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

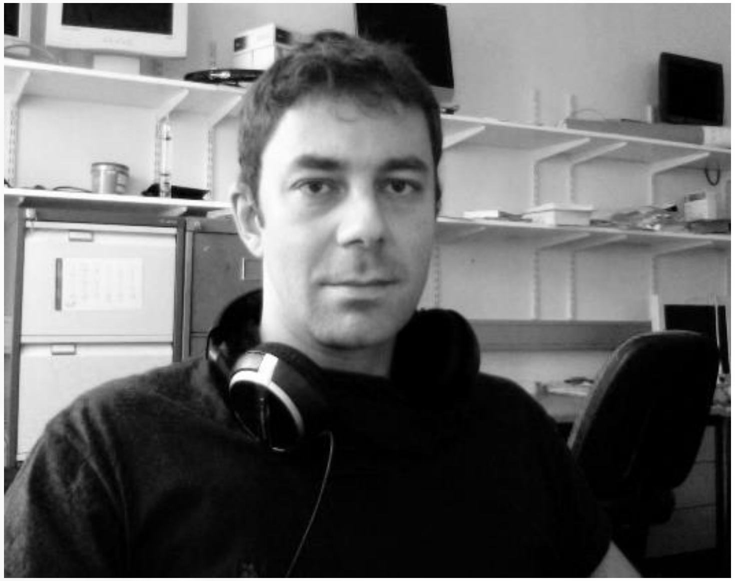
RNA测序显示后生动物细胞中基因表达水平分为两类

daniel hebenstreit

Theoretical and Computational Biology (TCB) group

at

MRC Laboratory of Molecular Biology



Interest in

mechanisms and statistics of transcriptional regulation in metazoan cells

- Hebenstreit D, Gu M, Haider S, Turner DJ, Lio' P, Teichmann SA.

EpiChIP: Gene-by-gene quantification of epigenetic modication levels.

Nucleic Acids Res. 2011 Mar;39(5):e27. Featured article.

- Hebenstreit D, Teichmann SA.

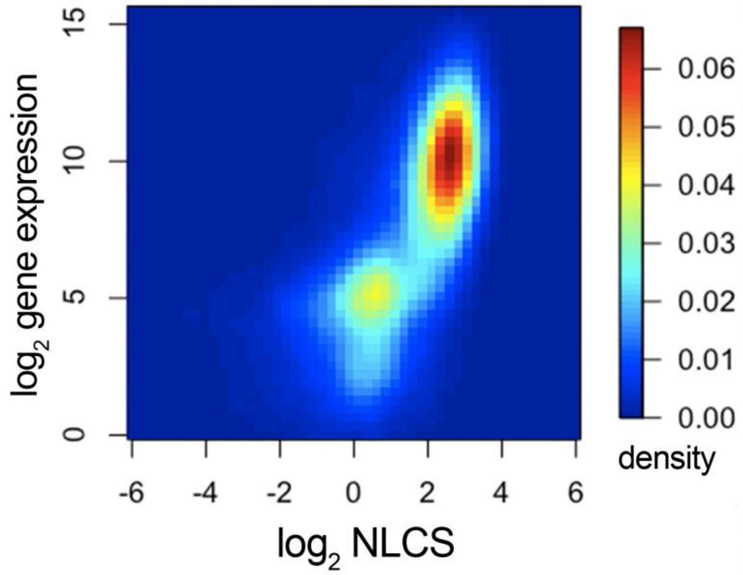
Analysis and simulation of gene expression profiles in pure and mixed cell populations.

Phys Biol. 2011 Jun;8(3):035013.

