Supplementary Material to: On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data

Stephanie C. Hicks[1,2], Mingxiang Teng[1,2], and Rafael A. Irizarry[1,2]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
[2]Department of Biostatistics, Harvard School of Public Health

August 25, 2015

Contents

1 Supplemental Methods

1.1 Obtaining processed single-cell RNA-Seq data

For each published study, we downloaded the processed single-cell RNA-Seq data provided by the authors on GEO (summary provided in Table 1 in manuscript). In all of the studies, we used the processed expression data available on GEO, applied principal components analysis on the `log2` transformed values (adding 1 to avoid logs of 0), and computed the proportion of detected genes from the same data set with the exception of one study. In Patel et al. (2014), the processed expression data available on GEO excluded most non-detected genes and the data were pre-standardized by the authors by removing the gene-specific mean in each row. In this case, we used the computed the proportion of detected genes from the raw data.

1.2 Obtaining raw reads for single-cell RNA-Seq data

We downloaded the raw reads from the Sequence Read Archive (SRA) on NCBI [1]. The SRA files were convereted to FASTQ files using `fastq-dump` in the SRA Toolkit. For this analysis, we used the *Homo sapiens* and *Mus musculus* ENSEMBL genome builds (GRCh38 and GRCm38, respectively) and corresponding transcript annotation files.

We used STAR read aligner (`https://github.com/alexdobin/STAR`) (version 2.3.1) [2] to map the raw reads to their respective genomes. We first used the `genomeGenerate` run mode to generate the genome index files. This was done only once per genome/annotation combination (one for human, one for mouse). Then we mapped the reads in the `alignReads` mode with all default options. The mapped reads were returned in a BAM file format.

We obtained the transcript database for the human and mouse ENSEMBL genes from BiomaRt [3, 4] using the `GenomicFeatures` R/Bioconductor package [5]. The BAM files were read into R ussing the `Rsamtools` R/Bioconductor package. We used the `summarizeOverlaps()` function in the `GenomicAlignments` R/Bioconductor package [5] to count the number of reads that overlap with exons.

1.3 Identifying batch information in study design

We reconstructed the study design from the sequence identifiers provided in the FASTQ files [6]. We extracted the first line in each FASTQ file header using `sed` and then parsed the sequence identifier line using the `stringr` R package [7]. The sequence identifier contained the machine identifier, run number, flow cell identifier, and flow cell lane number.

2 Supplemental Tables and Figures

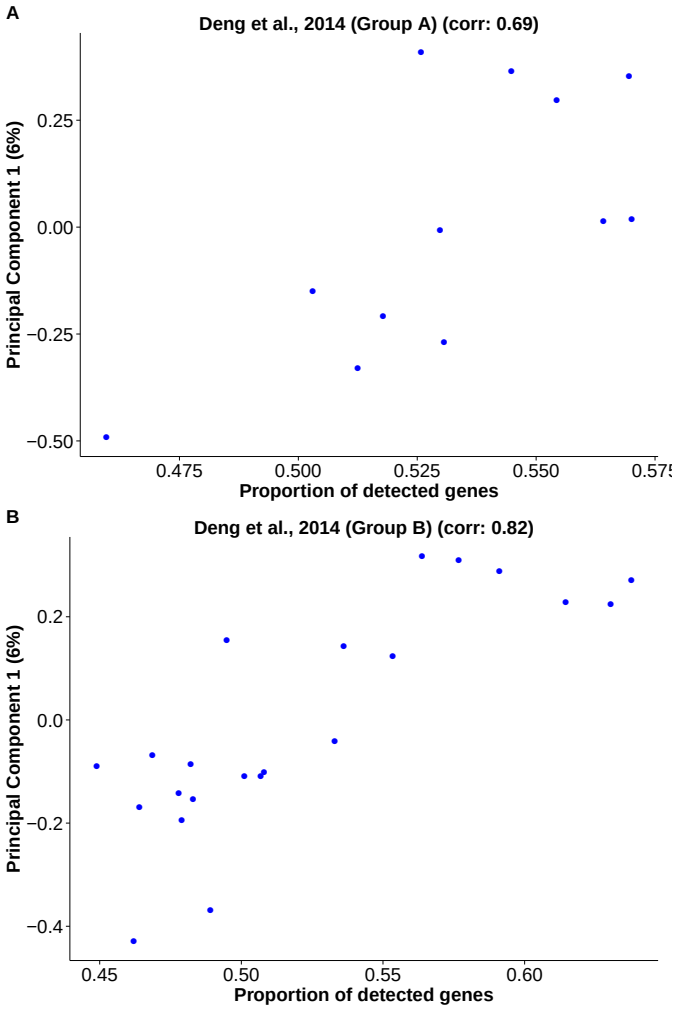

Figure 1: scRNA-Seq data from Deng et al. (2014). The biological groups defined by the authors (Supplementary Table 3) were seperated into four groups: Group A (Zygote, Early 2-cell), Group B (Mid 2-cell, Late 2-cell), Group C (4-cell, 8-cell, 16-cell), and Group D (Early, Mid and Late blastocyst). This is a plot comparing the relationship between the proportion of detected genes and the first principal component in Groups A and B. Groups C and D are shown in Figure 2 in the manuscript.

Patel et al. (2014) [8] performed paired-end single-cell RNA-Seq in 430 cells from five glioblastoma tumors (GSE57872). The study design is provided in Table 1. In four tumors, the individual cells sequenced were processed in seperate batches and in the fifth tumor, the cells were processed in two batches.

The processed data available was previously filtered by the authors for a composite gene expression either across all cells combined (average log2(TPM) > 4.5) or within a single tumor (average log2(TPM) > 6 in at least one tumor) and excluded the majority of non-detected genes.

