Published in final edited form as:
Nature. 2014 March 20; 507(7492): 381–385. doi:10.1038/nature12974.

Two independent transcription initiation codes overlap on vertebrate core promoters

Vanja Haberle[#1,2], **Nan Li**[#3], **Yavor Hadzhiev**[3], **Charles Plessy**[4,5], **Christopher Previti**[6,*], **Chirag Nepal**[6,$], **Jochen Gehrig**[3,£], **Xianjun Dong**[6,%], **Altuna Akalin**[6,&], **Ana Maria Suzuki**[4,5], **Wilfred F.J. van IJcken**[7], **Olivier Armant**[8], **Marco Ferg**[8], **Uwe Strähle**[8], **Piero Carninci**[4,5,+], **Ferenc Müller**[3,+], and **Boris Lenhard**[2,9,+]

[1]Department of Biology, University of Bergen, Thormøhlensgate 53A, N-5008 Bergen, Norway [2]Institute of Clinical Sciences and MRC Clinical Sciences Center, Faculty of Medicine, Imperial College London, Hammersmith Hospital, Du Cane Road, London W12 0NN, United Kingdom [3]School of Clinical and Experimental Medicine, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK [4]RIKEN Omics Science Center, Yokohama, Kanagawa, 230-0045 Japan (ceased to exist on 01 April 2013 due to RIKEN reorganisation) [5]RIKEN Center for Life Science Technologies, Division of Genomic Technologies, RIKEN Yokohama Campus, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan [6]Computational Biology Unit, Uni Computing, Uni Research AS, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway [7]Erasmus Medical Center, Center for Biomics, Room Ee679b, Dr Molewaterplein 50, 3015 GE Rotterdam, The Netherlands [8]Institute of Toxicology and Genetics, Karlsruhe Institute of Technology, Postfach 3640, 76021 Karlsruhe, Germany [9]Department of Informatics, University of Bergen, Thormøhlensgate 55, N-5008 Bergen, Norway

[+]Corresponding authors: Piero Carninci (carninci@riken.jp) Ferenc Müller (f.mueller@bham.ac.uk) Boris Lenhard (b.lenhard@imperial.ac.uk) .
[*]Present address: German Cancer Research Center (DKFZ), Genomics & Proteomics Core Facility (GPCF), Im Neuenheimer Feld 580/TP3, Heidelberg 69120, Germany
[$]Present address: Broegelmann Research Laboratory, The Gade Institute, University of Bergen, The Laboratory Building, Haukeland University Hospital, N-5021 Bergen, Norway
[£]Present address: Acquifer AG, Sophienstraße 136, 76135 Karlsruhe, Germany
[%]Present address: Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston MA, USA
[&]Present address: Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, 4058 Basel, Switzerland

Author contributions: B.L., F.M, P.C. and V.H. conceived the study. N.L., Y.H., J.G., and M.F. performed experiments. O.A. and W.v.I. performed sequencing. V.H., C.P., C.N., X.D. and A.A. performed computational analyses. 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 manuscript with input from P.C, W.v.I. and U.S.

Extended Data is linked to the online version of the paper at www.nature.com/nature.

High throughput sequencing data has been deposited at the NCBI Sequence Read Archive (SRA) under accession numbers SRA097279 and SRA104816.

The authors declare no competing financial interests.

Data accession: Processed data is 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

Europe PMC Funders Author Manuscripts

[#] These authors contributed equally to this work.

Europe PMC Funders Author Manuscripts

Abstract

A core promoter is a stretch of DNA surrounding the transcription start site (TSS) that integrates regulatory inputs[1] and recruits general transcription factors to initiate transcription[2]. The nature and causative relationship of DNA sequence and chromatin signals that govern the selection of most TSS by RNA polymerase II remain unresolved. Maternal to zygotic transition (MZT) represents the most dramatic change of the transcriptome repertoire in vertebrate life cycle[3-6]. 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 10th cell cycle, marking the *midblastula transition* (MBT)[7]. 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 (CAGE)[8] and determined the positions of H3K4me3-marked promoter-associated nucleosomes[9]. We show that the transition from maternal to zygotic transcriptome is characterised 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 selection grammar characterized by broader patterns of dinucleotide enrichments, precisely aligned with the first downstream (+1) nucleosome. The developmental dynamics of the H3K4me3-marked nucleosomes reveals their DNA sequence-associated positioning at promoters prior to zygotic transcription and subsequent transcription-independent adjustment to the final position downstream of zygotic TSS. The two TSS-defining grammars coexist often in physical overlap in core promoters of constitutively expressed genes to enable their expression in the two regulatory environments. The dissection of overlapping core promoter determinants represents a framework for future studies of promoter structure and function across different regulatory contexts.

Mapping of transcription start sites using CAGE[8] identified two major promoter classes with respect to the TSS precision[10,11]: “sharp” promoters with one predominant TSS often associated with a TATA-box that determines the TSS selection, and “broad” promoters with a wider distribution of TSSs often overlapping a CpG island. Even with recent reports of prevalence of known core promoter elements in human promoters[12], the actual mechanism for choosing TSSs within vertebrate promoters in various cell types and conditions remains unknown.

To address the 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 unfertilised egg to organogenesis[13] (Fig. 1a). The data revealed numerous cases of promoter dynamics, where maternal mRNAs 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 (pre-MBT) stages (blue)

reflecting maternally inherited transcripts, as opposed to those activated in early (orange) or later (red) zygotic stage (post-MBT). Additional clusters included constitutively present TSSs (green) and TSSs with peak activity at the transitional stages (yellow), confirming major changes in the zebrafish transcriptome initiated at MBT[3,14]. 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 population of TSSs with very heterogeneous relative usage during development (Fig. 1a, Extended Data Fig. 1b-d).

