and a supplement zip file for the R input (available in Zenodo: http://dx.doi.org/10.5281/zenodo.17606)). We performed within-column normalization to remove the GC bias in the data, indicated by the initial quality assessment. To this end, we applied the 'withinLaneNormalization' function from the EDASeq package v2.0.0[17] to each column in the matrix, using the gene GC values for the species associated with the column. Next, we used the 'calcNormFactors' from the edgeR package v3.8.6[18], with the trimmed mean of M-values (TMM) method[19], to calculate normalization factors for the library sizes for the samples. We used these normalization factors in the depth normalization of the columns (using the column sums of the original, unfiltered, counts matrix as a proxy for library sizes). The normalized data were $\log_2$-transformed (after adding '1' to each value in the matrix to avoid undefined values).

We then considered how to account for the fact that the assignment of samples to sequencing flowcells and lanes was nearly completely confounded with the species annotations of the samples (Figure 1). The consideration of 'batch effect' was the most important difference between the analysis that recapitulated the patterns reported by the mouse ENCODE papers (the previous 'Results' section) and the current reanalysis effort. Specifically, we accounted for the sequencing study design batch effects using the 'ComBat' function from the sva package v3.12.0[20], with a model that includes effects for batch, species and tissue. For this purpose the samples were classified into five batches, based on the sequencing study design (see methods and Figure 1).

To visualize the data, we used the function 'prcomp' (with the 'scale' and 'center' options set to TRUE) to perform principal component analysis (PCA) of the transposed log-transformed matrix of 'clean' values (after removal of invariant columns, i.e. genes), and the ggplot2 package[13] to generate scatter plots of the PCA results. None of the first five principal components (accounting together for 56% of the variability in the data) support the clustering of the gene expression data by species (Figure 3a and Figure S4–Figure S5). However, the sixth principal component, which accounts for 6% of

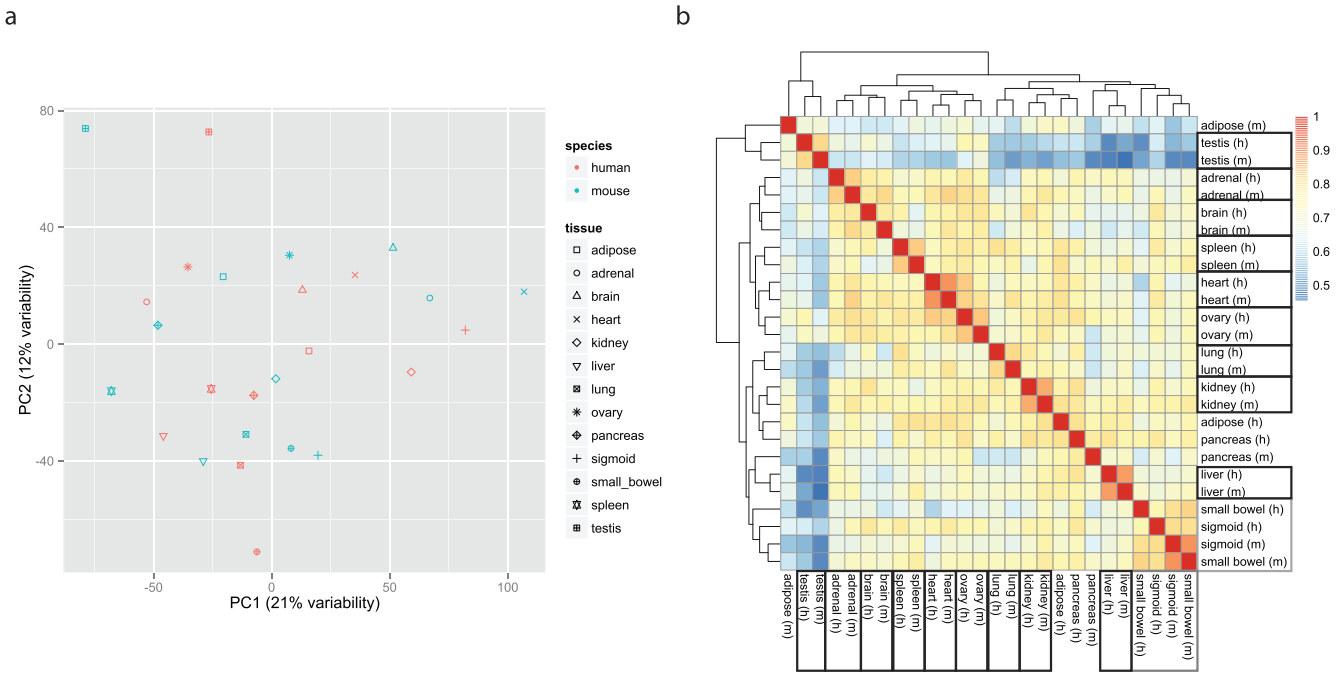

Figure 3. Clustering of data once batch effects are accounted for. a. Two-dimensional plots of principal components calculated by applying PCA to the transposed matrix of batch-corrected log-transformed normalized fragment counts (from 10,309 orthologous gene pairs that remained after the exclusion steps described in the results) for the 26 samples, after removal of invariant columns (genes). **b**. Heatmap based on pairwise Pearson correlation of the expression data used in panel **a**. We used Euclidean distance and complete linkage as distance measure and clustering method, respectively.

the variability in the data, does support such a clustering, suggesting that even though the 'species' and 'batch' variables are confounded, accounting for 'batch' does not remove completely the variability due to 'species' (Figure S5). We also plotted a heatmap of the matrix of Pearson correlations between the 26 samples, using the pheatmap function from the pheatmap package v1.0.2[14] with default settings (i.e. complete linkage hierarchical clustering using the Euclidean distances). This time the heatmap shows considerable clustering of the comparaive gene expression data by tissue (Figure 3b).

Discussion

In our reanalysis we have made a number of specific choices, including the exclusion of a certain subset of lowly expressed genes, the specific approach we chose to summarize the count data, the standardization and normalization methods we used (for example, we chose to standardize by the total count of reads that mapped to the ortholog gene pairs), the approach we used to account for the GC content bias, and the method we used to account for the sequencing design batch effect. Moreover, we excluded the sequencing data from 12 mitochondrial genes from both species, a step that – to the best of our ability to determine – was not taken by the original studies. In addition, our definition of ortholog gene pairs differs slightly from that of the original study, as we discussed in the methods. In practice, only the correction for the sequencing design batch effect had a drastic impact on the results. For example, without accounting for batch, using per-gene raw fragment counts instead of FPKM values does not seem to impact the degree to which the uncorrected data support clustering by species (Figure S6).

Visualizing or plotting the data is another important area where different choices can sometime lead to quite distinct conclusions. We chose to display, in addition to the PCA plots, heatmaps based on the correlations among the samples. We note that if the actual data (not pairwise correlations) are clustered, the observed patterns (by species or by tissues, in the respective analyses), seem practically identical (Figure S7). The heatmaps shown in the main figures are based on Pearson pairwise correlations, which provide the highest level of clustering by tissue in the analysis that takes into account batch effects. Alternative heatmap plots based on either Spearman pairwise correlations or other distance measures and clustering methods look similar in principle (Figure S8 to Figure S10), but the clustering by tissue is somewhat less pronounced (clustering by species, when batch is not accounted for, is more pronounced).

It is important to note that most of the analysis and plotting decisions we have made contributed to a somewhat better clustering of the expression data by tissue, both visually and empirically. We have made these – mostly standard - analysis and plotting choices regardless of the end result (namely, we believe that these are objectively reasonable choices). Importantly, we made identical choices for the clustering analysis and plot types for the data with and without batch correction, and our conclusions are robust with respect to a wide range of possible alternative approaches (Figure S7–Figure S10).

That said, we do acknowledge that we find the clustering of the data by tissue to be a more intuitive pattern. In other words, we believe that the clustering of comparative gene expression data by species – a result that contradicts previous observations – is a surprising outcome. Hence, we would have intuitively accepted as more correct most reasonable choices of analysis pipelines and data visualizations that supported a greater degree of clustering by tissue.

As we mentioned above, most of the choices we made resulted in little difference to the overall pattern. It was only the correction for the sequencing design batch effect that had a profound impact. Once we accounted for the batch effect by using ComBat, the comparative gene expression data no longer clustered by species, and instead, we observed a clear tendency for clustering by tissue. This is not surprising, as the sequencing batch, which we corrected for, was nearly entirely confounded with species. It stands to reason that some individual gene expression levels do cluster by species and some by tissue (see for example, Figure S5). While previous data sets strongly support a general clustering of gene regulatory phenotypes by tissue[21], we expect the degree of clustering of the gene expression data to differ somewhat across tissues. Yet, in this particular case, by removing the confounding sequencing batch effect we also removed most of the species effect on gene expression levels (a similar case of confounding batch and main effect of interest was discussed a few years ago, with respect to gene expression differences between human populations[22]).

