Research Article

For reprint orders, please contact: reprints@futuremedicine.com

Epigenomics

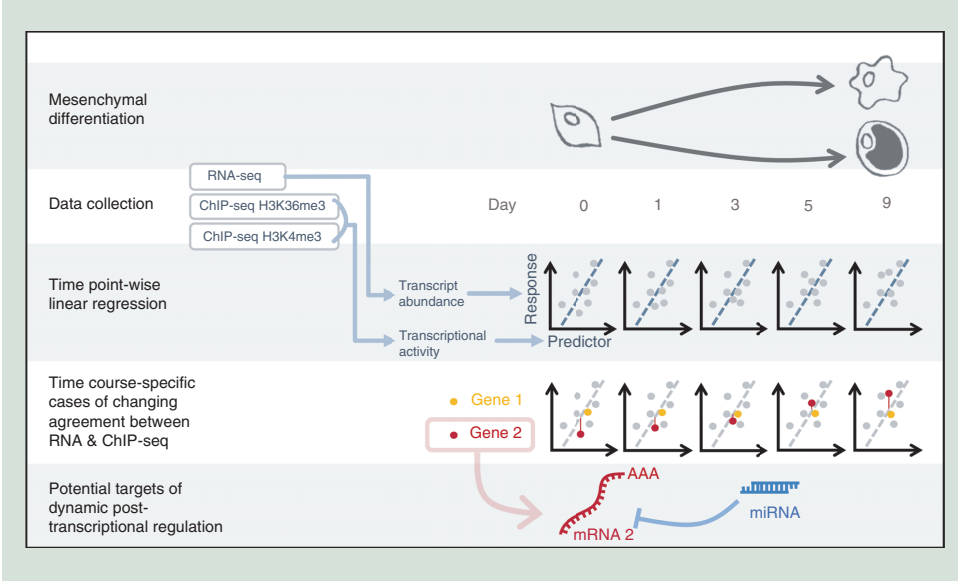
Identification of genes under dynamic post-transcriptional regulation from time-series epigenomic data

Julia C Becker[1], Deborah Gérard[1], Aurélien Ginolhac[1], Thomas Sauter[1] & Lasse Sinkkonen*,[1]

[1]Life Sciences Research Unit, University of Luxembourg, Belvaux, Luxembourg

*Author for correspondence: lasse.sinkkonen@uni.lu

Aim: Prediction of genes under dynamic post-transcriptional regulation from epigenomic data. **Materials & methods:** We used time-series profiles of chromatin immunoprecipitation-seq data of histone modifications from differentiation of mesenchymal progenitor cells toward adipocytes and osteoblasts to predict gene expression levels at five time points in both lineages and estimated the deviation of those predictions from the RNA-seq measured expression levels using linear regression. **Results & conclusion:** The genes with biggest changes in their estimated stability across the time series are enriched for noncoding RNAs and lineage-specific biological processes. Clustering mRNAs according to their stability dynamics allows identification of post-transcriptionally coregulated mRNAs and their shared regulators through sequence enrichment analysis. We identify miR-204 as an early induced adipogenic microRNA targeting *Akr1c14* and *Il1rl1*.



First draft submitted: 15 June 2018; Accepted for publication: 18 December 2018; Published online: 2 May 2019

Keywords: adipocyte • ChIP-seq • chromatin • differentiation • epigenomics • linear regression • microRNA • osteoblast • post-transcriptional regulation • RNA-seq

The signal for histone modifications associated with active and accessible chromatin has been known to correlate with respective gene expression levels since the earliest genome-wide chromatin immunoprecipitation (ChIP) experiments [1]. More recent work has described in further details which histone modifications, and in what combinations, are associated with high levels of transcription [2–5]. Currently, any measurement of chromatin

Future Medicine

accessibility combined with transcription factor (TF) binding motif data can allow accurate predictions of gene expression levels [6]. Many approaches ranging from linear regression to various machine learning approaches have been applied to prediction of gene expression from epigenetic data (see Singh *et al.* [7] for a more detailed description). The currently existing tools allow improved predictions by including the use of nonlinear models and combinatorial interactions with the latest deep-learning approaches currently achieving the most accurate predictions [7]. As a simple and easily accessible but still relatively robust approach, linear regression has been shown to model gene expression levels with good correlation to measured expression levels, already when using ChIP-seq data of only a few histone modifications, depending on the used histone mark and the tested genomic region [8–10]. Moreover, based on these linear regression models, the chromatin level measurements were shown to explain as much as 80% of the differences in mRNA levels, suggesting that regulation of mRNA stability through post-transcriptional mechanisms has only a limited effect on the genome-wide level. However, the same models could also show that individual genes under post-transcriptional control for example by microRNAs (miRNAs) can be identified as less stable and having lower actual expression levels than predicted by the linear regression model [9,10]. Moreover, in addition to the impact of post-transcriptional control, the deviations of predicted gene expression levels from measured expression levels can also depend on the exact input data used for prediction models. While expression of majority of the genes can be captured by measuring the levels of canonical histone modifications at promoters and gene bodies, unusually wide signals [11], absence of canonical histone marks [12] or for example cell-to-cell variation in gene expression [13] might influence the accuracy of the predictions.

The final outcome of protein expression levels is quite tightly linked to the levels of produced mRNA molecules and typically a change in mRNA levels leads to a change of similar magnitude in the protein levels, although the capacity of the mRNAs to translate into proteins varies greatly between the individual mRNAs [14]. This is consistent with the findings that miRNAs mainly regulate their target genes at the level of mRNA stability [15,16]. Despite their moderate impact at the genome-wide level, miRNAs play an important role in most developmental processes, as evidenced in mammals by the embryonic lethality of miRNA-depleted mice [17,18] and inability of miRNA-deficient embryonic stem cells (ESCs) to exit self-renewal and differentiate [19–22]. Moreover, miRNAs have now been shown to impact the terminal differentiation of most cell lineages, including the commitment to adipocytes and osteoblasts [23–25]. Most miRNAs show cell type-specific expression profiles, often switching on or off upon induction of cell differentiation and provide robustness to the differentiation processes by in parallel targeting and destabilizing multiple nodes within a gene regulatory circuit, and thereby reinforcing the new cell state [26,27]. Consequently, mRNAs important for cellular differentiation change in their stability during the differentiation process due to their dynamic targeting by miRNAs and other post-transcriptional regulators [28,29]. Conversely, genome-wide identification of changes in mRNA stabilities over time during differentiation or other cellular processes could allow identification of novel genes and regulatory interactions important for the new cell state.

Here, we use our recently generated time-series epigenomic data [30] of several histone modifications from the differentiation of multipotent bone marrow stromal progenitor cells (MSCs) into two terminally differentiated cell types, adipocytes and osteoblasts, to identify RNAs likely to be under dynamic post-transcriptional regulation in both lineages. To achieve this, we have used a linear regression model for each time point of differentiation to estimate the RNA expression levels from H3K4me3 and H3K36me3 signals obtained from the ChIP-seq experiments. We use the discrepancy between the predicted expression values and the actual expression values obtained by RNA-seq analysis (indicated by each genes residual) as a measure of RNA stability. Following the dynamics of this discrepancy across the time-series, we identify RNAs under changing levels of post-transcriptional regulation per lineage in an unbiased manner. We find that the RNAs with greatest changes in their stability during differentiation are enriched for noncoding RNA (ncRNA) species and for lineage-specific biological processes. By clustering mRNAs according to their stability dynamics and performing cluster-specific motif enrichment analysis, we identify the putative miRNAs targeting mRNAs with similar dynamics. We find miR-204 to be an early-induced adipocyte-specific regulator of lineage commitment capable of targeting its predicted targets, providing biological validation to our approach.

Materials & methods

Cell culture

The mouse bone marrow stromal cell line ST2 established from Whitlock-Witte type long-term bone marrow culture of BC8 mice [31] was used during all experiments. Cells were grown in Roswell Park Memorial Institute

(RPMI) 1640 medium (Gibco, Life Technologies, 32404014) supplemented with 10% fetal bovine serum (Gibco, Life Technologies, 10270-106, lot #41F8430K) and 1% L-Glutamine (Lonza, BE17-605E) and kept in a constant atmosphere of 37°C and 5% CO_2. In order to differentiate the mouse bone marrow stromal cell line into adipocytes and osteoblasts, ST2 cells were seeded 4 days before differentiation, reached 100% confluency after 48 h and were further cultivated for 48 h postconfluency (D0). Adipogenic differentiation was subsequently initiated on D0 by adding differentiation medium I consisting of growth medium, 0.5 mM isobutylmethylxanthine (Sigma-Aldrich, I5879, MO, USA), 0.25 µM dexamethasone (DEXA) (Sigma-Aldrich, D4902) and 5 µg/ml insulin (Sigma-Aldrich, I9278). From day 2 on differentiation medium II consisting of growth medium, 500 nM rosiglitazone (Sigma-Aldrich, R2408) and 5 µg/ml insulin (Sigma-Aldrich, I9278) was added and replaced every 2 days until 9 days of differentiation. Osteoblastic differentiation was induced with growth medium supplemented with 100 ng/ml bone morphogenetic protein-4 (BMP-4) (PeproTech, 315-27). Same media was replaced every 2 days until 9 days of osteoblastogenesis.

Small RNA transfections

The ST2 cells were seeded in six-well plates and transfected 48 h after reaching confluency using Lipofectamine RNAiMAX (Life Technologies, 13778150) following manufacturer's recommendations. Lipofectamine reagent and small RNAs were diluted in RPMI 1640 medium without glutamine, phenol red and FCS (Gibco, Life Technologies, 32404014). The small RNAs – siGENOME Lamin A/C Control siRNA (Dharmacon D-001050-01-05; si-*Lamin*) or siGENOME nontargeting siRNA Pool #2 (Dharmacon D-001206-14-05; si-Ctrl) or miRIDIAN miRNA Mouse mmu-miR-204-5p Mimic (Dharmacon C-310460-07-0005; miR-204) or miRIDIAN miRNA Mimic Negative Control #1 (Dharmacon CN-001000-01-05; miR-Ctrl) – were diluted to 50 nM, and Lipofectamine was diluted 9 µl in 150 µl dilution. The two solutions were combined 1:1 and incubated at room temperature for 5 min prior to transfection. After 3.5 h, the media was changed to standard growth medium (see above). Total RNA was collected 24 and 48 h following the medium change.

RNA extraction & cDNA synthesis

Total RNA was isolated from ST2 cells using TRIsure (Bioline, BIO-38033). Medium was discarded and 1 ml of TRIsure was added to each six well. The RNA was separated from DNA and proteins using 200 µl of chloroform (Carl Roth, 6340.1) and precipitated from the aqueous phase with 400 µl of 100% isopropanol (Carl Roth, 6752.4) and further incubated at -20°C overnight. The cDNA was synthetized using 1 µg of total RNA, 0.5 mM dNTPs (ThermoFisher Scientific, R0181), 2.5 µM oligo dT-primer (Eurofins MWG GmbH, Germany), 1 U/µl Ribolock RNase inhibitor (ThermoFisher Scientific, EO0381) and 1 U/µl M-MulV Reverse transcriptase (ThermoFisher Scientific, EP0352) for 1 h at 37°C or with RevertAid Reverse Transcriptase (Thermo Scientific, EP0442) for 1 h at 42°C. The PCR reaction was stopped by incubating samples at 70°C for 10 min.