The observed differential TSS utilisation 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 underlying signatures guiding differential promoter interpretation by the maternal and zygotic transcription machinery, we further dissected the maternal- and zygotic-specific promoter usage. We first identified a subset of promoters similar to the example in Figure 1a, showing a significant degree of shifting between maternally and zygotically utilized 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 bp (Extended Data Fig. 2c). Their preferred maternal and zygotic TSS displayed 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 maternal TSS, hinting at the presence of a functional TATA-like element[15] (Fig. 2a, Extended Data Fig. 2e). In contrast, zygotic TSS did not show TATA-like signal in the expected position, but a sharp SS|WW boundary in local C/G and A/T dinucleotide enrichment precisely aligned ~50 bp downstream of zygotic TSS (Fig. 2a, Extended Data Fig. 2e). This suggests 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 canonical TATA-box, but equally positioned 30 bp upstream of the maternal TSS (Fig. 2c). In contrast, zygotic TSS did not show presence of 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, *i.e.* the existence of two major, independent mechanisms for defining transcription initiation acting on the same core promoter.

We hypothesized that the uncovered rules for maternal and zygotic TSS selection may apply generally for all constitutively expressed genes, even in the absence of clear TSS shifting. The dinucleotide analysis on all 8369 constitutively expressed promoters showed the same precise positioning of the W-box signal upstream of the maternal dominant TSS, and the alignment of zygotic TSS with downstream SS|WW boundary, as seen in the “shifting”

promoters (Fig. 3a, 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 regions.

Finally, we showed that exclusively maternal and exclusively zygotic promoters also utilise 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 maternal to zygotic transition, with fundamental difference 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 the W-box-dependent initiation in the oocyte predicts “sharp” TSS configuration[10]. 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 fixed ~-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[10], with majority (>70%) containing only one broad TC (Extended Data Fig. 5a-c). Constitutively used promoters changed their shape accordingly, from single or multiple “sharp” TSS configuration in the maternal stages, to “broad” in zygotic stages (Extended Data Fig. 5d,e). Thus, the switch between maternal and zygotic TSS is accompanied by a global change in the promoter architecture within the same region.

To functionally validate the observed TSS selection grammars, we identified the W-boxes and dinucleotide frequency patterns in a constitutively active promoter (*sf3a2*) and mutated all W-boxes associated with maternal TSSs (Fig. 3b) for analysis in transgenic zebrafish. Fluorescence reporter activity and 5′ RACE assays demonstrated that removal of all W-boxes did not influence zygotic transgene activity or zygotic TSS selection (Extended Data Fig. 6), confirming W-box independent promoter usage in the embryo. To validate the W-box dependent TSS selection in the oocyte, we analysed maternal TSS selection in early F1 embryos from stable transgenic lines with wild type or mutated variant of the *sf3a2* promoter. We developed a novel method (single locus CAGE; SL-CAGE; Supplementary Table 2) for detection and relative quantification of TSS usage at 1 bp resolution within targeted promoter. TSS usage patterns of the wild type *sf3a2* transgene promoter in early embryos of several transgenic lines were highly reproducible and perfectly correlated with the maternal TSS usage of the endogenous *sf3a2* gene as well as with that seen by CAGE (Fig. 3b, Extended Data Fig. 7). The removal of W-boxes severely reduced the use of associated downstream positions as TSSs in the mutant transgenic lines compared to the wild type and led to an aberrant TSS usage pattern (Fig. 3b, Extended Data Fig. 7d; $P < 0.01$), confirming that the selection of these TSSs depends on W-box signal in the oocyte. These results strongly support two independent TSS selection mechanisms used by the oocyte and the embryo within a single promoter.

To address the relationship between stage-specific TSS selection and chromatin configuration, we analysed the positioning of H3K4me3 and H2A.Z containing nucleosomes

at core promoters by ChIP-seq. It was shown that H3K4me3 marking on promoter associated nucleosomes precedes gene transcription during zebrafish genome activation[5,6]. The data revealed precise positioning of the first downstream (+1) nucleosome ~50 bp from the preferred zygotic TSS, but no fixed spacing to the maternal TSS for all constitutively active promoters in the zygotic (prim 6) stage (Fig. 3c, Extended Data Fig. 3d). We observed a less sharp nucleosome alignment to the zygotic TSS in earlier stages, including the 512 cells stage, which precedes the onset of zygotic transcription. In contrast, no alignment of H3K4me3-marked nucleosomes to the maternal TSS was detected in any stage (Fig. 3c). These results revealed a positional interdependency between zygotic TSS and the +1 nucleosome in the embryo as a feature of zygotic TSS selection grammar, independent of the W-box motif-guided TSS selection in the oocyte. Promoter-associated nucleosome alignment corresponded with alternating WW|SS patterns downstream of zygotic TSS (asterisks in Fig. 3a,c), providing internucleosomal position signal. At a higher resolution the nucleosome-occupied DNA downstream of zygotic TSS displayed a 10 bp periodicity in AA and TT dinucleotide enrichment (Fig. 3d, Extended Data Fig. 3e), previously identified as intranuclesomal positioning signal[16]. The strong association of zygotic, but not maternal TSS, with these nucleosome positioning signals argues that TSS selection in a vertebrate oocyte is independent of inter- and intra-nucleosomal DNA signals.

