Two independent transcription initiation codes overlap on vertebrate core promoters

Vanja Haberle1,2*, Nan Li3*, Yavor Hadzhiev1, Christopher Previti4†, Chirag Nepal4†, Jochen Gehrig3†, Xianjun Dong6†, Altuna Akalin7†, Ana Maria Suzuki4,5, Wilfred F. J. van IJcken2, Olivier Armand8, Marco Ferg8, Uwe Strähle8, Piero Carninci9,9, Ferenc Müller3 & Boris Lenhard2,9

A core promoter is a stretch of DNA surrounding the transcription start site (TSS) that integrates regulatory inputs and recruits general transcription factors to initiate transcription. The nature and causative relationship of the DNA sequence and chromatin signals that govern the selection of most TSSs by RNA polymerase II remain unresolved. Maternal to zygotic transition represents the most marked change of the transcription repertoire in the vertebrate life cycle. Early embryonic development in zebrafish is characterized by a series of transcriptionally silent cell cycles regulated by inherited maternal gene products: zygotic genome activation commences at the tenth cell cycle, marking the mid-blastula transition. This transition provides a unique opportunity to study the rules of TSS selection and the hierarchy of events linking transcription initiation with key chromatin modifications. We analysed TSS usage during zebrafish early embryonic development at high resolution using cap analysis of gene expression, and determined the positions of H3K4me3-marked promoter-associated nucleosomes. Here we show that the transition from the maternal to zygotic transcription is characterized by a switch between two fundamentally different modes of defining transcription initiation, which drive the dynamic change of TSS usage and promoter shape. A maternal-specific TSS selection, which requires an A/T-rich (W-box) motif, is replaced with a zygotic TSS which maternal messenger RNAs were initiated from different positions than zygotic transcripts, often with shifting of TSS positions within a single promoter.

To address developmental-stage-specific promoter usage throughout early embryonic development, we analysed a nucleotide-resolution map of transcription initiation events in the zebrafish genome, generated by CAGE across 12 stages from the unfertilized egg to organogenesis (Fig. 1a). The data revealed numerous cases of promoter dynamics in which maternal messenger RNAs were initiated from different positions than zygotic transcripts, often with shifting of TSS positions within a single promoter (Fig. 1a).

Clustering of individual TSSs by expression profile revealed several major classes of TSS dynamics (Fig. 1b): TSSs present preferentially in maternal (before the mid-blastula transition (MBT)) stages, reflecting maternally inherited transcripts, as opposed to those activated in early or later zygotic stages (post-MBT). Additional clusters included constitutively present TSSs and TSSs with peak activity in the transitional stages, confirming major changes in the zebrafish transcriptome initiated at the MBT. An equivalent clustering of entire promoters revealed a similar pattern (Extended Data Fig. 1a). However, promoters with no change in the overall expression level often contained a population of TSSs with very heterogeneous relative usage during development (Fig. 1a and Extended Data Fig. 1b–d).

The observed differential TSS utilization between inherited (maternal) and de novo transcribed (zygotic) mRNAs suggested distinct rules for TSS selection acting within the same promoter in the oocyte and the embryo. To reveal the underlying signatures guiding differential promoter interpretation by the maternal and zygotic transcription machinery, we further dissected maternal- and zygotic-specific promoter usage. We first identified a subset of promoters similar to the example in Fig. 1a, showing a significant degree of shifting between maternally and zygotically used TSSs (Extended Data Fig. 2a,b). Our set contained 911 ‘shifting’ promoters whose CAGE signal in pre- and post-MBT stages overlapped by less than 40% (Supplementary Table 1). The TSS shift happened in either direction, but mainly within a narrow window of up to 100 base pairs (bp) (Extended Data Fig. 2c). The preferred maternal and zygotic TSS of those promoters showed antagonistic developmental dynamics, with degradation of inherited maternal transcripts and gradual activation of zygotic ones (Extended Data Fig. 2d).

Aligning sequences of shifting promoters by their maternal dominant TSS revealed a clear enrichment of T- and/or A-containing (WW) dinucleotides ~30 bp upstream of the maternal TSS, hinting at the presence of a functional TATA-like element (Fig. 2a and Extended Data Fig. 2e). By contrast, the zygotic TSS did not show a TATA-like signal in the expected position, but had a sharp SS–WW boundary in local

---

1Department of Biology, University of Bergen, Thormehlensgate 53A, N-5008 Bergen, Norway. 2Institute of Clinical Sciences and MRC Clinical Sciences Center, Faculty of Medicine, Imperial College London, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK. 3School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. 4RIKEN Omics Science Center, Yokohama, Kanagawa 230-0045, Japan. 5RIKEN Center for Life Science Technologies, Division of Genomic Technologies, RIKEN Yokohama Campus, 1-7-2 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. 6Computational Biology Unit, Uni Computing, Uni Research AS, University of Bergen, Thormehlensgate 55, N-5008 Bergen, Norway. 7Erasmus Medical Center, Department of Biomedics, Room Ee679b, Dr Molewaterplein 50, 3015 GE Rotterdam, The Netherlands. 8Institute of Toxicology and Genetics, Karlsruhe Institute of Technology, Postfach 3640, 76021 Karlsruhe, Germany. 9Department of Informatics, University of Bergen, Thormehlensgate 55, N-5008 Bergen, Norway. *Present addresses: German Cancer Research Center (DKFZ), Genomics & Proteomics Core Facility (GPCF), Im Neuenheimer Feld 580/TP3, Heidelberg 69120, Germany (C.P.); Bragelmann Research Laboratory, The Gade Institute, University of Bergen, The Laboratory Building, Haukeland University Hospital, N-5021 Bergen, Norway (C.N.); Acquifer AG, Sophienstraße 136, 76135 Karlsruhe, Germany (J.G.); Department of Neurology, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA (X.D.); Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, 4058 Basel, Switzerland (A.A.).