Batch					Biological Group				
	mi	runID	fc	lane	Group 1 (MGH28)	Group 2 (MGH29)	Group 3 (MGH30)	Group 4 (MGH31)	Group 5 (MGH26)
Batch 1	HISEQ	717	H799JADXX	2	94	0	0	0	0
Batch 2	HISEQ	643	H110YADXX	2	0	75	0	0	0
Batch 3	HISEQ	644	H11YBADXX	1	0	0	73	0	0
Batch 4	HISEQ	718	H14TNADXX	2	0	0	0	70	0
Batch 5	GLPB22-B5C	556	H0PFYADXX	1	0	0	0	0	53
Batch 6	HISEQ	704	H759HADXX	2	0	0	0	0	65

Table 1: Study design by Patel et al. (2014). Number of cells sequenced from individual tumors and across batches. Abbreviations: machine identifier (mi), flow cell (fc).

Treutlein et al. (2014) [9] performed paired-end single-cell RNA-Seq in 198 mouse lung cells at four different developmental stages of lung epithelium (GSE52583). The authors used FPKMs for normalization. The study design in is provided in Table 2. Individual cells from four developmental stages of the mouse lung epithelium include three biological replicates in stage ED18.5.

Batch					Biological Group			
	mi	runID	fc	lane	Group 1 (ED14.5)	Group 2 (ED16.5)	Group 3 (ED18.5)	Group 4 (Adult)
Batch 1	HISEQ	418	C2FP5ACXX	1	1	0	0	0
Batch 2	HISEQ	418	C2FP5ACXX	2	44	0	0	0
Batch 3	NA	NA	NA	NA	0	27	0	0
Batch 4 (rep 1)	DJG84KN1	381	C1WAVACXX	4	0	0	20	0
Batch 5 (rep 2)	DJG84KN1	404	D2B8YACXX	2	0	0	34	0
Batch 6 (rep 3)	HISEQ	413	D21AUACXX	7	0	0	26	0
Batch 7	HISEQ	418	C2FP5ACXX	3	0	0	0	46

Table 2: Study design by Treutlein et al. (2014). Number of cells sequenced across four developmental stages and across batches. Abbreviations: machine identifier (mi), flow cell (fc), NA (missing information in FASTQ header). The phenotypic information provided on GEO states Batch 3 was run on an Illumina MiSeq and the other batches were run on a Illumina HiSeq 2000, but the sequence identifer information in the FASTQ file was missing.

Deng et al. (2014) [10] performed single-end single-cell RNA-Seq in 286 mouse developmental cells (ranging from zygote to late blastocyst) to study monoallelic expression (GSE45719). The authors used RPKMs for normalization. The study design is provided in Table 3. Within each runID, cells were processed across multiple flow cell lanes.

For the purpose of this manuscript, the biological groups were seperated into four groups: Group A (Zygote, Early 2-cell), Group B (Mid 2-cell, Late 2-cell), Group C (4-cell, 8-cell, 16-cell), and Group D (Early, Mid and Late blastocyst).

RunID	Time-course									
		2-cell						blastocyst		
	Zygote	Early	Mid	Late	4-cell	8-cell	16-cell	Early	Mid	Late
	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10
Run0040	0	0	6	0	0	0	23	0	0	0
Run0044	0	0	0	0	0	0	0	0	42	0
Run0081	4	4	0	0	0	0	0	0	0	0
Run0083	0	4	6	1	7	0	0	0	0	0
Run0084	0	0	0	8	7	7	0	0	0	0
Run0085	0	0	0	1	0	11	0	0	0	0
Run0088	0	0	0	0	0	10	27	0	0	0
Run0095	0	0	0	0	0	0	0	15	18	0
Run0099	0	0	0	0	0	0	0	26	0	19
Run00100	0	0	0	0	0	0	0	2	0	11
Run00192	0	0	0	0	0	10	0	0	0	0
Run00193	0	0	0	0	0	9	8	0	0	0

RunID	Flow cell Lane							
	1	2	3	4	5	6	7	8
Run0040	6	6	6	0	6	5	0	0
Run0044	6	6	6	5	4	6	6	3
Run0081	0	0	4	4	0	0	0	0
Run0083	0	0	0	2	2	2	6	6
Run0084	1	5	4	3	3	4	1	1
Run0085	0	0	0	0	4	4	4	0
Run0088	5	5	5	4	5	5	4	4
Run0095	5	5	5	3	5	5	5	0
Run0099	5	6	5	6	6	6	6	5
Run00100	5	4	4	0	0	0	0	0
Run00192	0	0	0	10	0	0	0	0
Run00193	11	6	0	0	0	0	0	0

Table 3: Study design by Deng et al. (2014). Number of cells sequenced across the developmental stages and across batches.

Trapnell et al. (2014) [11] performed paired-end single-cell RNA-Seq in 372 primary human myoblasts taken over a time-course of serum-induced cell differentiation (GSE52529). The data was processed using FPKMs. The study design is provided in Table 4. In a four-stage time-course investigating cell differentiation, the individual cells sequenced from each time point were processed in seperate batches.

	Batch				Biological Group			
	mi	runID	fc	lane	Group 1 (Hour 0)	Group 2 (Hour 24)	Group 3 (Hour 48)	Group 4 (Hour 72)
Batch 1	HWI-ST1233	229	H0L2PADXX	1	96	0	0	0
Batch 2	HWI-ST1233	226	H0NF2ADXX	1	0	96	0	0
Batch 3	HWI-ST1233	231	H0KLJADXX	1	0	0	96	0
Batch 4	HWI-ST1233	NA	C268PACXX130720	4	0	0	0	84

Table 4: Study design by Trapnell et al. (2014). Number of cells sequenced across the time-course and across batches. Abbreviations: machine identifier (mi), flow cell (fc), NA (missing information in FASTQ header).

