nature biotechnology

Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation

Dimos Gaidatzis[1,2,4], Lukas Burger[1,2,4], Maria Florescu[1–3] & Michael B Stadler[1,2,4]

RNA-seq experiments generate reads derived not only from mature RNA transcripts but also from pre-mRNA. Here we present a computational approach called exon-intron split analysis (EISA) that measures changes in mature RNA and pre-mRNA reads across different experimental conditions to quantify transcriptional and post-transcriptional regulation of gene expression. We apply EISA to 17 diverse data sets to show that most intronic reads arise from nuclear RNA and changes in intronic read counts accurately predict changes in transcriptional activity. Furthermore, changes in post-transcriptional regulation can be predicted from differences between exonic and intronic changes. EISA reveals both transcriptional and post-transcriptional contributions to expression changes, increasing the amount of information that can be gained from RNA-seq data sets.

Cellular RNAs are regulated at multiple stages, including transcription, RNA maturation and degradation. Several analytic methods have been developed to measure these processes on a transcriptome-wide scale. For example, global run-on sequencing (GRO-seq)[1] uses incorporation of a nucleotide analog to enrich for nascent RNA. In Nascent-seq[2,3], newly transcribed RNAs are isolated by purification of their complex with proteins and the DNA template. Cellular fractionation techniques[4] have also been adapted to measure nascent transcripts, which are enriched in the nucleus. mRNA half-lives have been determined, for example, by blockage of transcription followed by transcriptional profiling[5]. RNA sequencing, the most widely used method for transcriptome analysis, has been applied in numerous studies to determine steady-state mRNA levels[6,7] and alternative splicing events[8] and to identify previously unknown transcripts and noncoding RNAs[9–11]. In general, these protocols aim to enrich for mature mRNA by selection of polyadenylated RNA or by depletion of ribosomal RNA.

Many computational methods (reviewed in refs. 12,13) have been developed for the analysis of RNA-seq data, to enable spliced alignment[14,15], transcript assembly[16,17], transcript quantification[14,18] and differential expression analysis[19–21]. Although RNA-seq mostly generates reads that map to exons, it also captures less abundant intronic sequences[6]. However, their interpretation has remained controversial. Some have suggested that they originate from DNA contamination and can thus be used as a quality metric for RNA-seq data[22] (see also RNA-seq guidelines of the Roadmap Epigenomics Consortium, http://www.roadmapepigenomics.org/), whereas others have hypothesized that they stem from unknown exons or intronic enhancers[6,7]. In a study based on exon arrays, probes mapping to introns were used to investigate pre-mRNA dynamics[23]. Three recent studies based on RNA-seq provided evidence that intronic reads might correlate with transcriptional activity. In two of these, the read coverage along introns was related to nascent transcription in combination with co-transcriptional splicing events[24] and later was used to fit a detailed transcriptional model within a single sample[25]. More recently, levels of exonic reads were found to lag 15 min behind levels of intronic reads for a set of oscillating transcripts during *Caenorhabditis elegans* development, suggesting that intronic levels are a proxy for nascent transcription[26].

Here we analyze 17 published RNA-seq data sets covering a wide range of cell types and perturbations. We find that changes in intronic read counts are not technical artifacts but instead directly measure changes in transcriptional activity. Using EISA to compare intronic and exonic changes across different experimental conditions allows the separation of transcriptional and post-transcriptional contributions to observed changes in steady-state RNA levels, without the need for additional experiments such as GRO-seq[1] or Nascent-seq[2,3]. This approach opens up the possibility of using standard RNA-seq experiments to determine whether expression changes observed in gene knockout, knock-down or overexpression experiments are caused by transcriptional or post-transcriptional mechanisms, thereby providing information pertinent to gene function.

RESULTS

Quantifying expression changes in exons and introns

Given that exonic reads essentially reflect mature cytoplasmic mRNAs, we determined whether *in silico* separation of exonic and intronic reads from a standard RNA-seq experiment could be used to separate transcriptional and post-transcriptional contributions to expression changes. In our approach, expression changes across conditions were quantified separately for exonic (Δexon) and intronic

[1]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. [2]Swiss Institute of Bioinformatics, Basel, Switzerland. [3]University of Basel, Basel, Switzerland. [4]These authors contributed equally to this work. Correspondence should be addressed to D.G. (d.gaidatzis@fmi.ch) or L.B. (lukas.burger@fmi.ch) or M.B.S. (michael.stadler@fmi.ch).

Received 10 April 2014; accepted 22 May 2015; published online 22 June 2015; doi:10.1038/nbt.3269

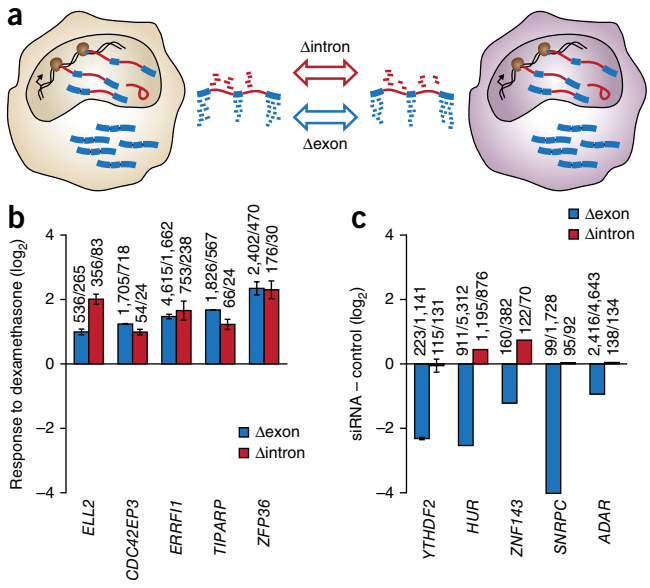

Figure 1 Comparison of exonic and intronic changes for single genes under controlled transcriptional and post-transcriptional perturbations. (**a**) Illustration of the computational approach. mRNA is displayed in the form of nascent unspliced or partially spliced transcripts (nucleus) and mature mRNAs (cytoplasm). RNA-seq reads that map to annotated transcripts are separated into exonic (blue) and intronic reads (red). The differences in exonic (Δexon) and intronic (Δintron) read counts between experimental conditions are then quantified and compared to each other. (**b**) Δexon and Δintron after treatment of human A549 cells with the glucocorticoid dexamethasone. Five target genes of the glucocorticoid receptor known to be upregulated in a receptor-dependent manner[27] are shown. Numbers indicate the exonic and intronic read counts with and without treatment. (**c**) Same as in **b** for siRNA-target genes in five different knock-down experiments.

reads (Δintron) and the relationship between Δexon and Δintron was investigated (**Fig. 1a**). To test the feasibility of the approach, we monitored the dynamics of intron and exon levels of single genes under well-characterized transcriptional and post-transcriptional perturbations. We first calculated expression changes of glucocorticoid receptor target genes upon stimulation of human A549 cells with glucocorticoid dexamethasone[27]. From a set of genes shown previously to be transcriptionally induced in a receptor-dependent manner[27], we selected a subset of five genes and determined their RNA levels by counting either reads mapping to exons (which is the common approach) or only reads mapping to introns. As expected, the selected genes were upregulated on the exonic level upon stimulation (**Fig. 1b**). Notably, a similar upregulation was observed on the intronic level, suggesting that the intronic reads may provide information about the transcriptional changes that occur upon treatment. However, intronic changes do not always mirror exonic changes. We performed the same analysis for five genes knocked down in different short interfering RNA (siRNA) experiments (**Supplementary Table 1**). In all cases, we observed exonic downregulation of the targeted genes whereas intronic levels remained virtually unchanged (**Fig. 1c**). The correlated intronic and exonic changes during direct transcriptional activation (**Fig. 1b**) and the lack of intronic changes upon post-transcriptional perturbations (**Fig. 1c**) indicate that a comparison of exonic and intronic expression changes can separate transcriptional and post-transcriptional effects. Performing such an analysis on a genome-wide scale poses challenges concerning intron annotation and read coverage. We conservatively define the intronic part of a gene as all nucleotides in the gene body that do not overlap an exon from any known transcript isoform. In addition, we consider only genes that have sufficient exonic and intronic coverage and do not overlap with other genes (**Supplementary Data 1**). To account for differences in the exonic/intronic ratio between different samples, we performed library-size normalization for exons and introns separately (Online Methods). Although it has been shown that the intronic yield is decreased in polyA RNA compared to total RNA[24], we also observed many polyA-enriched data sets with sufficient intronic coverage for quantification (see below). We thus considered both total RNA and polyA-RNA data sets for further analysis.

