Ge *BMC Genomics* (2017) 18:200
DOI 10.1186/s12864-017-3566-0

BMC Genomics

RESEARCH ARTICLE Open Access

Exploratory bioinformatics investigation reveals importance of "junk" DNA in early embryo development

Steven Xijin Ge

Abstract

Background: Instead of testing predefined hypotheses, the goal of exploratory data analysis (EDA) is to find what data can tell us. Following this strategy, we re-analyzed a large body of genomic data to study the complex gene regulation in mouse pre-implantation development (PD).

Results: Starting with a single-cell RNA-seq dataset consisting of 259 mouse embryonic cells derived from zygote to blastocyst stages, we reconstructed the temporal and spatial gene expression pattern during PD. The dynamics of gene expression can be partially explained by the enrichment of transposable elements in gene promoters and the similarity of expression profiles with those of corresponding transposons. Long Terminal Repeats (LTRs) are associated with transient, strong induction of many nearby genes at the 2-4 cell stages, probably by providing binding sites for Obox and other homeobox factors. B1 and B2 SINEs (Short Interspersed Nuclear Elements) are correlated with the upregulation of thousands of nearby genes during zygotic genome activation. Such enhancer-like effects are also found for human Alu and bovine tRNA SINEs. SINEs also seem to be predictive of gene expression in embryonic stem cells (ESCs), raising the possibility that they may also be involved in regulating pluripotency. We also identified many potential transcription factors underlying PD and discussed the evolutionary necessity of transposons in enhancing genetic diversity, especially for species with longer generation time.

Conclusions: Together with other recent studies, our results provide further evidence that many transposable elements may play a role in establishing the expression landscape in early embryos. It also demonstrates that exploratory bioinformatics investigation can pinpoint developmental pathways for further study, and serve as a strategy to generate novel insights from big genomic data.

Keywords: Single-cell RNA-seq, Exploratory data analysis, Pre-implantation development, Early embryogenesis, Transposons, Repetitive DNA

Background

Hypothesis testing is the foundation of the scientific methodology [1]. In biomedical research, focused and hypothesis-driven projects are especially encouraged due to the enormous complexity of the field. But little attention has been paid to how such hypotheses are generated. Beyond intuition, a quantitative, evidence-based methodology for generating research hypotheses will increase the tempo of biomedical research. The idea of leveraging the "big data" of tens of thousands of large genome-wide datasets has been discussed [2, 3]. Such data-driven, inductive methods can complement hypothesis-driven approaches to form an iterative process of ongoing research [3]. Our knowledge about many fundamental biological processes remains limited and fragmented. We can take advantage of the massive genome-wide datasets to learn about biological systems, akin to the study of an unknown planet in space exploration.

First proposed by Tukey, the goal of exploratory data analysis (EDA) [4] is to ***explore the data and find what it can tell us***. It is open-ended and much broader than testing pre-defined hypotheses or building desired models.

Correspondence: gexijin@gmail.com
Department of Mathematics and Statistics, South Dakota State University, Box 2225, Brookings, SD 57110, USA

EDA is a philosophically different approach to statistical analysis, not a new set of tools. In addition to data visualization, all statistical techniques can be used to ask various questions [5, 6]. EDA can assess distributions, structures and dependencies that are useful for modeling. More importantly, EDA can reveal important characteristics and trends, which can help formulate new hypotheses for further investigation [4]. Indeed, many big scientific discoveries are made accidentally [7]. Open-minded exploration of scientific data should be an essential part of a research project.

EDA can serve as a strategy to generate new hypotheses from biomedical data. The infrastructure is largely in place due to many open, collective efforts for sharing of data, annotation and software. The main challenge is to organically combine multiple datasets to gain a holistic understanding of complex biological systems. This requires interpretation of results within biological contexts and the ability to survey the literature while analyzing massive data. Such broad, data-driven effort aimed at general understanding of a biological process can be referred to as *exploratory bioinformatics investigation* (EBI).

This study represents an attempt of EBI on gene regulation in pre-implantation mouse embryos. In the initial stage of mammalian development, zygotes undergo maternal-to-zygotic transition (MZT) [8], during which maternal factors are eliminated while mRNA and protein synthesis using zygotic genome are initiated to take control of embryo development. A highly coordinated cascade of regulatory mechanisms unfolds rapidly to give rise to several cell lineages and the formation of blastocysts [9]. Understanding the complex process of pre-implantation development (PD) is important for both fertility related interventions as well as manipulation of embryonic stem cells (ESCs).

Many gene expression studies of the early mammalian embryos has been carried out using expressed sequence tags (ESTs) [10–13], DNA microarrays [14–18], RNA-sequencing (RNA-seq) [19], and, more recently, single-cell RNA-seq [20–22]. These studies documented the dynamic waves of gene expression at different stages. Using the powerful single-cell RNA-seq technique [21, 23], Deng et al. analyzed hundreds of cells from embryos of mixed background mice and found strong evidence for random and widespread monoallelic expression [21]. With sequence-level detail at single-cell resolution, this and other similar datasets [20, 22, 24] can be used for in-depth study of gene regulation during PD. The epigenetic remodeling of maternal and paternal genomes was also revealed by DNA methylation profiling [25–27]. In Zebra fish, transcription factors (TFs) such as POU5F1 (POU domain, class 5, transcription factor 1), NANOG (Nanog homeobox), and SoxB1 (transcription factor SoxB1) were found to activate zygotic gene expression [9, 28]. But the molecular mechanisms of PD remain poorly understood.

Retrotransposon expression is a defining event in genome reprogramming during PD [29]. Transposable elements (TEs) cover 30–50% of mammalian genomes [30]. Some are actively transcribed [31, 32], even retrotransposed, as much of the genome is briefly hypomethylated in early embryos. The expression of retrotransposons is dynamic and stage-specific [27, 31, 33–35]. Evsikov et al. showed that retrotransposons, especially long terminal repeats (LTRs), are abundantly represented in transcripts derived from mouse embryo at 2-cell stage [32]. Class III LTRs such as MERV-L family LTRs are transcribed at extremely high levels, accounting for about 3% of total transcriptional output at this stage [36]. Human endogenous retroviruses (ERVs) HERVK LTR are expressed at zygotic genome activation (ZGA) at eight-cell stage [33]. Long interspersed elements (LINEs), another major type of retrotransposon, are also expressed during PD [37], similar to short interspersed nuclear elements (SINEs). Interestingly, some endogenous retroviral activities have been found to be associated with and can serve as markers of pluripotency and totipotency [34, 35, 38] in stem cell populations, underlying the importance of studying retrotransposon expression.

Transcription of retrotransposons can directly influence the expression of neighboring genes. By analyzing EST libraries, Peaston et al. analyzed the expression of different types of TEs during PD and found that some TEs, especially MERV-L family LTRs, provide alternative 5′ first exons to 41 chimeric transcripts in 2-cell mouse embryos [39]. Besides LTRs, LINEs and SINEs also contributed a small number of chimeric transcripts. LINEs have also been found to initiate fusion transcripts in other studies [40]. More broadly, by analyzing cap-selected 5′ end of mouse and human transcripts from various embryonic and adult tissues, Faulkner et al. [31] found that 6–30% of all transcripts initiate from TEs. These transcripts are often tissue-specific. Studies on long noncoding RNAs (lncRNAs) also identified ~30,000 TEs critical for the biogenesis of about 30% of total lncRNAs sequences [41].

TEs can also influence gene expression indirectly by shaping the epigenetics landscape [25–27, 42–44], as GC-rich SINEs tend to be found near genes while LINEs has the opposite distribution. Some transcripts from TEs give rise to small RNAs that involve in post-transcriptional regulation during PD [45–47]. LINE-1 RNA can regulate its own expression [27]. Chip-Seq data in human cells shows that TEs contribute 25% of binding sites for TFs critical for embryonic stem cells such as POU5F1, NANOG and CTCF [41]. Xie et al. [16] compared expression dynamics of PD in human, mouse and bovine and found substantial difference in co-expression network, which could be partly due to the cis-regulatory modules provided by species-specific transposons. Recently,

Töhönen et al. used single-cell RNA-seq to analyze 348 single-cells from early human development [22]. They found that the promoters of 32 genes upregulated early at 4-cell stage are enriched with Alu elements that harbor motifs, including a "TAATCC" core motif bound by PITX and OTX homeobox family [22]. These studies provide substantial evidence for TEs' broad role in regulating gene expression during PD.

In this study, we use EBI approach to study the dynamic expression of both regular genes and TEs in mouse PD. Starting from the large single-cell RNA-seq data of Deng et al. [21], *our approach is to systematically observe the gene expression dynamics to help develop a broad understanding of the regulatory mechanisms of this complex process*. Our goal is to produce insights that can lead to novel, testable hypotheses on early embryo development. The data was analyzed alongside other expression and epigenetic studies of PD, using various tools and annotation databases (See flowchart in Fig. 1a). Our analyses provide evidence for co-regulation of transposons and regular genes in early embryo development in mouse, followed up with similar observation in human, bovine and zebrafish. This adds to existing evidence that the expansion of species-specific TEs may help rewire developmental pathways [41, 48]. Motif analysis of promoter sequences of stage-specific genes identified many homeobox domain TFs as potential regulators of PD, which could be experimentally tested. We also examined the non-random distribution of TEs in the mouse genome and the enrichment of SINEs in the promoters of genes specifically expressed in ESCs. Finally, we discussed the evolutionary benefits of transposons in promoting genetic diversity, especially in slow-reproducing animals.

Results

To study allelic-specific transcription, Deng et al. [21] analyzed individual cells from mixed lineage mouse embryos (CAST/EiJ females mated with C57BL/6 male) using the single-cell RNA-seq technique [49]. Excluding technical control samples, 259 RNA-seq libraries were generated, representing cells from zygote to blastocyst stages. The 2-cell (2C) stage is divided into early, middle (mid) and late phases. Each library contains about 22 million reads, mostly 43 base pair (bp) long. With a total of about 6 billion reads, this massive data enables in-depth EBI on the temporal and spatial regulation of transcription during PD.

Similar expression patterns for regular genes and retrotransposons

The raw sequence reads were re-analyzed to quantify gene expression using Tophat and Cufflinks programs [50] with updated genome annotation from Ensembl [51]. Gene expression data is available in Additional file 1: Table S1. Hierarchical clustering was used to analyze the expression pattern of 12,000 genes with sufficient change in expression across 259 individual cells at various developmental stages. In addition to temporal dynamics, this data can also reveal variability among cells at the same stage. See Additional file 2 for more details. Ten gene clusters were defined according to the similarity in temporal expression patterns (Fig. 1b), similar to previous efforts based on DNA microarray [14]. See Additional file 1: Table S2 for gene lists. Cluster A includes 3310 genes highly expressed in oocytes and quickly reduced at 2-cell (2C) and 4-cell (4C) stages. These are mostly maternal mRNAs undergoing degradation, which is evident from allele-specific mapping (Additional file 2: Figures S1–S4). ZGA occurs during the 2C stage, evidenced by marked changes in gene expression between early and mid 2C stages (Additional file 2: Figure S5). The 777 transcripts in cluster B are exclusively expressed at 2C and 4C stages. Cluster D genes are induced at mid 2C, but their transient expression lasts until the 16-cell (16C) stage. Cluster F genes are gradually upregulated at the 2C stage and downregulated in the blastocyst. Genes in clusters G to J are activated at various stages and remain highly expressed. Our goal is to find the regulatory mechanism behind these different patterns of expression.