Shalek et al. (2014) [12] performed paired-end single-cell RNA-Seq in 383 primary mouse dendritic cells stimulated with the LPS experimental condition across four time points (GSE48968). We focused on the time course only the LPS experimental condition in this analysis. The authors used TPMs for normalization. The study design is provided in Table 5.

	Batch				Biological Group			
	mi	runID	fc	lane	Group 1 (Hour 1)	Group 2 (Hour 2)	Group 3 (Hour 4)	Group 4 (Hour 6)
Batch 1	NA	D1588ACXX120910	NA	1	96	0	0	0
Batch 2	NA	D1588ACXX120910	NA	4	0	96	0	0
Batch 3	NA	D15C0ACXX120910	NA	1	0	0	95	0
Batch 4	NA	D15C0ACXX120910	NA	4	0	0	0	96

Table 5: Study design by Shalek et al. (2014). Number of cells sequenced across the time-course and across batches. Abbreviations: machine identifier (mi), flow cell (fc), NA (missing information in FASTQ header).

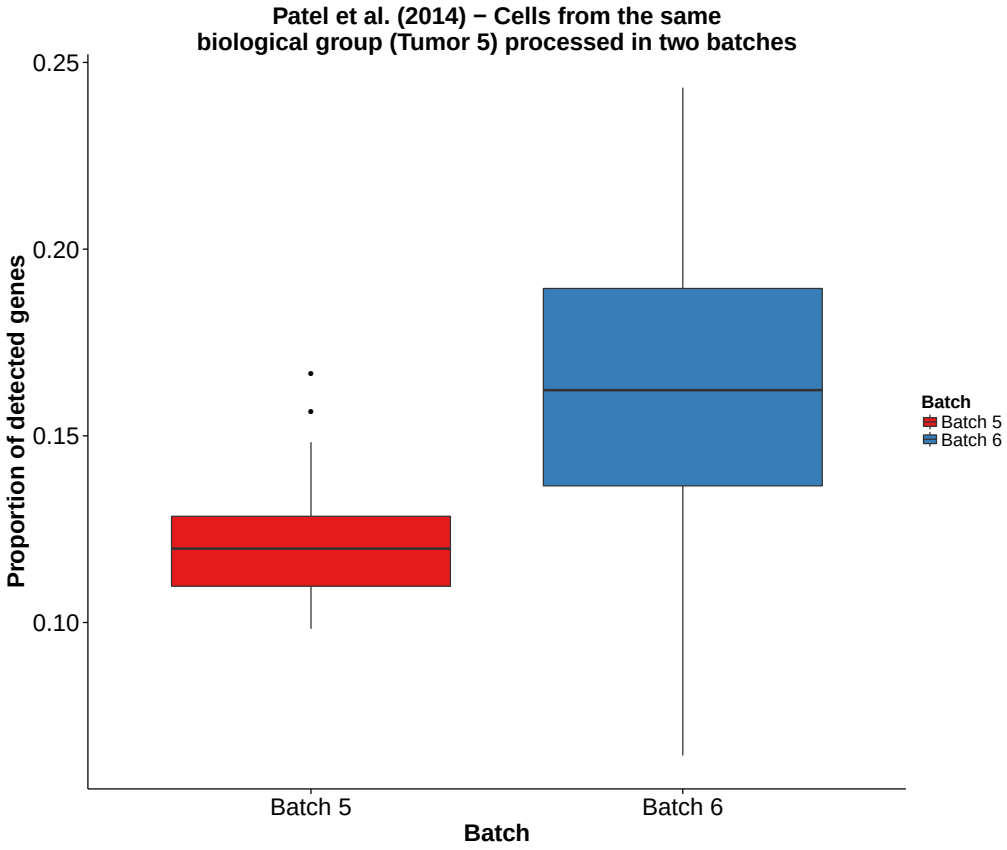

Figure 2: scRNA-Seq data from Patel et al. (2014). Cells from one biological group (Tumor 5) were processed two batches (Batch 5 and Batch 6). Each batch of processed cells has a different distribution of proportion of detected genes. The other biological groups were completely confounded with batch.

