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

Daniel Hebenstreit, Miaoqing Fang, Muxin Gu, Varodom Charoensawan, Alexander van Oudenaarden, Sarah Teichmann,

Corresponding authors: Daniel Hebenstreit, or Sarah Teichmann, MRC Laboratory of Molecular Biology

Review timeline:

Submission date:	18 November 2010
Editorial Decision:	31 December 2010
Resubmission:	01 March 2011
Editorial Decision:	07 April 2011
Revision received:	18 April 2011
Accepted:	19 April 2011

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 31 December 2010

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the three referees whom we asked to evaluate your manuscript. As you will see from the reports below, the referees raise substantial concerns on your work, which, I am afraid to say, preclude its publication.

The reviewers had substantial concerns that the observed bimodality of gene expression could be an artifact of the computational analysis, and felt that the generality of these observations across different cell types remained largely unclear. The editor notes three particularly important concerns:

1. The first reviewer was concerned that the choice of bowtie read mapping settings use in this work could lead to potential cross-mapping issues, which could potentially explain the bimodal distribution.

2. The second reviewer was concerned that the existence of the bimodal distribution rests heavily on the kernel density estimation based graph and the subsequent BIC-based model likelihood, and was concerned that different tests or parameter choices could lead to different conclusions.

3. The reviewers were not convinced that this bimodal distribution was a general phenomenon relevant to other cell types/populations. The reviewers felt it would be important to clarify why this distribution had not been observed in previous experiments. On a somewhat related topic the last reviewer was less than fully convinced that the Th2 cell population used here was truly pure.

Since these concerns raise substantial doubts regarding the biological relevance of the bimodal distribution observed in this work, and because the second reviewer indicated clearly that they could not support publication of this work, we feel we have no choice but to return this work with the message that we cannot offer to publish it.

Nevertheless, the reviewers did recognize that, if correct, the results presented in this work could be potentially important. Moreover, while the reviewers raised substantial concerns, they also suggested constructive analyses that could potentially help to allay their concerns. As such, we would like to suggest that we may be willing to reconsider a substantially revised work that convincingly addresses the reviewers' key concerns. This would have a new number and receipt date. We recognize that this may involve further experimentation and analysis, and we can give no guarantee about its eventual acceptability. However, if you do decide to follow this course then it would be helpful to enclose with your re-submission an account of how the work has been altered in response to the points raised in the present review.

I am sorry that the review of your work did not result in a more favorable outcome on this occasion, and I am sorry to send negative news during the holidays. Nonetheless, I hope you find these reviews helpful and that you will not be discouraged from sending your work Molecular Systems Biology in the future.

Thank you for the opportunity to examine this work.

Yours sincerely,

Editor

Molecular Systems Biology

____

Reviewer #1 (Remarks to the Author):

This manuscript claims that the expression profiles of cells are essentially bimodal, i.e. genes are either on or off. The off (or "LE") state would appear to be somewhat leaky, such that the distribution of reads/intensities is below one copy per cell, but still well above the intergenic background, and forms its own peak in the distribution of RNA abundance.

This observation is so straightforward that it is hard to believe that it hasn't been previously noticed. Here, however, it is supported in several ways - RNA-seq, microarrays, ISH, qPCR, and correlation with H3K9. Given a few caveats below which must be addressed, it may be correct. I am not aware of any previous papers making such a claim. I believe that this finding will be of widespread interest, provided it is true.

The paper also contains a few other treasures that will be very useful to researchers in the field - for example, the estimate that 1 RPKM ~= 1 copy per cell.

I have one major technical concern that needs to be addressed in order to show that the data is really correct as shown. There are additional points that are also troubling, which should also be fixed, because otherwise the paper doesn't make sense, fails to address key issues, compare its findings to previous literature, etc.

(1) The technical concern is the mapping method, which is bowtie with default settings to map the reads to the genome. Used this way, each read is mapped to the position of the first valid hit. This means that it is possible that a read is mapped to a position with two mismatches rather than a better position with zero mismatches, simply because bowtie considered the former position first. Second, and more importantly, bowtie also reports non-unique reads when using the default settings (again on the basis of the first valid read). This means that there is a potential for cross-mapping in the RNA-seq data (equivalent to cross-hybridization on arrays), which might explain the shoulder observed on the left in the read count per transcript distribution. This is especially relevant considering that the paper already suggests that cross-hybridization is the most logical explanation for the bimodal distribution of the array data. Therefore, the paper needs to show that the same distribution is obtained when they only consider 1) the best mapping reads (with the bowtie options --best and --strata) and 2) uniquely mapped reads only (with the -m option). If the shoulder disappears, then the distribution will not be bimodal, the two-populations argument is invalid, and many of the claims in the paper are untrue.

(2) Another potential issue is that the new RNA-Seq data in the paper is not strand-specific, so it's hard to determine whether the LE signal actually comes from the sense strand. Was the Cloonan et al. SOLID data mapped and considered in a strand-specific manner? This issue needs to be discussed. There may be other data sets in the literature that are better suited for this analysis.

(3) The paper should explain why previous groups have not previously seen the bimodal distribution that is so obvious here. For example, the Ramskˆld 2009 paper that is cited here states in the Abstract that "no support was found for the concept of distinct expression classes of genes". The evidence is found in Figure 2 of that paper and it is a different kind of graph - is this the explanation? Why are the Wang 2008 data (which were used in Ramskˆld 2009) not included in the analysis of other data sets here? This would allow direct comparison with the conclusions of Ramskˆld 2009. If the same type of analysis is applied and the bimodal distribution is not seen, then the findings in the present paper should be qualified to state that the bimodal distribution may be specific to certain cell types. Did Ramskˆld map the reads in a different way?

(4) On P. 3, it is stated that "Figure S1 shows expression of a marker protein in the cells". I'm actually not sure what Figure S1 is supposed to show. There are four markers, no? The magenta lines aren't explained. The units are arbitrary. What is the purpose of this figure, even?

(5) The explanation for the discrepancy between RNA-seq and microarrays doesn't completely make sense to me and it would be good to describe it in more "layman's terms" - i.e., something that can be understood by the average person with a PhD in molecular biology. The main thing I can't get my head around is why the sampling problems of RNA-seq should matter much, relative to the obvious false-positive problems of microarrays (and even qPCR on occasion). If the gene is really "off" but firing at a low level, is it important what that level is? It should still be classified into the low-but-not-off bin. If it has zero reads, where does it end up in Figure 1A and 2D? Is it excluded? Perhaps it would be better to put it into a catch-all bin which represents some practical lower limit of detection?

In any case, Figure S7 could be presented in a more straightforward way, or perhaps described in a more straightforward way. Why are there two different axes on Figure S7B (on top and bottom)? (I think the one on top is correct, based on comparison to Figure 1A). I would suggest some sort of model of the actual transcriptome abundance, and then run the Poisson test on that - aren't the results dependent on the actual distribution, which is unknown? Or is that what the splits are supposed to address? I suppose it is, but it would be nice to know for sure. In any case the black line would be "model" or "actual data", the red line would be "expected detection with 25M reads", and the red line would be labelled as it is already. And for the red line, instead of "Density", it would be good to plot "percent of genes at this level that are undetected".

In any case, it looks like all the problems are for genes that are expressed at less than one copy per cell, and that the likelihood of going undetected is below 5% up until about an order of magnitude lower. This suggests that the read depth problem really isn't an issue unless there are zero transcripts in the vast majority of cells. Something that might deserve discussion.

(6) The point above relates to the appearance of Figure 2D, right panel. I suspect that the faintness of the blob in the lower left (relative to the same graph in Figure 2E) may be due to the fact that the transcripts with zero reads are either omitted or more dispersed due to the "Poisson effect", relative to microarray data where they pile up at the background level. (Or, alternatively, the blob in the lower left may be due entirely to cross-hyb, and its modest appearance in the RNA-seq data due to the fact that the bowtie procedure results in less "cross-hyb" than the microarray data - so the real distribution would have a much more dispersed blob in the lower left). If the zero-read genes are assigned to some background level as I suggest above, does the phenomenon become more clear? I would expect the graph to pick up a bright vertical line in the blob in the lower left. Perhaps the intronic reads could be included. Anyhow, the Figure 2D and E right panels are critical to the claims of the paper but I can't tell if their appearance is mainly due to cross-hyb and data processing.

(7) There is a paragraph on P. 5 that begins "We also tested the extent to which genomic DNA may be contamination our polyA-purified mRNA sample...., and it ends with a sentence "....which means that genes with RPKM above this level are at least 90% probable to be expressed". This doesn't make sense to me. The way this is written, it sounds like the intergenic fragments were chosen because they shouldn't be expressed, i.e. they are a negative control. (and if so, why aren't they chosen to be at least 10 kb away from genes, as the materials and methods state? The text here says 5 kb). If they are used to determine the background expression of the genome, then this is a whole different ball game and the paragraph needs to be rewritten and the background/false detection level set in some other (independent) way. The way I would interpret the outcome shown would be that some portion of the polyA-purified sample could in fact be explained by genomic contamination, but it is a very small proportion. Anyhow, this paragraph should be revamped - it would be very interesting to know what proportion of interegenic DNA is detected, of course, and why. But it's not clear what is the real purpose of the analysis here, and what can be claimed from it.

(8) Figure 2C is kind of disturbing - it actually looks like the intercept of the top three points is around 20 mRNAs/cell = 0 RPKM. One wonders whether the genes at the origin simply don't work by in situ, or whether there is some issue with the point at (38, 25) which would straighten things out. I think this issue deserves some commentary, because the calibration between RPKM and copies per cell is likely to be widely cited. As it is, the graph could be used to argue that RNA-seq cannot detect genes expressed at less than 20 copies per cell. (As an aside, is 1 copy ~= 1 RPKM consistent with the estimates one could make given the number of RNAs per cell and their lengths?)