Intronic changes measure changes in transcription

To test whether changes in intronic reads can be used to measure changes in transcription on a genome-wide scale, we reanalyzed data from three experiments in which transcription (using Nascent-seq or GRO-seq) as well as mRNA expression had been profiled simultaneously.

In one system, mouse bone marrow–derived macrophages were stimulated with lipid A, and RNA was sequenced from fractionated (chromatin, cytoplasm) and unfractionated cells (polyA RNA) after 0, 30, 60 and 120 min[28] (**Supplementary Data 2**). **Figure 2a** shows the time evolution of expression levels of exons and introns in unfractionated RNA-seq as well as the chromatin and cytoplasmic fractions separately for all originally identified clusters of lipid-A-induced transcripts[28]. We hypothesized that if intronic reads in standard unfractionated RNA reflect changes in transcription, the intronic expression patterns should more closely follow the chromatin fraction, which represents nascent RNA[2,3], than the cytoplasmic fraction. This is indeed the case when considering all genes in all clusters ($P = 1.03 \times 10^{-9}$, one-sided paired *t*-test). Considering each cluster separately, we obtained significant *P*-values (<0.02) for the first four clusters A1, A2, B and C. As the clusters D, E and F contained genes that change less abruptly over time, we speculate that in these cases it is more difficult to detect differences between transcription and mRNA levels in the cytoplasm. As expected, the exonic changes of the unfractionated pool closely followed the cytoplasmic fraction (**Fig. 2a**, yellow and green). Taken together, these findings suggest that a single experiment using unfractionated RNA gives insight into both nascent transcription and cytoplasmic RNA levels.

We then investigated an experiment on rhythmic expression in mouse liver[29]. RNA levels were determined in 4-h intervals over 48 h, using both total RNA-seq and Nascent-seq, in which nascent RNA was extracted using a similar approach as in the macrophage activation study[28,29]. In mouse liver, roughly 10% of genes show rhythmic expression with a 24-h period[29–31]. Using Nascent-seq as a reference, we applied stringent selection criteria and identified 799 genes with rhythmic transcription (Online Methods and **Supplementary Script 1**). These genes displayed phases over the entire 24 h (**Fig. 2b**). When sorted according to the Nascent-seq phase, a clear oscillatory pattern was also evident in both exonic and intronic levels from the total RNA-seq experiment (**Fig. 2b**). Close inspection revealed that intronic peak activity was generally shifted to earlier time points with respect to exonic peak activity. For example, genes with late phases had their peak intronic levels at 6 or 10 h, whereas exonic levels reached their maxima at 10 or 14 h (**Fig. 2b**, bottom). To globally quantify these shifts in activity, we also inferred phases for the oscillating genes using only intronic or exonic reads. Compared to the exonic phases, intronic phases are significantly shifted toward earlier time points ($P = 1.6 \times 10^{-20}$), and, importantly, did not shift systematically with respect to