*These authors contributed equally to this work.

©2014 Macmillan Publishers Limited. All rights reserved
C/G and A/T dinucleotide enrichment precisely aligned ~50 bp downstream of the zygotic TSS (Fig. 2a and Extended Data Fig. 2e). This suggests that there are two fundamentally different sequence signals guiding transcription initiation in the oocyte and the embryo.

Only a small fraction of maternal TSSs (<10%) had a canonical TATA-box motif (Fig. 2b), whereas the majority contained other A/T-rich pentamers (Extended Data Fig. 2f). Motif discovery revealed the presence of an A/T-rich motif (W-box) with lower information content than the canonical TATA-box, but equally positioned 30 bp upstream of the maternal TSS (Fig. 2c). By contrast, the zygotic TSS did not show the presence of a TATA-box or W-box in the expected upstream region (Fig. 2b). This reveals a shift from W-box motif-dependent TSS selection in the maternal transcriptome to zygotic W-box-independent TSS selection; that is, the existence of two major, independent mechanisms for defining transcription initiation acting on the same core promoter.

We conjectured that the uncovered rules for maternal and zygotic TSS selection may apply generally to all constitutively expressed genes, even in the absence of clear TSS shifting. Dinucleotide analysis of all 8,369 constitutively expressed promoters showed the same precise positioning of the W-box signal upstream of the maternal dominant TSS, and alignment of the zygotic TSS with the downstream SS–WW boundary, as seen in the shifting promoters (Fig. 3a and Extended Data Fig. 3a–c). This confirmed a promoterome-wide distinction between determinants that govern TSS selection in the oocyte and the embryo, and demonstrated that complex TSS patterns in constitutively expressed promoters represent readouts of two independent grammars intertwined in the same core promoter.

Finally, we showed that exclusively maternal and exclusively zygotic promoters also utilize the corresponding stage-specific TSS selection signals (Extended Data Fig. 3f–h). These results confirm a global change in promoter interpretation that constitutes a central part of the maternal to zygotic transition, with fundamental differences in the TSS selection mechanism used by the transcription machinery in the oocyte and the embryo.

Fixed spacing between the motif and the TSS imposed by W-box-dependent initiation in the oocyte predicts the sharp TSS configuration\(^9\). The set of maternal broad promoters, which seemingly contradicted the imposed constraints, revealed a novel promoter architecture composed of multiple individual relatively sharp CAGE tag clusters (TCs), each with its associated W-box at a fixed approximately ~30 bp position (Extended Data Fig. 4). On the other hand, the exclusively zygotic promoters showed a less constrained distribution of TSSs, revealing the familiar shape of a broad promoter\(^9\), with the majority (>70%)...
To validate W-box-dependent TSS selection in the oocyte, we analysed Fig. 6, confirming W-box-independent promoter usage in the embryo. Enucleation zygotic transgene activity or zygotic TSS selection (Extended Data (RACE) assays demonstrated that removal of all W-boxes did not influence fluorescence reporter activity and 5′ nucleosome-positioning-signal-related transcription initiation is pervasive. The switch between the maternal and zygotic TSS is accompanied by a global change in the promoter architecture within the same region.

To validate functionally the observed TSS selection grammars, we introduced point mutations disrupting them in red. SL-CAGE TSSs in stable transgenic lines for endogenous sf3a2, the wild-type sf3a2 transgene and mutant sf3a2 transgene are shown in different shades of purple. **P < 0.01, ***P < 0.001, one-tailed Welch’s two sample t-test; mutant, n = 4; wild type, n = 3. c. Subtracted H3K4me3 coverage (Δ) of reads mapping to the plus (+) and minus (−) strand (schematic on top) in three developmental stages at the same set of promoters as in a, d. Density of AA dinucleotide in a ±100 bp region for promoters in a.
suggests that TSS selection in a vertebrate oocyte is independent of inter- and intranucleosomal DNA signals. Recent efforts to identify sequence-based signals for nucleosome positioning and dynamic nucleosome organization at promoters highlight the epigenetic and chromatin mechanisms that, together with the DNA sequence, direct transcription initiation. The association of nucleosome positioning signals with zygotic promoter activity described here raises the question of whether promoter-associated nucleosome positioning contributes to regulation of positioning of transcription initiation, or is merely a consequence of transcription at the predefined position. To investigate this relationship we analysed the DNA sequence underlying +1 nucleosome positioning in the transcriptionally silent pre-(512 cells) and active post-MBT (prim 6) stages (Fig. 4a and Extended Data Fig. 8). In the pre-MBT stage, H3K4me3-marked nucleosomes occupied a CG/GC enriched region and centred at the peak of highest CG/GC enrichment, often directly overlapping the TSSs of the maternal transcripts, supporting the idea that H3K4me3 initially appears at CpG islands before transcription. In the post-MBT stage, the +1 nucleosome was positioned just downstream of the SS–WW enrichment boundary at ~50 bp from the zygotic TSS, occupying a WW-enriched region, with a small local GC/GC peak at the nucleosome midpoint (Fig. 4a and Extended Data Fig. 8b). Additional downstream nucleosomes followed a similar pattern of WW-enriched bound DNA alternating with internucleosomal SS enrichment. Local GC/GC enrichment at the nucleosome midpoints is in accordance with previously described nucleosome positioning preferences; however, the total sequence preference complexity and its relation to TSS in different developmental stages were not reported so far. The results show that the initial positioning of promoter-associated nucleosomes, which correlates with a broad internucleosomal phasing pattern, changes in later stages to a final precise positioning, which correlates tightly with the zygotic TSS and intranucleosomal phasing patterns, suggesting interdependence of the final nucleosome positioning and transcription.

