RNA sequencing reveals two major classes of gene expression levels in metazoan cells

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.

Running title: Two classes of gene expression levels

Character count: 27,218

Subject categories: Bioinformatics, Chromatin & Transcription

Correspondence to:

D. H.
danielh@mrc-lmb.cam.ac.uk
Phone: +44 (0) 1223 402479
Fax: +44 (0) 1223 213556

S. A. T.
sat@mrc-lmb.cam.ac.uk
Phone: +44 (0) 1223 252947
Fax: +44 (0) 1223 213556

Abstract

The expression level of a gene is often used as a proxy for determining whether the protein or RNA product is functional in a cell or tissue. Therefore, it is of fundamental importance to understand the global distribution of gene expression levels, and to be able to interpret it mechanistically and functionally. Here we use RNA sequencing of mouse Th2 cells, coupled with a range of other techniques, to show that all genes can be separated, based on their expression abundance, into two distinct groups: one group comprising of lowly expressed and putatively non-functional mRNAs, and the other of highly expressed mRNAs with active chromatin marks at their promoters.

Key words: expression levels/RNA-seq/ChIP-seq/RNA-FISH/bimodal

Introduction

Expression level is frequently used as a way of characterizing gene function, by Northern blotting, PCR, microarrays, and, more recently, RNA-sequencing (Wang et al, 2009a) (RNA-seq). Therefore, it is a central issue in molecular biology to know how many transcripts are expressed in a cell at what levels. This question was studied very early in the history of molecular biology using methods such as reassociation kinetics (Hastie & Bishop, 1976), which indicated the existence of distinct abundance classes, and recently, we pointed out that separate peaks are visible in the abundance distributions of a number of microarray data sets (Hebenstreit & Teichmann, 2011). At the same time, microarrays or RNA-seq data have been described as displaying broad, roughly lognormal distributions of expression levels with no clear separation into discrete classes (Hoyle et al, 2002; Lu & King, 2009; Ramskold et al, 2009). There are several reasons for this: many samples are heterogeneous in terms of cell type (Hebenstreit & Teichmann, 2011) or are based on a previous generation of less sensitive microarrays, many are from unicellular organisms rather than animals, and finally, data processing and plotting methods can obscure the presence of distinct abundance classes. Here, we provide experimental and computational support for two overlapping major mRNA abundance classes.

Results and Discussion

We based our analysis on murine Th2 cells (Zhu et al, 2010) as these cells can be obtained in large quantities *ex vivo* and can be prepared as a pure and homogeneous cell population. Furthermore, there is a well characterized set of genes whose proteins are known to be expressed and functional in Th2 cells, as well as a set of genes known to be not expressed in these cells (Table S1 lists the genes we used in our study, Figure S1 shows expression of two marker proteins in the cells).

We generated Th2 RNA-seq data for two biological replicates (Table S2 gives the number of reads and mappings we obtained) and calculated gene expression levels using the standard measure of Reads Per Kilobase per Million (RPKM) (Mortazavi et al, 2008). The expression levels of the biological replicates are highly correlated (r^2 = 0.94, Figure S2). We then calculated the mean RPKMs of the two samples for all genes and $\log_2$ transformed these values.

Displaying the distribution of all gene expression levels as a kernel density estimate (KDE) reveals an interesting structure: the majority of genes follow a normal distribution which is centered at a value of ~4 $\log_2$ RPKM (~16 RPKM), while the remaining genes form a shoulder to the left of this main distribution (Figure 1A, solid line). This was conserved under different KDE bandwidths (Figure S3, left panel) or different histogram representations (Figure S3, right panel). As genes with zero reads cannot be included on the log scale, we prepared an alternative version of Figure 1A where we assigned low RPKM values to these. This helps to illustrate the fraction of zero read genes (Figure S4). As a comparison, we studied microarray data for the same cell type from a recent publication (Wei et al, 2009). The correlation between the microarray and the RNA-seq data was very good and highly statistically significant (Pearson $r^2 = 0.83$, Spearman $\rho = 0.84$, Figure S5). Surprisingly, displaying the distribution of microarray expression levels results in a clearly bimodal distribution (Figure 1B). Again, the appearance of the distribution was insensitive to the KDE bandwidth choice or histogram bin size (Figure S6). The bimodality was conserved when alternative normalization and processing schemes were used, independent of KDE bandwidths (Figure S7).