Recent efforts to identify sequence-based signals for nucleosome positioning[17,18] and dynamic nucleosome organisation at promoters[19,20] highlight the epigenetic and chromatin mechanisms[21,22] that, together with DNA sequence, direct transcription initiation. The association of nucleosome positioning signals with zygotic promoter activity described here raises the question 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) stage (Fig. 4a, Extended Data Fig. 8). In pre-MBT stage, H3K4me3-marked nucleosomes occupied 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 prior to transcription[23]. In the post-MBT stage the +1 nucleosome was positioned just downstream of the SS|WW enrichment boundary at ~50bp from the zygotic TSS, occupying a WW-enriched region, with a small local GC/CG peak at the nucleosome midpoint (Fig. 4a, Extended Data Fig. 8b). Additional downstream nucleosomes followed a similar pattern of WW-enriched bound DNA alternating with internucleosomal SS enrichment. The local GC/CG enrichment at the nucleosome midpoints is in accordance with previously described nucleosome positioning preferences[16]; however, the additional sequence preference complexity and its relation to TSS in different developmental stages were not reported so far. The results show that initial positioning of promoter-associated nucleosomes, which correlates with a broad internucleosomal phasing pattern, changes in later stages to final precise positioning, which correlates tightly with zygotic transcription initiation site and intranucleosomal phasing patterns, suggesting interdependence of 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 precision of nucleosome positioning was found, arguing against transcriptionally aided nucleosome readjustment and instead suggesting a pre-transcriptional process in repositioning of nucleosomes to their final position, in agreement with transcription-independent positioning of nucleosomes at promoters in human cells[24]. Consistently, H3K4me3 ChIPseq in TBP knock-down 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 transcription initiation machinery or active transcription.

The absence of nucleosome-positioning sequence signature, as well as of precise nucleosome positioning at promoters with canonical TATA-box in other systems[20,25], 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 TATA-box suggests that transcription initiation in the oocyte may be mediated by the oocyte-enriched transcription nucleating factor TBP2[26,27]. Conversely, early zygotic grammar prefers TSS position 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 TFIID recruitment by H3K4me3-TAF3 interactions[28] and emphasizes the interdependence of nucleosome configuration at promoters with promoter type and physiological state in vertebrates[20,25] and yeast[19,29].

Different TSS selection grammars deployed at separate promoters have been associated with different types of genes[19,20] and a handful of promoters were shown to switch between TATA-dependent and independent initiation[30]. Here we show for the first time that the two grammars co-exist in close proximity or in physical overlap genome-wide and are differentially utilised 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

CAGE tags mapping and CTSS calling

Sequenced CAGE tags (27 bp) from Nepal *et al.*[13] were mapped to a reference zebrafish genome (Zv9/danRer7 assembly) using Bowtie[31] with default parameters allowing up to 2 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 CAGE protocol[32], was removed in cases where it did not map to the genome. All unique 5′ ends of tags were considered as CAGE tag-defined transcriptional start sites (CTSSs) and the number of tags supporting each CTSS was counted. Raw tag count was normalized to a referent power-law distribution based on total 10^6 tags and $\alpha=-1.25$ as described in Balwierz *et al.*[33] resulting in normalized tags per

million (tpm). All analyses were done in R statistical computing environment[34] (http://www.R-project.org/) using Bioconductor[35] (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, tag clusters (TCs) were created for each stage individually using simple distance-based approach with a maximum allowed distance of 20bp between two neighbouring CTSSs. Next, for each TC we calculated a cumulative distribution of CAGE signal and determined the positions of 10^{th} and 90^{th} percentile to obtain more robust boundaries of a TC. TCs across all developmental stages within 100bp 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 5tpm 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. Self-organizing map[36] (SOM) unsupervised learning algorithm was applied to distribute CTSSs (or promoter regions) across $5 \times 5 = 25$ expression profiles.

Dinucleotide patterns analysis

To visualize dinucleotide composition patterns of sequences flanking TSS we first created an occurrence matrix ($n \times m$; where n = number of sequences and m = length of sequences) for each individual dinucleotide, by placing 1 if the given dinucleotide is present at given position or 0 if it is not. Values in the matrix were then smoothened: 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 2D Gaussian kernel with bandwidth = 3 (in both dimensions) at the central position. Matrix of smoothened values (densities) was visualized using different shades of blue in a map-like representation. Extended Data Figure 10 illustrates how the calculation and visualization was done.

TATA-box motif analysis

TATA-box position weight matrix (PWM) was obtained from JASPAR database[37] (http://jaspar.genereg.net/) and used to scan the region −35 bp to −22 bp upstream of TSS (expected position for a TATA-box according to Ponjavic *et al.*[15]). For each promoter sequence a maximal detected match (%) to TATA-box PWM was reported for maternal and zygotic dominant TSS separately. Distribution of obtained values across all promoters was visualized by histograms. In addition, the frequency of the top 10 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 TSS using

MEME[38] (http://meme.sdsc.edu/) with default parameters. Only motifs with E-value ≤ 0.01 were selected as significant.