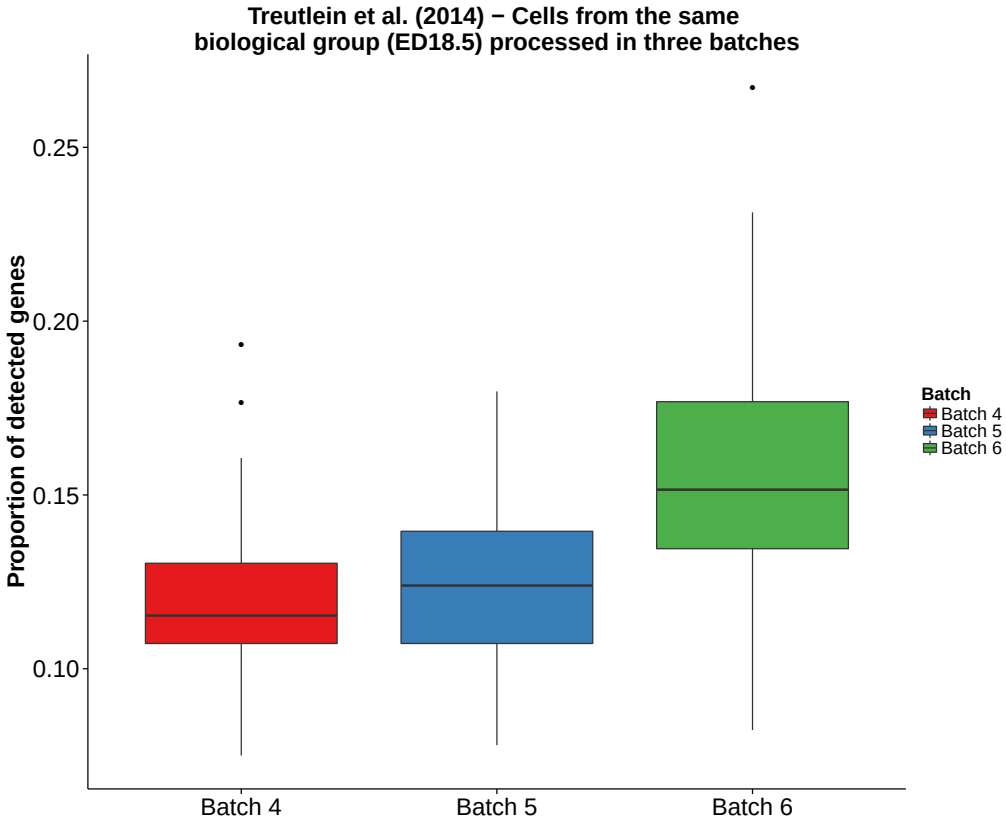

Figure 3: scRNA-Seq data from Treutlein et al. (2014). Cells from one biological group (ED18.5 or Group 3 as listed in Table 2) were processed three batches (Batch 4, Batch 5 and Batch 6). Each batch of processed cells has a different distribution of proportion of detected genes. The other biological groups were completely confounded with batch.

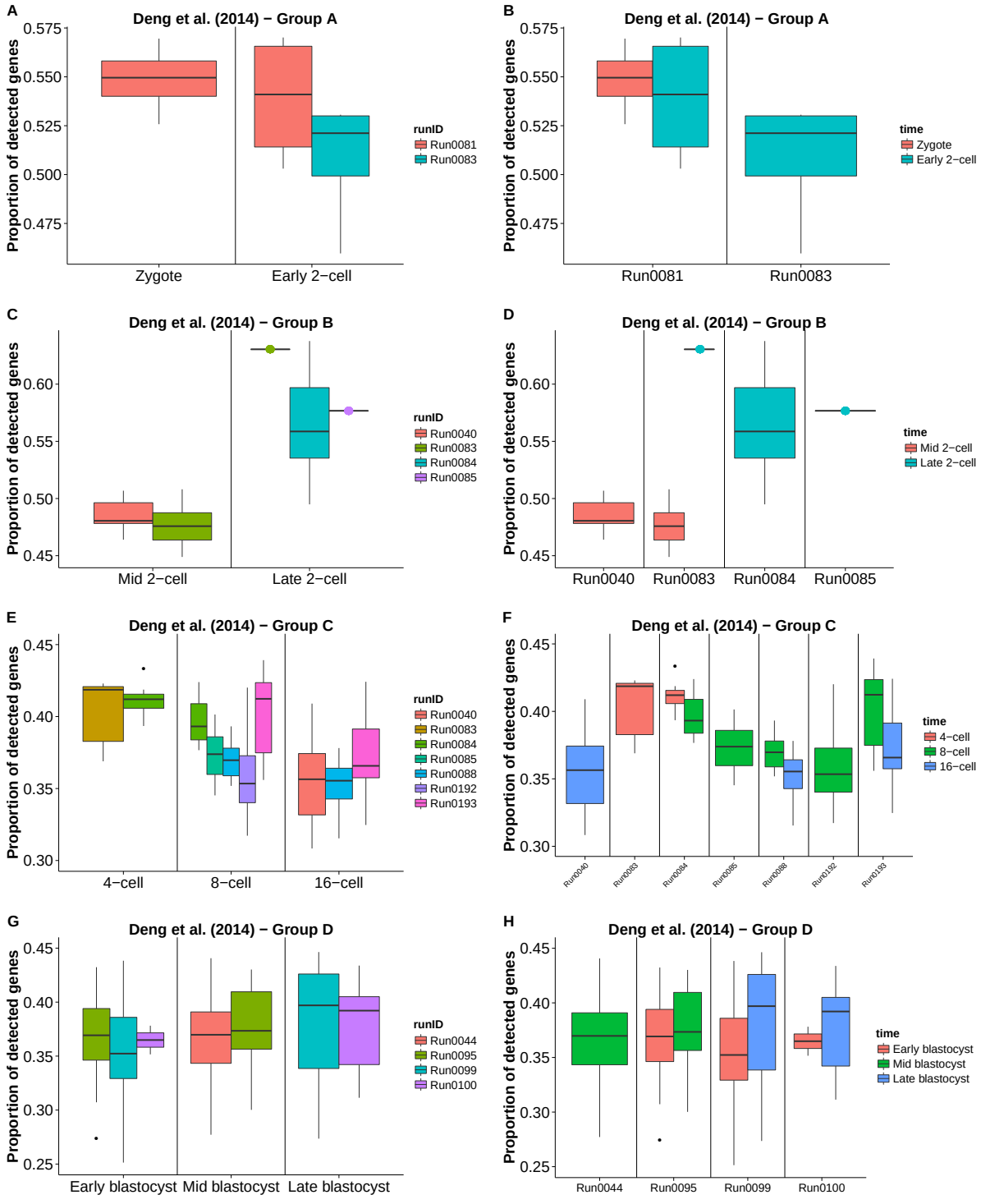

Figure 4: scRNA-Seq data from Deng et al. (2014). The biological groups defined by the authors (Supplementary Table 3) were seperated into four groups: Group A (Zygote, Early 2-cell), Group B (Mid 2-cell, Late 2-cell), Group C (4-cell, 8-cell, 16-cell), and Group D (Early, Mid and Late blastocyst). The left column contains boxplots of the proportion of detected genes grouped by biological groups and colored by batch. The right column contains boxplots of the proportion of detected genes grouped by batch and colored by biological group.