One could potentially employ more sophisticated modeling approaches to try and estimate separately the batch and species effects. One idea would be to rely on the fact that there are five sequencing batches, but only two species. This, however, is complicated by the fact that the two sequence batches specific to the human samples share the same run and flowcell (potentially a smaller batch effect), while the two sequence batches specific to the mouse samples extend over different instruments (potentially a larger batch effect). In any case, we feel that such modeling is beyond the scope of this reanalysis effort. Instead, we conclude that the study design used by the mouse ENCODE consortium was flawed with respect to the questions they set out to address.

In summary, we believe that our reanalysis indicates that the conclusions of the Mouse ENCODE Consortium papers pertaining to the clustering of the comparative gene expression data are unwarranted. In the narrow context of our reanalysis effort, we state that their conclusions are unwarranted, not wrong, because the study design was simply not suitable for addressing the question of 'tissue' vs. 'species' clustering of the gene expression data. That said, a large body of independent previous work supports general clustering of comparative gene expression data by tissue.

Finally, we note that in this reanalysis effort, we have only focused on the RNA sequencing data collected by the mouse ENCODE consortium. We have not considered information with respect to the study design used to collect the many other types of data reported by this consortium. Given our findings, we believe that it is appropriate to call for a careful review of these other data sets as well.

Data availability

All data are available from the Mouse ENCODE consortium; see Table S1 for specific source URLs and accession numbers.

Software availability

We provide supplementary files of the python codes used to process and prepare the data for analysis with R, and the data files for the python codes. We also provide the R codes we used to perform the different analyses as supplementary files, as well as the input for the R codes.

Archived software files as at the time of publication

Zenodo. Data files and codes used in the reanalysis of the mouse encode comparative gene expression data. DOI: 10.5281/zenodo.17606

License

These codes are provided under the MIT license.

Competing interests

The authors have no conflicts of interest or competing interests to disclose.

Grant information

This work was supported by NIH grant MH084703.

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgment

We thank J. Lieb, V. Lynch, J. Novembre, L. Pachter, D. Graur, M. Stephens, N. Banovich, and J. Leek, for comments on the manuscript. We also thank 5 additional colleagues, who asked us not to reveal their names, for their valuable comments.

Supplementary tables

Table S1. Source of RNA-Seq data.

Table S2. Summary of test scores for the 52 FASTQ files analyzed.

Table S3. Overrepresented sequences in read1 human files.

Table S4. Overrepresented sequences in read2 human files.

Table S5. Overrepresented sequences in read1 mouse files.

Table S6. Overrepresented sequences in read2 mouse files.

Supplementary figures

https://f1000researchdata.s3.amazonaws.com/supplementary/6536/43196de3-53a7-4a8f-9093-71ad32f461e4.pdf

References

1. Yue F, Cheng Y, Breschi A, *et al.*: **A comparative encyclopedia of DNA elements in the mouse genome.** *Nature.* 2014; **515**(7527): 355–364.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Lin S, Lin Y, Nery JR, *et al.*: **Comparison of the transcriptional landscapes between human and mouse tissues.** *Proc Natl Acad Sci U S A.* 2014; **111**(48): 17224–17229.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Cunningham F, Amode MR, Barrell D, *et al.*: **Ensembl 2015.** *Nucleic Acids Res.* 2015; **43**(Database issue): D662–669.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Harrow J, Frankish A, Gonzalez JM, *et al.*: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res.* 2012; **22**(9): 1760–1774.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Wu YC, Bansal MS, Rasmussen MD, *et al.*: **Phylogenetic Identification and Functional Characterization of Orthologs and Paralogs across Human, Mouse, Fly, and Worm.** *bioRxiv.* 2014.
Publisher Full Text
6. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res.* 2002; **12**(4): 656–664.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Kim D, Pertea G, Trapnell C, *et al.*: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol.* 2013; **14**(4): R36.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics.* 2009; **25**(9): 1105–1111.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Trapnell C, Roberts A, Goff L, *et al.*: **Differential gene and transcript expression**

analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 2012; **7**(3): 562–578.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics.* 2014; **30**(7): 923–930.
PubMed Abstract | Publisher Full Text

12. Team RC: **R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing. 2015.

13. Wickham H: **ggplot2: elegant graphics for data analysis.** Springer, New York, 2009.
Publisher Full Text

14. Kolde R: **pheatmap: Pretty Heatmaps.** R package version 1.0.2 ed. 2015.
Reference Source

15. Dillies MA, Rau A, Aubert J, *et al.*: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.** *Brief Bioinform.* 2013; **14**(6): 671–683.
PubMed Abstract | Publisher Full Text

16. Li S, Labaj PP, Zumbo P, *et al.*: **Detecting and correcting systematic variation in large-scale RNA sequencing data.** *Nat Biotechnol.* 2014; **32**(9): 888–895.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Risso D, Schwartz K, Sherlock G, *et al.*: **GC-content normalization for RNA-Seq data.** *BMC Bioinformatics.* 2011; **12**: 480.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1): 139–140.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol.* 2010; **11**(3): R25.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Leek JT, Johnson WE, Parker HS, *et al.*: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments.** *Bioinformatics.* 2012; **28**(6): 882–883.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Chan ET, Quon GT, Chua G, *et al.*: **Conservation of core gene expression in vertebrate tissues.** *J Biol.* 2009; **8**(3): 33.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Akey JM, Biswas S, Leek JT, *et al.*: **On the design and analysis of gene expression studies in human populations.** *Nat Genet.* 2007; **39**(7): 807–808; author reply 808–809.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Referee Status: 

Version 1

Referee Report 30 June 2015

doi:10.5256/f1000research.7019.r8710

Lior Pachter
Department of Mathematics, University of California, Berkeley, Berkeley, CA, USA

The article "A reanalysis of mouse ENCODE comparative gene expression data" by Gilad and Mizrahi-Man examines a claim, recently published in the pair of papers

1. Yue F, Cheng Y, Breschi A, *et al.*: A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014; 515(7527): 355–364.

2. Lin S, Lin Y, Nery JR, *et al.*: Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A. 2014; 111(48): 17224–17229.

that expression data from human and mouse cluster more by species than by tissue.

The Gilad--Mizrahi-Man paper consists of three "results":

1. A report of the experimental design in Lin *et al*.

2. An attempt to reproduce the results of Yue *et al.* and Lin *et al*. that pertain to the claim about species vs. tissue clustering of expression data.

3. A re-analysis of the Lin *et al.* data in a manner that addresses shortcomings in the original experimental design.

The first is the observation that Lin *et al*. improperly designed their experiment by confounding species with batches sequenced, thereby leading to a possible "batch effect" affecting their results. This observation was already published as a preprint by the first author on the pre-print server Twitter (see https://twitter.com/Y_Gilad/status/593088451462963202).

Having noted "a flaw in their [Lin *et al*.] study design" Gilad--Mizrahi-Man turn to the question of whether the flaw affected the conclusions in Yue *et al*. and Lin *et al*. about expression differences as pertaining to tissues vs. species. To this end, the authors attempted to reproduce the analysis of Lin *et al*.

It is evident that while the Lin *et al*. results may, in some technical sense, be "reproducible" they were certainly not "usable" as published. Gilad--Mizrahi-Man carefully expose a vast number of choices in software and processing options poorly described in Lin *et al*., and whose effect on the final result(s) is unclear. To quote just one example, they write that "An exception was the mouse pancreas sample, for which the mapping process stalled consistently at the same stage", a problem that led them to use TopHat v1.4.1 instead of TopHat v2.0.11. One may wonder whether software choices and other decisions in analysis affect final results, and Gilad--Mizrahi-Man address this question (although only partly).

For example, one fundamental analysis choice is whether to quantify abundances of genes by summing raw "fragment counts" from alignments to gene regions, or via the summing of abundances as quantified by probabilistic assignment of ambiguously mapped reads. Gilad--Mizrahi-Man cite a paper by Dillies *et al* . (and the French StatOmique Consortium) suggesting that "FPKM values were not optimal for clustering analysis" to argue for using "fragment counts". I strongly disagree with this choice because transcript abundances are necessary to accurately estimate gene-level abundances, a point that Dillies *et al.* fail to realize. As pointed out in my own paper on Cufflinks 2 (Trapnell *et al.* 2012) wrong does not cancel wrong for differential analysis, nor does it for the purpose of clustering.