sf3a2 promoter reporter constructs

Region spanning 500 bp upstream and 200 bp downstream of the dominant zygotic TSS in the *sf3a2* promoter was chosen for validation of TSS selection grammar. *sf3a2* promoter carries both maternal and zygotic promoter determinants and exhibits TSS shifting. The selected sequence, which ends within the first intron of *sf3a2*, was fused to a sequence containing 3′ end of the zebrafish *txnipa* first intron and splice acceptor fused to a mCherry reporter (Extended Data Fig. 6a). Genomic DNA of AB* zebrafish strain was used for PCR amplification using Advantage HD DNA Polymerase Mix (Clontech). The amplified fragments were cloned into pDB896 vector (kindly provided by D. Balciunas, Temple University, Philadelphia, PA and subsequently modified by replacing the *γ-crystallin:gfp* with mCherry) using In-Fusion® PCR Cloning System (Clontech) following the manufacturer's instructions. The expression cassette is flanked by Tol2 transposon arms for Tol2 transgenesis[39]. A polymorphic nucleotide (G->T) was identified in the promoter sequence at chr2:58,656,711 (Ensembl, Zebrafish Assembly Zv9/danRer7). The wild-type promoter reporter vector was used as a template for in site-directed mutagenesis PCR to introduce the mutation in the W-boxes as indicated in Fig. 3d. The resulting mutated PCR fragment was cloned in the same reporter vector using the In-Fusion® PCR Cloning System. Sequences of all primers are provided in Extended Data Figure 6b.

Microinjection and transgene expression analysis

1.5-2 nl of injection solution containing 20 ng/μl reporter plasmid DNA and 15 ng/μl Tol2 transposase mRNA, supplemented with 0.1% Phenol red (injection marker), was injected into zebrafish eggs within 10-15 min after fertilization. In the automated imaging and expression analysis experiments *ecfp* mRNA was added (30 ng/μl). The mCherry reporter activity was measured at prim 20 stage, in both wild-type and mutated *sf3a2* promoter construct injected embryos by automated imaging as described[40]. Embryo images were analyzed with Zebrafish Miner software[40]. The level of reporter expression was measured as pixel intensity value and normalized to the intensity of the ECFP signal (injection control) and averaged for all embryos in the experiment. In addition a percentage of expressing embryos was calculated from the total number of ECFP positive embryos with ECFP signal equal or above the detection threshold of the Zebrafish Miner software. Embryo images were wrapped onto reference embryo shape and overlaid by summing pixel intensity values.

5′ RACE

The 5′ RACE was performed with FirstChoice® RLM-RACE Kit (Life Technologies) following the manufacturer's protocol. Total RNA was isolated at prim6 stage from ~100 phenotypically normal looking and reporter gene expressing zebrafish embryos injected with either wild type or mutated *sf3a2* promoter reporter construct, using TRIZOL (Life Technologies) 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 TAP" (Tobacco Acid Pyrophosphatase) treated sample

was carried through adapter ligation, reverse transcription and PCR. Sequences of primers used in 5′ RACE are provided in Extended Data Figure 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 above. 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 presence of reporter expressing offspring. Transgene expression and TSS usage was analysed in F1 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 at 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 (normalisation control). Technical triplicates were run for each primer pair. The Ct values were determined by the SDSv2.4 software (Applied Biosystems), using 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 Ct values across technical triplicates and both primer pairs. The sequence of the primers used in qPCR is provided in Extended Data Figure 7c.

SL-CAGE

We have introduced a novel method for quantitative high-resolution detection of TSSs and their usage within a targeted promoter, called Single Locus deep CAGE. The method combines generation of 5′ complete cDNAs transcribed from capped mRNAs as described in the CAGE protocol[7] with the amplification of targeted cDNAs using gene-specific primers and subsequent high throughput paired-end sequencing of the single locus (typically single promoter region) based library. Supplementary Table 2 describes main steps of the protocol and provides sequences of all primers used in different steps. Paired-end sequenced reads (34 bp + 35 bp) were mapped to either spliced sequence of *sf3a2:mCherry* transgene or endogenous *sf3a2* and TSS usage was reconstructed as described above for CAGE tags.

Chromatin immunoprecipitation (ChIP)

ChIP experiments were carried out using the ChIP-IT Express Enzymatic kit (Active Motif) in line with the manufacturer's instructions. Chromatin was prepared using ~5000 and ~3000 embryos for 512 cell and oblong stage, respectively. Embryos were dechorionated enzymatically using pronase and fixed in 1.85% formaldehyde in Hanks Media for 20 min at room temperature. The embryos were washed once with PBS and the fixation was stopped by incubating in 1x Glycine for 10 min at room temperature followed by 3 washes with ice-cold PBS. Embryos were resuspended in 1 ml ice-cold lysis buffer, incubated on ice for 20 min, transferred to a pre-cooled dounce homogenizer and dounced by 10 strokes. Nuclei

were collected by centrifugation, resuspended in 200 µl digestion buffer and incubated at 37°C for 5 min. Chromatin was sheared by adding 10 µl of enzymatic shearing cocktail working stock (200 U/ml) and incubating for 10 min at 37°C. Shearing efficiency was checked by gel electrophoresis according to manufacturer's instructions. The reaction was stopped by adding 5 µl ice-cold 0.5M EDTA and incubating on ice for 10 min and sheared chromatin was cleared by centrifugation. For ChIP reactions 70 µl of sheared chromatin were mixed with 25 µl Protein G magnetic beads, 20 µl ChIP buffer 1, 1µl 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 200µl. 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 4h at 65°C. Samples were Proteinase K and RNase A treated and purified using pheonol chloroform extraction.

