SCIENTIFIC REPORTS

OPEN

Received: 5 April 2017
Accepted: 7 June 2017
Published online: 27 July 2017

High resolution temporal transcriptomics of mouse embryoid body development reveals complex expression dynamics of coding and noncoding loci

Brian S. Gloss[1,2], Bethany Signal[1,2], Seth W. Cheetham[3], Franziska Gruhl[4], Dominik C. Kaczorowski[1], Andrew C. Perkins[5] & Marcel E. Dinger[1,2]

Cellular responses to stimuli are rapid and continuous and yet the vast majority of investigations of transcriptional responses during developmental transitions typically use long interval time courses; limiting the available interpretive power. Moreover, such experiments typically focus on protein-coding transcripts, ignoring the important impact of long noncoding RNAs. We therefore evaluated coding and noncoding expression dynamics at unprecedented temporal resolution (6-hourly) in differentiating mouse embryonic stem cells and report new insight into molecular processes and genome organization. We present a highly resolved differentiation cascade that exhibits coding and noncoding transcriptional alterations, transcription factor network interactions and alternative splicing events, little of which can be resolved by long-interval developmental time-courses. We describe novel short lived and cycling patterns of gene expression and dissect temporally ordered gene expression changes in response to transcription factors. We elucidate patterns in gene co-expression across the genome, describe asynchronous transcription at bidirectional promoters and functionally annotate known and novel regulatory lncRNAs. These findings highlight the complex and dynamic molecular events underlying mammalian differentiation that can only be observed though a temporally resolved time course.

Over the past decade, transcriptomic investigations into the of nature embryonic stem cell (ESC) differentiation have elucidated key biochemical features of stemness and differentiation. Increasingly, it has become apparent that understanding the dynamics and coordination of gene expression signatures over time during the key phases of differentiation is critical to adequate characterization of fundamental biological processes.

Mouse ESC differentiation is a highly complex cascade of gene expression changes that allow single pluripotent cells in culture to progress to an organoid composed of cells reflecting three germ lineages within only five days. The spontaneous differentiation of these cells in culture has provided key insights into the developmental processes underlying the generation of the primary germ cell layers[1]. Microarray and RNA sequencing have provided a means to characterize the molecular transitions in gene expression underlying ESC biology and more recently single cell transcriptomic studies have provided the first glimpses into the molecular history of these cells[2]. However, it is clear that much more of the transcriptional landscape of ESC remains to be elucidated[3].

Access to new technologies, such as massively parallel sequencing (MPS), has led to a dramatic increase in our knowledge of the mammalian transcriptome. Early genomic tiling array analysis indicated that most of the genome was transcribed into RNA[4]. MPS of the transcriptome validated this observation and revealed that the majority of the mammalian genome is pervasively transcribed as interlaced and overlapping RNAs[5], many of

[1]Garvan Institute of Medical Research, Sydney, Australia. [2]St Vincents Clinical School, Faculty of Medicine, UNSW, Sydney, Australia. [3]The Gurdon Institute and Department of Physiology, Development, and Neuroscience, University of Cambridge, Cambridge, United Kingdom. [4]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [5]Mater–UQ Research Institute, The University of Queensland, Translational Research Institute, Brisbane, Australia. Correspondence and requests for materials should be addressed to M.E.D. (email: m.dinger@garvan.org.au)

which lack protein-coding potential[6]. The large number of long-noncoding transcripts (lncRNA) has become the focus of significant interest due to their exquisite cell type specific expression[7], potent biological function[8, 9], and rapid transactivation of cellular processes[10]. However, in general, lncRNAs are lowly expressed and short lived[11], possibly because, unlike mRNAs that require translation, are able to exert their function directly. These qualities obfuscate their identification and characterization with traditional approaches that are tuned to the properties of mRNAs[12]. Owing to the relative infancy of the field, the vast majority of noncoding transcripts are of unknown function[13]. Additionally, the expression patterns of these genes imply that their function is dependent on cellular context and likely regulatory[8], thus the identification of these molecules and the context in which they act remains a research priority[14].

Various expression profiling studies, using both microarrays and RNA-seq[15–18], have been used to explore the molecular changes occurring during ES cell development, typically at 24-hourly or more. This potentially has lead to incomplete gene expression relationships through the phenomenon of temporal aggregation bias whereby each time point is assumed to represent all the signaling changes occurring in that time window[19]. In contrast to single cell based approaches- which provide insight into the state of individual cells - examinations of whole cell populations provides system-wide behavior and a practical means to explore gene expression dynamics across time. The combination of these techniques has recently shed light on the molecular framework of cellular differentiation[20]. Higher temporal resolution has also shown rapid induction (within two hours of retinoic acid stimulation) of lncRNAs associated with the HOX locus[21]. Furthermore, high temporal resolution has provided valuable insights into transcriptional annotation and regulation in drosophila[22, 23], Xenopus[24] and C.elegans[25].

Here we show that additional temporal resolution of the global transcriptome in spontaneously differentiating mESC cells following LIF withdrawal enables the capture of the rapid and complex dynamic regulatory and noncoding changes occurring during ES development. We analyzed the transcriptome of differentiating mouse ESCs at six-hourly intervals over a five-day period, over which time the three primordial germ layers are specified. Using this fine-resolution temporal sampling approach, we identify significant transitions in the transcriptome and large-scale shifts in observable transcription factor activities that could not be observed at 24 hourly sampling periods. Moreover, we identify entirely novel coding and noncoding transcripts that are expressed only within specific sub-24-hour window. By leveraging the high sampling frequency of the data, we are able to both accurately recapitulate known regulatory cascades in ES development and predict and refine others. Finally, using correlative approaches, we can infer functions for uncharacterized lncRNAs and predict the regulatory centers across the genome that coordinate early development.

Results

The dynamic transcriptome of mESC differentiation at high temporal resolution.

A median 42-million, paired-end 100-bp reads (Supplementary Fig. S1A) were mapped from stranded, poly-A derived cDNA libraries derived from biological duplicate, six-hourly time courses of mESC differentiation over five days where key differentiation programs occur (0–120 hours, Fig. 1A). Transcript-level expression data was generated as previously described[26], then normalized for library size and transformed for data visualization and differential gene expression analysis. Evaluation of 24 hourly time points indicated that our data was comparable to previously published data in a similar model[27] (Supplementary Fig. S1B). An interactive gene expression portal was created to visualise this data (https://betsig.shinyapps.io/paper_plots).

