SUPPLEMENTARY INFORMATION

Daniel Hebenstreit[1], Miaoqing Fang[2], Muxin Gu[1], Varodom Charoensawan[1], Alexander van Oudenaarden[3], Sarah A. Teichmann[1]

[1]Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge, CB20QH, UK

[2]Department of Biological Engineering, Massachusetts Institute of Technology, MA02139, USA.

[3]Department of Physics, Massachusetts Institute of Technology, MA02139, USA.

Table of contents

Supplementary figures

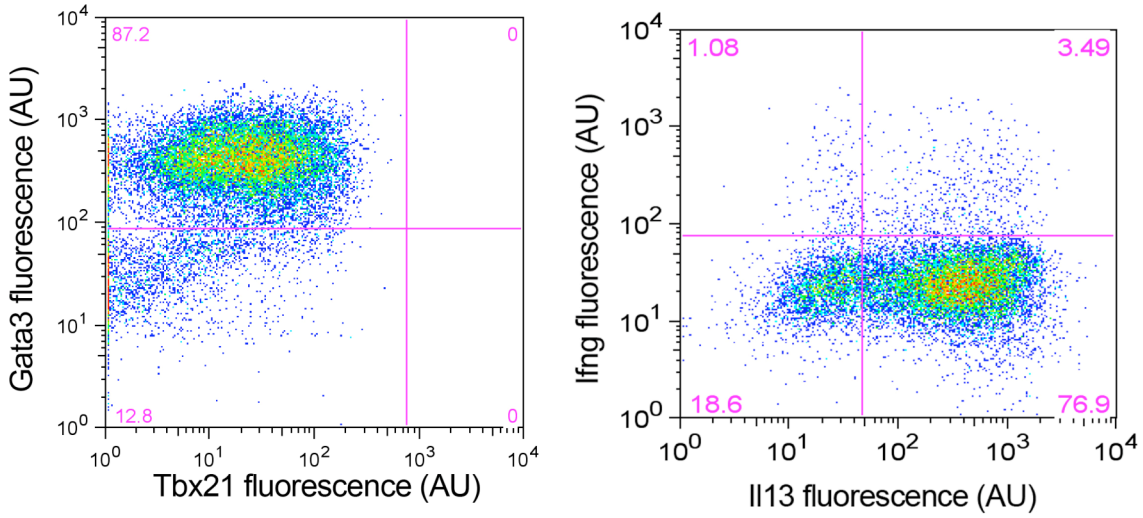


Figure S1. Th2 cells were stained by intracellular staining with anti-Gata3, anti-Tbx21, anti-Ifng, and anti-Il13 antibodies and analyzed by FACS. Gata3 and Il13 are markers of Th2 differentiation, so a high proportion of Gata3 and Il13 expressing cells indicates a high level of Th2 homogeneity in the cell population. Tbx21 and Ifng are markers of Th1 cells, and are shown as a control. Each dot represents a single cell with fluorescence intensities for the two antibody stains on the x- and y-axes. Overlapping dots change color to indicate the density of cells at that point. The purple lines separate the plots into four regions each, depending on whether cells are expressing or the proteins or not. ~80 to 90% purity was routinely achieved, indicating successful Th2 differentiation.

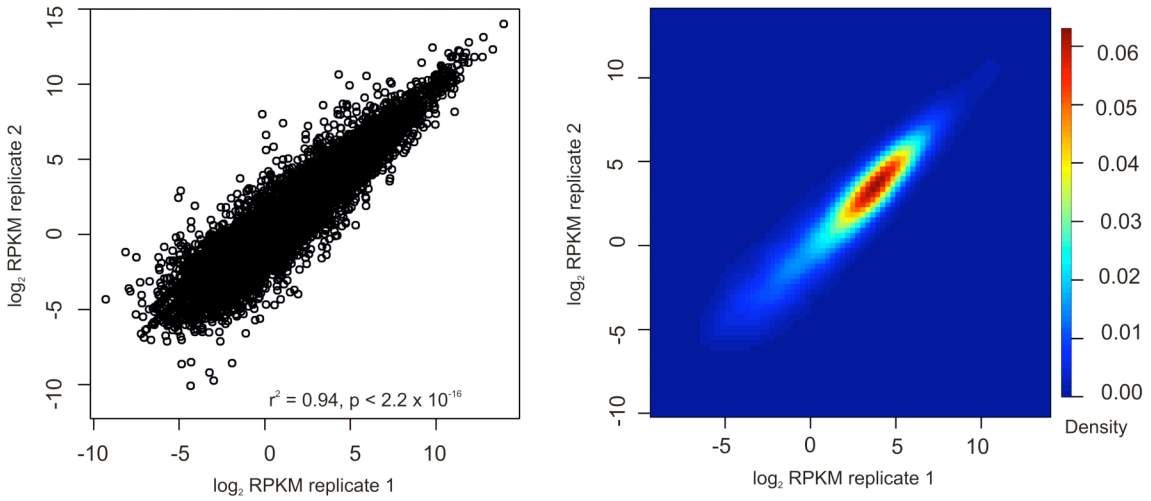


Figure S2. Correlation between two RNA-seq replicates. A scatter plot (left) and a 2-D kernel density estimation are shown (right). Correlation coefficient and significance of correlation are inset in the left panel.