TBP knock-down

One-cell stage embryos were injected with either *β-actin:yfp*[41] (1.7nl, 42pg/nl) or 1.3*ntl:yfp*[42] (1.7nl, 68pg/nl) constructs. The injected embryos were then split into four groups. One was kept as non-injected control and the other three groups were further injected with one of the two *tbp*-targeting morpholinos (1.7nl, 2.5mM) or with a mismatch morpholino described in Ferg *et al.*[43]. All embryos were kept in E3 medium at 28.5°C until non-injected group reached 30% epiboly stage (4.7 hpf) and were then analyzed under fluorescence stereoscope (Nikon SMZ1500). Arrest of epiboly movements, loss of *β-actin:yfp* (TBP dependent) and retention of *ntl:yfp* (TBP independent) reporter activities[42] were used as marker for assessing and sorting TBP morphants for ChIP analysis. Approximately 1500 non-injected, 1500 mismatch morpholino injected embryos, 1200 *tbp mo1* morphants and 1000 *tbp mo2* morphants were used for ChIP as described above. We used a previously published set of genes downregulated in zebrafish TBP morphants[43]. Genes with log fold change ≤-1.5 were selected as TBP-dependent and were aligned with respect to dominant TSS of the nearest promoter detected by CAGE for H3K4me3 ChIP-seq signal visualisation.

Deep sequencing of chromatin DNA

ChIP-seq was performed as described before[44]. 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 Bowtie[31] with default parameters allowing up to 2 mismatches and keeping only uniquely mapped reads. Coverage was calculated for plus and minus strands separately using unextended reads and taking max. 20 reads mapping to the exactly same position. Minus strand coverage was subtracted from plus strand coverage to obtain subtracted

coverage, which was used for visualisation and nucleosome midpoint estimation. Significantly enriched regions (peaks) were detected using MACS[45] (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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors are grateful to Laszlo Tora, Emma Kenyon, Guilhem Chalancon and Julie Chih-yu Chen for comments on the manuscript, and to Laura O'Neill, for technical advice. V.H., C.N., C.P., A.A. and X.D. were supported by grants from Norwegian Research Council (YFF) and Bergen Research Foundation awarded to B.L. F.M., U.S. and B.L. acknowledge support from EU FP6 integrated project EuTRACC and FP7 integrated project ZF Health. B.L. was additionally supported by Medical Research Council UK, F.M. and P.C. by EU FP7 project Dopaminet and U.S. by EU FP6 project NeuroXSys. C.Pl, A.M.S and P.C. were supported by a Research Grant from MEXT to RIKEN CLST.

References

1. D'Alessio JA, Wright KJ, Tjian R. Shifting Players and Paradigms in Cell-Specific Transcription. Molecular Cell. 2009; 36:924–931. [PubMed: 20064459]
2. Kadonaga JT. Perspectives on the RNA polymerase II core promoter. Wiley Interdiscip Rev Dev Biol. 2011; 1:40–51. [PubMed: 23801666]
3. Mathavan S, Lee S, Mak A, et al. Transcriptome Analysis of Zebrafish Embryogenesis Using Microarrays. PLoS Genet. 2005; 1:e29.
4. Tadros W, Lipshitz HD. The maternal-to-zygotic transition: a play in two acts. Development. 2009; 136:3033–3042. [PubMed: 19700615]
5. Vastenhouw NL, et al. Chromatin signature of embryonic pluripotency is established during genome activation. Nature. 2010; 464:922–926. [PubMed: 20336069]
6. Lindeman LC, et al. Prepatterning of Developmental Gene Expression by Modified Histones before Zygotic Genome Activation. Developmental Cell. 2011; 21:993–1004. [PubMed: 22137762]
7. Kane DA, Kimmel CB. The zebrafish midblastula transition. Development. 1993; 119:447–456. [PubMed: 8287796]
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. U.S.A. 2003; 100:15776–15781. [PubMed: 14663149]
9. Barski A, et al. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell. 2007; 129:823–837. [PubMed: 17512414]
10. Carninci P, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nature Genetics. 2006; 38:626–635. [PubMed: 16645617]
11. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature Reviews Genetics. 2012; 13:233–245.
12. Venters BJ, Pugh BF. Genomic organization of human transcription initiation complexes. Nature. 2013; 502:53–58. [PubMed: 24048476]
13. Nepal C, et al. Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. Genome Research. 2013; 23:1938–1950. [PubMed: 24002785]
14. Giraldez AJ, et al. Zebrafish MiR-430 Promotes Deadenylation and Clearance of Maternal mRNAs. Science. 2006; 312:75–79. [PubMed: 16484454]