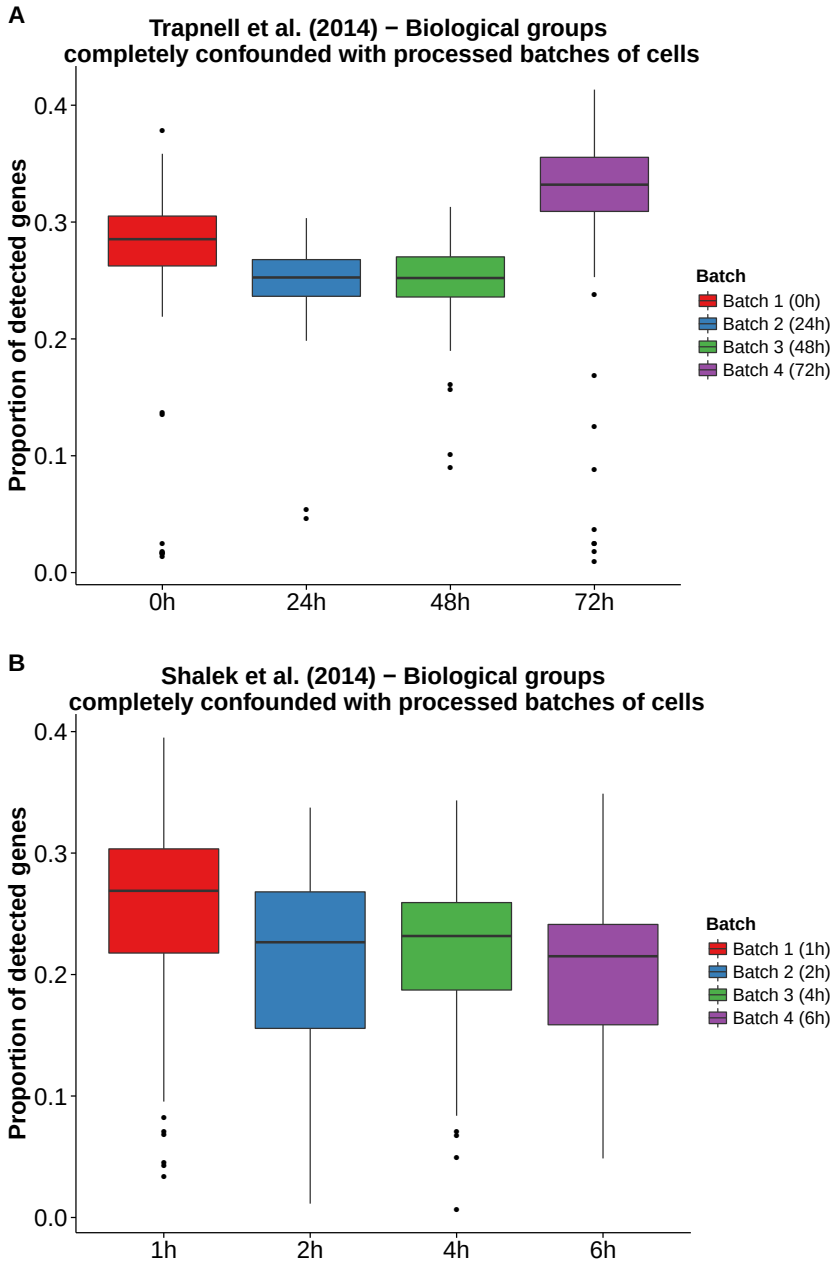

Figure 5: Boxplots of the proportion of detected genes from Trapnell et al. (2014) and Shalek et al. (2014). In both studies, biological groups were completely confounded with batch.