Quantitative PCR

Real-time quantitative PCR (qPCR) was performed in an Applied Biosystems 7500 Fast Real-Time PCR System and using Thermo Scientific Absolute Blue qPCR SYBR Green Low ROX Mix (ThermoFisher Scientific, AB4322B). For each reaction 5 µl of cDNA, 5 µl of primer pairs (2 µM) and 10 µl of the Absolute Blue qPCR mix were used. The PCR reactions were carried out at the following conditions: 95°C for 15 min followed by 40 cycles of 95°C for 15 sec, 55°C for 15 sec and 72°C for 30 sec. To calculate the gene expression level the $2^{-(\Delta\Delta Ct)}$ method was used where $\Delta\Delta Ct$ is equal to $(\Delta Ct_{[target\ gene]} - \Delta Ct_{[housekeeping\ gene]})_{tested\ condition} - (\Delta Ct_{[target\ gene]} - \Delta Ct_{[housekeeping\ gene]})_{control\ condition}$. *Rpl13a* was used as housekeeping gene and D0 or control siRNA/miRNA were used as control condition depending on the experiment. Sequences of the primer pairs are listed in Supplementary Table 1.

RT-qPCR of mature miRNAs

Mature miRNAs were reverse transcribed using the TaqMan® MicroRNA Reverse Transcription kit (Applied Biosystems, 4366597) and the quantification of mature miRNAs was done by using the TaqMan MicroRNA Assays (Applied Biosystems, 4440040) according to manufacturer's instructions an Applied Biosystems 7500 Fast Real-Time PCR System. The PCR amplification reaction was prepared using the Applied Biosystems TaqMan 2X Universal PCR Master Mix II, with UNG. To calculate the gene expression level the $2^{-(\Delta\Delta Ct)}$ method were used

10.2217/epi-2018-0084

where ΔΔCt is equal to $(\Delta Ct_{[target\ miRNA]} - \Delta Ct_{[U6]})_{tested\ condition} - (\Delta Ct_{[target\ miRNA]} - \Delta Ct_{[U6]})_{control\ condition}$. *U6* snRNA was used as control gene and D0 as control condition.

Chromatin immunoprecipitation followed by qPCR

To validate the methylation profiles of H3K4me3 and H3K36me3 observed over the time course of differentiation in the ChIP-seq analysis, ChIP-qPCR analysis of biological replicates for selected genomic regions was performed. The ChIP analysis was performed as previously described [30]. Briefly, the chromatin was crosslinked with formaldehyde (Sigma-Aldrich, F8775-25 ML) at a final concentration of 1% for 8 min and quenched with glycine (Carl Roth, 3908.3). The cells were lysed in 1.7 ml of lysis buffer, followed by sonication in shearing buffer containing protease inhibitors with a sonicator (Bioruptor® Standard Diagenode, UCD-200TM-EX) during 20 or 25 cycles at high intensity (30s off and 30s on) for undifferentiated cells or differentiating cells, respectively. For each immunoprecipitation 10 μg (for an antibody against H3K4me3 [Millipore, 17-614]) or 15 μg (for an antibody against H3K36me3 [Abcam, ab9050]) of sheared chromatin and 4 μg as input were used. For each immunoprecipitation, 5 μg of antibody was used. The immunocomplexes were captured using 25 μl of PureProteome™ Protein A Magnetic Bead System (Millipore, LSKMAGA10). The Protein A Magnetic beads were washed with increasing salt concentration and protein–DNA complexes eluted. The crosslinks were reversed overnight at 65°C in the presence of 10 μg of RNase A (ThermoFisher, EN0531) and 20 μg of proteinase K (ThermoFisher, EO0491). Immunoprecipitated DNA was purified using a MinElute Reaction Cleanup Kit (Qiagen, 28206) and the purified DNA used as template in qPCR together with the respective input samples.

Primer sequences for the ChIP-qPCR at transcription start site (TSS) and transcription end site (TES) of *Akr1c14* gene and *Spon2* gene were designed by the PrimerBlast software (ABI). Primers producing a DNA amplicon at the TSS spanning region were used for the H3K4me3 enriched samples, and those producing an amplicon at the TES spanning region were used for the H3K36me3 enriched samples. Input samples were included as an enrichment control. As a negative control we included primers for the TSS of *Pou5f1*, an ESC specific gene that is silenced in ST2 cells. ChIP-qPCR was performed on three biological replicates of immunoprecipitated genomic DNA unless otherwise indicated. The PCR reaction itself was performed as described above (section 'quantitative PCR'), with the following exception: for each reaction 1 μl of genomic DNA, 1 μl of primer pairs (2 μM) and 10 μl of the Absolute Blue qPCR mix were used, and 8 μl water topped up each reaction to a volume of 20 μl.

Data preprocessing

Gene expression from time-series transcriptomic data from differentiation of multipotent bone MSCs into adipocytes and osteoblasts was assembled and quantified using Cufflinks suite version 2.2.1 [32–34] with the mm10 reference genome with the options: GTF guide which use the supplied annotation file and the RABT assembly internal to Cufflinks to discover novel genes and isoforms, frag-bias-correct which improves the accuracy of the estimated abundance of transcripts, library-type set to fr-unstranded which indicates that the RNA-Seq library is unstranded and multiread-correct which corrects for multireads mapping, enabled. Cuffmerge was then used to build a master transcriptome from adipocyte and osteoblast individual transcriptomes with the options: ref-gtf which merge the provided gene assembly with novel genes and isoforms detected and ref-sequence which compares the newly made gene assembly with known transcripts, enabled. Then, Cuffquant was used to compute gene expression profiles with the frag-bias-correct which improves the accuracy of the estimated abundance of transcripts, multiread-correct which corrects for multireads mapping and library-type set to fr-unstranded which indicates that the RNA-Seq library is unstranded, enabled. Finally, Cuffnorm was used to normalize the expression levels using library-type set to fr-unstranded which indicates that the RNA-Seq library is unstranded, library-norm-method geometric which scale the fragments per kilobase of transcript per million mapped reads (FPKM) using the median of the geometric means over all RNA-Seq libraries and compatible-hits-norm which counts fragments that are compatible with transcripts in the reference gene annotation, enabled. The analysis of the ChIP-seq data has been described in Gerard *et al.* [30]. Prior to using them as variables in a linear regression model, the four data types from each time point (total of nine time points from two lineages) were preprocessed as indicated in the workflow in Supplementary Figure 1A. The reads per million (RPM) values for the time-series epigenomic data of H3K4me3 and H3K27ac were calculated over 4 kilobase (kb) windows around the TSS as inferred by the Cufflinks suite and, given the defined width of the region, were normalized only to library size. In case of multiple alternative TSS of one genomic locus, the 4 kb TSS-spanning region of the first TSS was used. Finally, reads per kilobase million (RPKM) from the time-series epigenomic data of H3K36me3 were calculated from the TSS to the TES as

inferred by the Cufflinks suite and were also normalized to both library size and gene length. The distributions of the normalized, log-transformed and noise-filtered read counts for each data type from both lineages are shown in Supplementary Figure 1D and E.

Reads per million, RPKM and FPKM values from epigenomic and transcriptomic data were subsequently $\log_2$-transformed after adding a pseudo-count of 0.001 to obtain closer to normal distributions. Genes were filtered using a distribution-based approach. For this, a filtering cutoff was selected based on the data distribution of each of the data collection time points separately by locating low points in each density distribution data series of *y*-axis coordinates and all data points below the cutoff were excluded from the dataset. Importantly, any data point that was below cutoff even for one time point of one of the variables, was excluded from the entire dataset (Supplementary Figure 1B–E).

Statistical computing was performed using R (R Core Team [2017]. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria [www.R-project.org/]) in Rstudio (RStudio Team [2016]. RStudio: integrated development for R. RStudio, Inc., Boston, MA, USA [www.rstudio.com/]) and with *tidyverse* packages (Hadley Wickham [2017]. Tidyverse: easily install and load 'tidyverse' packages. R package version 1.1.1. https://CRAN.R-project.org/package=tidyverse

Linear regression

Using the processed data as an input, a linear regression model was fitted for each time point of the time courses. The RNA-seq measurements were assigned to represent the response variable, and ChIP-seq measures the predictor variables for the regression. The model equation used for each time point's data was of the following format:

$$mRNA\ abundance = \beta_0 + \beta_1 \times H3K4me3_{intensity} + \beta_2 \times H3K36me3_{intensity} \quad (1)$$

Where β_0 represents the intercept, β_1 and β_2 represent the two slopes. During model fitting, residual plots were examined for single high-leverage outliers and four genes in the adipogenesis (*RP23-81C12.2*, *Gm10800, Gm10801* and *Yam1*), all of which consistently exceeded a H3K36me3 $\log_2$-RPKM value of four across all time points, were excluded. The same genes, with the exception of *Gm10801*, were also excluded in the osteoblastogenesis dataset for the same reason. Scatter, density and bar plots were produced with *ggplot2* [35].

Gene ontology & motif enrichment analysis

Prior to motif enrichment, the selected gene lists were clustered by similarity of the temporal profiles of their assigned raw residuals – reflecting the relative agreement between their measured transcript abundance and their predicted transcription rate using the the short time-series expression miner [36]. Using the resulting groups as an input, motif enrichment of 3′-UTRs was carried out, to identify common putative regulators within groups of similar temporal profiles as previously described [22,37]. In brief, the input groups (foreground) were tested for significantly more frequently occurring motifs compared with the list of 3′-UTRs of all expressed genes of the respective differentiation experiment (background). The enrichment of a motif was considered significant if the associated posterior probability was above 0.9. The 3′-UTR sequences were retrieved from the Ensembl database from the mm10 version (GRCm38.79). Not all genes have available 3′-UTR information. Hence, these cases could not be included in the analysis. Further, for some genes more than one possible 3′-UTR sequence is available. In such cases, the longest alternative was included as a representative 3′-UTR. Finally, in some cases, several genes are summarized in the representation of one data point in the model (and therefore one profile within the the short time-series expression miner analysis). These cases originate from the data preprocessing stage, where the Cufflinks analysis fails to separate individual genes of close genomic proximity and a single XLOC-id is assigned to multiple genes. The UTR-sequences for each of those summarized genes were included in the background and foreground lists for the motif enrichment. In a second step, the reverse complements of the enriched motifs were mapped to known miRNA sequences to identify enrichment of potential miRNA seed regions. The sequences of mature miRNAs were downloaded from miRbase (11 October 2016) [38].

In addition to motif enrichment, EnrichR [39,40] was used to perform gene ontology term enrichment analysis and to retrieve enriched terms of the type biological process.

Metagene analysis

Metagene plots (Supplementary Figure 5) were generated as follows. Matrices of scaled gene bodies were computed using the tool 'compute Matrix' in 'scale-regions' mode [41]. Common options were set as '-m 10000 -b 2500 -a 2500 -bs 25 –skipZeros– metagene'. So all gene bodies were normalized to 10 kb, and 2.5 kb up- and down-stream were considered. Binning was fixed to 25 bp. Bigwig inputs were from both lineages (i.e., osteoblast and adipocyte differentiation time points), plus the day 0 (D0) from the ST2 progenitor cell line. The reference gene sets were the following three GTF files:

- All genes, using the gencode file release M5 of the mouse genome GRCm38.p3.
- Top 5% of adipocyte lineage genes, after ranking them using the sdR standard in the adipocyte lineage (647 genes).
- Top 5% of osteoblast lineage genes, after ranking them using the sdR standard in the osteoblast lineage (668 genes).