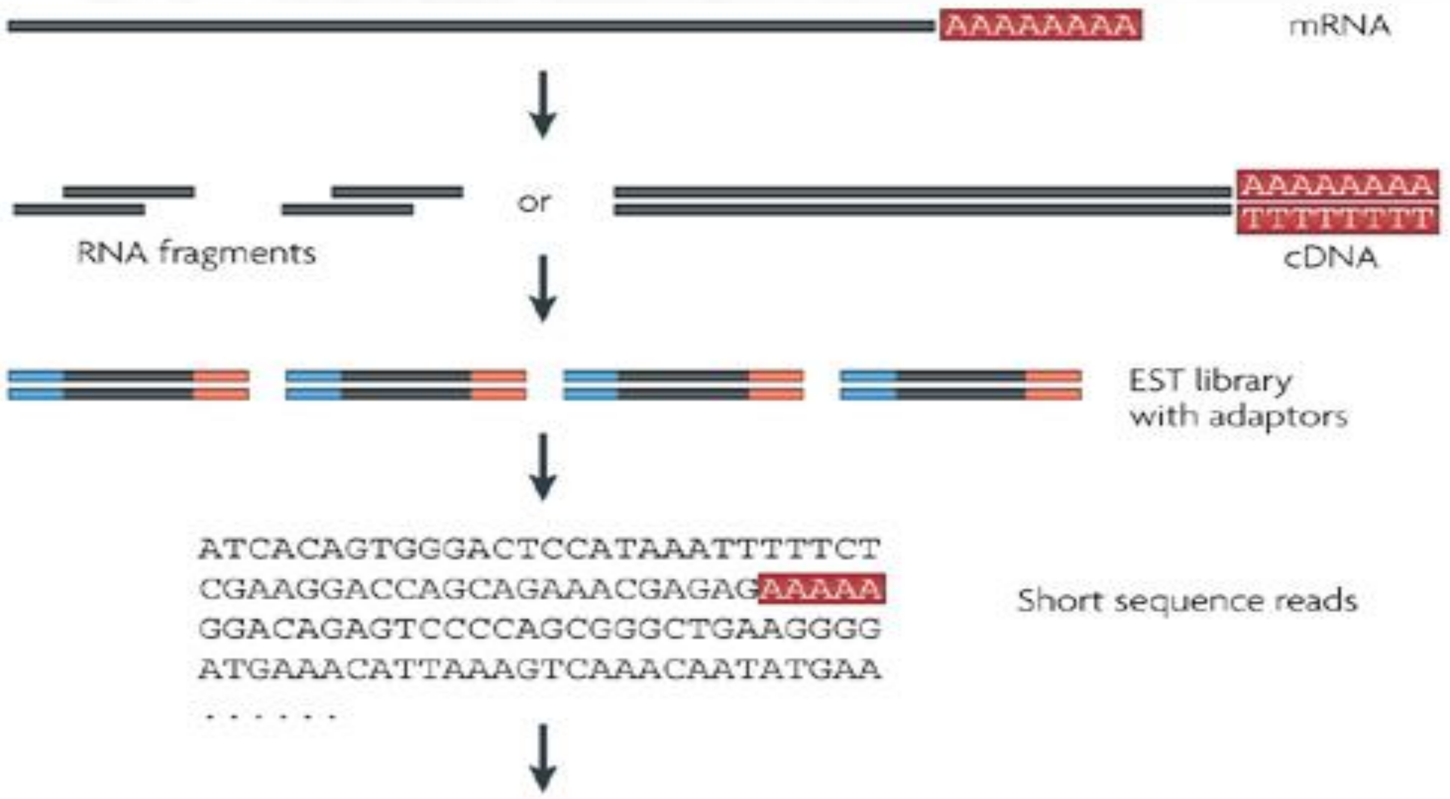
- Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA.