To assess the reproducibility and provide confidence in the biological validity of the global transcriptome trends, a principle components analysis (PCA) was performed on the 2,000 most variable genes (Fig. 1B). This analysis indicated that biological replicates clustered closely, indicating that synchrony was retained, and that the major contributor to the determination of variance was explained by time. Deconvolution of the dimensions yielded time-dependent expression (in the first dimension) of genes enriched in focal adhesion/ECM interactions KEGG pathways. Interestingly, the second dimension deconvolution (PC2), in which undifferentiated ESCs resemble the more differentiated embryoblast, yielded genes enriched in MAPK-signaling and cancer pathways, implying that the process of differentiation involves a partial reacquisition of a mitogenic signaling. In the third component (PC3), in which the undifferentiated ES cell is separate, the axon-guidance pathway was enriched. We then evaluated expression patterns of genes associated with pluripotency, primitive streak formation and cell specialization (Fig. 1C). We observed that, although the gene expression patterns were broadly consistent with published studies (Supplementary Fig. S1B), there were changes in expression on less than 24 hourly timeframes that could not be attributed to stochastic expression changes (within the top 5% of deviation of all genes from expression values loess-smoothed over 24 hours). Similarly, we observed that no single 24-hourly measure was representative of the average expression over that day (Mann-Whitney U p. adj. $<<<0.0001$) and that more than 1,000 genes displayed a more than 2-fold difference mostly in the first 24 hours of differentiation (Supplementary Fig. 1C,D). Therefore we conclude that 24-houly expression profiles are unable to capture the intervening expression changes and that 6-hourly measurements reduce the phenomenon of temporal aggregation bias[19], providing enhanced resolution of transcriptional changes in this system.

To evaluate characteristics of sub-24 hour gene expression patterns in in the transcriptome of developing ESCs, we observed that, compared to 24 hour time points, 417 more genes had counts data considered sufficient for differential gene expression analysis (>1 CPM in at least two samples); this was associated with a relative increase in detected noncoding genes (13% (588 vs. 520); defined as antisense, lincRNA and processed transcript biotypes) over protein coding, (2% (13336 vs. 13036) despite being underrepresented in the total pool (chi-squared p value < 0.001, Supplementary Table S1 and Supplementary Fig. S1E)). The additional time points allowed the assembly of 58% more novel multiexonic intergenic, antisense and intronic noncoding RNAs from the data - indicating that a substantial proportion of noncoding transcripts are present on timescales much shorter than 24 hours. These results indicate that enhanced temporal resolution reduces the phenomenon of temporal aggregation bias and allows the observation of more distinct cell expression states than typical time-courses.

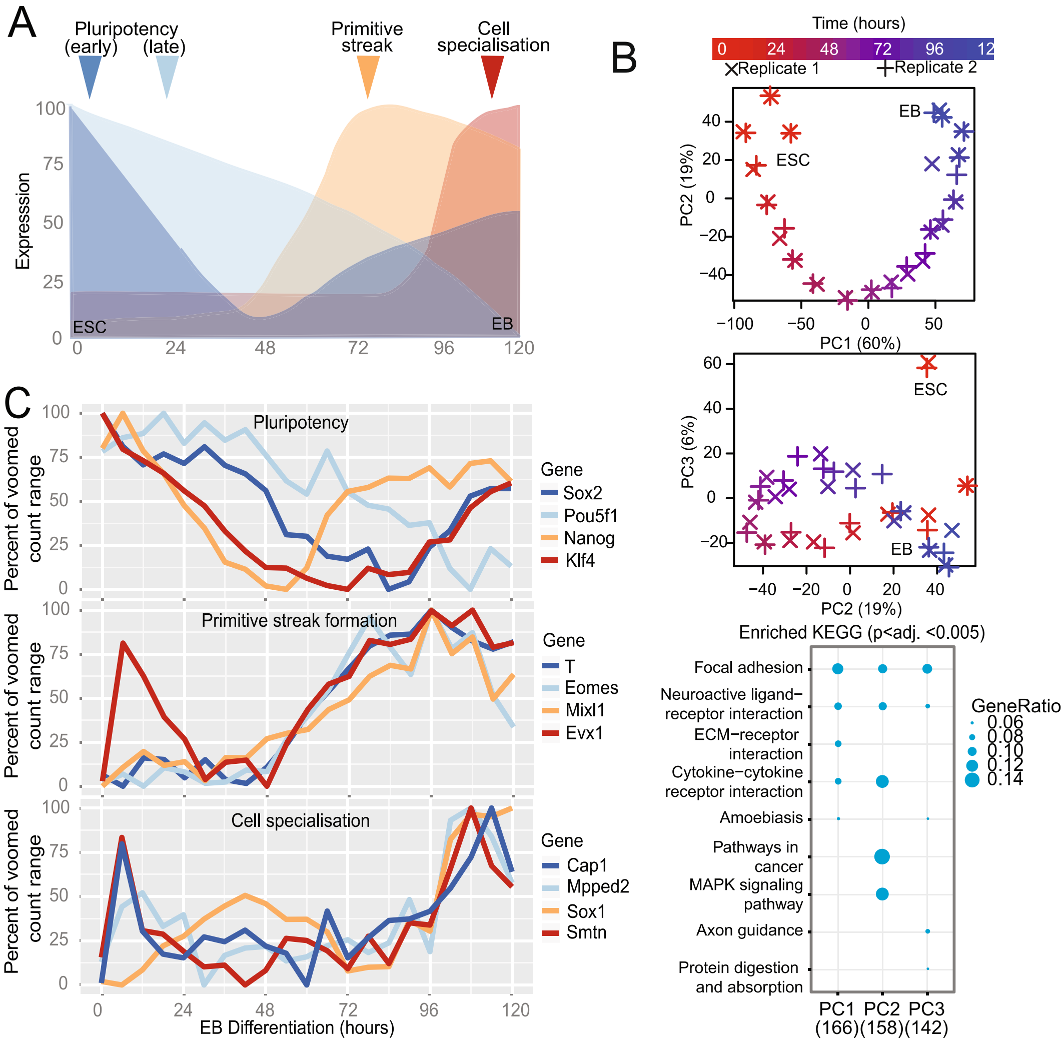

Figure 1. Global and gene-specific evaluation of augmented temporal resolution in mES differentiation. (**A**) Schematic of mouse embryonic stem cell (ESC) differentiation into embryoid bodies (EB) over the time course evaluated here. (**B**) Analysis of the top three principle components (PCs) based on the 2,000 most variable genes from biological duplicate-6 hourly transcriptomes and KEGG pathway enrichment for 500 genes contributing most to each of the top three PCs. (**C**) Expression profiles of genes associated with pluripotency, primitive streak formation and cell specialization.

An improved signaling cascade described by higher temporal resolution. Increased sampling frequency can provide a powerful insight into understanding of the contribution of gene regulatory networks to cellular differentiation[22]. We utilized the DREM v2 analysis tool[28] to evaluate transcription factor (TF) target gene expression patterns. Divergence of gene targets responsive to groups of TF at each time point, either 24-hourly or 6-hourly (Fig. 2A,B) was shown if the overall difference was significant at $p < 0.001$. Compared to 24-hourly, the observed complexity was significantly higher, especially in the first 48 hours. We observed that significant changes in gene regulation occurred continuously within the 24-hour windows. Most notably, first 24 hours following depart from pluripotency resembles an ordered cascade of TF activity (Figs 2A and S2A) with large-scale changes in TF activity at 12, 18 and 24 hours; of which little can be deduced measuring at just 24 hours (Figs 2B and S2B). Focusing on the interplay between two key transcription factors (Otx2 and Pou5f1/Oct4[29], Fig. 2A), we observed a rapid rise in Otx2 activity in the first six hours and stable Pou5f1activity for the first 24 hours (Red Box). Otx2 activity did not coincide with mRNA expression of the factor itself (Fig. 2A vs. C), although previous studies have observed increased in Otx2 protein expression within 3-hours of differentiation[29], however periodic drops in *Pou5f1* mRNA expression appeared to coincide with decreases in POU5F1 target genes, we calculated the time taken for *Pou5f1* expression to result in changes in highly positively correlated ($r > 0.8$) target genes using a cross-correlation approach similar to ref. 30. We then evaluated how these "delays" enriched for certain Reactome pathways (Fig. 2D). We found rapid effects for targets enriched for "gene expression"- and a delayed effect on "cell cycle" pathways compared to a null distribution produced by 500 random "target" selections (grey). These were similarly observed in the DREM GO-term enrichment tool for Pou5F1 targets decreasing in expression at 42 (early- Transcription Factor Activity) and 54 hours (late- Epithelial Proliferation; Fig. 2A, Blue Box &

