Supplementary Material to: Missing Data and Technical Variability in Single-Cell RNA-Sequencing Experiments

Stephanie C. Hicks[1,2], F. William Townes[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

September 8, 2017

Contents

1 Supplemental Methods

1.1 Inclusion criteria for single-cell RNA-Seq data

In July 2015 we performed a search on GEO [1] for single-cell RNA-Seq data sets with a search criteria of "single-cell RNA-Seq" and selected studies with a listed sample size of larger than $n = 200$ single cells, which led to 12 data sets [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. This list of studies included a range of different single-cell RNA-Seq protocols for the sequencing including Tang et al. (2009) [15], STRT-Seq [16, 17], and SMART-Seq / SMART-Seq2 [18, 19]. We also included three additional data sets that were that were used in methodological papers illustrating new single-cell RNA-Seq protocols: MARS-Seq [21], Drop-Seq [22], and Chromium Single Cell from 10X Genomics [14]. A summary of these data sets is provided in Table 1 in the manuscript.

1.2 Obtaining processed single-cell RNA-Seq data

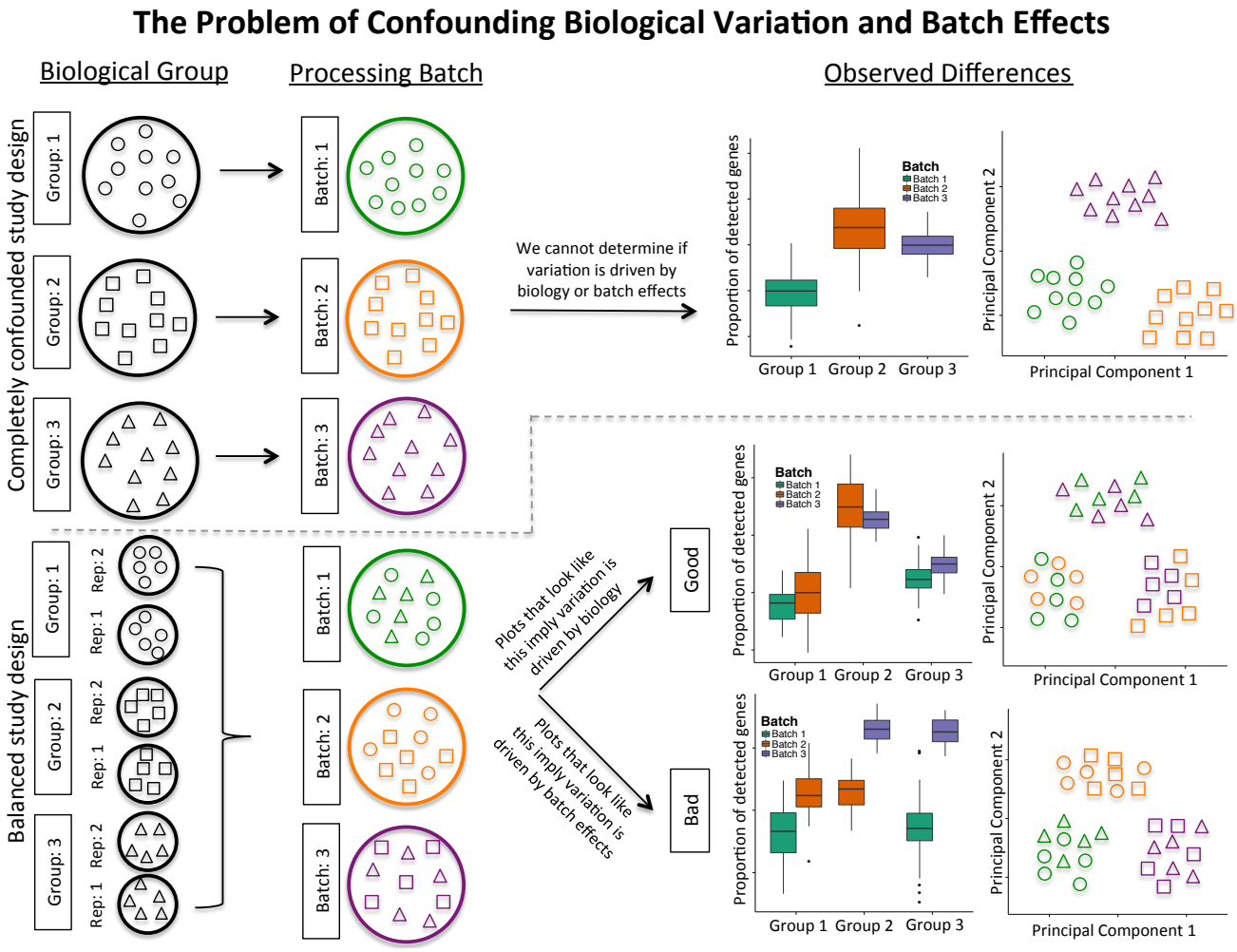
For each published study, we downloaded the processed single-cell RNA-Seq data provided by the authors on GEO [1] with one exception. For Zheng et al. (2017) [14], we used the matrix of 20K randomly sample cells available on the 10X Genomics website. In all of the studies, we used the processed expression data, applied principal components analysis on the `log2` transformed values (adding 1 to avoid logs of 0), and computed the detection rate from the same data set with the exception of the following studies:

- In Patel et al. (2014) [6], 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 detection rate from the count table derived from the raw reads.
- In Kowalczyk et al. (2015) [5] and Kumar et al. (2014) [4] the processed expression data was already provided on the transformed `log` scale.
- In studies that used unique molecular identifiers (UMIs) or barcoding for molecule counting [12, 13, 21, 22, 14], we applied a normalization previously suggested [22]. We which normalizes each UMI count matrix by dividing by the total number of UMIs per cell, multiplies by 10^6 and transforms the normalized counts to the `log2` scale (adding 1 to avoid logs of 0).

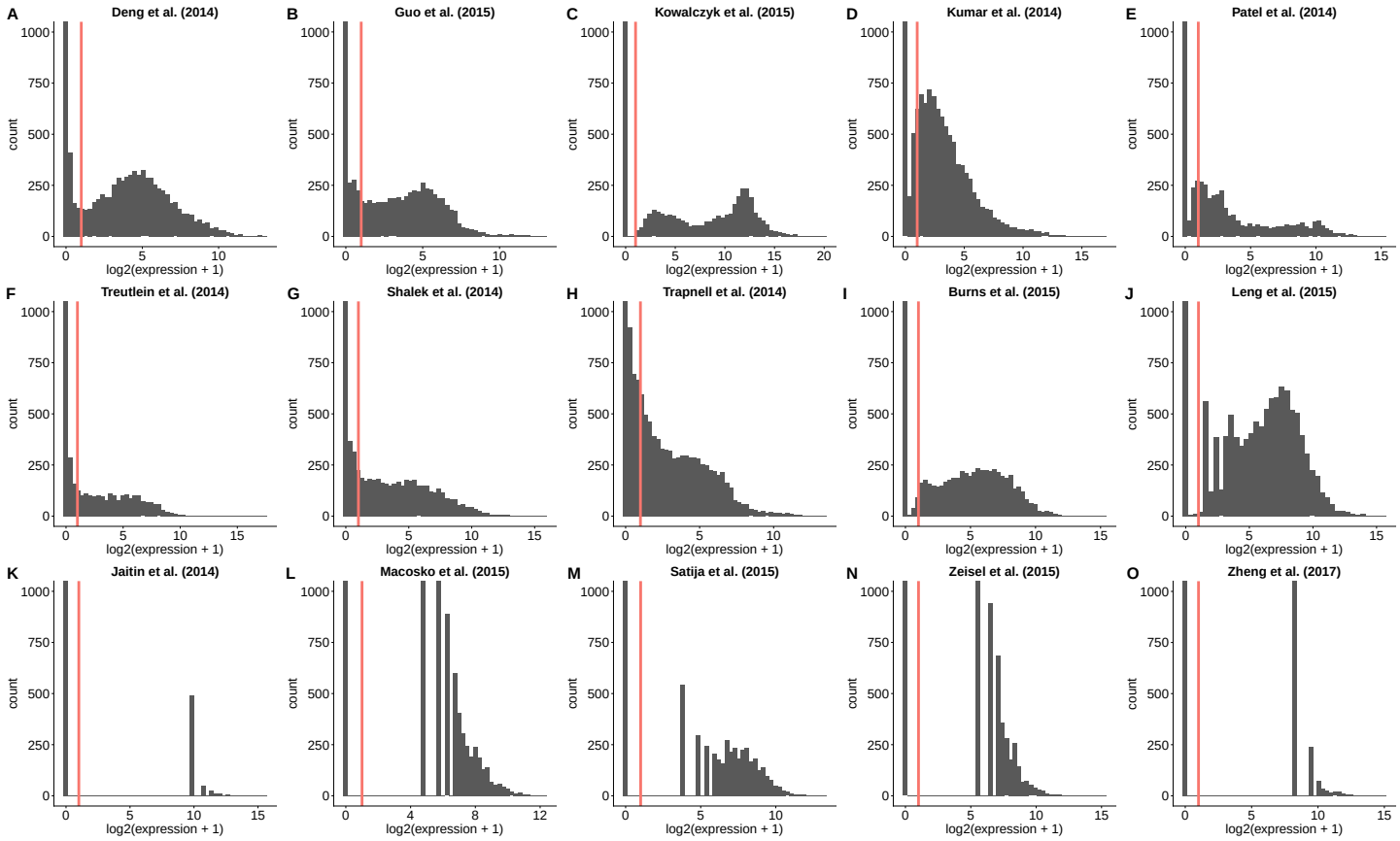
1.3 Identifying batch information in study design

We reconstructed the study design from the sequence identifiers provided in the FASTQ files [24], which contain the raw reads from the Sequence Read Archive (SRA) on NCBI [25]. The SRA files were converted to FASTQ files using `fastq-dump` in the SRA Toolkit. We extracted the first line in each FASTQ file header using `sed` and then parsed the sequence identifier line using the `stringr` R package [26]. The sequence identifier contained the machine identifier, run number, flow cell identifier, and flow cell lane number.