Visual inspection of both microarray and RNA-seq data thus reveals two overlapping main components of the distribution of gene expression levels. Quantifying this by curve fitting confirms a good fit to two distributions: the goodness-of-fit (measured by Akaike Information criterion, AIC (Akaike, 1974), Bayesian Information Criterion, BIC (Schwarz, 1978) or Likelihood ratio tests (Casella & Berger, 2001)) shows strong increases for both microarray and RNA-seq data when two-component models are fit by expectation-maximization (compared to single- or more-component models) (Figure S8). We designate the two groups of genes as the lowly expressed (LE) and highly expressed (HE) genes (Figure 1C), because we will present evidence below that the LE genes are expressed rather than simply being experimental background. Our findings are not limited to Th2 cells and hold for virtually all recently published metazoan RNA-seq datasets (e.g. (Marioni et al, 2008; Mortazavi et al, 2008; Mudge et al, 2008; Wang et al, 2008), Figure S9 and (Cloonan et al, 2008), Figure S10A, B) and all microarray data sets (e.g. (Cui et al, 2009), GNF Atlas 3 (Lattin et al, 2008), (Chintapalli et al, 2007), Figure S11) we have studied. The existence of further, minor groups of genes cannot be excluded, but is not clear at this point due to the diverse curve-fitting results for the different datasets if higher-order (more than two components) models are considered.

The difference between the microarray and RNA-seq distributions is explained by the fact that the microarrays yield a signal for all genes, part of which is due to cross-hybridization of oligo-nucleotide probes if the gene is not strongly expressed. RNA-seq on the other hand yields a signal for a gene only if at least one sequencing read is found. The accuracy of RNA-seq is biased towards longer and more highly expressed genes, e.g. 5 % of all genes account for 50 % of all reads in our data as well as in other datasets (Bullard et al, 2010; Mortazavi et al, 2008; Oshlack & Wakefield, 2009).

To explore how this accuracy bias affects the shape of the LE distribution, we studied the RNA-seq detection limit. We first plotted the number of genes with zero reads as a function of the total number of reads (taking subsets of the total reads). The number of genes without reads decreases slowly, with no change in slope and hence no indication of reaching a plateau. Even at a total of 25 million reads, ~30% of all genes are undetected (Figure 2A). We further estimated the numbers of genes remaining undetected at each expression level by assuming Poisson-distributed read numbers (Jiang & Wong, 2009) and determining the expected frequency of zeroes. This confirms the sensitivity drop at the lower end of the LE peak (Figure 2B). Extrapolating the numbers of expressed genes including the undetected ones reveals an emerging LE peak (Figure 2B). Thus the smaller portion of LE genes in the RNA-seq data compared to the microarray data is at least partially due to the RNA-seq detection limit, although this only becomes a problem for genes at less than ~ -3 to -4 $\log_2$ RPKM. It should be noted that these low expression levels correspond to an absence of transcripts in the majority of cells, as we demonstrate further below.

To further confirm that the LE genes correspond to low expression and not experimental noise, we performed realtime PCRs. We tested amplification by exon spanning primers of a set of genes that are known to be expressed or not expressed in Th2 cells, plus five random genes that we detected between -3.7 and -5 $\log_2$ RPKM in the RNA-seq experiment (Table S1). We were able to successfully PCR-amplify all genes with high specificity. The expressed genes map to the HE peaks, while the unexpressed genes map to the LE peaks, if we align the PCR results with the microarray/RNA-seq data (Figure 2C).

We also tested the extent to which genomic DNA can be detected in our polyA-purified mRNA sample, as proposed by Ramskold et al (Ramskold et al, 2009) as a means of quantifying experimental background. We randomly selected intergenic fragments with the same length distribution as genes, 10 kb away from genes. The resulting RPKM distribution contains a high number of zero-RPKM fragments (79 %) while the majority of non-zero fragments peaks slightly left of the LE shoulder (Figure 1A). The 90 % quantile of this intergenic background distribution is found at -4.97 $\log_2$ RPKM, which means that we can be quite confident that genes with an RPKM value above this level are truly expressed rather than representing experimental background noise (Figure 1A). We cannot rule out that detection of intergenic DNA corresponds to transcription as well, which would make the case for transcription of LE genes even stronger.