Of note, some XLOC_IDs (46 and 47 for adipo and osteo, respectively) had more than one gene symbol associated. Those genes were manually curated to retain one. Once the matrices were obtained, they were loaded into R (v3.5.1 R Core Team 2018) and plotted with ggplot2 [35] using the mean of the average intensity per bin.

Data deposit

The used RNA-seq and ChIP-seq data have been published and are described in detail in Gerard *et al.* [30] and the raw FASTQ and BAM files have been deposited in the European Nucleotide Archive with the accession number PRJEB20933.

Results

Identification of genes under dynamic post-transcriptional regulation from time-series epigenomic data using linear regression

Previous work has shown the usability of linear regression models in predicting gene expression levels from genome-wide ChIP-seq data [8–10]. We apply this approach for the transcriptome-wide identification of RNAs undergoing changes in their stability during the dynamic process of cellular differentiation. To this end, we have generated and taken advantage of parallel time-series transcriptomic and epigenomic profiles of differentiation of two mesenchymal cell types, adipocytes and osteoblasts, from their shared MSC precursors (Figure 1A) [30]. The transcriptomic profiles were produced using RNA-seq while the epigenomic profiles consist of ChIP-seq data for three histone modifications: H3K4me3 (marking open TSS), H3K36me3 (marking actively transcribed gene bodies) and H3K27ac (marking active enhancers and the most active TSSs).

As shown in Supplementary Figure 2A and B, in both lineages all three histone modifications are positively correlated with the transcriptome-wide RNA-seq measurements. Between the histone modifications, a particularly high correlation is found between H3K4me3 and H3K27ac signals, suggesting that these two modifications contribute relatively redundant information. Indeed, when fitting a linear regression model with RNA-seq measured transcript levels as a response variable and each possible combination of the histone modifications signals as predictor variables, the adjusted r^2-statistic for each modifications' predictive power indicated that H3K4me3 and K3K27ac did not provide additional value over each other (Figure 1B, Supplementary Figure 2C & D). Therefore, in the remaining analysis, we focused on a model with H3K4me3 and H3K36me3 as the sole predictor variables (Figure 1B). The resulting model fits for each time point of both differentiations are shown in Supplementary Figure 3 and reflect the relationship of the predicted transcript levels over the actual RNA-seq measured transcript levels.

The transcripts under post-transcriptional control from different regulators, such as miRNAs, are known to deviate in their predicted expression level from the actual measured expression value [9,10]. This deviation can be represented by each gene's absolute raw residual value ($r_{gene\ 1}$) obtained from the linear regression model (Figure 1C). To identify the transcripts under dynamic post-transcriptional regulation, we obtained the residual values for each gene from all time points, and calculated the standard deviation of those residuals per gene and time course (*sdR*, not to be confused with root mean squared error or RMSE), to arrive to an estimate of each gene's change in post-transcriptional control across time as illustrated in Figure 1C. Distribution of the obtained *sdR* values shows that most genes have a low *sdR*, indicating little change in the deviation from the predicted expression levels

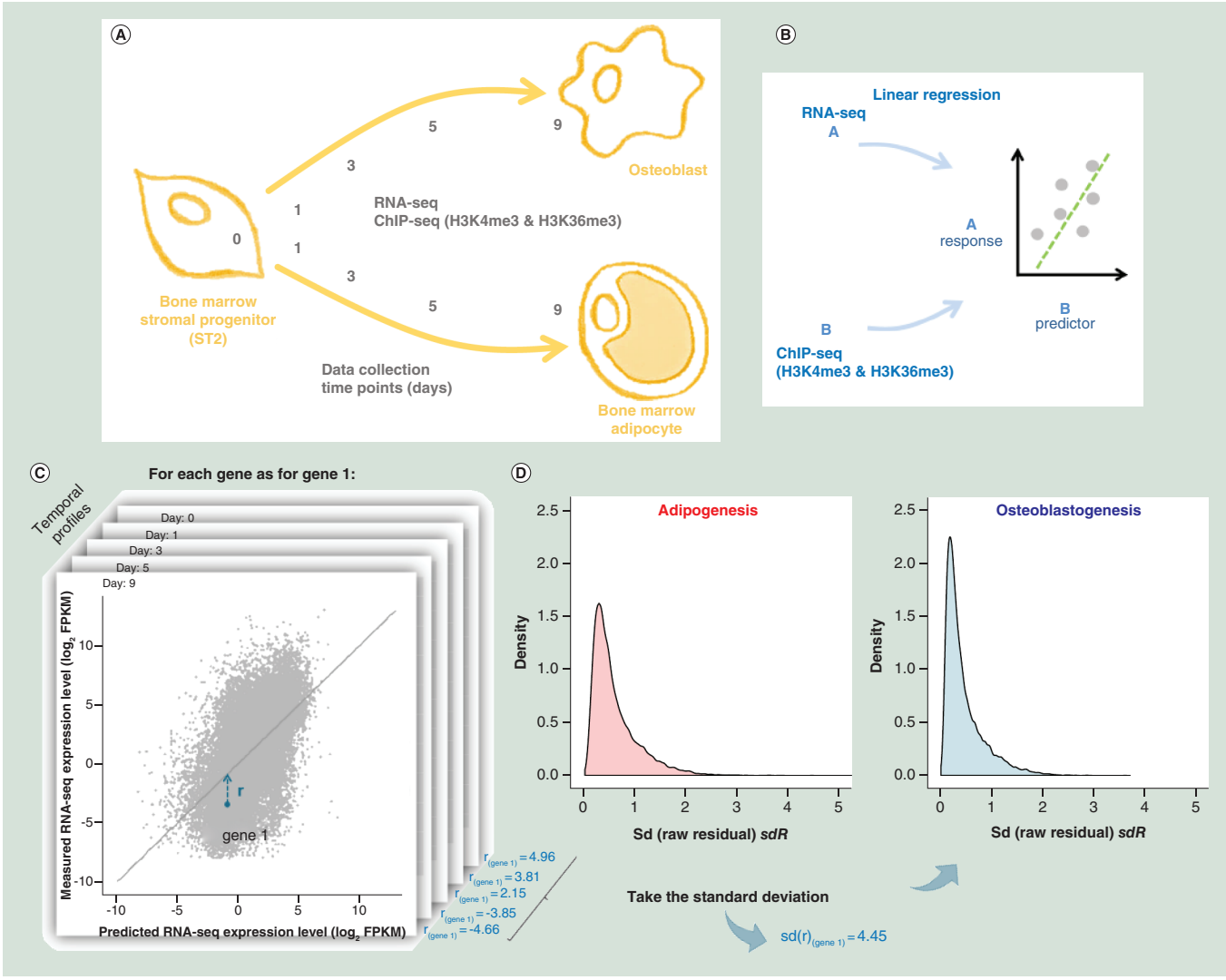


Figure 1. Prediction of genes under dynamic post-transcriptional regulation from time-series epigenomic data of mesenchymal differentiation using linear regression. (A) ST2 mesenchymal progenitors were differentiated to osteoblast and adipocyte phenotypes. At specified time points along both time courses, RNA-seq and ChIP-seq experiments were performed (for detailed description see Gerard *et al.* [30]). **(B)** The resulting data were integrated using a linear regression approach, with H3K4me3 and H3K36me3 ChIP-seq measurements serving as the predictor and RNA-seq as the response variables. **(C)** One regression fit per data collection time point allowed to look at each gene's measure of fit deviation (the residual) over time. **(D)** To approximate a gene's change in predictability, the standard deviation of its residual values resulting from the time point-wise regression fits (*sdR*) was taken. Shown is the density distribution of the *sdR* values assigned to each data point in the model series fit for each of the time course experiments. **(E–L)** A large *sdR* indicates a dynamic change of the agreement between a gene's measured and predicted RNA-seq level during the time course in question, while a small *sdR* indicates a low change in the level of this agreement. Data associated with two example genes of each time course experiment is shown. The upper subplots display the last time point's global model, highlighting each example gene and a trace of its previous positions respective to the fit. The measured and predicted RNA-seq values over time are shown in the smaller subplots. FPKM: Fragments per kilobase of transcript per million mapped reads; RNA-seq: RNA sequencing.

over time (Figure 1D). However, these density plots are positively skewed, with a number of transcripts changing in their deviation from the predicted expression levels during differentiation. As an example of a gene with a large *sdR* value, Figure 1E shows the migration of the measured expression levels of Complement factor D (*Cfd*) mRNA from being lower than predicted by the model in the undifferentiated cells (indicating presence of post-transcriptional repression) to being several folds higher than predicted from the chromatin level data (indicating increased mRNA stability). For clarity, the same data are also depicted as a line graph in Figure 1F. At the same time CLIP-associating protein 1 (*Clasp1*) shows very little deviation from the predicted expression levels with no changes

