

Supplementary Figure 1. **The definition of pan-ZGA genes, intersection of pan-ZGA genes with orphan nuclear receptors-regulated, OBOX-regulated and DUX-regulated genes, and association of repetitive elements upstream of genes with chromatin accessibility and transcription changes.**

**(A)** Venn diagram of ZGA list intersections grouped by mouse strain background. The core intersection consists of 200 ZGA genes. Overlaps comprising the pan-ZGA genes list are indicated by white circles. The pan-ZGA genes list consists of 542 genes (see methods below). **(B)** Pan-ZGA genes are partially controlled by orphan nuclear receptors (OrphNRs), OBOX and DUX. Pan-ZGA genes intersect with 85% of overlapped downregulated genes upon OrphNRs, OBOX or DUX perturbation, leaving 15% of genes regulated by currently unknown factors. The former category is comprised of 78% regulated by OrphNRs, 40% regulated by OBOX, and 4.6% regulated by DUX. OrphNRs and OBOX together regulate 34% of pan-ZGA genes. **(C)** The intersection between pan-ZGA genes and OrphNRs-regulated, OBOX-regulated and DUX-regulated genes. Notably, NR5A2, FOXR1, and SPIC TFs are present in the complete intersection between all four categories (N=16). The split bar is indicated by the tilted double line. To determine the overlap between pan-ZGA genes and OrphNRs-controlled, OBOX-controlled and DUX-controlled genes, the pan-ZGA list was intersected with publicly available perturbation records using published thresholds for downregulated genes. Downregulated genes in NR5A2 knockdown and OrphNRs chemically inhibited with SR1848 2-cell embryos[1](https://paperpile.com/c/rzyHYs/AfGf5) were defined as OrphNRs-controlled genes. *OBOX* knockout[2](https://paperpile.com/c/rzyHYs/8OT7H) and *DUX* knockout[3](https://paperpile.com/c/rzyHYs/dxoHw) downregulated genes were used without modifications. **(D)** **1.** The accumulation of *SINE B1* upstream to pan-ZGA genes. *SINE B1* abundance 8000 (8k) bp upstream to ZGA genes TSSs. To allow unbiased comparisons between non-ZGA and pa6n-ZGA gene groups, a 10-fold split with randomly sampled 52 genes in each batch was performed. The folds were centered by a median. The diamond represents a mean. *P* values are reported for two-sided Welch's t-test for unequal variances (N=52). A comparably lower significant *ERVL* accumulation within 8k bp upstream to pan-ZGA genes TSSs was observed (on average 1.4 retrotransposable element (RE) upstream of pan-ZGA compared to 0.3 RE upstream of non-ZGA, *P* value = 6.49e-08). According to this, the fraction RE upstream of pan-ZGA genes to non-ZGA genes is similar for *SINE B1* and *ERVL* with the latter being slightly more abundantly located: 3.3 and 4.7, respectively. No enrichment was observed for *ERVL-MaLR.***2.** The distribution of distances between centers of closest NR5A2 CUT&Tag and OBOX3 Stacc-seq peaks. Please note that OBOX3 Stacc-seq was performed on OBOX3-FLAG overexpressing 2-cell embryos[2](https://paperpile.com/c/rzyHYs/8OT7H) and NR5A2 CUT&Tag was performed to detect endogenous NR5A2 in wild-type 2-cell embryos[1](https://paperpile.com/c/rzyHYs/AfGf5). No distant associations are observed beyond 1000 bp cut-off. **3.** The overlap of the 2299 peak pairs identified in (2) with repetitive elements. **4.** TFIIIC motifs (boxes A and B) overlap with OBOX (motif #5) and NR5A2 (motif #1) transcription factor binding sites[4](https://paperpile.com/c/rzyHYs/32hV3). **(E)** Correlation between normalized number of repetitive elements (REs) (*SINE B1*, *SINE B2*, *LINE-1 (L1)*, *ERVL*, *ERVK*, *ERV1*) and chromatin accessibility identified by Omni-ATAC-seq[1](https://paperpile.com/c/rzyHYs/AfGf5) 8000 bp upstream of pan-ZGA (orange) and non-ZGA (black) genes. First, the number of REs was normalized to the total number of REs of the type. Next, both scales were min-max normalized. RepeatMasker[5](https://paperpile.com/c/rzyHYs/nl0tP) annotation for mm10, *Mus musculus* GRCm38 annotation, and pybedtools 0.9.0 (Ref.[6](https://paperpile.com/c/rzyHYs/QCVQ)) were used to estimate the accumulation of REs. Omni-ATAC-seq data was analyzed as previously described[1](https://paperpile.com/c/rzyHYs/AfGf5). **(F)** Correlation between the number of REs (*SINE B1*, *SINE B2*, *LINE-1 (L1)*, *ERVL*, *ERVK*, *ERV1*) with 2-cell to zygote log2 fold change of all transcripts identified by developmental RNA-seq[1](https://paperpile.com/c/rzyHYs/AfGf5) (data is binned). The number of REs was normalized to the total number of REs of the type. Spearman correlation coefficient ρ with associated *P* value is reported for each RE in the legend. The coefficient was calculated using Spearmanr function from the Stats module of SciPy 1.9.3 (Ref.[7](https://paperpile.com/c/rzyHYs/RhY2r)). Analogous counting of repetitive elements with pybedtools as in (D) was applied to examine the correlation of the number of REs with RNA-seq transcription data**.**