To test this, we ranked throughout-active genes by the timing of onset of their zygotic transcription and analysed their H3K4me3-marked nucleosome positioning patterns (Fig. 4b). No association between the timing of transcription activation and the precision of nucleosome positioning was found, arguing against transcriptionally aided nucleosome readjustment and instead suggesting a pre-transcriptional process for repositioning nucleosomes to their final position, in agreement with transcription-independent positioning of nucleosomes at promoters in human cells. Consistently, H3K4me3 ChIP-seq in TATA-box-binding protein (Tbp) knockdown embryos (Extended Data Fig. 9) showed no change in the overall H3K4me3 recruitment and nucleosome positioning at Tbp-dependent genes (Fig. 4c), demonstrating that H3K4me3-marked nucleosome positioning at these genes does not require Tbp-dependent recruitment of the transcription initiation machinery or active transcription.

The absence of a nucleosome-positioning sequence signature, as well as no precise nucleosome positioning at promoters with a canonical TATA-box in other systems, together with narrow TSS peaks, argues in favour of the W-box as the overriding determinant of maternal TSS selection. The similarity of the W-box to the TATA-box suggests that transcription initiation in the oocyte may be mediated by the oocyte-enriched transcription nucleating factor Tbp2 (also known as Tbp2). Conversely, early zygotic grammar prefers the TSS position to be at a fixed range from the precisely positioned +1 nucleosome, suggesting a mechanism in which the initiation complex chooses initiator-like sequences within a ‘catchment area’ determined by the nucleosome position (Fig. 4d). This model is compatible with motif-independent TFII D recruitment by H3K4me3–TAF3 interactions and emphasizes the interdependence of nucleosome configuration at promoters with promoter type and physiological state in vertebrates and yeast.

Different TSS selection grammars deployed at separate promoters have been associated with different types of genes, and a handful of promoters were shown to switch between TATA-dependent and transcription-independent initiation. We show for the first time, to our knowledge, that H3K4me3-marked nucleosome positioning reveals dynamic changes in underlying sequence signature and relation to TSS during maternal to zygotic transition. a. Frequency of dinucleotides centred on +1 nucleosome of constitutively active promoters in maternal (512 cells) and zygotic (prim 6) stages. Centres of nucleosomes were estimated from subtracted H3K4me3 coverage (grey). Density of maternal and zygotic TSSs is shown in light blue and light red, respectively. b, H3K4me3 signal at promoters of constitutively present transcripts sorted by the time of activation of their zygotic component. Horizontal lines separate groups of promoters that activate the zygotic component at a denoted developmental stage. c, H3K4me3 signal at Tbp-dependent promoters in non-injected embryos (top), embryos injected with mismatch morpholino (middle) or Tbp-targeting morpholino (bottom), sorted by expression fold change between Tbp knockdown (KD) and wild-type (WT) embryos. d, Summary of transcription initiation, TSS configuration and nucleosome positioning dynamics throughout maternal to zygotic transition.
that the two grammars co-exist in close proximity or in physical overlap genome-wide, and are differentially used at thousands of promoters active in both the oocyte and the embryo. The multiple layers of information embedded in the same short sequence, each representing a different aspect of a complex regulation, are part of the reason why promoter codes have been so difficult to detect. Our findings on overlapping promoter grammars have implications for future analyses of promoter content and function.

**METHODS SUMMARY**

TSSs defined by CAGE were clustered into TCs, and TCs into promoter regions, allowing a maximal distance of 20 and 100 bp, respectively. Dinucleotide density was calculated by applying two-dimensional Gaussian kernel smoothing on a dinucleotide occurrence matrix. TSS usage in the 5′u2 promoter was analysed in transgene-injected embryos by 5′ RACE and in stable transgenic zebrafish lines by SL-CAGE. ChiP was carried out on MNase-treated chromatin using the ChIP-IT Express Enzymatic kit (Active Motif), and libraries were sequenced for 36 or 56 bp on the HiSeq 2000 system. Minus-strand coverage was subtracted from plus-strand reads coverage for visualization and nucleosome midpoint estimation. Tbp was knocked down by injecting one-cell-stage zebrafish embryos with the tbp morpholino.