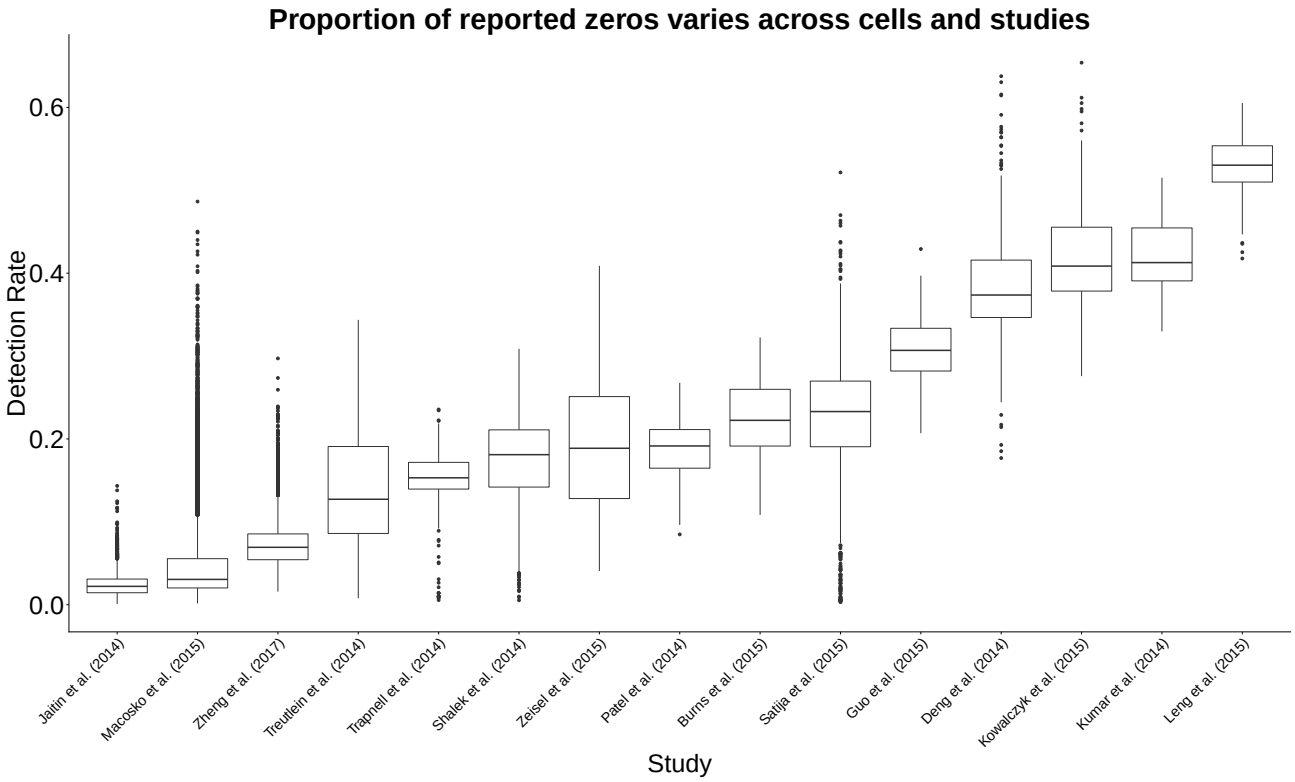
2 Supplemental Tables and Figures



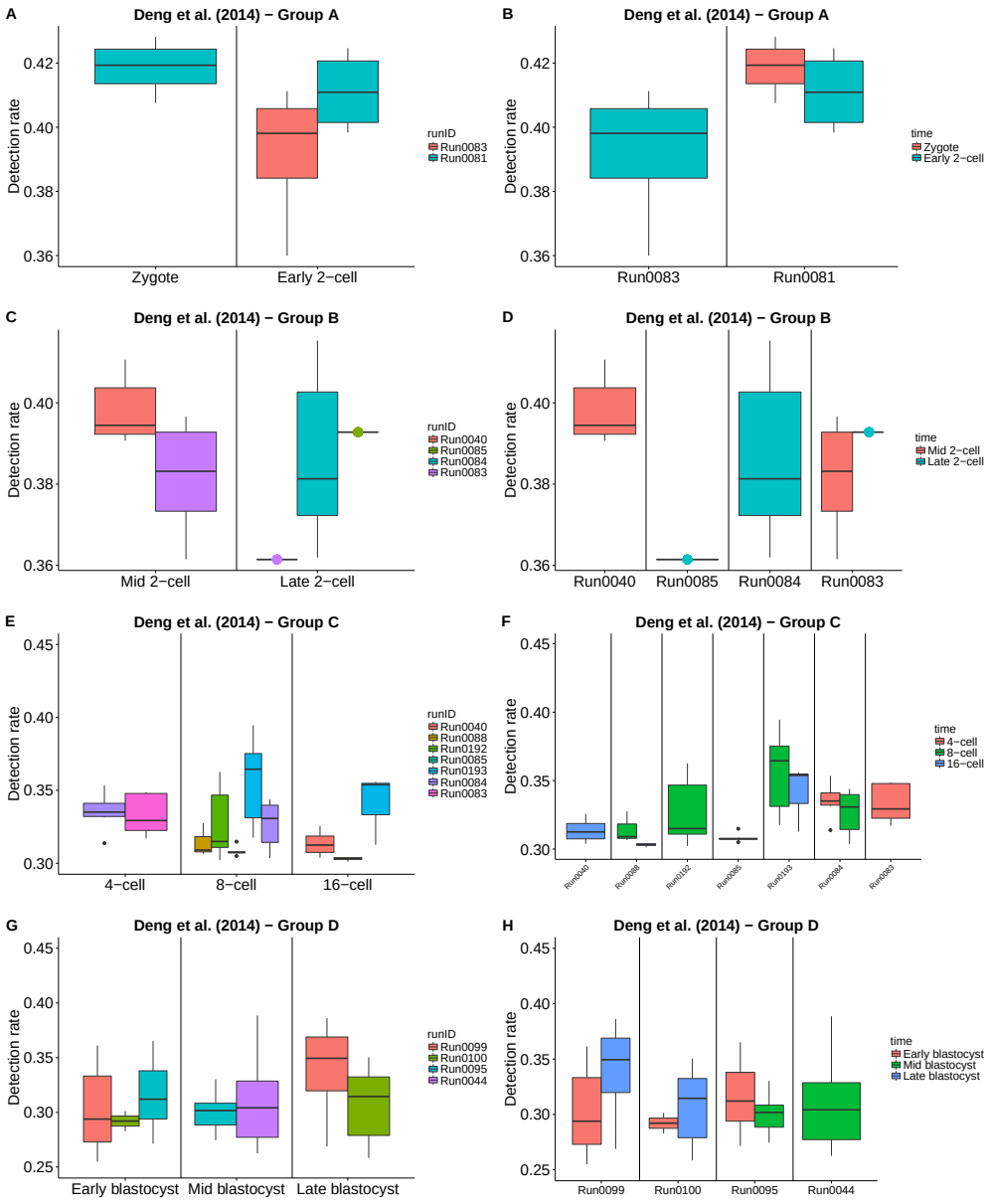
Supplemental Figure 1: **The problem of confounding biological variation and batch effects**. The top section depicts a completely confounded study design of processing individual cells from three biological groups (represented by shapes) in three separate batches (represented by colors). In this case, we cannot determine if biology or batch effects drive the observed variation. The bottom section depicts a balanced study design consisting of multiple replicates (rep) split and processed across multiple batches. The use of multiple replicates allows observed variation be attributed to biology (cells cluster by shape) or batch effects (cells cluster by color).



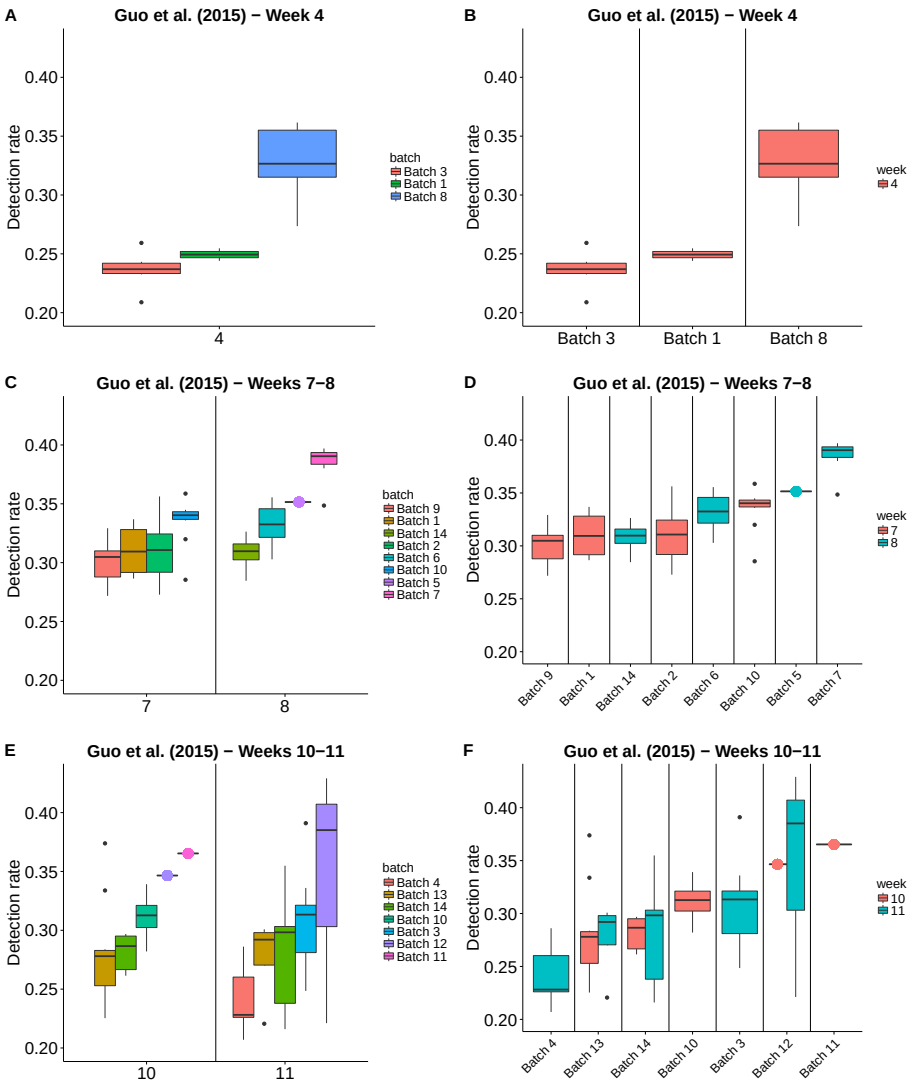
Supplemental Figure 2: Histograms of the gene expression from one cell in 15 studies. The top two rows of studies reported normalized expresssion values as TPMs, FPKMs or RPKMs. The bottom row of studies reported UMI counts, which we converted to CPMs (see description of data in main manuscript for details). The first cell in the data set provided on GEO was used here, but other cells had a similar pattern (plots not shown). Two modes can be seen in the histograms, which we interpreted to be associated with background noise and signal respectively, with the lower mode defined as values below a TPM, FPKM, RPKM or CPM threshold of $\delta = 1$ (red line). This is the threshold we used to define the detection rate in this paper, unless otherwise stated.