Analysis of the strand-specific RNA-sequencing data of Cloonan et al (Cloonan et al, 2008) yields similar conclusions. The experimental protocol selects for reads antisense to genes. In the distribution of 'sense' reads (i.e. the noise, reads that don't map to genes), more than 50 % of genic regions have zero reads. This noise distribution is unimodal and shifted by ~ 2 $\log_2$ RPKM with respect to the LE distribution (Figure S10A).

We next determined the distribution of RPKM within introns, again using fragments with the same length distribution as transcripts. (Please note that our intronic read densities are not enriched at 5' or 3' ends of the intronic regions (Figure S12).) The resulting intronic distribution is significantly higher than the intergenic background (two-sided Wilcoxon rank sum test, $p < 2.2 \times 10^{-16}$) and peaks at roughly -1 $\log_2$ RPKM (Figure 1A). Introns thus have one- to two orders of magnitude lower read density than exons. This suggests that we are detecting incompletely processed transcripts at a low but significant and uniform level across all the whole range of transcript abundances.

Since introns are one- to two orders of magnitude longer than exons, introns should be detectable with roughly the same accuracy as exons, if the full-length set of introns of a gene is used. If we plot the RPKM in exonic regions *versus* the RPKM in intronic regions for each gene, there is significant correlation ($r^2 = 0.86$, $p < 2.2 \times 10^{-16}$) across the whole spectrum of expression levels. Calculating the correlation for lowly and highly expressed genes separately yields only slightly lower correlations among LE genes compared to HE genes, and both correlations are highly significant

(Figure 2D). This provides evidence that confirms that LE genes are transcribed rather than experimental background: there would not be such a high correlation between introns and exons, particularly in the low abundance region, if their detection were due to noise.

We next studied gene expression using a single cell approach by performing single molecule RNA-FISH (Raj et al, 2008) for five genes that are expressed at different levels according to the literature and our RNA-seq data. The distributions of mRNA numbers per cell were very broad for expressed genes (e.g. Gata3), while low mRNA numbers from ‘not-expressed’ genes (e.g. Tbx21) were still detected (Figure 3A). All genes had Fano factors (σ^2/μ) larger than 4, indicating that they had extra-Poisson variation (a Poisson random variable would have $\sigma^2/\mu = 1$) and therefore burst-like transcription (Raj & van Oudenaarden, 2009) (Table S3). Importantly, cells expressing Tbx21 were not anti-correlated with cells expressing Gata3 (Figure 3B), meaning that we do not have a sub-population of Th1 cells in our Th2 cell populations. This further demonstrates that LE expression is not due to a contaminating cell type, as the same cells express groups of genes at HE and others at LE levels.

Since the RPKM as measured by RNA-seq should be proportional to the mean mRNA numbers per cell, we can use the RNA-FISH results to estimate how our RPKM values translate into mRNA numbers. We find that one RPKM corresponds to an average of roughly one transcript per cell in our Th2 data set (Figure 3C). Please note that the value of one RPKM/one transcript on average per cell serves as an estimate only as it is based on a limited number of data points.

It should be noted that the two groups of genes at high versus low expression levels cannot result from a mixture of different cell types. Mixing of different cell types leads to gene expression levels for each gene that are an average across cell types. Hence such distributions will become more unimodal, not less so (following the central limit theorem).

To study the nature of the LE and HE groups in more detail, we prepared Th2 ChIP-seq data for the activating H3K9/14 acetylation histone modification (Roh et al, 2005; Wang et al, 2009b) (H3K9/14ac) and one IgG control. We calculated the histone modification level at each gene by identifying a globally enriched window around the transcription start sites of genes, and using reads in this window as a measure of each gene's modification level, normalized by the total reads (giving the normalized locus specific chromatin state, NLCS, as used in (Hebenstreit et al, 2011)). Thus we were able to plot histone modification levels of each gene against expression levels from the RNA-seq or microarray data using a heatmap representation (Figure 3D, RNA-seq, Figure 3E, microarrays). Figure S13 is an alternative version of this figure, where we randomly assigned low RPKM values to the zero-read genes.

This strikingly confirms the two groups of gene expression levels, as there is a very good agreement between LE genes and absence of histone marks on one hand, and HE genes and presence of H3K9/14ac marks on the other hand (Figure 3D-E). This is seen for both the microarrays as well as the RNA-seq data. This extends previous findings of the relationship between H3K9/14ac and transcriptional activation by revealing an on/off-type of correlation between this histone mark and the LE/HE groups of genes. It should be noted that there is a very weak correlation within the LE and HE groups. The strongest correlation is within the RNA-seq HE group with a correlation coefficient $r^2 = 0.29$ in log space and $r^2 = 0.097$ on linear space.