**Code availability**

All the code and secondary data generated in this study are available on GitHub: <https://github.com/Pavel-Kravchenko/Rise_and_SINE>.

**Data availability statement**

The RNA-seq datasets used in this study were obtained using the following GSE accession numbers: GSE178638 (Ref.[1](https://paperpile.com/c/rzyHYs/AfGf5)), GSE215813 (Ref.[2](https://paperpile.com/c/rzyHYs/8OT7H)), GSE121746 (Ref.[3](https://paperpile.com/c/rzyHYs/dxoHw)), GSE45719 (Ref.[8](https://paperpile.com/c/rzyHYs/UiyCm)), GSE71257 (Ref.[9](https://paperpile.com/c/rzyHYs/dQ9uw)), GSE66582 (Ref.[10](https://paperpile.com/c/rzyHYs/gldOK)), GSE71434 (Ref.[11](https://paperpile.com/c/rzyHYs/jUtQr)), GSE134832 (Ref.[12](https://paperpile.com/c/rzyHYs/4PMVo)). The ATAC-seq dataset was obtained using GSE178234 (Ref.[1](https://paperpile.com/c/rzyHYs/AfGf5)) accession number. The secondary data generated in this study, supporting the findings, are available on GitHub.

**Data processing and visualization**

Publicly available RNA-seq datasets were downloaded from GEO database with NCBI sratoolkit 3.0.7 (Ref.[13](https://paperpile.com/c/rzyHYs/5pdn6)). RNA-seq data analysis was performed as previously described[1](https://paperpile.com/c/rzyHYs/AfGf5). The data were trimmed by TrimGalore 0.6.10 (Ref.[14](https://paperpile.com/c/rzyHYs/ksxGj)) with default settings and --quality 20, --trim-n flags. Trimmed paired-end and single reads were pseudo-aligned by Kallisto 0.46.2 (Ref.[15](https://paperpile.com/c/rzyHYs/vstsb)) with -b 100 flag to *Mus musculus* (mm10) Ensembl v96 transcriptome. Reads and abundances were imported to R 4.2.1 by tximport 1.24.0 (Ref.[16](https://paperpile.com/c/rzyHYs/FLSna)) and compared between developmental stages with DESeq2 1.36.0 (Ref.[17](https://paperpile.com/c/rzyHYs/rLdII)).