In any case, Gilad--Mizrahi-Man do examine whether quantification by EM affects results and in a later statement they state that "using per-gene raw fragment counts instead of FPKM values does not seem to impact the degree to which the uncorrected data support clustering by species", a result summarized in their Figure S6. While I applaud them for checking the dependence of results on this choice, without further analysis the question remains of whether other analysis choices affect results (although to be fair to the authors, the number of tests that would have to be conducted is enormous and quite possibly practically intractable). Nevertheless, it would be interesting if, for the purpose of future transcriptomics analyses, Gilad--Mizrahi-Man were to investigate some key steps as to their effect (e.g. annotation, an issue discussed recently by Ongen and Dermitzakis, or mapper choice).

The final result of Gilad--Mizrahi-Man is a re-analysis of the Lin *et al*. data from which they observe that a basic correction for batch effect removes the strong clustering of expression profiles by species touted in Yue *et al*. and Lin *et al.* The question of the relative species/tissue contribution to expression profile is of course fundamental and interesting, and obviously further data, carefully curated and analyzed, will answer the question definitively. As far as the Lin *et al.* paper goes, the Gilad--Mizrahi-Man paper certainly casts doubt on the suitability of the data for answering the question. For one thing, the term "batch effect" is unfortunately rather generic and in this specific case that has become a problem. After initial posting of the preprint by Gilad on Twitter, the authors of Lin *et al*. resequenced their libraries in a different configuration, but further investigation of the experimental design by Gilad *et al*. (subsequent to initial posting of a preprint of the article I'm reviewing) appears to have revealed additional problems. For example, tissues in human and mouse were selected from males vs. females respectively (except in ovaries and testis) resulting in another potential bias that would skew expression profiles to cluster by species rather than tissue. In other words, with their paper and subsequent analysis Gilad and Mizrahi-Man have convinced me to be skeptical of the data and conclusions of Lin *et al*. (and insofar as it pertains to the results in Lin *et al*., the paper by Yue *et al*.).

But that is not really the point. What matters now are the carefully documented serious shortcomings in the computational and experimental methodology of Lin *et al*. and for this reason I have approved the Gilad--Mizrahi-Man manuscript. Hopefully the issues raised will be properly addressed by Lin *et al*. (in a manner equally rigorous to that of Gilad--Mizrahi-Man).

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 25 June 2015

doi:10.5256/f1000research.7019.r8832

Mick Watson
The Roslin Institute, University of Edinburgh, Edinburgh, UK

The study is carried out well and the results support the conclusion.

The paper would benefit from including some of the discussion points made in the comments made to v1 of the paper. Lin *et al.* have re-sequenced the samples and removed the sequence lane batch effect, and reproduced the same result; however, the samples themselves are confounded, in that they were treated differently prior to sequencing. This discussion should be added to the paper.

I would also like to see a discussion of artifacts which are discoverable within the data, for example:

- the human samples have significant numbers of rRNA reads compared to the mouse samples. This should not happen with mRNA-Seq which includes a polyA selection.
- the human samples have a hugely varied number of reads per sample, compared to the mouse samples
- one of the mouse samples has over 1.8M reads that map to a single rRNA transcript. This is an outlier for mouse, as the other mouse samples have low numbers of rRNA reads
- The mitochondrial genes are turned on in one species but not the other

These points are all indicators of different sample extraction techniques, which also confound with species.

The authors may also wish to discuss use of FPKM, which may not be the most useful measure of gene expression in this study, as the human and mouse orthologues have different lengths.

See https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: It may be apparent from the review, but we have also been re-analysing these data.

Referee Report 22 June 2015

doi:10.5256/f1000research.7019.r8942

Michael Eisen
Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, CA, USA

In this paper Gilad and Mizrahi-Man reanalyze a high-profile dataset from the ENCODE consortium that was used to argue that gene expression levels are more similar for different tissues from the same species than the same tissues from different species, a somewhat counterintuitive result that contradicts earlier claims (including those by Gilad).

The main result of this new work is a simple observation: in the original experiment described in Lin *et al*.,

samples from the same species were run in the same sequencing batch. Since there are well-known batch effects, this is poor experimental design that calls into question the claim by Lin *et al*. that data clusters by species, since the data could instead by clustering by sequencing batch.

There is always a bit of a challenge in figuring out what to do with an observation. As the authors point out, this aspect of the experimental design effectively renders the data useless for asking questions about the relative contribution of species and tissue to gene expression variation. Yet it would seem like too light of a paper to simply say "The original authors messed up. Their claims are therefore invalid. QED."

So this paper contains a few analyses designed to ILLUSTRATE the point they make. They are not exactly results, since, once you realize that batch and species are completely confounded, correcting for batch will inevitably remove the species signal. In this context, it's a bit weird to present such an analysis as if it is a result, but I don't really see a way around it, and the authors are forthright in pointing out that their main observation is that the data from Lin *et al* are useless for addressing this issue, and that they can make no claims, even after correcting for batch effects, about what the data actually do say.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 26 May 2015

doi:10.5256/f1000research.7019.r8732

Rafael Irizarry
Department of Biostatistics, Harvard School of Public Health, Boston, USA

In this PNAS paper is is found that the first three principal components obtained from mouse and human gene expression data correlate with species and not with tissue. This is interpreted to imply that "tissues appear more similar to one another within the same species than to the comparable organs of other species".

Gilad and Mizrahi-Man (the authors) downloaded all the data from this paper and reanalyzed it carefully. The majority of their F1000Research article is dedicated to describing, in full detail, how they analyzed the data. The choices made all seem sound and they are able to reproduce the figures of the original PNAS article.

An important discovery made and reported by the authors is that mice and human samples were run in different lanes or different instruments. The confounding was near perfect (see Figure 1). The authors then apply a linear model (ComBat) to account for the batch effect and find that, after the correction, samples cluster almost perfectly by tissue (see Figure 3). They conclude that "Once we accounted for the batch effect by using ComBat, the comparative gene expression data no longer clustered by species, and instead, we observed a clear tendency for clustering by tissue. This is not surprising, as the sequencing batch, which we corrected for, was nearly entirely confounded with species."

There are three issues I recommend the authors consider:

1. As the authors suggest, with the observed level of confounding, if there is in fact a species effect, applying an approach that models batch as a linear effect will also account for species. Although

pointing out that there is almost perfect confounding is an important contribution, I don't see why ComBat should be applied here. If a model that removes species is applied, it is no surprise that the data will no longer cluster by species.

2. As mentioned, with the existing study design it is impossible to completely tease out species from batch. However, there is a relatively simple data analysis that can be performed to explore the possibility that instrument or lane are a large enough source of variability to overcome the tissue effect, which is know to be large. The analysis is:

   i) perform the same PCA analysis on the mouse data and compare the two instruments and then

   ii) perform PCA analysis on the human data and compare the two lanes.

   If in fact lane and instrument are a large sources of variability we should see it here. Of course, there is still the possibility that the instruments used for humans was very different to the one used for mice, while the two instruments used for mice were similar. Due to confounding we won't know for sure, but the analysis described here will at least give us at least a lower bound on how large these effects can be.

3. There is a comment in the F1000Research article from the first author of the PNAS article describing a second experiment in which confounding with instrument or lane was not present. In this analysis species continues to be the first few PCs. In a second version of this article, the authors can perhaps comment on this, as well as some of the other comments that suggest other possible sources of variability that may be confounded with species.

As a final remark, I am interested in reading the authors/readers thoughts on the biological interpretations that are being assigned to mathematical (euclidean) distance. Specifically, what does the word "similar" mean exactly. I understand what means in mathematics, but I am not sure what it means in biology when points are log (FPKM + 1) values for thousands of genes.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 26 May 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

Dr. Irizarry,

Thank you for spending the time to provide a review of our work. We agree with you that given the study design used by the mouse ENCODE consortium, applying a batch correction is futile. Indeed, we explicitly explain that in our discussion (you referred to that section of the text in your review).

We further agree that it would be intellectually interesting to research the extent of the batch effect further – for example, by following your suggestion on how to test for the effect of instrument and lane.

However, we feel that this additional effort is beyond the scope of our study. The mouse ENCODE consortium papers did not discuss (or account for) the sequencing study design. **We spent considerable effort tracking the details that allowed us to reconstruct their design**. We pointed out in our paper that given this study design, the unusual biological result reported by the mouse ENCODE consortium might have a technical explanation. We believe it is the responsibility of the mouse ENCODE consortium authors to provide evidence that excludes this technical possibility, rather than us having to prove that it is indeed the likely explanation.