15. Ponjavic J, et al. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. Genome Biology. 2006; 7:R78. [PubMed: 16916456]
16. Segal E, et al. A genomic code for nucleosome positioning. Nature. 2006; 442:772–778. [PubMed: 16862119]
17. Ioshikhes I, Hosid S, Pugh BF. Variety of genomic DNA patterns for nucleosome positioning. Genome Research. 2011; 21:1863–1871. [PubMed: 21750105]
18. Segal E, Widom J. What controls nucleosome positions? Trends in Genetics. 2009; 25:335–343. [PubMed: 19596482]
19. Rhee HS, Pugh BF. Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature. 2012; 483:295–301. [PubMed: 22258509]
20. Rach EA, et al. Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level. PLoS Genet. 2011; 7:e1001274. [PubMed: 21249180]
21. Cairns BR. The logic of chromatin architecture and remodelling at promoters. Nature. 2009; 461:193–198. [PubMed: 19741699]
22. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes & Development. 2011; 25:1010–1022. [PubMed: 21576262]
23. Thomson JP, et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature. 2010; 464:1082–1086. [PubMed: 20393567]
24. Fenouil R, et al. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Research. 2012; 22:2399–2408. [PubMed: 23100115]
25. Nozaki T, et al. Tight associations between transcription promoter type and epigenetic variation in histone positioning and modification. BMC Genomics. 2011; 12:416. [PubMed: 21846408]
26. Bártfai R, et al. TBP2, a Vertebrate-Specific Member of the TBP Family, Is Required in Embryonic Development of Zebrafish. Current Biology. 2004; 14:593–598. [PubMed: 15062100]
27. Akhtar W, Veenstra G. TBP2 is a substitute for TBP in Xenopus oocyte transcription. BMC Biol. 2009; 7:45. [PubMed: 19650908]
28. Lauberth SM, et al. H3K4me3 Interactions with TAF3 Regulate Preinitiation Complex Assembly and Selective Gene Activation. Cell. 2013; 152:1021–1036. [PubMed: 23452851]
29. Zaugg JB, Luscombe NM. A genomic model of condition-specific nucleosome behavior explains transcriptional activity in yeast. Genome Research. 2012; 22:84–94. [PubMed: 21930892]
30. Davis W Jr. Schultz RM. Developmental Change in TATA-Box Utilization during Preimplantation Mouse Development. Developmental Biology. 2000; 218:275–283. [PubMed: 10656769]
31. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009; 10:R25. [PubMed: 19261174]
32. Kodzius R, et al. CAGE: cap analysis of gene expression. Nature Methods. 2006; 3:211–222. [PubMed: 16489339]
33. Balwierz PJ, et al. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biology. 2009; 10:R79. [PubMed: 19624849]
34. The R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2013:1–3079.http://www.R-project.org/
35. Gentleman RC, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology. 2004; 5:R80. [PubMed: 15461798]
36. Toronen P, Kolehmainen M, Wong G, Castren E. Analysis of gene expression data using self-organizing maps. FEBS Letters. 1999; 451:142–146. [PubMed: 10371154]
37. Portales-Casamar E, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Research. 2009; 38:D105–D110. [PubMed: 19906716]
38. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994; 2:28–36. [PubMed: 7584402]
39. Abe, G.; Suster, ML.; Kawakami, K.; Detrich, H William; W., M.; Zon, LI. The Zebrafish: Genetics, Genomics and Informatics. Vol. 104. Academic Press; 2011. p. 23-49.

40. Gehrig J, et al. Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. Nature Methods. 2009; 6:911–916. [PubMed: 19898487]
41. Higashijima S-I, Okamoto H, Ueno N, Hotta Y, Eguchi G. High-Frequency Generation of Transgenic Zebrafish Which Reliably Express GFP in Whole Muscles or the Whole Body by Using Promoters of Zebrafish Origin. Developmental Biology. 1997; 192:289–299. [PubMed: 9441668]
42. Ferg, M. PhD dissertation. Heidelberg University; 2008. Large scale- and functional analysis for the requirement of TBP-function in early zebrafish development.
43. Ferg M, et al. The TATA-binding protein regulates maternal mRNA degradation and differential zygotic transcription in zebrafish. EMBO J. 2007; 26:3945–3956. [PubMed: 17703193]
44. Soler E, et al. A systems approach to analyze transcription factors in mammalian cells. Methods. 2011; 53:151–162. [PubMed: 20705139]
45. Zhang Y, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008; 9:R137. [PubMed: 18798982]

Europe PMC Funders Author Manuscripts

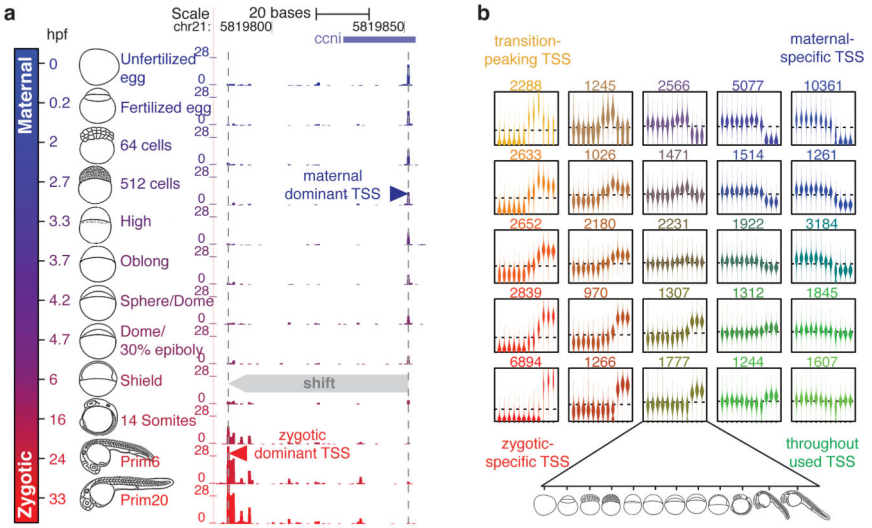

Figure 1. Dynamics of transcription initiation at 1bp resolution throughout zebrafish early embryonic development

a, CAGE signal at “shifting” promoter of cyclin 1 (*ccni*) gene. Colouring from blue to red reflects maternal to zygotic transition. Corresponding zebrafish developmental stages are depicted on the left, with timescale denoting hours past fertilization (hpf). **b**, Expression profiles obtained by self-organizing map (SOM) clustering of individual CAGE transcription start sites (CTSS). Each box represents one cluster, with beanplots showing distribution of relative expression at different time points for all CTSSs belonging to that cluster (number above the box). The developmental stages at x-axis in all boxes are shown at the bottom.