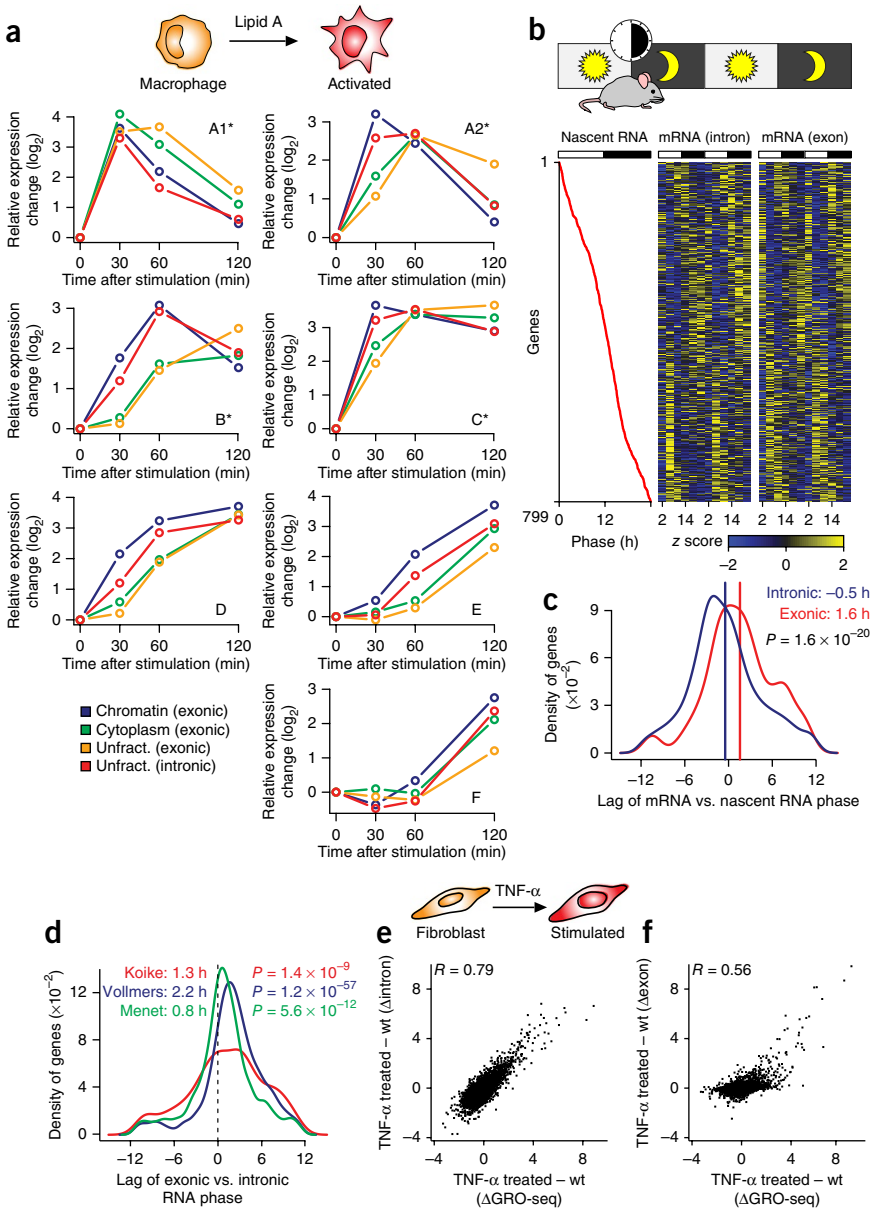

Figure 2 Intronic changes reflect changes in transcriptional activity. (**a**) Time-course experiment of RNA levels from fractionated and unfractionated cells after lipid A stimulation of mouse bone marrow-derived macrophages[28]. Shown are the average expression changes for all seven originally identified clusters of genes with similar expression profiles (A1 to F). The time point 15 min from the original fractionated data set was not used because it was not profiled in the unfractionated time course. To test if the unfractionated intronic expression levels were significantly more similar to the chromatin fraction than the cytoplasmic fraction, we performed a one-sided, paired *t*-test comparing the absolute differences between the unfractionated intronic and chromatin levels (d_{cr}) to the absolute differences between the unfractionated intronic and cytoplasm levels (d_{cy}). These were calculated as follows: d_{cr} = abs(unfract.intronic – chromatin) and d_{cy} = abs(unfract.intronic – cytoplasm), where d_{cr} and d_{cy} contain paired expression values of the respective genes at any of the probed time points. *, $P < 0.02$. (**b**) Changes in nascent, intronic and exonic RNA levels of circadian genes during day-night cycles in mouse liver[29]. Samples were taken in 4-h intervals over 48 h (*x* axis, starting at 0 h for Nascent-seq and 2 h for RNA-seq). Oscillating genes have been identified and sorted by phase (left panel) based on Nascent-seq data. The middle and right panels show *z*-scores of intronic and exonic expression levels for the same genes based on total RNA-seq data. (**c**) Distribution of time lags between intronic (blue) or exonic (red) and nascent phases. Mean phase differences are indicated in the key and by vertical lines. The shift between the two distributions was assessed by a Wilcoxon rank sum test. (**d**) Distribution of time lags between exonic and intronic phases for the three circadian data sets analyzed[29,31,32]. Mean phase differences are indicated in the legend. Wilcoxon signed rank tests were performed to assess if the distributions are significantly shifted with respect to zero. (**e**) Comparison of changes in GRO-seq to changes in intronic RNA levels in IMR90 fibroblasts 1 h after stimulation by TNF-α. (**f**) Same as **e** but comparing changes in GRO-seq to changes in exonic RNA levels. The *P* value (see text) was obtained from a test for the difference between two correlated correlations using the r.test function in the R package psych. This test determines whether two correlation coefficients are significantly different from one another, given that two variables are compared to a third common variable.

the phases inferred from Nascent-seq (**Fig. 2c**). Similar results have been obtained from two additional studies of circadian rhythm in mouse liver[31,32]. In these cases, intronic and exonic phases were directly compared to each other, as no matching Nascent-seq data were available. In all cases, a consistent and significant lag of exonic versus intronic phases was found (**Fig. 2d**, all $P < 1.4 \times 10^{-9}$).

An alternative method to measure transcription is GRO-seq[1], in which nascent transcripts are labeled by incorporation of a nucleotide analog, enriched and sequenced. We reanalyzed GRO-seq as well as RNA-seq data from IMR90 fibroblasts before and after stimulation by TNF-α (ref. 33). The transcriptional changes upon stimulation as measured by GRO-seq and the changes inferred from intronic read counts from the RNA-seq experiment were highly correlated (**Fig. 2e**; $R = 0.79$), in contrast to exonic changes, which show a lower similarity with GRO-seq (**Fig. 2f**; $R = 0.56$). The difference is highly significant ($P < 10^{-100}$), based on a test of difference between two correlated correlations[34]. Many genes that were upregulated according to GRO-seq did not change at the exonic level. Given that the profiling was done at an early time point (1 h after stimulation), it is likely that some transcriptional changes had not yet percolated to the mature mRNA level. Taken together, our findings support the idea that intronic reads can be used as a measure for changes in nascent transcription.

EISA recovers the role of transcription during neurogenesis

We applied EISA to transcriptome changes in a well-characterized, highly homogeneous differentiation system, in which mouse embryonic stem cells (ESCs) are differentiated into terminal neurons[35] (**Supplementary Data 3**). It was previously shown that most of the variation in gene expression in ESCs and terminal neurons can be predicted from chromatin marks, indicating that most changes in

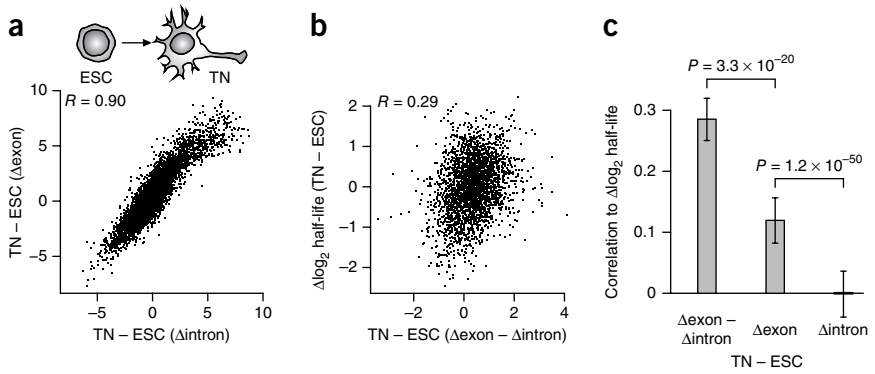

Figure 3 EISA recovers dominant role of transcriptional changes in neuronal differentiation. (**a**) Comparison of intronic (Δintron) and exonic (Δexon) expression changes in the differentiation of mouse ESCs to terminal neurons (TN). *R* indicates the Pearson correlation coefficient. (**b**) Post-transcriptional effects, as measured by Δexon – Δintron, versus the changes in RNA half-life between terminal neurons and ESCs, determined by transcriptional inhibition with actinomycin D[36]. (**c**) Pearson correlation coefficient for a comparison of changes in half-life with Δexon – Δintron, Δexon and Δintron, respectively. Error bars represent 95% confidence intervals. In all panels, Δexon and Δintron were averaged over two replicates. The *P* values were calculated as in **Figure 2e**,**f**.

expression are transcriptionally driven[36]. When readdressing this finding by quantifying both exonic and intronic changes from ESCs and terminal neurons using total RNA[36,37], we found very high agreement between exonic and intronic changes (**Fig. 3a**; $R = 0.9$). This provides direct evidence from standard RNA-seq alone that the majority of the changes in expression can be attributed to transcriptional changes. Nonetheless, we wondered if the small differences between Δexon and Δintron could reflect post-transcriptional regulation. Under a simple model of mRNA regulation, steady-state mRNA levels (m) are equal to the ratio of transcription (β) and degradation rate (α), $m = \beta/\alpha$ or $\log_2(m) = \log_2(\beta) - \log_2(\alpha)$[23]. In a differential setup, changes in degradation rate thus correspond to changes in steady-state mRNA levels minus changes in transcription rate, that is, $\Delta\log_2(m) = \Delta\log_2(\beta) - \Delta\log_2(\alpha)$, or Δexon – Δintron = $-\Delta\log_2(\alpha)$, assuming that changes in steady-state mRNA levels correspond to Δexon and changes in transcriptional rate to Δintron. mRNA half-life, on the other hand, is inversely proportional to the degradation rate ($t_{1/2} \sim 1/\alpha$) and thus, in a differential setup, $\Delta\log_2(t_{1/2}) = -\Delta\log_2(\alpha)$. Thus, altered post-transcriptional regulation of mRNAs should be manifested as altered mRNA stability, $\Delta\log_2(t_{1/2})$ = Δexon – Δintron. We therefore compared Δexon – Δintron to changes in $\log_2$ mRNA half-lives[36] from ESCs to terminal neurons (**Fig. 3b**) and found that they were significantly correlated (**Fig. 3c**; $R = 0.29$, $P < 2.2 \times 10^{-16}$). The fact that the correlation coefficient is not particularly high is not surprising given the high similarity of Δexon and Δintron and the substantial noise present in half-life measurements[38]. Notably, the correlation is reduced significantly ($P = 3.3 \times 10^{-20}$, test of difference between two correlated correlations[34]) when comparing half-lives to Δexon ($R = 0.12$) and absent when comparing them to Δintron ($R = -0.002$, $P = 1.2 \times 10^{-50}$). This provides direct evidence that Δexon is a composite measure of transcriptional and post-transcriptional changes and that the purely post-transcriptional component can be recovered by subtracting intronic from exonic changes.