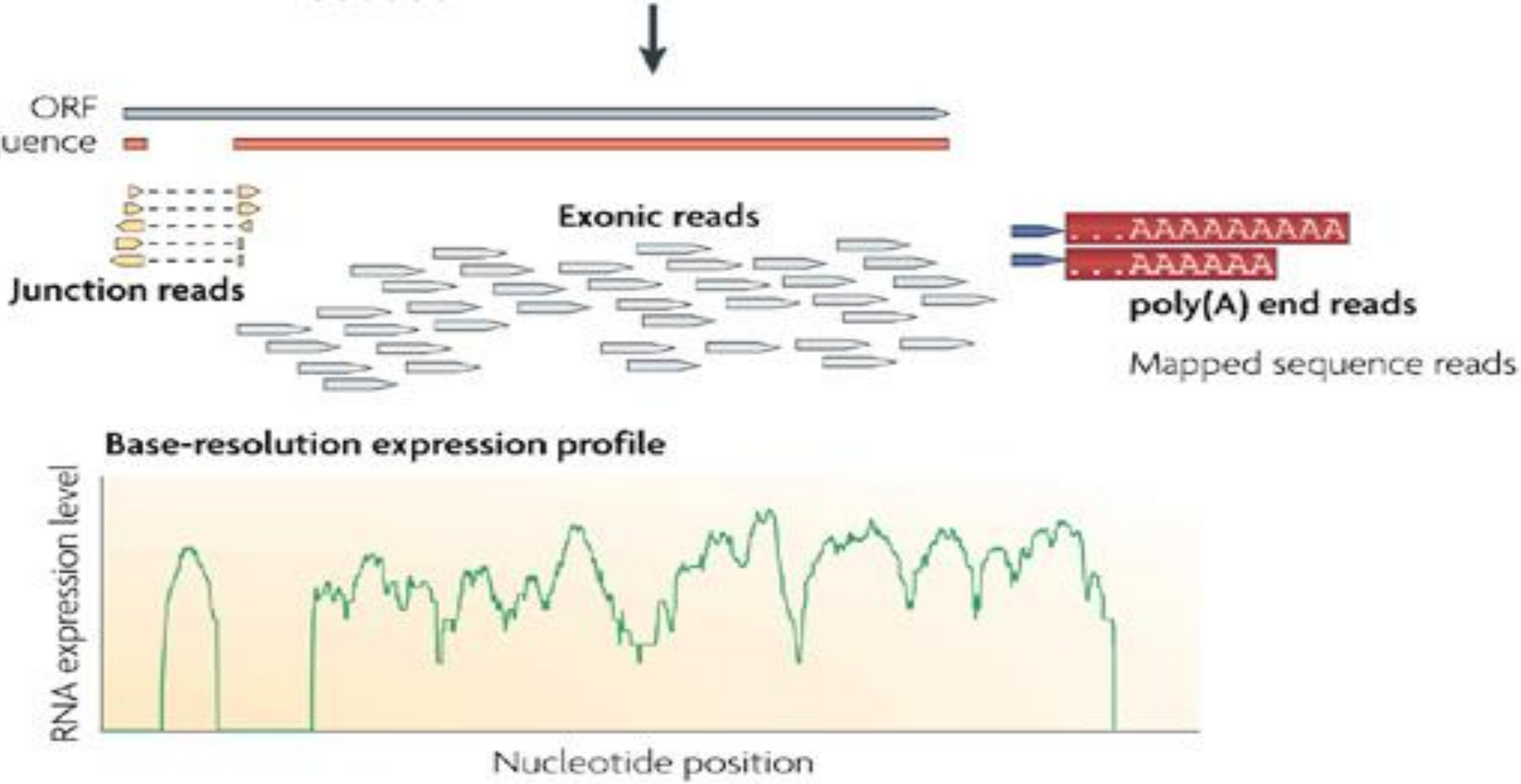
RNA sequencing reveals two major classes of gene expression levels in metazoan cells.

The heatmap shows the gene expression (on the y-axis) and activating histone modification (as normalized locus specific chromatin status, NLCS, on the x-axis) levels for all genes in a murine Th2 cell. The structure clearly reveals two main groups of genes, lowly expressed ones without activating histone marks, and highly expressed ones with activating histone marks. This suggests a fundamental switch-like mechanism underlying transcriptional regulation.