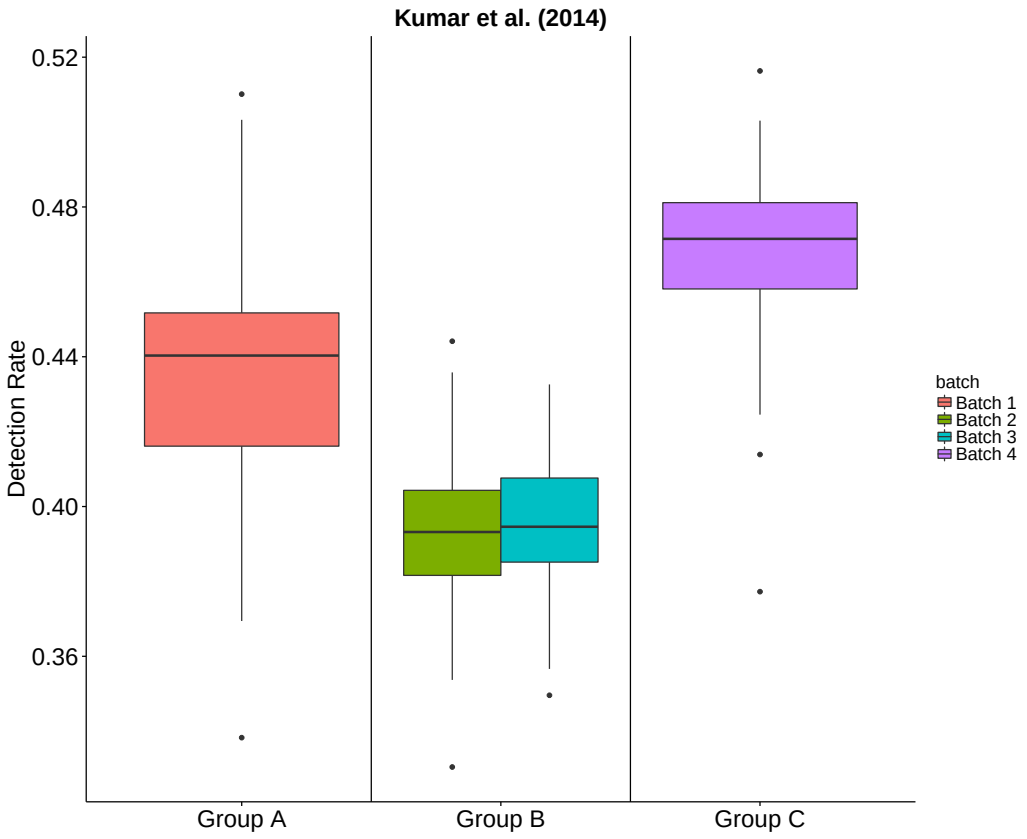
Supplemental Figure 3: Boxplots of the detection rate, or the proportion of genes in a cell reporting the normalized expression values greater than a predetermined threshold (e.g. proportion of genes in a cell such that $Z > \delta$ where $\delta = 0$ and Z is the normalized expression values).



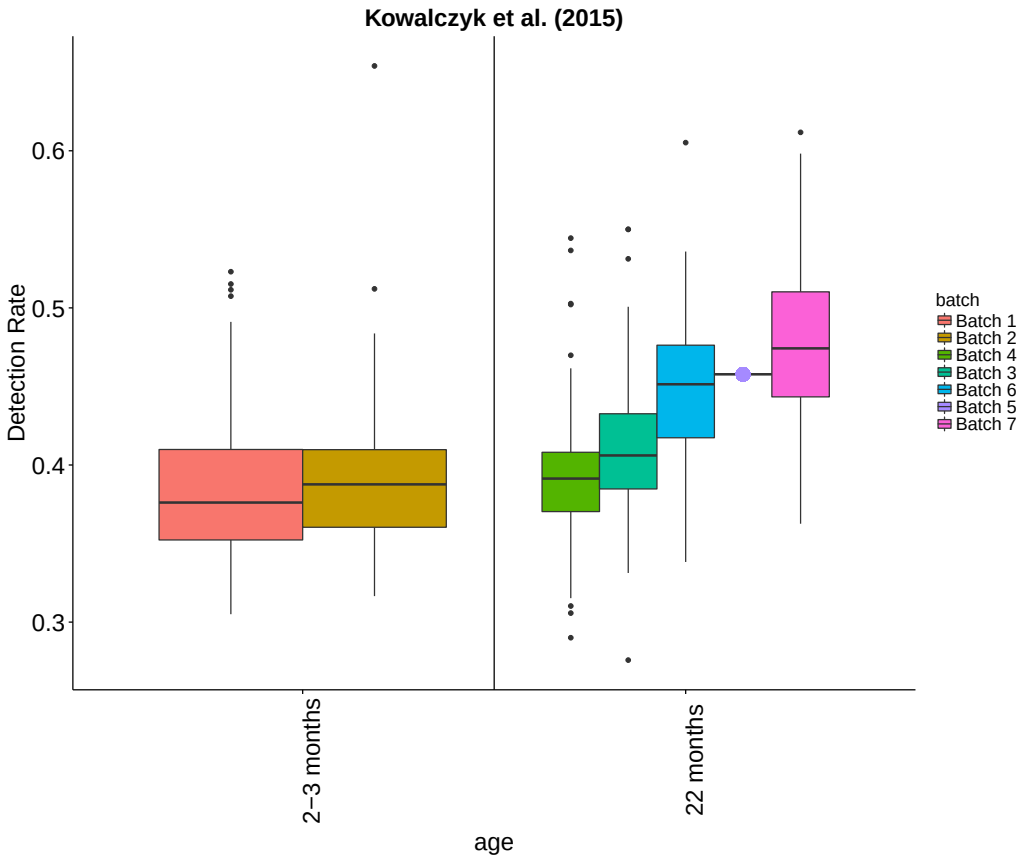
Supplemental Figure 4: Distribution of detection rate from Deng et al. (2014) [2]. The left column contains boxplots of the detection rate grouped by biological groups and colored by batch. The right column contains boxplots of the proportion of detected genes grouped by batch and colored by biological group.



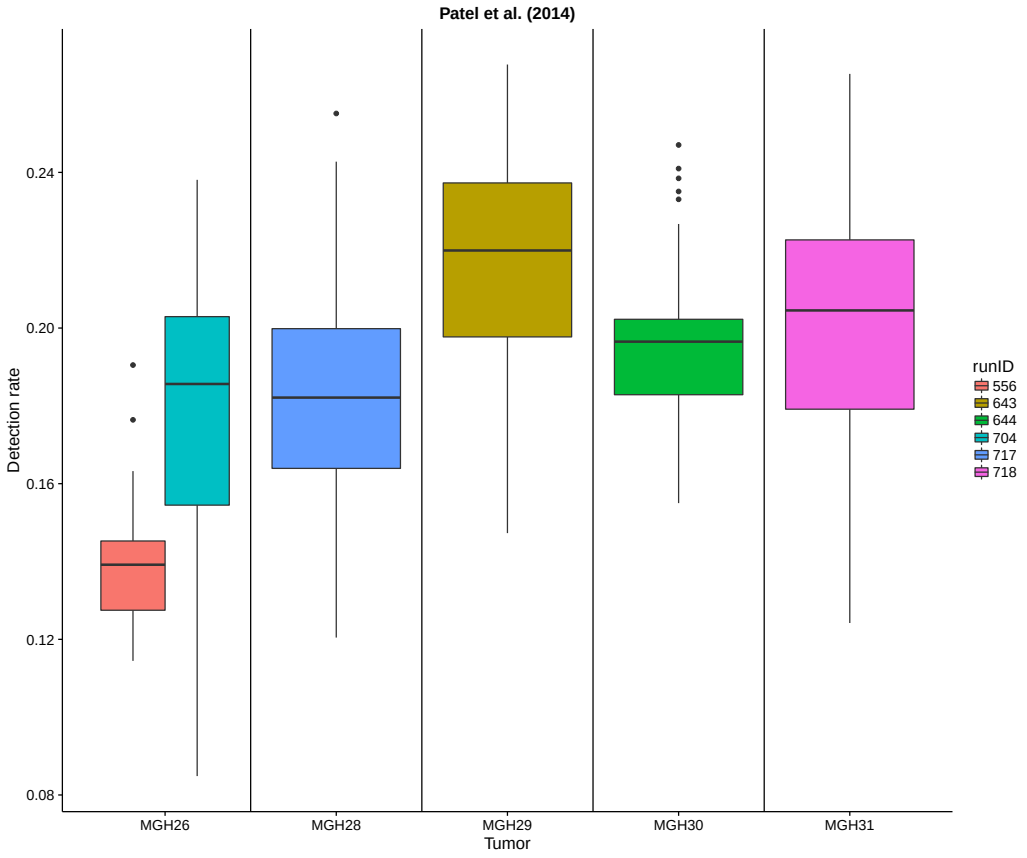
Supplemental Figure 5: Distribution of detection rate from Guo et al. (2015) [3]. The left column contains boxplots of the detection rate grouped by biological groups and colored by batch. The right column contains boxplots of the proportion of detected genes grouped by batch and colored by biological group.



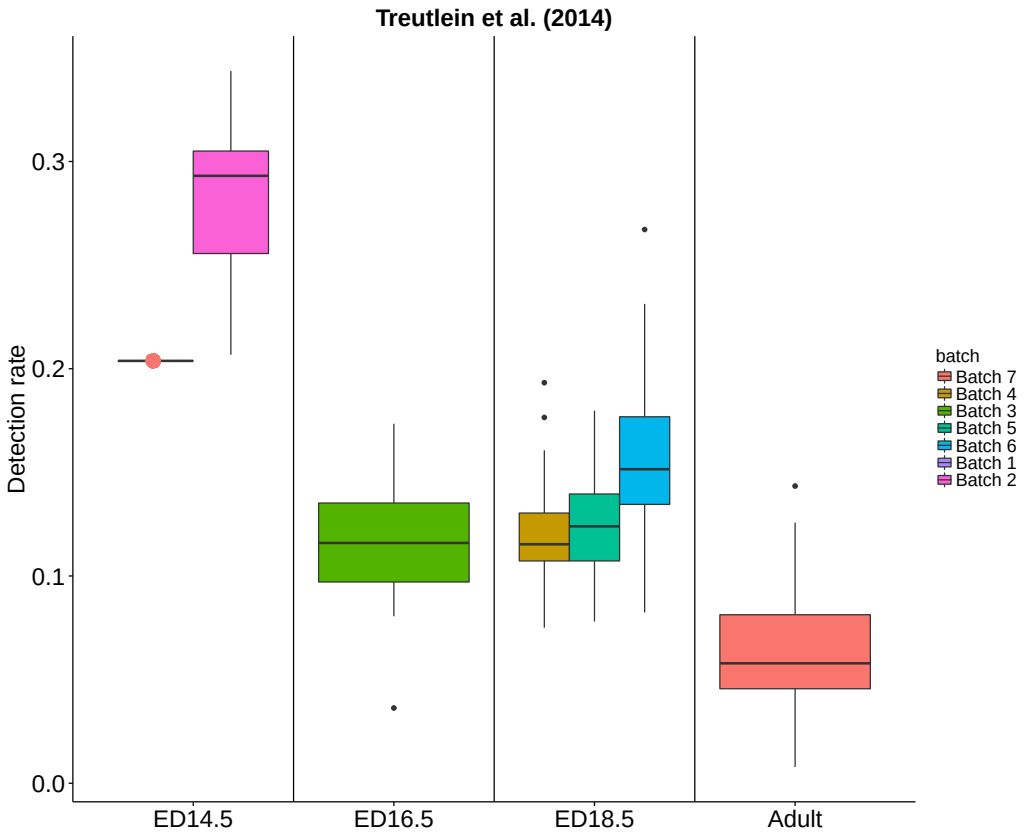
Supplemental Figure 6: Distribution of detection rate from Kumar et al. (2014) [4] colored by batch.