**Online Content** Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 6 February; accepted 23 December 2013.**

**Published online 16 February; corrected online 19 March 2014 (see full-text HTML version for details).**

1. D’Alessio, J. A., Wright, K. J. & Tjian, R. Shifting players and paradigms in cell-specific transcription. Mol. Cell 36, 924–931 (2009).
2. Kadonaga, J. T. Perspectives on the RNA polymerase II core promoter. Nature 418, 1938–1950 (2002).
3. Mathavan, S. et al. Transcriptome analysis of zebrafish embryogenesis using microarrays. PLoS Genet. 1, e29 (2005).
4. Tadros, W. & Lipshitz, H. D. The maternal-to-zygotic transition: a play in two acts. Development 136, 3033–3042 (2009).
5. Vastenhouw, N. L. et al. Chromatin signature of embryonic pluripotency is established during genome activation. Nature 464, 922–926 (2010).
6. Lindeman, L. C. et al. Prepatternning of developmental gene expression by modified histones before zygotic genome activation. Dev. Cell 21, 993–1004 (2011).
7. Kane, D. A. & Kimmel, C. B. The zebrafish midblastula transition. Development 119, 447–456 (1993).
8. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl Acad. Sci. USA 100, 15776–15781 (2003).
9. Barski, A. et al. High-resolution profiling of histone methylation in the human genome. Cell 129, 823–837 (2007).
10. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nature Genet. 38, 626–635 (2006).
11. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature Rev. Genet. 13, 233–245 (2012).
12. Venter, J. S. & Pugh, B. F. Genomic organization of human transcription initiation complexes. Nature 502, 53–58 (2013).
13. Nepal, C. et al. Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. Genome Res. 23, 1958–1950 (2013).
14. Giraldez, A. J. et al. Zebrafish mir-430 promotes deadenylation and clearance of maternal mRNAs. Science 321, 75–79 (2006).
15. Ponte, J. C. et al. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. Genome Biol. 7, R78 (2006).
16. Segal, E. et al. A genomic code for nucleosome positioning. Nature 442, 772–778 (2006).
17. Ishikihara, H. et al. Variety of genomic DNA patterns for nucleosome positioning. Genome Res. 21, 1863–1871 (2011).
18. Segal, E. & Widom, J. What controls nucleosome positions? Trends Genet. 25, 335–343 (2009).
19. Naeve, H. S. et al. Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature 483, 295–301 (2012).
20. Racht, E. A. et al. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. PLoS Genet. 7, e1001274 (2011).
21. Caims, B. R. The logic of chromatin architecture and remodelling at promoters. Nature 461, 193–198 (2009).
22. Lardon, A. M. & Bird, A. CpG islands and the regulation of transcription. Genes Dev. 25, 1010–1022 (2011).
23. Thomson, J. P. et al. CpG islands influence chromatin structure via the Cspg-binding protein Cfp1. Nature 464, 1082–1086 (2010).
24. Fenouil, R. et al. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. 22, 2399–2408 (2012).
25. Nozaki, T. et al. Tight associations between transcription promoter type and epigenetic variation in histone positioning and modification. BMC Genomics 12, 416 (2011).
26. Bärtai, R. et al. TB2P, a vertebrate-specific member of the TBP family, is required in embryonic development of zebrafish. Curr. Biol. 14, 593–598 (2004).
27. Akhtar, W. & Veenstra, G. TB2P is a substitute for TBP in Xenopus oocyte transcription, BMC Biol. 7, 45 (2009).
28. Lauberth, S. M. et al. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. Cell 152, 1021–1036 (2013).
29. Zaug, J. B. & Luscombe, N. M. A genomic model of condition-specific nucleosome behavior explains transcriptional activity in yeast. Genome Res. 22, 84–94 (2012).
30. Davis, W. J. & Schultz, R. M. Developmental change in TATA-box utilization during preimplantation mouse development. Dev. Biol. 218, 275–283 (2000).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The authors are grateful to L. Tora, E. Kenyon, G. Chalancon and J. C. Y. Chen for comments on the manuscript, to M. Reischl for Zebrafish Miner software, and to L. O’Neill for technical advice. V.H., C.N., C.Pr., A.A. and X.D. were supported by grants from the Norwegian Research Council (YFF) and the Bergen Research Foundation, awarded to B.L. F.M. and U.S. by the EU FP7 project Dopaminet and U.S. by the EU FP6 project NeuroXSys. C.Pr., C.N., C.Pl. and A.M.S. developed and performed SL-CAGE with input from P.C. V.H., B.L. and F.M. analysed the data and wrote the Supplementary Information. B.L., F.M., P.C. and V.H. conceived the study. N.L., Y.H., J.G. and D.C. performed microarrays. SRA104816. Reprints and permissions information is available at www.nature.com/ reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.L. (b.lenhard@imperial.ac.uk), F.M. (f.mueller@bham.ac.uk) or P.C. (carninci@riken.jp).
METHODS

CAGE tags mapping and CTSS calling. Sequenced CAGE tags (27 bp) previously described13 were mapped to a reference zebrafish genome (Zv9/danRer7 assembly) using Bowtie34 with default parameters allowing up to two mismatches and keeping only uniquely mapped reads. An additional G nucleotide, which is often attached to the 5’ end of the tag by the template-free activity of the reverse transcriptase in the cDNA preparation step of the CAGE protocol25, was removed in cases in which it did not map to the genome. All unique 5’ ends of tags that were considered CAGE tag-defined TSSs (CTSSs) and the number of tags supporting each CTSS was counted. Raw tag count was normalized to a referent power-law distribution based on a total of 10^6 tags and x = −1.25, as described previously33, resulting in normalized tags per million (tpm). All analyses were done in the R statistical computing environment16 (http://www.R-project.org/) using Bioconductor17 (http://www.bioconductor.org/) software packages and custom scripts. CTSS clustering into TCs and promoter regions. CTSSs supported by at least 1 tpm in at least one of the 12 developmental stages were clustered at two levels. First, TCs were created for each stage individually using a simple distance-based approach with a maximum allowed distance of 20 bp between two neighbouring CTSSs. Next, for each TC we calculated a cumulative distribution of the CAGE signal and determined the positions of the 10th and 90th percentile to obtain more robust boundaries of a TC. TCs across all developmental stages within 100 bp of each other were aggregated into a single promoter region. Only promoter regions supported by at least 5 tpm in at least one developmental stage were used in further analyses.