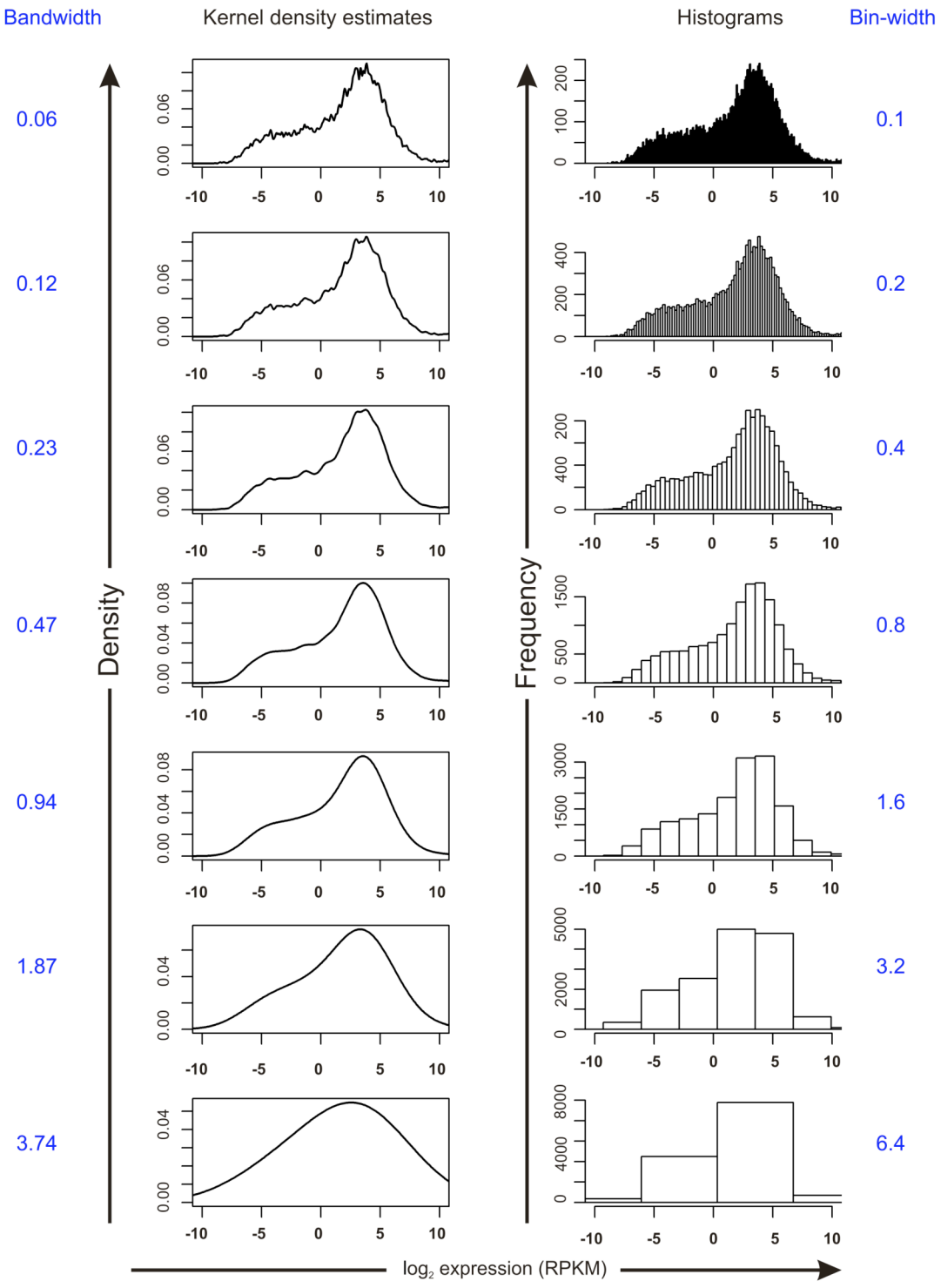


Figure S3. Examples of how different visualization methods affect the appearance of the RNA-seq data. The left panel corresponds to kernel density estimates (KDE). To demonstrate that the structure of the data is conserved under different settings, the

bandwidth (corresponding to the standard deviation of the Gaussian kernel) was increased in 2-fold steps from top to bottom (blue, left side). The bandwidth in the center corresponds to Silverman's 'rule of thumb'. The right panel shows histograms with different bin-sizes (indicated in blue on the right side). The structure of the data is conserved if the bin-size is less than the distance between the two peaks.

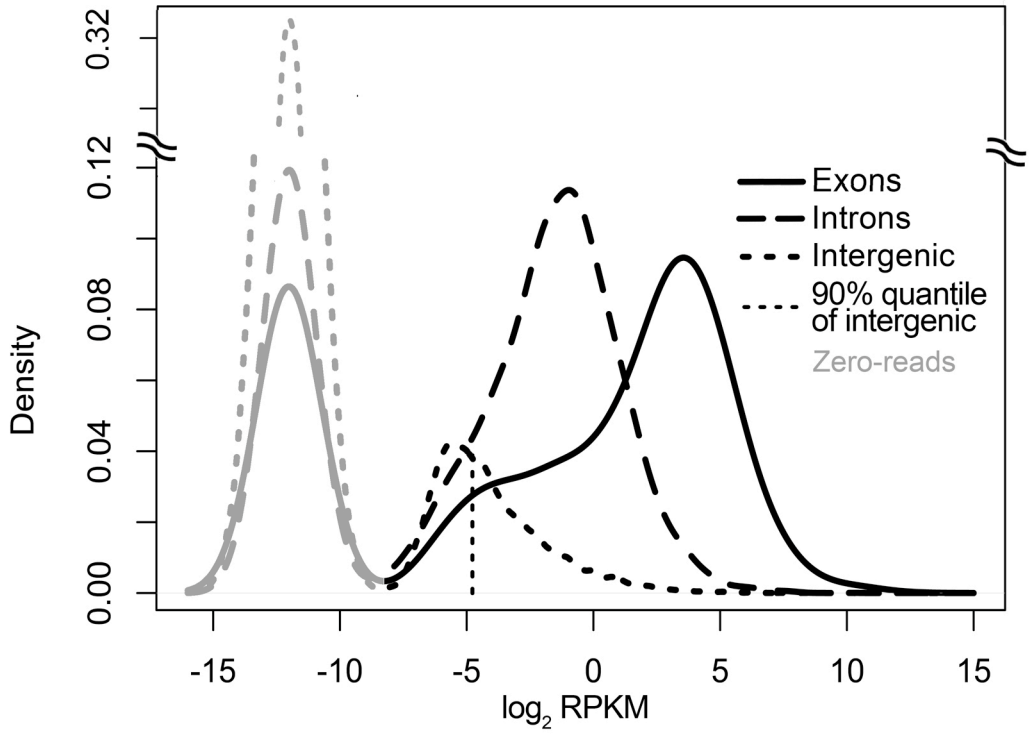


Figure S4. Kernel density estimates of RPKM distributions of RNA-seq data within exons, introns and intergenic regions as in Figure 1A. To indicated the fractions of fragments/genes with zero reads (grey), they were assigned random RPKM values, drawn from a normal distribution with mean = -12 and standard-deviation = 1 on the $\log_2$ scale.

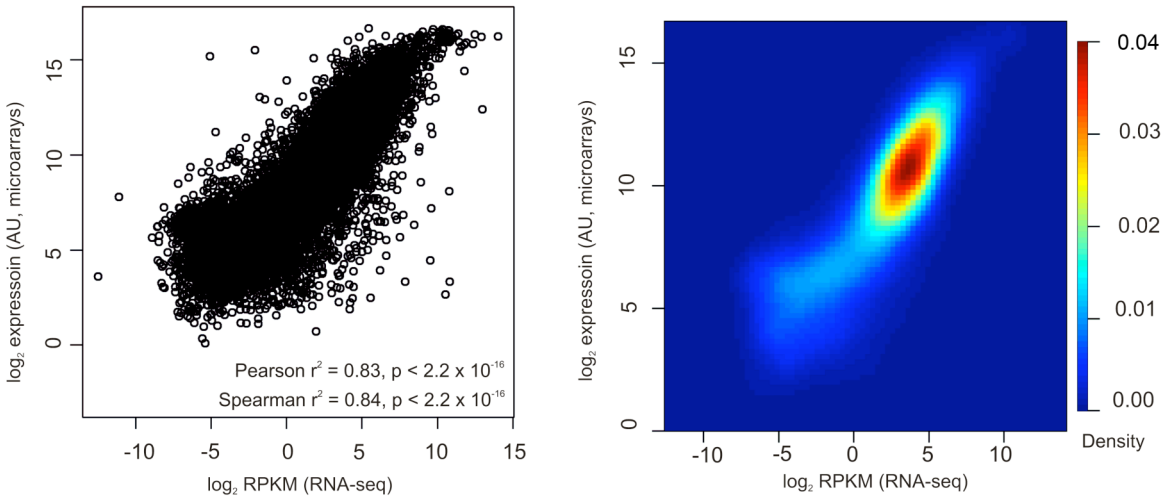


Figure S5. Correlation between RNA-seq and microarray data (Wei et al, 2009). A scatter plot (left) and a 2-D kernel density estimation are shown (right). Correlation coefficients and significance of correlations are inset in the left panel.