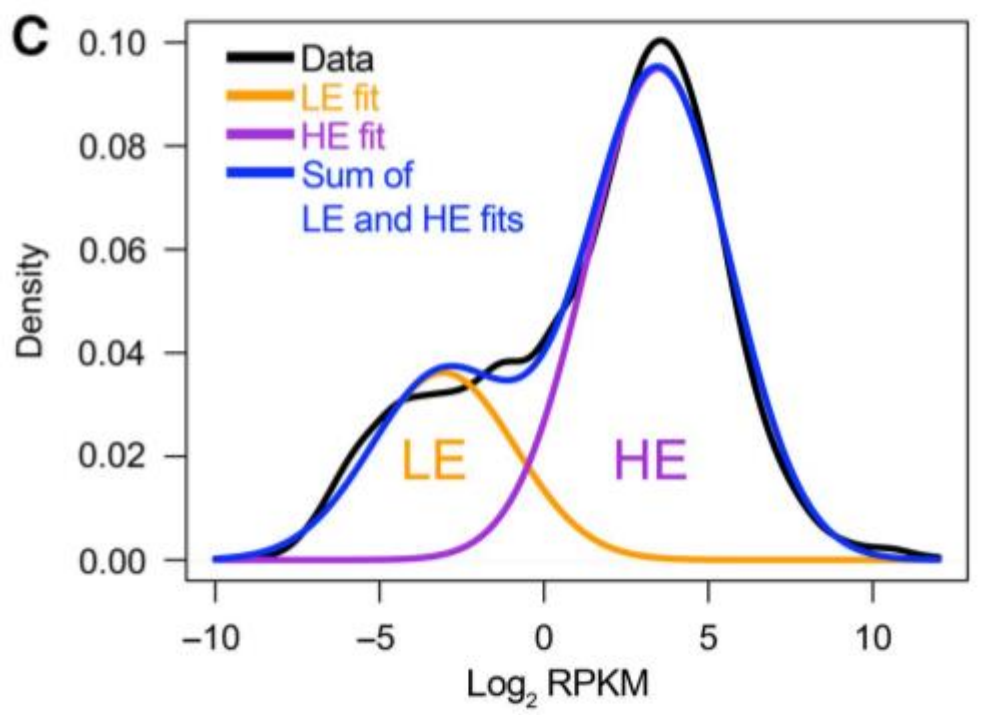


RNA-seq, also called "Whole Transcriptome Shotgun Sequencing", refers to the use of high-throughput sequencing technologies to sequence cDNA in order to get information about a sample's RNA content.