Expression profiling. Expression profiling was done at two levels: individual CTSSs, and entire promoter regions. To minimize the noise from weakly supported CTSSs, we selected only CTSSs with at least 5 tpm in at least one developmental stage. Normalized tpm values across 12 developmental stages for each CTSS (or promoter region) were divided by their standard deviation to obtain scaled expression measures. A self-organizing map (SOM) unsupervised learning algorithm was applied to distribute CTSSs (or promoter regions) across 5 × 5 = 25 expression profiles.

Dinucleotide pattern analysis. To visualize dinucleotide composition patterns of sequences flanking TSSs, we first created an occurrence matrix (n × m; where n = number of sequences and m = length of sequences) for each individual dinucleotide, by placing 1 if the given dinucleotide is present at a given position or 0 if it is not. Values in the matrix were then smoothed: at each position in the matrix the weighted average dinucleotide occurrence was calculated by taking into account surrounding positions. Weights of the surrounding positions were assigned by centring a two-dimensional Gaussian kernel with bandwidth = 3 (in both dimensions) at the central position. The matrix of smoothed values (densities) was visualized using different shades of blue in a map-like representation. Extended Data Fig. 10 illustrates how the calculation and visualization was done.

TATA-box motif analysis. A TATA-box PWM was obtained from the JASPAR database36 (http://jaspar.genereg.net/) and used to scan the region −35 bp to −22 bp upstream of the TSS (expected position for a TATA-box was determined as previously described33) for a maximal detected match (%) to the TATA-box PWM was reported for the maternal and zygotic dominant TSS. Normalized tpm values across all promoters was visualized by histograms. In addition, the frequency of the top ten most abundant pentamers found in the scanned sequences was shown. De novo motif discovery was performed on a set of 14 bp long sequences spanning the region from −35 bp to −22 bp upstream of the TSS using MEME37 (http://meme.sdsu.edu/) with default parameters. Only motifs with an E value ≤ 0.01 were selected as significant.

sf3a2 promoter reporter constructs. The region spanning 500 bp upstream and 200 bp downstream of the dominant zygotic TSS in the sf3a2 promoter was chosen for validation of the TSS selection grammar. The sf3a2 promoter carries both maternal and zygotic promoters determined and exhibits TSS shifting. The selected sequence, which ends within the first intron of sf3a2, was fused to a sequence containing the 3’ end of the zebrafish tknpi promoter first intron and splice acceptor fused to an mCherry reporter (Extended Data Fig. 6a). Genomic DNA of AB strain was used for PCR amplification using Advantage HD DNA Polymerase Mix (Clontech) following the manufacturer’s instructions. The PCR products of the expected size from the nested (inner) PCR reaction were purified from agarose gel and sequenced to identify the TSS. To demonstrate that the generated 5’ RACE products are specific to the 5’ ends of de-capped RNA, a ‘minus tobacco acid pyrophosphatase’ (TAP)–treated sample was put through adaptor ligation, reverse transcription and PCR. Sequences of primers used in 5’ RACE are provided in Extended Data Fig. 6f. Transgenic zebrafish lines. Transgenic zebrafish lines with the sf3a2 promoter (wild-type and mutated) reporter constructs were generated by microinjection of the corresponding construct into zebrafish zygotes as described earlier. The reporter-positive (mCherry) embryos were grown to adulthood and germline-transmitting female individuals were identified by crossing to wild-type zebrafish and selecting for the presence of reporter-expressing offspring. Transgene expression and TSS usage was analysed in F1 generation embryos. Experiments were carried out under licence by the Home Office Licence Number 40/3681 and PPL 40/3131.

Quantification of the reporter mRNA levels by qPCR. RNA from reporter-expressing embryos in the high/sphere stage was isolated using GeneElute Total RNA extraction kit (Sigma–Aldrich), following the manufacturer’s instructions. qPCR was performed using the SYBR Green detection method on 7900HT Fast Real-Time PCR System (Applied Biosystems). Two primer pairs were used for both the mCherry reporter and the endogenous sf3a2 gene (normalization control). Technical triplicates were run for each primer pair. The C values were determined by the SDS v.2.4 software (Applied Biosystems), using a manual threshold of 0.2 and automatic base line. Expression levels of the transgene were calculated relative to the endogenous sf3a2 in the same sample, using the average C values across technical triplicates and both primer pairs. The sequence of the primers used in qPCR is provided in Extended Data Fig. 7c.

SL-CAGE. We introduced a novel method for quantitative high-resolution detection of TSSs and their usage within a targeted promoter, called single-locus deep CAGE (SL-CAGE). SL-CAGE utilizes CAGE tags and their usage within a targeted promoter, called single-locus deep CAGE (SL-CAGE). SL-CAGE utilizes CAGE tags and their usage within a targeted promoter, called single-locus deep CAGE (SL-CAGE). SL-CAGE utilizes CAGE tags and their usage within a targeted promoter, called single-locus deep CAGE (SL-CAGE).
25 μl Protein G magnetic beads, 20 μl ChIP buffer 1, 1 μl protein inhibitor cocktail, 4 μg of anti-H3K4me3 (Abcam ab8580) or anti-H2A.Z (Abcam ab4174) antibody or an equivalent volume of water (no antibody control), respectively, and water to a final volume of 200 μl. ChIP was performed in duplicates for each stage. ChIP reactions were then incubated overnight at 4 °C while rotating. Magnetic beads were washed and incubated in elution buffer. After addition of reverse crosslinking buffer samples were decrosslinked for 4 h at 65 °C. Samples were treated with Proteinase K and RNase A and purified using phenol chloroform extraction.