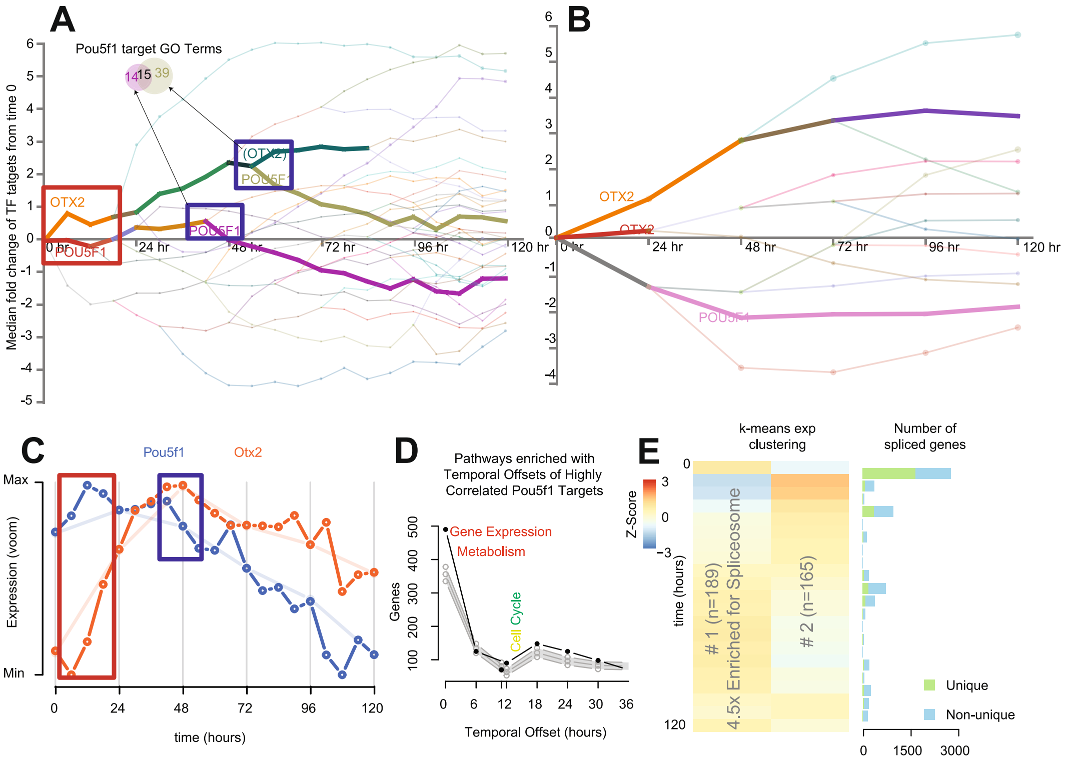

Figure 2. Insights into regulatory and gene expression kinetics. (**A** and **B**) Observable regulatory network dynamics at 24- and 6-hourly measures with Otx2 and Pou5f1 target containing profiles annotated and in bold, See Supplementary Fig. S2 for full figure. Transcriptomes at 24- (top) and 6-hourly (bottom) were subjected to DREM analysis of mouse TF/target gene interactions. Lines represent the median fold change (relative to time 0) of grouped TF target genes- representing activity of the TF itself, line colors are assigned by branch and are not comparable between panels. A p-value cutoff of 0.001 was applied to calculating divergent TF activity (splits). (**C**) Expression of the key transcription factors *Pou5F1* (*Oct4*) and *Otx2*. Red and blue boxes correspond to the time points highlighted in part A. (**D**) Distribution of the number of genes and the time delay required to meet a maximum correlation (>0.8) between gene targets of Pou5f1 and *Pou5f1* itself compared to 95% quantiles of 500 random gene selections. (**E**) Two k-means clusters of short-lived RNA (slRNA) genes displaying differential expression without changes at 24-hourly time points (adj. $p < 0.0001$).

Supplementary Fig. S2C) and associated with the decrease in *Pou5F1* expression (Fig. 2C, Blue Box). Importantly, Pou5F1 mRNA and protein expression are temporally correlated[29]. This result implies that TF-target genes may be activated in an ordered- time dependent fashion. To explore this more broadly, we evaluated other TF-target gene temporal dynamics for other TFs that exhibited strong positive or negative correlations between the TF and their target genes. We found evidence of highly structured TF-target expression patterns in time for negatively correlated Pou5f1 and Suz12 targets, as well as positively correlated Nanog, Myc, Sox2 and Suz12 targets (Supplementary Fig. S3).

These observations of precise temporal ordering of transcriptional events emphasize the importance of factoring time delays into understanding gene regulatory networks[31]. This highlights the capacity of increased temporal resolution to directly identify regulator-target gene interactions instead of relying on inference; which is common in cross-correlation approaches.

Increased temporal resolution identifies genes with previously uncharacterized expression patterns (Short-lived (slRNA) & Cycling (cycRNA)).

Having established that the increased temporal resolution markedly improves the molecular framework for evaluating the contribution of gene expression to ES differentiation, we next sought to identify gene expression signatures previously unable to be resolved using lower temporal resolution. For each 24-hour period, we identified genes that were differentially expressed between 0 and 6, 12 and 18 hours but not between any 24-hourly measures (Supplementary Fig. S2D). We identified 1,135 genes with significant changes in gene expression that were unchanged between any 24-hourly comparison (adjusted $p < 0.0001$). Of these, 354 were differentially expressed for more than half of the corresponding 24-hour window, mostly in the first and last 24-hour periods. These genes were described as short-lived RNAs (slRNAs). slRNA expression patterns over the first 24 hours of differentiation were found to be positively correlated with the same time window of retinoic acid directed differentiation[21] (Supplementary Fig. S2E) implying that these genes may form part of the early response to differentiation signals. K-means clustering and KEGG pathway analysis of the expression profiles of these genes (Fig. 2E) revealed enrichment in genes associated with the spliceosome (adjusted $p < 0.05$) dramatically decreasing in expression over the first 24 hours before returning