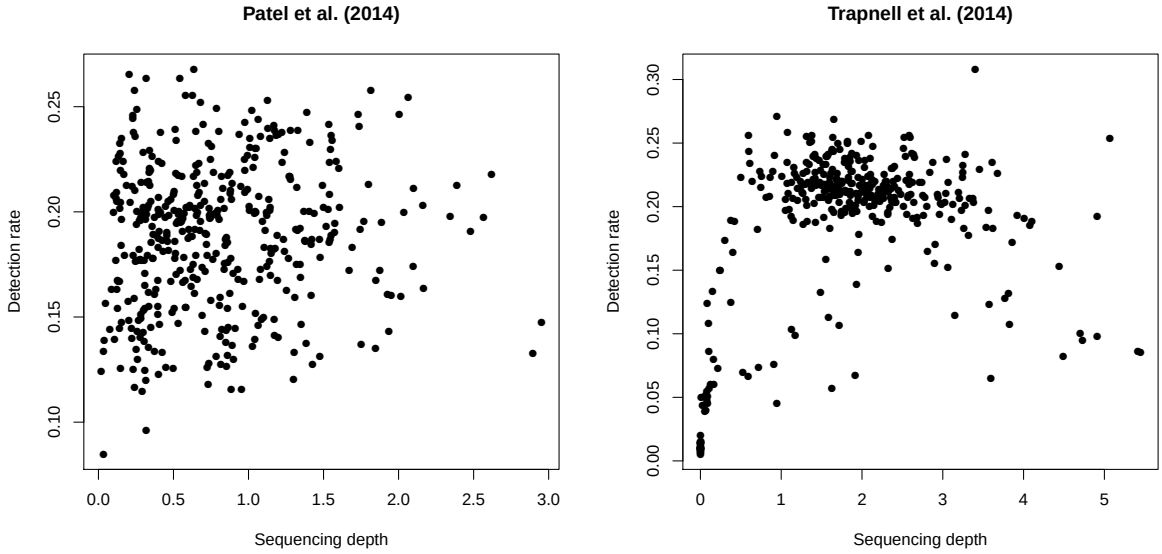
Supplemental Figure 7: Distribution of detection rate from Kowalczyk et al. (2015) [5] colored by batch.



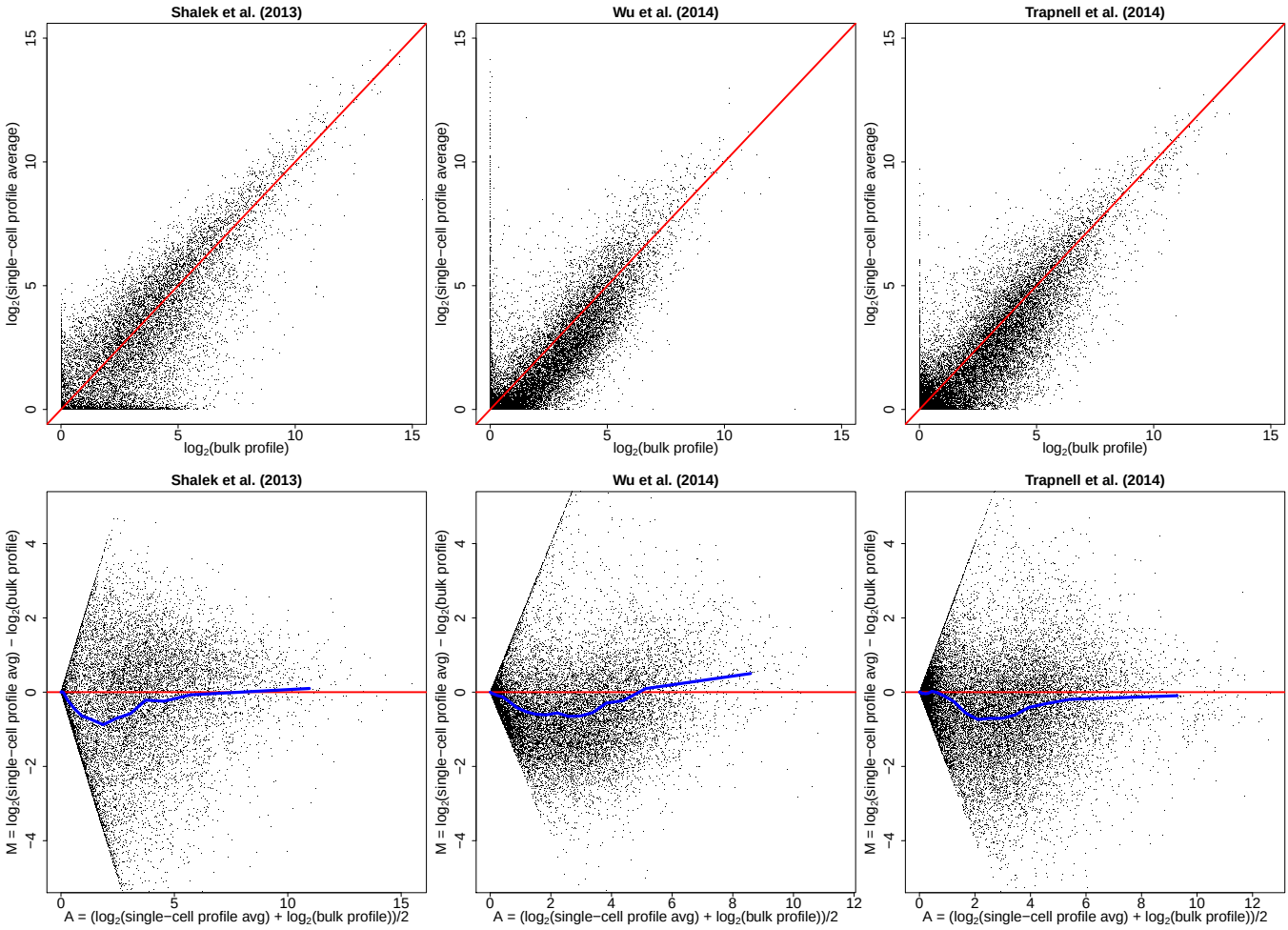
Supplemental Figure 8: Distribution of detection rate from Patel et al. (2014) [6] colored by batch.



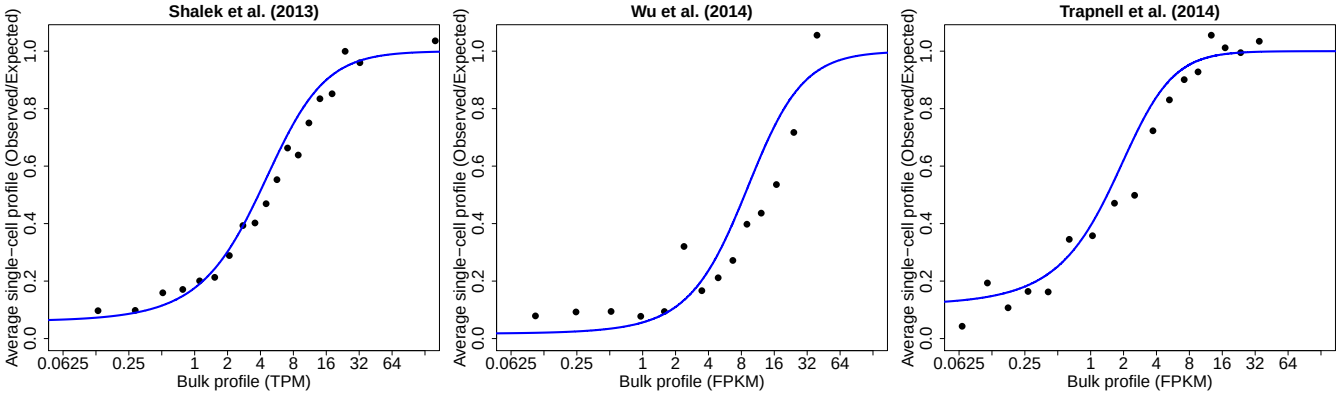
Supplemental Figure 9: Distribution of detection rate from Treutlein et al. (2014) [7] colored by batch.



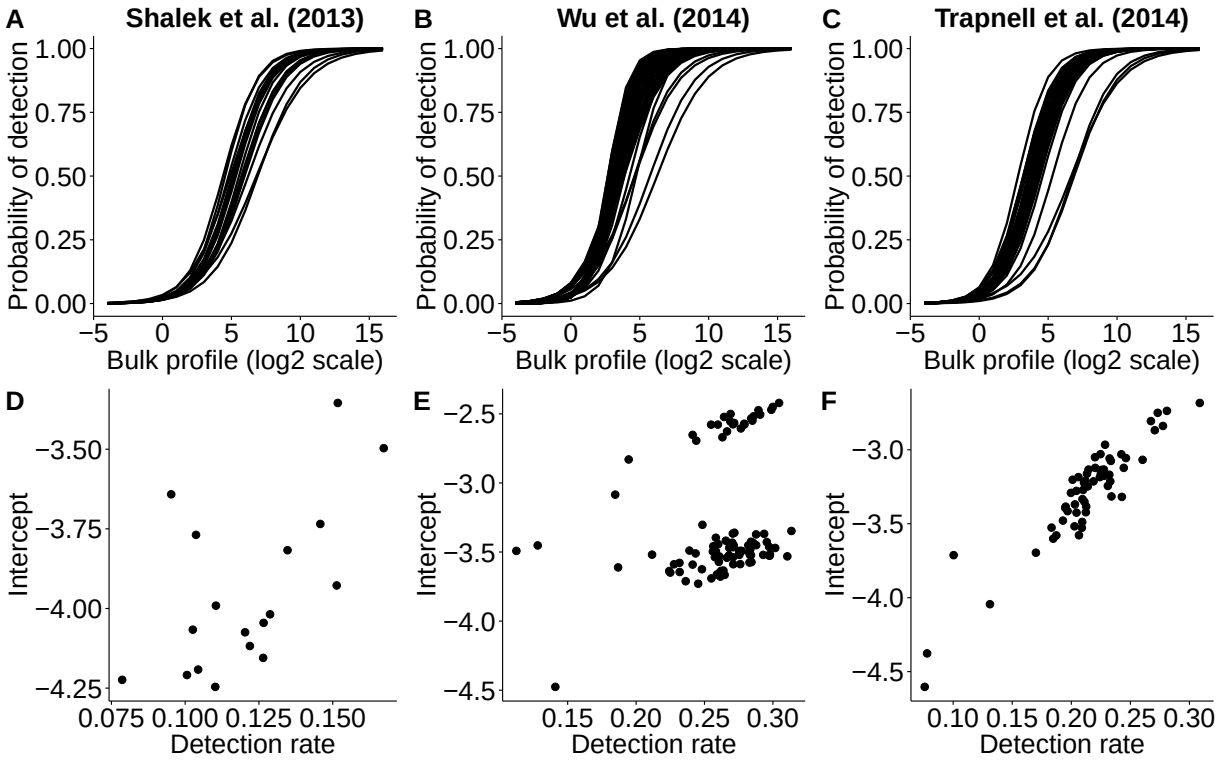
Supplemental Figure 10: Scatter plots comparing the detection rate and the sequencing depth for two scRNA-seq data sets[6, 9]. The counts and TPMs from Patel et al. (2014) were obtained from the patel2014gliohuman Github repository and the counts and TPMs from Trapnell et al. (2014) were obtained from the conquer data repository (GSE52529-GPL11154, GSE52529-GPL16791). The detection rate was calculating using TPMs $> \delta$ where $\delta = 1$. Sequencing depth was calculated as the sum of the counts across genes fore each cell.