**Tbp knockdown.** One-cell-stage embryos were injected with either β-actin:yfp13 (1.7 nl, 42 pg nl⁻¹) or 1.3 kb upstream of the ntl TSS ntl:yfp25 (1.7 nl, 68 pg nl⁻¹) constructs. The injected embryos were then split into four groups. One was kept as a non-injected control and the other three groups were further injected with one of the two Tbp-targeting morpholinos (1.7 nl, 2.5 mM) or with a mismatch morpholino described previously40. All embryos were kept in E3 medium at 28.5 °C until the non-injected group reached the 30% epiboly stage (4.7 hpf) and were then analysed under a fluorescence stereoscope (Nikon SMZ1500). Arrest of epiboly for 36 bp or 56 bp (1.7 nl, 42 pg nl⁻¹) constructs. The injected embryos were then split into four groups. One was kept as a non-injected control and the other three groups were further injected with one of the two Tbp-targeting morpholinos (1.7 nl, 2.5 mM) or with a mismatch morpholino described previously. All embryos were kept in E3 medium at 28.5 °C until the non-injected group reached the 30% epiboly stage (4.7 hpf) and were then analysed under a fluorescence stereoscope (Nikon SMZ1500). Arrest of epiboly movements, loss of β-actin:yfp (Tbp-dependent) and retention of ntl:yfp (Tbp-independent) reporter activities42 were used as markers for assessing and sorting Tbp morphants for ChIP analysis. Approximately 1,500 non-injected, 1,500 mismatch-morpholino-injected embryos, 1,200 ntl mol morphants and 1,000 ntl mol2 morphants were used for ChIP, as described earlier. We used a previously published set of genes downregulated in zebrafish Tbp morphants43. Genes with log fold change ≤ −1.5 were selected as Tbp dependent and were aligned with respect to the dominant TSS of the nearest promoter detected by CAGE for H3K4me3 ChIP-seq signal visualization.

**Deep sequencing of chromatin DNA.** ChIP-seq was performed as described before44. In brief, 10 ng of ChIP DNA was end-repaired, ligated to single read adaptors, size selected and amplified for 18 cycles according to Illumina’s ChIP-seq protocol. Cluster generation was performed according to the Illumina Cluster Reagents preparation protocol (http://www.illumina.com/). Samples were sequenced for 36 bp or 56 bp (Tbp morphants and controls) on the HiSeq 2000 system. ChIP-seq data analysis. Sequenced reads were mapped to the reference zebrafish genome (Zv9/danRer7 assembly) using Bowtie21 with default parameters allowing up to two mismatches and keeping only uniquely mapped reads. Coverage was calculated for plus and minus strands separately using unextended reads and taking maximum 20 reads mapping to exactly the same position. Minus-strand coverage was subtracted from plus-strand coverage to obtain subtracted coverage, which was used for visualization and nucleosome midpoint estimation. Significantly enriched regions (peaks) were detected using model-based analysis for ChIP-seq (MACS2) (http://liulab.dfci.harvard.edu/MACS/) with default parameters. Midpoints of nucleosomes within significantly enriched regions (FDR ≤ 0.01) were estimated from subtracted coverage and the nearest CAGE signal was used to determine strand specificity and relative position of the first downstream nucleosome.

**Data accession.** Processed data are available for download at http://promshift.genereg.net/zebrafish/. Tracks can be visualized as annotated custom tracks in the UCSC Genome Browser using the following URLs: http://promshift.genereg.net/zebrafish/CAGE_and_nucleosome_tracks.txt and http://promshift.genereg.net/zebrafish/Transgenic_lines_sf3a2_SL-CAGE_tracks.txt.
Extended Data Figure 1 | Promoters contain TSSs with very heterogeneous developmental expression profiles. 

a, Developmental expression profiles of CAGE-defined promoters obtained by self-organizing map (SOM) clustering. Each box represents one SOM cluster, with a series of beanplots showing distribution of scaled expression (logarithm of normalized number of CAGE tags per million) at different time points for all promoters belonging to that cluster (number of promoters denoted above the box). Different classes of promoters marked in different colours were used in all downstream analyses.

b, Projection of maternal- and zygotic-specific CTSSs (from Fig. 1b) onto the SOM classes of promoters shown in a. 

c, Fraction of promoters containing either maternal-specific (cluster 4,4 in Fig. 1b) or zygotic-specific (cluster 0,0 in Fig. 1b) CTSS, or both, for different classes of promoters established in a.

d, Example of a throughout-active promoter that contains a very heterogeneous population of CTSSs (colour coding of CTSSs as in Fig. 1b: maternal-specific shown in blue, throughout-expressed shown in green, and zygotic-specific shown in red).
Extended Data Figure 2 | Definition and properties of shifting promoters. 