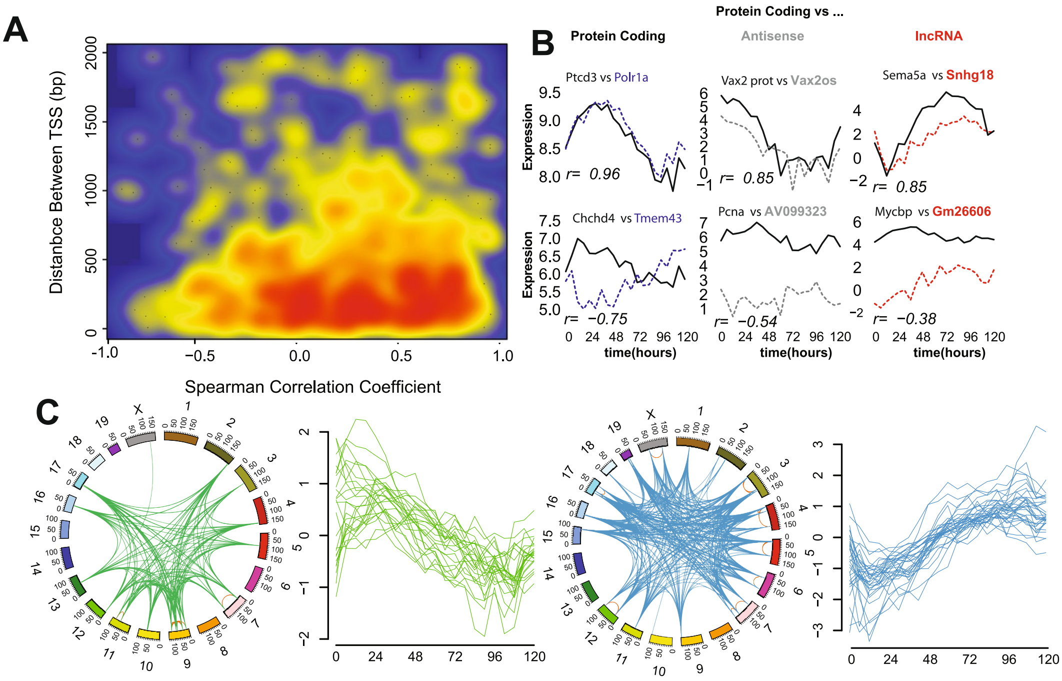

Figure 3. Analysis of gene co-expression patterns using augmented temporal resolution. (**A**) Smoothed scatter plot showing the correlation coefficient across the time course vs. distance between transcriptional start sites (TSS) of bidirectional gene pairs. Blue indicates no gene pairs; yellow and red indicate increasing numbers of pairs sharing similar properties. (**B**) Expression patterns of example bidirectional genes of the same or different gene biotype. Spearman's correlation coefficient is reported for each pair. (**C**) Genomic location (circos) and expression pattern (line plot) of two independent co-expressed groups of 5 or more contiguous genes sharing correlated expression ($r > 0.5$).

slowly to baseline. To examine whether this impacted gene-splicing patterns, we employed a differential exon (DEX) analysis between consecutive six-hourly time points and counted the number of genes displaying DEX usage (adjusted p value < 0.01 Fig. 2E). Consistent with previous studies, the alternate splicing was most highly associated with cell differentiation[32] (Fig. 2E). Increased temporal resolution has elucidated that these changes happen very rapidly (majority of changes in the first six hours), and that slRNAs may be involved in suppressing the alternate splicing of genes and limiting transcriptional plasticity.

Some slRNAs appeared to have periodic expression profiles. We thus sought to uncover periodic expression patterns genome-wide, by applying a fast-Fourier transformation to our data (see Methods). Periodogram analysis was utilized to ascertain the dominant cycling period for each gene. We found 137 genes, which we termed cycling RNAs (cycRNAs), sharing the same dominant cycling period of less than 36 hours in both biological replicate experiments (Supplementary Table S2). Supporting the efficacy of the approach, we found *Clock*, which encodes a key regulator of circadian rhythm in mammals, to have a period of 24.2 hours. We identified 20 genes that displayed characteristics of both slRNAs and cycRNAs (Supplementary Fig. S2F), including *Ewsr1* and *Clk1*, involved in gene splicing[33,34] as well as five uncharacterized lncRNAs. Given the highly specific expression patterns in this context, we propose these genes may similarly have roles in maintaining or establishing biological rhythms. Together these investigations show that the augmented temporal resolution approach provides access to gain insights from regulatory pathways by identifying transitions in expression that would otherwise have remained hidden.

Increased temporal resolution gives insight into local gene regulation in the genome.

Evaluating gene transcription at high temporal resolution in a highly dynamic process such as ES development, we anticipated that it might be feasible to dissect structural gene regulation within a given locus. To explore this possibility, we examined expression arising from transcripts that are oriented head-to-head as so-called bidirectional pairs[35,36]. Interestingly, we observed that the antisense transcript for *Evx1* (Fig. 1C) displayed a previously unobserved[15] increase in expression in the first 24 hours after departure from pluripotency that was reflected in its paired protein coding gene *Evx1* (Supplementary Fig. S4A), highlighting the increased power of frequent sampling over time. In total, we identified 1,251 gene pairs with bidirectional transcriptional start sites (TSS) within 2,000 bp and evaluated correlation coefficients across the time course, distance between TSS and median expression values. Consistent with other studies, we found expression correlation more positive for bidirectional gene pairs than random transcript pairs[35] (Supplementary Fig. S4B). We were also able to show that the distance between TSS of highly correlated bidirectional gene promoters is typically less than 500 bp (Fig. 3A), consistent with a common regulatory domain. Highly correlated or anti-correlated genes pairs displayed differences in total gene expression, particularly with discordant gene biotypes (Fig. 3B). We found that protein coding gene pairs were more likely to be of similar expression levels and positively correlated (Mann-Whitney $p < 0.05$) than protein coding/noncoding pairs (Supplementary Fig. S4C). Applying a variant of the temporal offset analysis used to measure TF- gene target delays, we calculated the time taken and defined the apparent driver gene type for peak correlation in coding/noncoding bidirectional pairs (Supplementary Fig. S4D,E). This did not reveal a generalized bias in either time taken or particular "driving" gene type. However, this approach shows that the lncRNA *Hotairm1*, required for activation of Hoxa1[37], appears to have a six-hour delay between its expression changes and HoxA1. We present evidence of other examples of lncRNA-led expression of protein coding genes in small numbers of bidirectional pairs (Supplementary Fig. S5).

To investigate whether the strong correlative potential between gene pairs could facilitate the identification of regions of the genome that are coordinately regulated[38], we scanned across the genome for regions containing five or more contiguous genes that were co-expressed ($r > 0.5$). This revealed 59 regions with a mean size of 821 kb -each containing 5–14 genes (mean of 6) genes. To examine the higher order chromatin architecture of these regions, we compared these regions to published data on topological associated domains (TADs) for mouse ESCs[39]. We found that the majority of the regions were each contained within a single TAD (Supplementary Fig. S4F), increasing the likelihood for a common regulatory architecture. Evaluation of gene-expression patterns across these regions revealed evidence of high co-expression at both the inter- and intra-chromosomal levels (Supplementary Fig. S4G). We assembled a map of regions of the mouse genome displaying high levels of clustered co-expression (Fig. 3C) by comparing the expression profiles of the regions. Two independent modules were identified with distinct decreasing (green)- and increasing (blue) expression patterns with differentiation. Given the independent location and expression patterns of these clusters, we suggest they may form core expression-factories of cellular differentiation. In support of this notion, this analysis identified the a region -associated with the "increasing module"- containing the imprinting locus of *H19, Igf2, Tnn3* and *Mrpl23*[40] (Supplementary Fig. S4H); previously shown to be activated in concert during early stem cell differentiation[41].