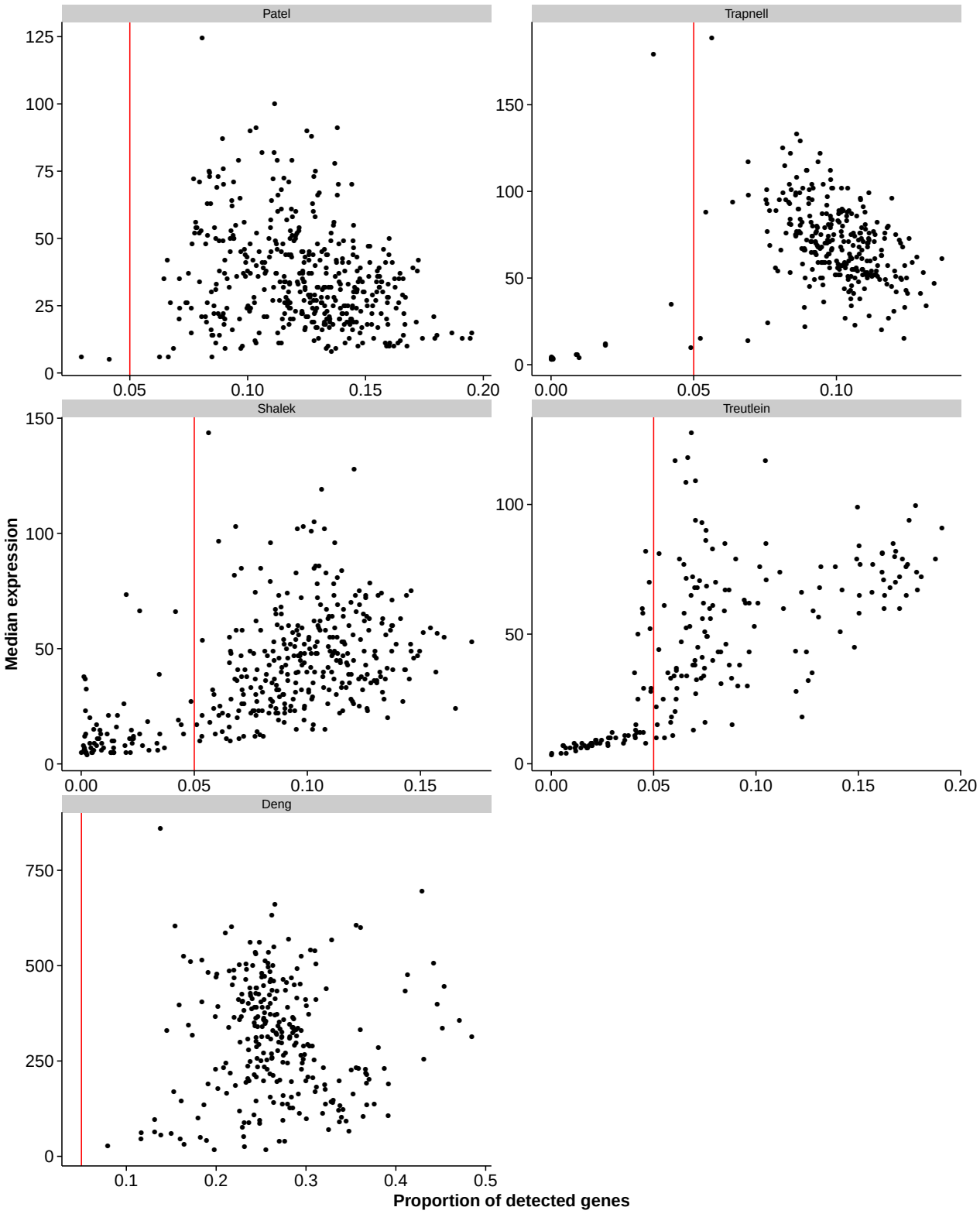

Figure 6: We removed a small set of cells with a proportion of detected genes less than 0.05 (red line). This was because these cells appeared to have resulted in failed experiments: very low detection rates and very low median expression values.

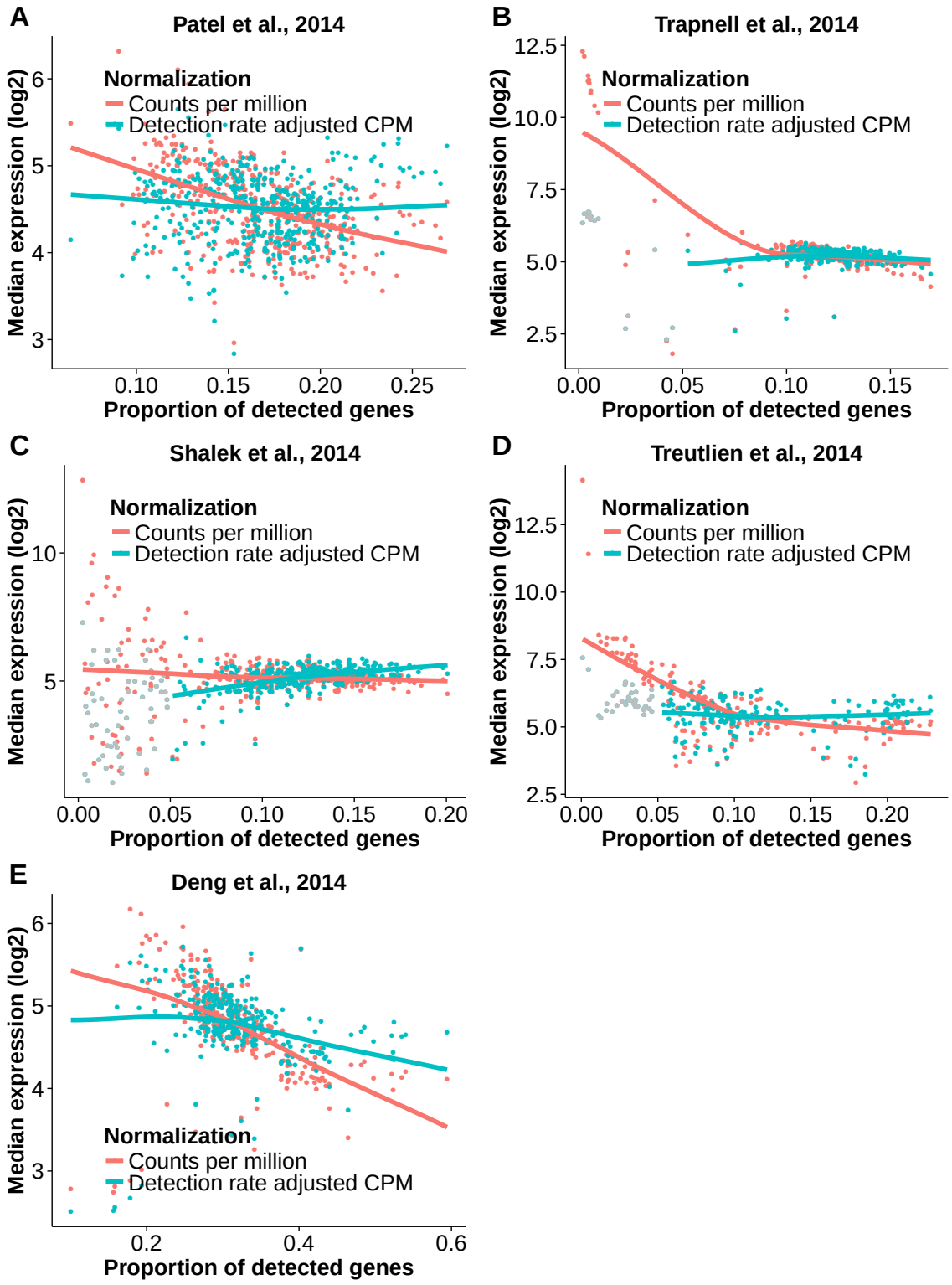

Figure 7: Median gene expression of detected genes in scRNA-Seq samples with varying proportions of detected genes using (1) counts per million (red) and (2) counts per million adjusting (gray and blue) for differences in proportions of detected genes. For the second normalization, a small set of cells with a proportion of detected genes less than 0.05 (Supplementary Figure 6) were first removed (gray points) and then the detection rate adjusted CPM was calculated (blue). Failure to account for differences in differences of the proportion of detected genes between cells over-inflates the gene expression of cells with a low proportion of detected genes. Used locally weighted scatter plot smoothing (loess) with a degree of 1 and span of 1 for all figures. scRNA-Seq data from (A) Patel et al. (2014). (B) Trapnell et al. (2014). (C) Shalek et al. (2014). (D) Treutlein et al. (2014). (E) Deng et al. (2014). Note: the range of the x-axis is different for each study.