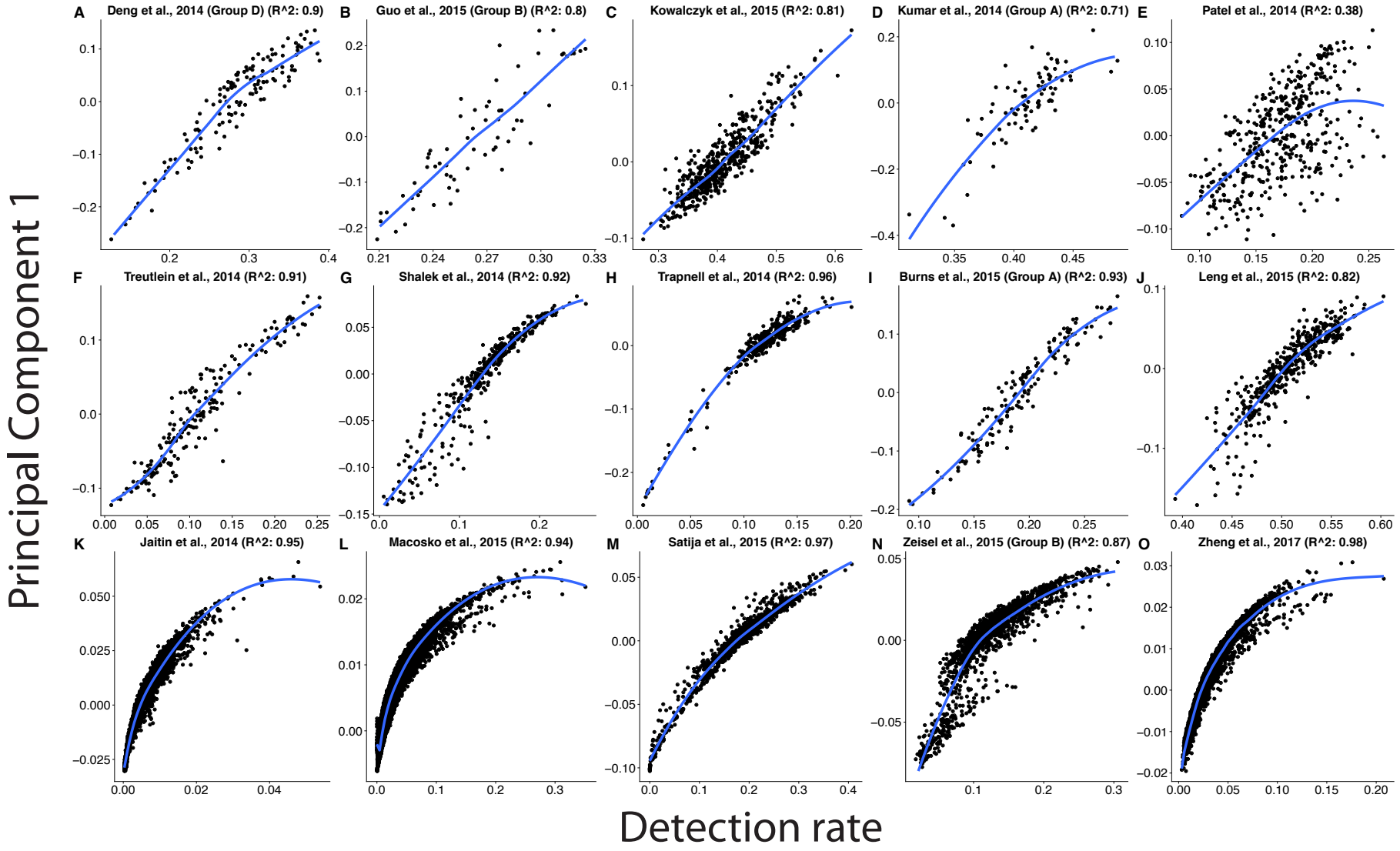
Supplemental Figure 11: **Row 1**: We reproduced scatter plots from three studies which compared bulk RNA-Seq (x-axis) to the single cell expression profile averaged across cells (y-axis). **Row 2**: M-A plot demonstrating the single cell profile averaged across cell is smaller than the bulk profile for the lowly expressed genes.



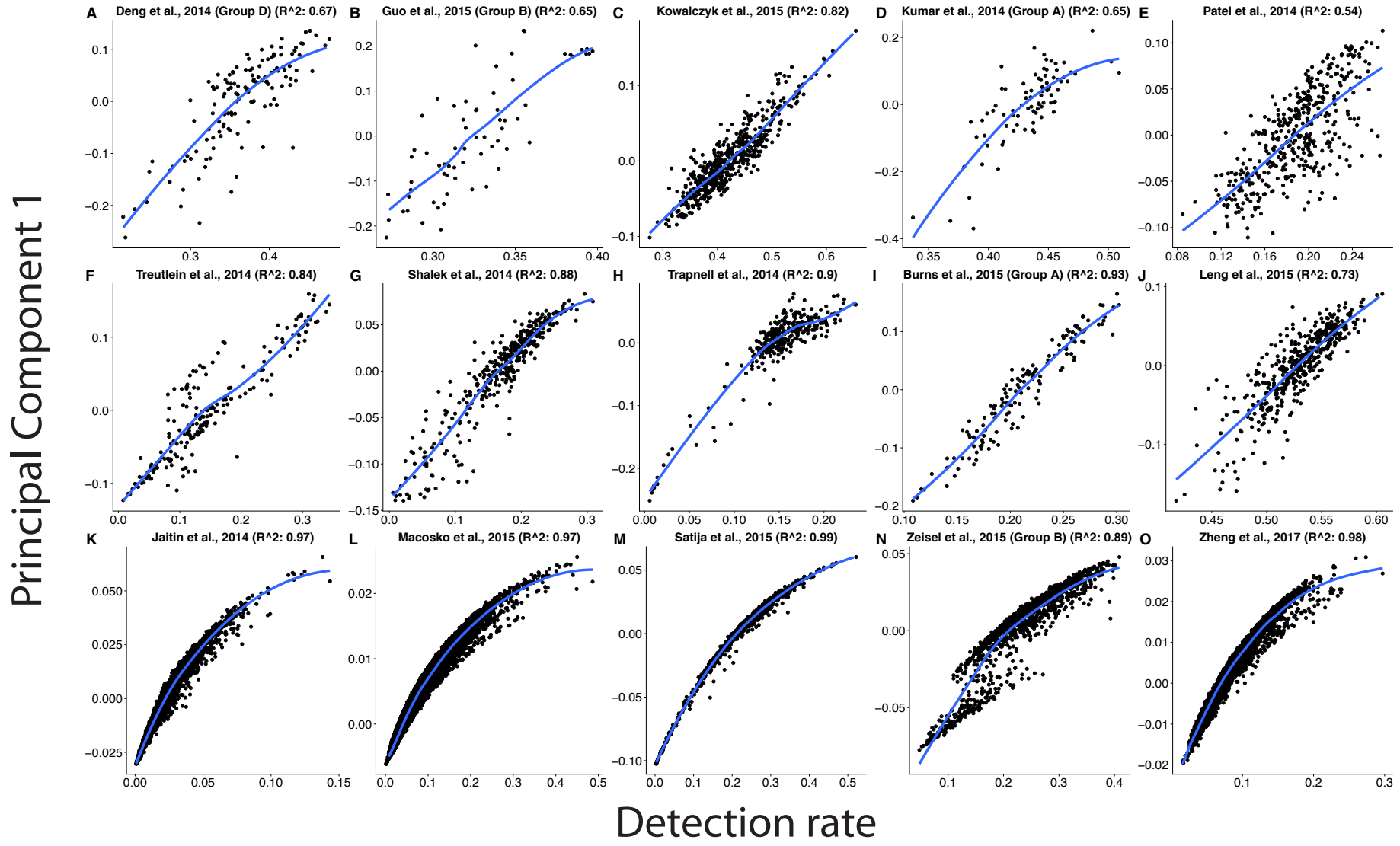
Supplemental Figure 12: Plots comparing bulk RNA-Seq (x-axis) to the ratio of the observed to expected single cell RNA-Seq profile averaged across cells (y-axis). In the lowly expressed genes, the observed scRNA-Seq profile averaged across cells is smaller than what it expected. Data was obtained from three publicly available scRNA-Seq studies that included a matched bulk RNA-Seq sample measured on the same population of cells. For each study, this technical bias can be approximated with a logistic curve (blue).



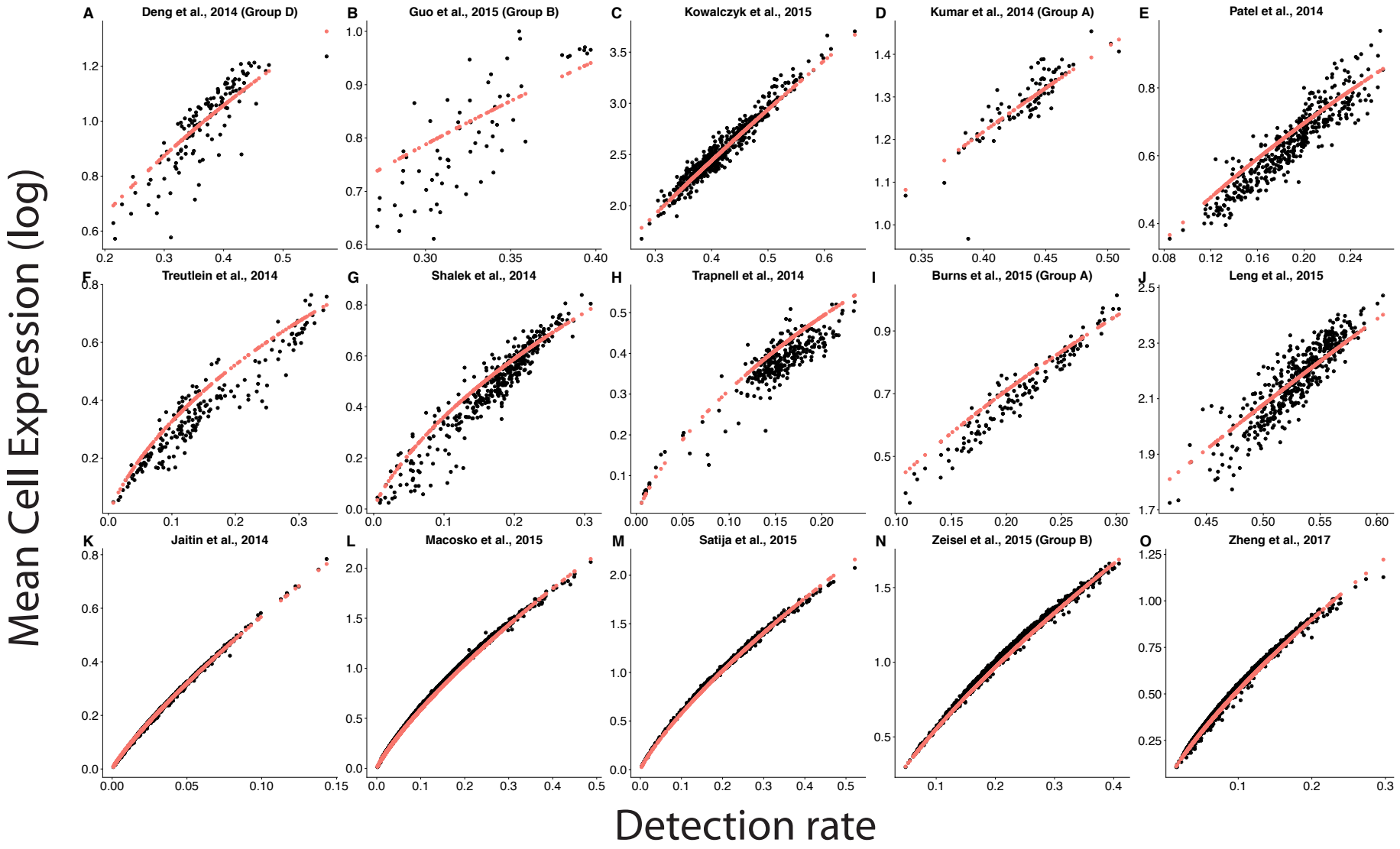
Supplemental Figure 13: Plots demonstrating the probability of a gene being detected varies cell to cell. Data was obtained from three publicly available scRNA-seq studies that included a matched bulk RNA-Seq sample measured on the same population of cells. For each cell in a single cell RNA-Seq data set, (A-C) a logistic regression model is fit to estimate the probability of a gene being detected using the bulk RNA-Seq expression. The probability that a given gene is detected estimated using logistic regression is not to be confused with the cell-specific detection rate. (D-F) However, the estimated intercepts from the logistic regression model are strongly associated to the cell-specific detection rate.