These investigations illustrate how analysis of high-resolution temporal transcriptomic data provides an independent and convenient approach (relying only RNA-Seq) to guide the partitioning of the genome into regulatory domains.

Increased temporal resolution refines the noncoding landscape of mESC differentiation.

Having shown that rapid changes in lncRNAs are a key feature of ES differentiation, and that co-expression analysis is a powerful tool for understanding gene regulation with augmented temporal resolution, we sought to unravel the roles that lncRNAs might play in ESC differentiation.

Analysis of gene annotations yielded confident expression data for 588 lncRNA genes at six-hourly resolution (520 for 24-hourly, Supplementary Table S1). Indeed, added temporal resolution increased information of all noncoding transcript biotypes indicating that a proportion of these genes were only present for a short duration in this system. It is important to note that these do not represent the entirety of lncRNA expression in this process since the depth of sequencing was not standard for lncRNA coverage[42] and the poly-A selection for cDNA would have missed enhancer RNAs, miRNA precursors and Nuclease P processed lncRNAs[43]. Clustering the observed lncRNA expression patterns with time-dependent protein coding gene expression showed that lncRNAs were enriched at lower expression levels and shared related expression profiles to protein coding genes (Fig. 4A). This relationship was further examined whereby K-means clustering of these expression profiles compared to clustering of a similar number of time-dependent protein coding genes (Figs 4B and S6A) revealed clusters of lncRNA genes resembling gene expression patterns associated with stemness (cluster a) primitive streak formation (cluster b) and WNT signaling (cluster c)[15]. As co-expression has been illustrated to provide valuable insight into lncRNA function[18], the additional correlative strength afforded by this study is anticipated to more reliably guide the functional association of these lncRNAs with these processes. The dynamics of lncRNA expression observed here indicate that future studies using RNA capture sequencing or higher sequencing depth of Ribosome depleted RNA will provide more comprehensive insights into the role of lncRNA in the molecular events underlying cell differentiation.

As lncRNAs often exert their function through guiding or assembling transcriptional machinery, we sought to identify potential regulatory lncRNAs in this system. We selected 50 highly or variably expressed lncRNAs (Fig. 4A) and tested for evidence of gene regulatory behavior across the transcriptome. The temporal resolution allowed the use of time to resolve precedence, thus adding weight to a potential causal relationship. Since lncRNAs typically exert their function as a transcript, we set a maximum time offset of 18 hours to avoid secondary (altered protein level) effects and examined patterns in the predicted gene targets of lncRNAs ($r > 0.8$, divided by positive or negative associations). Reactome pathway analysis revealed that 11 of these lncRNAs (including well characterized lncRNAs, Supplementary Fig. S6B and C) were potentially involved in regulating networks of genes associated with key developmental processes (p.adj <0.05, Supplementary Fig. S6C). These analyses assigned target gene networks consistent with characterized lncRNA biological functions for *Malat1* (oncogenic)[44], *Neat1* & *Rian* (association with gene repression)[45] and *Meg3* (tumour suppressor)[46]. Interestingly, these data suggest that the pro-tumorigenic function of *Malat1* may be mediated through facilitating the increase of MAPK signaling molecules. Importantly, these data also provide testable evidence for seven previously uncharacterized lncRNAs role in ES development and describes a map of regulatory interactions potentially driven by lncRNAs (Fig. 4C) whereby lncRNAs expression may impact coding gene expression across the genome. The identification of lncRNAs with a predicted biological role is important for unraveling lncRNA function, providing candidate functional lncRNAs and providing a level of molecular detail that is currently lacking in many lncRNA studies.

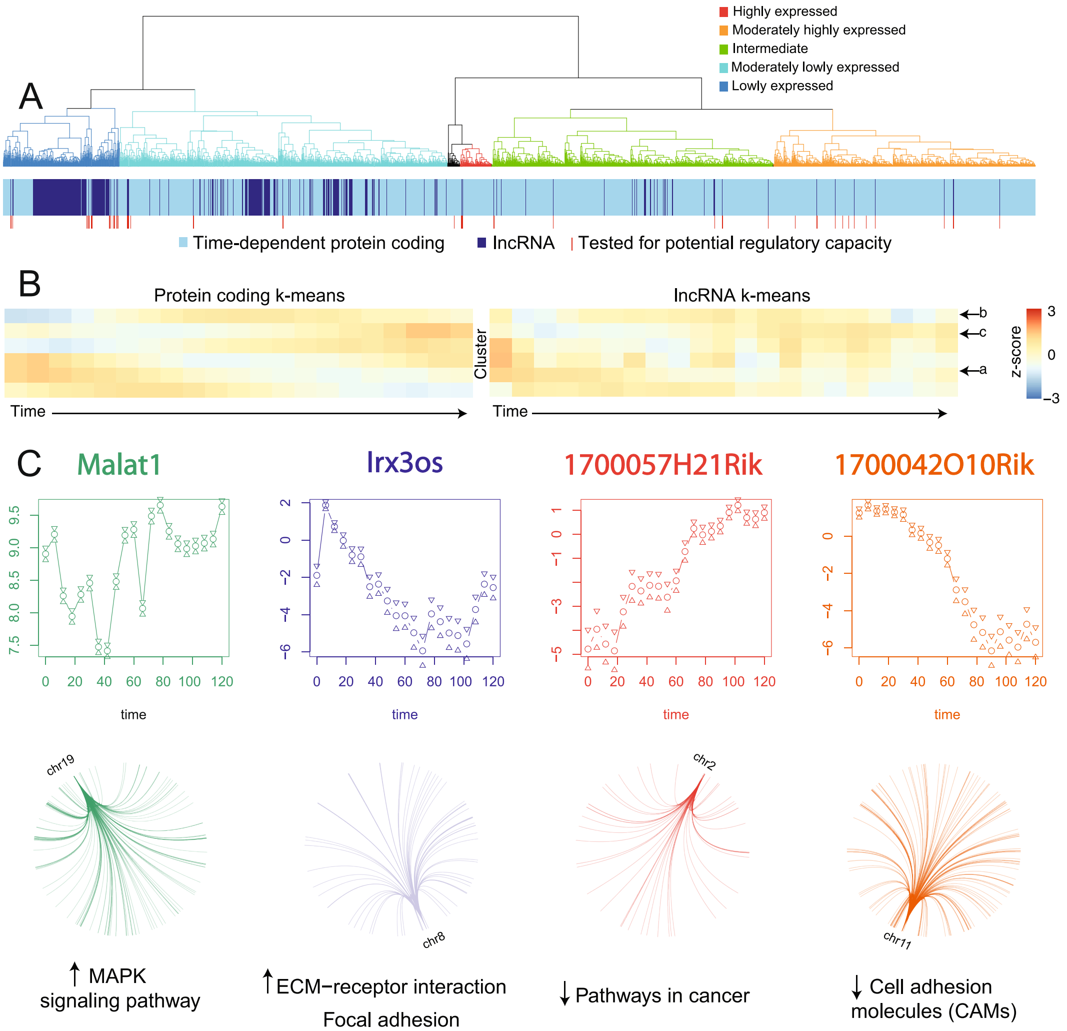

Figure 4. Augmented temporal resolution of ncRNA expression in cellular differentiation. (**A**) Hierarchical clustering of lncRNAs (dark blue) with time-dependent protein coding genes (light blue) by their expression patterns over time. Dendrogram was manually colored to reflect gene expression levels of the top-level clusters. (**B**) K-means clustered expression profiles of protein coding genes compared to the same number of lncRNA gene expression clusters. Common profiles are marked with arrows. (**C**) Expression profiles of four lncRNAs predicted to have regulatory roles in ES development as well as the genome location & pathways enriched in their gene targets. *Malat1* and *IRX3os* display a positive association with their targets, whereas *1700057H21Rik* and *1700042O10Rik* have a putative repressive impact.