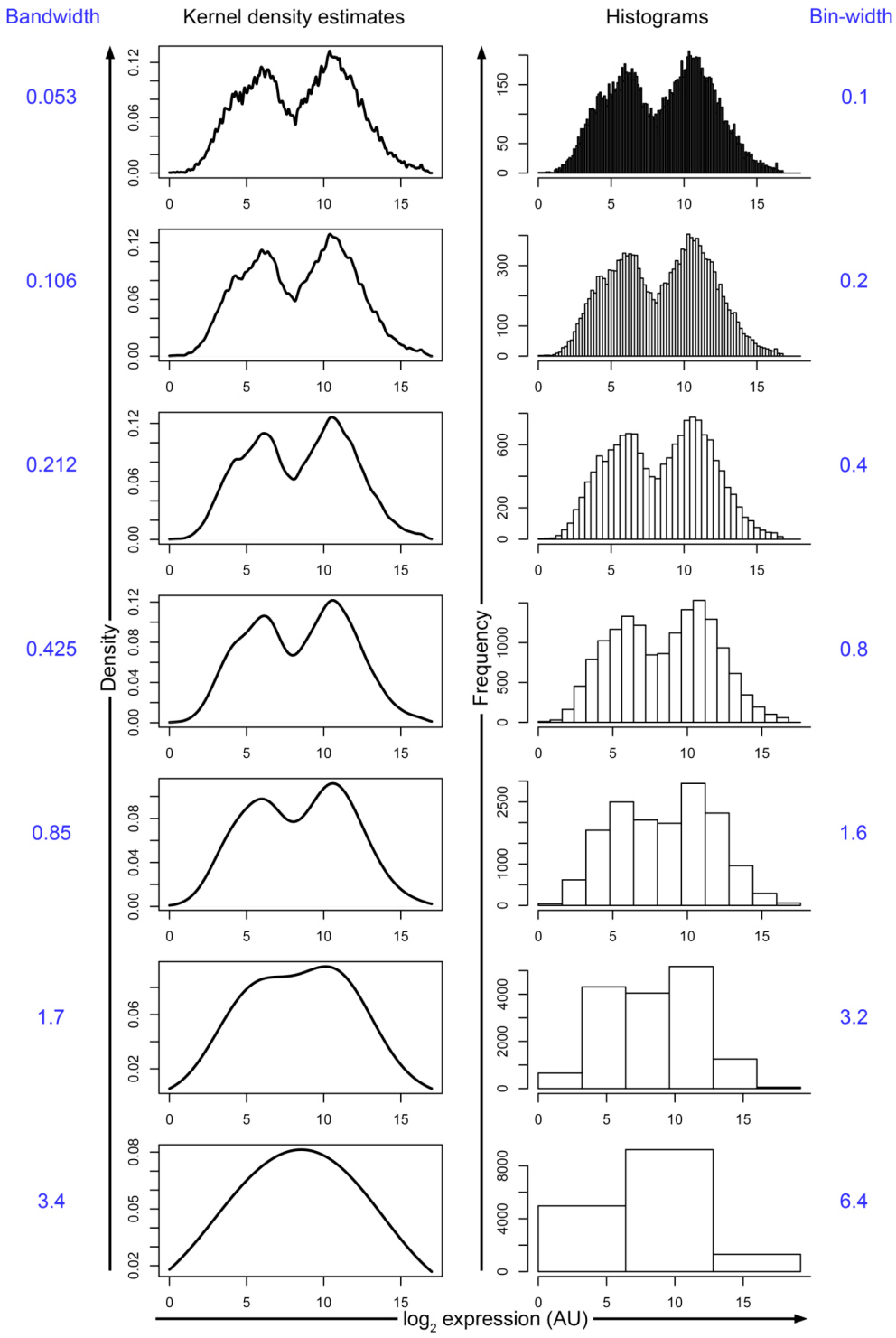
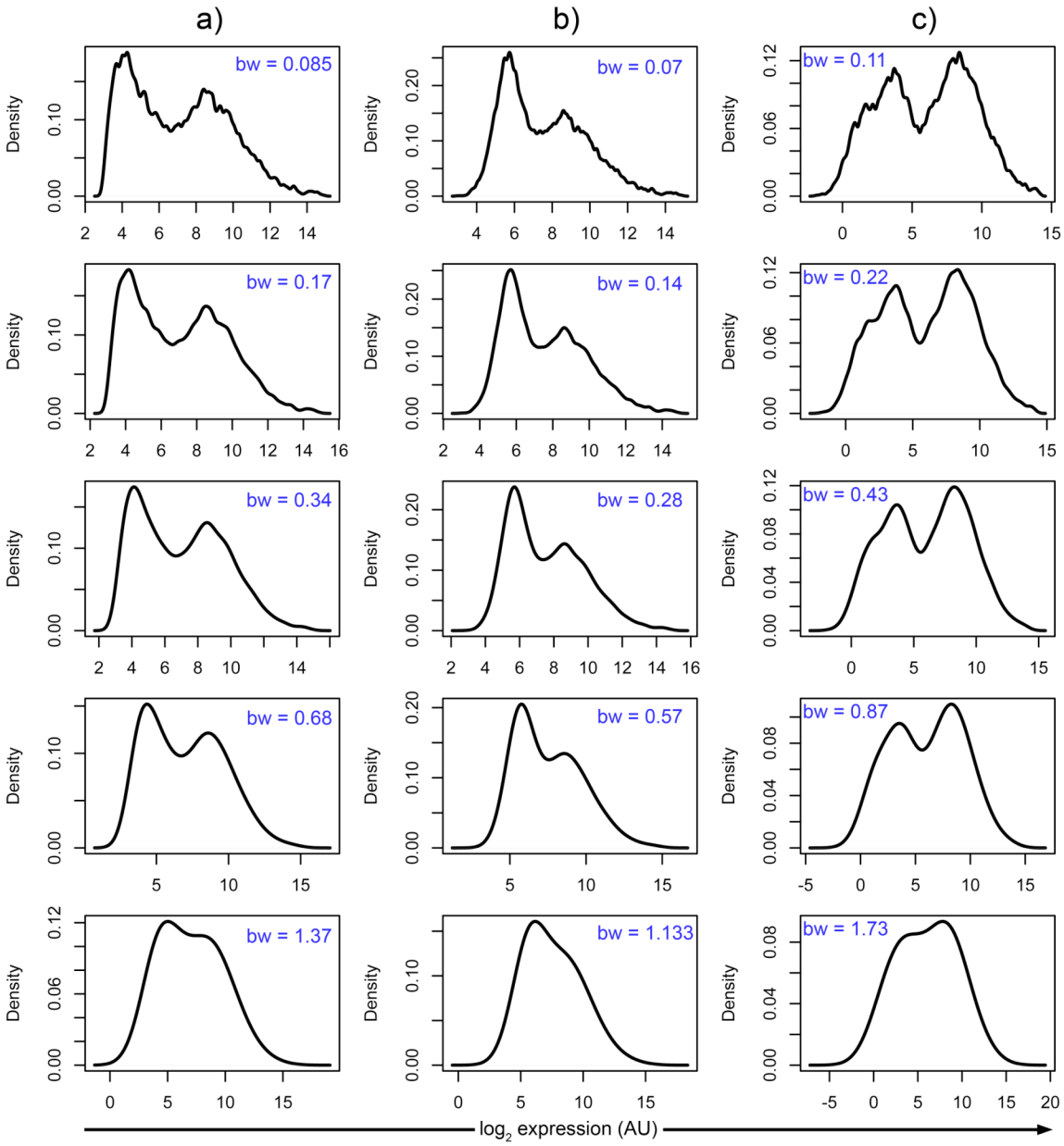


Figure S6. Examples of how different visualization methods affect the appearance of the microarray data ((Wei et al, 2009). The left panel corresponds to kernel density estimates (KDE). To demonstrate that the structure of the data is conserved under different settings,

the bandwidth (corresponding to the standard deviation of the Gaussian kernel) was increased in 2-fold steps from top to bottom (blue, left side). The bandwidth in the center corresponds to Silverman's 'rule of thumb'. The right panel shows histograms with different bin-sizes (indicated in blue on the right side). Bimodality is conserved if the bin-size is less than the distance between the two peaks.



	a)	b)	c)
Background correction	RMA	MAS	MAS
Normalization	Quantile	Quantile	QSpline
PM correction	PM only	PM only	MAS
Summarization	Median polish	avgdiff	Median polish

Figure S7. Examples for three further processing schemes in addition to MAS5 used in the main text. The raw data of (Wei et al, 2009) were processed by schemes a), b), and c) as indicated in the table and on top of the figure. PM, perfect match, RMA, robust multi-

chip average, MAS, microarray suite (Affymetrix). See the R Vignette of the 'affy' library for explanations of the individual methods and algorithms. Kernel density estimates (KDE) of the gene expression level distributions are shown. To demonstrate that the structure of the data is conserved under different KDE settings, the bandwidth (corresponding to the standard deviation of the Gaussian kernel) was increased in 2-fold steps from top to bottom (the bandwidth is given as 'bw =' in blue). The bandwidth in the center corresponds to Silverman's 'rule of thumb'.