(9) I am confused by the last four sentences of the paper. The idea is that transcription factors mediate the variation among expressed genes. But don't transcription factors also control which genes are on and off in the different cell types? i.e. the "key decision about whether a gene becomes switched on and expressed" should also be determined by transcription factors that are recruiting the epigenetic/chromatin factors. What else could it be? The chromatin proteins by definition don't have sequence specificity, otherwise we would call them transcription factors.

Reviewer #2 (Remarks to the Author):

The central premise of this manuscript is that gene expression levels in homogeneous cell populations can be divided into two distinct groups: the lowly expressed (or LE) group, contains genes that are expressed below one stable mRNA / cell (typically one stable mRNA / eight cells); and the highly expressed (or HE) group that contains mRNAs with marks of active chromatin at their promoters. The manuscript claims that these observations challenge a prevailing view that the LE group represents genes previously thought to be not expressed at all and that there is no distinct HE group, i.e., that expressed mRNAs have a wide range of different expression levels.

This manuscript is in the unenviable position of having to provide strong and convincing arguments that the current view is inaccurate. Although the main observation is potentially interesting and significant, I suspect that the observed bimodality in the distribution of gene expression levels (the source of the manuscript's claim) is an artifact of how the analysis was done. Even if this bimodality is real, I am not convinced that this is a general phenomenon - the manuscript only presents two examples, it should be straightforward to find and present similar analyses on microarray data from a much larger range of homogeneous cell populations.

First, let me address the observations regarding the LE mode. Unlike the authors, I do not think that this claim is controversial. Although these genes are often referred to as unexpressed, I doubt that many would be surprised if there was some leaky transcription and that these transcriptional products were processed into mature mRNAs. Presumably one role of miRNAs is to repress the expression of some of these unwanted transcripts. As the manuscript points out, there is no evidence that these LE mRNAs are translated. I would be surprised if this was a new observation: There is an ongoing debate regarding how much of the genome is transcribed and stably expressed, I suspect one of the participants in this debate has already made this claim.

On the other hand, the existence of a distinct HE group is surprising. I have looked at a large number of microarray intensity histograms and I have never seen anything like the bimodal distribution in figure 1B. Albeit, most of the time I was looking at data from mixed cell populations. The manuscript presents two sources of evidence that the intensity distribution is bimodal. The first is visual; there are two clear modes in figure 1B. My concern with this picture is that it was generated using kernel density estimation (KDE) and I could find no description anywhere in the manuscript as to how the bandwidth of the kernel was selected (or even what the kernel is, but I suspect it is Gaussian). This is a glaring omission and this is an important piece of information because by varying the bandwidth in KDE, one can generate anywhere from a unimodal distribution to a very multimodal one. To strengthen this part of their argument, the authors need to show that the bimodality is relatively insensitive to the choice of the bandwidth and furthermore that it appears when using other techniques are used to estimate the distribution (i.e. a histogram). The manuscript also provides a Gaussian mixture model analysis to support the bimodality. However, this analysis actually points to more than two modes; in figure S5, the maximum of the BIC-adjusted likelihood occurs at four, not two modes. The fact that there is a large increase in BIC-adjusted likelihood going from one to two modes is not surprising, figure 1A is clearly non-Gaussian, but it does not support the existence of exactly two modes, as required to support the manuscript's premise. Note also that BIC is simply one of a large number of different model selection methods, and the authors need to justify this choice and/or show that the same observations hold if they use AIC, a likelihood ratio test, or the Bayes factor.

So in summary, in my opinion, the main novel observation in this manuscript rests on what appears to be flawed computational analysis. As such, I cannot support its publication.

Reviewer #3 (Remarks to the Author):

RNA sequencing reveals two major classes of gene expression levels

Hebenstreit et al.

SUMMARY

In this short paper, Hebenstreit et al. use RNA-sequencing of mouse Th2 cells to observe that expressed genes can be separated into two classes: lowly expressed and highly expressed. They contrast this with some recent observations from RNA-sequencing data that gene expression levels are more continuous. They support their observations with ChIP-seq of a histone modification that is a strong indicator of transcription and single-cell RNA-FISH.

Overall, I found this paper well written and enjoyable to read. Whilst not strictly original, since the ideas tested were first proposed many years ago, the authors do make some relevant points and present new data that contributes to this important issue in our understanding of gene expression regulation. Thus, while I feel that this manuscript does require some revision, I think that it is generally suitable for publication in Molecular and Systems Biology.

MAJOR COMMENTS (in no particular order)

1. I think the justification for there not being a mixture of two cell types provided in the third paragraph on page 4 leaves something to be desired. In particular, I think the argument relies upon the somewhat pathological assumption that if multiple cell types were present that they would be so in the same ratio (i.e., 50% cell type 1, 50% cell type 2). In the far more likely case that you have an overwhelming amount of cell type 1 (80% say) and a much smaller proportion of cell type 2 (20%) then I think the argument is much harder to make. Also, I would move this paragraph to later on in the manuscript when the single-cell RNA-FISH data are discussed. In a similar vein, I am not sure about the argument on page 6 that the cell population you analyzed has no Th1 cells present. In particular, Zheng and Flavell (Cell, 1997) state: "Gata3 is necessary and sufficient for Th2 cytokine gene expression in CD4 T Cells". How do you reconcile this with the small number of cells that do not express Gata3, or that express it at a very low level? Do you think this is a limitation of the RNA-FISH data? Or is it a sign of low-level contamination? I think this possibility can not be excluded on the basis of the data presented and this should be reflected in the manuscript.

2. Did you separate the intronic regions depending upon their vicinity to the 3' end of the gene? Given the relative speed of the splicing machinery and the DNA polymerase, I would expect to see more reads towards the 3' end of each gene. It would be interesting to perform this analysis and to determine whether information about the rate at which splicing occurs could also be modeled.

3. In Figure 2D and 2E (right hand panel) there does seem to be a positive correlation between H3K9/14ac and expression across all expression ranges. I know that the authors state that the signal for lowly expressed genes is consistent with background, but the relation is rather striking (at least to this reviewer). Hence, I think the assertion that "H3K9/14ac marks are associated with promoters of highly expressed genes" should perhaps be toned down. Alternatively, the cutoffs/methods for ChIP-seq analysis need to be justified better. In particular, it was not clear to me that the ChIP-seq data were normalized - therefore, that the cutoff between signal and background was the same for all four ChIPs seemed somewhat surprising to me. This could do with some more explanation I think.

4. I found the method used for calculating the expected number of reads for each gene somewhat strange. In particular, the use of RPK in this instance instead of the RPKM seemed clumsy. Why did you just not include gene-length and the total number of reads in the lane as offsets in the Poisson model? This would allow all of the analysis to be done on the RPKM scale and would offer much easier interpretation I think.

MINOR COMMENTS

1. On page 4 it is stated that the two-component model shows "strong increases or maximal values for both microarray and RNA-seq data". In Figure S5, whilst there is clearly strong evidence for a two-component model as opposed to a uni-modal distribution, in no case is the two-component model the best fitting as adjudged by the BIC. The text should be adjusted to reflect this.

2. On page 6, "non-Poissonian distributions (which would have...." should be replaced with "that they had extra-Poisson variation (a Poisson random variable would have..."

3. On page 7, you should quantify the statement in the second paragraph that "...many genes that are characterized as not expressed and non-functional in Th2 cells...". What proportion of genes is this?

4. In the supplement, on page 3, third line you should refer to Table S2, not Table S4.

5. I found Figure S6C difficult to interpret - can you remove the black rectangles that surround the whole of each bar please - they add nothing to the figure and only ended up confusing this reviewer!

Resubmission 01 March 2011

After substantial revisions and additional analyses, we would now like to re-submit our manuscript (MSB-10-2556) entitled ‘RNA sequencing reveals two major classes of gene expression levels in metazoan cells’ for publication as a ‘Report’ in Molecular Systems Biology.

Reiterating the main conclusion, our work showed that the distribution of mRNA abundance levels is bimodal, revealing two major expression levels. The three key points to note are as follows:

1 - Four separate methods for quantifying mRNA abundance support the separation into lowly versus highly expressed genes: microarrays, mRNA-sequencing, qRT-PCR and single molecule RNA-FISH. This means that though all genes are expressed, there is a distinction between stochastic, leaky background expression and a functionally relevant expression level above about one mRNA per cell on average.

2 - An entirely orthogonal approach, ChIP-sequencing of an activating histone mark, also recapitulates two distinct abundance classes, illustrating that the switch-like transition from low to high expression goes hand-in-hand with a change in chromatin status.

3 - These findings have technical implications in terms of methods for analysis of RNA-sequencing data, and more importantly, broad implications for our understanding of gene expression regulation.

Previously, the referees recognized the potential impact of the insights summarized above: Referee 1 noted "I believe that this finding will be of widespread interest, provided it is true. The paper also contains a few other treasures that will be very useful to researchers...." and Referee 3 said "the authors...present new data that contributes to this important issue in our understanding of gene expression regulation...I think that it is generally suitable for publication in Molecular and Systems Biology". The third referee had technical concerns that we have now addressed.

With respect to the major queries noted previously,

1 - Settings used in the read mapping software could potentially lead to cross-mapping (referee 1). We have now shown that this is not the case using an alternative mapping strategy suggested by the referee. 2 - The methods used for data representation and model-fitting could affect the conclusions (referee 2). We have now recalculated the kernel density estimates and model fitting using a wide range of parameters, and the conclusions still hold.