During PD, much of the genome is de-methylated and transposons are actively transcribed. Previous studies have shown that their expression patterns are dynamic and stage-specific [27, 31, 33, 34]. To estimate their abundance, we used TETranscripts software [52] with a special index file derived from RepeatMasker data available at UCSC genome browser website [53]. Additional file 2: Figure S6 shows the expression pattern of some of the highly transcribed transposons. The Additional file 1: Table S3 contains the detailed expression levels of all TEs. The most highly expressed is a retrovirus-like element MT-int (RepBase ID: MTAI), a LTR element of the ERVL-MaLR family. It is expressed from the zygote to the 4C stage, similar to Cluster A genes. MERV-L (RepBase ID: MT2_Mm) is an ERVL family LTR that is sharply induced by more than 500-fold at mid 2C before decreasing to low levels at 8C, showing an expression pattern similar to cluster B genes. This is in agreement with a previous report that MERV-L accounts for about 3% of total transcriptional output at this stage [36]. There are several other LTR elements with this type of expression, including MT2C_Mm, ORR1A2, ORR1A3, and MT2B2. Intracisternal A particle (IAP) elements are also transcribed between 2C and 16C as expected [54], similar to cluster D genes. Besides LTRs, LINEs and SINEs are also expressed as expected. SINE transcripts increase modestly at the 2C stage and remain at that level through the blastocyst stage, similar to Cluster F and G genes. Retrotransposons are transcribed in a highly regulated manner

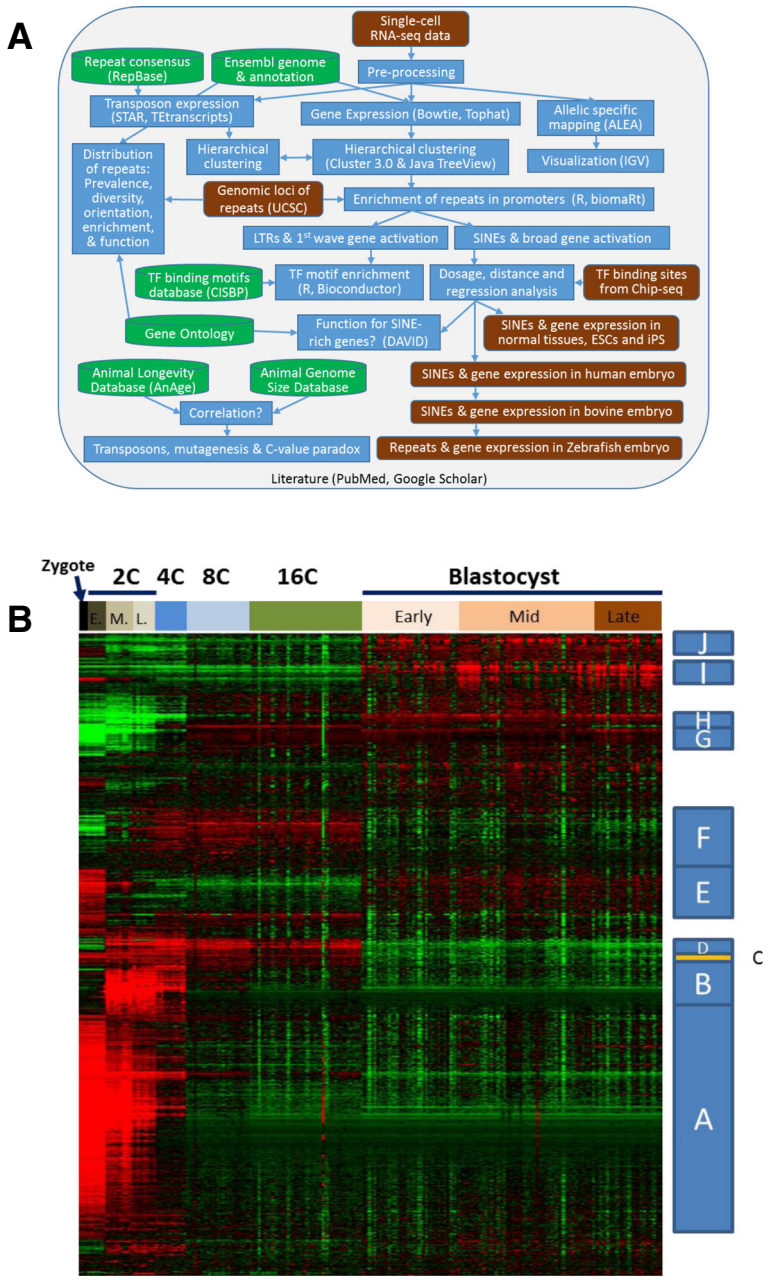

Fig. 1 a Exploratory bioinformatics investigation on gene regulation in mouse pre-implantation development. **b** Hierarchical clustering of gene expression during PD. Each of the 12,000 rows represents a gene. Columns correspond to samples labeled by developmental stages (E: early, M: middle, and L: late). *Red* indicates expression levels higher than average for the row. Expression lower than average is shown in *green*

during PD, with expression patterns mirroring those of regular genes.

Some MERV-L elements seem to give rise to a microRNA (miR-1194) from several genomic loci (Gm23215, Gm23551, Gm23943, Gm24617, Gm25042, and Gm26475). Low level expression of miR-1194 in ESCs has been detected [55]. Further study is needed to determine whether this microRNA aids the clearance of maternal mRNAs or transcripts from transposons.

Figure 1b also indicates the heterogeneity in expression profiles among cells of the same stage. This could be due to technical variations. However, there is a 16C sample identified as 16cell_4-2 with expression profile similar to early 2C stage or zygote, characterized by higher expression of oocyte-specific markers such as OOG3 (oogenesin 3), OOG4 (oogenesin 4). Few reads (1%) of this library map to paternal allele. There is another cell (mid-blast_2-11) at mid-blastocyst with the expression of

such genes. Reads from this cell contains substantial amount of paternal reads (33%), but these oocyte-specific genes seems to mostly originate from maternal alleles. Some variations are also observed in the expression patterns of TEs (Additional file 3: Figure S6). These may also be early signs of cell fate leading to different cell lineages [56, 57].

Enrichment of transposable elements in promoters

We systematically studied the distribution of all repetitive elements (REs) in the mouse genome (see Additional file 1: Table S4 for a list) based on RepeatMasker data available at UCSC genome browser website [53]. Covering 44% of the genome, the 5,147,736 REs are classified into 1554 different types with unique names and consensus sequences in the RepBase database [58]. They belong to 47 repeat families which are grouped into 16 repeat classes.

Figure 2a shows highly enriched REs in the 2 kb promoters of 10 gene clusters. B1 and B2 SINE elements are overrepresented in the promoters of Cluster C, D, F and G genes. Most of these genes are upregulated at the 2C stage (Fig. 1b). This is congruent with the fact that transcription of B1 increases at the 2C stage. For example, 32.7% of genes in cluster C contain at least one B1_Mus2 element in their promoter, which is much higher than the percentage (15.7%) in cluster A. Note that B1_Mus2 is just one of many forms of B1 elements.

Promoters of Cluster D genes are 9.6-fold enriched with MT2C_Mm LTR compared with other genes with false discovery rate (FDR) [59] of 3.02×10^{-11} (analysis of variance, ANOVA). The MT2C_Mm retrotransposon itself is transcribed between 2C and 16C, similar to Cluster D genes. MT2B2 and MLTR31D are also over-represented in the promoters of cluster D genes. Although relatively enriched, these LTRs are only found near a small proportion of Group D genes. As shown in Additional file 1: Table S5, about 5% of Cluster D genes contain MT2C_Mm elements, which is a 6.7-fold enrichment compared to all other clusters combined.

Promoters of Cluster B genes are enriched with other LTR elements, namely MERVL-int, ORR1A2-INT, ORR1A3-

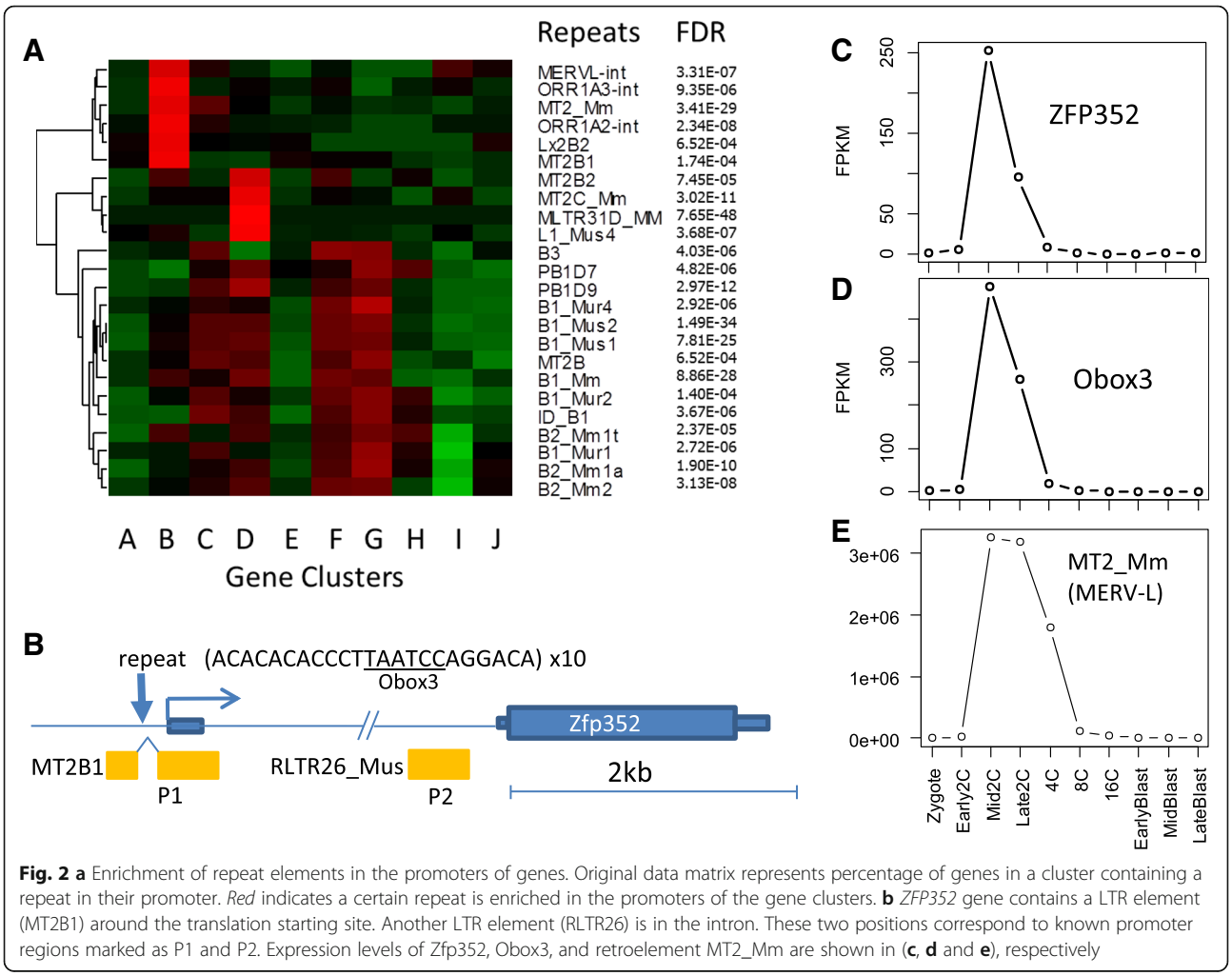

Fig. 2 a Enrichment of repeat elements in the promoters of genes. Original data matrix represents percentage of genes in a cluster containing a repeat in their promoter. *Red* indicates a certain repeat is enriched in the promoters of the gene clusters. **b** *ZFP352* gene contains a LTR element (MT2B1) around the translation starting site. Another LTR element (RLTR26) is in the intron. These two positions correspond to known promoter regions marked as P1 and P2. Expression levels of Zfp352, Obox3, and retroelement MT2_Mm are shown in (**c**, **d** and **e**), respectively

INT, MT2_Mm, and MT2B1. The most significant is a 9.6-fold enrichment of MERV-L (MT2_Mm) with FDR < 3.41×10^{-29} (*T*-test) compared with genes in other groups combined. This retrotransposon itself is transcribed at very high levels only in 2C-4C stages (Fig. 2e), similar to Cluster B genes, suggesting a shared mechanism of regulation. This agrees with previous reports [39]. Additional file 1: Table S6 includes 117 genes that contain ERVL elements in promoters and show an expression pattern similar to these LTRs.

As an example, transcription of *Zfp352* (zinc finger protein 352) starts at the middle of the LTR element MT2B1 (Fig. 2b). This gene is sharply upregulated at mid 2C stage, before quickly decreasing to very low levels at the 4C stage (Fig. 2c). This gene has been studied experimentally by Liu et al. [60], who found one major promoter P1 and an alternative, weaker promoter P2 in the intron. P1 actually lies within the MT2B1 element and P2 is in another LTR in the intron (Fig. 2b). *Zfp352* is likely generated by retrotransposition, similar to homologous pseudogene Zfp353-ps [61], where mRNA sequences are reverse-transcribed and inserted into the genome. Retrocopies of mRNA typically have no intron and are not expressed due to the lack of a functional promoter. But a subsequent or preceding insertion of LTR elements upstream of the retrocopy can provide a promoter. The expression of 117 such genes (See Additional file 2: Figure S8 for another example, Gm9125) was observed during PD when LTRs are actively transcribed. These expressed retrogenes can be further studied.

Early genes share motifs for Obox homeobox factors

We scanned the proximal promoter region (-300, 50 bp) of all the genes for transcription factor binding sites (TFBS) using the comprehensive CIS-BP database [62], which includes thousands of binding motifs determined using protein binding microarrays or inferred based on shared protein domains. Out of the 1823 binding motifs scanned, four were found enriched in the most highly induced gene group at mid 2C (Fig. 3a). Sharing the same "TAATC" core sequence, these four motifs can be bound by OBOX1, OBOX3, OTX1, PITX2, and other factors. Noticeably, the *Obox3* gene is upregulated by more than 100-fold at mid 2C, and quickly downregulated at 4C (Fig. 2d). This is similar to the expression pattern of its potential target genes such as *Zfp352* (Fig. 2c), suggesting a potential regulatory role of Obox factors.