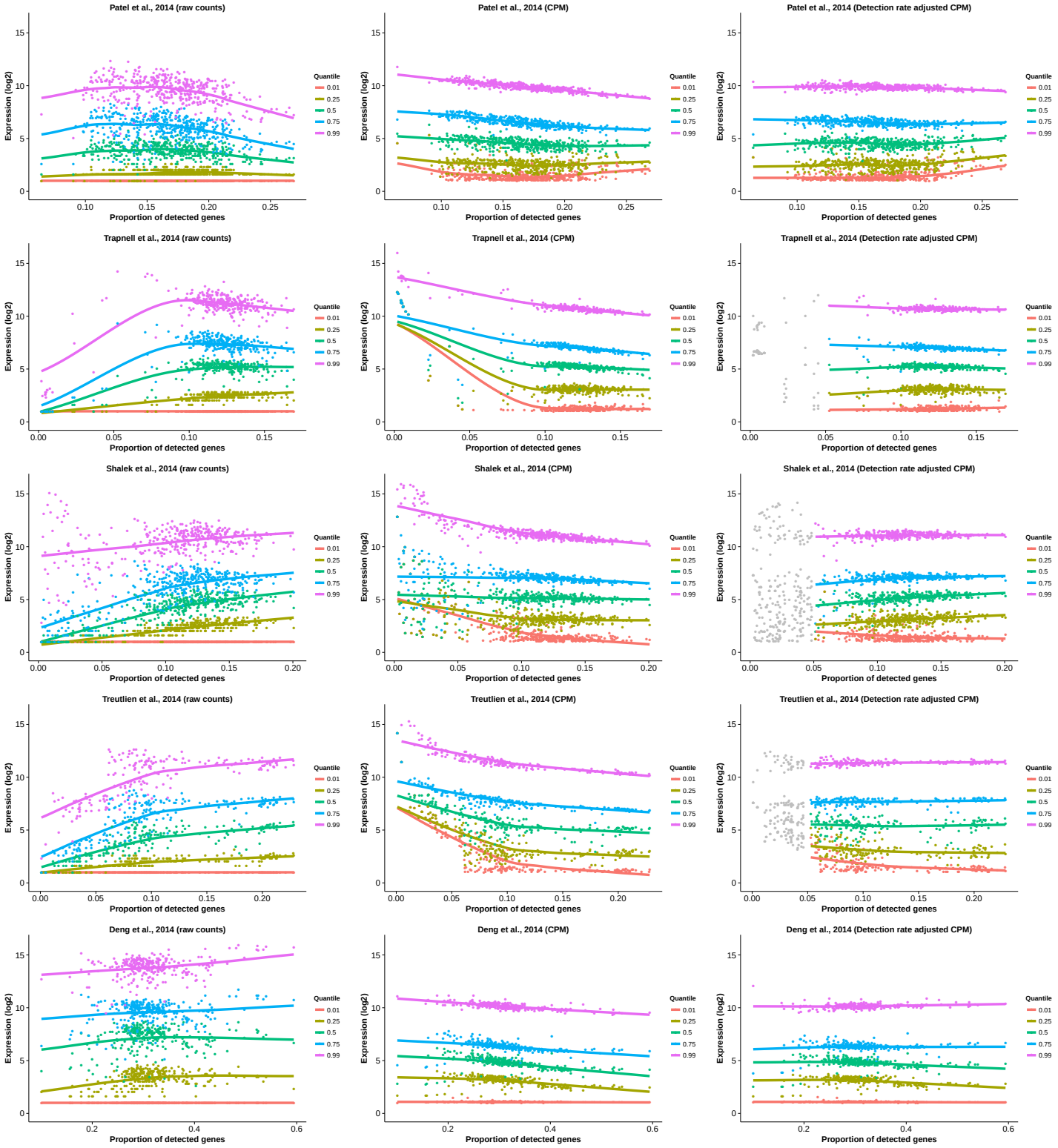

Figure 8: Five quantiles (0.01, 0.25, 0.5, 0.75, 0.99) of gene expression of detected genes in scRNA-Seq samples with varying proportions of detected genes using (1) raw counts (2) counts per million (CPM) and (3) detection rate adjsuted CPM. For the last normalization, a small set of cells with a proportion of detected genes less than 0.05 (Supplementary Figure 6) were first removed (gray points) and then the detection rate adjusted CPM was calculated. Failure to account for differences in differences of the proportion of detected genes between cells over-inflates the gene expression of cells with a low proportion of detected genes. Single cells from (Row 1) Patel et al. (2014). (Row 2) Trapnell et al. (2014). (Row 3) Shalek et al. (2014). (Row 4) Treutlein et al. (2014). Used locally weighted scatter plot smoothing (loess) with a degree of 1 and span of 1 for all figures.