Discussion and Conclusions

Transcriptional regulation of key biological events is a key feature in understanding the complexity of cellular processes. Here we describe a detailed transcriptomic resource for research in cellular development, a framework for unraveling this detail and identifying new targets for analysis. We also present a comprehensively detailed survey of noncoding transcripts throughout early stem cell development. We have identified many previously uncharacterized noncoding RNAs with potentially pivotal roles in cellular differentiation. This will provide a valuable tool for researchers unraveling the transcriptional complexity of cellular differentiation.

Increased interpretive power. The understanding of molecular events underlying the departure from pluripotency has been determined by the extant knowledge of how biological functions are exerted – often measured at 24 hourly or greater intervals. We hypothesized that interpretations of this model were missing detail in light of evidence indicating the unforeseen dynamics in RNA biology and regulation. By probing this detail with finer time distinctions, we show that gene expression profiles of well-characterized genes display significant variation of expression levels and that more detail can still be gleaned with increased sampling frequency (Supplementary Fig. S7A and B). Importantly, these variations are manifest in a significantly more complex gene regulatory framework. This is consistent with a reduction in temporal aggregation bias[19] and highlights early array-based investigations in yeast demonstrating the importance of sufficient temporal resolution in understanding gene expression patterns[47]. As such, much detail is likely missing from other systems that involve a change in phenotype or cellular behavior. With large-scale transcriptomic analyses becoming increasingly accessible, it is opportune to revisit other well-studied transitions with the view of improving understanding and applicability of their results rather than relying on presuppositions about gene expression patterns[48].

Insights into short bursts of transcription. We have shown the benefit of frequent sampling over time in observing the transcription of genes that are observable only within sub-24 hourly windows. This approach highlights the importance of taking into account the presence of short-lived transcripts and shows that cells express more of the transcriptome in a time-dependent fashion. To this end, we have identified rapid changing and periodically expressed genes, which we term short-lived (slRNA) and cycling (cycRNA), that were unobservable outside this framework. That many slRNAs exhibited changes in expression over the first 24 hours of differentiation is consistent with rapid initial cellular response to stimuli[21, 49]. Indeed, it is likely that significant gene expression changes- especially noncoding- occur on timeframes shorter than those presented that may not be amenable to optimal time point prediction strategies[48]. By probing deeper into time-dependent gene transcription-possibly by interpolating available datasets-[47] it will be possible to uncover further complexity underlying cellular plasticity and gene regulation. These observations reinforce the concept that adequate temporal resolution is vital for describing biological transitions- for example in dissecting primary from follow on effects in gene knockdown studies – and that end-point analysis likely does not reflect the complex biology of phenotype changes.

Insight genome organization and regulation. Similarly, by using time to separate the order of gene transcription, we have been able to predict local gene regulation across the genome. We have been able to observe concerted gene expression (in *trans*) of hundreds of genes separated by large genome differences (in *cis*). Typical studies of this nature involve correlative analysis requiring large samples sizes and resources[50]. We have instead leveraged the time axis to achieve these as well as discriminate driver from passenger molecular events. This has allowed the estimation of the time delay for changes in expression of regulatory molecules to manifest in changes in their target gene transcription and we have been able to unravel a potentially complex network of gene profiles responding to lncRNA transcription. Finally, we have been able to use an integrated biological system to draw strong associations in trans relationships with bidirectional promoters. Typically these associations are observed by using thousands of gene expression profiles, yet here we have been able to do so with only 42 transcriptomes (duplicate time courses of 21 points each).

Methods

Sample Generation and Library Preparation. Biological duplicate, low passage number (P18) W9.5 ESCs were cultured and differentiated as described previously[15, 51]. Cultures were harvested every six hours from the induction of differentiation to 120 hours post differentiation induction. Total RNA from cultures was purified using Trizol (Life Technologies) and DNase treatment was performed by RQ1 DNase (Promega) according to the manufacturer's instructions. RNA integrity was measured on a Bioanalyzer RNA Nano chip (Agilent). RNA-Seq library preparation and sequencing of Poly-A-NGS libraries generated from 500 ng total RNA using SureSelect Strand Specific RNA Library Preparation Kit (Agilent) were performed according to the manufacturer's instructions at the same time to minimize batch effect. Paired-end libraries were sequenced to the first 100 bp on a HiSeq 2500 (Illumina) on High Output Mode.

Quality control and read mapping. Library sequencing quality was determined using FastQC (Babraham Bioinformatics) and FastQ Screen (Babraham Bioinformatics). Illumina adaptor sequence and low quality read trimming (read pair removed if <20 base pairs) was performed using Trim Galore! (Babraham Bioinformatics: www.bioinformatics.babraham.ac.uk/). Tophat2[52] was used to align reads to the December 2011 release of the mouse reference genome (mm10) as outlined by Anders *et al.*[26]. Read counts data corresponding to GENCODE vM2 transcript annotations were generated using HTSeq[53]. *de novo* transcript assembly was performed on each merged BAM file using Cufflinks' reference annotation based transcript (RABT) assembly[54], using the Gencode vM2 transcriptome[55] as a guide (options: -u -I 500000 -j 1.0 -F 0.005-trim-3-dropoff-frac 0.05 -g gencode.vM2.annotation.gtf–library-type fr-firststrand). Transcript assemblies were then merged using Cuffmerge[56] using default parameters, and compared to the Gencode vM2 reference transcriptome using Cuffcompare[56]. Novel transcripts with a Cuffcompare class code of j, i, o, u or x were filtered using three steps to find novel lncRNAs. First, a Browser Extensible Data (BED) format file was generated using a python script (https://gist.github.com/davidliwei/1155568) and any single exon transcripts were removed. Second, the FASTA-formatted sequence for each transcript was obtained using BEDTools[57], the nucleotide (nt) length and open reading frame (ORF) size found using Perl scripts, and those with a length less than 200 nt or a ORF size greater than 300 nt were removed. Lastly, transcript sequences were submitted to Coding Potential Calculator (CPC)[58], and those with a coding potential of >0 were removed.