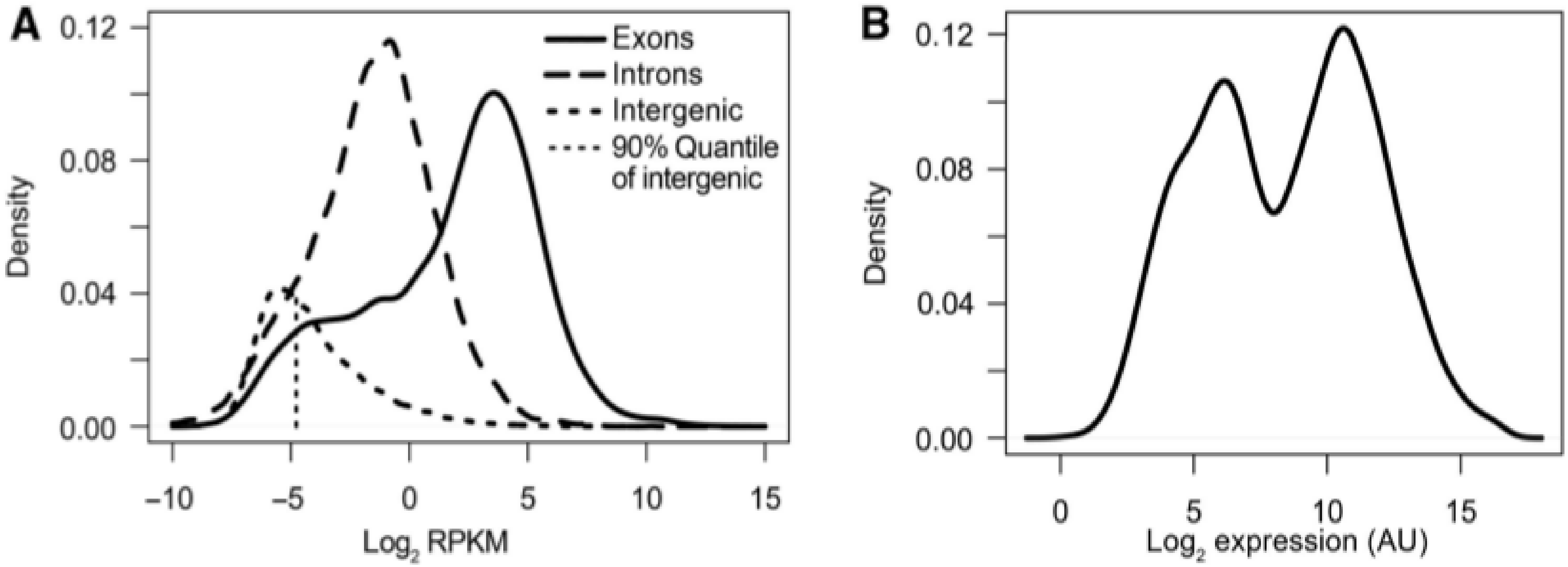

AAAAAAAA
mRNA
RNA fragments
or
AAAAAAAA
TTTTTTTT
cDNA
EST library
with adaptors
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
....
Short sequence reads


ORF
Coding sequence
Exonic reads
Junction reads
...AAAAAAAAA
...AAAAAA
poly(A) end reads
Mapped sequence reads
Base-resolution expression profile
RNA expression level
Nucleotide position

- how many transcripts are expressed in a cell at what levels?
- microarrays or RNA-seq data have been described as displaying broad, roughly lognormal distributions of expression levels with no clear separation into discrete classes.
- two overlapping major mRNA abundance classes



Kernel density estimates of RPKM distributions of murine Th2 poly(A)+ RNA-seq data