a, Schematic representation of shifting promoter score calculation between two stages. Forward and reverse scores are calculated using forward (F1, F2) and reverse (F1', F2') cumulative sums of CAGE tags along the promoter region, and the final shifting score is the maximum of the two. 
b, Heatmap showing the number of shifting promoters detected by applying the method shown in a (using a 0.6 score threshold) to all possible pairs of developmental stages. The majority of shifting was detected between the first six stages (predominantly maternal) and last three stages (predominantly zygotic). Promoters that qualified as shifting in at least four of these early versus late pairs were selected for a final set of 911 shifting promoters. 
c, Distribution of distance between maternal and zygotic dominant TSS (shift) for 911 shifting promoters. 
d, Developmental expression dynamics of maternal (blue) and zygotic (red) dominant TSS in shifting promoters. Average expression for the entire promoter is shown in black. e, Density of AA, AT, TT and CG dinucleotides for shifting promoters sorted by magnitude of shift. Promoters were centred at either the maternal (left) or the zygotic (right) dominant TSS. Blue arrowhead marks the AA/AT/TT enrichment ~30 bp upstream of the maternal TSS and red arrowhead marks the boundary between CG and AA/AT/TT enrichment ~50 bp downstream of the zygotic TSS. f, Relative frequency of top ten most frequent TATA-like pentamers in the region ~35 to ~22 bp upstream of maternal (blue) and zygotic (red) dominant TSS. Percentage of match to TATA-box PWM is denoted in the brackets.
Extended Data Figure 3 | Genome-wide sequence and chromatin signature of maternal versus zygotic TSSs. a, Dinucleotide density for all throughout-active promoters sorted by orientation-sensitive distance between the maternal and zygotic dominant TSS. Promoters were centred at either the maternal (left) or the zygotic (right) dominant TSS. b, Distribution of match (%) to TATA-box PWM in the region −35 to −22 bp upstream of the maternal (blue) and zygotic (red) dominant TSS in constitutively active promoters. P value of the two-tailed Wilcoxon rank-sum test is shown. c, Relative frequency of top ten most frequent TATA-like pentamers upstream of the maternal (blue) and zygotic (red) dominant TSS. Percentage of match to TATA-box PWM is denoted in the brackets. d, H2A.Z signal in the prim 6 stage at the same set of promoters as in a. Subtracted coverage of reads mapping to the plus and minus strand is shown. e, Density of TT dinucleotide in a ±100 bp region for the same set of promoters as in a. f, WW dinucleotide density for maternal- (left) and zygotic-specific (right) promoters centred at their dominant TSS and sorted by decreasing match to the TATA-box (white dashed lines). Average density of individual dinucleotides is shown below. g, Distribution of match (%) to TATA-box PWM in the region −35 to −22 bp upstream of dominant TSS for maternal- (blue) and zygotic-specific (red) promoters. P value of the two-tailed Wilcoxon rank-sum test is shown. h, Relative frequency of the top ten most abundant TATA-like pentamers in maternal (blue) and zygotic (red) promoters. Percentage of match to TATA-box PWM is denoted in the brackets.
Extended Data Figure 4 | Composite multiple sharp architecture of maternal-specific promoters. a, Distribution of promoter interquantile width ($q_{0.1} - q_{0.9}$; central promoter region that contains $\geq$80% of CAGE tags) for 1,944 maternal-specific promoters. b, Number of TCs within the promoter region for maternal broad promoters (interquantile width $>20$ bp) across early developmental stages. Bubble size reflects relative number of promoters that contain given number of TCs. c, Comparison of dinucleotide frequencies around single dominant TSS in promoter (left) and dominant TSSs from all TCs within the same promoter (right). d, Composite architecture of a maternal broad promoter showing multiple sharp TCs with their associated W-box at the expected $-30$ bp upstream position.
Extended Data Figure 5 | Change of promoter architecture during maternal to zygotic transition. 

**a**, Examples of maternal-specific, zygotic-specific and throughout-active promoters showing difference in architecture between maternal and zygotic stage. 

**b**, Number of TCs within promoter region for maternal-specific, zygotic-specific and constitutively active promoters. Bubble size reflects relative number of promoters that contain given number of TCs and different colours correspond to different developmental stages. 

**c**, Distribution of distance between neighbouring TCs within the same promoter region. 

**d**, Heatmap showing $k$-means clustering ($k = 15$) of throughout-active promoters, based on the number of TCs that they contain throughout 12 developmental stages. 

**e**, Representative examples for three selected clusters: highly expressed sharp promoter whose architecture does not change and retains only one TC (left); promoter in which number of TCs decreases from five clearly separated sharp TCs in maternal stages to one fairly broad TC in zygotic stages (middle); promoter that changes its architecture from two clearly separated sharp TCs to one broader TC after maternal to zygotic transition (right).
**Extended Data Figure 6** | Validation of sfa2 promoter code in transgene injected embryos. 

**a**, Schematic of reporter construct; the dashed lines indicate the spliced out intronic sequence. SA, splice acceptor; SD, splice donor. 

**b**, Primers used for sfa2 construct. 

| Primer set                        | Sequence                        |
|----------------------------------|---------------------------------|
| sfa2 promoter region wt          | F: agattctaccctgatcctgac         |
|                                  | R: tcaadea3cgaacacagatracacacac  |
| tempa intron                     | F: gctgctctgtaaaagcgagaggg      |
|                                  | R: ctgctctgtaaaagcgagaggg       |
| sfa2 promoter region site-directed mutagenesis | F: TTRTGUGCAACTCTGACCGCTGCGTGCAAGTACAGGGGAGAATTGCAACTGACGGCGAGCG |
|                                  | R: AGTCAAGTACGAGGATAGTTTC       |

**c**, Reporter gene expression in the wild-type and mutant sfa2 construct-injected embryos (36 hpf). Bright field (BF) and red channel fluorescent images (mCherry) are shown. Scale bar, 250 μm. The panels on the right show red channel projection overlay of 24 wild-type (WT, top) and 31 mutant (mut, bottom) embryos. 