3 - Our findings are possibly not of general nature and are affected by impure cell populations (referee 3). We have now included many further datasets to illustrate the generality of our conclusions, and have appended a paper of ours that addresses the effect of mixtures of cell types.

Our resubmission addresses these concerns and all the other questions raised by the referees in detail (Please see the point-by-point reply below). In particular, the text has been strengthened in content and clarity, and there is one new main figure, five new supplementary figures, and analyses of four additional RNA-seq datasets, three additional microarray datasets and many more analyses of the data already previously in the paper.

We hope that two manuscripts of ours which are now in press help to clarify (i) the method we use for processing the ChIP-sequencing data, which is a software called EpiChIP (Hebenstreit et al., Nucleic Acids Res, in press, available online) and (ii) the effects of heterogeneous populations of cells (Hebenstreit et al., Physical Biology, in press, proofs to this resubmission). In this paper, we simulate the effect of mixing cell types based on microarray data, and also use two lognormal distributions to describe data from individual homogeneous cell populations. Although we briefly introduce the concept of bimodal distributions, we do not investigate the nature of these and do not present any new experimental data. Thus this is a much more limited and focused piece of work than the current manuscript, which aims to interpret the entire range of gene expression levels in terms of transcriptional status, biological function and histone modification.

We hope you agree that the technical issues are now resolved and the impact of the results merit reconsideration at Molecular Systems Biology, and look forward to hearing from you,

Reviewer #1 (Remarks to the Author):

(1) The technical concern is the mapping method, which is bowtie with default settings to map the reads to the genome. Used this way, each read is mapped to the position of the first valid hit. This means that it is possible that a read is mapped to a position with two mismatches rather than a better position with zero mismatches, simply because bowtie considered the former position first. Second, and more importantly, bowtie also reports non-unique reads when using the default settings (again on the basis of the first valid read). This means that there is a potential for cross-mapping in the RNA-seq data (equivalent to cross- hybridization on arrays), which might explain the shoulder observed on the left in the read count per transcript distribution. This is especially relevant considering that the paper already suggests that cross-hybridization is the most logical explanation for the bimodal distribution of the array data. Therefore, the paper needs to show that the same distribution is obtained when they only consider 1) the best mapping reads (with the bowtie options --best and --strata) and 2) uniquely mapped reads only (with the -m option). If the shoulder disappears, then the distribution will not be bimodal, the two-populations argument is invalid, and many of the claims in the paper are untrue.

We had previously excluded non-uniquely mapping reads using our own custom script that parses the column of the bowtie output reporting the number of matched regions. We have now repeated the mapping of our data with the options -m 1 --best --strata . This did not alter the structure of the data, and did not change the distribution of RPKM values.

We have now changed all figures so that the newly mapped data are displayed, and have altered the following sentence in the RNA-seq data processing section of Materials and Methods by adding (underlined) and removing words (strikethrough):

"Reads were mapped to the mouse genome (mm9) with Bowtie (Langmead et al, 2009) using the default command options -m 1 --best -strata --solexa1.3-quals, and were assigned to exons of RefSeq genes using custom perl scripts."

(2) Another potential issue is that the new RNA-Seq data in the paper is not strand-specific, so it's hard to determine whether the LE signal actually comes from the sense strand. Was the Cloonan et al. SOLID data mapped and considered in a strand-specific manner? This issue needs to be discussed. There may be other data sets in the literature that are better suited for this analysis.

Thank you for this suggestion; the usefulness of strand-specific analysis had escaped our notice. We analyzed the Cloonan et al data separately for reads mapping in sense and antisense directions with respect to genes. This clearly demonstrates that the antisense reads (i.e. the signal, the reads that map to genes) display a clear shoulder (Fig. S10A). In the sense read data (i.e. the noise, reads that don t map to genes), more than 50 % of genic regions have zero reads, and the distribution is unimodal and shifted by ~ 2 log2 RPKM with respect to the LE distribution.

We have adapted the Material and Methods section and figure S10, and added the following text to the the Results and Discussion section (2nd paragraph on page six):

"Analysis of the strand-specific RNA-sequencing data of Cloonan et al 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 log2 RPKM with respect to the LE distribution (Figure S10A). "

(3) The paper should explain why previous groups have not previously seen the bimodal distribution that is so obvious here.

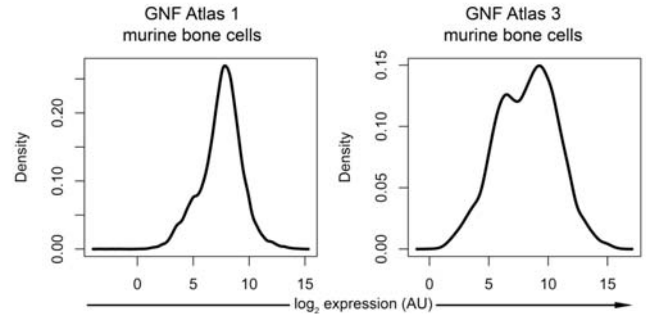
The lack of earlier reports of bimodality in microarray-based studies is due to four major reasons:

i) lowsensitivityofmicroarrays:onlynewerdatasetsshowbimodality. ii) strongly transforming ways of plotting the data iii) the organism the data is from: we only observe bimodality in multicellular as

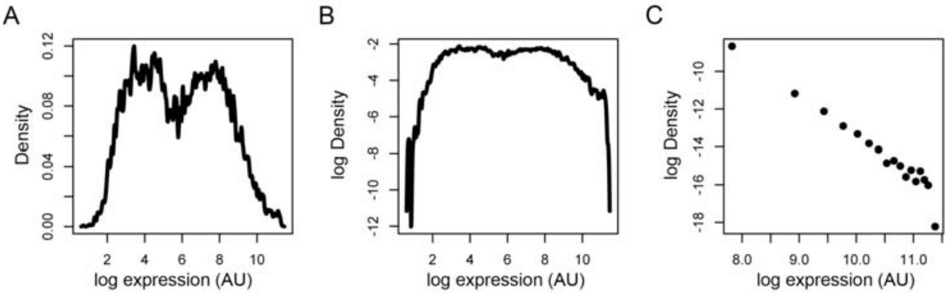
opposed to unicellular organisms. We think this is related to the absence of tissue-specific genes in unicellular organisms such as E. coli and S. cerevisiae, which effectively express most of their genes all the time.

iv) many samples consist of very heterogeneous cell types/tissues rather than a fairly homogeneous population of cells

With respect to point (i) above, we have compared the same tissue in a newer and older version of the GNF Atlas, which is a widely used microarray resource for a large number of different mouse and human tissues (Su et al, PNAS, 2004). As the figure below demonstrates, only the newer GNF Atlas 3 exhibits bimodality (we have included this dataset in our manuscript now).



With respect to point (ii) above, we have processed data for human Cd36+ cells (Cui et al, Cell Stem Cell, 2009) in different ways that are commonly found in literature to illustrate how plotting the data affects the perceived distribution:



A: similar to the kernel density estimates used in our manuscript, i.e. straightforward log transformation of expression levels B: additional log transformation of the y-axis. The dip in the middle of the two peaks has almost disappeared

C: data is binned in linear space, then log transformed on both axes. As a result, the whole left peak up to log expression level ~7.5, which includes part of the right peak, is combined into a single data point (the first data point)

With respect to point (iv) above, we discuss this thoroughly in a manuscript that is related to the current one, and which deals with the effects of cell type mixtures (please see our response to point

(1) of reviewer three). This publication is currently in press at Physical Biology and will appear in June 2011. We have attached the page proofs to this submission.

Please note that the focus of this paper (Hebenstreit & Teichmann, 2011) is to address the consequences of cell type mixing on distributions of gene expression levels from microarray data. In it, we briefly introduce the concept of bimodal distributions and mathematically model these as mixtures of two lognormal distributions in a similar way as we do in the current manuscript. However, this is the full extent of the overlap between the two manuscripts: the purpose of the Physical Biology paper is solely to investigate mixing of cell types by simulations (without presenting new experimental data) based on microarray data and we are not investigating the nature of the bimodality. In contrast, the current manuscript aims to establish and investigate the underlying form of the distribution of gene expression levels in a homogeneous cell population through a variety of experimental and computational analyses. In the current work, we have established that the lower distribution corresponds to "leaky", stochastic transcription, and provide functional and epigenetic interpretations of the LE vs HE groups of genes.

In summary, if we study publications that are based on microarrays and that address expression level distributions (e.g. Kuznetsov, Genetics, 2002, Ueda, PNAS, 2004 or Lu & King, Bioinformatics, 2009), it becomes apparent that in virtually all cases one or more of the aforementioned points apply.

Thus we have now made the following changes (additions underlined, deletions strikethrough) to the Introduction section:

"Microarray and mRNA-sequencing 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; Ramskoeld et al., 2009). There are several reasons for this: many samples are heterogeneous in terms of cell type 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 groups of expression levels. In contrast to these recent reports and more in agreement with the early publications, we present here evidence in support of two overlapping main abundance classes of mRNAs. Here, we provide experimental and computational support for two overlapping major mRNA abundance classes."

We have also changed the title of the manuscript to more precisely describe the organisms for which we observe bimodality:

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

For example, the Ramskoeld 2009 paper that is cited here states in the Abstract that "no support was found for the concept of distinct expression classes of genes". The evidence is found in Figure 2 of that paper and it is a different kind of graph - is this the explanation?