10.2217/epi-2018-0084

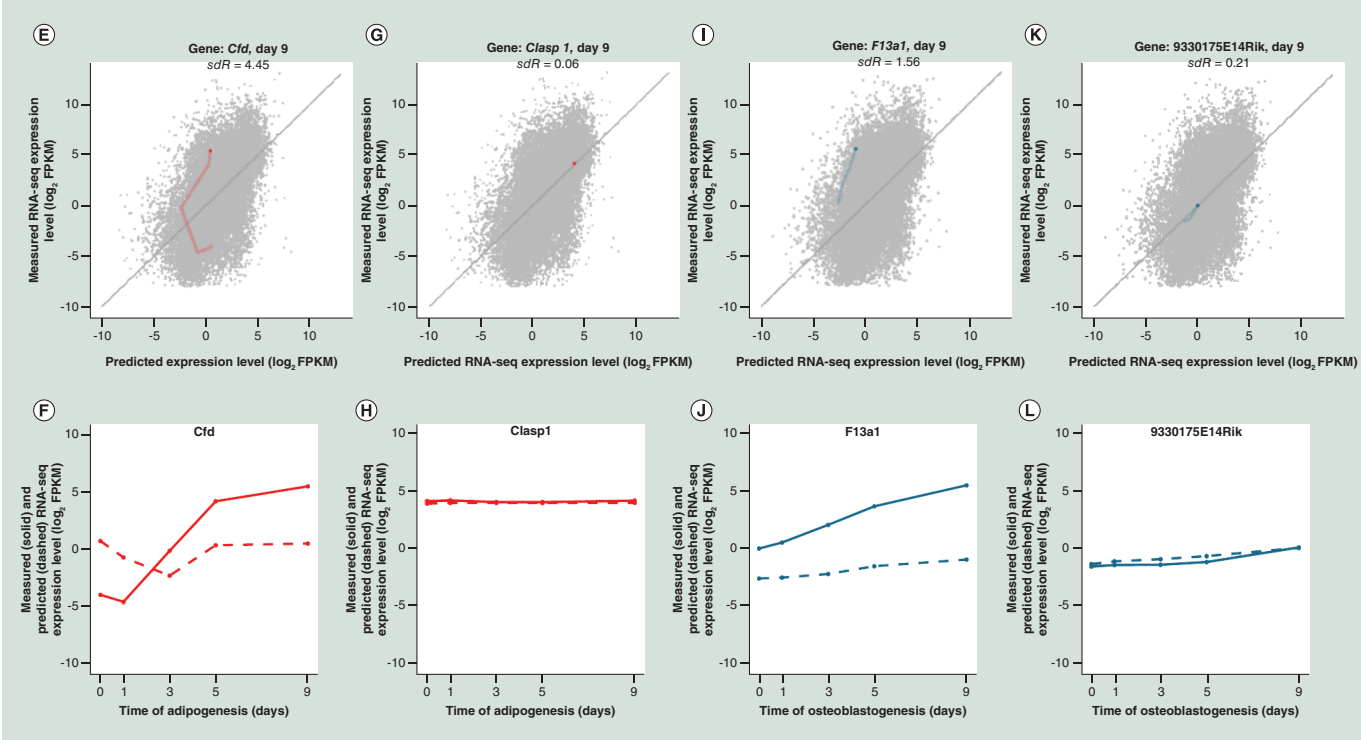


Figure 1. Prediction of genes under dynamic post-transcriptional regulation from time-series epigenomic data of mesenchymal differentiation using linear regression (cont.). (A) ST2 mesenchymal progenitors were differentiated to osteoblast and adipocyte phenotypes. At specified time points along both time courses, RNA-seq and ChIP-seq experiments were performed (for detailed description see Gerard *et al.* [30]). **(B)** The resulting data were integrated using a linear regression approach, with H3K4me3 and H3K36me3 ChIP-seq measurements serving as the predictor and RNA-seq as the response variables. **(C)** One regression fit per data collection time point allowed to look at each gene's measure of fit deviation (the residual) over time. **(D)** To approximate a gene's change in predictability, the standard deviation of its residual values resulting from the time point-wise regression fits (*sdR*) was taken. Shown is the density distribution of the *sdR* values assigned to each data point in the model series fit for each of the time course experiments. **(E–L)** A large *sdR* indicates a dynamic change of the agreement between a gene's measured and predicted RNA-seq level during the time course in question, while a small *sdR* indicates a low change in the level of this agreement. Data associated with two example genes of each time course experiment is shown. The upper subplots display the last time point's global model, highlighting each example gene and a trace of its previous positions respective to the fit. The measured and predicted RNA-seq values over time are shown in the smaller subplots. FPKM: Fragments per kilobase of transcript per million mapped reads; RNA-seq: RNA sequencing.

in expression over the adipogenesis, consistently with a very low *sdR* value of 0.06 (Figures 1G–H). Similarly, during osteoblastogenesis, coagulation factor XIII A chain (*F13a1*) is strongly induced but without a major increase in the histone modifications, obtaining *sdR* of 1.56 (indicating progressive increase in mRNA stability), while another gene, *9330175E14Rik*, is increasing both in the measured and in the predicted expression levels (*sdR* = 0.21; suggesting a lack of major change in post-transcriptional regulation; Figure 1I–L). See also Supplementary Figure 4 for additional examples of genes with high *sdR* in osteoblastogenesis.

Taken together, combining time point-specific linear regression models predicting RNA expression levels from time-series epigenomic data, and more specifically, calculating the *sdR* values for all genes across differentiation, allows identification of RNAs with changes in their expression predictability, thus identifying them as putative targets of dynamic post-transcriptional control.

Genes with greatest changes in the deviation over time from predicted expression levels are enriched for ncRNAs & lineage specific functions

To study the genes with biggest changes in their post-transcriptional regulation either during adipogenesis or osteoblastogenesis, we extracted the top 5% of transcripts with highest *sdR* values from both datasets (Figure 2A & B). Overall the H3K36me3 and H3K4me3 signals for these top transcripts showed similar profiles as for all other genes across the time points, with an increasing enrichment of H3K36me3 toward the end of the genes and the strongest enrichment of H3K4me3 at the TSS (Supplementary Figure 5). While signals for H3K4me3 appeared

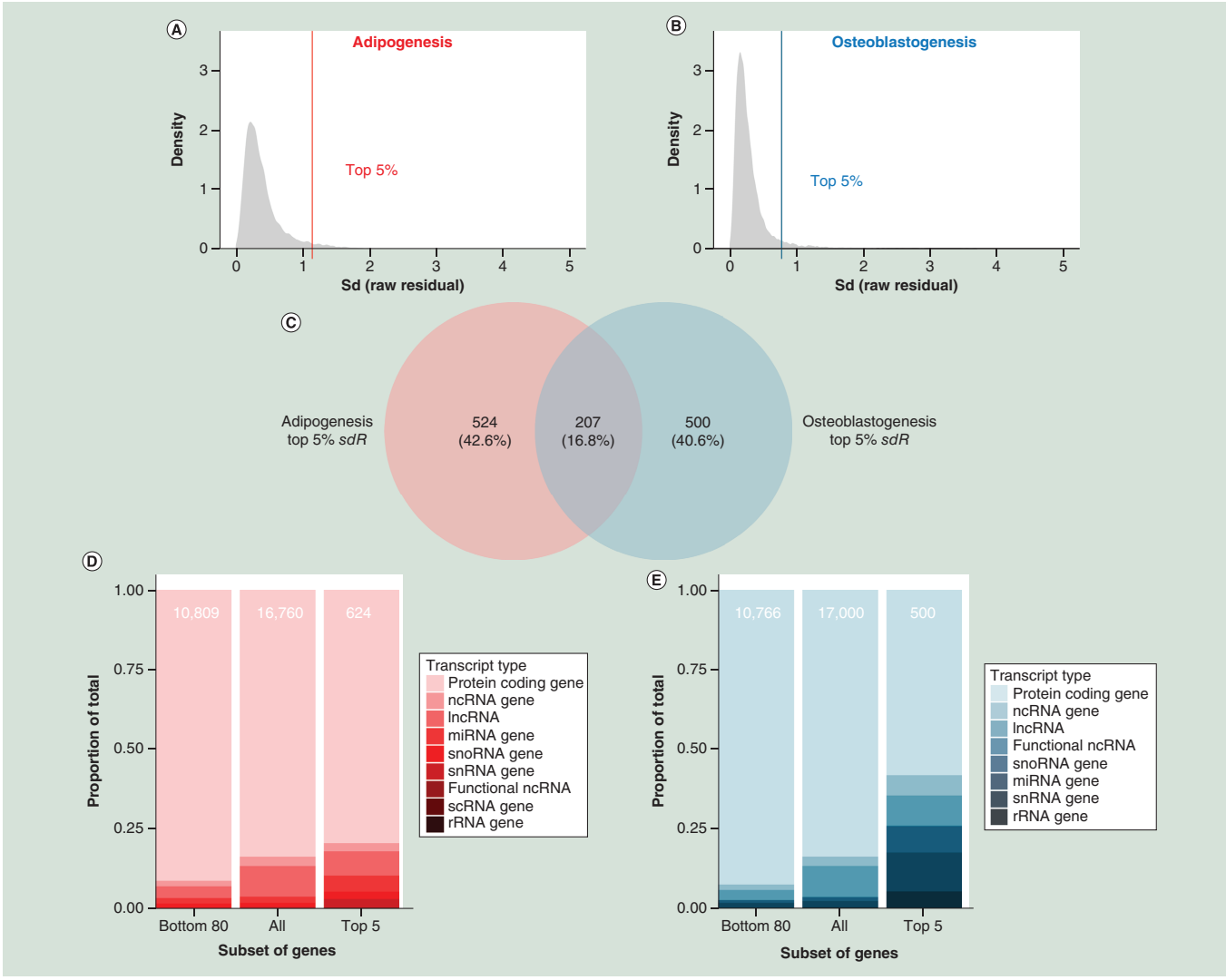


Figure 2. Use of *sdR* to subset for potential targets of post-transcriptional regulation in two differentiation time courses. (A & B) To subset for likely targets of dynamic post-transcriptional regulation the genes with the highest 5% *sdR* values were subset for both time course datasets. Shown are density distributions of *sdR* values within the genes that reach above 0 $\log_2$ FPKM at least once in of the two time courses. **(C)** Venn analysis of the resulting gene lists exhibited a relatively small overlap and for both **(D)** the adipocytes and **(E)** the osteoblasts elevated proportions of noncoding transcript types when compared with the entire set of transcripts expressed in the experimental set. For additional insight, the group of transcripts associated with the 80% lowest *sdR* values is plotted indicating a depletion of noncoding transcript types beyond that of all expressed genes. The transcript type information was retrieved from MouseMine [67]. **(F–H)** Different GO terms of the types of biological process that were enriched among **(F)** adipogenesis- and **(G)** osteoblastogenesis-specific as well as **(H)** lineage-overlapping predicted targets of post-transcriptional regulation.
FPKM: Fragments per kilobase of transcript per million; GO: Gene ontology.

similar between the gene groups, interestingly the H3K36me3 signal was typically lower for the high *sdR* genes compared with all genes. This is consistent with the observation that post-transcriptional regulators like miRNAs typically target genes expressed at intermediate rather than high levels [42].

The obtained gene lists of top 5% transcripts show a relatively low level of overlap (approximately 17% of all the top genes) between the two lineages, indicating that a large portion of the most dynamic post-transcriptional regulation is lineage specific (Figure 2C). A clear majority of all expressed transcripts in both lineages are protein coding genes with only around 15% of transcripts corresponding to other RNA species such as long ncRNAs, miRNAs, and small nuclear and nucleolar RNAs (snRNAs and snoRNAs; Figure 2D–E). Moreover, this proportion is further decreased to just a few percent of transcripts among a control set of bottom 5% of transcripts with lowest *sdR* values (Supplementary Figure 6). However, among the top transcripts with high *sdR* values the proportion