Europe PMC Funders Author Manuscripts

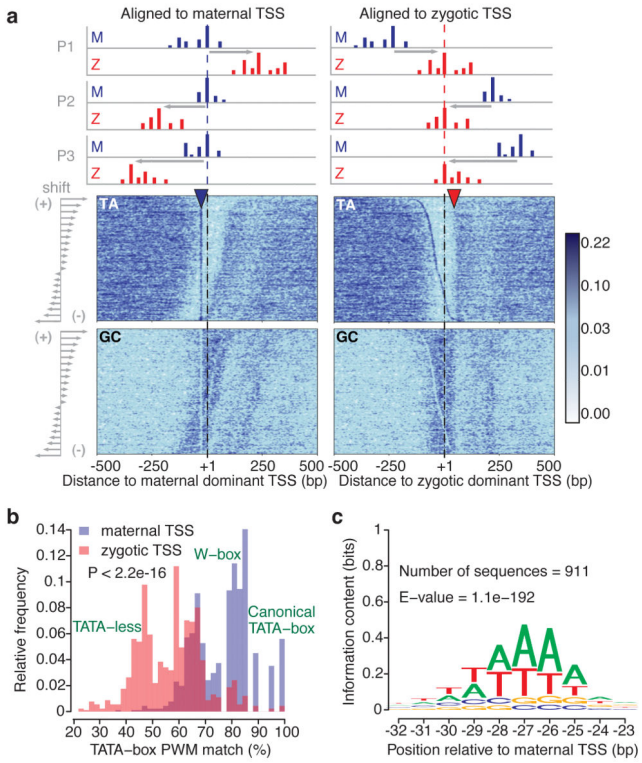

Figure 2. Sequence signature of a large set of "shifting" promoters changes dramatically during maternal to zygotic transition

a, Dinucleotide density (see Extended Data Fig. 10) at 911 "shifting" promoters sorted and aligned according to the distance and orientation of the TSS shift (schematics on the top; P1, P2, P3 – individual promoters; M – maternal stage; Z – zygotic stage). Promoters were centred at either maternal (left) or zygotic (right) dominant TSS. Blue arrowhead: TA enrichment at the expected position of the TATA-box; red arrowhead: boundary between GC and TA enrichment ~50bp downstream of zygotic TSS. **b**, Distribution of match (%) to TATA-box in the region −35 to −22 bp upstream of maternal (blue) and zygotic (red) dominant TSS (P-value - two-tailed Wilcoxon rank-sum test). **c**, Motif obtained by motif discovery upstream of maternal dominant TSS.