EISA recovers the post-transcriptional impact of microRNAs

We next applied EISA to experiments with purely post-transcriptional perturbations mediated by microRNAs (miRNAs). The impact of a particular miRNA on its target messages can be probed by a transfection or inhibition experiment and subsequent expression profiling[39,40]. To investigate the influence of transfected miRNAs on exonic and intronic RNA levels, we analyzed polyA-RNA-seq data from a miR-1 transfection experiment in HeLa cells profiled after 12 and 32 h[41] (**Supplementary Data 4**). After 12 h, several genes were downregulated at the exonic level, but were virtually unchanged in introns (**Fig. 4a**). Because intronic changes measure transcriptional effects, we can directly deduce from this figure alone that miR-1 is a post-transcriptional regulator, without any prior knowledge of miRNA function. At 32 h, additional expression changes started to emerge, which were correlated between introns and exons and thus likely reflect transcriptional changes (**Fig. 4b**). We hypothesize that these are secondary effects of the miRNA transfection. Because direct targets of miRNAs are enriched in miRNA seed matches in their 3′UTR[39], we can test this hypothesis by monitoring the localization of miR-1 seed-containing genes within **Figure 4a**,**b**. Indeed this reveals that genes that were downregulated at the exonic as well as at the intronic level were not enriched for miRNA seed sites in their 3′UTRs (**Fig. 4a**,**b**, insets). We tested statistically if the shift between miRNA targets and nontargets was different in Δexon and Δintron (two-way ANOVA; **Supplementary Fig. 1a**) and obtained *P* values of 7.2×10^{-87} at 12 h and 8.0×10^{-59} at 32 h. Thus, EISA recovered the post-transcriptional effect of miRNAs. This opens up the possibility of separating primary from secondary effects in miRNA transfection or inhibition experiments.

To generalize these findings, we performed similar analyses on a wide range of additional transfection experiments, representing two different miRNAs and four different cell types[42] (**Fig. 4c** and **Supplementary Data 4**), all profiled at 24 h after transfection using polyA-RNA-seq. Overall, these experiments revealed a similar picture as obtained for miR-1. Although the profiling was done at a fixed time after transfection, the amount of exonic changes varied substantially between experiments, presumably owing to variable transfection efficiencies and/or kinetics of target inhibition. We sorted the experiments in **Figure 4c** by the amount of exonic variation from left to right. This revealed progressively increasing secondary effects. In miRNA transfection experiments, targets are typically identified by selecting the most downregulated genes (Δexon). However, our analysis suggests that intronic information should help to reduce the number of false positives due to secondary effects. To test this, we sorted genes according to Δexon – Δintron, Δexon or Δintron and calculated the fraction of seed-containing genes using increasing numbers of top-ranked genes (**Fig. 4d**). In the presence of strong secondary effects (HeLa cells in **Fig. 4d**), Δexon – Δintron clearly improves identification of strong targets over Δexon. For example, considering the top 50 genes in the miR-124 transfection, 70% of genes contained a seed match if ranked by Δexon, but 92% if ranked by Δexon – Δintron. A similar improvement is observed for miR-155 (50% versus 86%). In data sets with weak secondary effects (IMR90 and HEK293 cells in **Fig. 4d**), Δexon and Δexon – Δintron performed equally well for selecting targets, whereas in the absence of secondary effects (Huh7 cells in **Fig. 4d**), target selection on Δexon performed best. We believe that in cases where there are no secondary effects to be accounted

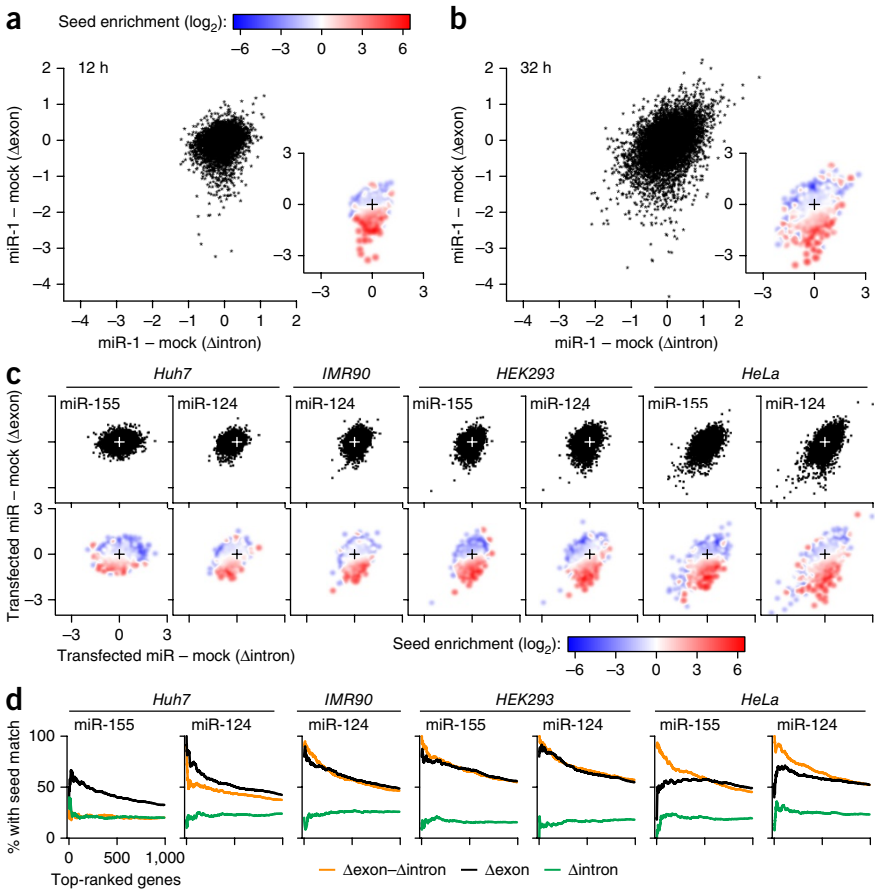

Figure 4 EISA recovers post-transcriptional mechanism of miRNAs. (**a**,**b**) Comparison of Δintron and Δexon 12 h (**a**) and 32 h (**b**) after miR-1 transfection in HeLa cells. Insets show localized enrichments for miR-1 seed-containing genes considering conserved as well as nonconserved predicted target sites in 3′UTRs. We estimated two-dimensional densities (bandwidth = 0.1) separately for seed-containing and nonseed-containing genes and divided the two to calculate seed enrichments (after adding a pseudo count of 10^{-3}). (**c**) Same as in **a**,**b** for miR-155 and miR-124 in Huh7, IMR90, HEK239 and HeLa cells, ordered by increasing amounts of exonic variation. Replicate experiments were averaged. (**d**) Fraction of genes with a seed match in their 3′UTR (*y* axis) for increasing numbers of top-ranked genes (*x* axis). Genes were sorted by Δexon – Δintron (orange), Δexon (black) or Δintron (green).

for, subtracting Δintron provides no benefit and only introduces additional measurement noise. Δintron by itself does not enrich for miRNA target genes, as expected from a purely transcriptional measure. We again statistically assessed whether the shift between miRNA targets and nontargets was different between Δexon and Δintron and obtained similar results as for miR-1 (**Supplementary Fig. 1b**). Taken together, these findings indicate that EISA can be used to estimate the extent of secondary effects in miRNA transfection experiments and to improve the selection of direct miRNA targets.