Since the LE group of genes is still expressed at low levels and contains at least five genes that are characterized as not expressed and non-functional in Th2 cells, it seems likely that the HE group of genes represents the active and functional transcriptome of cells. This is supported by SILAC proteomics data (Graumann et al, 2008) which is available for the embryonic stem cell data we presented earlier (Figure S10) and which indicates protein expression of HE genes only (Figure S10C). The tight correlation recently observed between RNA and protein levels in three mammalian cell lines also supports this (Lundberg et al, 2010).

Gene ontology (GO) analysis of LE and HE genes in the Th2 cells supports the notion that HE comprises the functional transcriptome, as many T cell specific processes (e.g. GO:0050863, GO:0045582, GO:0042110) and housekeeping processes are enriched (Table S4). On the other hand, many GO terms referring to differentiation of other celltypes (e.g. ear development GO:0043583, neuron fate commitment GO:0048663) are enriched among the LE set of genes (Table S5).

In conclusion, our data shows that two large groups of genes can be discriminated based on the distribution of expression levels. RNA-FISH indicates that the boundary between the groups is found at an expression level of roughly one transcript per cell. In addition, H3K9/14ac marks are associated with the promoters of highly expressed genes only (Figure 3F). It thus seems likely that the LE/HE groups reflect different transcription kinetics depending on the chromatin state or *vice versa*. The LE group is likely to correspond to 'leaky' expression, producing non-functional transcripts. The majority of LE genes are expressed at less than one copy per cell on average, and it would be interesting to know whether such stochastic expression has any function, e.g. in cell differentiation, or any deleterious effects. There may be a trade-off between the cost of repressing expression entirely and unwanted consequences of stochastic expression.

Regulation of gene expression is mostly mediated by transcription factor binding events at promoters and enhancers, e.g. (Heintzman et al, 2009). Often, differential regulation induces only small changes in expression levels, probably serving to fine-tune expression and shifting genes within the HE group. Our data suggests that in addition to this, there is a key decision about whether a gene becomes "switched on" and expressed which coincides with a boost in both transcription and H3K9/14ac histone modification.

References

Akaike H (1974) New Look at Statistical-Model Identification. *Ieee T Automat Contr* **Ac19:** 716-723

Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11:** 94

Casella G, Berger RL (2001) *Statistical Inference*, 2nd edn.: Duxbury Press.

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

Hastie ND, Bishop JO (1976) The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9:** 761-774

Hebenstreit D, Gu M, Haider S, Turner DJ, Lio P, Teichmann SA (2011) EpiChIP: gene-by-gene quantification of epigenetic modification levels. *Nucleic Acids Res*

Hebenstreit D, Teichmann S (2011) Analysis and simulation of gene expression profiles in pure and mixed cell populations. *Physical Biology*

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M et al (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*

Hoyle DC, Rattray M, Jupp R, Brass A (2002) Making sense of microarray data distributions. *Bioinformatics* **18:** 576-584

Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25:** 1026-1032

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

Lu C, King RD (2009) An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems. *Bioinformatics* **25:** 2020-2027

Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, Algenas C, Lundeberg J, Mann M, Uhlen M (2010) Defining the transcriptome and proteome in three functionally different human cell lines. *Mol Syst Biol* **6:** 450

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

Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* **4:** 14

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5:** 877-879

Raj A, van Oudenaarden A (2009) Single-molecule approaches to stochastic gene expression. *Annu Rev Biophys* **38:** 255-270

Ramskold D, Wang ET, Burge CB, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* **5:** e1000598

Roh TY, Cuddapah S, Zhao K (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* **19:** 542-552

Schwarz G (1978) Estimating Dimension of a Model. *Ann Stat* **6:** 461-464

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

Wang Z, Gerstein M, Snyder M (2009a) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10:** 57-63

Wang Z, Schones DE, Zhao K (2009b) Characterization of human epigenomes. *Curr Opin Genet Dev*

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

Acknowledgments

We would like to thank Guilhem Chalancon and Joseph Marsh for reading the manuscript and making valuable suggestions, Ines de Santiago and Ana Pombo for helpful and interesting discussions, Lucy Colwell for her role in establishing a fruitful collaboration, and Jonathon Howard for reminding us of the importance of absolute numbers.