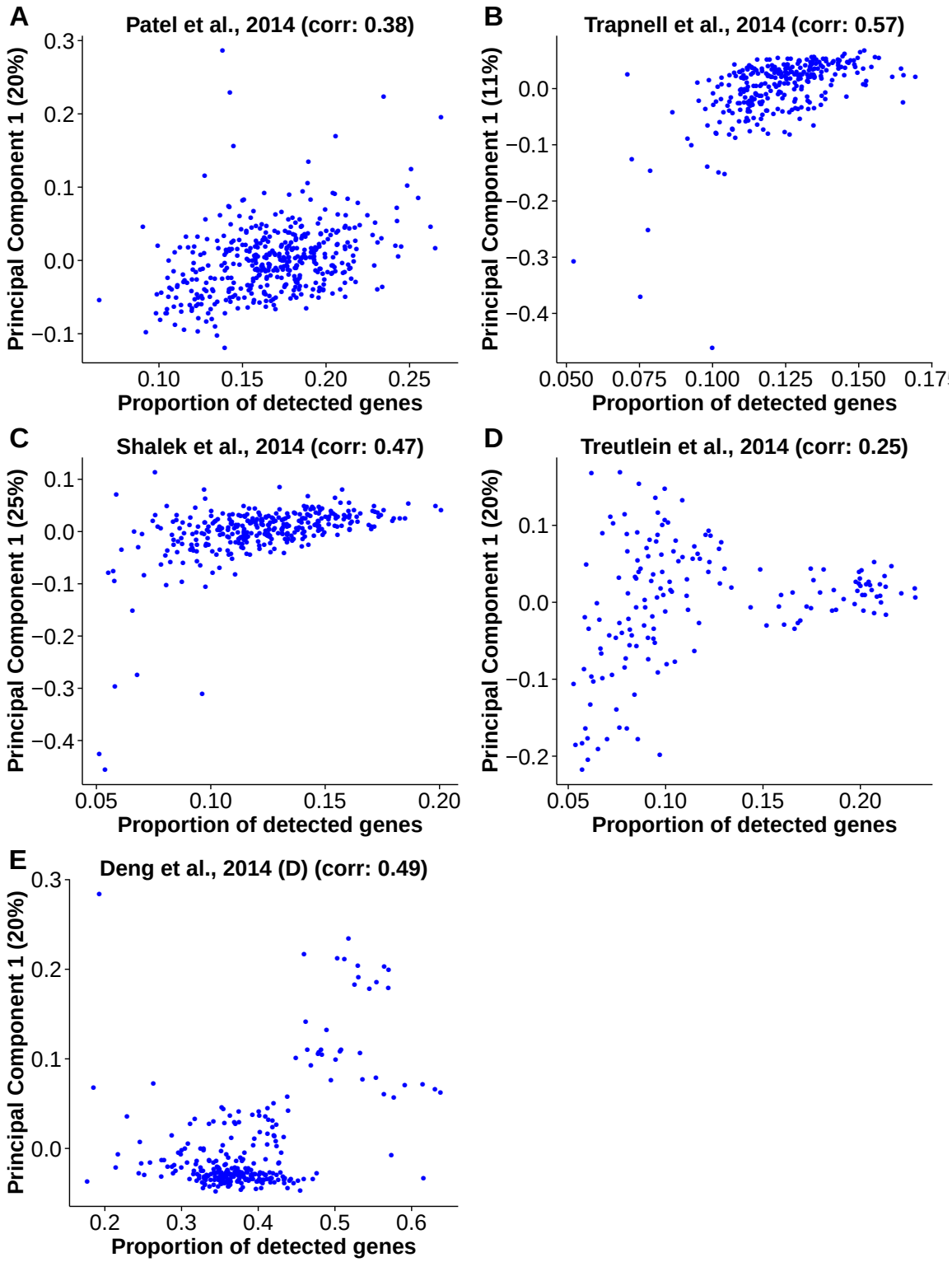

Figure 9: Strong correlation between the first principal component and the proportion of detected genes across cells after removing cells with a proportion of detected genes less than 0.05 (Supplementary Figure 6) and removing all genes with at least one cell not expressing that gene. In each study, the first principle component and the proportion of detected genes was calculated on the raw data normalized by counts per million to adjust for library size. Similar results were found using the processed data available on GEO (but not shown here).

References

[1] Yuichi Kodama et al. "The Sequence Read Archive: explosive growth of sequencing data". In: *Nucleic Acids Res* 40.Database issue (2012), pp. D54–6. DOI: 10.1093/nar/gkr854.

[2] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

[3] Steffen Durinck et al. "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis". In: *Bioinformatics* 21.16 (2005), pp. 3439–40. DOI: 10.1093/bioinformatics/bti525.

[4] Steffen Durinck et al. "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt". In: *Nat Protoc* 4.8 (2009), pp. 1184–91. DOI: 10.1038/nprot.2009.97.

[5] Michael Lawrence et al. "Software for computing and annotating genomic ranges". In: *PLoS Comput Biol* 9.8 (2013), e1003118. DOI: 10.1371/journal.pcbi.1003118.

[6] Yoav Gilad and Orna Mizrahi-Man. "A reanalysis of mouse ENCODE comparative gene expression data". In: *F1000Res* 4 (2015), p. 121. DOI: 10.12688/f1000research.6536.1.

[7] Hadley Wickham. "stringr: modern, consistent string processing". In: *The R Journal* 2.2 (2010), pp. 38–49.

[8] Anoop P Patel et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma". In: *Science* 344.6190 (2014), pp. 1396–401. DOI: 10.1126/science.1254257.

[9] Barbara Treutlein et al. "Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq". In: *Nature* 509.7500 (2014), pp. 371–5. DOI: 10.1038/nature13173.

[10] Qiaolin Deng et al. "Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells". In: *Science* 343.6167 (2014), pp. 193–6. DOI: 10.1126/science.1245316.

[11] Cole Trapnell et al. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". In: *Nat Biotechnol* 32.4 (2014), pp. 381–6. DOI: 10.1038/nbt.2859.

[12] Alex K Shalek et al. "Single-cell RNA-seq reveals dynamic paracrine control of cellular variation". In: *Nature* 510.7505 (2014), pp. 363–9. DOI: 10.1038/nature13437.