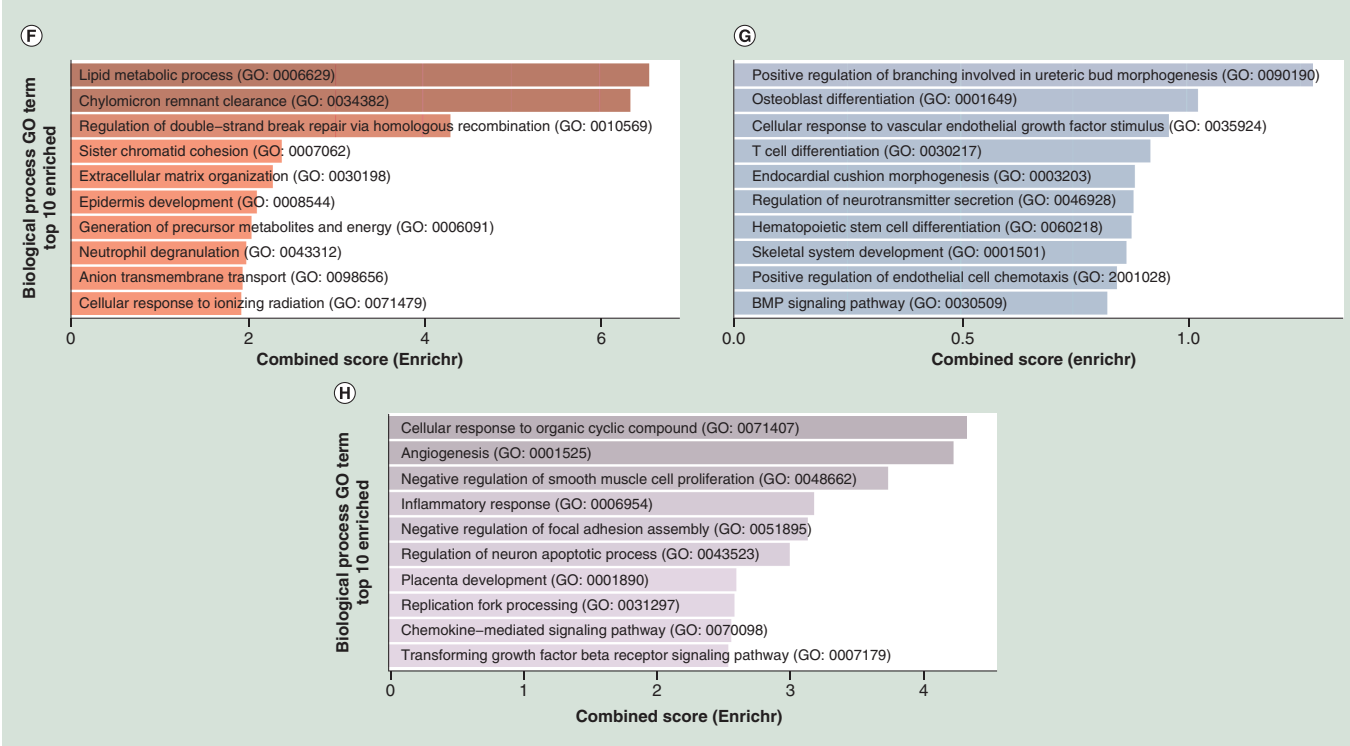


Figure 2. Use of *sdR* to subset for potential targets of post-transcriptional regulation in two differentiation time courses (cont.). (A & B) To subset for likely targets of dynamic post-transcriptional regulation the genes with the highest 5% *sdR* values were subset for both time course datasets. Shown are density distributions of *sdR* values within the genes that reach above 0 $\log_2$ FPKM at least once in of the two time courses. **(C)** Venn analysis of the resulting gene lists exhibited a relatively small overlap and for both **(D)** the adipocytes and **(E)** the osteoblasts elevated proportions of noncoding transcript types when compared with the entire set of transcripts expressed in the experimental set. For additional insight, the group of transcripts associated with the 80% lowest *sdR* values is plotted indicating a depletion of noncoding transcript types beyond that of all expressed genes. The transcript type information was retrieved from MouseMine [67]. **(F–H)** Different GO terms of the types of biological process that were enriched among **(F)** adipogenesis- and **(G)** osteoblastogenesis-specific as well as **(H)** lineage-overlapping predicted targets of post-transcriptional regulation.
FPKM: Fragments per kilobase of transcript per million; GO: Gene ontology.

of the different ncRNAs was increased to as high as 20 and 40% in the adipocytes and osteoblasts, respectively (Figure 2D–E). This finding is consistent with the fact that the different ncRNAs are typically derived from their precursor transcripts through multistep processing that often leads to fast turnover and lower abundance of the initial transcript compared to what could be assumed from the transcription rates. Moreover, many ncRNA molecules such as miRNAs are incorporated into protein complexes that significantly increase their half-life.

Next, we tested whether the top 5% and bottom 5% transcripts exhibited differences in their 3′-UTR length or number of miRNA binding sites within their 3′-UTRs compared with all transcripts on average. As shown in Supplementary Figure 7A and B, in both lineages the top 5% transcripts harbored shorter 3′-UTRs with less miRNA binding sites while bottom 5% transcripts showed the opposite behavior with longer 3′-UTRs, compared with all transcripts on average. This is most likely reflecting the proportion of ncRNAs in respective groups. Moreover, the transcripts within the bottom 5%, albeit not predicted to be dynamically regulated in adipo- or osteoblastogenesis, still include many known targets of post-transcriptional regulation in other biological contexts, further contributing to the result.

Next, we asked whether the genes under dynamic post-transcriptional control would be enriched for specific biological processes. Indeed, the transcripts with altered stability during adipogenesis were enriched for processes such as lipid metabolism and chylomicron clearance that are relevant for adipocytes (Figure 2F). The processes enriched in osteoblastogenesis included various differentiation and developmental processes consistent with the on-going lineage-specification, including osteoblast differentiation and skeletal system development (Figure 2G). Notable among the biological processes over-represented among the predicted targets of post-transcriptional regulation that are shared between the two lineages, is the negative regulation of smooth muscle cell proliferation, which

represents a third alternative cellular destination of the mesenchymal differentiation (Figure 2H). In contrast, the bottom 5% of transcripts with lowest *sdR* values were enriched for many housekeeping functions such as rRNA and tRNA transcription (Supplementary Figure 6F–H).

Clustering of transcripts according to their stability dynamics

miRNAs typically target multiple mRNAs in parallel with a typical conserved miRNA seed sequence predicted to recognize hundreds of targets in mammals [43]. In order to allow identification of specific miRNAs imposing particular dynamics on their targets, we clustered the most dynamic mRNAs (top 5%) from our regression models according to their stability dynamics, as depicted by their *sdR* values (Figure 3A & B, see materials and methods for details). The clustering identified nine significant dynamic profiles for adipogenesis and five profiles for osteoblastogenesis, with each cluster containing between 31 and 67 transcripts. The profiles ranged from steady or late increase in stability (profiles 39 and 22, respectively) to rapid but transient drop on day 1 of differentiation (for example profiles 33 and 44). Importantly, the observed dynamics across the profiles were largely due to changes in the transcript levels (Supplementary Figure 8A), as expected for changes in transcript stability, while no major changes in the H3K4me3 or H3K36me3 levels could be observed for most of the high *sdR* genes at the respective loci (Supplementary Figure 8B & C).

For each profile, we extracted the 3′-UTRs of the putative target mRNAs and performed an enrichment analysis of all possible heptamer sequences within each putatively coregulated group. The enriched heptamer motifs (posterior probability > 0.9) matching a reverse complement of known miRNA seed regions, and the corresponding mature miRNAs, are listed next to each profile in Figure 3. Interestingly, some of the miRNAs identified for the different profiles of adipogenesis have already been linked to regulation of adipocyte differentiation, including miR-466 and miR-204 [44]. Thus, it might be possible to identify novel targets of post-transcriptional regulation, and their regulators, through our linear regression model analysis.

miR-204/211 family miRNAs are induced early in adipocyte differentiation

A striking feature of several of the significant clusters identified in adipocyte differentiation is an early strong decrease in stability already by day 1 of differentiation (Figure 3A). One of the heptamers enriched in the 3′-UTRs of genes with such behavior in profile 44 is the sequence AAGGGAA, a reverse complement of the seed sequence of miR-204 that has already been linked to control of *Runx2* expression during mesenchymal commitment toward adipocytes. To confirm whether miR-204 is indeed induced already after day 1 of adipogenesis and could decrease the stability of its targets specifically in adipogenesis, we used RT-qPCR to measure the mature miR-204, and the related miR-211, in early adipocytes and osteoblasts (Figure 4A). Consistently with the prediction, both miRNAs were undetectable in the undifferentiated cells but became induced and expressed already on the day 1 of adipogenesis, while no reproducible expression could be observed in day 1 osteoblasts. Consistently, miR-204 remained undetectable and miR-211 barely detectable also in differentiated day 9 osteoblasts while both miRNAs continued to be expressed at >100-fold higher levels in adipocytes. Thus, miR-204/211 family miRNAs are early induced adipocyte-specific regulators and likely to contribute to the rapid decrease of their target mRNA stabilities upon lineage commitment.

Akr1c14 & *Il1rl1* are novel putative miR-204 targets with sharp decrease in stability upon adipocyte differentiation

Given the sudden induction of miR-204 and miR-211 upon adipocyte differentiation and the enrichment of their target sequence within the 3′-UTRs of the transcripts of profile 44, that show sharp decrease in stability in parallel, we aimed to investigate if any of the predicted targets could indeed be repressed by miR-204. From the seven transcripts carrying miR-204 target sequence within profile 44, three genes (*Akr1c14*, *Trim12c* and *Il1rl1*) are harboring multiple putative miR-204 binding sites in their 3′-UTRs, making them the targets with most potential. Parallel inspection of the RNA-seq measured expression profiles and the predicted expression levels derived by the linear regression model from the ChIP-seq measurements of adipogenesis confirms the deviation between the measured and predicted dynamics for all three genes (Figure 4B). All three genes show high expression in the undifferentiated D0 cells, followed by a sharp decrease within the first 24 h of differentiation that is not supported to the same extent by the model predictions, indicating that the downregulation involves a post-transcriptional component. For both *Akr1c14* and *Il1rl1,* the downregulation is between ten and 30-fold while *Trim12c* decreases

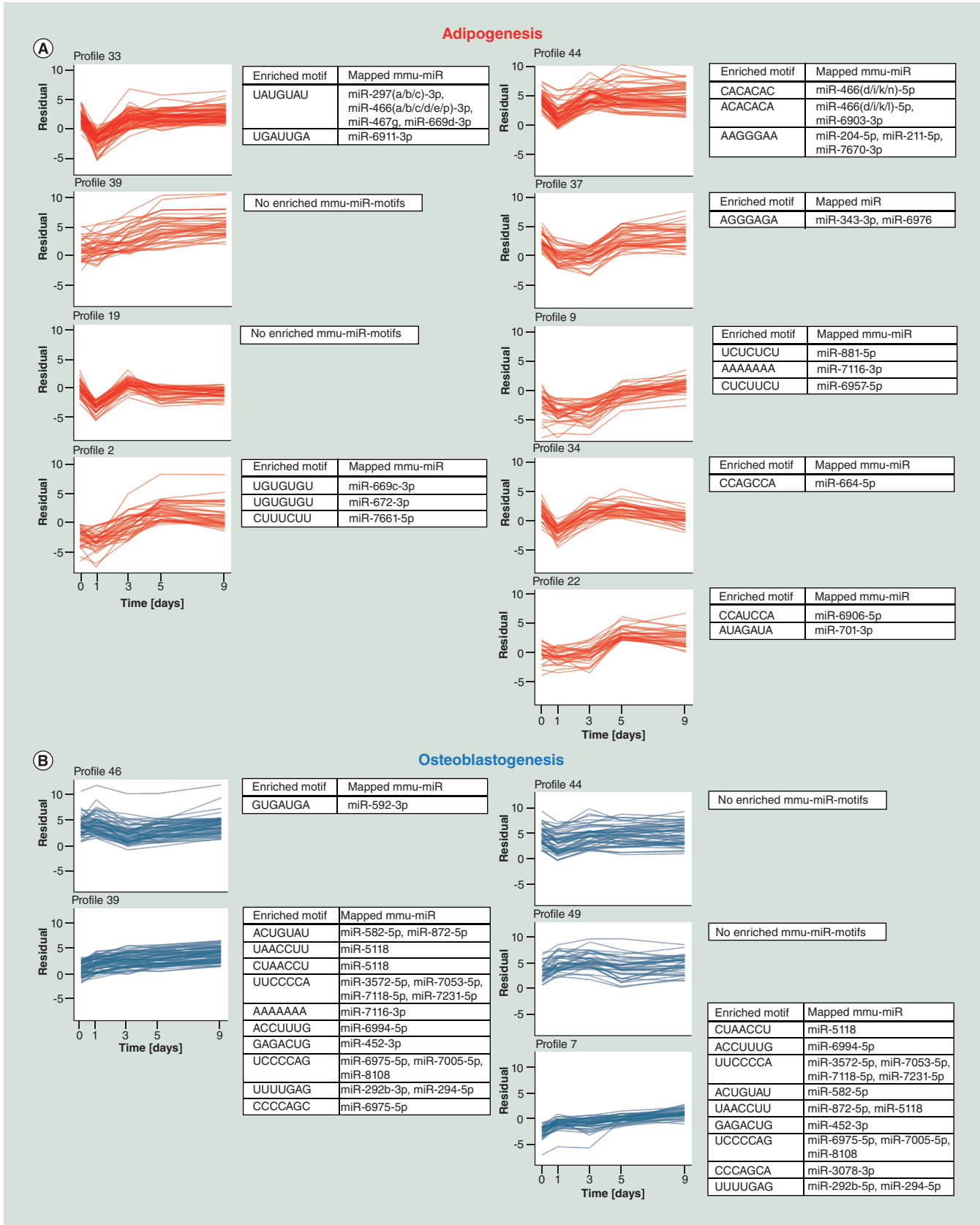


Figure 3. Clustering genes by temporal profile of fit deviation for identification of putative targeting microRNAs. (A & B) The lineage-wise predicted targets of dynamic post-transcriptional regulation (the subsets with top 5% *sdR* values) were clustered by similarity in their temporal profile of model fit deviation (their residual values) using STEM [36]. Each resulting cluster of mRNAs from **(A)** adipogenesis and **(B)** osteoblastogenesis was tested for enrichment of heptamer motifs in associated 3′-UTR sequences. Any enriched motifs were mapped to seed sequences of microRNAs to identify potential common mediators of post-transcriptional regulation.