Oocyte specific homeobox (Obox) family TFs have not been studied in detail because they seem to be rodent-specific and are only expressed during PD [63]. They are highly regulated in oocytes and cleavage stage embryos

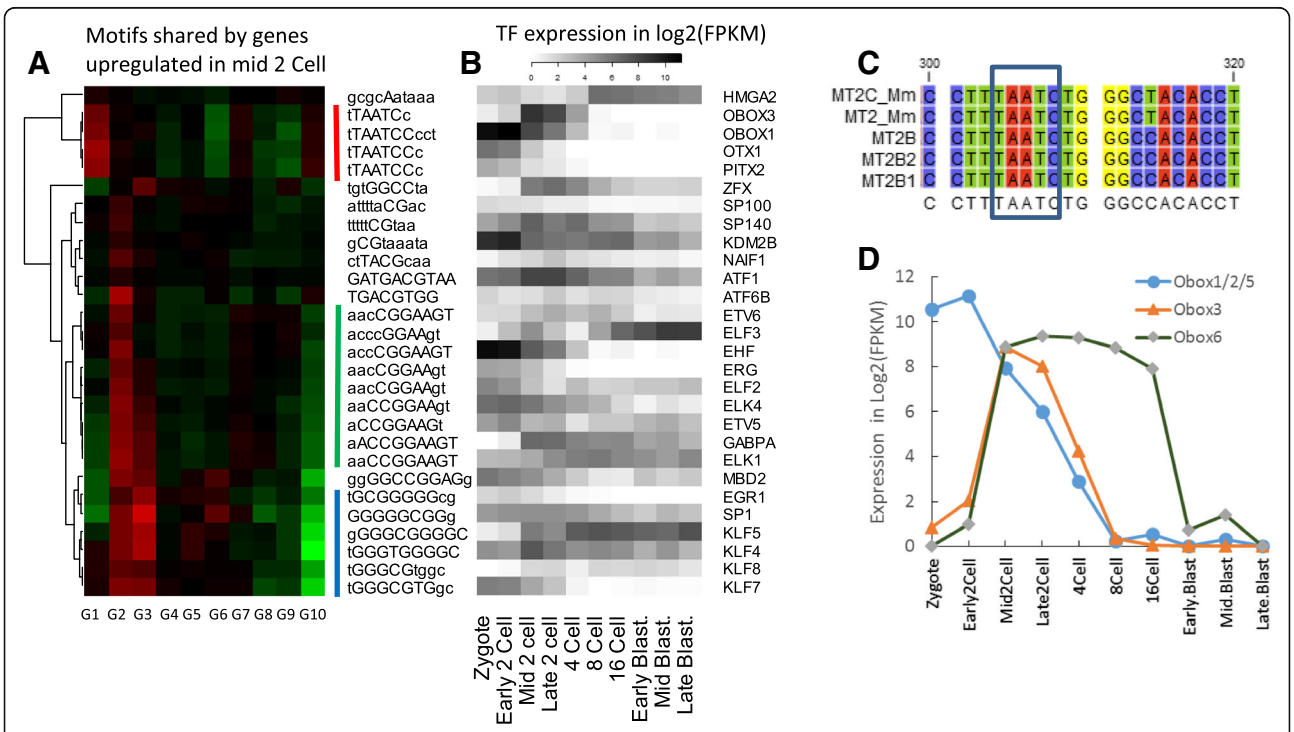

Fig. 3 Enrichment of transcription factor binding motifs in genes highly induced at mid 2-cell stage compared with early 2-cell stage. **a** Enriched motifs in the promoters of genes with highest fold-change at mid 2C. G1-G10 represents 10 groups of 500 genes sorted by fold-change. *Red* indicates enrichment of motifs. **b** The expression patterns of TFs that bind to the corresponding motifs shown in (**a**). **c** Binding motifs by Obox families in LTRs belong to the ERVL family. **d** Expression patterns of Obox family homeobox factors are highly regulated during PD

(Fig. 3d). They are among the top 10 TFs when ranked by variance in gene expressions during PD. Since these TFs bind to similar motifs (Additional file 2: Figure S9), it is difficult to distinguish the true driver of gene expression. Three members (Obox1/2/5) are similar in sequence, and are treated as one group in our RNA-seq mapping; they are highly expressed in oocytes and are maternally derived. Due to their higher expression at 2C and shared binding motifs, Obox1/2/5 could work together with Obox3 to regulate gene expression. Experimental studies are needed to confirm whether these factors are redundant or have differences. Obox6 expression is elevated from 2C to 16C, and low in the blastocyst stage (Fig. 3d), in agreement with a previous study [64].

Obox binding sites are enriched in ERVL family LTR elements that are overrepresented in the promoters of cluster B genes (Fig. 3c). For example, the 493 bp consensus sequence of MERV-L contains three Obox binding sites, while the 521 bp MT2B1 contains four. Although not as highly transcribed as MERV-L, MT2B1 is one of reliable predictors of nearby gene expression among LTRs, as the corresponding transcripts almost always start right from the repeat, similar to *Zfp352*. Interestingly, the MT2B1 element in the main promoter of *Zfp352* is interrupted by a (10x) tandem repeat of 21 bp sequence, which contains the Obox binding motif (Fig. 2b). These tandem repeats may be selectively retained during evolution. Furthermore, genes with multiple Obox binding motifs are more highly induced at mid 2C when compared to genes with one or no such motifs (Additional file 2: Figure S10). The retrovirus-like LTRs contain their own promoters and enhancers. These promoters may be bound by Obox TF families to drive the expression of both retrotransposons and nearby genes.

Among the genes with transient 2C expression is the Zscan4 family (Zscan4b, Zscan4c, Zscan4d, Zscan4e, and Zscan4f). Zscan4 proteins are involved in telomere elongation and genomic stability in ESCs [65], and have been reported to restore developmental potency in ESCs [66] and to transiently activate embryonic genes in induced pluripotent cells (iPSCs) [67]. Among the genes transiently expressed at 2C, Zscan4 genes ranked highest in terms of number of Obox3 binding sites, with three binding sites at 90% similarity level in the 350 bp around the TSS. Further study is needed to verify if Obox family TFs are upstream regulators of Zscan4 genes, and whether induction of Obox gene expression in adult cells could promote pluripotency.

Thus, there is evidence that the poorly-characterized Obox family induces gene expression at mid 2C to jump start ZGA. Even though Obox6 mutants develop normally [64], loss of function of other family members may be lethal to the embryo, due to potential effects on hundreds of downstream genes and even LTR transposons.

We also examined the expression of these genes in another single-cell RNA-seq dataset [20], and found that the patterns are similar, except that Obox3 is expressed lower (Additional file 2: Figure S11). Using this this dataset, we also identified the enrichment of the same "TAATC" motif among genes upregulated in 2C stage (Additional file 2: Figure S12).

The overrepresented "TAATC" motif can also bound by other homeobox TFs like OTX1 and PITX2 (Fig. 3b and c). These TFs may also regulate gene expression, as their transcripts are also present in oocyte, even though at a much lower level (Fig. 3b). Recently, a similar motif was discovered by Töhönen et al. [22] in the promoters of 32 human genes upregulated in early ZGA using single-cell RNA-seq. This agreement may suggest a conserved mechanism of ZGA, and warrants experimental confirmation.

SINEs associated with broad genome activation

To further study expression change in late 2C, we ranked genes by their fold-changes from late 2C compared to mid 2C, and then divided them into 24 groups of 500 genes. As shown in Fig. 4a–c, the REs differentially distributed in the promoters of these gene groups are mainly SINE elements, including the Alu family (mainly B1) and B2 family. The most significantly associated elements are B1_Mm, B1_Mus1, and B1_Mus2 (Fig. 4c). These B1 elements are highly similar in their sequences, with only a few base-pair differences in most cases (Additional file 2: Figure S13). With over 400,000 copies, B1 is one of the most prevalent retrotransposons in the mouse genome. Similar to human Alu elements, B1 originated from initial duplication of the 7SL RNA [68]. The distribution of SINE, but not LINE, is conserved across species [69]. B2 elements originated from tRNAs [68]. It is possible that B1 and B2 elements play a role in gene regulation during PD, and their fast expansion may have benefited mammalian development.

Figure 4b also shows that genes downregulated at the 2C stage are depleted in these elements. The average fold-change observed in these groups is highly associated with the percentage of genes containing Alu family elements in their promoters. As shown in Fig. 4d, the Pearson's correlation coefficient (PCC) is 0.90 ($P < 3.0 \times 10^{-9}$, test of association) for one of such element B1_Mm. About 36% of the 1500 genes that are upregulated by more than 2 fold contain B1_Mm elements, which is much higher than the 18% observed in gene downregulated by 2 fold. In addition to B1_Mm elements, other B1 and B2 elements also show this trend. Using a different dataset of single-cell RNA-seq data [20], we were able to confirm this remarkably linear and consistent correlation (Fig. 4e).

Multiple B1 elements in promoters are associated with stronger upregulation in a dosage-dependent manner (Fig. 5a–d). As shown in Fig. 5c, the 496 genes with five

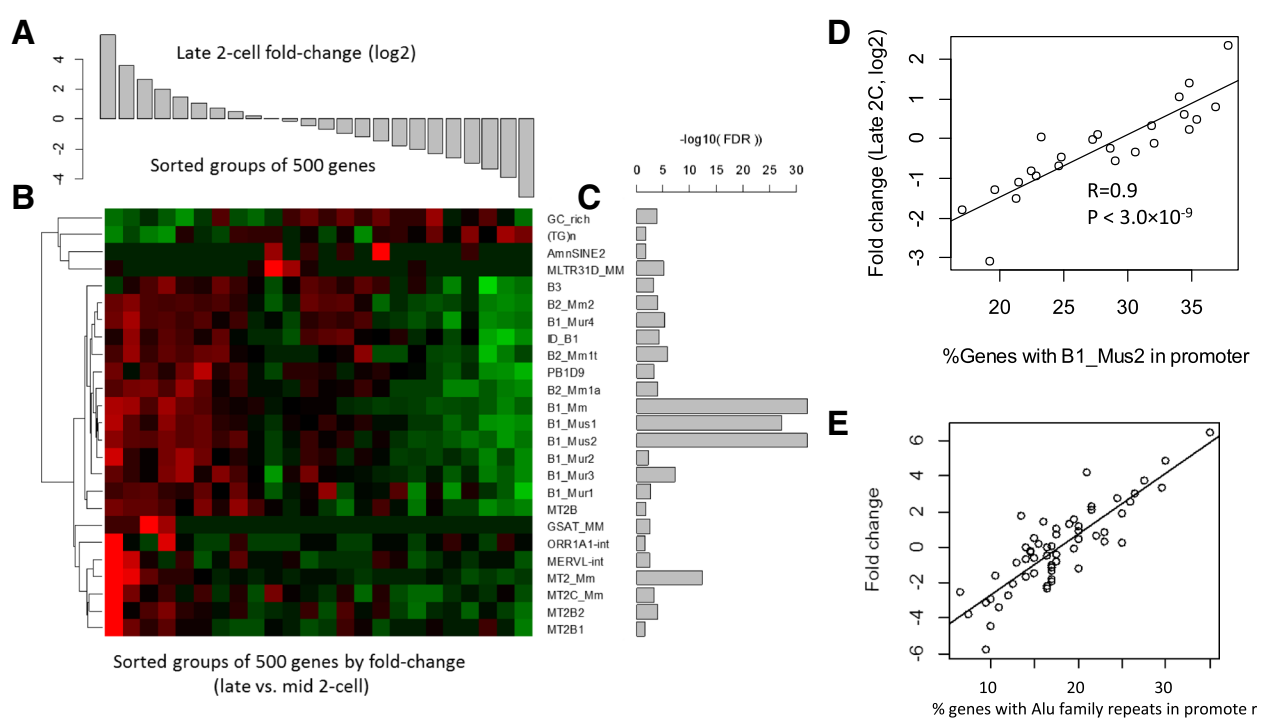

Fig. 4 Enrichment of SINE elements in genes upregulated in the late 2-cell stage. **a** Average fold-change for 24 groups of 500 genes. **b** Relative enrichment (*red*) or depletion (*green*) of repeat elements in the promoters of genes in these groups. **c** FDR values derived from Chi-square tests of independency of repeat element frequency and gene groups. **d** Correlation of fold-change in 2C with the presence of B1 elements in gene groups. Each point in the plot represents one group of 500 genes. The *vertical axis* represents the percentage of genes in the group with B1_Mus2 in their promoters, while the horizontal axis represents average fold-change. **e** Association of repeats with fold-change during the late 2-cell stage. The average fold change of genes with one or more repeats within 2 kb on either sense or antisense strand

B1 elements or more (some are partial) in promoters are upregulated by 2.1-fold on average, which is significantly higher than the 1.65-fold observed in 568 genes with 4 B1 elements ($P < 0.0064$, *T*-test). This is in turn higher than genes with three B1 elements ($P < 0.027$, *T*-test). The effect of B1 elements is surprisingly linear. Each additional B1 element is associated with an approximately 20–40% upregulation (Additional file 2: Table S7). Among genes that are upregulated more than 2 fold at late 2C stage, 59% contain Alu family repeats in promoter, which is much higher than the 41% observed other genes.

The effects of the B1 elements are also dependent on the distance to TSS (Fig. 5d). The correlation is stronger when B1 elements are located closer to TSS. Further upstream, the effect of Alu elements is weaker, diminishing after 5–7 kb. This is similar to what was observed in the MT2 elements (Fig. 5b). B1_Mm and other mouse-specific B1 elements are more efficient than other Alu family repeats common in muridae (B1_Mur1-4) or rodentia (PB1D9). But it is difficult to isolate a specific type of element, as different subtypes of Alu family repeats often co-appear.