Which leads us to your third point: Indeed, the mouse ENCODE consortium authors commented that they have now collected additional sequence data, using a different design, and that their results held. In that sense, we believe that this means that the mouse ENCODE consortium authors accepted our claim that their original design was flawed.

Yet, as mentioned in a few other comments here, there is an additional technical batch effect that was not yet excluded – related to tissue extraction and sample preparation. We plan to discuss this additional technical batch effect in a revised version of the text (we will wait to see additional reviews before we provide a revised version of the paper).

Again, thank you for your time and thoughts.

Competing Interests: No competing interests were disclosed.

Discuss this Article

Version 1

Author Response 09 Jul 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

Anshul,

Thank you for your thoughtful comment (and for the discussion over Twitter). We appreciate the time and effort you put into this.

You raise several issues, but the most important one (we believe), is that neither the mouse ENCODE paper, nor ours, adequately addressed the question of 'species' vs. 'tissue' clustering. We agree with you!

Our paper, however, was ***not*** about this question. Our claim – mind you – is ***not*** that tissue (or species) cluster better, but rather that the conclusions of the mouse ENCODE paper are ***unwarranted***. This is an important point!

Our goal is not to prove anything other than raise doubt regarding the conclusions of the mouse ENCODE paper. Your concerns and the issues you have raised can be seen as additional reasons to doubt those conclusions. It is important to remember that the authors of the mouse ENCODE paper need to defend their conclusions, not us. Our 'conclusions' are merely that the study design and analysis of the mouse ENCODE paper are flawed.

Now more to the point: We agree that the results from PCA are not very insightful in this case. We provide PCA plots because it is the framework of analysis used by the mouse ENCODE authors. Our goal was to show that even using this framework, we couldn't recapitulate the (somewhat incoherent) basis for their biological conclusions. We could try to perform the analysis you suggested based on clustering of samples using the top N PCs, but would the results truly be more insightful? As you pointed out – and we agree - PCA is not really the right tool to answer these questions.

So we might turn to ANOVA. As you wrote, this is a much more intuitive tool with which to address the question. Yet, analyzing all genes together makes little sense, we - again - agree. It does not take into account differences between genes. In our mind this 'combined analysis' actually makes as much sense as using PCA… It provides some vague estimate of the proportion of variance explained by 'species' and 'tissue' for the entire data set. For whatever its worth, we find that species explains very little in this type of analysis (in the new data), while 'tissue' and 'tissue by species' explain a bit more. Note that in the slides we reported the % of 'explained variance' (in fact, most of the variance, in either the old or new data, remains unexplained). This may have been unclear or even a bit misleading and we will change it. Yet – again – there is certainly no evidence for robust clustering by 'species' (Actually, in either the 'old' or 'new' data).

What we really need, is a gene-based analysis. However, the study design makes it difficult to effectively use ANOVA in this case. The analysis by gene is hopeless. With exactly **one sample from each variety** (one sample from each tissue in each species), one can't estimate a gene-specific interaction effect, one can't effectively estimate the error term, and tissue and species are not always orthogonal either. Combine this with the fact that tissue samples came from different (and unbalanced) individuals, sexes, and ages, and it's a nightmare for analysis. For whatever its worth, we have performed this analysis anyway, ignoring all of the obvious caveats. We will add the slides to the ppt. You will see that, in the new data, a higher proportion of variation in many more genes is explained by 'tissue'.

To us, the most important aspect of your comment is that you raise additional questions about the analysis (both ours and the original mouse ENCODE paper). We believe that everyone will intuitively agree that 'some' genes are expressed more similarly across species and 'some' genes are expressed more similarly across tissues. The conclusion of the mouse ENCODE paper, however, was that "…in general, differences dominate similarities between the two species." (quote from the abstract of the mouse ENCODE PNAS paper).

Yet, the study design, as we stated multiple times, is flawed. One simply can't effectively address the question of the relative contribution of 'tissue' vs' 'species' using this flawed design. **That is our only conclusions.**

Do you not agree, based on our observations, this discussion, as well as your own concerns about the framework of the analysis, that the conclusion of the mouse ENCODE paper is unwarranted?

Competing Interests: No competing interests were disclosed.

Reader Comment 07 Jul 2015

Anshul Kundaje, Stanford University School of Medicine, USA

These comments are specifically wrt. to the slides that Yoav posted a few weeks ago on twitter goo.gl/YPNQ4H. The slides perform several interesting analyses on the "old" (sequencer confounded) and "new" (no sequencer confounding) data measuring RNA-seq expression in 12-13 mouse and human tissues. The main analyses include (i) some interesting PCA comparisons; (ii) clustering results on the samples ; (iii) An ANOVA based expression variance partitioning. Each of these analyses are performed on 3 different processed versions of the RNA-seq data.

There are 2 issues that I believe this analysis tries to address
(Qi) Is the new data different from the older data.
(Qii) Does tissue or species dominate the variation in new data

Regarding (Qi), I think there is no debate that the data are (and should be) different. I believe the PNAS paper authors also agree with this statement.
(Qii) is where I think there is still disagreement between the PNAS authors and Gilad et al.

We have atleast 3 main factors contributing to gene expression variation in this dataset - genes, species and tissues. And then several other confounding factors e.g. age, sex. Lets ignore these for now since the analyses in the slides don't directly model these.

(1) The PCA is decomposing the correlation/normalized covariance across the samples featurized by genes via an orthogonal transformation. What we obtain is a projection of the samples into a new orthogonal subspace of PCs (metagenes). Whats nice about PCA is that a small number of PCs (5-10) can potentially explain a large proportion(>80-90%) the covariance captured by 1000s of features (genes) in the original space. This dimensionality reduction allows (arguably) interpretable visualization of the samples relative to each other and can also potentially (not always) help get around the curse of dimensionality that plagues analysis of high-dimensional data (more on this in the next few paragraphs).

I very much like the visualization of the projection of the samples on to each PC as it gives a nice intuitive feel for what the PCs are doing. E.g. Slide 10 shows projections of the new samples on PC1. Several matched tissues from human and mouse project onto the same point (or very close to each other) on PC1 with a few notable exceptions i.e. testis, spleena and pancreas. So one could surmise that PC1 is representing some metagene signature that is largely tissue-associated but does have a species-related component as well (due to the exceptions). Slide 11 shows projections on PC2. Wrt. PC2 the samples from the same species can be seen to be closer to each other and one can obtain a clean separation of species. So PC2 can be interpreted to represent some metagene signature that is species-associated. This may lead one to conclude that since PC2 and not PC1 (which explains more variability than PC2) is more associated with separating species then maybe species < tissue. However, IMHO I don't think one can conclude that especially since PC1 and PC2 only capture 20% and 13% of the variability. So there is still a huge amount of variability to be explained. PC3 for example is also more species associated. PC4 is more tissue-associated and so on. This also makes it rather dangerous to visualize the samples using only the 2 or 3 PCs at a time as any conclusion from such a visualization is incomplete. If we really want to figure out whether samples from the same species or tissue are closer to each other (cluster together) in the subspace of PCs, we should be using the top N PCs that explain a large proportion (90-95%) of variability. Then compute explicit distances (euclidean distance would be very justifiable in this orthogonal subspace) between samples in this reduced subspace. One could then compute within species/tissue vs. between species/tissue distances or cluster the samples using the distance matrix computed in this space of PCs. Neither the PNAS paper nor Gilad *et al*. do this.

In slide 19 and 20, Gilad *et al*. present two types of clustering. In slide 19, they first compute Pearson

correlation between all pairs of samples using all genes. Then they use the correlation values of each sample with all other samples as new features to compute a euclidean distance between pairs of samples. They use complete linkage hierarchical clustering with this distance matrix to cluster the sames. This clustering shows the samples largely cluster by species with a few exceptions. In slide 20, they compute a distance matrix (correlation?) across samples using all genes and then use this directly as a distance matrix in complete linkage clustering. In this version of clustering, there is a sorta 50-50 split between species and tissues. IMHO, neither of these clustering approaches have a reasonable justification. There is significant literature explaining the curse of dimensionality in high-dimensional clustering and a distance measure based on 1000s of genes (with lots of correlated structure between them) is almost meaningless. This wikipedia article summarizes some of the issues https://en.wikipedia.org/wiki/Clustering_high-dimensional_data (Sorry I didn't have time to find a real reference!). Moreover, Pearson correlation is not really a distance metric (although it is routinely used in clustering) and more importantly it focuses very strongly on highly expressed genes. I think Rafa Irizzary has also highlighted this issue in his review. Lior Pachter has written about this issue in a blog post as well (https://liorpachter.wordpress.com/tag/mahalanobis-distance/) with several suggested alternatives such as the mahalanobis distance. As I mentioned in the previous paragraph, since we are in PCA land, why not cluster the samples projected onto the small set of PCs that explain most of the variability. The mahalanobis distance goes one set further adjusting for the variance explained by each PC. Either of these approaches would potentially get around the curse of dimensionality in that atleast the distances are meaningful. There are of course caveats here as as well. E.g. PCA projects the samples in a single subspace and clusters could exist in different subspaces. But it would certainly be better than clustering based on distances computed using all genes as features and directly link the clustering with the PCA.