**Selection of ZGA genes**

A gene list for every dataset was obtained from gene expression comparison between zygotes (20-28 hours post hCG injection) and to 2-cell stage (39-48 hours post hCG injection) embryos. Genes were initially filtered by FPKM to assure genuine robust expression comparison with FPKM>2 and FPKM>5 for the zygote and the 2-cell stage, respectively. ZGA genes were selected as genes with a minimum 4-fold increase (FDR=0.05) in expression from the zygote to the 2-cell embryonic stage. Mouse strain-specific datasets were pre-united in the case of pure background and pre-intersected in the case of mixed background. The core ZGA list was obtained by an intersection of upregulated genes in all strains. The pan-ZGA list was obtained by uniting the core list with a union of all intersections of ZGA genes in all datasets except one, analogous to leave-one-out strategy[18](https://paperpile.com/c/rzyHYs/7vnox) (see below, ‘Formula’). The described approach allowed to account for strain or experiment variation and to integrate ZGA genes that would have been excluded with a direct intersection. Supporting tables with accession numbers and gene lists are provided as Supplementary table 1 and on GitHub.

**Formula. Selection of pan-ZGA genes**

The union of the intersection of all individual ZGA gene lists in 𝔸={Z1, Z2, …, Zn} with the union of intersections of all individual ZGA gene lists but one indexed as *k*, where *k* is every ZGA list from 𝔸. Pan-ZGA genes of 2, 3, …, n-1 degree could be produced similarly, excluding 2, 3, …, n-1 *k* datasets, respectively. The first argument of the intersection of all individual ZGA gene lists was isolated to introduce the core ZGA list explicitly.

**TF binding colocalization analysis**

NR5A2 2-cell CUT&Tag[1](https://paperpile.com/c/rzyHYs/AfGf5), OBOX3-FLAG 2-cell Stacc-seq[2](https://paperpile.com/c/rzyHYs/8OT7H) and DUX in mESC ChIP-seq[19](https://paperpile.com/c/rzyHYs/JZQOr) peaks with published thresholds were used to search for colocalized binding of TFs as it was described in[20](https://paperpile.com/c/rzyHYs/GEtBj). OBOX data was converted from mm9 to mm10 using UCSC LiftOver[21](https://paperpile.com/c/rzyHYs/EmXzT). Pybedtools 0.9.0 (Ref.[6](https://paperpile.com/c/rzyHYs/QCVQ)) was used to calculate absolute distances between closest pairs over all peak midpoints. 1000 coordinate permutations were performed to construct a background distribution of absolute inter midpoints distance. The background distribution was used to derive a *P* value based on the Poisson distribution for the observed number of pairs of peaks at each distance.

**Intersection with repetitive elements**

RepeatMasker[5](https://paperpile.com/c/rzyHYs/nl0tP) table for mm10, *Mus musculus* GRCm38 annotation and pybedtools 0.9.0 (Ref.[6](https://paperpile.com/c/rzyHYs/QCVQ)) were used to estimate the accumulation of repetitive elements. To count *SINE B1* retrotransposons before genes, 8000 bp regions before pan-ZGA and non-ZGA genes TSSs were selected. The regions were split into 10 cross-validation groups to account for variation. Non-ZGA genes were size-equally randomly sampled 100 times and averaged. To annotate co-bound by TFs genomic regions, CUT&Tag, Stacc-seq, and ChIP-seq peaks with midpoints closer than 100 bp were intersected with coordinates of all repetitive elements. Analogous counting of repetitive elements with pybedtools was applied to examine the correlation between the number of RE with RNA-seq or Omni-ATAC-seq data.

**Statistical tests**

The hypotheses of absence of difference between means of ZGA and non-ZGA groups, and ZGA and shuffled control groups (Fig. S1D1) were tested using two-sided Welch's t-test for unequal variances implemented in the Stats module of SciPy 1.9.3 (Ref.[7](https://paperpile.com/c/rzyHYs/RhY2r)). Spearman correlation coefficient with associated *P* value (Fig. S1F) was calculated using Spearmanr function from the same module.