Application of EISA to different cell lines and tissues

To further study the relationship between transcriptional and post-transcriptional regulation, we investigated RNA-seq data from nine human ENCODE cell lines[9] (**Supplementary Data 5**). H1 ESCs provide an example of intronic versus exonic changes, relative to the average overall cell lines (**Fig. 5a**). The changes were highly correlated in all nine cell lines (**Fig. 5a**,**b**), whereas Δintron and Δexon were not correlated between different cell lines (**Fig. 5b**, off-diagonal). We repeated our analysis on transcriptomic data of 16 human tissues made publicly available by Illumina. This analysis yielded very similar results, with correlations around 0.8 between intronic and exonic changes of the same tissue, shown for brain in **Figure 5c** and for all tissues in **Figure 5d**. These findings indicate that as a general rule, most differences in RNA levels between cell types are accounted for by transcription. This is in agreement with earlier studies showing that gene expression levels can be accurately predicted using only chromatin modifications[36,43] or in addition based on transcription-factor binding data[44,45].

For tissue-specific miRNAs, expression is anti-correlated with mRNA expression (corresponding to Δexon) of the cognate target genes[46,47]. As EISA allows direct quantification of the post-transcriptional differences across tissues, this association should be even more pronounced when replacing Δexon by Δexon – Δintron. Using a linear regression approach[46,48] (Online Methods), we statistically assessed for all miRNAs if their targeting patterns were predictive for Δexon – Δintron, Δexon or Δintron in each tissue. A directional *P*-value (**Fig. 5e**) quantifies both the significance (absolute value) and the direction (sign) of the prediction. A negative sign indicates that the targets of a particular miRNA had values below the average (Δexon – Δintron, Δexon or Δintron), which is expected only in tissues where the miRNA is highly expressed and therefore represses its target genes. We selected five miRNAs known to be specifically expressed in brain, heart/skeletal muscle or liver[47]. For Δexon – Δintron, the strongest signal is always found in the tissue with high expression of the respective miRNA (**Fig. 5e**, black bars, tissue with high expression indicated by arrowhead). Although qualitatively similar results can be obtained through prediction of Δexon (dark gray bars) in accordance with previous findings[47], Δexon – Δintron leads to higher significance for four of the five miRNAs, and equal significance for the remaining one. Notably, Δintron contains no information about the activity of miRNAs (light gray bars) despite its high correlation to Δexon. In summary, these findings show that EISA can detect small post-transcriptional differences even in systems that are mostly characterized by transcriptional changes.

Technical and statistical considerations

Several technical and statistical issues should to be taken into consideration when performing an EISA. First, the number of genes that can be quantified above a minimal number of reads strongly depends on the size of a data set, for example measured by the total number of uniquely mapped reads. To estimate these requirements, we randomly subsampled reads from each of the 17 data sets used in this study (**Supplementary Table 1**), separately for reads mapping to exons and to introns (**Fig. 6a**,**b**). All samples contain more exonic than intronic reads resulting in a larger number of quantifiable genes on the exonic level. For example, ten million reads mapping to gene bodies allowed quantification of about 11,000 genes based on exonic

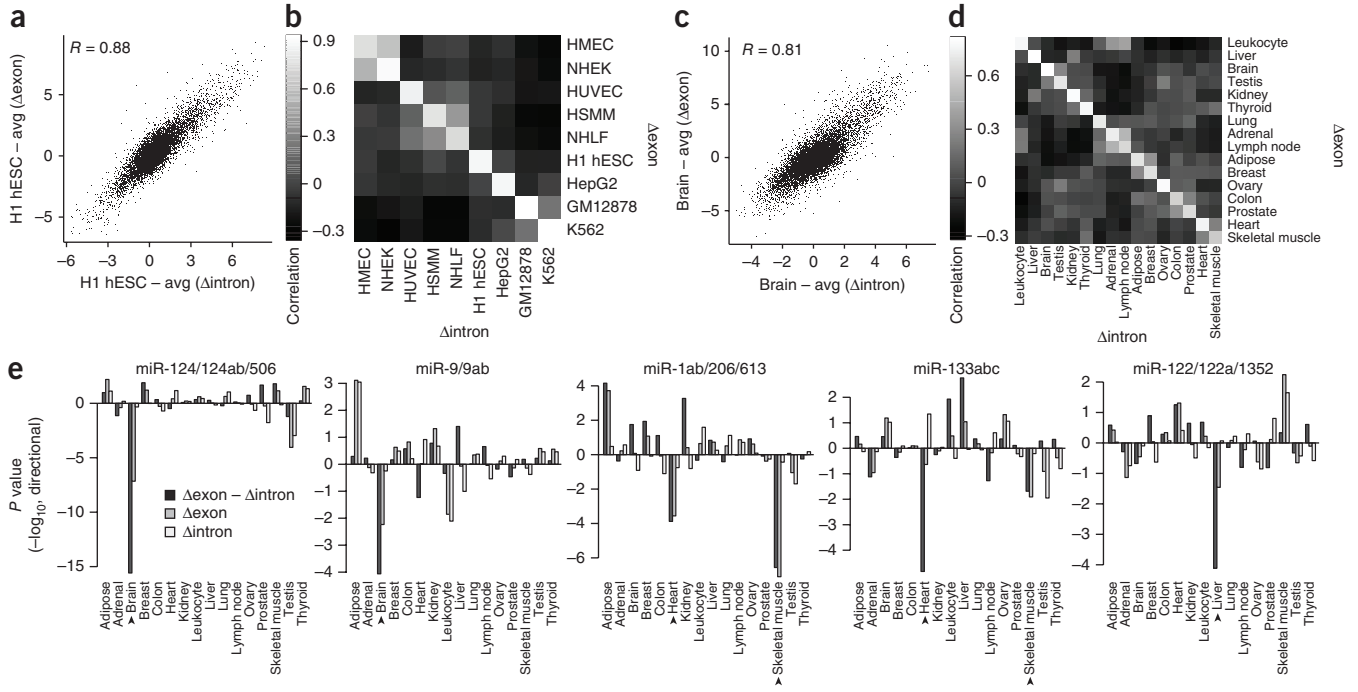

Figure 5 Quantification of transcriptional and post-transcriptional changes. (**a**) Comparison of Δintron and Δexon for H1 human embryonic stem cells (hESC) relative to the average over all ENCODE cell lines[9] (avg). *R* indicates the Pearson correlation coefficient. (**b**) Heatmap of Pearson correlations between Δintron and Δexon for all pairs of ENCODE cell lines. Cell lines were hierarchically clustered according to their exonic changes. (**c**) Same as in **a** for brain relative to the average (avg) over all tissues in the Illumina BodyMap 2.0 data set (ENA archive: ERP000546, **c**–**e**). (**d**) Same as in **b** for all tissues of the Illumina BodyMap 2.0. (**e**) Inference of tissue-specific miRNAs. For five miRNAs with known tissue-specific expression (tissue with high expression indicated by arrowheads), we tested whether their targets are predictive for Δexon – Δintron, Δexon and Δintron in each tissue of the Illumina BodyMap 2.0. The importance of a miRNA in a given tissue is displayed by a directional *P*-value (*y* axis), which quantifies the significance (absolute value) and the direction (sign) of the prediction (negative for target downregulation).

reads (**Fig. 6a**), and around 7,000 based on intronic reads (**Fig. 6b**). It has been reported that compared to total RNA-seq, polyA-RNA-seq results in only a small fraction of intronic reads[24,49]. Indeed, we find on average a higher fraction of intronic reads in total RNA data sets (24.4%) compared to polyA RNA data sets (14.8%). However, in the 17 data sets analyzed here there is only a weak association between the RNA isolation protocol and the fraction of intronic reads. It is likely that additional experimental factors influence the relative amounts of intronic and exonic reads that may differ even between experiments of the same protocol type. As a result, also polyA-RNA-seq data sets are amenable to EISA as illustrated by the many polyA data sets used in this study (**Supplementary Table 1**).