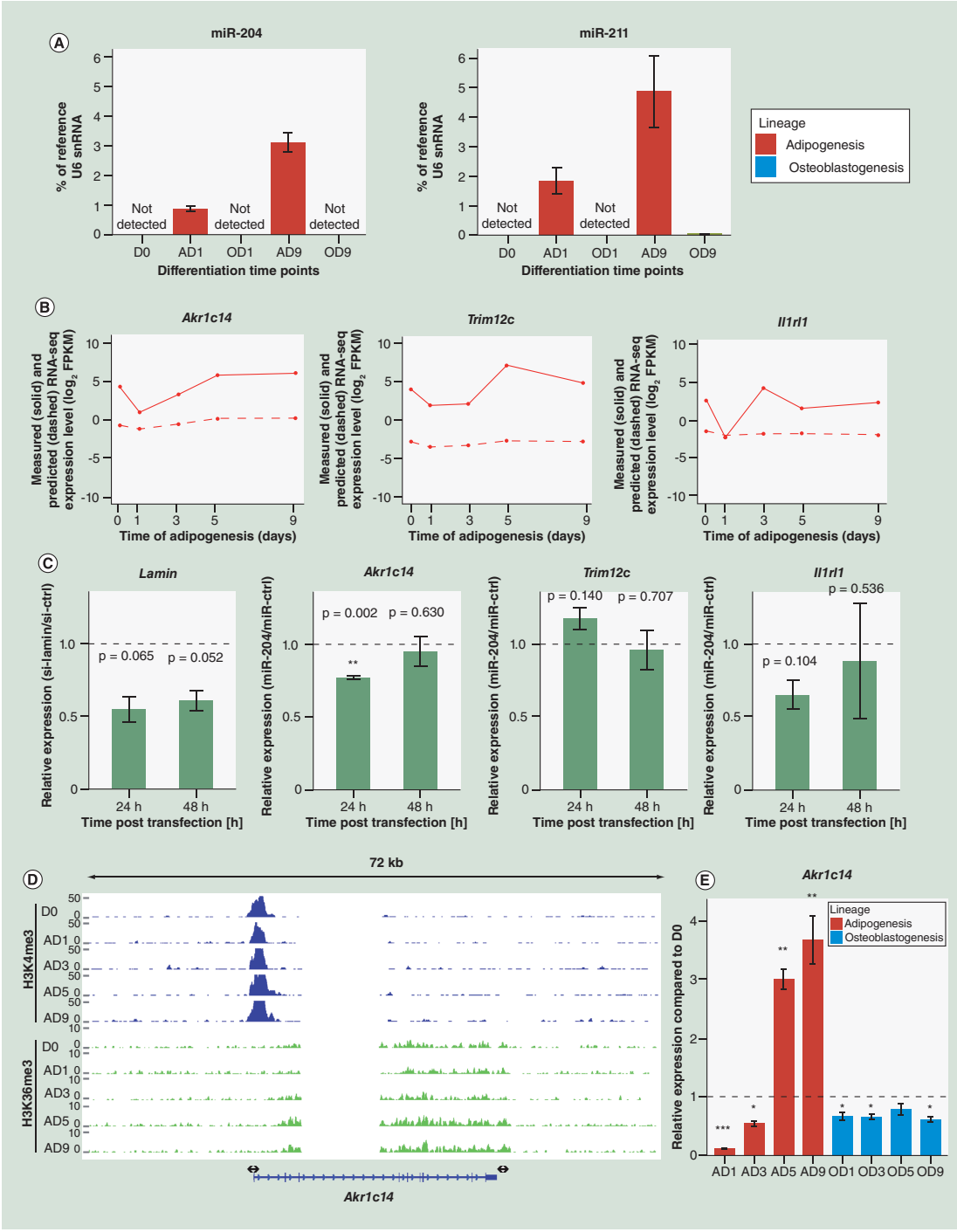


Figure 4. Exploration of miR-204 as a post-transcriptional regulator in adipogenesis. (A) Both mature miR-204 and miR-211 are more abundant after just day 1 of adipogenesis and reach even higher levels by day 9. Plotted on the y-axis is the miRNAs' expression as percentage of the control U6 snRNA at indicated differentiation time points including the undifferentiated state as measured by RT-qPCR. The values represent the mean of three biological replicates +/- standard error of mean. **(B)** The measured and model-predicted RNA-seq values (log_2 FPKM) over time of adipogenesis are displayed for the genes *Akr1c14*, *Trim12c* and *Il1rl1* **(C)** Transcript levels of the same three genes following miR-204 mimic transfection of undifferentiated ST2 cells. *Lamin* transcript quantification after si-*Lamin* transfection was included as a control for transfection conditions. Plotted on the y-axis are the *Rpl13a*-normalized expression fold changes at 24 and 48 h post-transfection relative to siCtrl or miR-Ctrl. The values represent the mean of three biological replicates +/- standard error of mean. **p-value $<$ 0.01 (one sample t-test). *Rpl13a* relative expression change upon transfection conditions is displayed in Supplementary Figure 11. **(D)** Histone modifications H3K4me3 and H3K36me3 at the genomic locus of the *Akr1c14* gene at different time points of adipocyte differentiation. **(E)** *Akr1c14* transcript abundance during adipogenic and osteoblastogenic differentiation as measured by RT-qPCR. Plotted on the y-axis are the expression fold changes at indicated time points of differentiation relative to the undifferentiated state. Red bars and blue bars indicate adipocyte and osteoblast differentiation, respectively. The values represent the mean of three biological replicates +/- standard error of mean.
*p-value $<$ 0.05, **p-value $<$ 0.01, ***p-value $<$ 0.001 (one sample t-test).
FPKM: Fragments per kilobase of transcript per million; RNA-seq: RNA sequencing.

more moderately by approximately fourfold (Figure 4B). Moreover, while *Akr1c14* and *Il1rl1* expression levels recover already by D3 of differentiation, *Trim12c* follows different dynamics recovering only after D3.

To directly test whether increased miR-204 levels could repress these target mRNAs, we transfected miR-204 mimics into the undifferentiated D0 cells and collected RNA 24 and 48 h post-transfection. Transfection efficiency was controlled by a parallel transfection of si-*Lamin* and, as shown in Figure 4C, led to an approximately 50% decrease in *Lamin* mRNA levels. Importantly, the transfection of miR-204 led to a reproducible and significant decrease in *Akr1c14* mRNA levels at 24 h time point when compared with the cells similarly transfected with an unspecific control mimic, followed by a gradual recovery by 48 h (Figure 4C). Similarly, also *Il1rl1* mRNA responded to miR-204 transfection at 24 h time point, albeit with a higher p-value, indicating that *Akr1c14* and *Il1rl1* can indeed be repressed by increased levels of miR-204. Curiously, there was no significant effect of miR-204 on *Trim12c* levels, despite it being the target mRNA with highest number of putative binding sites in its 3′-UTR.

Based on the above results, we propose *Akr1c14* and *Il1rl1* to be putative novel targets of miR-204 in early adipogenesis and that their rapid decrease upon differentiation is enhanced by destabilization triggered by the miRNA. Indeed, the tenfold and 30-fold reductions in *Akr1c14* and *Il1rl1* mRNA levels at day 1 of adipogenesis were also validated by RT-qPCR but were not supported by the inspection of H3K4me3 and H3K36me3 signals obtained at the locus by ChIP-seq, indicating a limited contribution from transcriptional regulation (Figure 4D & E; Supplementary Figure 9). However, the approximately threefold inductions in expression levels later in the differentiation were also accompanied by a twofold increase in the signal of H3K36me3 for *Akr1c14*, and to a lesser extent in H3K4me3 for *Il1rl1*, suggesting that this eventual upregulation is mainly driven by increased transcription. The lack of major changes in the H3K4me3 and H3K36me3 signals at the *Akr1c14* TSS and TES were also validated by ChIP-qPCR using biological replicates (Supplementary Figure 10A & B). Using the same samples, *Spon2*, a gene showing a similar expression profile to *Akr1c14,* but predicted not to be under post-transcriptional regulation, could be validated to show a clear decrease and subsequent increase also in the enrichment of H3K4me3 and H3K36me3, in parallel with the changes in the mRNA levels (Supplementary Figure 10B & D). Suggesting that the lack of changes in these histone modifications at genes with dynamic expression might be unique to genes with high *sdR* values, consistent with their putative role as targets of post-transcriptional control.

The increased transcription of both *Akr1c14* and *Il1rl1* after D3 is also accompanied by an increase in the predicted mRNA stability, despite the remaining high levels of miR-204 and miR-211. This could be due to other factors counteracting the destabilizing effect of the miRNAs, or, as suggested by the behavior of other mRNAs of the profile 44, due to a collective increase in the overall target pool of these miRNAs, thereby dispersing the miRNAs to more targets and diluting their impact on individual targets.

Discussion

Efforts to describe mRNA decay rates across species have identified a broad distribution of mRNA half-lives, suggesting that majority of transcripts are under some level of post-transcriptional control [45]. Consistently, majority of human mRNAs carry conserved miRNA binding sites in their 3′-UTRs [43]. However, such regulation is often context specific, changing during cell state transitions and having a limited impact at the genome-wide level (predicted to explain <12% of overall mRNA levels) [9].

Previous work has already shown that predicting gene expression levels from chromatin level data with high accuracy is possible using linear regression, other statistical methods, as well as more advanced approaches such as deep learning [7–10,46,47]. Moreover, these studies have highlighted that the deviation of the measured gene expression levels from those predicted by the models from chromatin data can be used to infer transcript stabilities, and thereby learn whether they are under post-transcriptional control. This is further supported by the presence of longer 3′-UTRs with more miRNA binding sites within those transcripts that are not well predictable by the models. However, even with improved models, it remains difficult to use this information to directly identify specific regulator-target mRNA interactions, and their dynamics during biological processes such as cellular differentiation, given the sheer number of identified transcripts and the different putative regulators. In fact, for our goal of identifying RNAs under dynamic post-transcriptional control, a model that captures only well-predictable genes and allows deviation between measured and predicted values is preferred.