Bioinformatics. All analyses were performed in the R Statistical Environment[59]. Briefly, counts data were background corrected and normalized for library size using edgeR[60], then transformed using voom[61] for differential expression analysis using LIMMA[62]. Transcription Factor (TF) activity was inferred from gene expression data using DREM[28] with a branching P-value of 0.001 based on curated TF-target gene lists associated with mouse ESC differentiation from ChEA[63]. TF-target gene was calculated by maximal Pearson's correlation coefficient of >0.8 using a custom autocorrelation analysis and verified with the "ccf" function in R. Gene differential exon (DEX) usage was analyzed by DEXSeq[64] on vM2 gene annotations using default settings and an adjusted p value cutoff of 0.001 for DEX between biological duplicates at each consecutive time-point. Genome position analyses were performed using genomic ranges[65] based on vM2 annotations imported with 'rtracklayer'[66] and Pearson's correlation coefficient of gene expression Bidirectional genes were defined as two genes with expression data on opposing strands with <2000 bp between the transcriptional start sites (TSS). Co-expressed gene clusters were defined as >5 contiguous genes with expression data displaying a Pearson's Correlation Coefficient of >0.5

with neighbouring genes. Cluster co-expression data was visualized with corrplot[67] and Cytoscape (v3.1.0)[68], location of related clusters was visualized by Circos[69]. Gene expression periodicity was measured on 120 interpolated expression values[70] for each replicate time series using GeneCycle[71], candidate periodically expressed genes were identified as having the same calculated dominant cycling frequency between biological replicates. Time-dependent expression signatures were established using maSigPro[72] with a replicate correlation coefficient cutoff of 0.8. Target genes of potential regulatory (top 50 most highly and/or variably expressed) lncRNAs were identified using the GeneReg package[73] on 100 point-interpolated expression data based on fitted expression values between duplicates and setting a maximum time delay of 18 hours and a global correlation coefficient of 0.9 and visualized using Cytoscape. Gene lists were functionally annotated with KEGG and Reactome pathways (adjusted p value < 0.05) using the clusterProfiler and ReactomePA packages[74].

Availability of data and material. Data has been deposited into GEO under accession number GSE75028.

References

1. Martello, G. & Smith, A. The nature of embryonic stem cells. *Annual review of cell and developmental biology* **30**, 647–675, doi:10.1146/annurev-cellbio-100913-013116 (2014).
2. Liu, N., Liu, L. & Pan, X. Single-cell analysis of the transcriptome and its application in the characterization of stem cells and early embryos. *Cellular and molecular life sciences: CMLS* **71**, 2707–2715, doi:10.1007/s00018-014-1601-8 (2014).
3. Rosa, A. & Brivanlou, A. H. Regulatory non-coding RNAs in pluripotent stem cells. *International journal of molecular sciences* **14**, 14346–14373, doi:10.3390/ijms140714346 (2013).
4. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science (New York, N.Y.)* **306**, 2242–2246, doi:10.1126/science.1103388 (2004).
5. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108, doi:10.1038/nature11233 (2012).
6. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227, doi:10.1038/nature07672 (2009).
7. Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F. & Mattick, J. S. Specific expression of long noncoding RNAs in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 716–721, doi:10.1073/pnas.0706729105 (2008).
8. Bonasio, R. & Shiekhattar, R. Regulation of transcription by long noncoding RNAs. *Annual review of genetics* **48**, 433–455, doi:10.1146/annurev-genet-120213-092323 (2014).
9. Fatica, A. & Bozzoni, I. Long non-coding RNAs: new players in cell differentiation and development. *Nature reviews. Genetics* **15**, 7–21, doi:10.1038/nrg3606 (2014).
10. Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Molecular cell* **43**, 904–914, doi:10.1016/j.molcel.2011.08.018 (2011).
11. Clark, M. B. *et al.* Genome-wide analysis of long noncoding RNA stability. *Genome research* **22**, 885–898, doi:10.1101/gr.131037.111 (2012).
12. Signal, B., Gloss, B. S. & Dinger, M. E. Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs. *Trends in genetics: TIG* **32**, 620–637, doi:10.1016/j.tig.2016.08.004 (2016).
13. Quek, X. C. *et al.* lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic acids research* **43**, D168–173, doi:10.1093/nar/gku988 (2015).
14. Gloss, B. S. & Dinger, M. E. The specificity of long noncoding RNA expression. *Biochimica et biophysica acta* **1859**, 16–22, doi:10.1016/j.bbagrm.2015.08.005 (2016).
15. Dinger, M. E. *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome research* **18**, 1433–1445, doi:10.1101/gr.078378.108 (2008).
16. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods* **5**, 613–619, doi:10.1038/nmeth.1223 (2008).
17. Bruce, S. J. *et al.* Dynamic transcription programs during ES cell differentiation towards mesoderm in serum versus serum-freeBMP4 culture. *BMC genomics* **8**, 365, doi:10.1186/1471-2164-8-365 (2007).
18. Bergmann, J. H. *et al.* Regulation of the ESC transcriptome by nuclear long noncoding RNAs. *Genome research* **25**, 1336–1346, doi:10.1101/gr.189027.114 (2015).
19. Bay, S. D., Chrisman, L., Pohorille, A. & Shrager, J. Temporal aggregation bias and inference of causal regulatory networks. *Journal of computational biology: a journal of computational molecular cell biology* **11**, 971–985, doi:10.1089/cmb.2004.11.971 (2004).
20. Chu, L. F. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology* **17**, 173, doi:10.1186/s13059-016-1033-x (2016).
21. De Kumar, B. *et al.* Analysis of dynamic changes in retinoid-induced transcription and epigenetic profiles of murine Hox clusters in ES cells. *Genome research* **25**, 1229–1243, doi:10.1101/gr.184978.114 (2015).
22. mod, E. C. *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science (New York, N.Y.)* **330**, 1787–1797, doi:10.1126/science.1198374 (2010).
23. Arbeitman, M. N. *et al.* Gene expression during the life cycle of Drosophila melanogaster. *Science (New York, N.Y.)* **297**, 2270–2275, doi:10.1126/science.1072152 (2002).
24. Tan, M. H. *et al.* RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. *Genome research* **23**, 201–216, doi:10.1101/gr.141424.112 (2013).
25. Boeck, M. E. *et al.* The time-resolved transcriptome of C. elegans. *Genome research* **26**, 1441–1450, doi:10.1101/gr.202663.115 (2016).
26. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature protocols* **8**, 1765–1786, doi:10.1038/nprot.2013.099 (2013).
27. Hirst, C. E. *et al.* Transcriptional profiling of mouse and human ES cells identifies SLAIN1, a novel stem cell gene. *Developmental biology* **293**, 90–103, doi:10.1016/j.ydbio.2006.01.023 (2006).
28. Schulz, M. H. *et al.* DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC systems biology* **6**, 104, doi:10.1186/1752-0509-6-104 (2012).
29. Yang, S. H. *et al.* Otx2 and Oct4 drive early enhancer activation during embryonic stem cell transition from naive pluripotency. *Cell reports* **7**, 1968–1981, doi:10.1016/j.celrep.2014.05.037 (2014).
30. Li, H., Luan, Y., Hong, F. & Li, Y. Statistical methods for analysis of time course gene expression data. *Frontiers in bioscience: a journal and virtual library* **7**, a90–98 (2002).
31. Chen, H., Mundra, P. A., Zhao, L. N., Lin, F. & Zheng, J. Highly sensitive inference of time-delayed gene regulation by network deconvolution. *BMC systems biology* **8**(Suppl 4), S6, doi:10.1186/1752-0509-8-S4-S6 (2014).
32. Salomonis, N. *et al.* Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 10514–10519, doi:10.1073/pnas.0912260107 (2010).
33. Paronetto, M. P., Minana, B. & Valcarcel, J. The Ewing sarcoma protein regulates DNA damage-induced alternative splicing. *Molecular cell* **43**, 353–368, doi:10.1016/j.molcel.2011.05.035 (2011).