Finally, personally I don't like the use of PCA (in the PNAS paper or this analysis) to answer the key question here - what is the relative contribution of species and tissue to variance of expression across all genes. It is incapable of explicitly answering this question.

Which brings me to the ANOVA analysis.

(2) The ANOVA analysis is attempting to directly partition the variance in the expression data as a function of species and tissue. However, the figures in the slides don't make much sense to me. Slide 18 would lead us to conclude that tissue accounts for 94.8% of the total variance. This to me is impossible. I am assuming here that the linear model being used is expression ~ species + tissue + species:tissue. Yoav hinted on twitter that the reason this looks odd is that the interaction term accounts for a significant proportion of variance. If that is the case, it indicates a dependence between species and tissue and one cannot simply look at the main effects to conclude species > tissue or not. Species may be heavily contributing through the interaction term.

However, I have a bigger issue with the ANOVA analysis. If I understood it correctly, this ANOVA model assumes a single linear model for expression variation across all 3 factors i.e. genes, species and tissues .. but uses only 2 of the factors in the model. It makes little sense to me to have a single linear model for all genes across all tissues and species with no term for gene identity. There is no way to explain differences between genes. It is also assuming fixed effects (I assume) which I don't believe is appropriate for this analysis. IMHO, a random effects model with an explicit term for gene identity is far more appropriate. An alternative is the variance partitioning analysis using random effects for species and tissues shown in Fig 2B. of the mouseENCODE Nature paper for each gene separately. If the analysis is on each gene separately then only having a species and tissue term is reasonable. But across all genes, I don't see how this ANOVA model makes sense. I may be misunderstanding something.

I look forward to your comments.

Competing Interests: I am part of the ENCODE consortium and a co-author on the mouseENCODE Nature paper (but not on the PNAS paper). Prof. Snyder and I are colleagues in the Genetics Department at Stanford University and active collaborators. However, these views are entirely my own and do not represent the views of any other members of consortia, organizations or papers that I am part of.

Author Response 25 Jun 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

For those with little time to read the entire comment, and those who are not invested enough in this area to study all the details of the papers and back and forth discussions, here is a quick summary:

- Mouse ENCODE data were collected using a flawed sequencing design. We uncovered the flaws and discussed the issue in our paper.
- In response, Mouse ENCODE collected new data from the same samples using a corrected sequencing design and reported that their conclusions are unchanged.
- However, we have now found a clear difference between the original and new data, which was not mentioned by the Mouse ENCODE authors.

Now in more detail:

Shin et al collected new sequencing data from the same samples they used in the original study. In contrast to the ***original*** gene expression data, the ***new data*** were collected using a sequencing study design that does not confound species with sample assignment to lanes and flowcells (see table included in the comment by Shin). The mouse ENCODE authors wrote that after analyzing the new data, they "*arrive at species-specific clustering as previously reported*." Yet, the figure provided by Shin is a 3D PCA plot, rotated in a way that makes it difficult to see PC1.

We have performed an analysis of the ***new data*** as well. Our findings are not consistent with the statement made by the mouse ENCODE authors.

Specifically, **we found that in the new data, gene expression levels in tissues from human and mouse cluster significantly better than in the old data**. There are a number of ways to show this, but the conclusion is robust. For example, if one chooses the approach used by the mouse ENCODE authors (using fpkm values and examining the clustering of samples using PCA plots), we can visualize the difference between the original and new sequencing data by considering the correlation between the PC1 values of the corresponding mouse and human tissue samples. I will post the figures on twitter (@Y_Gilad), but here are the numbers:

For the original data (where batch effect is confounded with species) the Pearson correlation between the PC1 values of the corresponding human and mouse tissues is 0.40 ($P = 0.18$).

For the new data (collected using a new sequencing design), the Pearson correlation between the PC1 values of the corresponding human and mouse tissues is 0.64 ($P = 0.02$). Moreover, two samples

(pancreas and spleen) have a number of severe and obvious QC issues (for example, the mapping % is much lower than in all other samples). If one excludes those samples, the Pearson correlation between the PC1 values of the corresponding human and mouse tissues in the new data is 0.89 ($P = 0.0006$).

As I mentioned, this finding - of differences in clustering properties between the original and new data - are robust with respect to choices in methodology and plotting. In other words, anyone who wishes to download and analyze the original and new data can easily observe the differences between the data sets (for example, whether one uses fpkm or raw counts, normalize the data in different ways or not at all, correct for GC content or not, and whether one plots PCA results or heatmaps, etc…).

The difference between the data sets is due to the effect of sequencing batch on the patterns observed with the original gene expression data. In the new data, species clustering is weaker and tissue clustering is stronger.

I have discussed the difference between the two data sets with Mike Snyder and all other co-authors of the Shin et al. paper before I posted this comment. I did this because it was strange to me that Shin did not comment on the difference between the data sets. Indeed, based on Shin's comment one might assume that Orna and I were wrong, and sequence batch does not actually have a noticeable effect on the data. Yet, as I have shown, the effect of sequence batch is evident by comparing clustering patterns between the original and new data.

By discussing this issue with the mouse ENCODE authors I have learned that they acknowledge that a difference between the two data sets exist, but they did not feel that it warranted an explicit discussion. The authors also declined to post an additional comment clarifying this. They wrote to me: "the point of collecting the new data was to see if the clustering conclusion still stands when samples from different species were mixed in the same lane and in our opinion it did."

We agreed to disagree.

Competing Interests: No competing interests were disclosed.

Author Response 08 Jun 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

It should be noted that on June 7th, ENCODE authors have fixed the sample annotation with respect to sex (details available on the ENCODE website, or on twitter...). Unfortunately the corrected study design still makes little sense with respect to the question of tissue/species. The corrected design results in a only partial confounding of tissue and sex, yet the authors did not comment on how sex was modeled or taken into account.

Competing Interests: No competing interests were disclosed.

Reader Comment 05 Jun 2015

Nicolas Robine, New York Genome Center, USA

We downloaded the original dataset supporting the claims in Shin et al (files listed here in Table S1) and processed them through our standard QC pipeline. This include mapping to the genome with STAR, quantifying genes with featureCounts (we use the gencode v18 GTF annotation) and computing a number of QC metrics and QC plots. One analysis we do is “Xist vs chrY” (inspired by ’t Hoen *et al.* 2013, Figure 2C). Xist is a female-specific long-non coding RNA (in mammals). For the vast majority of our samples, we can check pretty accurately if the samples are male OR female (high level of both XIST and chrY genes are usually indicative of sample pooling or sample mix-up, more often than chromosomal disorder of the individual).
Simply counting the number of reads mapping to XIST and the protein-coding genes from chromosome Y, we can infer the sex of the individual whose sample is from. (see attached table https://www.dropbox.com/s/v2vw1kbqv47x7ek/mouseENCODE_XIST_chrY_table.xlsx?dl=0)
We see that for 7 tissues (adipose, adrenal, brain, kidney, liver, sigmoid colon and spleen) the sex of the samples is not matching. It seems to me that in order to study how transcriptional landscapes compares between mouse and human tissues, one would want to reduce potential source of variability, such as male-female differences. Apparently, this has not been done here.

Competing Interests: No competing interests were disclosed.

Author Response 04 Jun 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

While the genders of the donors were not reported in the original paper, we found that information in the biosample file on the ENCODE site. From this file, we learned that other than for the ovary and testes samples, the gender of the human and mouse donors were different for all the other tissues. In other words, for 11 tissue types, every time a certain gender was sampled from mouse, the **opposite gender** was sampled from human. This is remarkable. The probability this would happen by chance is truly, very small indeed.

Competing Interests: No competing interests were disclosed.

Reader Comment 04 Jun 2015

Lenny Teytelman, Protocols.io, USA

As an advocate of post-publication peer review and discussion, I find this exchange between the Gilad and Snyder groups fascinating. It is precisely the kind of discussion that is illuminating and helpful for anyone following up on the work.

As someone not familiar with the prior literature, I also have a quick question regarding the prior studies. The abstract mentions "This observation was surprising, as it contradicted much of the comparative gene regulatory data collected previously." And in the introduction, "Indeed, previous comparative studies reported that gene expression data from human and mouse (and across other species more generally) tend to cluster by tissues, not by species." If possible, including references to these studies in the next version of this article would be useful.