We have used linear regression as a simple but robust method that can provide good predictive power based on only a few well-selected histone modifications to predict gene expression levels. The analysis was performed across mesenchymal differentiation into two lineages from shared precursor cells with five time points separated by one or more days from each lineage and focused on gene loci with open TSSs. While the used RNA-seq data are based

on three biological replicates, the ChIP-seq data were generated from one sample at each time point. However, the gene expression and the histone modification signals across the time courses showed good correlation for most genes and behavior of individual loci was also validated using ChIP-qPCR from several biological replicates, suggesting that the quality of the individual ChIP-seq samples is overall comparable. This is also supported by the metagene analysis in Supplementary Figure 5.

By generating a model for each time point separately, and thereby deriving a residual value for each gene per each time point, we were able to focus on the deviation of those residuals (*sdR*) as an estimate of RNA stability across time (Figure 1). Focusing on RNAs that change in their estimated stability over time, in other words, are likely to be under dynamic post-transcriptional regulation, allows to significantly narrow the list of transcripts of interest. Moreover, clustering of these transcripts according to their stability dynamics enables identification of potentially coregulated gene sets that can be matched to their putative regulators through enrichment analysis, combined for example with co-expression analysis (in cases where also regulator levels are altered). Our approach is well suited for the identification of miRNA-mediated regulation as miRNAs mostly affect their targets through mRNA destabilization [48], miRNA target recognition is relatively well defined [49,50], and miRNA activity is largely controlled by their level of expression to control processes like cell state transitions [51]. However, although not specifically addressed in this manuscript, our approach could equally help to identify important RNA-binding proteins and elucidate their influence on the target mRNA stabilities over time.

Interestingly, the transcripts that showed changes in their stability were enriched for different ncRNA species in both adipocytes and osteoblasts (Figure 2). Their deviation from predicted expression levels at a given time point is readily explained by the various maturation processes the different RNA species go through, as well as the typically increased stabilities introduced by their incorporation into different RNA–protein complexes, when compared with the protein-coding mRNAs that mainly serve as the basis of used model. Thus, for long ncRNAs that are incorporated as structural molecules in different ribonucleoprotein complexes, the stabilities are likely to be underestimated from chromatin data. On contrary, different small RNA species, such as miRNAs, could not be detected in their mature form by regular RNA-seq and are instead sequenced at the level of their primary miRNAs (pri-miRs) that are processed cotranscriptionally and exhibit very fast turn-over [52], leading to an overestimation of the stability of these unprocessed transcripts. However, identification of ncRNAs as an enriched group among the genes with most changes in their stability suggests that our approach might allow detection of altered RNA processing event during differentiation. Conceptually similar regulation of RNA processing has been described for example during differentiation of ESCs that are transcribing both of the two miRNA genes giving rise to let-7 and miR-17–92 cluster, respectively, in their pluripotent state. However, upon ESC differentiation the mature let-7 increases and mature miRNAs of the miR-17∼92 cluster decrease without corresponding changes in transcription, due to cell state-dependent post-transcriptional control of the processing of the immature precursor molecules [53,54]. Future work might reveal similar regulatory mechanisms also in other cell types such as adipocytes and osteoblasts.

Our enrichment analysis for shared miRNA binding sites within 3′-UTRs of genes with similar stability dynamics upon adipo- and osteoblastogenesis revealed in total more than 20 enriched heptamer motifs, ranging from no enriched motifs for some stability profiles up to ten different enriched motifs for the transcripts with gradually increasing stability during osteoblastogenesis (Figure 3). All of the motifs could be matched to reverse complement of at least one known mouse miRNA seed sequence. None of the miRNA motifs enriched in osteoblastogenesis profiles were previously directly associated with bone formation or differentiation. However, among the miRNA motifs enriched in osteoblastogenesis profiles 46 and 39 were miRNAs miR-592-3p and miR-452-3p, respectively (Figure 3). Interestingly, miR-592 expression has previously been shown to depend on PPARG, the master regulator of adipogenesis known to be downregulated in osteoblastogenesis [55]. Moreover, the ascending strand of the miRNA duplex, miR-592-5p, has been shown to target *Sox9*, a marker of osteo-progenitors [56,57]. Similarly, miR-452-3p, reverse complement of which is enriched in the group of upregulated transcripts with high *sdR* in profile 39 was recently shown to target *Gsk3b* [58], inhibition of which is relevant for stimulation of osteoblastogenesis via BMP and Wnt signaling [59].

The differences in dynamics between the two lineages is likely to be related to the used differentiation protocol. While adipocyte differentiation is induced by several compounds on D0 of differentiation, followed by a change to a different combination of compounds from day 2 onward, the osteoblast differentiation relies on only one protein throughout the differentiation (see methods for details). The most intriguing patterns of stability changes were exhibited by many of the transcripts in adipogenesis where they first showed sharp decrease in stability which followed by different dynamics of recovery. In two of such profiles of adipogenesis with comparable dynamics

10.2217/epi-2018-0084

(profiles 33 and 44), the heptamer motifs corresponding to miR-466 family of miRNAs were found to be enriched. Interestingly, in profile 33, the enrichments correspond to seed sequence of the miRNA from the descending strand (3p) of hairpin while those from profile 44 corresponded to the ascending strand (5p). This suggests not only that miR-466 family is likely to play a role in the observed regulation, but also that both strands of this family might be loaded in the RISC-like complex and be functional in early adipogenesis. Whether the strand selection depends on family member and which family members are active in adipocytes remains to be determined. The miR-466 family miRNAs are expressed from multiple different loci in the mouse genome and several of which are transcribed based on our epigenomic analysis (data not shown). Consistently, Huang *et al*. have shown that miR-466, and related miR-467, are expressed in ST2 cells and become upregulated upon adipocyte differentiation [44], consistently with our predictions. Moreover, computational analysis of miRNAs contributing to browning of adipocytes in mice has suggested miR-466 family to target many important genes for brown adipocyte biology [60].

In our analysis, we focused on miR-204 and miR-211, identified through enrichments in profile 44 that share their seed sequence and thus are from the same miRNA family. We focused on these miRNAs since miR-204 was already shown to promote adipogenesis in mouse MSCs and in human adipose-derived MSCs, and to inhibit osteoblastogenesis [44,61], suggesting this to be both a relevant and conserved regulatory mechanism. Moreover, inhibition of miR-204 was shown to inhibit adipogenesis, suggesting its induction is necessary for normal differentiation [61]. Therefore, it was interesting to find putative miR-204 targets to be enriched among transcripts early repressed in adipogenesis and we set out to test whether our predicted targets could indeed be repressed by miR-204 overexpression (Figure 4). Based on the results, we have identified *Akr1c14* and *Il1rl1* to be putative novel targets of miR-204 in early adipogenesis.

Akr1c14 is a metabolic enzyme implicated in reduction of aldehydes and ketones and has been shown to become downregulated in adipocytes in an adipose tissue inflammation model [62] and was therefore suggested to play a role in maintenance of normal adipocyte metabolism. Indeed, using a mouse model of increased adipocyte browning, Milet *et al*. showed *Akr1c14* to be reduced in selected adipocyte depots with increased browning, further suggesting a role in white adipocyte specific metabolism [63].

Also *Il1rl1* has been linked to adipocyte metabolism and differentiation. *Il1rl1* encodes for a cytokine receptor that can be bound by the cytokine IL-33, both of which are expressed in human adipocytes [64]. The protein product of *Il1rl1*, ST2, has been shown to mediate IL-33 signaling in white adipocytes and to have an inhibitory effect on lipid metabolism and adipogenesis [65]. Moreover, IL-33 signaling appears to play a protective role on adipose tissue inflammation and insulin resistance in obese mice [65,66]. Thus, both *Akr1c14* and *Il1rl1* are likely to contribute to the effects of miR-204 on adipocyte biology.

Taken together, we apply a linear regression-based model for prediction of gene expression levels from time-series epigenomic data and use the change in the deviation of predicted expression levels from measured expression (*sdR*) to identify genes under dynamic post-transcriptional regulation. This approach allows identification of both protein coding genes under classic post-transcriptional control by RBPs or miRNAs, as well as noncoding genes that could be controlled during their processing in a lineage-specific manner. Analysis of the transcripts with similar stability dynamics can be used to obtain more detailed regulator-target gene interactions and the overall approach could be applied for various biological processes involving cell state transitions.

Limitation

The obtained model fits for our linear regression analysis could be higher than currently achieved. These model fits could be improved for example by including more ChIP-seq data on additional histone modifications beyond those currently used. In particular, including repressive marks such as H3K27me3 could aid exclusion of low abundance genes. However, our data preprocessing prior to the linear regression analysis was planned to be very stringent on excluding nonexpressed and very low expressed genes (Supplementary Figure 1). Indeed, >23,000 detected transcripts were removed as a consequence of the data preprocessing, leaving only the most abundant 47% of transcripts to be used as input in the linear regression model. Moreover, in the data preprocessing, we used also the signal for H3K27 acetylation, occurring on the same residue as H3K27me3, at gene TSSs at each time point of differentiation, thereby already excluding many of the H3K27me3 marked genes. In addition, the model is considering the H3K36me3 signal, deposition of which is coupled with RNA Polymerase II activity and therefore present only at active genes devoid of repressive marks such as H3K27me3.

While additional data are likely to improve the model fits, performing ChIP-seq analysis at several time points over a time course of a biological process for multitude of histone modifications can be laborious and costly.

Therefore, we are here aiming to provide an application relying on a minimal number of two histone modifications that can still provide useful predictions.

Future perspective

Integration of different omics data types will be increasingly important to understand complex biological processes and to harvest the full potential of these data sets. Here, we have used such an integrative approach to study the dynamic post-transcriptional regulation in two mesenchymal differentiation processes using the data on histone modifications H3K4me3 and H3K36me3, which are mainly studied in the current manuscript at hand. In the related work by Gerard *et al.* [30], we also focus on the analysis of transcriptional regulation and the enhancer regions identified using the H3K27ac mark in the same differentiation processes. Only by bringing together both the post-transcriptional and the transcriptional layers of regulation can a full overview of the gene regulatory control of cellular differentiation be achieved. In current work, we present an example of how the integration of the epigenomic and transcriptomic data can advise the understanding of both transcriptional and post-transcriptional regulation and future work will be needed to further combine the two in an informative manner.

Summary points

- Combining time point-specific linear regression models predicting RNA expression levels from time-series epigenomic data.
- Identification of putative targets of dynamic post-transcriptional control by use of gene-specific *sdR* values from the linear regression models.
- Genes with greatest changes in the deviation over time from predicted expression levels are enriched for noncoding RNAs and lineage specific functions.
- miR-204 and miR-211 are early induced adipogenic miRNAs.
- *Akr1c14* and *Il1rl1* are putative targets of miR-204 in adipogenesis.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/full/10.2217/epi-2018-0084

Acknowledgments

We would like to thank M Bouvy-Liivrand for help with establishing the ST2 cell culture and differentiation, P Berninger from the University of Basel who had created and kindly made available his script for motif enrichment analysis and EMBL Gene Core at Heidelberg for support with high-throughput sequencing. The analyses presented in this paper were carried out using the HPC facilities of the University of Luxembourg.