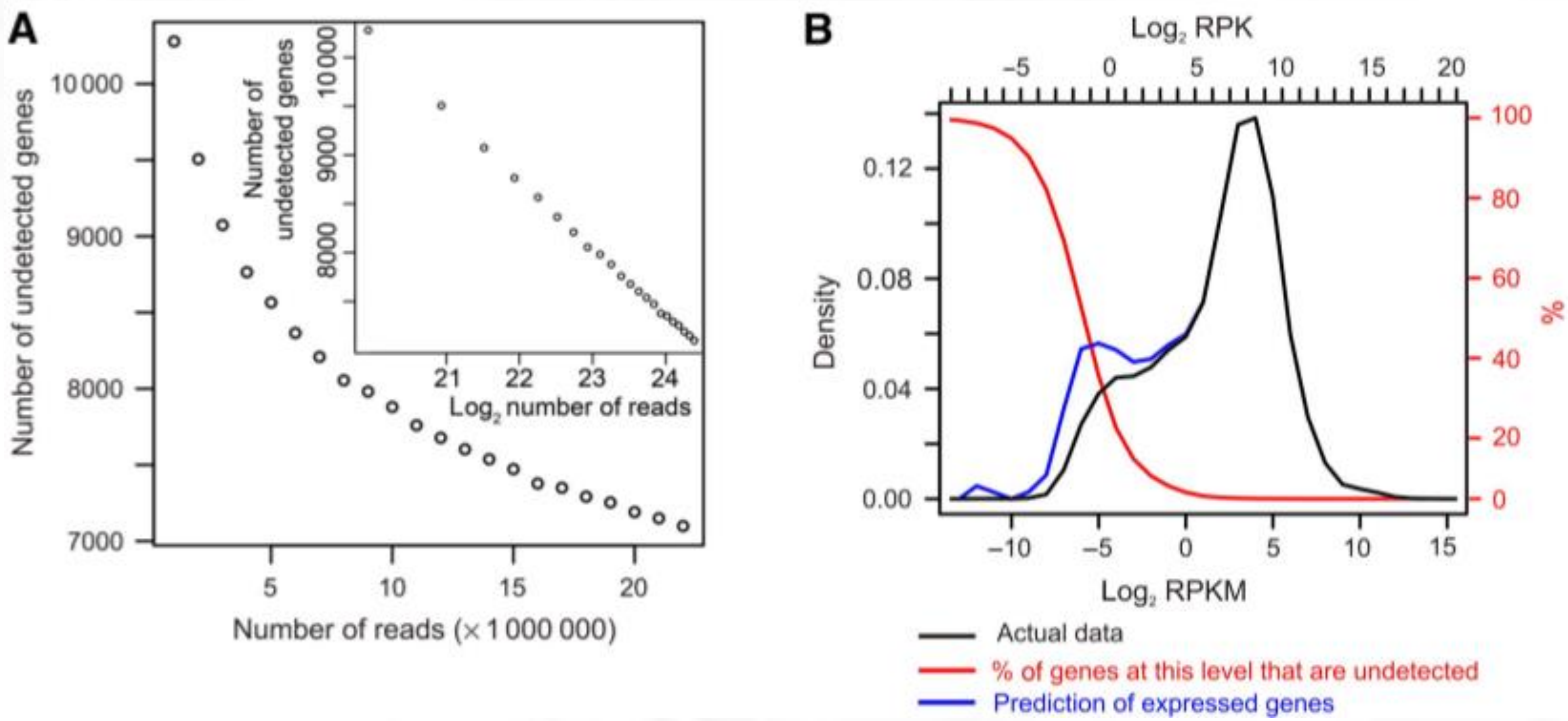


The fragments used to estimate intron and intergenic RPKM were based on randomizations using the same length distribution as the exonic parts of genes

- Kernel density estimate of expression level distribution of microarray data
- The bimodality was conserved when alternative normalization and processing schemes were used,independent of KDE bandwidths

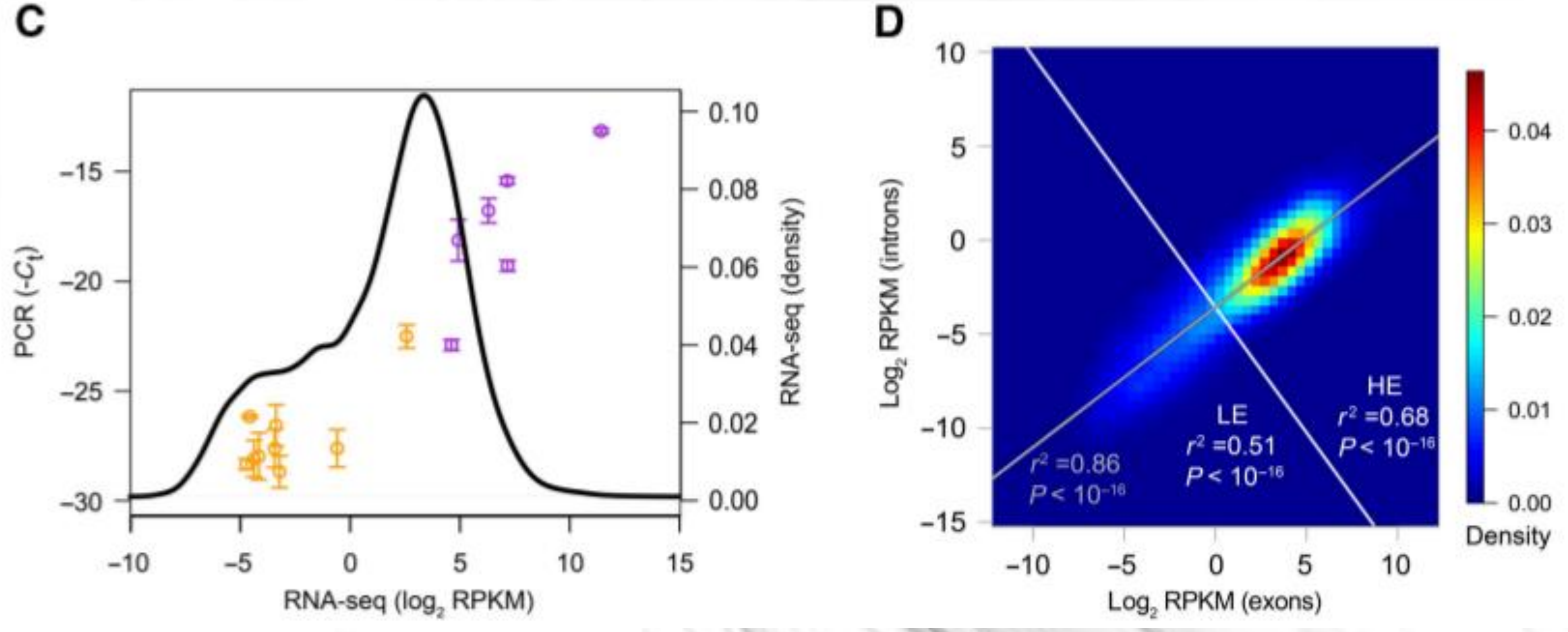
Visual inspection and curve fitting of both microarray and RNA-seq data thus reveals two overlapping main components of the distribution