A second important issue that arises in an EISA is the evaluation of the significance of Δintron or Δexon – Δintron values from replicate experiments to confidentially call transcriptional or post-transcriptional changes, respectively. The significance of Δintron can be assessed in an identical manner to as it is commonly done for changes on the exonic level, using statistical methods such as for example, edgeR[21], DESeq[19] or voom as part of the limma software[20]. Testing for significance of Δexon – Δintron is more complicated as it requires the integration of exonic and intronic counts into the same statistical model. Absolute exonic and intronic read counts within one sample have very different distributions and show only moderate correlation (data not shown), which may be due to varying transcript and intron lengths across genes as well as differences in the capture efficiency and/or the half-lives of the introns. To account for this aspect, we modeled Δexon – Δintron in the framework of a generalized linear model in edgeR by introducing a count type ("exon" or "intron") as an additional factor in addition to the experimental conditions. To calculate the statistical significance for Δexon – Δintron, we incorporated an interaction term between these factors (**Supplementary Script 2**). The significance of this interaction is calculated based on a likelihood ratio test between the full model and a reduced model containing all experimental factors except the interaction term. We show an example application of this approach to the mouse differentiation system from ESCs to terminal neurons (**Figs. 3a** and **6c**), where two replicate experiments per condition are available. As expected, for genes that changed significantly according to their false discovery rate (FDR < 0.05), there was a large distance from the diagonal (**Fig. 6c**). Given the beforementioned contrasting properties of intronic and exonic read counts, it is, however, unclear whether the statistical approach presented here is optimal. Future work may lead to a more comprehensive and detailed statistical description and thus potentially more accurate results.

The sequencing requirements for detecting significant post-transcriptional changes depend on multiple factors, such as the fraction of intronic reads, the number and quality of replicate experiments and the magnitude of post-transcriptional changes in a particular system. We addressed the sequencing requirements in the previously analyzed ESCs-terminal neurons system for which replicate data is available. We subsampled the reads for each replicate separately and determined the number of significantly changing genes (FDR < 0.05) with a minimal fold-change of two and four in Δexon – Δintron. While the number of significantly changing genes clearly increases with sequencing depth, a slow saturation is apparent, in particularly at a cut-off of four (**Fig. 6d**). While the relatively low sequencing depth of 8 million reads does not allow us to determine the continuation of the curve at higher depth, the analysis shows that several hundreds of significantly changing genes can already be detected with only four million reads.

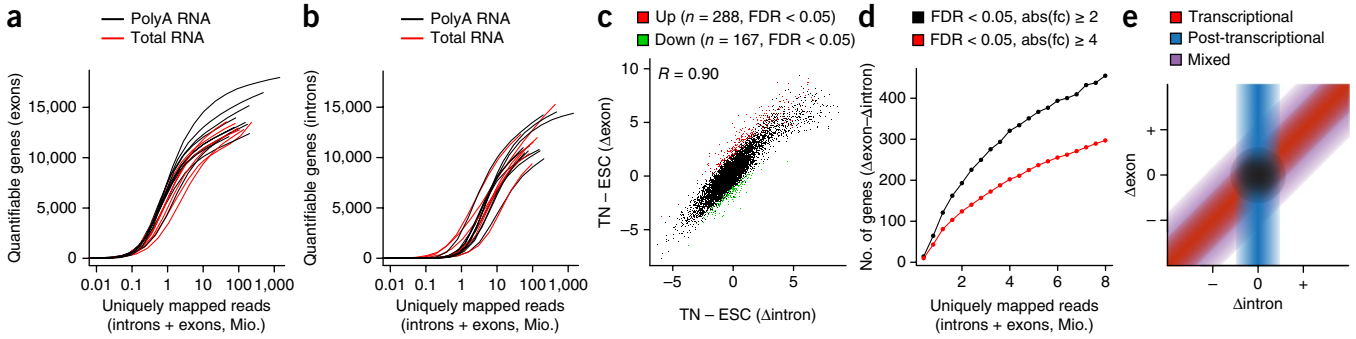

Figure 6 Read depth versus the number of quantifiable genes. (**a**,**b**) Reads were randomly subsampled from all data sets used in this study (**Supplementary Table 1**) and for each subset of reads, the number of quantifiable genes was estimated separately for reads in exons (**a**) or introns (**b**). The type of RNA enrichment is indicated by the color. (**c**) Changes in intronic and exonic levels between ESCs and terminal neurons (as in **Fig. 3a**). Mio., million. Significantly upregulated and downregulated genes on the level of Δexon – Δintron (FDR < 0.05) are shown in red and green, respectively. FDR was calculated by the topTags() function in edgeR using Benjamini & Hochberg adjustment. (**d**) For the same data set as in **c**, the number of significantly changing genes in Δexon – Δintron as a function of subsampled read depth and a minimal fold-change of two- and fourfold. (**e**) Schematic representation of the comparison between Δintron and Δexon used in **Figures 3**, **4** and **5**. Genes that reside in red and blue zones are predominantly regulated at the transcriptional and post-transcriptional levels, respectively. The purple zone contains genes that are regulated at both levels (mixed).

DISCUSSION

We show that application of EISA to standard RNA-seq data allows simultaneous quantification of changes in steady-state and nascent RNA levels across different experimental conditions (**Fig. 2**). Post-transcriptional effects can be measured by the difference in exonic and intronic changes across varying conditions, as demonstrated by its correlation to changes in mRNA half-lives during differentiation (**Fig. 3b**,**c**), improved identification of miRNA targets in the presence of secondary effects (**Fig. 4**) and improved inference of tissue-specific miRNA expression (**Fig. 5e**). As a result, RNA-seq experiments can be used to determine whether expression changes observed in a gene knockout, knock-down or overexpression experiment are caused by transcriptional or post-transcriptional mechanisms, providing information about the potential function of the gene under study (e.g., miR-1 in **Fig. 4a**). More generally, EISA can be used in any type of comparative study, for example, mutant versus wild type, treated versus untreated or diseased versus healthy, to gain insights into the regulatory mechanism responsible for the observed expression changes.

EISA should be used with caution in cases where proteins involved in the global regulation of the mRNA life cycle are perturbed. Extensive changes in transcript structures, caused, for example, by the perturbation of a splicing factor, can lead to misclassification of exonic and intronic reads. Similarly, perturbations of factors involved in the degradation of introns could lead to changes on the intronic level and would be wrongly interpreted as transcriptional by EISA.