34. Liu, Y. *et al*. Phosphorylation of the alternative mRNA splicing factor 45 (SPF45) by Clk1 regulates its splice site utilization, cell migration and invasion. *Nucleic acids research* **41**, 4949–4962, doi:10.1093/nar/gkt170 (2013).
35. Trinklein, N. D. *et al*. An abundance of bidirectional promoters in the human genome. *Genome research* **14**, 62–66, doi:10.1101/gr.1982804 (2004).
36. Yang, M. & Elnitski, L. Orthology-driven mapping of bidirectional promoters in human and mouse genomes. *BMC bioinformatics* **15**(Suppl 17), S1, doi:10.1186/1471-2105-15-S17-S1 (2014).
37. Zhang, X. *et al*. A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood* **113**, 2526–2534, doi:10.1182/blood-2008-06-162164 (2009).
38. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature genetics* **31**, 180–183, doi:10.1038/ng887 (2002).
39. Dixon, J. R. *et al*. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380, doi:10.1038/nature11082 (2012).
40. Kaffer, C. R., Grinberg, A. & Pfeifer, K. Regulatory mechanisms at the mouse Igf2/H19 locus. *Molecular and cellular biology* **21**, 8189–8196, doi:10.1128/MCB.21.23.8189-8196.2001 (2001).
41. Poirier, F. *et al*. The murine H19 gene is activated during embryonic stem cell differentiation *in vitro* and at the time of implantation in the developing embryo. *Development (Cambridge, England)* **113**, 1105–1114 (1991).
42. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics* **15**, 121–132, doi:10.1038/nrg3642 (2014).
43. Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nature reviews. Genetics* **17**, 47–62, doi:10.1038/nrg.2015.10 (2016).
44. Li, L. *et al*. Role of human noncoding RNAs in the control of tumorigenesis. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 12956–12961, doi:10.1073/pnas.0906005106 (2009).
45. Guttman, M. *et al*. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295–300, doi:10.1038/nature10398 (2011).
46. Zhang, X. *et al*. A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. *The Journal of clinical endocrinology and metabolism* **88**, 5119–5126, doi:10.1210/jc.2003-030222 (2003).
47. Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S. & Simon, I. Continuous representations of time-series gene expression data. *Journal of computational biology: a journal of computational molecular cell biology* **10**, 341–356, doi:10.1089/10665270360688057 (2003).
48. Rosa, B. A., Zhang, J., Major, I. T., Qin, W. & Chen, J. Optimal timepoint sampling in high-throughput gene expression experiments. *Bioinformatics* **28**, 2773–2781, doi:10.1093/bioinformatics/bts511 (2012).
49. Gasch, A. P. *et al*. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell* **11**, 4241–4257 (2000).
50. Prieto, C., Risueno, A., Fontanillo, C. & De las Rivas, J. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PloS one* **3**, e3911, doi:10.1371/journal.pone.0003911 (2008).
51. Bruce, S. J. *et al*. *In vitro* differentiation of murine embryonic stem cells toward a renal lineage. *Differentiation; research in biological diversity* **75**, 337–349, doi:10.1111/j.1432-0436.2006.00149.x (2007).
52. Kim, D. *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).
53. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169, doi:10.1093/bioinformatics/btu638 (2015).
54. Roberts, A. & Pachter, L. RNA-Seq and find: entering the RNA deep field. *Genome medicine* **3**, 74, doi:10.1186/gm290 (2011).
55. Harrow, J. *et al*. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**, 1760–1774, doi:10.1101/gr.135350.111 (2012).
56. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**, 511–515, doi:10.1038/nbt.1621 (2010).
57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842, doi:10.1093/bioinformatics/btq033 (2010).
58. Kong, L. *et al*. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research* **35**, W345–349, doi:10.1093/nar/gkm391 (2007).
59. R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (2013).
60. 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, doi:10.1093/bioinformatics/btp616 (2010).
61. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* **15**, R29, doi:10.1186/gb-2014-15-2-r29 (2014).
62. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**, Article 3, doi:10.2202/1544-6115.1027 (2004).
63. Lachmann, A. *et al*. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444, doi:10.1093/bioinformatics/btq466 (2010).
64. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome research* **22**, 2008–2017, doi:10.1101/gr.133744.111 (2012).
65. Lawrence, M. *et al*. Software for computing and annotating genomic ranges. *PLoS computational biology* **9**, e1003118, doi:10.1371/journal.pcbi.1003118 (2013).
66. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842, doi:10.1093/bioinformatics/btp328 (2009).
67. Wei, T. & Simko, V. Corrplot: Visualization of a correlation matrix v. R package version 0.73 https://CRAN.R-project.org/package=corrplot (2013).
68. Shannon, P. *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504, doi:10.1101/gr.1239303 (2003).
69. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639–1645, doi:10.1101/gr.092759.109 (2009).
70. Orlando, D. A. *et al*. Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature* **453**, 944–947, doi:10.1038/nature06955 (2008).
71. Ahdesmaki, M., Fokianos, K. & Strimmer. GeneCycle: Identification of Periodically Expressed Genes. http://CRAN.R-project.org/package=GeneCycle (2012).
72. Conesa, A. & Nueda, M. J. maSigPro: Significant Gene Expression Profile Differences in Time Course Microarray Data. http://bioinfo.cipf.es/ (2013).
73. Huang, T. GeneReg: Construct time delay gene regulatory network. http://CRAN.R-project.org/package=GeneReg (2012).
74. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* **16**, 284–287, doi:10.1089/omi.2011.0118 (2012).

Acknowledgements

The authors acknowledge Kenneth Sabir and Ruth Pidsley for reviewing the manuscript; the Garvan Foundation and the Peter Wills Bioinformatics Facility for providing facilities and Agilent Technologies for providing RNAseq kits. BG acknowledges constructive feedback from Aaron Statham, Mark Pinese, Nenad Bartonicek, Jesper Maag and Quek Xiucheng. B.G. is supported by Cancer Institute NSW Early Career Fellowship 13/ECF/1-45.

Author Contributions

B.G. wrote the manuscript, assisted study conception, performed the analyses and library preparations (assisted by D.K.). M.D. conceived the study and assisted writing the manuscript. B.S. performed de-novo assembly and P.C. deconvolution, designed the web portal, assisted with figure composition and reviewed the manuscript. S.C., F.G. and D.K. performed lab work and reviewed the manuscript. A.P. provided design input, biological samples and facilities.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-06110-5

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.