Although Ramskoeld et al find "no support for distinct expression classes of 4

genes", they do not investigate or quantify this in detail. Two figures in this paper are closely related to this issue: Fig. 1A and Fig. S3, which both show RPKM distributions. We have several comments about these plots, as explained in points a-d below:

a) Both figures only show a middle section of the whole density distributions. In fact, the sections shown (from 0.01 to 10 RPKM) would correspond to the LE peak and the ascending HE peak in our distributions. It is therefore hard to judge whether the distributions of Ramskoeld et al contain shoulders or not.

b) According to the legend of Fig. S3, to generate figures 1A and S3, the authors binned expressions of all genes across human tissues. Although it is not clear how the binning was done in detail, it is possible that the binning and averaging across different datasets and tissue types blurs the distinction between the LE and HE expression groups.

c) It is important to determine where the background noise level starts in the distributions. To this end, Ramskoeld et al perform a randomization on intergenic regions to determine the RPKM levels corresponding to background noise (and so do we - please see the response to point 7 below). This led them to select a cutoff of 0.3 RPKM to decide whether a gene is expressed or not. In our RPKM distributions, this cutoff would define roughly half of all LE group genes as not expressed. The authors themselves state that "it is very possible that the background was overestimated." (page 2, right column, line 13). We agree, because our studies have shown that (i) regions of ~ 5 to 10 kb adjacent to genes have to be excluded for determining intergenic background, as these still display increased levels of RPKM (presumbably corresponding to spurious transcription linked to the gene). (ii) gene expression at 0.05 RPKM and lower can still be detected by other means (exon-spanning PCR, RNA- FISH).

d) Ramskoeld et al do not link gene expression levels to ChIP-seq histone modification data. As we show in our manuscript, this is a crucial point in the analysis: the switch from absence to presence of H3K9/14ac marks directly coincides with the intersection of the LE and HE expression peaks (under the premise of clear expression levels distributions - see a) and b)).

Why are the Wang 2008 data (which were used in Ramskoeld 2009) not included in the analysis of other data sets here? This would allow direct comparison with the conclusions of Ramskoeld 2009. If the same type of analysis is applied and the bimodal distribution is not seen, then the findings in the present paper should be qualified to state that the bimodal distribution may be specific to certain cell types. Did Ramskoeld map the reads in a different way?

We have now analyzed samples from all datasets used in the Ramskoeld et al article (Wang et al, Marioni et al, Mudge et al and Mortazavi et al in addition to Cloonan et al which we had included in the previous version of the manuscript). We mapped the raw data of Marioni et al and Mortazavi et al (using the stringent Bowtie suggested by the referee in point (1) above), and we used the mapped data of Wang et al, and calculated RPKMs for these data sets. We further used the precalculated expression levels of Mudge et al. In all cases, whether we mapped the data ourselves or not, we observed LE shoulders (Fig. S9). We further included three additional microarray datasets (Cui et al, Cell Stem Cell, 2009, Lattin et al, Immunome Res, 2008 (GNF Atlas 3) and Chintapalli, Nat Genet, 2007), all of which are bimodal (Fig. S11).

We have now added references to the new datasets (underlined) to the following sentence of the Results and Discussion section (second paragraph on page four):

"Our findings are not limited to Th2 cells and hold for virtually all 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 recent 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.

We have now added the following text to the RNA-seq data processing section of Materials and Methods:

"RNA-seq data from (Mudge, 2008) was downloaded from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/), accession number GSE12297. We used the processed data for 'Cerebellar cortex 40 Control' directly and performed no further calculations except log transformation and kernel density estimation. The RNA- seq data for 'skeletal muscle' from (Wang, 2008) was downloaded from GEO (accession number GSE12946). We used the data that was mapped to the human genome (hg18), assigned it to RefSeq genes, and processed it similarly as described above. We further downloaded RNA-seq data from (Marioni, 2008) from the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra/). The data for human liver tissue was used (accession numbers SRX000571 and SRX000604). The two files were concatenated, mapped to the human genome (hg18) with Bowtie and processed further as described above. Finally, RNA-seq data for mouse brain (Mortazavi, 2008) was downloaded from SRA (accession numbers SRX000350 and SRX001866). As described above, the two files were concatenated, mapped to the mouse genome (mm9) with Bowtie and processed further."

The following text has now been added to the Microarray data section of Materials and Methods:

"We further downloaded microarray data for murine bone cells from the GNF Mouse GeneAtlas V3 (Lattin et al, 2008; GEO, GSE10246) and processed them as described above. Similarly, the processed microarray data for two replicates of human Cd133+ cells (Cui, 2009) was downloaded from GEO, accession number GSE12646 and processed (using Affymetrix build 28 annotations for the Affymetrix U133A chip). Finally, we downloaded from GEO (accession number GSE7763) microarray data for Drosophila eye tissue from the FlyAtlas (Chintapalli, 2007). We mapped the probesets to genes using Affymetrix probe annotations (build 28) for GeneChip Drosophila Genome 2.0 and processed the data the same way as for the other datasets."

(4) On P. 3, it is stated that "Figure S1 shows expression of a marker protein in the cells". I'm actually not sure what Figure S1 is supposed to show. There are four markers, no? The magenta lines aren't explained. The units are arbitrary. What is the purpose of this figure, even?

Thank you for pointing out that this was not clear. The figure displays the expression levels of four proteins, of which two are markers for Th2 cells (Gata3 and Il13) and the two others (Tbx21, Ifng) are markers for another Th cell type (Th1). The figure shows a FACS dot plot, which is very common in immunology and is the standard experimental readout for demonstrating successful T helper cell differentiation (e.g. compare the figures 1b, 3a, 4 and 5 in Tanaka et al, Nat Immunol, 2011). The dots in the figure indicate individual cells that are stained with fluorescently labeled antibodies and have fluorescence intensities that correspond to expression levels of the corresponding proteins in the individual cells. Overlapping dots change color to indicate the overlap (and give an idea of the quantity of the overlap). The fluorescence intensities are on arbitrary scales as the absolute numbers of proteins expressed cannot be determined and the absolute values further depend on the settings of the FACS machine. FACS plots usually reveal separate populations of cells - those that express a protein and those that don t (and have only background fluorescence, i.e. auto-fluorescence). If two proteins are stained for simultaneously (as in our case), four populations of cells can be theoretically expected: double negative, double positive and single negative/positive for either protein. It is therefore common practice to divide the FACS dot plots into four regions based on the visible populations , and to indicate the percentage of cells that are found in the different regions. This is what we indicate by the purple lines and the numbers in the corners of the plots in our manuscript. Due to the ambiguity in manually setting the regions, we are just using this type of experiment as a check for successful Th2 differentiation (a recent example is Tanaka et al, Nat Immunol, 2011).

We have improved the legend of figure S1 by adding the following text to it:

"Gata3 and Il3 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."

We have also changed the main text referring to this figure to mention "two marker proteins" (instead of one).

(5) The explanation for the discrepancy between RNA-seq and microarrays doesn't completely make sense to me and it would be good to describe it in more "layman's terms" - i.e., something that can be understood by the average person with a PhD in molecular biology. The main thing I can't get my head around is why the sampling problems of RNA-seq should matter much, relative to the obvious false-positive problems of microarrays (and even qPCR on occasion). If the gene is really "off" but firing at a low level, is it important what that level is?

Thank you for pointing out that this is not clear to the non-specialist. We agree that the actual level of OFF background expression isn t terribly important, and that the read depth problem is not an issue until stochastic expression becomes predominant at the low end. To us, an obvious issue was that the Th2 cell RNA- seq distribution is shouldered (solid black exon trace in Fig 1A), while the microarrays appear more sharply bimodal (Fig 1B). We investigate this difference by exploring the impact of sampling issues in the RNA-sequencing data. When we model the expression levels of genes with zero reads based on a Poisson process, the RNA-seq distribution becomes more clearly bimodal (we have now made this part of a main figure, Fig 2B, blue trace).

Please see our responses to the next five comments below, where we discuss in detail how we now address this and related issues in the manuscript.

It should still be classified into the low-but-not-off bin. If it has zero reads, where does it end up in Figure 1A and 2D? Is it excluded? Perhaps it would be better to put it into a catch-all bin which represents some practical lower limit of detection?

The two panels in Figure 2A show that there is no practical lower limit of detection; an extrapolation of the linear fit to this figure predicts that all genes have at least one read at a read depth of 23 billion total reads. This prediction, together with the Poisson-based simulation above, supports the idea that roughly 6000 genes with 0 reads are basically part of the LE group of genes rather than representing a separate group.

The referee s suggestion to put the genes into a separate catch-all bin is valuable in the sense that this is a way of visualizing the relative proportions of 0 read/LE/HE genes. We have now done this in Figure S4 in the Supplementary Material. We accomplished this by assigning small RPKM values to these genes. These values are randomly sampled from a Normal distribution (on log2 scale) with mean = -12 and standard-deviation = 1. We did not add the same value to all zero-read genes, as otherwise there would have been a very high spike at a single value, leading to scaling problems (due to the kernel density estimation) with respect to the other data. Now it is straightforward to study the fraction of zero-read genes as everything is on the same scale, and the areas under all curves can be directly compared (except the intergenic zero-reads, which are cut-off due to their height - we prefer to keep them on the same scale for clarity, and also to maintain the standard deviation of the Normal distribution).

We repeated the procedure for zero-read genes in the heatmap results of the ChIP-seq experiment (the RNA-seq zero-read genes were treated as before, the much lower number of ChIP-seq zero-read genes were assigned random log2 RPKM with mean = -3 (and the same standard deviation as above). The zero- read genes now appear as additional blobs in Figure S13. We have further changed the y-axis scaling of the figure - please see the response to point (3) of reviewer 3.

We added the following sentences to the Results and Discussion section (1st paragraph, page 4 and 1st paragraph, page 8, respectively):

"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)."