Supplemental Figure 14: First principal component is strongly associated with the detection rate using a detection threshold of $\delta = 1$. Here the detection rate is the proportion of genes in a cell reporting the normalized expression values greater than a predetermined threshold (e.g. $Z > \delta$ where $\delta = 1$ and Z is the normalized expression values).



Supplemental Figure 15: We also confirmed the relationship between the first principal component and the detection rate using a detection threshold of $\delta = 0$. Here the detection rate is the proportion of genes in a cell reporting the normalized expression values greater than a predetermined threshold (e.g. $Z > \delta$ where $\delta = 0$ and Z is the normalized expression values).



Supplemental Figure 16: Plots comparing detection rate (x-axis) to the cell-specific mean on the log scale. Even if the cell expression profiles have the same means on original scale, if we transform expression profiles to the log-scale, then the log-scale means will not be the same and it will depend on the detection rate (red). See below for mathematical details.

To explain this mathematically, consider a single cell experiment where we take a random sample of N cells from the population. Denote X_{gi} as the expression value for the g^{th} gene on i^{th} cell where

$$X_{gi} = \begin{cases} 0 & D_{gi} = 0 \\ A_{ij} & D_{gi} = 1 \end{cases}$$

Assume $A_{gi} \sim \text{logNormal}(\mu, \sigma^2)$ and denote $P[D_{gi} = 1] = p_i$ the margignal probability of detection for the i^{th} cell. Under these assumptions, $E[X_{gi}] = p_i \exp(\mu + \sigma^2/2)$. However, for the log-transformed data, we can use Taylor's series approximation to show that,

$$\begin{aligned} E[\log(X_{gi} + k)] &= (1 - p_i)\log(k) + p_i E[\log(X_{gi} + k)|D_{gi} = 1] \\ &= (1 - p_i)\log(k) + p_i E[\log(S_{gi})|D_{gi} = 1] \\ &\approx (1 - p_i)\log(k) + p_i \left[\log(\exp(\mu + \sigma^2/2) + k)\right] \end{aligned}$$

where $S_{gi} = X_{gi} + k$ is a shifted logNormal distribution with mean $E[S_{gi}] = E[X_{gi}] + k$ and variance $\text{Var}(S_{gi}) = \text{Var}(X_{gi})$ and

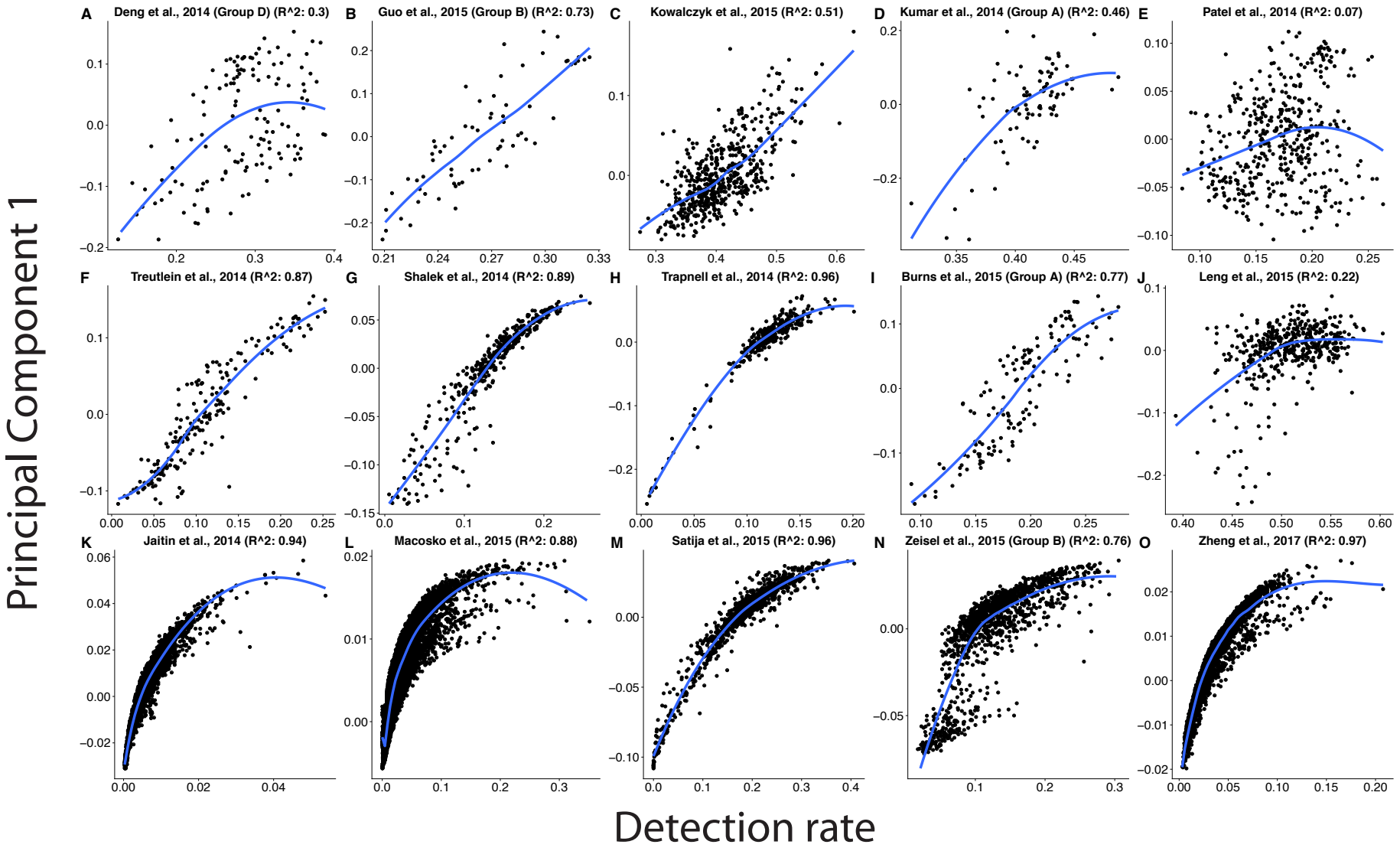
$$E[\log(S_g)] \approx \log(\mu_{S_g}) + \frac{E[S_g - \mu_{S_g}]}{\mu_{S_g}} = \log(\mu_{S_g}) + \frac{0}{\mu_{S_g}} = \log(\mu_{S_g}) \tag{1}$$

Because scRNA-Seq experiments are typically performed to assure $E(X_{gi}) = M/G$ for all cells, where G is the number of genes (or features) and M is sequencing depth (e.g. $M = 10^6$), then

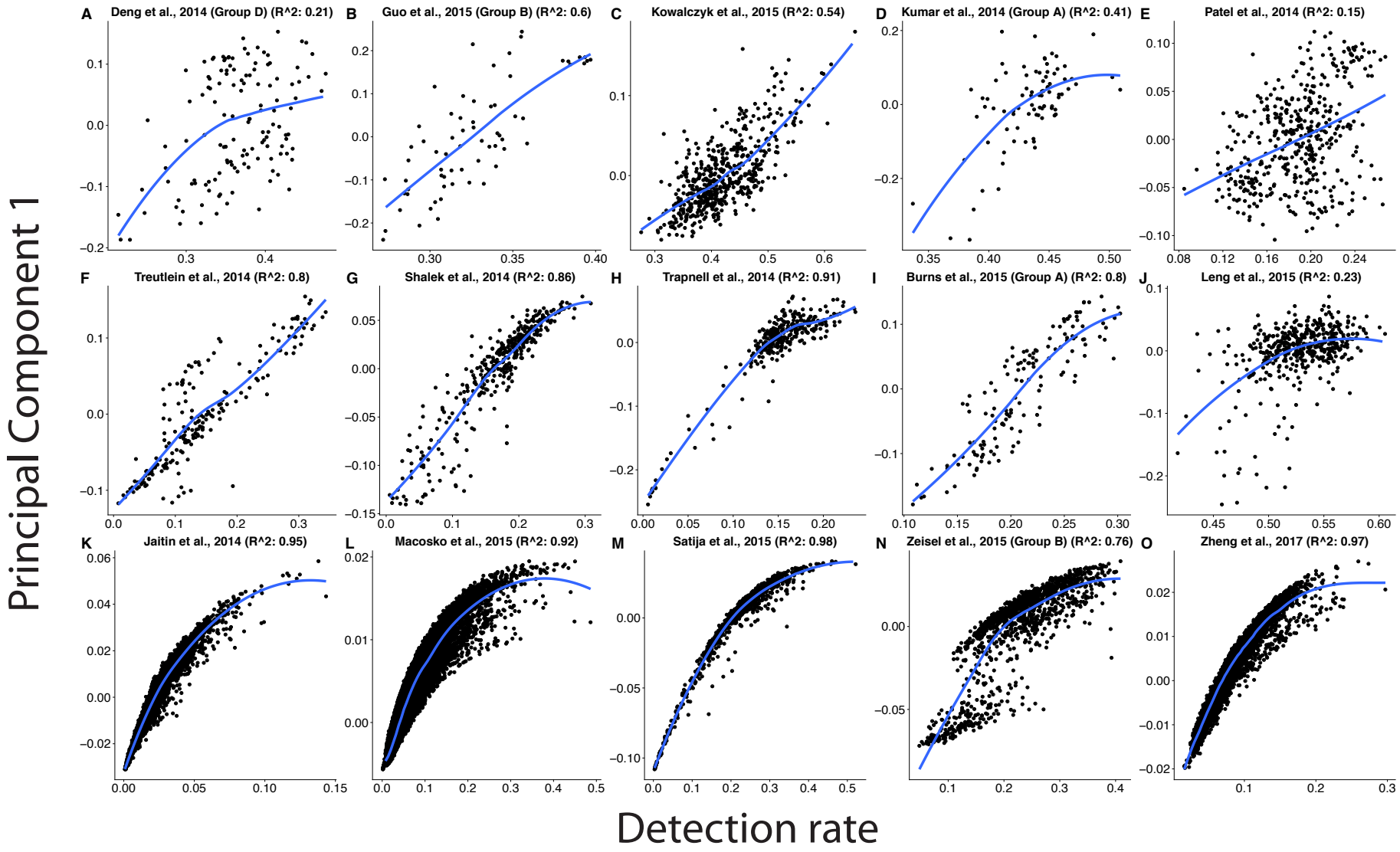
$$\mu = \log(M/G) - \log(p_i) - \sigma^2/2$$