**References**

1. [Gassler, J. *et al.* Zygotic genome activation by the totipotency pioneer factor Nr5a2. *Science* **378**, 1305–1315 (2022).](http://paperpile.com/b/9rF7w0/4XL3O)

2. [Ji, S. *et al.* OBOX regulates mouse zygotic genome activation and early development. *Nature* **620**, 1047–1053 (2023).](http://paperpile.com/b/9rF7w0/7r4Nx)

3. [Chen, Z. & Zhang, Y. Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development. *Nat. Genet.* **51**, 947–951 (2019).](http://paperpile.com/b/rzyHYs/dxoHw)

4. [Ciesla, M., Skowronek, E. & Boguta, M. Function of TFIIIC, RNA polymerase III initiation factor, in activation and repression of tRNA gene transcription. *Nucleic Acids Res.* **46**, 9444–9455 (2018).](http://paperpile.com/b/rzyHYs/32hV3)

5. Smit, AFA[. *et al.*](http://paperpile.com/b/9rF7w0/7r4Nx) [RepeatMasker. *RepeatMasker Open-3.0*](http://paperpile.com/b/OdBApH/pnInL) <http://www.repeatmasker.org> [(1996–2010).](http://paperpile.com/b/OdBApH/pnInL)

6. [Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).](http://paperpile.com/b/rzyHYs/QCVQ)

7. [Virtanen, P., Gommers, R., Oliphant, T. E. & Haberland, M. Fundamental algorithms for scientific computing in python and SciPy 1.0 contributors. SciPy 1.0. *Nat. Methods* (2020).](http://paperpile.com/b/rzyHYs/RhY2r)

8. [Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).](http://paperpile.com/b/rzyHYs/UiyCm)

9. [Yu, C. *et al.* BTG4 is a meiotic cell cycle-coupled maternal-zygotic-transition licensing factor in oocytes. *Nat. Struct. Mol. Biol.* **23**, 387–394 (2016).](http://paperpile.com/b/rzyHYs/dQ9uw)

10. [Wu, J. *et al.* The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652–657 (2016).](http://paperpile.com/b/rzyHYs/gldOK)

11. [Zhang, B. *et al.* Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature* **537**, 553–557 (2016).](http://paperpile.com/b/rzyHYs/jUtQr)

12. [Guo, M. *et al.* Precise temporal regulation of Dux is important for embryo development. *Cell Res.* **29**, 956–959 (2019).](http://paperpile.com/b/rzyHYs/4PMVo)

13. [Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19–21 (2011).](http://paperpile.com/b/rzyHYs/5pdn6)

14. [Krueger, F. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.](http://paperpile.com/b/9rF7w0/QFTCB) [www.bioinformatics.babraham.ac.uk/projects/trim\_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) [(2015).](http://paperpile.com/b/9rF7w0/QFTCB)

15. [Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).](http://paperpile.com/b/rzyHYs/vstsb)

16. [Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2015).](http://paperpile.com/b/rzyHYs/FLSna)

17. [Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).](http://paperpile.com/b/rzyHYs/rLdII)

18. [Lachenbruch, P. A. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics* **23**, 639–645 (1967).](http://paperpile.com/b/rzyHYs/7vnox)

19. [Hendrickson, P. G. *et al.* Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017).](http://paperpile.com/b/rzyHYs/JZQOr)

20. [Festuccia, N., Owens, N., Chervova, A., Dubois, A. & Navarro, P. The combined action of Esrrb and Nr5a2 is essential for murine naïve pluripotency. *Development* **148**, (2021).](http://paperpile.com/b/rzyHYs/GEtBj)

21. [Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–8 (2006).](http://paperpile.com/b/rzyHYs/EmXzT)