"Figure S13 is an alternative version of this figure, where we randomly assigned low RPKM values to the zero-read genes."

We further added the following sentences to the "RNA-seq data processing" and "ChIP-seq data analysis" sections, respectively, of the Materials and Methods section (Supplementary Material):

"We prepared alternative versions of Figure 1A and Figure 3D, where we assigned a random log2 RPKM value derived from a Normal distribution with = -12 and = 1 to each gene without sequencing reads (Figure S4 and S13)."

"For the alternative version of Figure 3D (Figure S13), we assigned a random log2 RPKM value derived from a Normal distribution with = -3 and = 1 to each gene without ChIP-seq sequencing reads."

In any case, Figure S7 could be presented in a more straightforward way, or perhaps described in a more straightforward way. Why are there two different axes on Figure S7B (on top and bottom)? (I think the one on top is correct, based on comparison to Figure 1A).

We agree that the description of Figure S7B (now Figure 2B) could be improved. We considered the following points for adding two different x-axes:

a) The Poisson model is based on the integer numbers of reads per DNA stretch (i.e. reads per kilobase, RPK, which is what we used). The RPK scale therefore allows one to directly study the input for the statistical model.

b) The RPK numbers change with the total number of sequencing reads, while RPKM does not. Therefore, the RPK scale makes it easier to compare data from different experiments with differing total read numbers.

c) For the prediction, we binned genes by expression levels and estimated percentages of undetected genes for each bin (it does not make much sense to predict what percentage of a single gene remains undetected). The distribution shown is therefore slightly different from the expression level distribution we show in main Figure 1.

For these reasons, we previously included both the RPK scale, which was used in the actual calculations, and the RPKM scale, which allows comparison with the other figures in our manuscript.

We agree that this is a bit confusing, so we have now put the RPKM scale on bottom and the RPK scale on top of the figure. We further slightly modified the legend of the figure (now Figure 2B) and added the following sentence to the legend:

"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."

I would suggest some sort of model of the actual transcriptome abundance, and then run the Poisson test on that - aren't the results dependent on the actual distribution, which is unknown? Or is that what the splits are supposed to address? I suppose it is, but it would be nice to know for sure.

Our prediction is based on the actual distribution and makes sense for expression levels at which genes are detected. Based on the (Poisson- distributed-) actual numbers of sequencing reads at the various expression levels, we can estimate what fraction of the total expressed genes the detected

genes represent (and thus, which fraction remains undetected).

Our calculations require at least some genes to be expressed at the level we are looking at. It is impossible to predict expression in a range where no reads are detected.

The splits indicate the bins we used for the calculation; please see c) above.

To make our calculations clearer, we have added the following example to the RNA-seq data sensitivity analysis section of Materials and Methods:

"For instance, at RPK = 1 we would expect two sequencing reads for a gene that is 2 kb long and one read for a 1 kb gene (giving the same expression level). Since the actual read numbers vary according to a Poisson distribution, not all genes that are expressed at that level will have exactly one or two reads, respectively, but some will have more and some none at all. The Poisson distribution gives the expected portion of zeros, which would be 37 % for the 1 kb gene and 13.5 % for the 2 kb gene. Thus, if we detect 150 1 kb genes and 250 2 kb genes at RPK = 1, we can estimate that a further 127 (= 150/(1 - 0.37) - 150 + 250/(1 - 0.135) - 250) genes of the same lengths are expressed at the same level but remain undetected."

In any case the black line would be "model" or "actual data", the red line would be "expected detection with 25M reads", and the red line would be labelled as it is already. And for the red line, instead of "Density", it would be good to plot "percent of genes at this level that are undetected".

Thank you for this good suggestion, we have now changed the labeling of the figure to make it clearer.

In any case, it looks like all the problems are for genes that are expressed at less than one copy per cell, and that the likelihood of going undetected is below 5% up until about an order of magnitude lower. This suggests that the read depth problem really isn't an issue unless there are zero transcripts in the vast majority of cells. Something that might deserve discussion.

We entirely agree with this (see first response to point 5, above). We have now tried to clarify our analyses and conclusions of this part of the manuscript by adding text and re-phrasing sentences.

We have added the following text (underlined) to the Results and Discussion section (third paragraph on page five):

"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 [...]"

We further removed the word ‘steep’ in the sentence "This confirms the steep sensitivity drop at the lower end of the LE peak".

We modified the last sentence of the same paragraph and added text to discuss the conclusions (underlined):

"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 log2 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."

In addition, we added the following text to the second to last paragraph of the Results and Discussion section:

"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."

(6) The point above relates to the appearance of Figure 2D, right panel. I suspect that the faintness of the blob in the lower left (relative to the same graph in Figure 2E) may be due to the fact that the transcripts with zero reads are either omitted or more dispersed due to the "Poisson effect", relative to microarray data where they pile up at the background level. (Or, alternatively, the blob in the lower left may be due entirely to cross-hyb, and its modest appearance in the RNA-seq data due to the fact that the bowtie procedure results in less "cross-hyb" than the microarray data - so the real distribution would have a much more dispersed blob in the lower left). If the zero-read genes are assigned to some background level as I suggest above, does the phenomenon become more clear? I would expect the graph to pick up a bright vertical line in the blob in the lower left. Perhaps the intronic reads could be included. Anyhow, the Figure 2D and E right panels are critical to the claims of the paper but I can't tell if their appearance is mainly due to cross-hyb and data processing.

Figure 2D (now Figure 3D): Yes, the fainter blob in the RNA-seq heatmap is due to genes that are not included because no sequencing reads map to them.

As mentioned above in point (5), we have now added the zero read genes in Figure S13 of the Supplementary Material, and they appear as additional blobs We used this approach instead of adding the intron reads, as it would be unclear how to weight the intron reads and/or the scale would not correspond to RPKM anymore. In addition, we now use a different scale on the y-axes of the heatmaps (please see our response to point (3) of reviewer 3).

Please note that the non-zero-read LE genes are very unlikely to be genomic background: Figure 1A shows that this probability is only 10% at a log2 RPKM value of as low as -5. Please also see our responses to point (1) above and point (7) below.

(7) There is a paragraph on P. 5 that begins "We also tested the extent to which genomic DNA may be contamination our polyA-purified mRNA sample...., and it ends with a sentence "....which means that genes with RPKM above this level are at least 90% probable to be expressed". This doesn't make sense to me. The way this is written, it sounds like the intergenic fragments were chosen because they shouldn't be expressed, i.e. they are a negative control. (and if so, why aren't they chosen to be at least 10 kb away from genes, as the materials and methods state? The text here says 5 kb). If they are used to determine the background expression of the genome, then this is a whole different ball game and the paragraph needs to be rewritten and the background/false detection level set in some other (independent) way. The way I would interpret the outcome shown would be that some portion of the polyA-purified sample could in fact be explained by genomic contamination, but it is a very small proportion. Anyhow, this paragraph should be revamped - it would be very interesting to know what proportion of interegenic DNA is detected, of course, and why. But it's not clear what is the real purpose of the analysis here, and what can be claimed from it.

Thank you for pointing out the 5/10kb discrepancy, and that the aims and conclusions of this analysis are unclear (the 5 kb value was an error, which we have now corrected).

This analysis was inspired by Ramskoeld et al (2009, PLoS Comp Biol). (Please see our response to point 3.) In this paper, detection of intergenic DNA was regarded as background noise, an assumption that is also frequently made in designing negative controls for RT-PCR, Ramskoeld et al. used this to determine the false discovery rate of RNA-seq.

Our aim was to show that the LE peak/shoulder does not correspond to this background noise, but instead is actual transcription, even if it is not functional. Therefore we used this readout to demonstrate for our data that intergenic DNA is detected at much lower levels than most LE genes. We agree that it cannot be ruled out that intergenic DNA is transcribed at some rate. This would provide even stronger support that LE genes are being transcribed as well.

To make the purpose of our analysis clearer, we have changed (underlined/strikethrough) the first sentence in the first paragraph on page six of the Results and Discussion section. The sentence now reads:

"We also tested the extent to which genomic DNA maybe contaminating can be detected in our polyA-purified mRNA sample, as proposed by Ramskold et al. as means for quantifiying experimental background."

We further added this sentence to the same paragraph:

"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."

(8) Figure 2C is kind of disturbing - it actually looks like the intercept of the top three points is around 20 mRNAs/cell = 0 RPKM. One wonders whether the genes at the origin simply don't work by in situ, or whether there is some issue with the point at (38, 25) which would straighten things out. I think this issue deserves some commentary, because the calibration between RPKM and copies per cell is likely to be widely cited. As it is, the graph could be used to argue that RNA-seq cannot detect genes expressed at less than 20 copies per cell. (As an aside, is 1 copy ~= 1 RPKM consistent with the estimates one could make given the number of RNAs per cell and their lengths?)

The translation of RPKM into numbers of mRNA depends on the cell type, as different cell types can have vastly different amounts of total mRNA. Based on estimates of RNA mass per cell and spiked-in probes, Mortazavi et al (Nat Methods, 2008) arrive at an estimate of 3 RPKM per transcript per liver cell, which seems reasonably close to our estimate for Th2 cells.

For a truly accurate determination of the RNA-seq detection limit in terms of mRNA numbers from this plot, it would be necessary to include many more RNA- FISH datapoints, say 20-100 more. Since the probeset optimization for this technique is extremely labour-intensive, taking roughly one month per gene, this is unfortunately outside the scope of this paper.

To comment on this, we have added a phrase (underlined) to the following sentence in third paragraph on page 7 of the Results and Discussion section:

"We find that one RPKM corresponds to an average of roughly one transcript per cell in our Th2 data set (Figure 3C). Please not 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."