Using this, we plug μ into the above equation to get

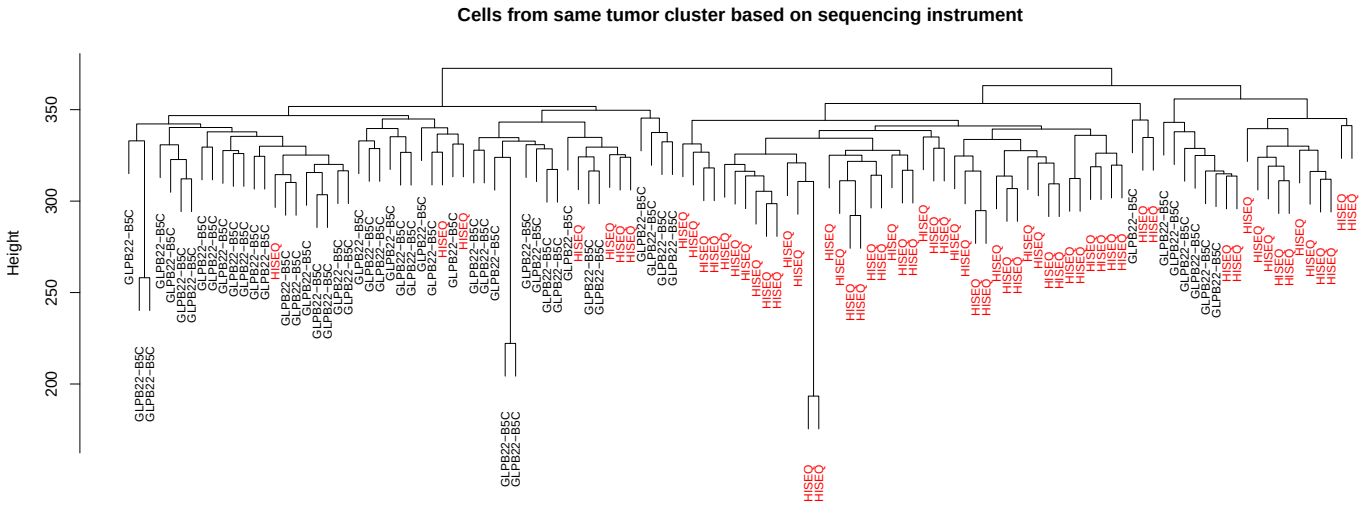
$$E[\log(X_{gi} + k)] \approx (1 - p_i) * log(k) + p_i \left[\log(\frac{M}{G * p_i} + k)\right] \tag{2}$$



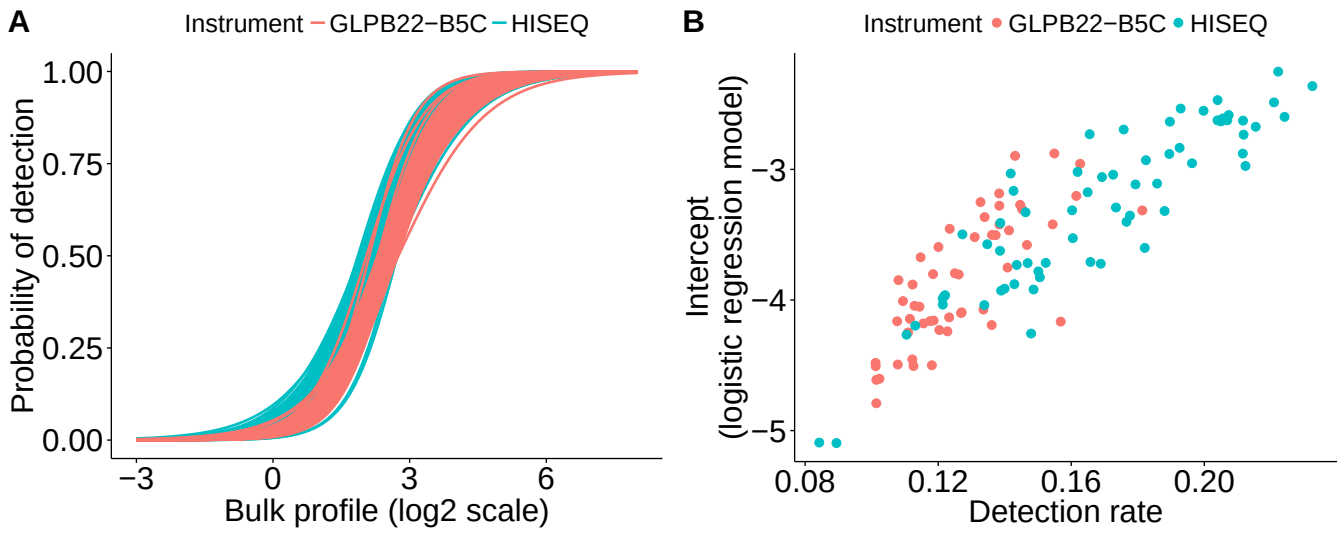
Supplemental Figure 17: First principal component is strongly associated with the detection rate (where the predetermined detection threshold is defined as $\delta = 1$) even after removing the cell-specific mean on the log-scale.



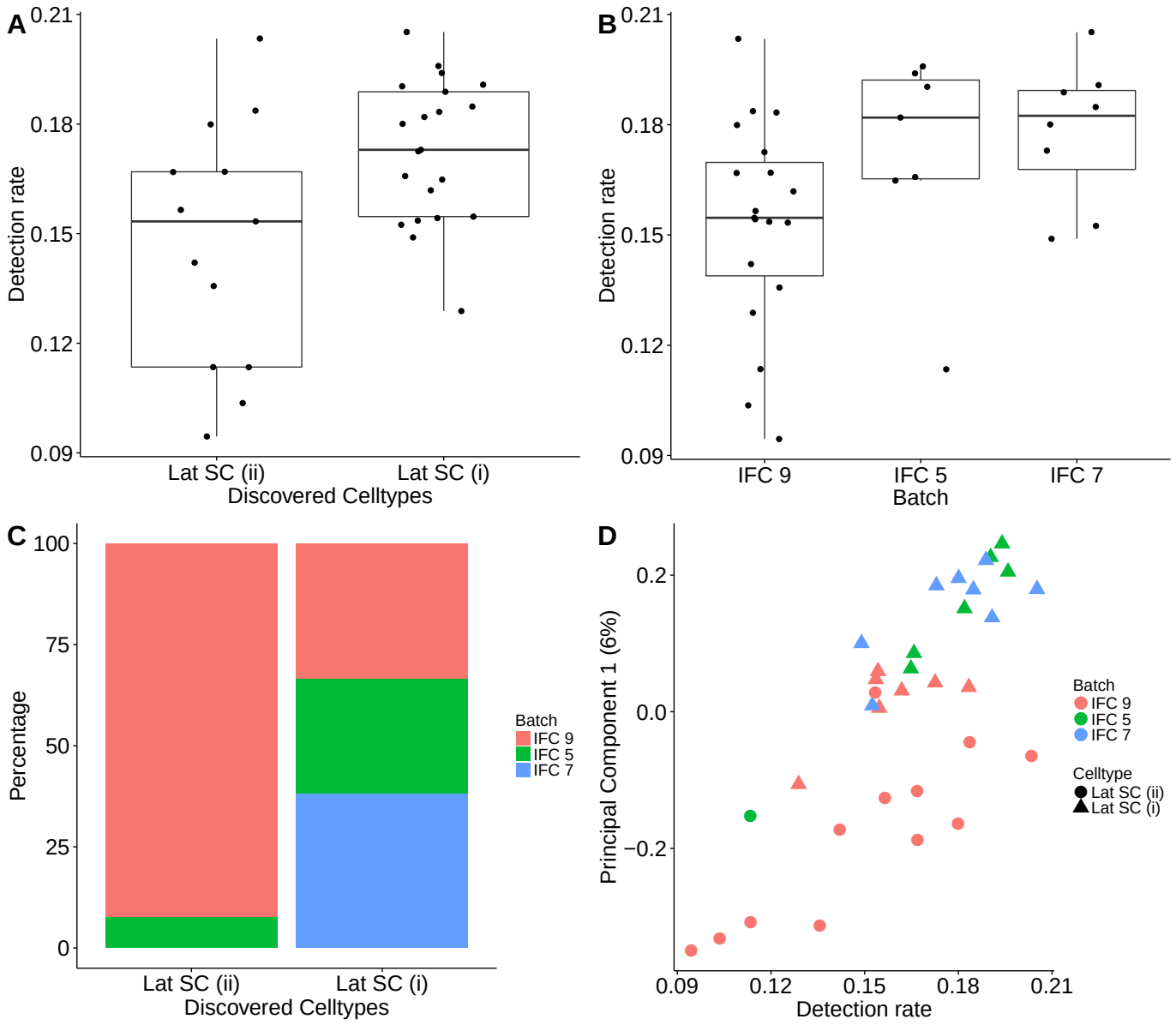
Supplemental Figure 18: First principal component is strongly associated with the detection rate (where the predetermined detection threshold is defined as $\delta = 0$) even after removing the cell-specific mean on the log-scale.



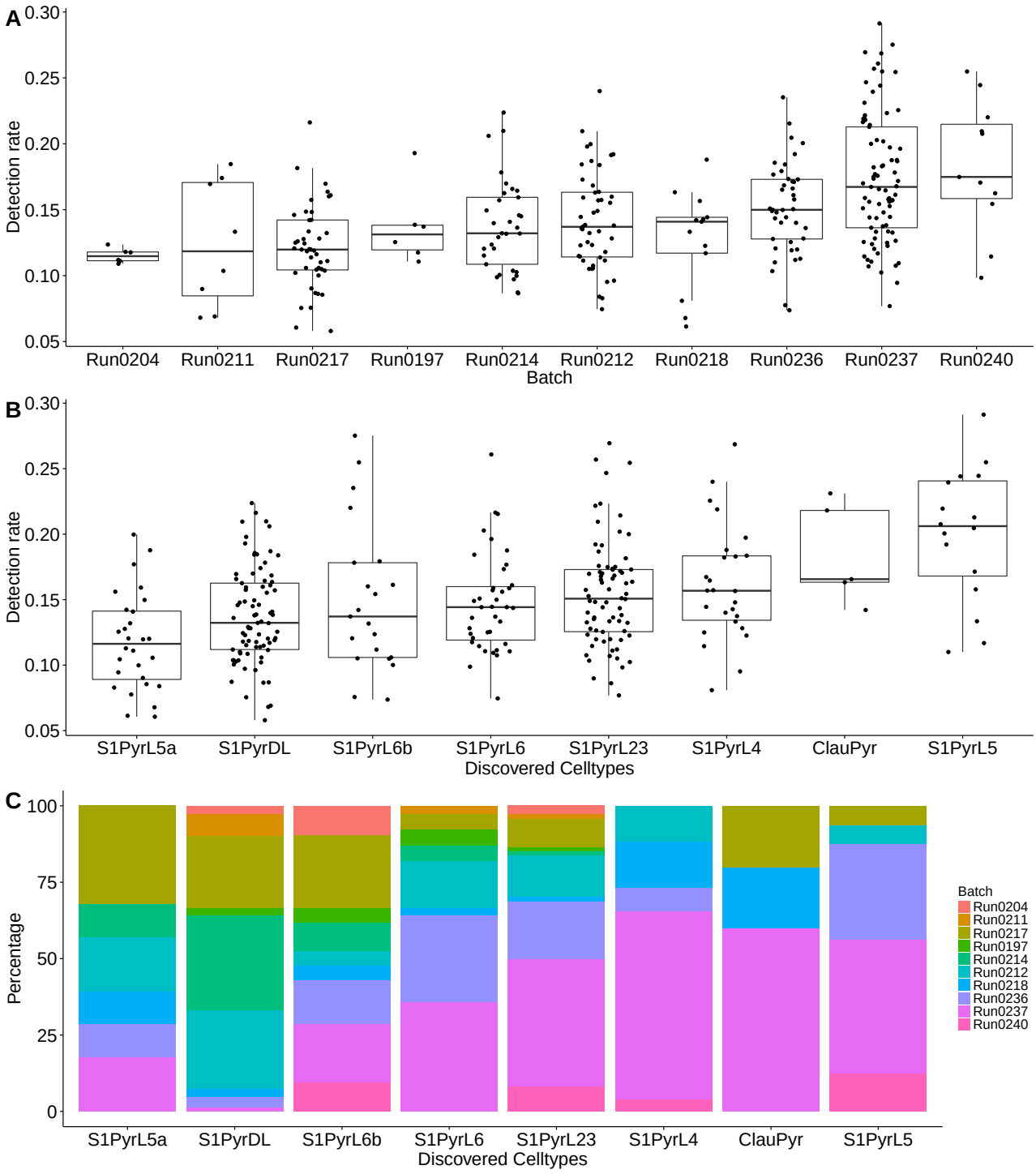
Supplemental Figure 19: Hierarchical clustering based on Euclidian distance of scRNA-Seq data measuring the expression level of 118 single cells isolated from one tumor, reveals two clusters of cells corresponding to the names of two sequencing instruments used ('GLPB22-B5C' and 'HISEQ').