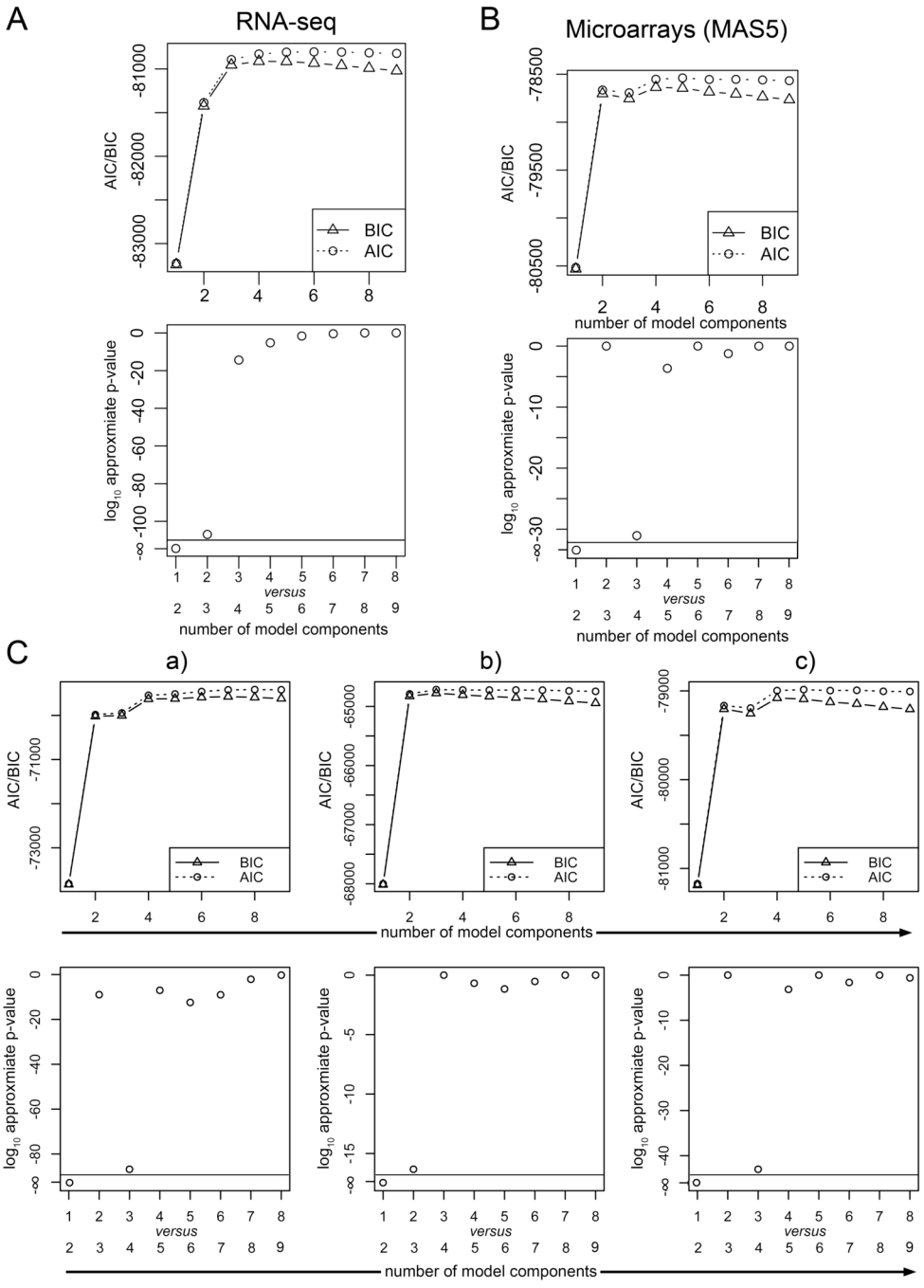


Figure S8. Goodness-of-fit tests for mixture models of one- to nine lognormal components fit to our RNA-seq data (A) and the microarray data of (Wei et al, 2009) (B, C) by expectation maximization. Tests for data normalized by MAS5 (B), as used in the main text, and by the three alternative normalization methods (C) as demonstated in Figure S7 (a), b) and c)) are shown as indicated. The tests used were the Akaike

Information criterion (AIC), the Bayesian information criterion (BIC), and a likelihood ratio test. For the latter, we compared each model to the next more complex one in terms of components. We numerically calculated the $\log_{10}$ p-values based on a χ^2 distribution. In the case that the numerical p-value was zero, we included it on the log scale as $-\infty$.