(9) I am confused by the last four sentences of the paper. The idea is that transcription factors mediate the variation among expressed genes. But don't transcription factors also control which genes are on and off in the different cell types? i.e. the "key decision about whether a gene becomes switched on and expressed" should also be determined by transcription factors that are recruiting the epigenetic/chromatin factors. What else could it be? The chromatin proteins by definition don't have sequence specificity, otherwise we would call them transcription factors.

Thank you for pointing out the unclear phrasing of these sentences. Our intention here was to make it clear that we do not regard a switch-like change from LE to HE as the only mode of gene regulation (and not that on/off switches are accomplished without transcription factors).

We re-phrased the four sentences at the end of the Results and Discussion section:

"Regulation of gene expression is extensively mostly mediated by transcription factor binding events at promoters and enhancers, e.g. (Heintzman et al., 2009). This regulation is extremely important in terms of Often, differential regulation induces only small changes in expression levels, probably serving to fine-tune expression and shifting genes within the HE group ing and is probably responsible for the wide distribution of HE expression levels. 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."

Reviewer #2 (Remarks to the Author):

This manuscript is in the unenviable position of having to provide strong and convincing arguments that the current view is inaccurate. Although the main observation is potentially interesting and significant, I suspect that the observed bimodality in the distribution of gene expression levels (the source of the manuscript's claim) is an artifact of how the analysis was done. Even if this bimodality is real, I am not convinced that this is a general phenomenon - the manuscript only presents two examples, it should be straightforward to find and present similar analyses on microarray data from a much larger range of homogeneous cell populations.

Yes, it is indeed straightforward to find more datasets featuring bimodality. We did not present these results previously for the sake of brevity. We now include many more datasets as described in point (3) of reviewer one above.

On the other hand, the existence of a distinct HE group is surprising. I have looked at a large number of microarray intensity histograms and I have never seen anything like the bimodal distribution in figure 1B. Albeit, most of the time I was looking at data from mixed cell populations. The manuscript presents two sources of evidence that the intensity distribution is bimodal. The first is visual; there are two clear modes in figure 1B. My concern with this picture is that it was generated using kernel density estimation (KDE) and I could find no description anywhere in the manuscript as to how the bandwidth of the kernel was selected (or even what the kernel is, but I suspect it is Gaussian). This is a glaring omission and this is an important piece of information because by varying the bandwidth in KDE, one can generate anywhere from a unimodal distribution to a very multimodal one. To strengthen this part of their argument, the authors need to show that *the bimodality is relatively insensitive to the choice of the bandwidth and furthermore that it appears when using other techniques are used to estimate the distribution (i.e. a histogram).*

We agree that the kernel density estimation was not described in much detail before. We perform the density estimations with the function density or kde2d (in the 2D case) of the statistical software R using default settings. To describe this in more detail, we have now added the following paragraph to the Materials and Methods:

"Kernel density estimation

Gene expression distributions were displayed as kernel density estimates in most cases. These were calculated using the function ‘density()’ of the freely available statistical software package ‘R’ (http://www.r-project.org/). We used default settings of this function unless stated otherwise. This means a Gaussian kernel and that the bandwidth equals 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power (corresponding to Silverman’s "rule of thumb", (Silverman 1986, page 48, eqn (3.31)) unless the quartiles coincide when a positive result will be guaranteed (R manual). For 2D kernel density estimations we used the function ‘kde2d()’ of the R library ‘MASS’ with the default bandwidth and a Gaussian kernel. This bandwidth is calculated based on a variation of above formula for the 1D case, where the factor 1.06 instead of 0.9 is used. Densities were estimated at 50 grid points in either direction and displayed as heatmaps."

We further varied the kernel bandwidth settings and explored histogram representations of both our RNA-seq data and the Wei et al microarray data and its alternative normalizations. We have added these as new figures S3, S6 and S7. These clearly demonstrate that the default kernel density estimates represent the densities very well, and different bandwidths do not affect the general structure of the data (and neither do the different microarray normalizations). Likewise, the histogram representations conserve the bimodal/shouldered appearance across a wide range of different bin-sizes. The distinction between LE and HE only starts to disappear when the bandwidth or the bin-size become large enough to bridge the dip.

To address the new figures, we have added the following text (underlined) to the Results and Discussion section (top paragraph, page four):

"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 log2 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 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 r2 = 0.84, Spearman r = 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)."

The manuscript also provides a Gaussian mixture model analysis to support the bimodality. However, this analysis actually points to more than two modes; in figure S5, the maximum of the BIC-adjusted likelihood occurs at four, not two modes. The fact that there is a large increase in BIC-adjusted likelihood going from one to two modes is not surprising, figure 1A is clearly non-Gaussian, but it does not support the existence of exactly two modes, as required to support the manuscript's premise.

We apologize for this misunderstanding, but we are not claiming that only two modes exist, only that there are two major modes, as mentioned in the title of the paper.

We have now clarified this in the Results and Discussion section (2nd paragraph, page four; additions underlined, deletions strikethrough):

"Quantifying this by curve fitting confirms a best good fit to two distributions: the goodness-of-fit [...] shows strong increases or maximal values for both microarray and RNA-seq data when two-component models are fit by expectation-maximization (compared to single- or more-component models)"

The main point of our manuscript is that the distribution is non-Gaussian and can be better described by two sub-distributions. This is, as the reviewer notes her/himself, an unenviable position as it implies that the current view is inaccurate . Whether further minor sub-populations exist is an interesting question that is not ruled out by our work, but is difficult to answer with the currently available data sets for the following reasons:

a) Studying the four microarray and six RNA-seq datasets we now present in the revised manuscript reveals that all microarray distributions are bimodal and all RNA-seq distributions have shoulders on the left sides. However, the shapes of the curves vary and depend on technical issues, such as microarray normalization strategies (which all conserve the bimodality though).

b) The goodness-of-fit criteria indicate much greater improvements upon extension from single- to two-component models compared to higher- order extensions in all data sets. The best fitting model varies from two- to nine-components depending on the goodness-of-fit criterion used (see below), and is also affected by different microarray normalization strategies.

Finally, we have added the following sentence to the Results and Discussion section (2nd paragraph, page four):

"The existence of further, minor groups of genes cannot be excluded, but is unclear at this point due to the diverse curve-fitting results for the different datasets if higher-order (more than two components) models are considered."

Note also that BIC is simply one of a large number of different model selection methods, and the authors need to justify this choice and/or show that the same observations hold if they use AIC, a likelihood ratio test, or the Bayes factor.

We chose the BIC criterion as it gives a rough approximation to the logarithm of the Bayes factor, is easy to use, and does not require evaluation of prior distributions (Kass RE & Raftery AE, Journal of the American Statistical Association, 1995).

We have now added AIC plots for all model fits and have performed likelihood ratio tests (LRT) between each model and the next more complex one (Figures S8 to S11).

The AIC results are similar to the BIC results (indicating the strongest improvement upon extension from one- to two-component models), apart from the fact that higher-order models are preferred (the more components the model has, the more the AIC diverges from BIC). This is not surprising as it is known that the AIC tends to overestimate the number of parameters needed (Kass RE & Raftery AE, Journal of the American Statistical Association, 1995). Again, the model that yields the best fit according to AIC depends on the dataset studied.

We further performed LRT tests for all models and the next more-complex ones, i.e. the model with n components as the null model and the one with n+1 components as the alternative model ($0 < n < 9$). We approximated the test statistics with 2 distributions and calculated the p-values with R. The results of these calculations lead to similar conclusions as those for other goodness-of-fit criteria. The R-generated p-values are lowest for the switch from one- to two- component models (and non-zero in only two cases). For more-component models, the transitions with low or high p-values vary from one dataset to another.

We have added the following text to the ‘Curve fitting’ section of the Materials and Methods to describe the new calculations:

"The log likelihood values output by Mclust were used to calculate AIC (Akaike, 1974), BIC (Schwarz, 1978) and likelihood ratio statistics (Casella, 2001). The latter were calculated for the model with n components as the null model and the one with n+1 components as the alternative model ($0 < n < 9$). We approximated the test statistics with 2 distributions and calculated the p-values with R."

Reviewer #3 (Remarks to the Author):

RNA sequencing reveals two major classes of gene expression levels Hebenstreit et al.

MAJOR COMMENTS (in no particular order)

1. I think the justification for there not being a mixture of two cell types provided in the third paragraph on page 4 leaves something to be desired.

We agree that the question of cell type mixtures is very important. The paragraph in question was not meant to provide a proof that there can t be a mixture of cell types. Instead we wanted to point out the (admittedly trivial) fact that the two peaks cannot be the consequence of cell type mixture. We have slightly rephrased this paragraph now (new text underlined, old strikethrough):

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

In particular, I think the argument relies upon the somewhat pathological assumption that if multiple cell types were present that they would be so in the same ratio (i.e., 50% cell type 1, 50% cell type 2). In the far more likely case that you have an overwhelming amount of cell type 1 (80% say) and a much smaller proportion of cell type 2 (20%) then I think the argument is much harder to make.

No, the argument does not rely on the assumption of a 1:1 ratio of cell types. It should be intuitively clear that any mixture of cell types results in an averaging of gene expression levels. In other words, genes with low expression would tend to have higher values, and genes with high expression lower values, so that the dip between the LE and HE groups disappears.

We have actually studied this rigorously using mathematical modeling based on microarray data in a paper in press at the journal Physical Biology and have attached the page proofs. The main findings are (i) that a bimodal distribution is conserved over a large parameter space with moderate contamination of one cell type by another, and (ii) that more extensive mixing of cell types involves a reduction in the bimodality.