Supplemental Figure 20: Illustration of how technical variation can drive differences in estimates of the probablity of detection. This is scRNA-Seq data measuring the expression level of 118 single cells isolated from one tumor and two sequencing instruments ('GLPB22-B5C' and 'HISEQ'). (A) Estimates of the probability of detection using a logistic regression model colored by the sequencing instrument. (B) The detection rates are related to the intercept estimate in the logistic regression model.



Supplemental Figure 21: (A) Boxplots of the detection rate by the discovered celltypes from Burns et al. (2015). (B) Boxplots of the detection rate by integrated fluidics circuit chips (IFCs) using a Fluidigm platform. (C) Percentage of cells from each IFC for the two discovered cell types. The colors respresent the IFCs and are ordered by the median of the detection rate for each IFC. (D) Relationship between the first PC and the detection rate colored by IFC and shape representing the discovered cell type.



Supplemental Figure 22: (A) Boxplots of the detection rate by the discovered celltypes from Zeisel et al. (2015). (B) Boxplots of the detection rate by runID in the header of FASTQ file. (C) Percentage of cells from each runID for the two discovered cell types.The colors respresent the runIDs and are ordered by the median of the detection rate for each runID.

Deng et al. (2014) [2] performed single-end single-cell RNA-Seq in 286 mouse developmental cells (ranging from zygote to late blastocyst) to study monoallelic expression (GSE45719). The main purpose of this study was to investigate monoallelic gene expression in mouse embryos, but here we consider the different developmental stages (oocyte to blastocyst) as the biological condition as an example. The authors used RPKMs for normalization. The study design is provided in Table 1. Within each runID, cells were processed across multiple flow cell lanes. For the purpose of this manuscript, the biological groups were binned 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 1: Study design by Deng et al. (2014). Number of cells sequenced across the developmental stages and across batches.

Guo et al. (2015) [3] performed paired-end single-cell RNA-Seq in 154 human primordial germ cells from the migrating stage to the gonadal stage (GSE63818). The authors used FPKMs for normalization. The study design is provided in Table 2. For the purpose of this manuscript, the stages were binned into three groups: Group A (4 weeks), Group B (7-8 weeks) and Group C (10-11 weeks). We excluded the 17-19 weeks.

	Batch				Biological Group		
	mi	runID	fc	lane	Group 1 (4 weeks)	Group 2 (7-8 weeks)	Group 3 (10-11 weeks)
Batch 1	D2SV54V1	212	HA63CADXX	1	2	6	0
Batch 2	D2SV54V1	214	HA5F6ADXX	1	0	12	0
Batch 3	D2SV54V1	221	H9V90ADXX	1	6	0	12
Batch 4	D2SV54V1	257	HBDFTADXX	1	0	0	9
Batch 5	HWI-ST1352	210	h9v15adxx	1	0	1	0
Batch 6	HWI-ST1352	212	HA5EKADXX	2	0	10	0
Batch 7	HWI-ST1352	214	HA5EGADXX	1	0	9	0
Batch 8	HWI-ST1352	219	H9V56ADXX	1	10	0	0
Batch 9	HWI-ST1352	221	H9V62ADXX	1	0	11	0
Batch 10	HWI-ST1352	228	HBD3JADXX	1	0	10	13
Batch 11	HWI-ST1352	255	HBDF3ADXX	1	0	0	1
Batch 12	HWI-ST1352	255	HBDF3ADXX	2	0	0	4
Batch 13	HWI-ST1352	260	hbdeaadxx	1	0	0	15
Batch 14	HWI-ST1352	260	hbdeaadxx	2	0	7	16

Table 2: Study design by Guo et al. (2015). Number of cells sequenced across stages and across batches. Abbreviations: machine identifier (mi), flow cell (fc), NA (missing information in FASTQ header).

Kowalczyk et al. (2015) [5] performed paired-end single-cell RNA-Seq in 533 short-term hematopoietic stem cells (HSCs) from young (2-3 months) and old (>22 months) C57BL/6 mice (GSE59114). We focused on the C57BL/6 mouse strain in this analysis. The authors used TPMs for normalization. The study design is provided in Table 3.

	Batch				Biological Group	
	mi	runID	fc	lane	Group 1 (2-3 months)	Group 2 (22 months)
Batch 1	NA	NA	C28P5ACXX130621	1	78	0
Batch 2	NA	NA	C28P5ACXX130621	2	85	0
Batch 3	NA	NA	C2JV4ACXX131001	3	0	95
Batch 4	NA	NA	C2JV4ACXX131001	4	0	93
Batch 5	NA	NA	C2JV4ACXX131001	5	0	1
Batch 6	SL-HDD	H7UB9ADXX140425	H7UB9ADXX	1	0	96
Batch 7	SL-HDD	H7UB9ADXX140425	H7UB9ADXX	2	0	95

Table 3: Study design by Kowalczyk et al. (2015). Number of cells sequenced across two ages and across batches. Abbreviations: machine identifier (mi), flow cell (fc), NA (missing information in FASTQ header). The machine id and runID in the sequence identifer information in the FASTQ files were missing in Batches 1-5.

Kumar et al. (2014) [4] performed paired-end single-cell RNA-Seq in 361 mouse embryonic stem cells (mESCs) from two mouse strains (v6.5 mESCs and Dgcr8 -/- mESCs) and two cultures (serum+LIF media and 2i+LIF media) (GSE60749). The authors used TPMs for normalization. The study design is provided in Table 4. For the purposes of this analysis, the combinations of mouse strains and culture conditions were considered seprately: Group A (Dgcr8 mouse with serum+LIF culture condition), Group B (v6.5 mouse with serum+LIF culture condition) and Group C (v6.5 mouse with 2i+LIF culture condition).

	Batch				Biological Group		
	mi	runID	fc	lane	Group 1 (Dgcr8, serum+LIF)	Group 2 (v6.5, serum+LIF)	Group 3 (v6.5, 2i+LIF)
Batch 3	NA	NA	C2A3PACXX130727	7	84	0	0
Batch 1	NA	NA	C2A3PACXX130727	1	0	90	0
Batch 4	SL-HDD	H8VC9ADXX140602	H8VC9ADXX	2	0	93	0
Batch 2	NA	NA	C2A3PACXX130727	5	0	0	94

Table 4: Study design by Kumar et al. (2014). Number of cells sequenced across strains/culture conditions and across batches. Abbreviations: machine identifier (mi), flow cell (fc), NA (missing information in FASTQ header).

Patel et al. (2014) [6] performed paired-end single-cell RNA-Seq in 430 cells from five glioblastoma tumors (GSE57872). The study design is provided in Table 5. 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 5: 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) [7] 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 6. 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 6: 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.

Shalek et al. (2014) [8] 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 7.

	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 7: 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).

Trapnell et al. (2014) [9] performed paired-end single-cell RNA-Seq in 306 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 8. 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 8: 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).

References

[1] Ron Edgar, Michael Domrachev, and Alex E Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic Acids Res* 30.1 (2002), pp. 207–10.

[2] 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.

[3] Fan Guo et al. "The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells". In: *Cell* 161.6 (2015), pp. 1437–52. DOI: 10.1016/j.cell.2015.05.015.

[4] Roshan M Kumar et al. "Deconstructing transcriptional heterogeneity in pluripotent stem cells". In: *Nature* 516.7529 (2014), pp. 56–61. DOI: 10.1038/nature13920.

[5] Monika S Kowalczyk et al. "Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells". In: *Genome Res* (2015). DOI: 10.1101/gr.192237.115.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Joseph C Burns et al. "Single-cell RNA-Seq resolves cellular complexity in sensory organs from the neonatal inner ear". In: *Nat Commun* 6 (2015), p. 8557. DOI: 10.1038/ncomms9557.

[11] Ning Leng et al. "Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments". In: *Nat Methods* 12.10 (2015), pp. 947–50. DOI: 10.1038/nmeth.3549.

[12] Rahul Satija et al. "Spatial reconstruction of single-cell gene expression data". In: *Nat Biotechnol* 33.5 (2015), pp. 495–502. DOI: 10.1038/nbt.3192.

[13] Amit Zeisel et al. "Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq". In: *Science* 347.6226 (2015), pp. 1138–42. DOI: 10.1126/science.aaa1934.

[14] Grace X Y Zheng et al. "Massively parallel digital transcriptional profiling of single cells". In: *Nat Commun* 8 (2017), p. 14049. DOI: 10.1038/ncomms14049.

[15] Fuchou Tang et al. "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nat Methods* 6.5 (2009), pp. 377–82. DOI: 10.1038/nmeth.1315.

[16] Saiful Islam et al. “Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq”. In: *Genome Res* 21.7 (2011), pp. 1160–7. DOI: 10.1101/gr.110882.110.

[17] Saiful Islam et al. “Quantitative single-cell RNA-seq with unique molecular identifiers”. In: *Nat Methods* 11.2 (2014), pp. 163–6. DOI: 10.1038/nmeth.2772.

[18] Daniel Ramsköld et al. “Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells”. In: *Nat Biotechnol* 30.8 (2012), pp. 777–82. DOI: 10.1038/nbt.2282.

[19] Simone Picelli et al. “Smart-seq2 for sensitive full-length transcriptome profiling in single cells”. In: *Nat Methods* 10.11 (2013), pp. 1096–8. DOI: 10.1038/nmeth.2639.

[20] Tamar Hashimshony et al. “CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification”. In: *Cell Rep* 2.3 (2012), pp. 666–73. DOI: 10.1016/j.celrep.2012.08.003.

[21] Diego Adhemar Jaitin et al. “Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types”. In: *Science* 343.6172 (2014), pp. 776–9. DOI: 10.1126/science.1247651.

[22] Evan Z Macosko et al. “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets”. In: *Cell* 161.5 (2015), pp. 1202–14. DOI: 10.1016/j.cell.2015.05.002.

[23] Sayantan Bose et al. “Scalable microfluidics for single-cell RNA printing and sequencing”. In: *Genome Biol* 16 (2015), p. 120. DOI: 10.1186/s13059-015-0684-3.

[24] 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.

[25] 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.

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