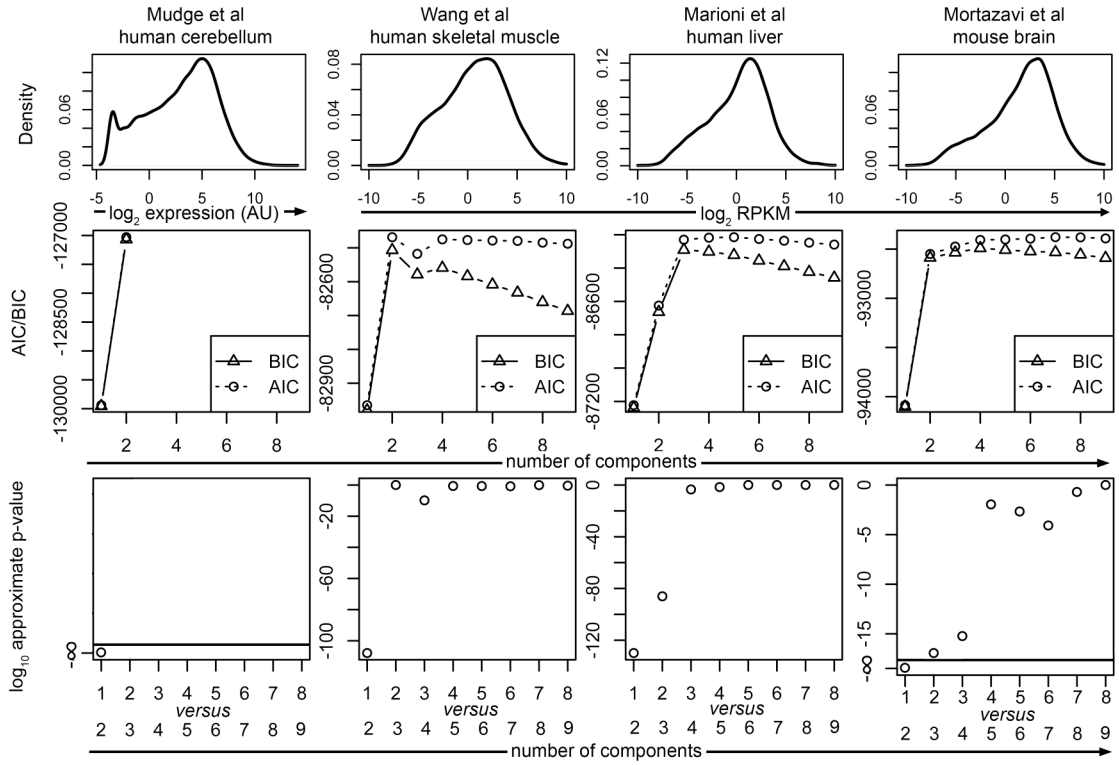


Figure S9. Kernel density estimates (KDE) and goodness-of-fit test for four additional RNA-seq datasets (Marioni et al, 2008; Mortazavi et al, 2008; Mudge et al, 2008; Wang et al, 2008). The KDE are shown on top using a Gaussian kernel and a bandwidth corresponding to Silverman's 'rule of thumb'. All distributions exhibit a shoulder on the left side. The goodness-of-fit tests used were the Akaike Information criterion (AIC), the Bayesian information criterion (BIC), and a likelihood ratio test. For the latter, we compared each model to the next more complex one in terms of components. We numerically calculated the $\log_{10}$ p-values based on a χ^2 distribution. In the case that the numerical p-value was zero, we included it on the log scale as $-\infty$.

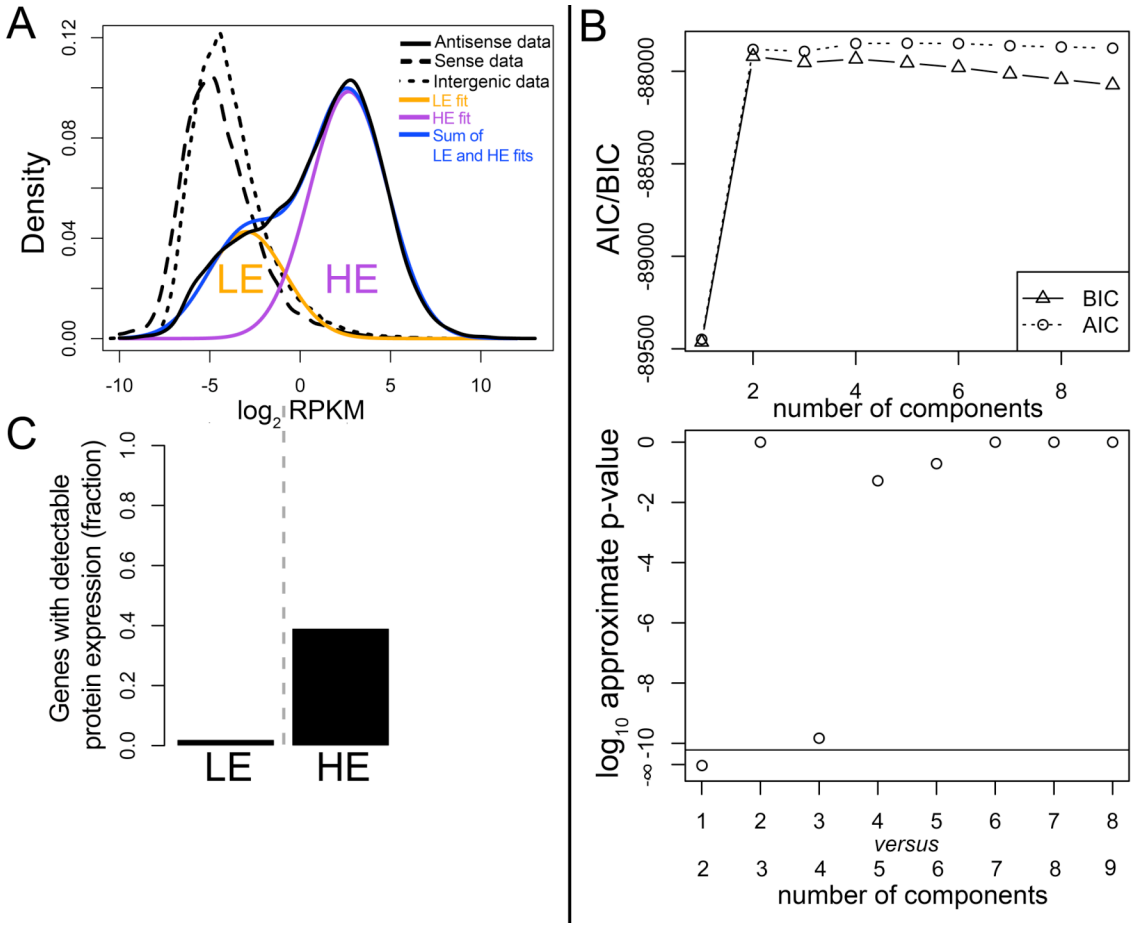


Figure S10. LE and HE groups in RNA-seq data of murine embryonic stem cells from (Cloonan et al, 2008). (A) The kernel density estimates (KDE) of expression levels are shown separately for genes in sense or antisense with reads mapping to them, since the data was prepared in a strand-specific manner (reads antisense to genes are selected by the experimental protocol), and for intergenic regions as indicated. The KDE use a Gaussian kernel and a bandwidth corresponding to Silverman's 'rule of thumb' (see Materials and Methods). Curve fitting was carried out as described for Figure 1C. (B) Plots of AIC, BIC and p-values of likelihood ratio tests as goodness-of-fits indicator for one- to nine-component normal distribution mixture models as described in Figure S8 and S9. (C) Genes were separated into LE and HE sets based on the expectation-maximization based curve fittings. SILAC protein expression data of murine embryonic stem cells (Graumann et al, 2008) was used to determine the fraction of genes that are expressed as proteins for the LE and HE sets separately.