Please refer to our response (3) to Referee 1 above with respect to the comparison between this paper and the present work.

Also, I would move this paragraph to later on in the manuscript when the single- cell RNA-FISH data are discussed.

This is a good suggestion; we have now moved the paragraph to follow the RNA- FISH paragraph.

In a similar vein, I am not sure about the argument on page 6 that the cell population you analyzed has no Th1 cells present. In particular, Zheng and Flavell (Cell, 1997) state: "Gata3 is necessary and sufficient for Th2 cytokine gene expression in CD4 T Cells". How do you reconcile this with the small number of cells that do not express Gata3, or that express it at a very low level? Do you think this is a limitation of the RNA-FISH data? Or is it a sign of low-level contamination? I think this possibility cannot be excluded on the basis of the data presented and this should be reflected in the manuscript.

Gata3, Il13 and Il4 are conventionally considered Th2 cell markers, and our FACS data in Figure S1 show that we have an 87% pure cell population according to Gata3 and 77% according to Il13. The union of both markers would boost the percent purity further. Please refer to our reply to point (4) by referee one above for further details on the FACS profiling.

Since Th2 cells are related to Th1 cells, we provide FACS data for the two Th1 cell markers Tbx21 and Ifng (Figure S1). This shows a 1% subpopulation of cells expressing Ifng but not Il13, suggesting at most a miniscule Th1 contamination. The absence of Th1 cells is also supported by the Gata3 vs Tbx21 RNA-FISH data (Figure 3B).

Even if there were contaminating cell types in our Th2 population, this would be an argument in favor of our findings, as the concept of two expression levels would thus be robust even under these conditions. It is worth emphasizing again that the two expression levels cannot be a consequence of a mixing of cell types, as this inevitably leads to averaging of gene expression levels. Please also refer to our answer to point (1) above.

2. Did you separate the intronic regions depending upon their vicinity to the 3' end of the gene? Given the relative speed of the splicing machinery and the DNA polymerase, I would expect to see more reads towards the 3' end of each gene. It would be interesting to perform this analysis and to determine whether information about the rate at which splicing occurs could also be modeled.

We have carried out this analysis now and have added the results as Figure S12. We do not see an increase of reads towards the 3 ends.

We have added the following sentence to the Results and Discussion section (third paragraph on page six):

"(Please note that our intronic read densities are not enriched at 5’ or 3’ ends of the intronic regions (Figure S12).)"

Further, we have added the following text to the RNA-seq data processing section of Materials and Methods:

"To test for a possible RPKM bias in 5’ or 3’ ends of intronic regions, the introns of each gene were lined up. If the intronic region was at least 6 kb in total, RPKM were separately determined for the most 5’ 2 kb, for the 2 kb in the center and for the most 3’ 2 kb."

3. In Figure 2D and 2E (right hand panel) there does seem to be a positive correlation between H3K9/14ac and expression across all expression ranges. I know that the authors state that the signal for lowly expressed genes is consistent with background, but the relation is rather striking (at least to this reviewer). Hence, I think the assertion that "H3K9/14ac marks are associated with promoters of highly expressed genes" should perhaps be toned down.

We have now calculated the correlation coefficients for expression level versus histone modification within the LE and HE groups of genes in the RNA-seq and microarray data sets (right panels in (now) Figs 3D and E). The highest correlation is within the HE genes in the RNA-seq dataset, which has an r2 of 0.097 in linear space and 0.29 in log space.

We feel that these weak correlations within the LE and HE groups pale in comparison to the over 4-fold increase in histone modification level from LE to HE groups, but we have now added the following sentence to the Results and Discussion section (second paragraph on page 8):

"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 r2 = 0.29 in log space and r2 = 0.097 on linear space."

Alternatively, the cutoffs/methods for ChIP-seq analysis need to be justified better. In particular, it was not clear to me that the ChIP-seq data were normalized - therefore, that the cutoff between signal and background was the same for all four ChIPs seemed somewhat surprising to me. This could do with some more explanation I think

Previously, the ChIP-seq data was not normalized, but we have now altered this and use our EpiChIP software package to determine the ChIP-seq cutoff. The EpiChIP software is available at epichip.sourcefourge.net and the paper is now available online (Hebenstreit et al, Nucleic Acids Res, in press).

We have added the following phrase (underlined) to the main text (1st paragraph, page 8):

"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)."

In addition, we added the following phrase (underlined) and sentence, respectively, to the ChIP-seq data analysis paragraph of the Materials and Methods section:

"The resulting window of -400 to +807 bp at transcriptional start sites was used to quantify the ChIP-seq signal for each gene (the area below the peaks within this window) which was normalized by the total (genomewide) area to yield the "normalized locus specific chromatin signal" (NLCS) (Hebenstreit, Nucleic Acids Res, 2010)."

"The threshold separating background from signal was determined with the curve fitting function of EpiChIP."

4. I found the method used for calculating the expected number of reads for each gene somewhat strange. In particular, the use of RPK in this instance instead of the RPKM seemed clumsy. Why did you just not include gene-length and the total number of reads in the lane as offsets in the Poisson model? This would allow all of the analysis to be done on the RPKM scale and would offer much easier interpretation I think.

Please see our response to point (5) of reviewer one.

MINOR COMMENTS

1. On page 4 it is stated that the two-component model shows "strong increases or maximal values for both microarray and RNA-seq data". In Figure S5, whilst there is clearly strong evidence for a two-component model as opposed to a uni- modal distribution, in no case is the two-component model the best fitting as adjudged by the BIC. The text should be adjusted to reflect this.

We have deleted the phrase "or maximal values" from the sentence.

2. On page 6, "non-Poissonian distributions (which would have...." should be replaced with "that they had extra-Poisson variation (a Poisson random variable would have..."

We have changed the sentence as suggested.

3. On page 7, you should quantify the statement in the second paragraph that "...many genes that are characterized as not expressed and non-functional in Th2 cells...". What proportion of genes is this?

We were referring to the genes in Table S1. Of these, five have been characterized in the literature as being not expressed in Th2 cells in the strictest sense; if we included the other genes which are expressed at LE levels according to our RNA-seq/PCR data, the number would be nine.

We replaced "many" in the quoted sentence with "at least five"

4. In the supplement, on page 3, third line you should refer to Table S2, not Table S4.

Thank you, we have corrected this error.

5. I found Figure S6C difficult to interpret - can you remove the black rectangles that surround the whole of each bar please - they add nothing to the figure and only ended up confusing this reviewer!

Thank you, we agree that this is confusing. We have removed the black rectangles.

2nd Editorial Decision 07 April 2011

Thank you again for submitting your revised work to Molecular Systems Biology. We have now heard back from the two referees who agreed to evaluate this study. As you will see, the referees were largely satisfied with the revisions made to this work, and were generally supportive. The first reviewer, however, has a series of small remaining concerns, and make suggestions for modifications, which we would ask you to carefully address in a revision of the present work.

In addition to the points raised by Reviewer #1, the editor would like to ask you address the following data and format issues:

1. It will be important the new RNA-seq data described here should be deposited in a public database. With the closure of SRA, GEO appears to be the best public repository for this type of data.

2. In addition to our capacity to host datasets in our supplementary information section, we allow readers to directly download the 'source data' associated with selected figure panels (e.g. <http://tinyurl.com/365zpej>). This sort of figure-associated data may be particularly appropriate for the scatter plot data in Fig. 2D and 3B. Guidelines have been pasted at the end of this email.

3. The current manuscript is somewhat longer than our standard Report format. We will be able to accommodate this length, but please be sure that the manuscript does not become longer during revision, and consider streamlining the text wherever possible. In addition, some or all of the the "Materials and Methods" section should be moved from the Supplementary information to the main manuscript. This section does not count toward the Report length restrictions.

4. Please provide a standfirst text, bullet points, and thumbnail image (more below).

Yours sincerely,

Editor

Molecular Systems Biology

REFEREE REPORTS

Reviewer #1 (Remarks to the Author):

The revised manuscript is substantially clearer, and includes additional analyses that address virtually all of the points raised by all the reviewers. I believe that it is suitable for publication with minor corrections and perhaps redrawing a couple of the figures to make them easier to see. Since it will likely be widely cited, and followed up, the authors may wish to make other minor modifications to address the points below. I do not feel that the paper requires re-review, but since it now has the potential to be used in arguments about pervasive transcription, antisense, etc, I do feel that it would be worth making a few tweaks now in order to avoid confusion and misinterpretation.

Specific points:

(1) Since the title now says "metazoan", the Abstract should explain why. E.g. add a sentence at the end that says "Similar observations are made in other data sets, including drosophila", or something to this extent. Same for the Introduction. And for the same reason, the Results and Discussion should probably start by saying "We initially based our analysis...." Instead of "we based our analysis....".

(2) P. 6, line 6. "...while the majority of non-zero fragments peaks slightly left of the LE shoulder...The 90% quantile of this intergenic background distribution is found at -4.97 log2 RPKM, which means that we can be quite confident that genes with an RPKM value above this level are truly expressed...." This conclusion seems remarkably unsophisticated and unconvincing in comparison to the rest of the manuscript. Is there some more quantitative degree of confidence that could be assigned? Or am I missing something here? I would think that a more effective quantitative demonstration would be to add the intergenic data to Figure 1C and 2B, and show that the LE is different from intergenic.

(3) At some point it might be worth pointing out in the main text that the samples analyzed are poly-A purified, so all conclusions bear this qualification. The number of intergenic transcripts are almost certainly underestimated here, and it would indeed be a reasonable argument that the elevation of LE genes above intergenic background is because they contain poly-A signals, not because they are really expressed higher. If the authors can argue against this, the argument might as well be raised now.