Competing Interests: none

Reader Comment 03 Jun 2015

J Michael Cherry, Department of Genetics, Stanford University, USA

Data for the Lin *et al.* experiments were added to the ENCODE Portal last week. See http://goo.gl/C3yUwg . Use the 'Download' link to retrieve URLs to retreive metadata and data files.

We, the ENCODE DCC, are working to provide more transparency & data provenance, plus metadata for pipelines and software created by the consortium. For an example follow: https://www.encodeproject.org/experiments/ENCSR307BCA/

These are also available from our REST API for programmatic access: https://www.encodeproject.org/help/rest-api/.

Competing Interests: PI, ENCODE DCC

Author Response 01 Jun 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

Shin *et al.* wrote on May 21st that they collected new data, using a different sequencing design, and that the new data still support their original claim. Regardless of our notion that additional batch effects need to be addressed (see other comments), we would like to be able to analyze the new data as well. For one, we find it difficult to assess PC1 in the new figures submitted by Shin *et al.* (see their comment), as the 3D plot is rotated somewhat compared to the original figure in their paper. Unfortunately, though Shin *et al.* wrote that their new data will be made available on the ENCODE website, this is not the case; we are unable to find these new data.

Competing Interests: No competing interests were disclosed.

Reader Comment 30 May 2015

Mick Watson, The Roslin Institute, UK

In Lin et al, the paper mentions RNA-Seq carried out at both Stanford and Salk.

The 26 datasets analysed here (13 mouse and 13 human), listed in Table S1, are they from Stanford, Salk or both?

Competing Interests: No competing interests were disclosed.

Reader Comment 27 May 2015

Michele Busby, Broad Institute, USA

My last comment regarding the gene ontology analysis may not hold water. From the description of the gene ontology analysis in the original paper, it appears that they did not use a tool, like GOSeq, specifically

designed for RNA Seq data.

In RNA Seq, you are more likely to call genes that produce lots of reads (because they are longer or have higher expression) differentially expressed than genes that produce fewer reads, even if they have the same change in effect size, because they are measured with lower variance. You need to control for this in your enrichment analysis. It is not clear from the paper that this happened.

I apologize for missing this. It is a very common error and I should have checked for it before I wrote my comment.

Competing Interests: No competing interests were disclosed.

Reader Comment 26 May 2015

Christopher Mason, Department of Physiology and Biophysics, Weill Cornell Medical College, USA

Instead of just re-sequencing the same libraries, a better idea would be to try and ribo-deplete their original RNAs to remove the potentially confounding effect of RNA degradation on the samples. This may not remove *all* the batch effects from RNA extraction, isolation, and preparation, but it would surely help (as shown here: http://www.nature.com/nbt/journal/v32/n9/abs/nbt.2972.html). Also, when we ribo-depleted RNA across many tissues across 10 species of primates (http://nar.oxfordjournals.org/content/43/D1/D737.long), we saw the samples cluster by tissue, not species. To be safe, we often perform both polyA and ribo- on large sets of samples, and questions like these highlight why it is a good idea to do so. If not for all samples (since 2X the cost), at least for some to ensure robust signal.

Competing Interests: No competing interests.

Reader Comment 25 May 2015

David Lovell, QUT, Australia

Since this study makes use of correlation as a measure of association between the logged gene expression levels of different tissues, I think it's important to point out that correlation is not valid for data that carry only relative information.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., & Bähler, J. (2015). Proportionality: A Valid Alternative to Correlation for Relative Data. PLoS Comput Biol, 11(3), e1004075. http://doi.org/10.1371/journal.pcbi.1004075

Fortunately, it should be pretty straightforward to use proportionality in place of correlation and I'd be very happy to help with that.

Competing Interests: No competing interests were disclosed.

Reader Comment 22 May 2015

Uri David Akavia, Department of Biochemistry, McGill University, Canada

Dear Yoav,

If you look at Table 1 of the GTEx pilot article in Science, May 8th (DOI: 10.1126/science.1262110), you can see RIN values of the relevant human tissues. I'm guessing that these are the same samples used in Shin *et al*.

In that case - the RIN values are provided, but unfortunately the RIN values (as well as ischemia time) seem to indicate that Shin et al might be measuring the effect of RNA degradation over time.

If these aren't the RIN values, I would appreciate pointing them out.

Uri David Akavia
McGill University

Competing Interests: No competing interests were disclosed

Author Response 22 May 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

Shin and ENCODE co-authors,

You have made an **extraordinary** claim in your papers. It was extraordinary because it was counter-intuitive, because it challenged a strong paradigm in biomedical research (and more generally in modern biology), and mostly – because it contradicted a dozen or so previous studies that addressed an identical question.

You must therefore see why it is reasonable to hold you to a high standard and require that no other (e.g., technical) considerations can provide alternative explanation to your observed patterns. In the case of the sequencing study design, you have confounded species with sample assignments. You might have assumed that this type of batch effect is minor, but many others – including GTEx – have shown that it is actually quite considerable.

Thus, regardless of your assumptions (or intuition), the confounding sequencing batch effects could have potentially explained your original observations. I think that you have recognized that there is no way to argue against this rationale (based on the original data), which is why you sequenced the samples again, using a different study design.

The problem is that through the Twitter discussion, our colleagues raised an alternative, possibly even more significant, technical explanation for your observations. Based on the details you provide in your methods, it seems that the human and mouse samples were collected using quite distinct protocols. In addition, based on my own experience (and given the description of the tissue collection protocols), I am guessing that the quality of RNA extracted from the human tissues is significantly different from the RNA quality of the mouse tissues (you have not provided RIN data).

So, once again, we suspect a technical confounding batch effect that could potentially explain your observations. One cannot simply assume that batch effects do not exist, or that they are minor (without

explicitly testing for them).

Indeed, in our world as scientists, technical explanations must always be excluded or they remain reasonable alternatives. This is especially the case when one's observations contradict those of previous studies. In such cases, in particular, we have a strong prior to favor technical explanations for the discrepancy in observations. It is the authors' responsibility to exclude all of those possible alternative explanations by providing the relevant data.

Competing Interests: No competing interests were disclosed.

Reader Comment (*Member of the F1000 Faculty*) 22 May 2015

Steven Salzberg, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, USA

Shin Lin et al: your additional experiment does help, in that you seem to have controlled for machine effects this time, where (as you write) you "have re-generated in a single experimental batch and re-sequenced 24 of the original 26 tissues using the multiplexing scheme in Table 1." However (and unfortunately), as Yoav Gilad pointed out yesterday (https://twitter.com/Y_Gilad/status/601079582733815808), there is a much more profound batch effect that re-sequencing these same samples cannot remove. The human data was from recently deceased organ donors, preserved and handled very differently from the mouse data, which was from 10-week-old littermates. This could create a far stronger batch effect that would appear to be a species effect. E.g., the mice were very young, so you might be seeing a "young vs old tissue" effect. Or it could have to do with the greater degeneration of the human samples which had been preserved. And so on.

Competing Interests: None.

Author Response 21 May 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

Shin *et al*.,

Thank you for posting the new data and analyses. For completion of records, can you please provide additional details on the tissue collection protocols, RNA extraction protocols, and the RIN data for each sample? Thank you.

Competing Interests: No competing interests were disclosed.

Reader Comment 21 May 2015

Shin Lin, Department of Genetics, Stanford University, USA

We continue our comment from May 19, 2015, in which we agreed to show new data to the scientific community. We have re-generated in a single experimental batch and re-sequenced 24 of the original 26 tissues using the multiplexing scheme in Table 1 (

https://www.dropbox.com/s/dod8fcp2ds9zj52/table1part2.jpg?dl=0). In this design, lane/flow cell/sequencing machine effect (to be referred as "lane" effect henceforth for simplicity) can be separated from species effect. When we eliminate mitochondrial reads and quantile normalize by lane, we arrive at species-specific clustering (Figure 1 https://www.dropbox.com/s/sw9h0zajhmfw5uj/fig1part2.jpg?dl=0), as previously reported.[1] Thus, we emphatically disagree with the conclusion from Gilad and Mizrahi-Man that our conclusions are “not warranted,” but rather we argue that objective normalization procedures allow the discovery of the clustering of transcriptomes by species.