Financial & competing interests disclosure

This work was supported by funding from the University of Luxembourg and FNR INTER project (grant number INTER/ANR/15/11191283). D Gérard was supported by fellowship from the National Research Foundation of Luxembourg (FNR) (AFR 7924045). The authors have no other relevant affiliations or financialinvolvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Schubeler D, Macalpine DM, Scalzo D *et al.* The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* 18(11), 1263–1271 (2004).
2. Mikkelsen TS, Ku M, Jaffe DB *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448(7153), 553–560 (2007).

3. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 130(1), 77–88 (2007).
4. Heintzman ND, Stuart RK, Hon G *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39(3), 311–318 (2007).
5. Ernst J, Kheradpour P, Mikkelsen TS *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345), 43–49 (2011).
6. Schmidt F, Gasparoni N, Gasparoni G *et al.* Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.* 45(1), 54–66 (2017).

• **Demonstrates the ability to predict gene expression levels directly from open chromatin data with transcription factor binding affinity predictions.**

7. Singh R, Lanchantin J, Robins G, Qi Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 32(17), i639–i648 (2016).

• **Comparative analysis of the different methods for predicting gene expression and a thorough introduction to different approaches.**

8. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA* 107(7), 2926–2931 (2010).
9. Tippmann SC, Ivanek R, Gaidatzis D *et al.* Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. *Mol. Syst. Biol.* 8, 593 (2012).

•• **A thorough analysis of linear regression for prediction on gene expression from histone modification chromatin immunoprecipitation-seq data.**

10. Wang C, Tian R, Zhao Q *et al.* Computational inference of mRNA stability from histone modification and transcriptome profiles. *Nucleic Acids Res.* 40(14), 6414–6423 (2012).

•• **Demonstration of mRNA stability prediction from histone modification chromatin immunoprecipitation-seq data with linear regression.**

11. Benayoun BA, Pollina EA, Ucar D *et al.* H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* 158(3), 673–688 (2014).
12. Perez-Lluch S, Blanco E, Tilgner H *et al.* Absence of canonical marks of active chromatin in developmentally regulated genes. *Nat. Genet.* 47(10), 1158–1167 (2015).
13. Faure AJ, Schmiedel JM, Lehner B. Systematic analysis of the determinants of gene expression noise in embryonic stem cells. *Cell Syst.* 5(5), 471–484.e474 (2017).
14. Edfors F, Danielsson F, Hallstrom BM *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* 12(10), 883 (2016).
15. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. *Nature* 455(7209), 58–63 (2008).
16. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466(7308), 835–840 (2010).
17. Bernstein E, Kim SY, Carmell MA *et al.* Dicer is essential for mouse development. *Nat. Genet.* 35(3), 215–217 (2003).
18. Yang WJ, Yang DD, Na S, Sandusky GE, Zhang Q, Zhao G. Dicer is required for embryonic angiogenesis during mouse development. *J. Biol. Chem.* 280(10), 9330–9335 (2005).
19. Kanellopoulou C, Muljo SA, Kung AL *et al.* Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev.* 19(4), 489–501 (2005).
20. Murchison EP, Partridge JF, Tam OH, Cheloufi S, Hannon GJ. Characterization of Dicer-deficient murine embryonic stem cells. *Proc. Natl Acad. Sci. USA* 102(34), 12135–12140 (2005).
21. Wang Y, Medvid R, Melton C, Jaenisch R, Blelloch R. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat. Genet.* 39(3), 380–385 (2007).
22. Sinkkonen L, Hugenschmidt T, Berninger P *et al.* MicroRNAs control *de novo* DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.* 15(3), 259–267 (2008).
23. Wang Q, Li YC, Wang J *et al.* miR-17-92 cluster accelerates adipocyte differentiation by negatively regulating tumor-suppressor Rb2/p130. *Proc. Natl Acad. Sci. USA*105(8), 2889–2894 (2008).
24. Mudhasani R, Puri V, Hoover K, Czech MP, Imbalzano AN, Jones SN. Dicer is required for the formation of white but not brown adipose tissue. *J. Cell. Physiol.* 226(5), 1399–1406 (2011).
25. Lian JB, Stein GS, Van Wijnen AJ *et al.* MicroRNA control of bone formation and homeostasis. *Nat. Rev. Endocrinol.* 8(4), 212–227 (2012).
26. Ebert MS, Sharp PA. Roles for microRNAs in conferring robustness to biological processes. *Cell* 149(3), 515–524 (2012).

27. Posadas DM, Carthew RW. MicroRNAs and their roles in developmental canalization. *Curr. Opin. Genet. Dev.* 27, 1–6 (2014).
28. T Hoen PA, Hirsch M, De Meijer EJ, De Menezes RX, Van Ommen GJ, Den Dunnen JT. mRNA degradation controls differentiation state-dependent differences in transcript and splice variant abundance. *Nucleic Acids Res.* 39(2), 556–566 (2011).
29. Houseley J, Tollervey D. The many pathways of RNA degradation. *Cell* 136(4), 763–776 (2009).
30. Gerard D, Schmidt F, Ginolhac A *et al.* Temporal enhancer profiling of parallel lineages identifies AHR and GLIS1 as regulators of mesenchymal multipotency. *Nucleic Acids Res.* doi:10.1093/nar/gky1240 (2018) (Epub ahead of print).
•• **Integrative time series analysis of transcriptional regulation from epigenomic data used also in the manuscript at hand.**
31. Ogawa M, Nishikawa S, Ikuta K *et al.* B cell ontogeny in murine embryo studied by a culture system with the monolayer of a stromal cell clone, ST2: B cell progenitor develops first in the embryonal body rather than in the yolk sac. *EMBO J.* 7(5), 1337–1343 (1988).
32. Trapnell C, Williams BA, Pertea G *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28(5), 511–515 (2010).
33. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* 12(3), R22 (2011).
34. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 27(17), 2325–2329 (2011).
35. Wickham H. ggplot2. Springer-Verlag NY, USA, doi:10.1007/978-0-387-98141–3 (2009) (Epub ahead of print).
36. Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinf.* 7, 191 (2006).
37. Ma J, Flemr M, Stein P *et al.* MicroRNA activity is suppressed in mouse oocytes. *Curr. Biol.* 20(3), 265–270 (2010).
38. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42(Database issue), D68–73 (2014).
39. Kuleshov MV, Jones MR, Rouillard AD *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44(W1), W90–97 (2016).
40. Chen EY, Tan CM, Kou Y *et al.* Enrichr: interactive and collaborative *HTML5* gene list enrichment analysis tool. *BMC Bioinf.* 14, 128 (2013).
41. Ramirez F, Ryan DP, Gruning B *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 44(W1), W160–165 (2016).
42. Iwama H, Kato K, Imachi H, Murao K, Masaki T. Human microRNAs preferentially target genes with intermediate levels of expression and its formation by mammalian evolution. *PLoS ONE* 13(5), e0198142 (2018).
43. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19(1), 92–105 (2009).
44. Huang J, Zhao L, Xing L, Chen D. MicroRNA-204 regulates Runx2 protein expression and mesenchymal progenitor cell differentiation. *Stem Cells* 28(2), 357–364 (2010).
45. Halbeisen RE, Galgano A, Scherrer T, Gerber AP. Post-transcriptional gene regulation: from genome-wide studies to principles. *Cell. Mol. Life Sci.* 65(5), 798–813 (2008).
46. Cheng C, Yan KK, Yip KY *et al.* A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol.* 12(2), R15 (2011).
47. Dong X, Greven MC, Kundaje A *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* 13(9), R53 (2012).
48. Eichhorn SW, Guo H, Mcgeary SE *et al.* mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol. Cell.* 56(1), 104–115 (2014).
49. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 136(2), 215–233 (2009).
50. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4, e05005 (2015).
51. Ivey KN, Srivastava D. MicroRNAs as regulators of differentiation and cell fate decisions. *Cell Stem Cell* 7(1), 36–41 (2010).
52. Morlando M, Ballarino M, Gromak N, Pagano F, Bozzoni I, Proudfoot NJ. Primary microRNA transcripts are processed co-transcriptionally. *Nat. Struct. Mol. Biol.* 15(9), 902–909 (2008).
53. Rybak A, Fuchs H, Smirnova L *et al.* A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. *Nat. Cell. Biol.* 10(8), 987–993 (2008).
54. Du P, Wang L, Sliz P, Gregory RI. A biogenesis step upstream of microprocessor controls miR-17 approximately 92 expression. *Cell* 162(4), 885–899 (2015).
55. Bengestrate L, Virtue S, Campbell M *et al.* Genome-wide profiling of microRNAs in adipose mesenchymal stem cell differentiation and mouse models of obesity. *PLoS ONE* 6(6), e21305 (2011).
56. Li Z, Li B, Niu L, Ge L. miR-592 functions as a tumor suppressor in human non-small cell lung cancer by targeting SOX9. *Oncol. Rep.* 37(1), 297–304 (2017).

57. Akiyama H, Kim JE, Nakashima K *et al.* Osteo-chondroprogenitor cells are derived from Sox9 expressing precursors. *Proc. Natl Acad. Sci. USA* 102(41), 14665–14670 (2005).
58. Li T, Jian X, He H *et al.* MiR-452 promotes an aggressive colorectal cancer phenotype by regulating a Wnt/beta-catenin positive feedback loop. *J. Exp. Clin. Cancer Res.* 37(1), 238 (2018).
59. Schoeman MA, Moester MJ, Oostlander AE *et al.* Inhibition of GSK3beta stimulates BMP signaling and decreases SOST expression which results in enhanced osteoblast differentiation. *J. Cell. Biochem.* 116(12), 2938–2946 (2015).
60. Arias N, Aguirre L, Fernandez-Quintela A *et al.* MicroRNAs involved in the browning process of adipocytes. *J. Physiol. Biochem.* 72(3), 509–521 (2016).
61. He H, Chen K, Wang F *et al.* miR-204-5p promotes the adipogenic differentiation of human adipose-derived mesenchymal stem cells by modulating DVL3 expression and suppressing Wnt/beta-catenin signaling. *Int. J. Mol. Med.* 35(6), 1587–1595 (2015).
62. Lin Q, Huang Y, Booth CJ *et al.* Activation of hypoxia-inducible factor-2 in adipocytes results in pathological cardiac hypertrophy. *J. Am. Heart Assoc.* 2(6), e000548 (2013).
63. Milet C, Bleher M, Allbright K *et al.* Egr1 deficiency induces browning of inguinal subcutaneous white adipose tissue in mice. *Sci. Rep.* 7(1), 16153 (2017).
64. Wood IS, Wang B, Trayhurn P. IL-33, a recently identified interleukin-1 gene family member, is expressed in human adipocytes. *Biochem. Biophys. Res. Commun.* 384(1), 105–109 (2009).
65. Miller AM, Asquith DL, Hueber AJ *et al.* Interleukin-33 induces protective effects in adipose tissue inflammation during obesity in mice. *Circ. Res.* 107(5), 650–658 (2010).
66. Han JM, Wu D, Denroche HC, Yao Y, Verchere CB, Levings MK. IL-33 reverses an obesity-induced deficit in visceral adipose tissue ST2+ T regulatory cells and ameliorates adipose tissue inflammation and insulin resistance. *J. Immunol.* 194(10), 4777–4783 (2015).
67. Motenko H, Neuhauser SB, O'Keefe M, Richardson JEAMouseMine: a new data warehouse for MGI. *Mamm. Genome.* 26(7-8), 325–330 (2015).