(4) P. 6, line 15-16. Instead of "(i.e. the noise, reads that don't map to genes)" I think it should be "(i.e. corresponding to antisense transcripts)". Unless I completely misunderstand. I do think it is worth comparing to non-genic reads, however. There is a lot of interest in antisense transcription. It would be worth clarifying how the distribution of the sense reads (i.e. antisense transcripts) in the Cloonan dataset relate to the intergenic background.

(5) It's very hard to see what's going on in Figure 2D. It would be better to show this one in a heat-map as in Figures 3D and E, using a line to show the boundary between LE and HE.

(6) Since there is a lot of interest in low-copy "pervasive" transcription, it would be illuminating to see Figures 3A, B, and C on a log scale (perhaps indicating zero on the left/bottom) in order to ask how accurate RNA-seq makes measurements at low copy numbers. I believe the LE genes would be expressed at 1 copy per 30 cells, or something like that, which I believe the paper claims is detectable by FISH (and shows are detectable by PCR in Figure 2C, although I am also concerned that random primers might also give signal after ~30 cycles - is this the case?).

(7) Several references are missing information, including the two Hebenstreit papers, Heintzman, and Wang.

(8) Figure S4, please state what the height of the peak is for zero-reads intergenic.

Reviewer #2 (Remarks to the Author):

They have addressed my concerns regarding their data analysis. I still find the result hard to believe but I guess that makes this a particularly interesting manuscript.

1st Revision - authors' response 18 April 2011

We are pleased that the reviewers were happy with our revision and would hereby like to resubmit our manuscript (MSB-11-2769) entitled ‘RNA sequencing reveals two major classes of gene expression levels in metazoan cells’ for publication as a ‘Report’ in Molecular Systems Biology.

In response to your four points to take into account for resubmission, please find our responses below:

1. It will be important the new RNA-seq data described here should be deposited in a public database. With the closure of SRA, GEO appears to be the best public repository for this type of data.

We have submitted our datasets to gene expression omnibus (GEO) and the accession number is: GSE28666. We mention this in the main text now as well.

2. In addition to our capacity to host datasets in our supplementary information section, we allow readers to directly download the 'source data' associated with selected figure panels (e.g. <http://tinyurl.com/365zpej>). This sort of figure-associated data may be particularly appropriate for the scatter plot data in Fig. 2D and 3B. Guidelines have been pasted at the end of this email.

As you will see from the online submission site, we have provided source data for the following figures:

Fig 1A (exonic regions)

Fig 1B

Fig 2C

Fig 2D

Fig 3AB

Fig 3C

3. The current manuscript is somewhat longer than our standard Report format. We will be able to accommodate this length, but please be sure that the manuscript does not become longer during revision, and consider streamlining the text wherever possible.

In addition, some or all of the the "Materials and Methods" section should be moved from the Supplementary information to the main manuscript. This section does not count toward the Report length restrictions.

We have now moved the Materials and Methods to the main manuscript. In an effort to shorten the manuscript, we now present Figure 2 in a more compact way.

In response to the first reviewer's additional comments, we performed one additional analysis, added two more supplementary figures, and modified some figures to make them clearer (see point-by-point response below).

Thank you very much for considering our manuscript, we look forward to hearing from you.

Reviewer #1 (Remarks to the Author):

(1) Since the title now says "metazoan", the Abstract should explain why. E.g. add a sentence at the end that says "Similar observations are made in other data sets, including drosophila", or something to this extent. Same for the Introduction. And for the same reason, the Results and Discussion should probably start by saying "We initially based our analysis...." Instead of "we based our analysis....".

Thank you for these helpful suggestions. We have added the following sentence to the end of the abstract: "These observations are confirmed in many other microarray and RNA-sequencing datasets of metazoan cell types."

We have added the following sentence to the end of the Introduction:

"Our findings hold for metazoan datasets including human, mouse and Drosophila sources." We inserted initially (underlined) into the first sentence of the Results and Discussion section: "We initially 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."

(2) P. 6, line 6. "...while the majority of non-zero fragments peaks slightly left of the LE shoulder...The 90% quantile of this intergenic background distribution is found at -4.97 log2 RPKM, which means that we can be quite confident that genes with an RPKM value above this level are truly expressed...." This conclusion seems remarkably unsophisticated and unconvincing in comparison to the rest of the manuscript. Is there some more quantitative degree of confidence that could be assigned? Or am I missing something here? I would think that a more effective quantitative demonstration would be to add the intergenic data to Figure 1C and 2B, and show that the LE is different from intergenic.

The 90% quantile means that the probability is 90% for genes above this PKM to be truly expressed, which is a quantitative degree of confidence. To make this clearer, we modified the relevant sentence (page 6, line 8) by adding as follows:

"The 90 % quantile of this intergenic background distribution is at -4.97 log2 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."

Instead of adding the intergenic data to Figs 1B and 2C as suggested by the referee, we prefer to add a new figure (Fig. S12) with normalized LE and intergenic distributions. We feel this is more appropriate, because the total areas under the LE and intergenic distributions are very different (the LE is smaller than the intergenic). Note again the high proportion of zero reads in the intergenic distribution, which are not displayed on the log scale. To introduce and discuss the new figure, we have added the following sentence to line 10, page 6 of the main text: "Further, the overlap between the intergenic and the normalized LE fit is small (Fig. S12)."

(3) At some point it might be worth pointing out in the main text that the samples analyzed are poly-A purified, so all conclusions bear this qualification. The number of intergenic transcripts are almost certainly underestimated here, and it would indeed be a reasonable argument that the elevation of LE genes above intergenic background is because they contain poly-A signals, not because they are really expressed higher. If the authors can argue against this, the argument might as well be raised now.

We forgot to mention this, thank you for the suggestion. We added the underlined word to the following sentence (page 3, 4th line from bottom):

"We generated Th2 poly(A)+ RNA-seq data for two biological replicates and calculated gene expression levels using the standard measure of Reads Per Kilobase per Million (RPKM). (Table S2 provides the numbers of reads and mappings data)."

Our key point about the existence of two main expression classes, LE and HE, only concerns annotated genes and mRNA, and is independent of the question of intergenic transcription. To address the question of intergenic and antisense transcription within the same framework as mRNA abundance levels would require total RNA-seq rather than poly-A purified mRNA-seq.

(4) P. 6, line 15-16. Instead of "(i.e. the noise, reads that don't map to genes)" I think it should be "(i.e. corresponding to antisense transcripts)". Unless I completely misunderstand. I do think it is worth comparing to non-genic reads, however. There is a lot of interest in antisense transcription. It would be worth clarifying how the distribution of the sense reads (i.e. antisense transcripts) in the Cloonan dataset relate to the intergenic background.

We have added the term as suggested. The sentence (page 6, line 17) reads now: "In the distribution of ‘sense’ reads (corresponding to antisense transcripts in genic regions), more than 50 % of genic regions have zero reads."

The suggestion to compare sense reads (i.e. antisense transcripts) to the intergenic background is excellent, and we have now added this to Fig S10. This plot shows that the intergenic data and sense data overlap almost perfectly – if anything, the intergenic regions are more highly expressed than the sense regions.

We have now modified the section describing our data analysis of the Cloonan et al. dataset to read as follows (additions underlined, deletions strikethrough):

"Analysis of the strand-specific mRNA-sequencing data of ES cells of Cloonan et al (Cloonan et al, 2008) yields similar conclusions. The polyA-purification protocol selects for reads antisense to genes (the antisense reads correspond to mRNA). In the distribution of ‘sense’ reads (corresponding to antisense transcripts in genic regions), more than 50 % of genic regions have zero reads. This noise distribution is unimodal and shifted by ~ 2 log2 RPKM with respect to the LE distribution, and overlaps almost perfectly with the distribution of reads in intergenic regions (Figure S10A)."

We have also slightly modified the Methods section and the figure legend to take account of the additional intergenic data.

(5) It's very hard to see what's going on in Figure 2D. It would be better to show this one in a heat-map as in Figures 3D and E, using a line to show the boundary between LE and HE.

We have modified the figure as suggested.

(6) Since there is a lot of interest in low-copy "pervasive" transcription, it would be illuminating to see Figures 3A, B, and C on a log scale (perhaps indicating zero on the left/bottom) in order to ask how accurate RNA-seq makes measurements at low copy numbers. I believe the LE genes would be expressed at 1 copy per 30 cells, or something like that, which I believe the paper claims is detectable by FISH (and shows are detectable by PCR in Figure 2C, although I am also concerned that random primers might also give signal after ~30 cycles - is this the case?).

We have included as new Figure S14 log transformed versions of Fig 3A, B and C.

To introduce the new figure, we have added the following sentence to the 6th line from the bottom of page 7 of the main text: "See Figure S14 for log transformed versions of Figure 3A-C."

We are using RNA-FISH as a rough estimate of the RNA-seq/transcript numbers correspondence only, as we also state in the main text ("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."). For a thorough analysis of the low end RNAseq accuracy, it will be necessary to perform many more RNA-FISH experiments. Please refer to our response to point (8) of referee one in our point-by-point reply in the previous revision.

To clarify the issue about genomic DNA contamination in the PCR amplification: all ten LE gene PCR primers are exon-spanning and six of these map to splice junctions. We have now added the information on splice junction spanning of the primers to Table S1. Furthermore, it is worth noting that the threshold cycles are all below 30 (Figure 2C).

(7) Several references are missing information, including the two Hebenstreit papers, Heintzman, and Wang.

Thank you, we have corrected this error.

(8) Figure S4, please state what the height of the peak is for zero-reads intergenic.

We have modified the figure so that the intergenic peak is included.