To systematically investigate the effects of various repeats on ZGA, we used multiple linear regression to model 2C fold change as a function of the numbers of various kinds of repeats within the 2 kb promoter. The model also included CpG islands and TFBS of 13 TFs determined by Chip-Seq [70]. The results (Fig. 5e) confirm the effect of SINE and LTR elements. While the effects of ERVL family LTRs are strong and only seen on the same strand, B1 elements have weaker effects on both strands. Genes with multiple c-Myc sites show a bigger fold-change. E2F1 binding sites are associated with weaker but significant upregulation in a larger number of genes. The effect of Alu elements and c-Myc and E2F1 binding sites can be confirmed using independent single-cell RNA-seq data [20] (Additional file 2: Figure S14).

SINEs are a major source of CpG dinucleotides in mammalian genomes. It is possible that the effect of Alu elements in promoters is through the contribution of CpG sites that affect epigenetic modifications. Linear regression analysis shows B1 elements have a much more significant ($P < 2.2\times10^{-16}$) correlation with ZGA fold-change than those of CpG dinucleotides ($P < 0.04$), or CpG islands ($P < 0.004$). Also, gene clusters defined by methylation of promoter regions during PD [26] (Additional file 2: Figure S15) have little in common with the gene clusters by expression (Fig. 1b). Therefore, the effect of B1 repeats cannot be fully explained by CpG sites.

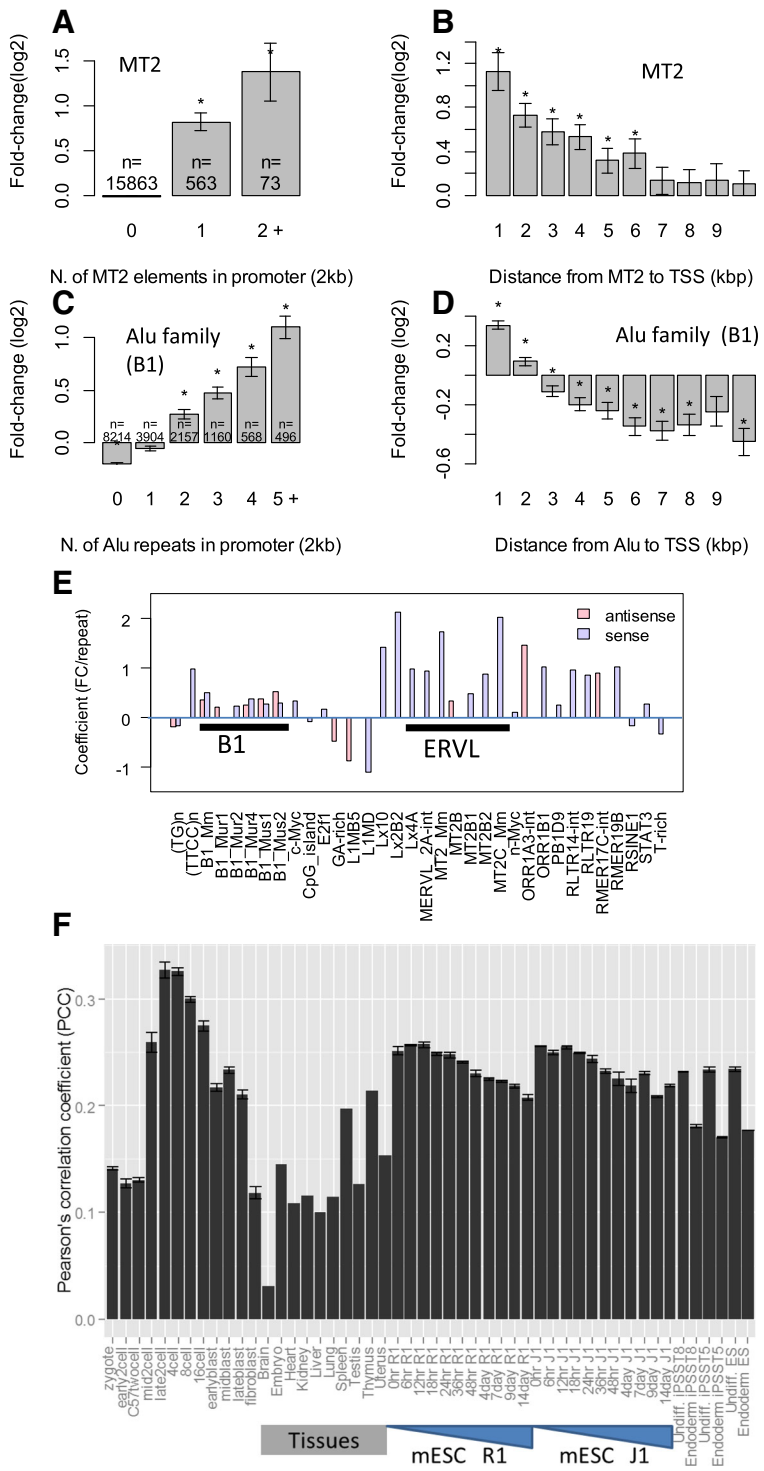

Fig. 5 Presence of some repeats strongly correlated with activation of gene expression during 2C stage in a dosage- and distance -dependent manner. The heights of the bars indicate average fold-change in late 2-cell stage compared to early 2-cell. Significant deviations from zero are indicated by stars. The numbers on the bar represent the number of genes affected. The MT2 repeats in (**a**) and (**b**) can be MT2B, MT2B1, MT2B2, MT2_Mm, or MT2C_Mm. In (**c**) and (**d**), Alu family repeats mostly represent B1 elements. **e** Result from multiple linear regressions shows that SINE and LTR elements are correlated with gene expression change. **f** Correlation of the number of Alu family repeats in promoters and broad gene expression in various cells and tissues. Error bar shows standard error calculated from replicates. R1 and J1 are mouse embryonic stem cell lines

SINEs correlate with gene expression in adult tissues and stem cells

It has been reported that Alu family repeats are enriched near housekeeping genes [71, 72], which are both broadly and highly expressed across tissue types. We calculated the correlation coefficient between expression levels and the number of Alu family repeats in promoters among genes. As shown in Fig. 5e, there is a significant positive correlation for all cells and tissues. The PCC dramatically increases from 0.13 at early 2C to 0.33 at late 2C, and gradually decreases to 0.21 at the late blastocyst stage. PCC is higher in embryonic cells than in adult tissues. This association is independent of CpG islands (Additional file 2: Figure S16).

Occurrences of Alu family repeats are a stronger predictor of gene expression in undifferentiated iPSCs compared with day 5 definitive endoderm [73]. A similar pattern of correlation is observed with B2 family repeats (Additional file 2: Figure S17). SINEs may play a role in regulating gene expression in both pre-implantation embryos and ESCs.

Urrutia et al. [74] found no association between Alu content and peak (maximum) gene expression level across tissues. However, we found a significant correlation ($R = 0.16$, $P < 2.2 \times 10^{-16}$, test of association) with average gene expression. In addition, even among genes expressed in all tissues, more Alu elements in the promoter are associated with higher average expression ($R = 0.13$, $P < 2.2 \times 10^{-16}$, test of association). The same is true for tissue-specific genes ($R = 0.16$, $P < 2.2 \times 10^{-16}$, test of association).

Promoters of ESC–specific genes are rich in SINEs and low in CpG_island

To further delineate the role of Alu family repeats in gene regulation, we re-analyzed RNA-seq data [75] of normal tissues from both fetal and adult mice, as well as ESCs. We divided 16,989 protein-coding genes into 25 groups using k-means clustering based on their expression pattern in various tissues/cells (Fig. 6a). In addition to ubiquitously-expressed housekeeping genes, we identified many clusters of tissue-specific genes. Cluster 13 contains 89 genes specifically expressed in mESC cells. As shown in Additional file 2: Figure S18, this includes known TFs (Nanog and Pou5f1), as well as many other factors such as Zscan10, Esrrb Foxn4, and Sox15. These genes contain many Alu family repeats in their

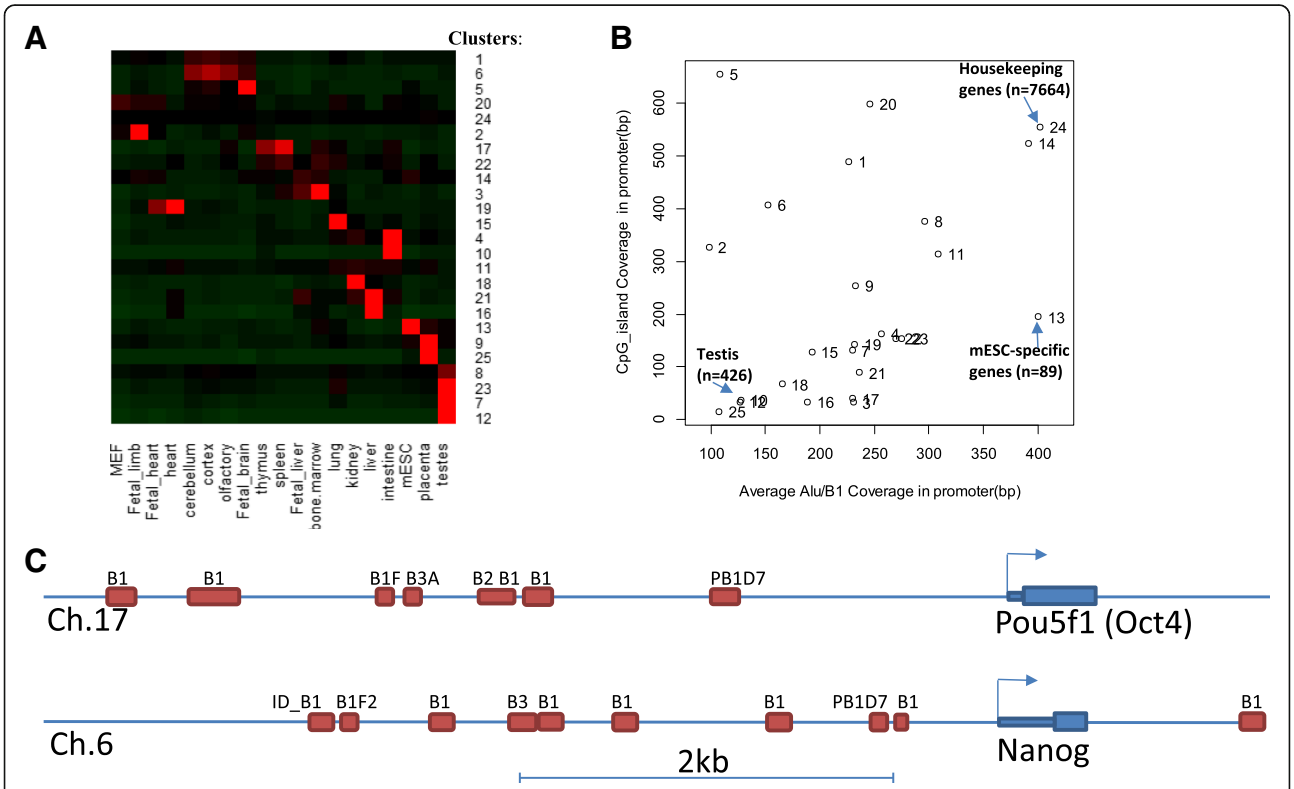

Fig. 6 Genomic content of promoters of mESC-specific genes. **a** Results of k-means clustering of normal tissue gene expression show housekeeping genes (Cluster 24) and tissue-specific genes. **b** Gene clusters are plotted by average coverage of Alu/B1 repeats and CpG island coverage. Housekeeping genes (Cluster 24) are high in both Alu/B1 coverage and CpG island. mESC specific genes (Cluster 13) are high in Alu/B1, but low in CpG island. Gene Clusters specifically expressed in testis, intestine, and placenta are lower in both. **c** Devoid of CpG islands, the promoters of Pou5f1 and Nanog are enriched with Alu/B1 elements

promoters (–5kbp to 1 kb). As shown in Fig. 6b, their average Alu/B1 coverages of the promoters of these genes are as high as housekeeping genes, much higher than other tissue-specific gene clusters. Unlike the promoters of housekeeping genes, the promoters of ESC-specific genes are less likely to have CpG islands. For example, the Alu-rich promoter regions of Nanog and Oct4 are shown in Fig. 6c. Human orthologs of these two genes are also enriched with Alu elements. SINE elements might be important for the expression of pluripotency related genes.

SINEs correlate with ZGA in other species

Single-cell RNA-seq data of early human embryogenesis [20] were used to investigate the correlation between transposons and ZGA, which occurs between 4- and 8-cell stages [76]. Regression analysis at the repeat family level shows a significant association between the presence of Alu family repeats (FDR = 1.74×10^{-50}) and gene expression (Additional file 2: Table S8). The most prevalent Alu element, AluJb, is significant on both sense and antisense strands. Presence of an Alu element is associated with 15–48% upregulation. Figure 7a shows that genes with multiple Alu elements in promoters are upregulated in a dosage-dependent manner, similar to what was observed in mouse. As suggested by Fig. 7a–b, genes with fewer than two Alu elements in the promoter are downregulated at 8C, similar to what was observed in mouse. This is in agreement with another single-cell RNA-seq analysis of human preimplantation embryos, which shows that Alu elements are enriched in upstream of TSS of the 129 genes upregulated during PD [22]. Alu elements were found to contain binding motifs for PITX1 and TBX1 [22].