Contributions

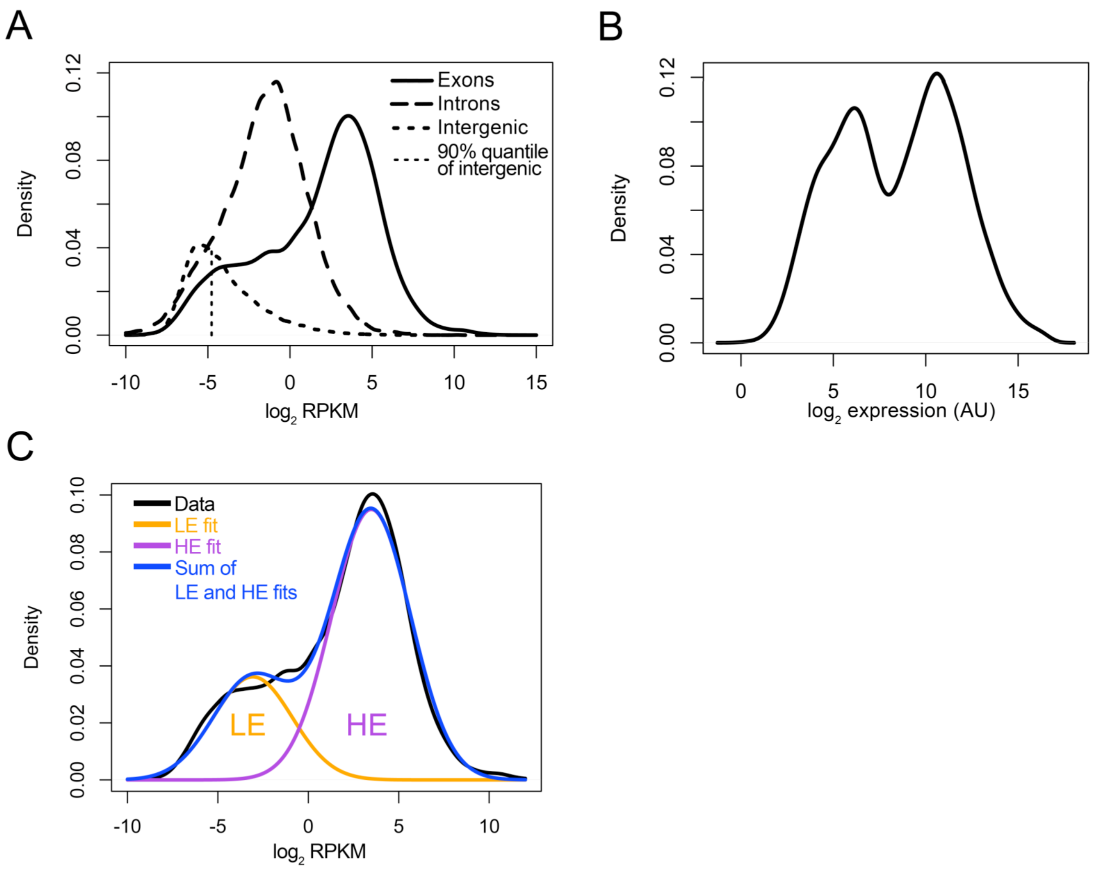
Experiments, with the exception of RNA-FISH, were carried out by DH. RNA-FISH staining and image processing were carried out by MF. Computational analyses were carried out by DH, with contributions from MG and VC. DH and SAT wrote the manuscript with contributions from MF and AVO.