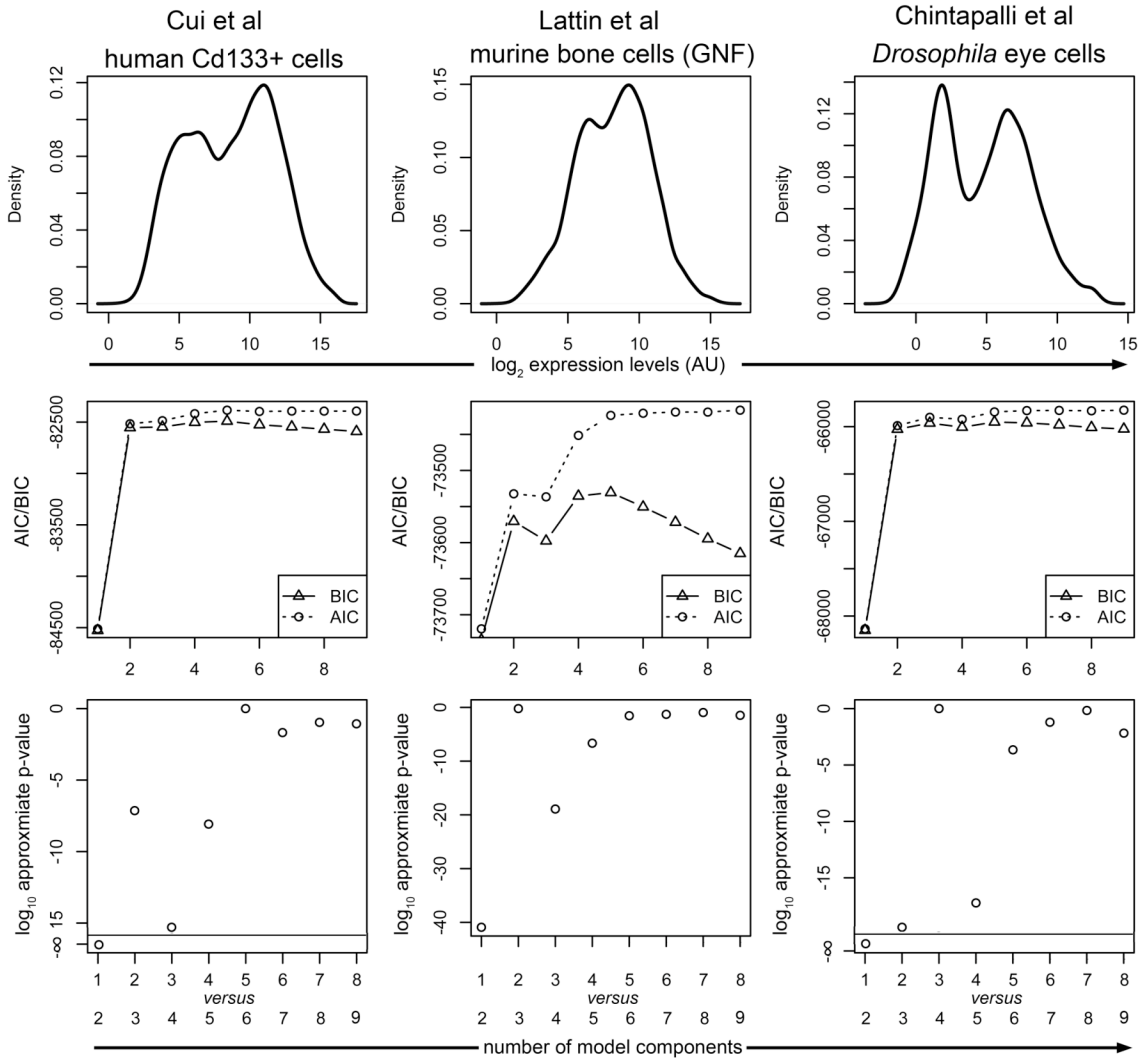


Figure S11. Kernel density estimates (KDE) and goodness-of-fit test for three additional microarray datasets (Chintapalli et al, 2007; Cui et al, 2009; Lattin et al, 2008). The KDE are shown on top using a Gaussian kernel and a bandwidth corresponding to Silverman's 'rule of thumb'. All distributions exhibit bimodality. The goodness-of-fit tests used were the Akaike Information criterion (AIC), the Bayesian information criterion (BIC), and a likelihood ratio test. For the latter, we compared each model to the next more complex one in terms of components. We numerically calculated the $\log_{10}$ p-values based on a χ^2 distribution. In the case that the numerical p-value was zero, we included it on the log scale as $-\infty$.

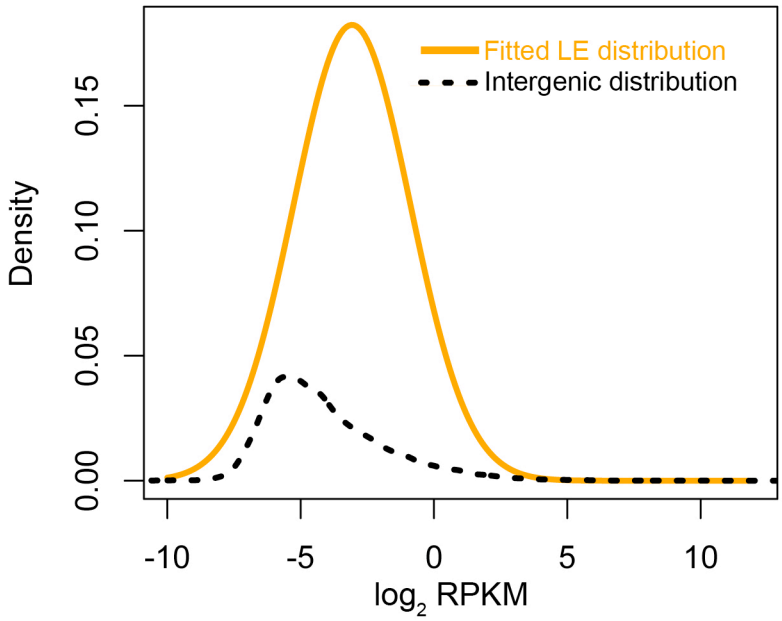


Figure S12. Distributions of RPKM for LE genes and intergenic regions. The fragments used to estimate intergenic RPKM were based on randomizations using the same length distribution as the exonic parts of genes. The area under the LE distribution is normalized to one (in contrast to Figure 1A where it is part of the total RPKM distribution within exons). The area under the intergenic distribution is less than one because of the fragments with zero reads (please see Figure S4).

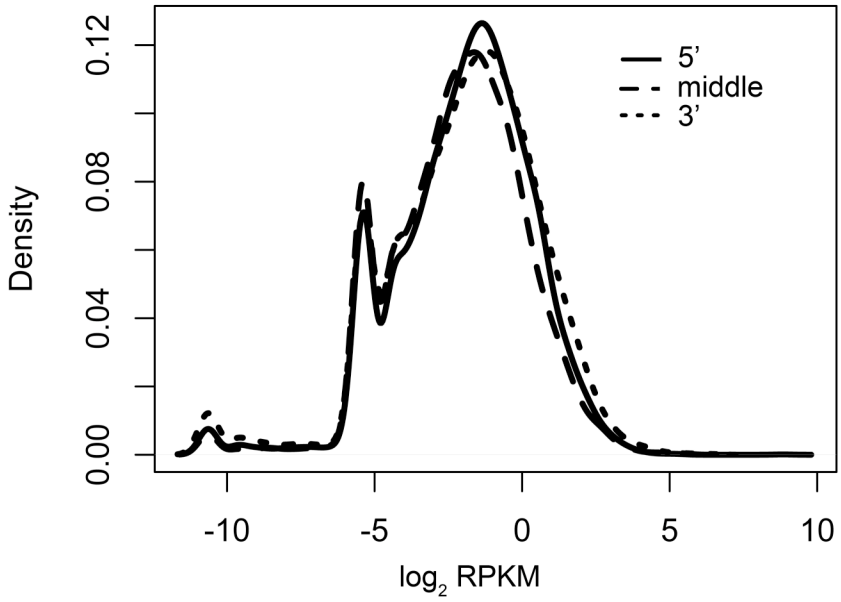


Figure S13. No RPKM bias in 5' or 3' ends of intronic regions. Introns of each gene were lined up. If the intronic region was at least 6 kb in total, RPKM were determined for the most 5' 2 kb, for the 2 kb in the center and for the most 3' 2 kb. The $\log_2$ RPKM distributions for all selected genes are shown and are almost identical.