We also compared the occurrences of human Alu and mouse B1 elements in promoters of orthologous gene pairs. The number of Alu elements in human gene promoters is highly correlated with the number of B1 elements in orthologous mouse gene promoters ($R = 0.57$). This has been noted as surprising, as B1 and primate Alu elements replicated in these genomes independently [77]. The differences in ZGA fold-changes between orthologous gene pairs in human and mouse are significantly associated with the differences in the number of Alu family repeats in their promoters ($P < 1 \times 10^{-14}$, regression analysis). The rapid expansion of B1 in the mouse and Alu in the human genome may contribute to gene expression divergence.

Bovine ZGA is associated with tRNA SINE repeats. We used RNA-seq data based on pooled embryos [19].

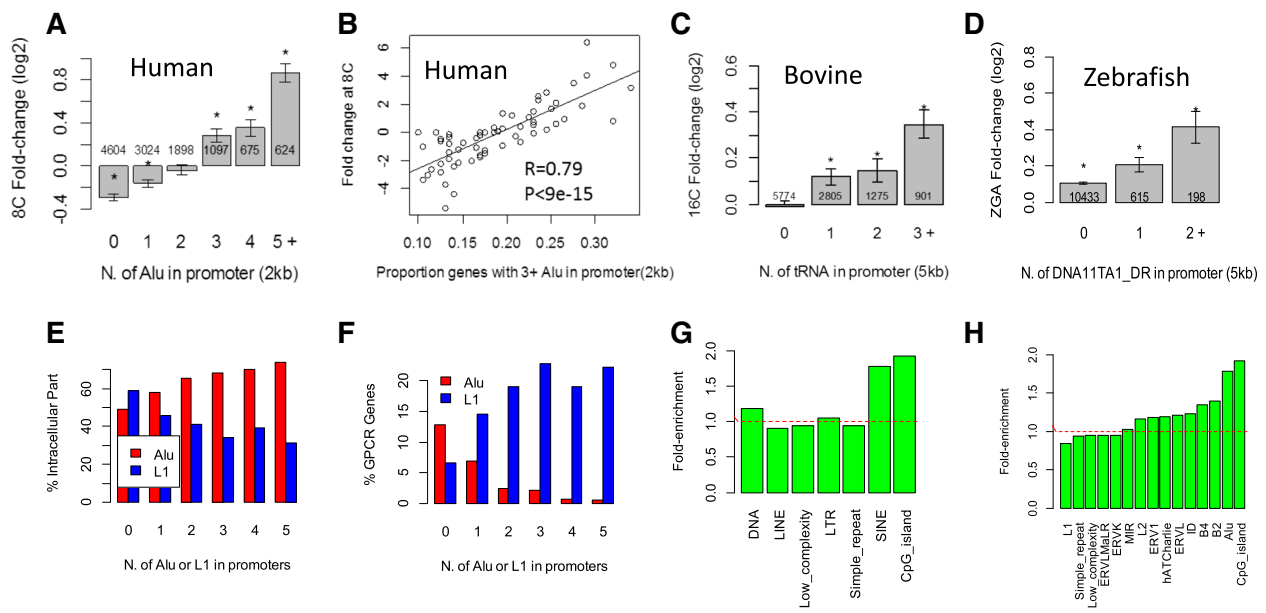

Fig. 7 **a** Genes with more Alus in promoter are upregulated in a dosage-dependent manner. *Stars* indicate significant difference from zero based on *t*-test. **b** Among gene groups defined by fold-change, highly expressed ones tend to contain more Alus in promoters in humans. **c** Association of gene expression with tRNA family of SINE repeats in the bovine genome. *Stars* indicate significant difference from zero. **d** DNA transposon, DNA11TA1_DR, is associated with gene upregulation during ZGA in zebrafish in a dosage-dependent manner. **e** Mouse genes with multiple Alu family repeats, mostly B1 elements, are associated with GO:0044224, intracellular part. Genes with L1 elements in their promoter, on the other hand, are depleted in genes related to intracellular part. **f** Genes with L1 elements are enriched with GPCR activity, while in genes containing Alu elements, such genes are depleted. **g** Mouse genes with SINE element in promoters are enriched in genes with yeast orthologs (>20% identity according to BioMart), compared to genes that does not contain such elements in promoters. **h** Among SINE elements, Alu family, mostly B1 elements, are enriched with genes with yeast orthologs

The repeat families associated with expression change during ZGA between 8C and 16C are ERV1, simple repeat, and tRNA. The tRNA family SINE repeats in bovine have a weaker but significant association (FDR < 1.1×10^{-5}) based linear regression. It is also dosage-dependent (Fig. 7c), as genes with three or more tRNA family repeats are more highly upregulated than genes with one or two such repeats ($P < 0.011$, T-test). The tRNA family repeats are associated with higher gene expression in both sense and antisense strands. The most significant repeats are SINE2-1_BT and SINE2-2_BT, which are 120 bp bovine-specific SINE repeats derived from tRNA. The association with tRNA family repeats can also be confirmed using DNA microarray data [16].

ZGA in zebrafish is associated with AT-rich DNA transposons (Additional file 2: Figure S19). The zebrafish (*Danio rerio*) genome [78] is dominated by more than 2 million DNA transposons. There are fewer retrotransposons compared with mammals. Using RNA-seq, Harvey et al. studied the zebrafish ZGA [79], which happens at about 3.5 h post-fertilization. Results from regression analysis (Additional file 2: Tables S9 and S10) show that some DNA transposons are highly associated with gene upregulation at ZGA. The most significant is DNA11-TA1_DR, a non-autonomous DNA transposon. There are 813 genes containing this repeat in the 5 kb promoter region, and their expression is significantly higher than other genes (FDR < 1.88×10^{-9}, T-test). The 198 genes with two or more DNA11TA1_DR elements are induced at significantly ($P < 0.03$, T-test) higher levels than the 615 genes with one element, which is in turn higher ($P < 0.013$, T-test) than genes without such an element (Fig. 7d).

The accumulation of SINEs near genes is likely to result from a positive selection process [42]. Since there is no known mechanism to remove SINEs from genomes, it is difficult to explain the lack of SINEs in gene-poor regions [42]. The transposons significantly correlated with ZGA are often prevalent in and specific to the host species. Specific transposons seem to be encouraged to expand during the course of evolution.

Non-random distribution of transposons in the genome

We studied the distribution of all repeats across the mouse genome using the EBI approach. While some REs like B1 are prevalent, others are only observed dozens of times. We found that the frequency follows lognormal distribution (Additional file 2: Figure S20). Lognormal distribution implies that growth rate is independent of existing occurrence [80]. The distribution of the distances between repeats is power-law like [81], which could be expected as transposons often "copy-and-paste" to nearby loci and form clusters on the genome.

Some repeats show enrichment and strand-preference near genes (Additional file 2: Figure S21). We found that SINEs are enriched in introns, promoters, and downstream regions, suggesting that SINEs are located near genes. On the contrary, LINEs are depleted from these regions and are away from genes. There are 97 types of LTRs that are specifically enriched in promoter regions. Some repeats demonstrate strand-preference. For example, RLTR10-int repeats align 4.6-times more frequently on the same strand relative to the nearby gene than on the opposite strand, which is highly unusual ($P < 1.4 \times 10^{-74}$, chi-squared test). There are a total of 14 repeats that are enriched in a strand-specific manner, including several prevalent ERVL-MaLR family members (MTC, ORR1D1, ORR1A2, ORR1A2-int), ERVK family members (RMER19B, MYSERV6-int, MYSERV-int, RLTR10, RLTR10-int, MLTR18A_MM, and RLTR9A3A), the ERVL family (MT2B1, RMER15-int), and the ERV1 family (LTR72_RN). This could be explained by new LTR elements generating new genes by activating retrogenes, as discussed. We have shown that MT2B1 contains Obox3 binding sites and is strongly associated with gene expression during PD. Other elements may regulate gene expression in other situations [82].

Surprisingly, most intronic retrotransposons are more likely found on the opposite strand of the host gene (Additional file 2: Figure S21). It is possible that intronic sequences, once spliced off transcripts in the nucleus, are spliced, reverse-transcribed, and inserted back into the genome, resulting in intronic retrotransposons on the antisense strand. DNA transposons do not show such a strong strand-specificity. More investigation on the intronic strand-specificity is needed to verify this possible mechanism.

Genes with multiple B1 in promoters form core cellular machinery

Based on Gene Ontology (GO) [83], we found that genes with multiple B1 elements in promoters are more likely to code for proteins that constitute intracellular parts (Fig. 7e) and less likely to be related to G-protein coupled receptor (GPCR) activity, an extracellular signaling process (Fig. 7f). On the contrary, genes with L1 LINE elements in their promoter are enriched in GPCR related genes and depleted of intracellular parts. Low complexity repeats are found to be enriched in promoters of genes related to the RNA metabolic process (Additional file 2: Figure S22A). Thus, the distribution of repeats in the genome is related to the function of genes. It is possible that the mouse embryo utilizes B1 elements to quickly establish core proteins during ZGA.

Mouse genes with SINEs in promoters are more evolutionarily conserved. Figure 7g–h suggests that mouse genes with SINEs in the promoter are much more likely to have yeast orthologs (>20% identity according to BioMart [84]). This is similar to, but independent of, the effect of CpG islands (Additional file 2: Figure S22C). A similar trend is observed in human genes (Additional file 2: Figure S23). Thus, the distribution of transposons is correlated with the function of nearby genes. The distribution of transposons may be regulated through selection.

Key transcription factors in early development

Taking advantage of high-resolution expression data [21], we also systematically analyzed TFBS in the promoters of genes co-regulated at other stages beyond 2C. Enriched TFBS and expression patterns of corresponding genes are shown in Additional file 2: Figures S24–S41, which are summarized in Fig. 8a.

We identified several TFs known to regulate embryo development, such as SOX2, OCT4 (POU5F1), and KLF4 [85, 86]. In addition, many TFs in Fig. 8a are upregulated at the same developmental stage as their potential target genes, thus giving more support to their involvement in gene regulation. This includes Obox3 and KLF4, which are upregulated at mid 2C, as well as NR2C2, Zscan10, and ELK1 at late 2C. Zscan10 is known to be expressed during PD [87] and is involved in maintaining pluripotency in ESCs [88]. Similarly,

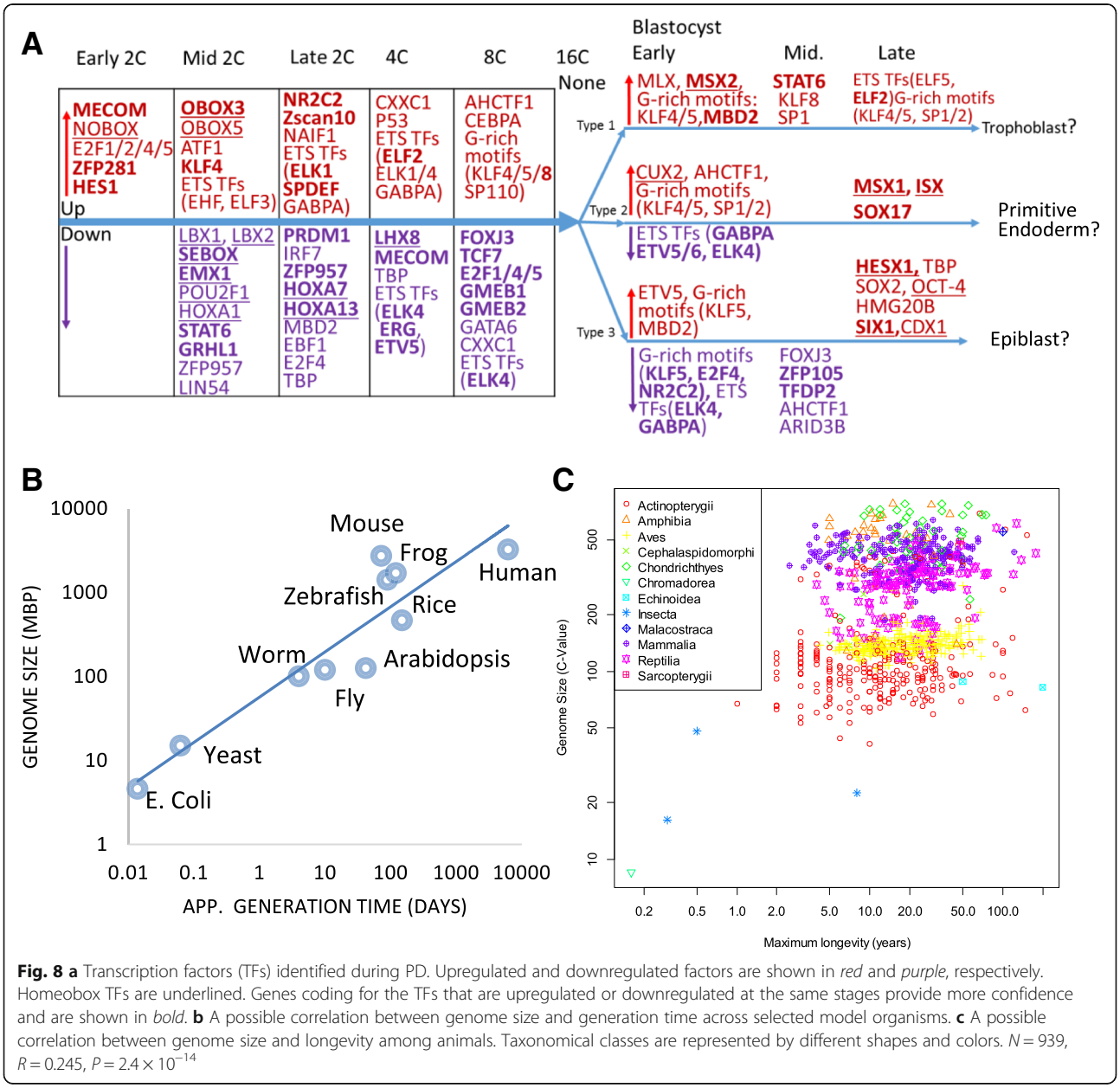

Fig. 8 a Transcription factors (TFs) identified during PD. Upregulated and downregulated factors are shown in *red* and *purple*, respectively. Homeobox TFs are underlined. Genes coding for the TFs that are upregulated or downregulated at the same stages provide more confidence and are shown in *bold*. **b** A possible correlation between genome size and generation time across selected model organisms. **c** A possible correlation between genome size and longevity among animals. Taxonomical classes are represented by different shapes and colors. $N = 939$, $R = 0.245$, $P = 2.4 \times 10^{-14}$