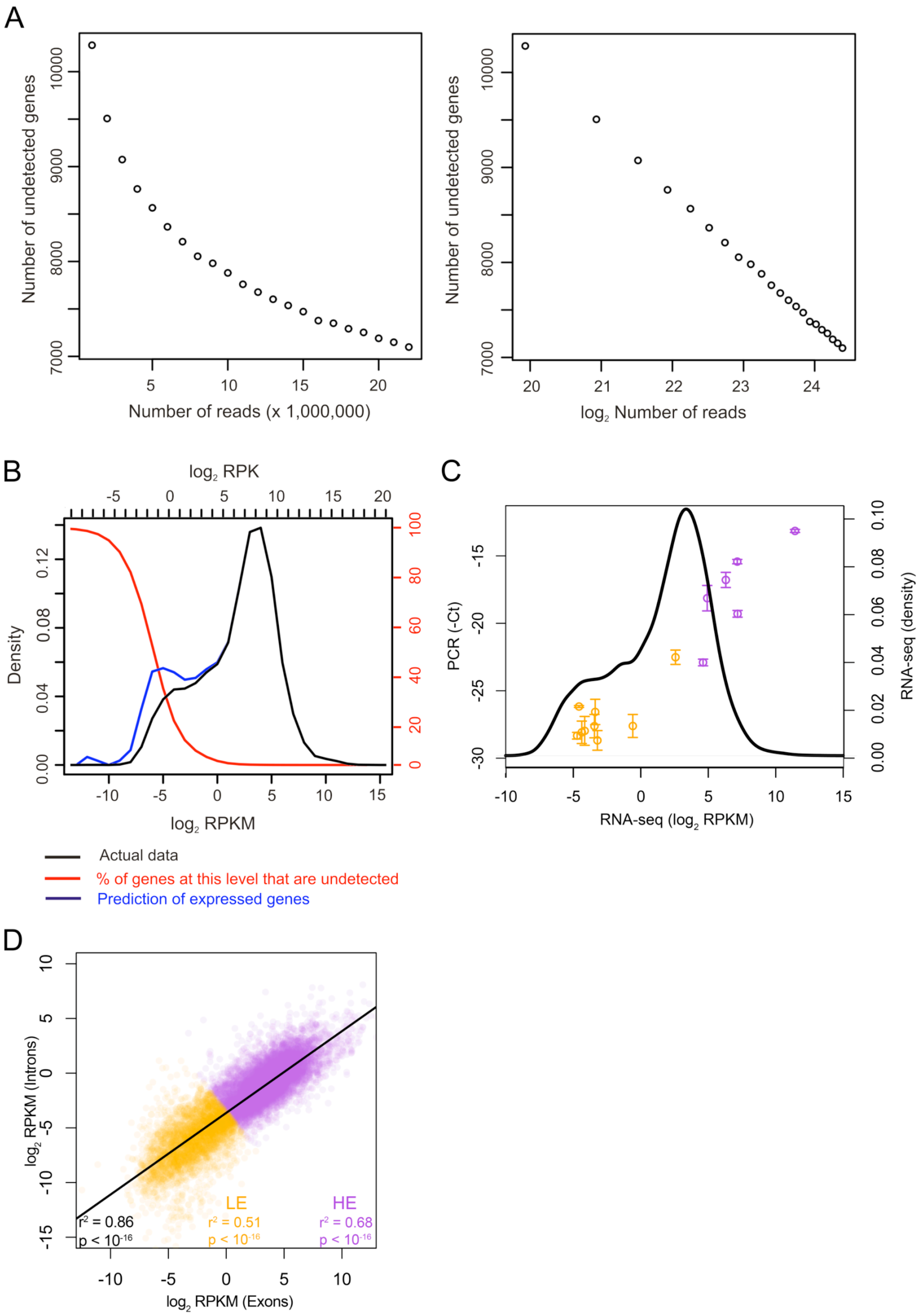
Figure legends

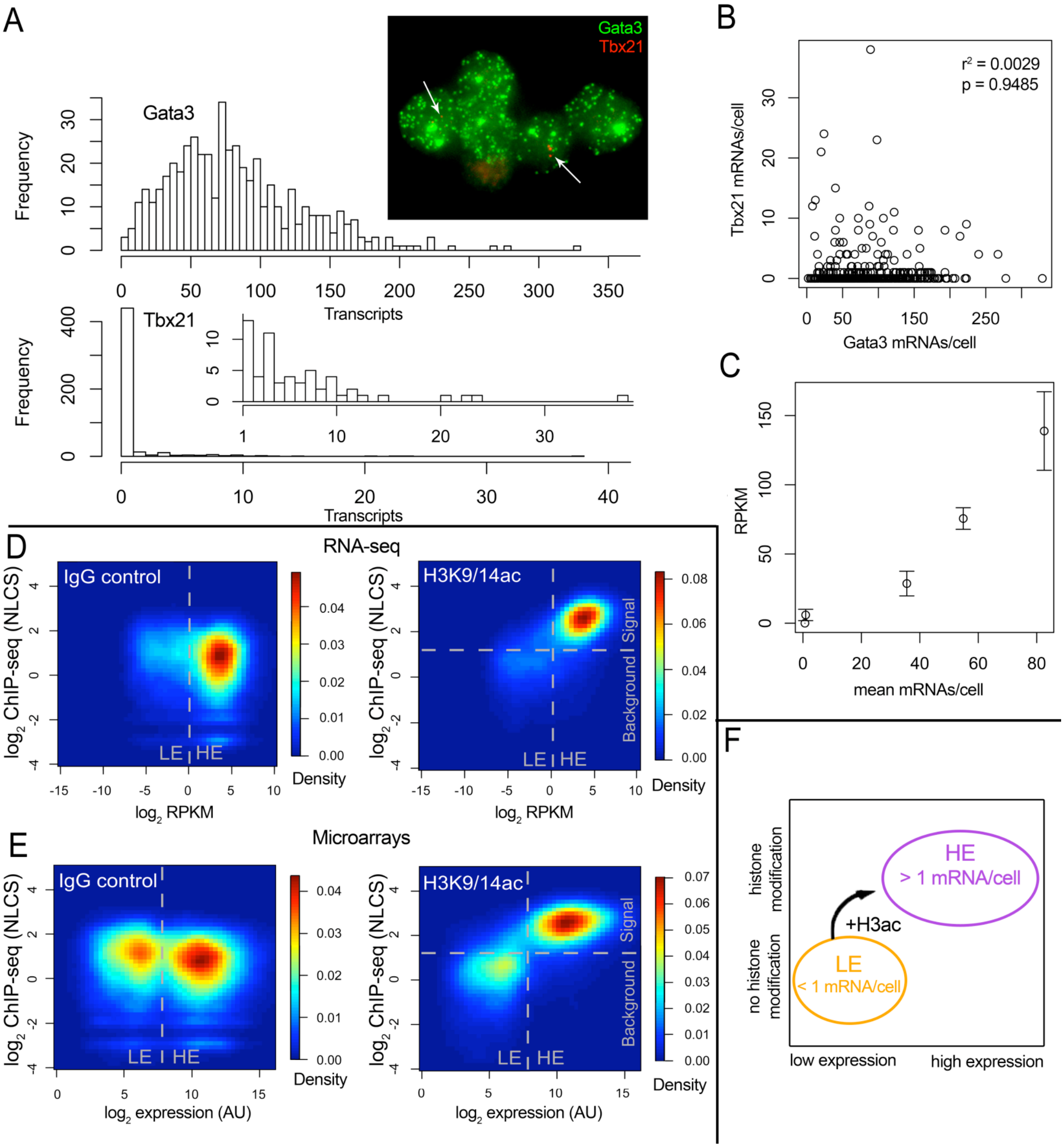
Figure 1. Distribution of gene expression levels. (A) Kernel density estimates of RPKM distributions of RNA-seq data within exons, introns and intergenic regions as indicated. The fragments used to estimate intron and intergenic RPKM were based on randomizations using the same length distribution as the exonic parts of genes. The 90% quantile of the intergenic distribution is indicated. (B) Kernel density estimate of expression level distribution of microarray data (Wei et al, 2009). (C) Expectation maximization based curve fitting of RNA-seq data of (A).

Figure 2. Sensitivity of RNA-seq. (A) Detection of genes in dependency of the total read numbers on linear scale (left) and $\log_2$ scale (right). Random subsets of the total reads for the two RNA-seq replicates were taken and the number of genes with zero reads were plotted *versus* the total read numbers used. The figure represents an average of five independent subsets for each data point. (B) Prediction of genes remaining undetected due to Poisson statistics underlying RNA-seq. The theoretically expected fraction of genes remaining undetected (red, y-axis on the right side of the figure in red) was determined for each expression level and was used to infer from the binned (small ticks on top indicate the bins) actual expression data (black) the expressed genes including the undetected ones (blue). In addition to the RPKM scale, the reads per kilobase (RPK) scale (without normalization to the total number of mapped reads) is shown (on top), which was used for the calculation of the (integer-) Poisson statistic and which, in contrast to the RPKM scale, depends on the total number of sequencing reads. (C) RT-PCR for the genes listed in Table S1. The RNA-seq expression levels of the genes are plotted *versus* the negative threshold cycles (C_t) of the PCRs. The plot is overlaid (with the same x-axis scaling) upon the kernel density estimate of the RNA-seq expression level distribution (black line) to show the positions of the genes in the total expression distribution. Genes either in the LE peak of the RNA-seq distribution or which have been previously characterized as not expressed in Th2 cells are shown in orange. Genes known to be expressed are shown in purple. Error bars indicate standard error of mean from three independent biological replicates. Please refer to Tables S1 and S3 for details of genes and PCR primers. (D) Correlation of RPKM within exons and introns from RNA-seq data of Figure 1A. Correlation and significance of correlation were calculated for the whole distribution (black) or for LE and HE genes separately (orange and purple, respectively). Division into LE and HE was performed along a line perpendicular to a fitted trendline (black), centered at Exon RPKM = 1.