Gilad and Mizrahi-Man's work focused on one particular dataset in Lin et al.[1] However, that paper contains a principal component analysis (PCA) on data from multiple sources: Stanford (human, mouse), Salk (human), HBM (human), LICR (mouse), and CSHL (mouse). There are undoubtedly many technical differences between the various sources. Yet, the clustering by species was seen in higher order principal components (PCs) (see Figure 1A in Lin et al.); clustering by tissues, in lower components (Figure 1B in Lin et al.) or by normalizing species separately (Extended Data Fig. 1c of Yue et al.[2]). The same behavior is seen in the Stanford-only data—both in Lin et al., which minimizes primer index effect (Figure 1C & 1D in Lin et al.) and now the newly generated results correcting for lane effect (Figure 1A & B https://www.dropbox.com/s/sw9h0zajhmfw5uj/fig1part2.jpg?dl=0). The latter are consistent with our earlier observation that experimental batch did not drive the species-specific clustering. Finally, as additional supportive evidence, we have also examined related data from an independent study. Using Riken Fantom 5 CAGE data[3] from 12 matched primary mouse and human cells (and replicates) as described in the mouse ENCODE main paper Yue et al.[2] Supplementary Information, we again find species-specific clustering (Figure 2 https://www.dropbox.com/s/g487schbjoya6eg/fig2part2.jpg?dl=0).

The recognition of global differences between the human and mouse transcriptomes is consistent with the experiences that many experimentalists have using the mouse model. Given the substantial differences in size and metabolic rates, we do not feel it is implausible that there are strong expression differences reflected in basic metabolic and cellular processes at the organism level. Rather than questioning the utility of the mouse model, which will assuredly continue as an invaluable tool, we propose that a better understanding of these differences between human and mouse may allow us to better utilize this model system as it pertains to the investigation of human diseases.

(***Data mentioned herein will be available for download at the Mouse ENCODE website shortly.)

Shin Lin[1,2], Yiing Lin[3], Michael A. Beer[4], Thomas R. Gingeras[5], Joseph R. Ecker[6,7], Michael Snyder[1]

[1] Department of Genetics, Stanford University, 300 Pasteur Drive, M-344 Stanford, California 94305; [2] Division of Cardiovascular Medicine, Stanford University, Falk Building, 870 Quarry Road Stanford, California 94304; [3] Department of Surgery, Washington University School of Medicine, 660 S. Euclid Ave., Campus Box 8109, St. Louis, Missouri 63110; [4] McKusick-Nathans Institute of Genetic Medicine and the Department of Biomedical Engineering, Johns Hopkins University, 733 N. Broadway, BRB 573 Baltimore, Maryland 21205; [5] Cold Spring Harbor Laboratory, Functional Genomics, 1 Bungtown Road, Cold Spring Harbor, New York 11742;[6] Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037; and [7] Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037.

Acknowledgement
We thank the other members of the Mouse ENCODE consortium in formulating this response.

References

1. Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman WC, Gingeras TR, Ecker JR, Snyder MP: Comparison of the transcriptional landscapes between human and mouse tissues. *Proc Natl Acad Sci U S A*. 2014; **111** (48): 17224-17229 PubMed Abstract | Free Full Text | Publisher Full Text

2. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See LH, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu YC, Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Ramalho Santos M, Herrero J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert R, Hansen RS, De Bruijn M, Selleri L, Rudensky A, Josefowicz S, Samstein R, Eichler EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang KH, Skoultchi A, Gosh S, Disteche C, Treuting P, Wang Y, Weiss MJ, Blobel GA, Cao X, Zhong S, Wang T, Good PJ, Lowdon RF, Adams LB, Zhou XQ, Pazin MJ, Feingold EA, Wold B, Taylor J, Mortazavi A, Weissman SM, Stamatoyannopoulos JA, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B: A comparative encyclopedia of DNA elements in the mouse genome. *Nature*. 2014; **515** (7527): 355-364 PubMed Abstract | Free Full Text | Publisher Full Text

3. The FANTOM Consortium, the RIKEN PMI, CLST(DGT): A promoter-level mammalian expression atlas. *Nature*. 2014; **507** (7493): 462-470 PubMed Abstract | Publisher Full Text

Competing Interests: We declare no conflicts of interest.

Author Response 21 May 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

Michele, this is a good point; thank you for your thoughts and for taking the time to write a comment.

While I'd argue that there are easy ways to avoid confounding sequence batches and species (or any other biological variable of interest), I also intuitively agree with you. Differences in sample preparation could have a much larger impact than sequence batches. Note that in fact, genes associated with apoptosis, or cell death, are not typically enriched in post-mortem samples (or RNA samples of low quality), but the point stands -- differences in sample prep can result in marked differences in gene expression estimates.

Yet, the ENCODE authors did not report in their paper many salient details on sample preparation (for example, time since death, time to freezing, time to ship, etc.), and did not report RIN scores for the RNA samples. While you are correct that these types of experiments are 'difficult', recording and reporting these details is - in fact - easy.

Since we don't have those details, we can't conclude with confidence that there is a batch effect related to sample preparation, though again -- I agree with you that this is a most reasonable assumption given the details that are provided in the method section of the ENCODE paper.

Still, without being able to conclude or recapitulate with confidence a batch effect related to sample preparation, we analyzed what data we were provided. This led to the paper in front of us.

So, I would say that it is possible that with more details, we could have detected additional confounding batch effects, perhaps more significant than the sequencing design batch we reported. Importantly, if that is the case, sequencing again the same samples using a different design will not help resolving the technical confounders.

Competing Interests: No competing interests were disclosed.

Reader Comment 21 May 2015

Michele Busby, Broad Institute, USA

I am growing more suspicious that the sequencing batches are a red herring.

As pointed out below, in the original paper's supplemental tables 1 and 2 there is enrichment in dozens of classes of housekeeping genes between human and mouse.
If there is an artifact from a sequencing run, I would expect it to affect genes with certain chemical properties (abundance, transcript length, gc content). Obviously if a certain class is enriched with transcripts with those properties we will see false enrichment. But this would appear to be a lot of artifact with a lot of correlation with biological classes.

Because "cell death" is in the list of enriched processes, I wonder if there is a biological difference between the human and mouse samples, but that it originated in the sample handling rather than evolution. Samples from human bodies are not taken right away. The donors lie in the hospital bed while their loved ones say goodbye. I would expect to see some signature of apoptosis in the human samples that is not present in the mouse, which could be biologically "real" and affecting the clustering on the original data.

Obviously, blocking for sequencing runs is ideal. But the confounders here are confounded and in this paper here, as it's written, you may be overstating the effects of the sequencing run versus all the other things that are difficult about analyzing human tissue samples. Given the attention it received, this could lead to confusion in the field and unwarranted doubts about other papers where the sequencing is also not striped across runs.

I will also say that, though I don't think the main conclusion in the original paper is likely to be true, I respect that measuring transcription across species is very, very difficult. Many of the papers in the field say strange things. I offer this comment only in the spirit of collaboration so we know how to design future projects, not to criticize the original authors.

Competing Interests: None

Author Response 21 May 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

Thanks for your kind note, Mike. To your question: Only one tissue was in common across the two species in the mixed batch (brain). The other two samples were from pancreas (human) and spleen (mouse)... So too little data to try anything in our minds.

Competing Interests: No competing interests were disclosed.

Reader Comment 21 May 2015

Joe Foley, McGill University, Canada

Shin Lin: *"Because batch effects assuredly occur, we sought to minimize biases generally. First, for 10 of 13 tissues, the corresponding mouse and human samples had matching indexing/barcode primers. ... It should be noted that our study design minimized library preparation and primer index effect."*

You minimized the index sequence effect by confounding it with tissue type instead of species.

The simplest solution to the problem of sequencing batch effects would be to give each library a unique index and sequence all the libraries together in every lane of every flow cell. But as you say, that wouldn't account for the index sequence effect. Maybe a better design would be to use a mixture of several unique index sequences for each library, and still sequence them all together - at least this way you could look naively at the variation between index sequences within each library.

Or try to do better than N = 1 in each experimental condition, which seems like the the buried lede here. Even if you're willing to ignore the possibility of biological variation (!), at least technical replicates would have solved a lot of the problems here.

Competing Interests: No competing interests were disclosed.

Reader Comment 21 May 2015

Mike White, Washington University in St. Louis School of Medicine, USA

Yoav & Orna, kudos to you for raising this issue and documenting your procedure in detail.

I don't know if this is useful, but here's my question: there were two tissues for both species run on the same lane. By analyzing just those data, can you estimate the effect of species vs. tissue? I'm guessing two tissues isn't enough, but what do you think? Did you try that analysis?

Competing Interests: No competing interests were disclosed.

Reader Comment 20 May 2015

Lee Elizabeth Edsall, Duke University, USA

I find it interesting that CASAVA v1.7 doesn't include the run number in the sequence identifier in a fastq file. The run number is included in the other file types (e.g. qseq format). That's a critical piece of information. I'm glad Illumina added it in version 1.8. By the way, the format is on the page numbered 74 of the CASAVA 1.7 user guide, which corresponds to page number 88 of the PDF file.