many motifs enriched in the promoters of downregulated genes are bound by TFs that are downregulated at the same time. Some of the TFs are likely maternally derived with high expression in the zygote and reduced at the 2C stage: LBX1, LBX2, SEBOX, ZFP959, POU2F1, STAT6, GRHL1, EMX1, PRDM1, HOXA7, LHX8, EBF1, and E2F4. SEBOX is known as one of the maternal effect genes (MEGs), and its RNA products are carried over from oocyte to regulation of gene expression in early PD [89]. Further study should verify whether other TFs in this list are MEGs.

The GGAA motif bound by the E26 transformation-specific (ETS) domain TFs is repeatedly identified as enriched at several stages. One of the TFs, GABPA, has been shown to be involved in early embryogenesis [90]. The most highly expressed ETS domain TFs in oocyte and 2C is EHF. ELF3 is highly expressed at the blastocyst stage. ETS domain TFs are a large and conserved family of TFs involved in a variety of developmental processes in animals [91, 92].

Similarly, the G-rich motifs are repeatedly identified at various stages. This motif can be bound by KLF4 or other TFs, such as KLF5, KLF8, SP1, SP2, ZBTB7B, etc. The most highly expressed is KLF5, which was reported to regulate lineage formation in the pre-implantation embryo [93], alongside KLF4 [94].

To account for the heterogeneity among cells in the blastocyst stage, we used hierarchical clustering to divide them into three types/lineages, which were then analyzed separately. Sox2 and OCT4 binding sites are enriched in promoters of genes upregulated in Type 3 cells at the late blastocyst stage. These markers indicate that type 3 cells likely correspond to the epiblast, which is derived from the inner cell mass (ICM) and leads to the embryo proper [95]. Figure 8a shows that two homeobox TFs, HESX1, and SIX1, are both upregulated together with their target genes in this type of cell. HESX1 is believed to be downstream of multiple pluripotency related pathways [96]. SIX1 is considered to be an oncogene and is known to be involved in embryonic muscle formation [97, 98].

In Type 2 cells, SOX17 binding sites are overrepresented in promoters of genes upregulated at the late blastocyst stage, when the SOX17 gene itself is also upregulated. SOX17 directly promotes differentiation towards extraembryonic cells, which leads to the primitive endoderm [99], which is also derived from ICM but develops into the yolk sac. Therefore, type 2 cells are likely committed to the primitive endoderm. MSX1 and ISX are two other homeobox TFs identified as inducing gene expression in these cells. Little is known about ISX in embryogenesis, but it is highly induced at the late blastocyst stage in type 2 cells. Interestingly, we found *MSX1* and *MSX2* are associated with primitive endoderm and trophoblasts, respectively. Their differential role should be further studied.

Type 1 cells may represent trophoblasts, which make up the outer layer of the blastocyst. In the promoter of genes highly induced in this type of cell, we failed to detect binding sites for CDX2, a key regulator for trophoblasts [95, 100]. But sites bound by MSX2 are enriched. MSX1 and MSX2 were shown to be critical for the interaction between the blastocyst and the uterus [101, 102]. These two proteins are highly conserved in mammals [103], and mutations of *Msx1* and *Msx2* lead to failures in implantation [101, 102].

Our analysis identified many other TFs that could potentially contribute to the complex gene regulatory network during PD. For example, *MECOM* is upregulated at early 2C and downregulated at 4C, along with its potential target genes. MECOM is highly expressed in the embryo, and mutant is embryonic lethal [104, 105]. Figure 8a also shows that we identified many homeobox TFs, which are believed to be regulators of morphogenesis and development [106]. In addition, many TFs in Fig. 8a have been studied in relation to embryogenesis and cancer. Further study of these TFs will elucidate their role in gene regulation in PD.

Discussion

Transposons, mutagenesis, and the C-value paradox

Some organisms in our ecosystem can finish a reproduction cycle in 20 min, while others require more than 10 years to reach sexual maturity. If genetic mutations happen stochastically at similar rates, a huge imbalance in how quickly organisms evolve and adapt would result. Slow-reproducing organisms, therefore, are under pressure to find ways to dramatically promote mutagenesis and genotype diversity. Transposition of mobile DNA elements may be a necessary "copy-and-paste" mechanism that promotes not only insertional mutations but also homologous recombination. We showed here that they may also regulate many coding genes during early development and thus will have substantial influence of morphogenesis. Some retrotransposons are even active in somatic cells and lead to genetic mosaicism within individuals [107]. There is additional evidence that TEs are drivers of genome evolution [39, 41, 48, 108], rather than just "junk" DNA.

Following this argument, we would expect slow-reproducing organisms to have more TEs in their genome, leading to larger genomes. Across model organisms, generation time seems to be proportional to genome size on a log-scale (Fig. 8b). Using larger datasets [109, 110], longevity and genome size are weakly but significantly correlated (PCC = 0.245, $P = 2.4 \times 10^{-14}$) across 939 animal species (Fig. 8c). Further study is needed to confirm the correlation in more organisms. But this may shed some light on the C-value (genome size) paradox [111, 112]: eukaryote haploid DNA contents vary greatly, but are unrelated to organismic

complexity. Even though TEs can be disruptive for individuals, they might be necessary for the adaptation and survival of the species, especially in slow-reproducing organisms. Otherwise, it is hard to imagine how the accumulation of tens of thousands of transposons near essential genes could be tolerated over millions of years. Our analysis show that these TEs may influence development from an early formative stage. Thus expansion of different TEs that contribute to the rewiring of developmental pathways may facilitate speciation and adaptation.

Gene regulation by transposons

TEs can be involved in epigenetic regulation, as they can recruit the silencing machinery [26, 27, 43]. Many examples have been reported that retrotransposons can serve as TF binding sites to promote nearby gene expression. In addition to LTRs, which contain promoters that can be used to drive expression of nearby genes [39], LINEs elements contain an antisense promoter that can be used by nearby genes [40, 113, 114]. MER20, a DNA transposon, was found to contribute to pregnancy-related gene network and its evolution [48]. In ESCs, Kunarso et al. [41] reported that about 25% of POU5F1, NANOG and CTCF binding sites are provided by TEs. Thus TEs is important in the regulatory network of ESCs [44]. More importantly, retroviral activity was found to be a hallmark of pluripotency [33, 34, 38, 115]. ERV-derived LTR elements may be contributing to the gene regulatory network of innate immunity [116]. Adding to these results, this study systematically investigated the correlation of TEs and the genomics reprogramming in PD. We show that TEs maybe play a more profound role than previously thought, affecting thousands of genes. We also provide some evidence for the potential role of SINEs in activating housekeeping and ESC-specific genes.

Possible mechanisms of B1 and Alu in gene regulation

Mouse B1 and human Alu elements originated from 7SL RNA [68] and contain RNA polymerase III promoters. The A-box and B-box included in Alu sequences are bound by a multi-subunit transcription factor TFIIIC, to form the Pol III complex. Although some microRNAs are shown to be transcribed by Pol III using upstream Alus [117], it is unlikely that Pol III would produce thousands of essential genes. This would predict a strand-specific correlation and alternative TSS', similar to LTRs. Alu family repeats also contain binding sites for many factors associated with RNA Pol II [118, 119], including p53 [120], retinoic acid receptors [121, 122], YY1 [123], PIT2 [124], etc. Our analysis shows that B1 elements also contain TFBS' for Obox family proteins, especially Obox3. These TFs may act upon Alu elements to drive gene expression.

In addition, it is well documented that TFIIIC, without the rest of the Pol III apparatus, has the so-called extra-transcriptional effects (ETC) ranging from nucleosome positioning, genome organization, and direct effect on Pol II transcription [125, 126]. Alu provides the majority of the TFIIIC binding sites in humans and mice [127]. B2 elements are also enriched with CTCF binding motifs defined by Chip-Seq [16, 26].

Some human Alus serve as estrogen receptor (ER)-dependent enhancers for BRCA1 [128]. Su et al. found that human Alu elements in the proximal upstream region are more conserved and show many properties of enhancers [129]. Indeed, similar to enhancer RNAs (eRNAs) [130], transcription of Alu transposons may boost the expression of downstream genes through an enhancer mechanism specific to embryonic cells. Further study is needed to investigate these possible mechanisms.

Transposons near genes should be treated more like regulatory elements. In order for a single TF to regulate a large number of genes, it can evolve to take advantage of existing mobile elements. This is more likely than the scenario where hundreds of genomic loci converge to the binding motif of an existing TF, which is especially true for highly specific motifs such as that of CTCF [131].

Conclusions

Guided by biological curiosity, exploratory bioinformatics analysis of the single-cell RNA-seq and related data yields many actionable insights. One of the surprising observations is that genes with similar expression patterns in early embryogenesis share specific transposons in their promoters. During ZGA, while LTRs are linked to transient, forceful and early induction of several hundred genes, SINE elements are associated with the upregulation of thousands of essential genes. The machinery that transcribes retrotransposons may also be used to establish the expression landscape of early embryos. This study also demonstrates the power of single-cell RNA-seq, especially when applied to the study of normal developmental processes.

Methods

Raw data for the single-cell RNA-seq were downloaded from NCBI's Short Read Archive with accession number PRJNA195938 using the fastq-dump program of SRA-tools suite. FastQC was used for the initial quality check [132]. Trimming of sequences was carried out using cutadapt [133]. Mouse genome sequence (GCRm38) and annotation were downloaded from ENSEMBL using the biomaRt [84] package on Bioconductor [134]. We used the Tophat and cufflinks programs [50] to map and quantify gene expression. Translation starting sites (TSSs) of genes were defined as the TSS of the highest expressed transcript isoforms across all the samples in this study.

Read mappings of RNA-seq data were generated by Integrative Genomics Viewer (IGV) [135]. ALEA software [136] was used to map the reads to maternal and paternal alleles based on single nucleotide polymorphisms (SNPs) derived from genome sequences of the two strains [137]. In order to estimate retrotransposon expression, we re-mapped reads using STAR [138] to allow more multiple-mapped reads using the following parameters: STAR –outFilterMultimapNmax 100 –winAnchorMultimapNmax 100 –outSAMmultNmax 100 –outSAMtype BAM SortedByCoordinate –outFilterMismatchNmax 3. The expression levels of retrotransposons were calculated using TEtranscripts [52], which was specially designed to estimate both gene and TE abundances by using an additional index of TEs based on UCSC repeatMasker files. The parameters used are: TEtranscripts –format BAM –mode multi –GTF genes.gtf –TE mm10_rmsk_TE.gtf -i 2 –stranded no. Additional file 3 gives all commands used in sequence analyses.

In the clustering analysis, genes with expression levels less than 5 FPKM across all samples were eliminated from analysis. The remaining genes were sorted by standard deviation and the top 12,000 were selected. Cluster 3.0 [139] was used for hierarchical clustering using Pearson's correlation as distance metrics and average linkage. Java TreeView [140] is used for visualization and interactively explore data.

To detect enriched TEs in promoters, frequencies of TEs in the promoters of genes in 10 clusters are tabulated. The repeatMasker file downloaded from UCSC is used and overlaps with promoters are computed using the genomicRanges [141] package from Bioconductor. *P* values are calculated using Chi-squared tests followed by FDR correction.