Figure 3. (A) Distribution of mRNA numbers among single cells. Histograms for Gata3 and Tbx21 (with an inset histogram starting from 1 instead of 0 to better illustrate higher expressions) and a sample fluorescence microscopy image are shown. Tbx21 transcripts are marked with white arrows to ease identification. (B) Correlation between Gata3 and Tbx21 expression. Correlation coefficient and significance are inset. (C) Plot of mean mRNA numbers per cell *versus* RNA-seq RPKM of five genes. Error bars indicate SEM from two RNA-seq biological replicates. (D-E) 2D kernel density estimates of gene expression level vs. ChIP-seq signal for each gene for RNA-seq (D) and microarray (E) data. Divisions between background and signal for the ChIP-seq component were determined by curve fitting with the software EpiChIP (Hebenstreit et al, 2011) and are indicated. Divisions between LE and HE groups of genes are indicated. (F) Scheme summarizing the results.


A
Density
0.00
0.04
0.08
0.12
-10
-5
0
5
10
15
log$_2$ RPKM
Exons
Introns
Intergenic
90% quantile of intergenic
B
Density
0.00
0.04
0.08
0.12
0
5
10
15
log$_2$ expression (AU)
C
Density
0.00
0.02
0.04
0.06
0.08
0.10
-10
-5
0
5
10
log$_2$ RPKM
Data
LE fit
HE fit
Sum of LE and HE fits
LE
HE


A
Number of undetected genes
7000
8000
9000
10000
5
10
15
20
Number of reads (x 1,000,000)
Number of undetected genes
7000
8000
9000
10000
20
21
22
23
24
$\log_2$ Number of reads
B
$\log_2$ RPK
-5
0
5
10
15
20
Density
0.00
0.04
0.08
0.12
0
20
40
60
80
100
-10
-5
0
5
10
15
$\log_2$ RPKM
Actual data
% of genes at this level that are undetected
Prediction of expressed genes
C
PCR (-Ct)
-30
-25
-20
-15
RNA-seq (density)
0.00
0.02
0.04
0.06
0.08
0.10
-10
-5
0
5
10
15
RNA-seq ($\log_2$ RPKM)
D
$\log_2$ RPKM (Introns)
-15
-10
-5
0
5
10
LE
HE
$r^2 = 0.86$
$p < 10^{-16}$
$r^2 = 0.51$
$p < 10^{-16}$
$r^2 = 0.68$
$p < 10^{-16}$
-10
-5
0
5
10
$\log_2$ RPKM (Exons)


A
Gata3
Tbx21
Frequency
Gata3
0
10
20
30
0
50
100
150
200
250
300
350
Transcripts
Frequency
Tbx21
0
200
400
0
5
10
1
10
20
30
0
10
20
30
40
Transcripts
B
r^2 = 0.0029
p = 0.9485
Tbx21 mRNAs/cell
0
10
20
30
0
50
150
250
Gata3 mRNAs/cell
C
RPKM
0
50
100
150
0
20
40
60
80
mean mRNAs/cell
D
RNA-seq
IgG control
$\log_2$ ChIP-seq (NLCS)
LE
HE
0.00
0.01
0.02
0.03
0.04
Density
$\log_2$ RPKM
-15
-10
-5
0
5
10
H3K9/14ac
Background
Signal
0.00
0.02
0.04
0.06
0.08
Density
E
Microarrays
IgG control
H3K9/14ac
$\log_2$ expression (AU)
0
5
10
15
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Density
F
histone modification
no histone modification
HE
> 1 mRNA/cell
+H3ac
LE
< 1 mRNA/cell
low expression
high expression