Europe PMC Funders Author Manuscripts

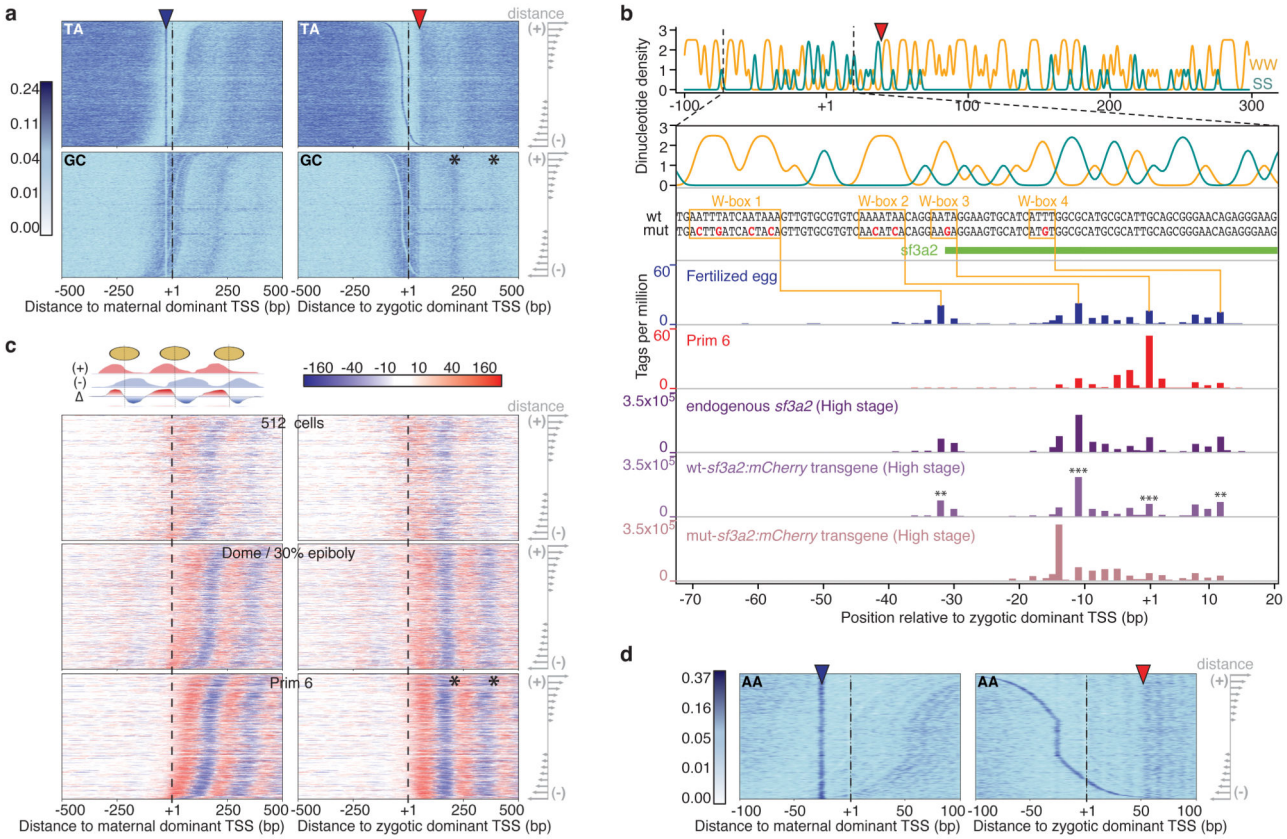

Figure 3. Transition from maternal W-box motif-dependent, to zygotic nucleosome positioning signal-related transcription initiation is pervasive

a, Dinucleotide density at 8369 constitutively expressed promoters sorted by the distance between maternal and zygotic dominant TSS. Promoters were centred at either maternal (left) or zygotic (right) dominant TSS. Blue arrowhead: position of maternal TSS-associated W-box; red arrowhead: SS|WW boundary ~50bp downstream of zygotic TSS; asterisks: GC enrichment in the internucleosomal region. **b**, Predicted maternal and zygotic codes in *sf3a2* promoter. Dinucleotide density and sequence of the wild-type (wt) and mutated (mut) *sf3a2* promoter is shown on top. TSSs detected by CAGE in wild type zebrafish in maternal and zygotic stage are shown in blue and red, respectively. The W-boxes associated with maternal TSSs are marked in orange, and the introduced point mutations disrupting them in red. Single locus CAGE TSSs in stable transgenic lines for endogenous *sf3a2*, 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, $n_{mut} = 4$, $n_{wt} = 3$). **c**, Subtracted H3K4me3 coverage (Δ) of reads mapping to (+) and (−) strand (schematic on top) in three developmental stages at the same set of promoters from panel a. **d**, Density of AA dinucleotide in +/− 100 bp region for promoters from panel a.

Europe PMC Funders Author Manuscripts

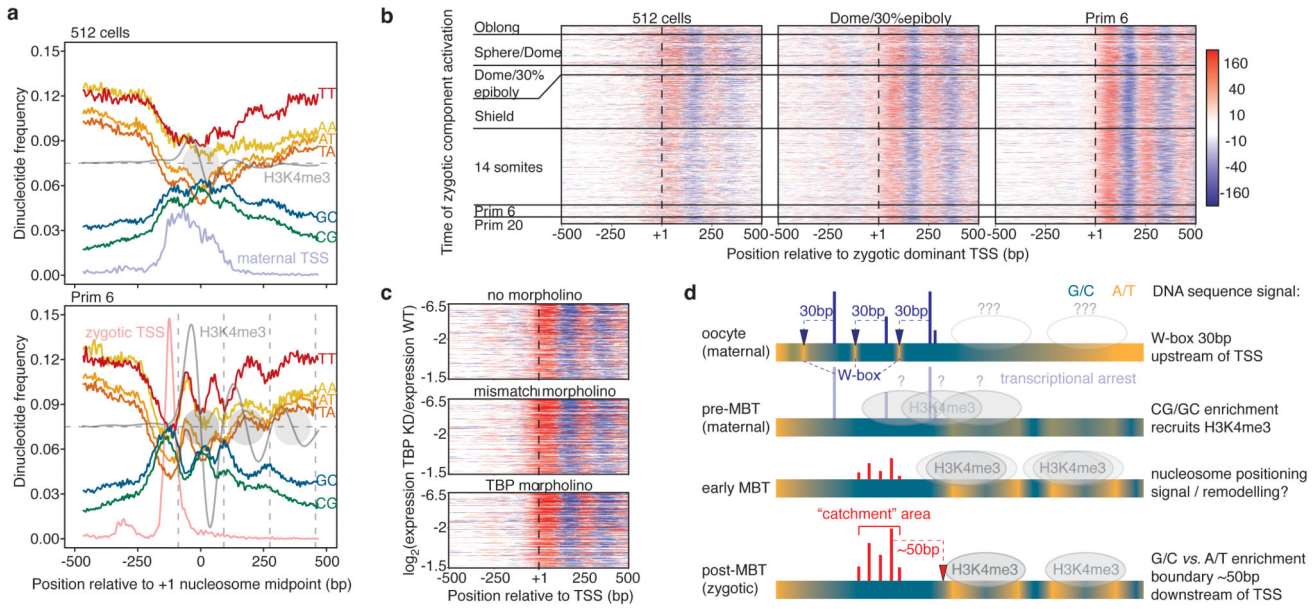

Figure 4. H3K4me3-marked nucleosome positioning reveals dynamic changes in underlying sequence signature and relation to TSS during MZT

a, Frequency of dinucleotides centred on +1 nucleosome of constitutively active promoters in maternal (512 cells) and zygotic (prim 6) stage. Centres of nucleosomes were estimated from subtracted H3K4me3 coverage (gray). Density of maternal and zygotic transcription start sites 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 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 TBP expression fold-change between knockdown and wild type embryos. **d**, Summary of transcription initiation, TSS configuration and nucleosome positioning dynamics throughout MZT.