how accuracy bias affects the shape of the LE distribution

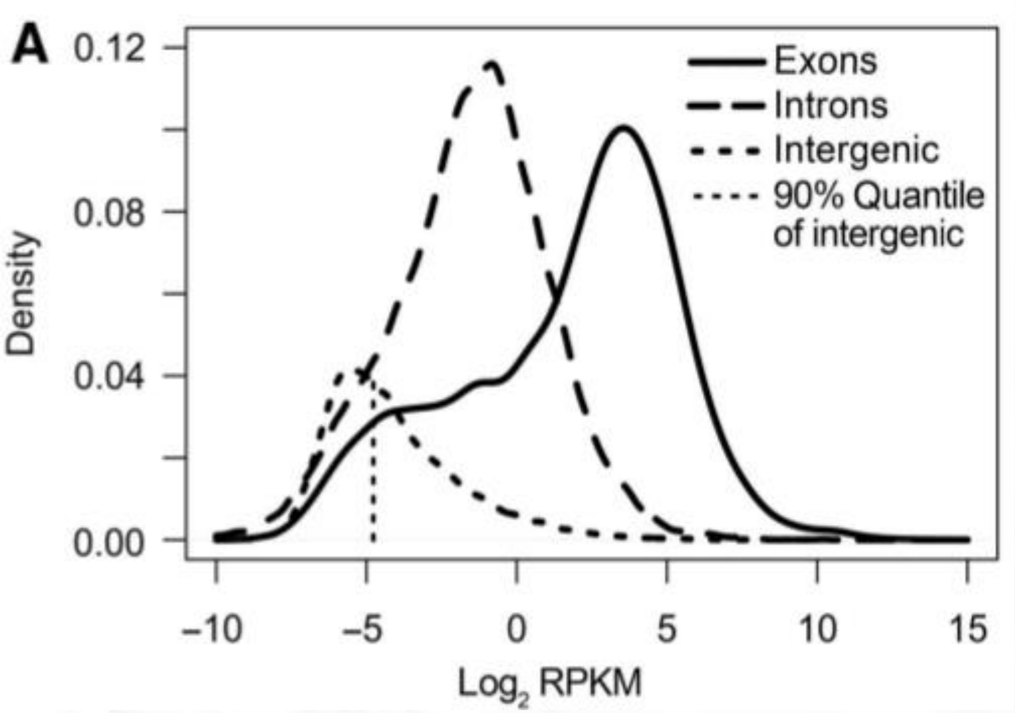


the LE genes correspond to low expression and not experimental noise

- The expressed genes map to the HE peak, while almost all unexpressed genes map to the LE peak.
- 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.