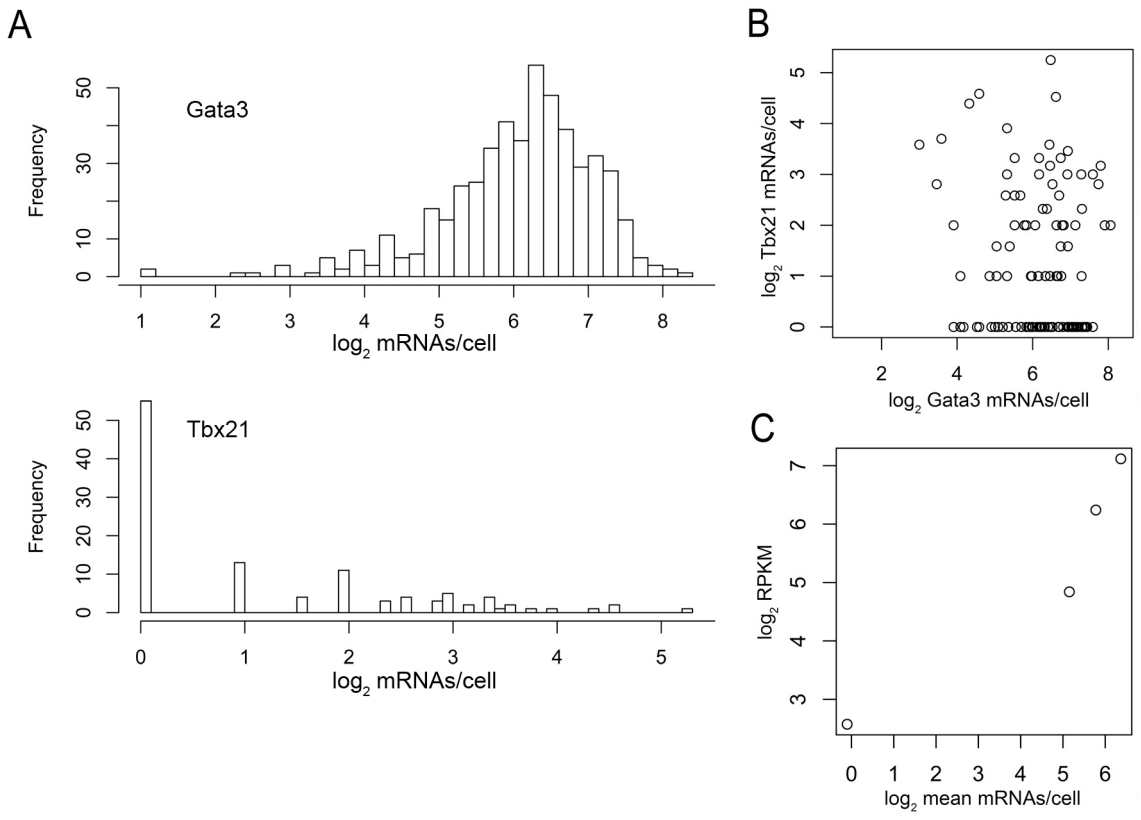


Figure S14. Log_2 transformed plots of Figure 3A, B and C.

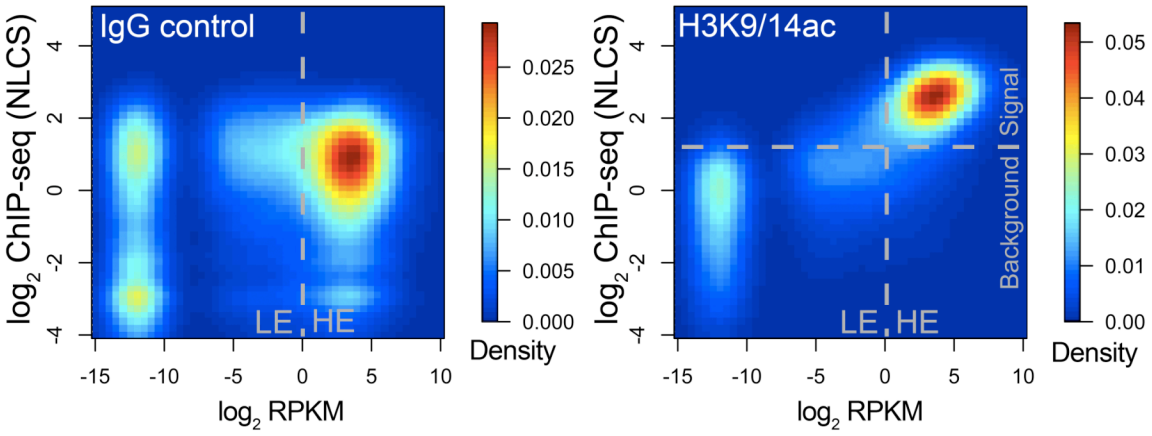


Figure S15. 2D kernel density estimates of RNA-seq gene expression level vs. ChIP-seq signal for each gene as in Figure 3D. To indicate the fractions of fragments/genes with zero RNA-seq or ChIP-seq reads, random RPKM value were assigned to them, drawn from normal distributions with mean = -12 or mean = -3, respectively, and standard-deviations = 1 (in both cases) on the $\log_2$ scale. These genes appear as additional blobs with respect to Figure 3D.

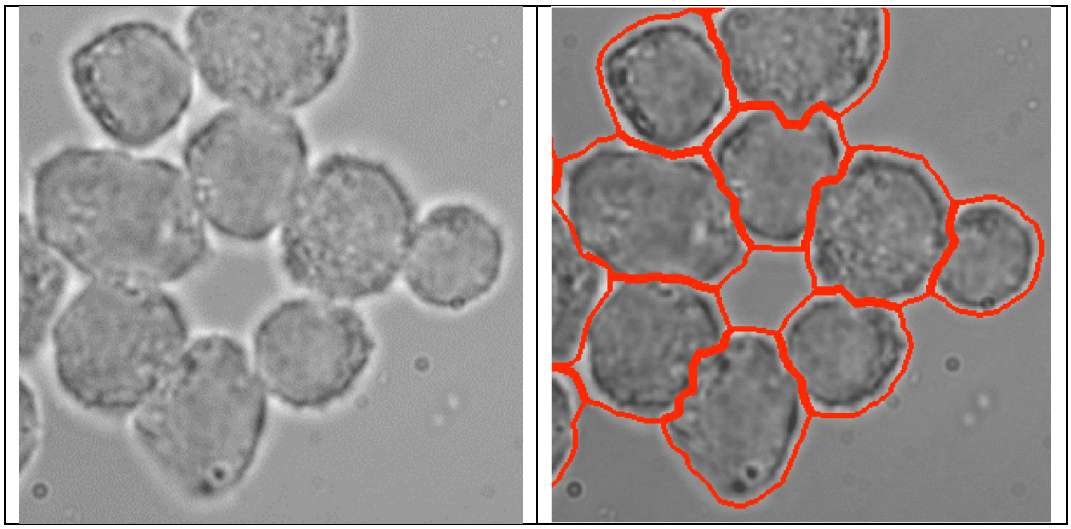


Figure S16. Segmentation of cells using bright-field images. The left panel is a bright-field image of the cells. The right panel is the segmented image.

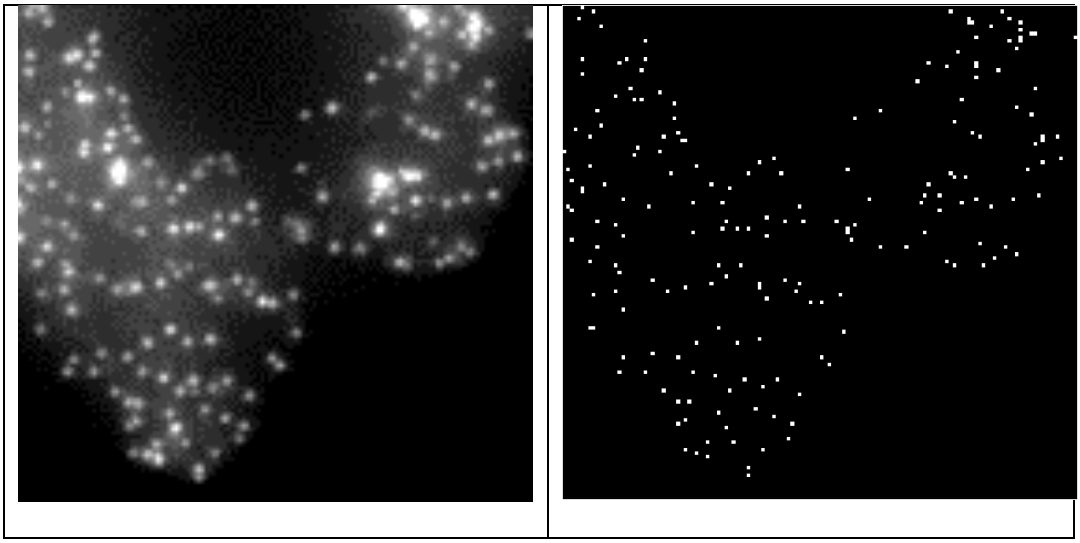


Figure S17. Analysis of mRNA spots. The left panel is a fluorescent maximum Z-projection image showing Gata3 transcripts in Th2 cells. The right panel is processed binary image showing each individual mRNA transcript as a single bright pixel.