The key features in our TF binding analysis are: 1) use of RNA-seq data for TSS location, 2) use position-weight matrices (PWMs) to scan promoter sequences and run ANOVA on the highest scores, which avoid arbitrary cutoff in deciding TF binding, 3) rank genes by fold-change to avoid cutoff in gene clusters, and 4) filter and prioritize using the expression pattern of TF genes.

In multiple linear regression analysis, each gene is characterized by ZGA fold-change (dependent variable) and the number of repeats by repeat type in their promoters (independent variables). TF binding motifs defined by Chip-Seq [70] are also included.

Processed single-cell RNA-seq data for human [20] was downloaded from NCBI using accessing number GSE44183. Accession numbers for bovine expression data are GSE52415 (RNA-seq data [19]), and GSE18290 (DNA microarray data [16]). Zebrafish data of Harvey et al. [79] is downloaded from supplementary data http://www.biologists.com/DEV_Movies/DEV095091/DEV095091TableS2.xls We also used Zebrafish data from Anes et al. [142], with accession number GSE22830.

Additional files

Additional file 1: Table S1. Stage-specific expression level of 36882 genes, calculated using the formula log2 (FPKM + 1). **Table S2.** Gene lists of 10 gene clusters defined in Figure 1B. **Table S3.** Expression levels of transposable elements in mouse embryos at different stages. **Table S4.** List of all repetitive elements (REs) in the mouse genome with repeat name, repeat family, and repeat class. **Table S5.** Frequencies of REs in the promoters of different gene clusters. **Table S6.** List of 117 genes that contain ERVL elements in promoters and show an expression pattern similar to these LTRs. (XLSX 4556 kb)

Additional file 2: Tables S7-S10 and **Figures S1-S41.** It also includes a list of main observations from our open-ended analysis. **Table S7.** Association of repeats and transcription factor binding sites with expression change in 2-cell mouse embryo. **Table S8.** Repeats significantly associated with zygote genome activation in human. **Table S9.** Repeat elements associated with gene expression in zebrafish. **Table S10.** Confirmation of association of repeats with ZGA gene expression in zebrafish. These are followed by **Figures S1-S41**. (PDF 1849 kb)

Additional file 3: Linux commands and shell scripts used to analyze short reads to quantify the expression of regular genes and transposons. (TXT 70 kb)

Abbreviations

16C: 16-Cell; 2C: 2-Cell; 4C: 4-Cell; bp: Base Pair; EBI: Exploratory Bioinformatics Investigation; EDA: Exploratory Data Analysis; ER: Estrogen Receptor; eRNAs: Enhancer RNAs; ERVs: Endogenous Retroviruses; ESCs: Embryonic Stem Cells; ESTs: Expressed Sequence Tags; ETC: Extra-Transcriptional Effects; ETS: E26 Transformation-Specific; FDR: False Discovery Rate; GO: Gene Ontology; GPCR: G-Protein Coupled Receptor; IAP: Intracisternal A Particle; ICM: Inner Cell Mass; IGV: Integrative Genomics Viewer; iPSCs: Induced Pluripotent Cells; LINEs: Long Interspersed Elements; lncRNAs: Long Noncoding Rnas; LTRs: Long Terminal Repeats; MEGs: Maternal Effect Genes; MZT: Maternal-To-Zygotic Transition; Obox: Oocyte Specific Homeobox; PCC: Pearson's Correlation Coefficient; PD: Pre-implantation Development; REs: Repetitive Elements; RNA-seq: RNA-Sequencing; SINEs: Short Interspersed Nuclear Elements; SNPs: Single Nucleotide Polymorphisms; SoxB1: transcription factor SoxB1; TEs: Transposable Elements; TFBS: Transcription Factor Binding Sites; TFs: Transcription Factors; TSSs: Translation Starting Sites; Zfp352: Zinc Finger Protein 352; ZGA: Zygotic Genome Activation

Acknowledgements

This research used computers managed by Administrative and Research Computing at South Dakota State University, and support from Brian Moore and Alan Carter.

Funding

The work benefited indirectly from support by National Institute of Health (GM083226), the National Science Foundation/EPSCoR Grant Number IIA-1355423, and the State of South Dakota.

Availability of data and materials

Not applicable. Some intermediate data from our meta-analysis is available as supplementary files.

Authors' contributions

SXG is the sole author and responsible for all aspects of this study.

Authors' information

As an open-ended bioinformatics investigation, this work is neither focused nor hypothesis-driven, hence unlikely to be fundable. But I wanted to know if we can critically re-analyze existing data to find anything useful to biologists, especially when developing fundable ideas. Benefiting from my training in solid state physics, I tried to obtain a broad understanding that can explain experimental data, and to gain as many actionable insights as possible.

Competing interests
The author declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Received: 13 October 2016 Accepted: 7 February 2017
Published online: 23 February 2017

References

1. Popper KR. Conjectures and refutations; the growth of scientific knowledge. New York: Basic Books; 1962.
2. Biesecker LG. Hypothesis-generating research and predictive medicine. Genome Res. 2013;23(7):1051–3.
3. Kell DB, Oliver SG. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. Bioessays. 2004;26(1):99–105.
4. Tukey JW. Exploratory data analysis, Reading. Massachusetts: Addison-Wesley Pub. Co; 1977.
5. Tufte ER. The visual display of quantitative information. Cheshire, Conn. (Box 430, Cheshire 06410): Graphics Press; 1983.
6. Velleman PF, Hoaglin DC. Applications, basics, and computing of exploratory data analysis. Boston: Duxbury Press; 1981.
7. Roberts RM. Serendipity: accidental discoveries in science. New York: Wiley; 1989.
8. Schier AF. The maternal-zygotic transition: death and birth of RNAs. Science. 2007;316(5823):406–7.
9. Lee MT, Bonneau AR, Giraldez AJ. Zygotic genome activation during the maternal-to-zygotic transition. Annu Rev Cell Dev Biol. 2014;30:581–613.
10. Marra M, Hillier L, Kucaba T, Allen M, Barstead R, Beck C, Blistain A, Bonaldo M, Bowers Y, Bowles L, et al. An encyclopedia of mouse genes. Nat Genet. 1999;21(2):191–4.
11. Rothstein JL, Johnson D, DeLoia JA, Skowronski J, Solter D, Knowles B. Gene expression during preimplantation mouse development. Genes Dev. 1992;6(7):1190–201.
12. Sasaki N, Nagaoka S, Itoh M, Izawa M, Konno H, Carninci P, Yoshiki A, Kusakabe M, Moriuchi T, Muramatsu M, et al. Characterization of gene expression in mouse blastocyst using single-pass sequencing of 3995 clones. Genomics. 1998;49(2):167–79.
13. Ko MS, Kitchen JR, Wang X, Threat TA, Wang X, Hasegawa A, Sun T, Grahovac MJ, Kargul GJ, Lim MK, et al. Large-scale cDNA analysis reveals phased gene expression patterns during preimplantation mouse development. Development. 2000;127(8):1737–49.
14. Hamatani T, Carter MG, Sharov AA, Ko MS. Dynamics of global gene expression changes during mouse preimplantation development. Dev Cell. 2004;6(1):117–31.
15. Sharov AA, Piao Y, Matoba R, Dudekula DB, Qian Y, VanBuren V, Falco G, Martin PR, Stagg CA, Bassey UC, et al. Transcriptome analysis of mouse stem cells and early embryos. PLoS Biol. 2003;1(3):E74.
16. Xie D, Chen CC, Ptaszek LM, Xiao S, Cao X, Fang F, Ng HH, Lewin HA, Cowan C, Zhong S. Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. Genome Res. 2010;20(6):804–15.
17. Vassena R, Boue S, Gonzalez-Roca E, Aran B, Auer H, Veiga A, Izpisua Belmonte JC. Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. Development. 2011;138(17):3699–709.
18. Zhang P, Zucchelli M, Bruce S, Hambiliki F, Stavreus-Evers A, Levkov L, Skottman H, Kerkela E, Kere J, Hovatta O. Transcriptome profiling of human pre-implantation development. PLoS One. 2009;4(11):e7844.
19. Graf A, Krebs S, Zakhartchenko V, Schwalb B, Blum H, Wolf E. Fine mapping of genome activation in bovine embryos by RNA sequencing. Proc Natl Acad Sci U S A. 2014;111(11):4139–44.
20. Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature. 2013;500(7464):593–7.
21. Deng Q, Ramskold D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science. 2014;343(6167):193–6.
22. Tohonen V, Katayama S, Vesterlund L, Jouhilahti EM, Sheikhi M, Madissoon E, Filippini-Cattaneo G, Jaconi M, Johnsson A, Burglin TR, et al. Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. Nat Commun. 2015;6:8207.
23. Xue L, Cai JY, Ma J, Huang Z, Guo MX, Fu LZ, Shi YB, Li WX. Global expression profiling reveals genetic programs underlying the developmental divergence between mouse and human embryogenesis. BMC Genomics. 2013;14:568.
24. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol. 2013;20(9):1131–9.
25. Guo H, Zhu P, Yan L, Li R, Hu B, Lian Y, Yan J, Ren X, Lin S, Li J, et al. The DNA methylation landscape of human early embryos. Nature. 2014;511(7511):606–10.
26. Smith ZD, Chan MM, Mikkelsen TS, Gu H, Gnirke A, Regev A, Meissner A. A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature. 2012;484(7394):339–44.
27. Fadloun A, Le Gras S, Jost B, Ziegler-Birling C, Takahashi H, Gorab E, Carninci P, Torres-Padilla ME. Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. Nat Struct Mol Biol. 2013;20(3):332–8.
28. Leichsenring M, Maes J, Mossner R, Driever W, Onichtchouk D. Pou5f1 transcription factor controls zygotic gene activation in vertebrates. Science. 2013;341(6149):1005–9.
29. Bui LC, Evsikov AV, Khan DR, Archilla C, Peynot N, Henaut A, Le Bourhis D, Vignon X, Renard JP, Duranthon V. Retrotransposon expression as a defining event of genome reprogramming in fertilized and cloned bovine embryos. Reproduction. 2009;138(2):289–99.
30. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011;7(12):e1002384.
31. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009;41(5):563–71.
32. Evsikov AV, de Vries WN, Peaston AE, Radford EE, Fancher KS, Chen FH, Blake JA, Bult CJ, Latham KE, Solter D, et al. Systems biology of the 2-cell mouse embryo. Cytogenet Genome Res. 2004;105(2-4):240–50.
33. Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, et al. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature. 2015;522(7555):221–5.
34. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 2012;487(7405):57–63.
35. Ishiuchi T, Enriquez-Gasca R, Mizutani E, Boskovic A, Ziegler-Birling C, Rodriguez-Terrones D, Wakayama T, Vaquerizas JM, Torres-Padilla ME. Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. Nat Struct Mol Biol. 2015;22(9):662–71.
36. Kigami D, Minami N, Takayama H, Imai H. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. Biol Reprod. 2003;68(2):651–4.
37. Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian Jr HH. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev. 2009;23(11):1303–12.
38. Wang J, Xie G, Singh M, Ghanbarian AT, Rasko T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014;516(7531):405–9.
39. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell. 2004;7(4):597–606.
40. Li J, Kannan M, Trivett AL, Liao H, Wu X, Akagi K, Symer DE. An antisense promoter in mouse L1 retrotransposon open reading frame-1 initiates expression of diverse fusion transcripts and limits retrotransposition. Nucleic Acids Res. 2014;42(7):4546–62.
41. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42(7):631–4.
42. Ichiyanagi K. Epigenetic regulation of transcription and possible functions of mammalian short interspersed elements, SINEs. Genes Genet Syst. 2013;88(1):19–29.
43. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8(4):272–85.

44. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell. 2012;148(1-2):335–48.
45. Bianchi E, Sette C. Post-transcriptional control of gene expression in mouse early embryo development: a view from the tip of the iceberg. Genes (Basel). 2011;2(2):345–59.
46. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature. 2008;453(7194):534–8.
47. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature. 2008;453(7194):539–43.
48. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011;43(11):1154–9.
49. Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell. 2010;6(5):468–78.
50. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78.
51. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):D662–669.
52. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. 2015;31(22):3593–9.
53. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.
54. Dupressoir A, Heidmann T. Germ line-specific expression of intracisternal A-particle retrotransposons in transgenic mice. Mol Cell Biol. 1996;16(8):4495–503.
55. Calabrese JM, Seila AC, Yeo GW, Sharp PA. RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. Proc Natl Acad Sci U S A. 2007; 104(46):18097–102.
56. Biase FH, Cao X, Zhong S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. Genome Res. 2014;24(11): 1787–96.
57. Shi J, Chen Q, Li X, Zheng X, Zhang Y, Qiao J, Tang F, Tao Y, Zhou Q, Duan E. Dynamic transcriptional symmetry-breaking in pre-implantation mammalian embryo development revealed by single-cell RNA-seq. Development. 2015; 142(20):3468–77.
58. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110(1-4):462–7.
59. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res. 2001;125(1-2):279–84.
60. Liu TY, Chen HH, Lee KH, Choo KB. Display of different modes of transcription by the promoters of an early embryonic gene, Zfp352, in preimplantation embryos and in somatic cells. Mol Reprod Dev. 2003;64(1):52–60.
61. Chen HH, Liu TY, Huang CJ, Choo KB. Generation of two homologous and intronless zinc-finger protein genes, zfp352 and zfp353, with different expression patterns by retrotransposition. Genomics. 2002;79(1):18–23.
62. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158(6):1431–43.
63. Rajkovic A, Yan C, Yan W, Klysik M, Matzuk MM. Obox, a family of homeobox genes preferentially expressed in germ cells. Genomics. 2002;79(5):711–7.
64. Cheng WC, Hsieh-Li HM, Yeh YJ, Li H. Mice lacking the Obox6 homeobox gene undergo normal early embryonic development and are fertile. Dev Dyn. 2007;236(9):2636–42.
65. Zalzman M, Falco G, Sharova LV, Nishiyama A, Thomas M, Lee SL, Stagg CA, Hoang HG, Yang HT, Indig FE, et al. Zscan4 regulates telomere elongation and genomic stability in ES cells. Nature. 2010;464(7290):858–63.
66. Amano T, Hirata T, Falco G, Monti M, Sharova LV, Amano M, Sheer S, Hoang HG, Piao Y, Stagg CA, et al. Zscan4 restores the developmental potency of embryonic stem cells. Nat Commun. 2013;4:1966.
67. Hirata T, Amano T, Nakatake Y, Amano M, Piao Y, Hoang HG, Ko MS. Zscan4 transiently reactivates early embryonic genes during the generation of induced pluripotent stem cells. Sci Rep. 2012;2:208.
68. Deininger P. Alu elements: know the SINEs. Genome Biol. 2011;12(12):236.
69. Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520–62.
70. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell. 2008;133(6):1106–17.
71. Eller CD, Regelson M, Merriman B, Nelson S, Horvath S, Marahrens Y. Repetitive sequence environment distinguishes housekeeping genes. Gene. 2007;390(1-2):153–65.
72. Kim TM, Jung YC, Rhyu MG. Alu and L1 retroelements are correlated with the tissue extent and peak rate of gene expression, respectively. J Korean Med Sci. 2004;19(6):783–92.
73. Christodoulou C, Longmire TA, Shen SS, Bourdon A, Sommer CA, Gadue P, Spira A, Gouon-Evans V, Murphy GJ, Mostoslavsky G, et al. Mouse ES and iPS cells can form similar definitive endoderm despite differences in imprinted genes. J Clin Invest. 2011;121(6):2313–25.
74. Urrutia AO, Ocana LB, Hurst LD. Do Alu repeats drive the evolution of the primate transcriptome? Genome Biol. 2008;9(2):R25.
75. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012;488(7409):116–20.
76. Braude P, Bolton V, Moore S. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. Nature. 1988;332(6163):459–61.
77. Tsirigos A, Rigoutsos I. Alu and b1 repeats have been selectively retained in the upstream and intronic regions of genes of specific functional classes. PLoS Comput Biol. 2009;5(12):e1000610.
78. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013; 496(7446):498–503.
79. Harvey SA, Sealy I, Kettleborough R, Fenyes F, White R, Stemple D, Smith JC. Identification of the zebrafish maternal and paternal transcriptomes. Development. 2013;140(13):2703–10.
80. Sutton J. Gibrat's legacy. J Econ Lit. 1997;35(1):40–59.
81. Klimopoulos A, Sellis D, Almirantis Y. Widespread occurrence of power-law distributions in inter-repeat distances shaped by genome dynamics. Gene. 2012;499(1):88–98.
82. Britten RJ. DNA sequence insertion and evolutionary variation in gene regulation. Proc Natl Acad Sci U S A. 1996;93(18):9374–7.
83. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
84. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–40.
85. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005;122(6):947–56.
86. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006;126(4):663–76.
87. Yoshikawa T, Piao Y, Zhong J, Matoba R, Carter MG, Wang Y, Goldberg I, Ko MS. High-throughput screen for genes predominantly expressed in the ICM of mouse blastocysts by whole mount in situ hybridization. Gene Expr Patterns. 2006;6(2):213–24.
88. Wang ZX, Teh CH, Kueh JL, Lufkin T, Robson P, Stanton LW. Oct4 and Sox2 directly regulate expression of another pluripotency transcription factor, Zfp206, in embryonic stem cells. J Biol Chem. 2007;282(17):12822–30.
89. Kim KH, Kim EY, Lee KA. SEBOX is essential for early embryogenesis at the two-cell stage in the mouse. Biol Reprod. 2008;79(6):1192–201.
90. Ristevski S, O'Leary DA, Thornell AP, Owen MJ, Kola I, Hertzog PJ. The ETS transcription factor GABPalpha is essential for early embryogenesis. Mol Cell Biol. 2004;24(13):5844–9.
91. Maroulakou IG, Bowe DB. Expression and function of Ets transcription factors in mammalian development: a regulatory network. Oncogene. 2000; 19(55):6432–42.
92. Remy P, Baltzinger M. The Ets-transcription factor family in embryonic development: lessons from the amphibian and bird. Oncogene. 2000; 19(55):6417–31.

93. Lin SC, Wani MA, Whitsett JA, Wells JM. Klf5 regulates lineage formation in the pre-implantation mouse embryo. Development. 2010;137(23):3953–63.
94. Aksoy I, Giudice V, Delahaye E, Wianny F, Aubry M, Mure M, Chen J, Jauch R, Bogu GK, Nolden T, et al. Klf4 and Klf5 differentially inhibit mesoderm and endoderm differentiation in embryonic stem cells. Nat Commun. 2014;5:3719.
95. Marikawa Y, Alarcon VB. Establishment of trophectoderm and inner cell mass lineages in the mouse embryo. Mol Reprod Dev. 2009;76(11):1019–32.
96. Li WZ, Wang ZW, Chen LL, Xue HN, Chen X, Guo ZK, Zhang Y. Hesx1 enhances pluripotency by working downstream of multiple pluripotency-associated signaling pathways. Biochem Biophys Res Commun. 2015;464(3):936–42.
97. Liu Y, Chakroun I, Yang D, Horner E, Liang J, Aziz A, Chu A, De Repentigny Y, Dilworth FJ, Kothary R, et al. Six1 regulates MyoD expression in adult muscle progenitor cells. PLoS One. 2013;8(6):e67762.
98. Wu W, Ren Z, Li P, Yu D, Chen J, Huang R, Liu H. Six1: a critical transcription factor in tumorigenesis. Int J Cancer. 2015;136(6):1245–53.
99. Niakan KK, Ji H, Maehr R, Vokes SA, Rodolfa KT, Sherwood RI, Yamaki M, Dimos JT, Chen AE, Melton DA, et al. Sox17 promotes differentiation in mouse embryonic stem cells by directly regulating extraembryonic gene expression and indirectly antagonizing self-renewal. Genes Dev. 2010;24(3):312–26.
100. Beck F, Erler T, Russell A, James R. Expression of Cdx-2 in the mouse embryo and placenta: possible role in patterning of the extra-embryonic membranes. Dev Dyn. 1995;204(3):219–27.
101. Daikoku T, Cha J, Sun X, Tranguch S, Xie H, Fujita T, Hirota Y, Lydon J, DeMayo F, Maxson R, et al. Conditional deletion of Msx homeobox genes in the uterus inhibits blastocyst implantation by altering uterine receptivity. Dev Cell. 2011;21(6):1014–25.
102. Nallasamy S, Li Q, Bagchi MK, Bagchi IC. Msx homeobox genes critically regulate embryo implantation by controlling paracrine signaling between uterine stroma and epithelium. PLoS Genet. 2012;8(2):e1002500.
103. Cha J, Sun X, Bartos A, Fenelon J, Lefevre P, Daikoku T, Shaw G, Maxson R, Murphy BD, Renfree MB, et al. A new role for muscle segment homeobox genes in mammalian embryonic diapause. Open Biol. 2013;3(4):130035.
104. Buonamici S, Chakraborty S, Senyuk V, Nucifora G. The role of EVI1 in normal and leukemic cells. Blood Cells Mol Dis. 2003;31(2):206–12.
105. Wieser R. The oncogene and developmental regulator EVI1: expression, biochemical properties, and biological functions. Gene. 2007;396(2):346–57.
106. Burglin TR, Affolter M. Homeodomain proteins: an update. Chromosoma. 2016; 125:497–521.
107. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011;479(7374):534–7.
108. Kazazian Jr HH. Mobile elements: drivers of genome evolution. Science. 2004;303(5664):1626–32.
109. de Magalhaes JP, Budovsky A, Lehmann G, Costa J, Li Y, Fraifeld V, Church GM. The Human Ageing Genomic Resources: online databases and tools for biogerontologists. Aging Cell. 2009;8(1):65–72.
110. Gregory TR: Animal Genome Size Database. 2015. http://www.genomesize.com. Accessed 20 Jan 2016.
111. Cavalier-Smith T. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J Cell Sci. 1978;34:247–78.
112. Fedoroff NV. Presidential address. Transposable elements, epigenetics, and genome evolution. Science. 2012;338(6108):758–67.
113. Matlik K, Redik K, Speek M. L1 antisense promoter drives tissue-specific transcription of human genes. J Biomed Biotechnol. 2006;2006(1):71753.
114. Jachowicz JW, Torres-Padilla ME. LINEs in mice: features, families, and potential roles in early development. Chromosoma. 2016;125(1):29–39.
115. Ohnuki M, Tanabe K, Sutou K, Teramoto I, Sawamura Y, Narita M, Nakamura M, Tokunaga Y, Nakamura M, Watanabe A, et al. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. Proc Natl Acad Sci U S A. 2014;111(34):12426–31.
116. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351(6277):1083–7.
117. Gu TJ, Yi X, Zhao XW, Zhao Y, Yin JQ. Alu-directed transcriptional regulation of some novel miRNAs. BMC Genomics. 2009;10:563.
118. Polak P, Domany E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. BMC Genomics. 2006;7:133.
119. Roman AC, Benitez DA, Carvajal-Gonzalez JM, Fernandez-Salguero PM. Genome-wide B1 retrotransposon binds the transcription factors dioxin receptor and Slug and regulates gene expression in vivo. Proc Natl Acad Sci U S A. 2008;105(5):1632–7.
120. Cui F, Sirotin MV, Zhurkin VB. Impact of Alu repeats on the evolution of human p53 binding sites. Biol Direct. 2011;6:2.
121. Laperriere D, Wang TT, White JH, Mader S. Widespread Alu repeat-driven expansion of consensus DR2 retinoic acid response elements during primate evolution. BMC Genomics. 2007;8:23.
122. Vansant G, Reynolds WF. The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. Proc Natl Acad Sci U S A. 1995;92(18):8229–33.
123. Humphrey GW, Englander EW, Howard BH. Specific binding sites for a pol III transcriptional repressor and pol II transcription factor YY1 within the internucleosomal spacer region in primate Alu repetitive elements. Gene Expr. 1996;6(3):151–68.
124. Hjalt TA, Amendt BA, Murray JC. PITX2 regulates procollagen lysyl hydroxylase (PLOD) gene expression: implications for the pathology of Rieger syndrome. J Cell Biol. 2001;152(3):545–52.
125. Donze D. Extra-transcriptional functions of RNA Polymerase III complexes: TFIIIC as a potential global chromatin bookmark. Gene. 2012;493(2):169–75.
126. Noma K, Cam HP, Maraia RJ, Grewal SI. A role for TFIIIC transcription factor complex in genome organization. Cell. 2006;125(5):859–72.
127. Moqtaderi Z, Wang J, Raha D, White RJ, Snyder M, Weng Z, Struhl K. Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. Nat Struct Mol Biol. 2010;17(5):635–40.
128. Norris J, Fan D, Aleman C, Marks JR, Futreal PA, Wiseman RW, Iglehart JD, Deininger PL, McDonnell DP. Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. J Biol Chem. 1995;270(39):22777–82.
129. Su M, Han D, Boyd-Kirkup J, Yu X, Han JD. Evolution of Alu elements toward enhancers. Cell Rep. 2014;7(2):376–85.
130. Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature. 2011;474(7351): 390–4.
131. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell. 2007;128(6):1231–45.
132. Simon A. FastQC: a quality control tool for high throughput sequence data. 2010.
133. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17(1):3.
134. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.
135. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
136. Younesy H, Moller T, Heravi-Moussavi A, Cheng JB, Costello JF, Lorincz MC, Karimi MM, Jones SJ. ALEA: a toolbox for allele-specific epigenomics analysis. Bioinformatics. 2014;30:1172–4.
137. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011;477(7364):289–94.
138. Dobin A, Gingeras TR. Optimizing RNA-Seq Mapping with STAR. Methods Mol Biol. 2016;1415:245–62.
139. de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics. 2004;20(9):1453–4.
140. Saldanha AJ. Java Treeview–extensible visualization of microarray data. Bioinformatics. 2004;20(17):3246–8.
141. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118.
142. Aanes H, Winata CL, Lin CH, Chen JP, Srinivasan KG, Lee SG, Lim AY, Hajan HS, Collas P, Bourque G, et al. Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res. 2011;21(8):1328–38.