It is known that long noncoding RNAs (lncRNAs) or enhancer RNAs (eRNAs) contribute to changes in intronic read counts[9–11,50]. Although this may affect specific genes of interest, EISA can be used on a genome-wide scale, probably because the intronic signal is accumulated over large genomic regions and may therefore not be strongly affected by localized events. The high correlation between intronic and exonic changes across many cell types (**Fig. 5b**,**d**) could only be explained by lncRNAs or eRNAs if their expression patterns generally and accurately followed the expression pattern of their host genes, which seems unlikely. In fact, although expression of exonic antisense lncRNAs is correlated with host gene expression, this correlation is much weaker between host gene expression and both sense and antisense intronic lncRNAs[11]. Additionally, lncRNAs and eRNAs are expressed at much lower levels than mRNAs[9,11,50] and are unlikely to produce a high number of intronic reads. Similarly, changes in splicing patterns are unlikely to explain the observed correlated intronic and exonic changes, as this would imply very strong constraints on splicing for the majority of genes. At the single gene level, however, alternative splicing or noncoding RNAs may lead to a misinterpretation of the intronic signal. This can potentially be mitigated by using the RNA-seq data to detect novel exons or small regions of high intronic read density, possibly reflecting noncoding RNAs. The resulting improvement in genome annotation is likely to translate to more accurate estimates of transcriptional and post-transcriptional changes.

EISA places genes into different zones on the Δintron versus Δexon plot based on a simple comparison of intronic and exonic changes (**Fig. 6e**). We have shown here that these zones separate genes that are under transcriptional control from genes that are regulated predominantly on a post-transcriptional level. This analysis can be performed with any RNA-seq data set, with no special experimental requirements or additional cost. Therefore, EISA increases the value of many existing and future RNA-seq data sets and provides a tool to study transcriptional and post-transcriptional contributions to expression changes.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Schübeler, A. Krebs, R. Ivanek and T.C. Roloff for feedback on the manuscript. We gratefully acknowledge funding from the Novartis Research Foundation.

AUTHOR CONTRIBUTIONS

D.G., L.B. and M.B.S. designed the study; D.G., L.B., M.F. and M.B.S. analyzed the data. L.B., D.G. and M.B.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
2. Wuarin, J. & Schibler, U. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol. Cell. Biol.* **14**, 7219–7225 (1994).
3. Khodor, Y.L. *et al.* Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in *Drosophila*. *Genes Dev.* **25**, 2502–2512 (2011).
4. Zaghlool, A. *et al.* Efficient cellular fractionation improves RNA sequencing analysis of mature and nascent transcripts from human tissues. *BMC Biotechnol.* **13**, 99 (2013).
5. Wang, Y. *et al.* Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. USA* **99**, 5860–5865 (2002).
6. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
7. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
8. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
9. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
10. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
11. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
12. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 91 (2013).
13. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).
14. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
15. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
16. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
17. Birol, I. *et al. De novo* transcriptome assembly with ABySS. *Bioinformatics* **25**, 2872–2877 (2009).
18. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
19. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
20. Law, C.W., Chen, Y., Shi, W. & Smyth, G.K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
21. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
22. DeLuca, D.S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
23. Zeisel, A. *et al.* Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Mol. Syst. Biol.* **7**, 529 (2011).
24. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18**, 1435–1440 (2011).
25. Gray, J.M. *et al.* SnapShot-Seq: a method for extracting genome-wide, in vivo mRNA dynamics from a single total RNA sample. *PLoS ONE* **9**, e89673 (2014).
26. Hendriks, G.J., Gaidatzis, D., Aeschimann, F. & Grosshans, H. Extensive oscillatory gene expression during *C. elegans* larval development. *Mol. Cell* **53**, 380–392 (2014).
27. Reddy, T.E. *et al.* Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.* **19**, 2163–2171 (2009).
28. Bhatt, D.M. *et al.* Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**, 279–290 (2012).
29. Menet, J.S., Rodriguez, J., Abruzzi, K.C. & Rosbash, M. Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. *Elife* **1**, e00011 (2012).
30. Storch, K.F. *et al.* Extensive and divergent circadian gene expression in liver and heart. *Nature* **417**, 78–83 (2002).
31. Koike, N. *et al.* Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science* **338**, 349–354 (2012).
32. Vollmers, C. *et al.* Circadian oscillations of protein-coding and regulatory RNAs in a highly dynamic mammalian liver epigenome. *Cell Metab.* **16**, 833–845 (2012).
33. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
34. Steiger, J.H. Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **87**, 245–251 (1980).
35. Bibel, M. *et al.* Differentiation of mouse embryonic stem cells into a defined neuronal lineage. *Nat. Neurosci.* **7**, 1003–1009 (2004).
36. Tippmann, S.C. *et al.* Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. *Mol. Syst. Biol.* **8**, 593 (2012).
37. Stadler, M.B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
38. Tani, H. *et al.* Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **22**, 947–956 (2012).
39. Lim, L.P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769–773 (2005).
40. Krützfeldt, J. *et al.* Silencing of microRNAs *in vivo* with 'antagomirs'. *Nature* **438**, 685–689 (2005).
41. Guo, H., Ingolia, N.T., Weissman, J.S. & Bartel, D.P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840 (2010).
42. Nam, J.W. *et al.* Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol. Cell* **53**, 1031–1043 (2014).
43. Karlić, R., Chung, H.R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. USA* **107**, 2926–2931 (2010).
44. Cheng, C. & Gerstein, M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.* **40**, 553–568 (2012).
45. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* **13**, R53 (2012).
46. Sood, P., Krek, A., Zavolan, M., Macino, G. & Rajewsky, N. Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc. Natl. Acad. Sci. USA* **103**, 2746–2751 (2006).
47. Farh, K.K. *et al.* The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310**, 1817–1821 (2005).
48. Bussemaker, H.J., Li, H. & Siggia, E.D. Regulatory element detection using correlation with expression. *Nat. Genet.* **27**, 167–174 (2001).
49. Cui, P. *et al.* A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* **96**, 259–265 (2010).
50. Kim, T.K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).

ONLINE METHODS

Processing of RNA-seq data. The RNA-seq data sets used in this study (**Supplementary Table 1**) were mapped in an annotation-free manner to the genome assemblies hg18 for human and mm9 for mouse using the R/Bioconductor package QuasR (version 1.2.2). QuasR[51] ties together all the tools necessary to obtain expression tables within R starting from the raw reads. It includes the aligners bowtie[52] and SpliceMap[53]. Bowtie was used to align all the samples with short reads (≤50), considering only uniquely mapping reads. The command used to perform the alignments was 'qAlign("samples.txt","BSgenome.Hsapiens.UCSC.hg18")' for human and 'qAlign("samples.txt","BSgenome.Mmusculus.UCSC.mm9")' for mouse, which internally instructs bowtie to create alignments with parameters "–m1 –-best –-strata –-phredNN-quals." For samples with longer reads, spliced alignments were performed with SpliceMap, using the command 'qAlign("samples.txt", "BSgenome.Hsapiens.UCSC.hg18", splicedAlignment = TRUE)' for human and 'qAlign("samples.txt", "BSgenome.Mmusculus.UCSC.mm9", splicedAlignment = TRUE)' for mouse, which internally instructs SpliceMap to perform spliced alignments with default parameters. Colorspace alignments for data sets from the SOLiD platform (GSE39978, SRA025656) were performed directly with bowtie using the parameters '-m 2–best–strata -C -S'. The miRNA transfection data sets (GSE21992, GSE52530) required the removal of the 3′ adaptor TCGTATGCCGTCTTCTGCTTG. This was performed using the function 'preprocessReads' from QuasR with default parameters. In the case of RNA-seq experiments with paired-end data only the first read was used.