Supplementary tables

Gene symbol	Expressed in Th2 cells (literature)?	Expressed in Th2 cells (our RNA-seq)?	Used in FACS stain?	Amplified in PCR?	Used in RNA-FISH?
Arbp	Yes (house keeping gene used as PCR control, e.g. (Hebenstreit et al, 2008))	Yes		Yes	
Cd4	Yes (Zhu et al, 2010)	Yes		Yes	Yes
Gata3	Yes (Zhu et al, 2010)	Yes	Yes	Yes	Yes
Il13	Yes (Zhu et al, 2010)	Yes	Yes	Yes	
Il4	Yes (Zhu et al, 2010)	Yes		Yes	
Il7r	Yes (Gregory et al, 2007)	Yes		Yes	Yes
Tbx21	No (Zhu et al, 2010)	Yes	Yes	Yes	Yes
Ifng	No (Zhu et al, 2010)	Yes (LE)	Yes	Yes	
Il17a	No (Zhu et al, 2010)	Yes (LE)		Yes	
Il2	No (Malek, 2008)	No		Yes	Yes
Rorc	No (Zhu et al, 2010)	Yes (LE)		Yes	
Pgf		Yes (LE)		Yes	
Ptprg		Yes (LE)		Yes	
Wdfy3		Yes (LE)		Yes	
Ripply3		Yes (LE)		Yes	
Glp1r		Yes (LE)		Yes	

Table S1. Genes examined in this study.

Sample	Read length	Total reads	Unique reads mapped to genome	Reads mapped to exons	Reads mapped to splice junctions
Replicate 1	41 bp	16,445,455	11,366,694	9,040,864	1,168,912
Replicate 2	36 bp	26,408,070	8,913,202	6,420,356	670,093

Table S2. RNA-seq sequencing read statistics.

Gene symbol	Median	Mean	Stdev	Fano factor
Cd4	39	54.86	67.83	83.88
Gata3	75	82.56	48.41	28.39
Il2	0	0.68	1.64	4.00
Il7r	24	35.55	36.89	38.29
Tbx21	0	0.93	3.15	10.64

Table S3. Single Molecule RNA-FISH statistics of five genes.

Gene symbol	fwd	rev	Exon spanning?	Junctions binding?
Arbp	AATCTCCAGAGGCACCATTG	ACCCTCCAGAAAGCGAGAGT	Yes	No
Cd4	AAGGGGCATGGGAGAAAGGAT	AAGGTCACTTTGAACACCCAC	Yes	Yes
Gata3	CCCTCCGGCTTCATCCTCT	CTGCACCTGATACTTGAGGC	No	
Il13	CCTGGCTCTTGCTTGCCTT	GGTCTTGTGTGATGTTGCTCA	No	
Il17a	CTCCAGAAGGCCCTCAGACTAC	AGCTTTCCCTCCGCATTGACACAG	Yes	No
Il2	TGAGCAGGATGGAGAATTACAGG	TGTTGTCAGAGCCCTTTAGTTTT	Yes	Yes
Il7r	TATGTGGGGCTCTTTTACGAGT	GCCTCGGCTTTAACTATTGTGT	Yes	Yes
Ifng	ATGAACGCTACACACTGCATC	CCATCCTTTTGCCAGTTCCTC	Yes	No
Pgf	TCTGCTGGGAACAACTCAACA	GTGAGACACCTCATCAGGGTAT	Yes	Yes
Ptprg	AGTCAGTCCGAGGGACAATTC	GGTGGCGTAGTCAAGGAGC	Yes	Yes

Rorc	CCGCTGAGAGGGCTTCAC	TGCAGGAGTAGGCCACATTACA	Yes	Yes
Tbx21	TTTCCAAGAGACCCAGTTCATTG	ATGCGTACATGGACTCAAAGTT	Yes	Yes
Wdfy3	CCACCATCGGGTTCATTAACA	GTGGGACAGAGATGCCTATGT	Yes	No
Ripply3	GGCCCGAAAGTTCCATTCCA	CTCCCGATGTGTGTTGGTCT	Yes	Yes
Glp1r	ACGGTGTCCCTCTCAGAGAC	ATCAAAGGTCCGGTTGCAGAA	Yes	No

Table S6. Primer sequences.

Supplementary references

Chintapalli VR, Wang J, Dow JA (2007) Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nat Genet* **39:** 715-720

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5:** 613-619

Cui K, Zang C, Roh TY, Schones DE, Childs RW, Peng W, Zhao K (2009) Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **4:** 80-93

Graumann J, Hubner NC, Kim JB, Ko K, Moser M, Kumar C, Cox J, Scholer H, Mann M (2008) Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol Cell Proteomics* **7:** 672-683

Gregory SG, Schmidt S, Seth P, Oksenberg JR, Hart J, Prokop A, Caillier SJ, Ban M, Goris A, Barcellos LF, Lincoln R, McCauley JL, Sawcer SJ, Compston DA, Dubois B, Hauser SL, Garcia-Blanco MA, Pericak-Vance MA, Haines JL (2007) Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. *Nat Genet* **39:** 1083-1091

Hebenstreit D, Giaisi M, Treiber MK, Zhang XB, Mi HF, Horejs-Hoeck J, Andersen KG, Krammer PH, Duschl A, Li-Weber M (2008) LEF-1 negatively controls interleukin-4 expression through a proximal promoter regulatory element. *J Biol Chem* **283:** 22490-22497

Lattin JE, Schroder K, Su AI, Walker JR, Zhang J, Wiltshire T, Saijo K, Glass CK, Hume DA, Kellie S, Sweet MJ (2008) Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res* **4:** 5

Malek TR (2008) The biology of interleukin-2. *Annu Rev Immunol* **26:** 453-479

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18:** 1509-1517

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621-628

Mudge J, Miller NA, Khrebtukova I, Lindquist IE, May GD, Huntley JJ, Luo S, Zhang L, van Velkinburgh JC, Farmer AD, Lewis S, Beavis WD, Schilkey FD, Virk SM, Black

CF, Myers MK, Mader LC, Langley RJ, Utsey JP, Kim RW et al (2008) Genomic convergence analysis of schizophrenia: mRNA sequencing reveals altered synaptic vesicular transport in post-mortem cerebellum. *PLoS ONE* **3:** e3625

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470-476

Wei G, Wei L, Zhu J, Zang C, Hu-Li J, Yao Z, Cui K, Kanno Y, Roh TY, Watford WT, Schones DE, Peng W, Sun HW, Paul WE, O'Shea JJ, Zhao K (2009) Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity* **30:** 155-167

Zhu J, Yamane H, Paul WE (2010) Differentiation of effector CD4 T cell populations (*). *Annu Rev Immunol* **28:** 445-489