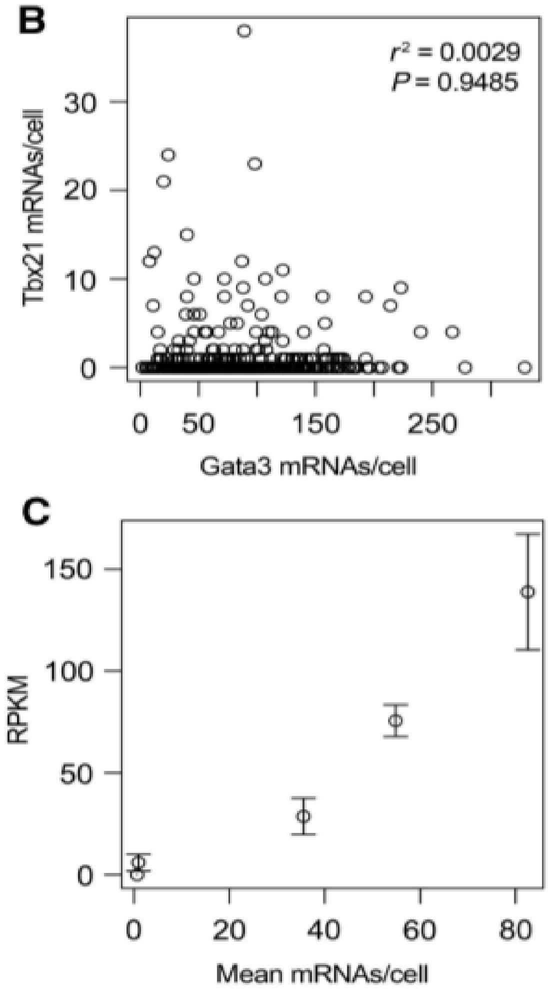
The 90% quantile of this intergenic background distribution is at -4.97 $\log_2$ RPKM, which means that we can be quite confident (with probability>90%) that genes with an RPKM value above this level are truly expressed rather than representing experimental background noise.



we are detecting incompletely processed transcripts at a low but significant and uniform level across the whole range of transcript abundances.

Result of single-molecule RNA-fluorescence in *situ* hybridization for five genes that are expressed at different levels according to the literature and our RNA-seq data

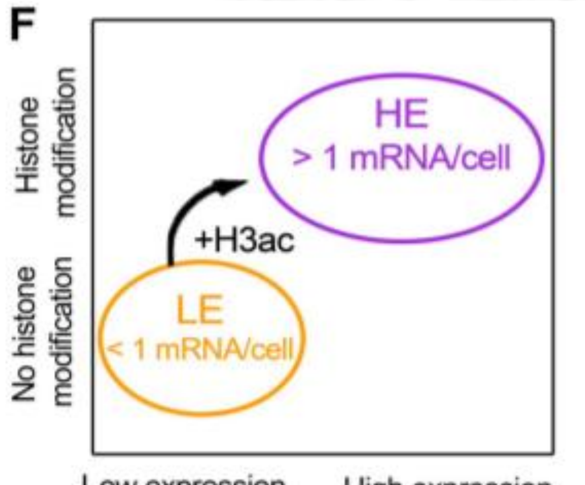
cells expressing Tbx21 were not anticorrelated with cells expressing Gata3, meaning that we do not have a subpopulation of Th1 cells in our Th2 cell populations. Thus, LE expression is not due to a contaminating cell type, as the same cells express groups of genes at HE and others at LE levels.



RNA-FISH indicates that one RPKM corresponds to an average of roughly one transcript per cell in our Th2 data set.

H3K9/14ac marks are associated with the promoters of HE genes only

- As 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.
- 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
- there is a very weak correlation within the LE and HE groups



- Two groups one transcript per cell
- It thus seems likely that the LE/HE groups reflect different transcription kinetics depending on the chromatin state or vice versa.
- 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.
- Often, differential regulation induces only small changes in expression levels, probably serving to fine-tune expression and shifting genes within the HE group.
- 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.

THANK YOU GUYs