**d**, Percentage of mCherry-expressing embryos from total ECFP-positive embryos (left) and overall reporter activity measured as average pixel intensity values from 24 wild-type and 31 mutant sfa2 promoter-injected embryos (right). 

**e**, 5' RACE of sfa2 promoter construct. Fragments amplified by PCR using a forward primer specific for the 5' RACE adaptor and two different inner reverse primers specific for the mCherry reporter. The arrowheads indicate the PCR products of expected size. A, 5' RACE adaptor; IF, inner forward primer; IR1, inner reverse primer 1; IR2, inner reverse primer 2; M, DNA marker; mut, embryos injected with construct containing sfa2 promoter with mutations disrupting W-boxes (as shown in f); TAP, tobacco acid pyrophosphatase; WT, embryos injected with construct containing wild-type sfa2 promoter. 

**f**, Primers used in 5' RACE. 

| Primer set                        | Sequence                        |
|----------------------------------|---------------------------------|
| Outer PCR forward primer         | GCTGAGCTGCAAGATGACACCTG         |
| Inner PCR forward primer         | CGCCGATCCACCTGCGTTGCGCTCPG      |
| Outer PCR mCherry reverse primer | TEACAGAAGACGCTGCGCTCGCGCTG       |
| Inner PCR mCherry reverse primer (IR1) | CTCACGCTGCAAGACGCTGCGCTG       |
| Inner PCR mCherry reverse primer (IR2) | CTCACGCTGCAAGACGCTGCGCTG       |

**g**, Genomic sequence of the sfa2 promoter with point mutations introduced to disrupt W-boxes (as shown in f) and 5' RACE of the wild-type and the mutated sfa2 promoter construct. Tags per million (TPM) on the y-axis. Position relative to zygotic dominant TSS (bp) on the x-axis.
Extended Data Figure 7 | Validation of sf3a2 promoter code in stable transgenic lines. a, sf3a2:mCherry transgene expression in wild-type (WT; top four panels) and mutant (MUT; bottom four panels) sf3a2 promoter transgenic embryos (high stage). Scale bar, 1 mm. b, Quantitative polymerase chain reaction (qPCR) results of transgene expression: ratio between abundance of sf3a2:mCherry trasgene mRNA and endogenous sf3a2 mRNA. Bar height shows mean value across six qPCR experiments (three replicas for each of the two different primer pairs) and error bars denote 95% confidence interval of the mean. c, Sequence of primers used in qPCR. d, TSSs and their relative usage detected by SL-CAGE in three wild-type and four mutant sf3a2 promoter transgenic lines. Promoter sequence is shown at the top, with point mutations introduced to disrupt W-boxes highlighted in red. For each transgenic line two tracks are shown: TSSs used by the endogenous sf3a2 promoter (upper track, more intensive colour) and TSSs used by the transgene (lower track, lighter colour). P value (one-tailed Welch’s two sample t-test) of the difference between mutated and wild-type promoter variants is shown for four maternal TSSs associated with disrupted W-boxes (dashed boxes).
Extended Data Figure 8 | H3K4me3-marked nucleosome positioning signal in maternal and zygotic stage. a, Venn diagrams showing number of promoters containing H3K4me3 peak within ±1 kb of TSS in early (512 cells) and late (prim 6) stage for distinct classes of promoters (as defined in Extended Data Fig. 1a). b, Frequency of selected dinucleotides centred at the position of the first downstream (+1) nucleosome for distinct classes of promoters. Only promoters containing an H3K4me3 peak in both 512 cells and prim 6 stages were used. Centres of nucleosomes were estimated from subtracted H3K4me3 coverage (shown in grey). Density of TSSs in maternal (512 cells) and zygotic (prim 6) stages is shown in light blue and light red, respectively. Note that maternal-specific promoters have only maternal TSSs and the distribution of these TSSs in the zygotic stage is shown only for orientation. The opposite is true for zygotic-specific promoters.
Extended Data Figure 9 | Validation of Tbp morphant phenotypes before ChIP-seq analysis. Tbp morphants show epiboly defects and differential loss of reporter activity as previously described. Embryos in the 30% epiboly stage showing epiboly delay and corresponding Tbp-independent (ntl:yfp) and Tbp-dependent (β-actin:yfp) reporter gene activities were used for H3K4me3 ChIP-seq analysis. Zebrafish fertilized eggs were injected with the reporter constructs as indicated on the right followed by injection of morpholino (Tbp 5 mismatch or Tbp mo1 morpholino, respectively). Fluorescence images on top, bright field images below. ntl, notail; yfp, yellow fluorescent protein. Scale bar, 400 μm.
Extended Data Figure 10 | Dinucleotide maps: calculation and visualization of dinucleotide patterns. Genomic sequences (of the same length) are sorted and aligned into a matrix-like representation (P₁–P₁₀ positions; S₁–S₁₀ sequences). Marking the presence of selected dinucleotide (for example, AA) by 1 and the absence by 0 creates an occurrence matrix. Next, a weighted average is calculated at each position by placing a two-dimensional Gaussian kernel at that position and assigning weights to surrounding positions. An example of calculating the value at position S₄P₇ is shown. Surrounding positions are coloured on the basis of the weights assigned to them by the Gaussian kernel (bandwidth = 1 in both dimensions, and covariance = 0 between the two dimensions). Averaged values are mapped to different shades of blue to visualize the dinucleotide density across the set of input sequences.