Competing Interests: No competing interests.

Author Response 20 May 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

I appreciate the author's intention to recollect some of their data using a different study design. Unfortunately, if data were to be collected from the same set of original samples, there are some additional concerns. Based on the SI Methods provided by the authors, it seems that the human RNA samples were extracted from tissues collected from deceased individuals and then flash frozen. Presumably, different tissues were collected from different individuals, and there is no information regarding the age, sex, cause of death, and quality of extracted RNA. Additional human RNA – for a subset of tissues – were purchased directly.

For mouse, all tissues were collected from a single strain – so even if there were multiple individuals, all genetic backgrounds were identical. Also, the mice were all of a similar age and it is unclear if the RNA samples were flash frozen. Again, RNA quality properties are not reported.

The description of the sample acquisition already indicates a profound batch effect that is impossible to distinguish from the species effect – even before we address the issue of the sequencing study design.

Competing Interests: No competing interests were disclosed.

Reader Comment 20 May 2015

Rafael Irizarry, Harvard / Dana Farber Cancer Institute, USA

I wrote this post for those interested in learning more about the mathematical and statistical considerations related to confounding in the context of this discussion:
http://simplystatistics.org/2015/05/20/is-it-species-or-is-it-batch-they-are-confounded-so-we-cant-know/

Competing Interests: No competing interests were disclosed.

Reader Comment 20 May 2015

Michele Busby, Broad Institute, USA

This is a nice discussion of important topics!

I am interested in whether the RIN scores differed substantially between the mouse and the human samples. I would expect the human samples to have lower RIN because it is more difficult to collect samples from human donors.

If this is the case, then you do have to normalize by GC content and also gene length before clustering as different degrees of degradation in the samples would lead to consistent loss of RNA from transcripts with specific characteristics across all tissues in the more degraded group.

Competing Interests: None

Reader Comment (*Member of the F1000 Faculty*) 20 May 2015

Steven Salzberg, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, USA

Y. Gilad makes a salient point - what the mouse ENCODE authors say in their response is misleading in that they use the term "lane effect" when Gilad and Mizrahi-Man clearly explain that the human and mouse samples were (mostly) run on different instruments. Because this batch effect is almost completely confounded with the main effect reported (the clustering by species), it's nearly impossible to separate the two. Thus even if a species-based clustering is present, the data in the PNAS paper don't support that claim.

Competing Interests: None.

Author Response 20 May 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

Why are you calling these 'lane effects'? The samples were sequenced on different instruments (different sequencers) and by default - different flow cells. Can you explicitly acknowledge that fact please?

Competing Interests: No competing interests were disclosed.

Reader Comment 20 May 2015

Shin Lin, Department of Genetics, Stanford University, USA

In our analysis comparing the various tissue transcriptomes between human and mouse using datasets collected from a number of laboratories, we reported that significant differences existed between matched tissues across the species such that globally, the tissues within one species were more similar to each other[1]. We found that this was due in large part to the vastly different proportions of tissue-specific genes which were expressed in various organs, and that “housekeeping genes” e.g. metabolic genes were the major drivers of this species-specific clustering. When this surprising observation was made initially, the ENCODE consortium underwent extensive discussions over a two-year period, and conducted analyses and generated additional data to address concerns pertaining to potential laboratory and batch effects. In an examination of a subset of our data, Mizrahi-Man *et al*. revisit observations previously encountered by the consortium.

After receiving our data, which we provided as part of the ENCODE consortium, Mizrahi-Man *et al*. deduced that our multiplexed libraries were pooled onto different lanes such that libraries from each

species were largely sequenced on the same lanes. Reasoning that sequencing on different lanes might introduce differences in the resultant data generated, they normalized the expression values generated from libraries sequenced on different lanes. Because batch effects assuredly occur, we sought to minimize biases generally. First, for 10 of 13 tissues, the corresponding mouse and human samples had matching indexing/barcode primers. Second, 22 of the libraries were constructed in a single batch and sequenced on four lanes. Four other samples—human brain, human pancreas, mouse brain, and mouse spleen—were prepared in a later batch; should batch effects dominate the pattern of clustering, these four samples would be expected to cluster together. However, we did not observe such an effect (Fig. 1c of Lin *et al*, 2014). When we quantile normalize the data by the two experimental batches, we continue to observe species-specific clustering (https://www.dropbox.com/s/rjh9l6fn9t0svnh/fig1.jpg?dl=0).

Why then, does the normalization performed by Mizrahi-Man *et al*. result in tissue-specific clustering? If one normalizes away the species-specific differences, of course one will not see them. In the course of its analyses, the consortium demonstrated that if the mouse and human datasets were separately normalized, the global expression comparisons resulted in tissue-specific clustering (see Extended Data Fig. 1c of Yue *et al*[2]). This made intuitive sense to us, as it effectively removed the overall expression differences between the species and made apparent the expression differences between tissues. Indeed, the normalization procedure performed by Mizrahi-Man *et al*. did just that. Because the sequencing libraries were multiplexed largely by species, their normalization was equivalent to intra-species normalization, which effectively removed the global differences between human and mouse gene expression. To reiterate, this normalization sequence and resultant patterns of data clustering were well known to us and detailed in the ENCODE consortium main paper (see Extended Data Fig. 1c of Yue *et al*[2]).

There remains the issue of our study design with respect to confounding of lane effect and species. It should be noted that our study design minimized library preparation and primer index effect. A recent GEUVADIS consortium study showed that both factors are each contributors to RNA-seq variance and of much greater effect than that of lane (see Fig. 3c of 't Hoen *et al.*[3]). No study design given the current constraints of multiplexing and lane organization can account for both primer index and lane effect simultaneously, but the study design published in Lin et al. accounts for the larger of these two effects. Although our experience is consistent with the GEUVADIS data showing that lane effect is not a large contributor to variance, we recognize it is better to have data. Thus, we are sequencing under a new pattern of pooled libraries, and soon, we will post the results for the community.

Shin Lin[1,2], Yiing Lin[3], Michael A. Beer[4], Thomas R. Gingeras[5], Joseph R. Ecker[6,7], Michael Snyder[1]

[1] Department of Genetics, Stanford University, 300 Pasteur Drive, M-344 Stanford, California 94305; [2] Division of Cardiovascular Medicine, Stanford University, Falk Building, 870 Quarry Road Stanford, California 94304; [3] Department of Surgery, Washington University School of Medicine, 660 S. Euclid Ave., Campus Box 8109, St. Louis, Missouri 63110; [4] McKusick-Nathans Institute of Genetic Medicine and the Department of Biomedical Engineering, Johns Hopkins University, 733 N. Broadway, BRB 573 Baltimore, Maryland 21205; [5] Cold Spring Harbor Laboratory, Functional Genomics, 1 Bungtown Road, Cold Spring Harbor, New York 11742;[6] Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037; and [7] Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037.

Acknowledgement
We thank the other members of the Mouse ENCODE consortium in formulating this response.

References

1. Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman WC, Gingeras TR, Ecker JR, Snyder MP: Comparison of the transcriptional landscapes between human and mouse tissues. *Proc Natl Acad Sci U S A*. 2014; **111** (48): 17224-17229 PubMed Abstract | Free Full Text | Publisher Full Text

2. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See LH, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu YC, Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Ramalho Santos M, Herrero J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert R, Hansen RS, De Bruijn M, Selleri L, Rudensky A, Josefowicz S, Samstein R, Eichler EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang KH, Skoultchi A, Gosh S, Disteche C, Treuting P, Wang Y, Weiss MJ, Blobel GA, Cao X, Zhong S, Wang T, Good PJ, Lowdon RF, Adams LB, Zhou XQ, Pazin MJ, Feingold EA, Wold B, Taylor J, Mortazavi A, Weissman SM, Stamatoyannopoulos JA, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B: A comparative encyclopedia of DNA elements in the mouse genome. *Nature*. 2014; **515** (7527): 355-364 PubMed Abstract | Free Full Text | Publisher Full Text

3. 't Hoen PA, Friedlander MR, Almlof J, Sammeth M, Pulyakhina I, Anvar SY, Laros JF, Buermans HP, Karlberg O, Brannvall M, GEUVADIS Consortium, den Dunnen JT, van Ommen GJ, Gut IG, Guigo R, Estivill X, Syvanen AC, Dermitzakis ET, Lappalainen T: Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol*. 2013; **31** (11): 1015-1022 PubMed Abstract | Publisher Full Text

Competing Interests: The authors declare no conflicts of interest.