Quantification of exonic and intronic levels. Using RefSeq mRNA coordinates from UCSC (genome.ucsc.edu, downloaded in October 2013) and considering only transcripts that map to a unique position in the genome, we quantified both the number of reads that started within any annotated exon of a gene (exonic) as well as the number of reads within the gene body that did not overlap any of the annotated exons (intronic). Exon coordinates were extended by ten basepairs on both sides to ensure that exonic reads close to the exon junctions were not counted as intronic reads. In GRO-seq experiments, we quantified the number of reads that started within the full gene body. Counting was performed using the function qCount from the Bioconductor package QuasR[51] considering read orientation in stranded experiments. For each data set, normalization for library size was performed by dividing each sample by the total number of reads and multiplying by the average library size. Notably, this normalization was performed separately for exonic and intronic reads as the exonic to intronic ratio can vary from data set to data set and from sample to sample. To minimize differences in expression across samples caused by genes with a small number of counts, $\log_2$ expression levels (exonic and intronic) were calculated after adding a pseudo-count of 8. Based on these expression levels we selected the genes with sufficient counts for downstream analysis by requiring an average $\log_2$ expression level of at least 5 (i.e., 24 counts) over all included samples and separately for exonic and intronic counts. Overlapping genes were not considered for analysis, as it is difficult to unambiguously assign the intronic reads to the respective genes. For stranded RNA-seq experiments, only genes on the same strand were considered as overlapping genes whereas for nonstranded experiments, genes on opposite strands were also considered as overlapping. Lists of nonoverlapping genes for human and mouse and for stranded and nonstranded analyses are available in **Supplementary Data 1**. Δexon (Δintron) was defined as the difference in $\log_2$ exonic (intronic) expression levels between the respective experimental conditions.

Analysis of circadian dynamics data sets. For each time point, $\log_2$ fold-changes of gene expression were calculated relative to the average over all time points. Phases and amplitudes of genes were then fitted separately for each type of RNA (nascent RNA, exonic total RNA and intronic total RNA) as follows (**Supplementary Script 1**). For each individual gene we fitted a cosine curve of the form $y = C\cdot\cos(\omega t + \varphi)$, where C is the amplitude and φ the phase, and the frequency $\omega = 2\pi/24$ h (known period of 24 h). Since $C\cdot\cos(\omega t + \varphi) = A\cdot\cos(\omega t) - B\cdot\sin(\omega t)$ with $A = C\cdot\cos(\varphi)$ and $B = C\cdot\sin(\varphi)$, we performed the fit using a linear regression using $\cos(\omega t)$ and $-\sin(\omega t)$ as regressors. The following lines of R code were used to perform the fit and calculate C and φ:

```
coeffs <– lm(y ~ cbind(cos(w*t),-sin(w*t)))$coefficients;
C <– sqrt(coeffs[2]^2 + coeffs[3]^2);
φ <– atan2(coeffs[3],coeffs[2]);
```

Oscillating genes were identified as genes with an amplitude in nascent RNA from data set GSE36872 greater than 0.35 (n = 799). For heatmap display, z-score expression values were calculated using scaled counts (normalized for library size) by subtracting the mean and dividing by the s.d. over time points for each gene.

RNA-seq analysis of ENCODE cell lines. For the ENCODE cell lines, separate RNA-seq data sets for polyA$^+$ and polyA$^-$ fractions are available. We combined both fractions for our analysis. To account for the different sequencing depths in the two fractions, for each sample separately, we normalized the total number of reads that map to gene bodies to the minimum total number and then combined the reads of the two fractions into a single fraction. Subsequently these samples were treated like every other data set for further normalization.

MiRNA expression inference in different tissues. Using predicted (conserved) miRNA target sites in 3′UTRs from TargetScan 6.2 (http://www.targetscan.org/) for 153 conserved miRNA families, we performed a linear regression for each tissue using the number of predicted miRNA target sites for each miRNA as regressors and different types of relative expression levels as the response. To control for possible technical confounders, we also included the AT content of the 3′UTRs as an additional predictor. In total the regression contained 154 variables for 10,968 data points and was performed for each tissue three times separately using Δexon – Δintron, Δexon or Δintron as the response variable.

Subsampling of RNA-seq read counts. To estimate the number of quantifiable genes as a function of sequencing depth, subsampling of RNA-seq read counts was implemented by sampling n elements from a multinomial distribution with k possible outcomes (k equals twice the number of genes; each gene represented separately for exons and introns) and probabilities corresponding to the observed fraction of reads per gene and read type (exons or introns). For each data set, values for n were selected between 1 and the sum of exonic and intronic counts in all samples of the data set. For each value of n, the sampling was repeated 20 times, and the average number of quantifiable genes as defined above was recorded separately for introns and exons.

To determine the sequencing requirements for detecting significant post-transcriptional changes in the ESC-terminal neurons differentiation system, read subsampling was implement by modeling the read counts with a multivariate hypergeometric distribution with k possible outcomes, with k corresponding to the number of genes. The total fraction of subsampled reads f was varied from 5% to 100% (in steps of 5%) of the total of number of reads in each replicate data set and for each read type (exonic or intronic). For a given fraction f, the resulting count table was mean-normalized, filtered for genes with a sufficient number of reads (mean $\log_2$ counts ≥ 5 for exons and introns separately) and the number of significantly changing genes in Δexon – Δintron was determined using edgeR[21] (see Technical and statistical considerations and **Supplementary Script 2**). For each f, subsampling was performed 20 times and the final number of significantly changing genes determined as the average over all repetitions.

Additional online material. Lists of nonoverlapping human and mouse genes, raw and normalized count tables for the main figures and R code example are also available online from: http://www.fmi.ch/groups/gbioinfo/EISA/EISA.html.

51. Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M.B. QuasR: quantification and annotation of short reads in R. *Bioinformatics* **31**, 1130–1132 (2015).
52. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
53. Au, K.F., Jiang, H., Lin, L., Xing, Y. & Wong, W.H. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.* **38**, 4570–4578 (2010).

doi:10.1038/nbt.3269

Erratum: Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation

Dimos Gaidatzis, Lukas Burger, Maria Florescu & Michael B Stadler
***Nat. Biotechnol.*; doi:10.1038/nbt.3269; corrected online 8 July 2015**

In the version of this article initially published online, several errors appeared in the HTML version. In the section "EISA recovers the role of transcription during neurogenesis," the expression "$(t1/2 = 1/\alpha)$" should have read "$(t1/2 \sim 1/\alpha)$" in the sentence "mRNA half-life, on the other hand, is inversely proportional to the degradation rate $(t1/2 = 1/\alpha)$." In the Online Methods, "Analysis of circadian dynamics data sets," the symbol "<-" was given as "≤" in two cases and as "≤<" in one case; the formulas "coeffs ≤ lm(y ~ cbind(cos(w*t),-sin(w*t)))$coefficients"; "C ≤< sqrt(coeffs[2]^2 + coeffs[3]^2)"; "ϕ ≤ atan2(coeffs[3],coeffs[2])" should have been "coeffs <- lm(y ~ cbind(cos(w*t),-sin(w*t)))$coefficients"; "C <- sqrt(coeffs[2]^2 + coeffs[3]^2)"; "ϕ <- atan2(coeffs[3],coeffs[2])." In addition, the corresponding authors are Dimos Gaidatzis, Lukas Burger and Michael Stadler, rather than Dimos Gaidatzis, Lukas Burger and Maria Florescu.The errors have been corrected in HTML version of this article.