THE EMBO JOURNAL

Embryonic heterogeneity arises by combinatorial action of transcription factors in open chromatin

Ann Rose Bright, Siebe van Genesen, Qingqing Li, Alexia Grasso, Siebren Frölich, Maarten van der Sande, Simon van Heeringen, and Gert Jan C. Veenstra
DOI: 10.15252/embj.2020104913

Corresponding author(s): Gert Jan C. Veenstra (g.veenstra@science.ru.nl)

Review Timeline:

Submission Date:	5th Mar 20
Editorial Decision:	13th May 20
Revision Received:	13th Nov 20
Editorial Decision:	10th Dec 20
Revision Received:	21st Dec 20
Accepted:	24th Dec 20

Editor: Stefanie Boehm

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. Depending on transfer agreements, referee reports obtained elsewhere may or may not be included in this compilation. Referee reports are anonymous unless the Referee chooses to sign their reports.)

Thank you for submitting your manuscript assessing open chromatin and transcriptomic changes during Xenopus embryogenesis for consideration by The EMBO Journal. Please excuse the delay in communicating this decision to you, which was due to referees not being able to review as planned also on account of the current SARS-CoV-2 pandemic and us therefore having to consult additional experts. We have now however received two reports on your study, which are included below for your information. In addition, we also consulted an external expert on the reports. Taking everything into consideration, we have decided to invite a revision of the study.

As you will see, the referees appreciate the potential value the results and data have to the field. However, they also raise several concerns which would need to be addressed in a revised version of the manuscript. Referee #3's concerns can likely be resolved by textual changes and adding further explanations regarding the experimental setup to the manuscript. In addition, it will be crucial that all data-sets are available to the reader, also including all points the referee indicates. Referee #4's main concern is that the conclusions drawn from bulk ATACseq should be validated by single cell ATACseq or analysis of specific regions. We agree that such validation, potentially based on at least some scATACseq, would further improve the study. The referee's specific comments 1- 4 should also be addressed and the manuscript revised accordingly. In addition, please carefully respond to any other points raised by either of the referees.

We recognize that addressing the referees' comments fully could require a significant amount of experimental work and time and also realize that an experimental revision may currently be impaired by the COVID-19/SARS-CoV-2 pandemic affecting labs worldwide. We can extend the revision time where needed and we have extended our 'scooping protection policy' to cover the period required for a full revision. However, it is nonetheless important to clarify any questions and concerns at this stage, and we strongly encourage you to contact us to discuss a revision plan and any potential issues may foresee for the revision.

Referee #3:

This paper integrates chromatin accessibility and gene expression datasets, including single cell analysis, across early development in Xenopus tropicalis. The authors discover that lineage-restricted genes, in this case those expressed in the organizer, are associated with accessible chromatin in the organizer, while those that are not expressed are not associated with accessible chromatin in the organizer, as might be expected. However organizer genes are also associated with accessible chromatin in animal cap cells, where they are not expressed, which raises the question of what activates their expression in the organiser region but not the animal cap, and how chromatin accessibility is regulated in these different regions. Using motif analysis on accessible regions around genes that are associated with specific cell groups in the embryo, the authors identify several candidate transcription factors that may be activating organiser genes, including Eomes and Foxb1, and validate experimentally that they act alone or in combination to up-regulate gene expression. In addition, the authors find that most of these transcription factors, which are zygotically expressed, act at promoters marked by maternal histone modifications.

This is an interesting and useful paper that probes how (combinatorial) transcription factor activity acting at open chromatin can regulate gene expression in a specific cell types. The results obtained and data within the manuscript also pave the way for more studies that will give insights into how genes are differentially regulated in different regions of the early embryo, leading to cell fate specification.

My main concern as a reader is that large amounts of the data that are discussed are not provided with the manuscript, meaning that readers will not be able to interrogate the data properly. This manuscript should be published with minor corrections (below) and the proviso that these missing data are made available.

Minor corrections:

Figure 1A: What do the numbers on the left of the peak plots refer to, please indicate in the legend? A scale bar should be included to allow the reader to judge the extent of the genomic region shown. This applies to all figures visualizing peak data. In addition, is the black line the sox2 gene, or is the gene located at the position of the sox2 label? This should be made clearer.
Pg 4: "... we analyzed the transcript levels of nearby genes". Make it clear how these nearby genes are annotated - it seems from the methods that this an intersect with GREAT analysis and each peak is linked to the nearest gene. Is this the case? If so, perhaps "...we analyzed the transcript levels of the nearest gene for each accessible chromatin region" would be better? In addition it is not clear what RNA-seq data were used for this analysis and should be made clear.
Figure 2D: This is not clearly explained. How were regulatory regions for pluripotent genes defined? Is one line of the heat map one regulatory region? Within what distance of a gene was a regulatory region deemed to belong to that gene? What other pluripotent genes are represented in the heatmap but not labelled? A table of this information would be useful. Were the same number and sizes of random regions chosen? In short, how was this analysis done?
Figure 5: Analysis was done to associate regulatory elements with hypervariable genes. Again, how was a regulatory elements linked to a gene? Can a gene have more than one regulatory element?

Motif analysis (Figure S2 and Figure 5): The motif analysis is not well explained and points the reader to a preprint (Bruse and Heeringen, 2018, Biorxiv) which does not yet appear to have been peer-reviewed, and which needs to be read in some detail to pick out the information. For the benefit of the reader of this submitted manuscript, the method should be summarised here as well, and specifically the concept of 'motif activity' explained in more detail as this is not at all clear. For instance, the workflow in Figure 5 is not well explained and because of this, it was difficult for this reviewer to interpret the data shown. In particular negative motif activity is not explained or commented on.

Missing data:

The current Supplementary Tables (ST1 and ST2) should include the a gene ID as well as a gene symbol, to allow genes to be more easily found if names become defunct (e.g. LOC100493862-like in cluster 0 (ST1) does not appear in a search of Xenbase, JGI, UCSC or NCBI).

A supplementary table showing metrics (e.g. total reads, % reads that map, etc...) for the all the -seq libraries generated in the study would be of benefit to the reader, so the quality the data can more easily be judged.

Tabulated data for the results presented in Figs 1 and 2 should be given, e.g a table of the ATAC-seq peaks with the nearest gene (if this is what was annotated) and distance to that gene, a table of intersected ATAC-seq (whole embryo and AC/DMZ), H3K4me3 and p300 peaks, a table of differential genes and linked ATAC-seq peak.

Tabulated data for the results presented in Fig 3 for the whole embryo should also be given, i.e. not just for the AC and DMZ.
Tabulated data for the results presented in Fig 5 should be given, e.g. a table of ATAC-seq peaks associated with the hypervariable genes, a table of regulated genes associated with MaD or ZyD promoters.

Referee #4:

The manuscript entitled "Combinatorial action of transcription factors in open chromatin contributes to early cellular heterogeneity and organizer mesendoderm specification" by Brihgt et al. reports transcriptome and open-chromatin profiles of Xenopus embryo during peri-gastrula stage by scRNA-seq and ATAC-seq. The authors analyzed the chromatin accessibility of the animal cap (AC) where the pluripotent ectoderm occurs and the dorsal marginal zone (DMZ) where the cells with organizer activity occur, and found that the pluripotent cells in AC had a more open chromatin state than the lineage-restricted cells in (DMZ), a finding consistent with the case in mammalian embryonic stem cells (ESCs) and the differentiated cells. The authors performed an integrated analysis of the scRNA-seq and ATAC-seq data, and found several transcription factor (TF)-binding motifs in open chromatin regions, and the functions of key TFs were evaluated by experiments injecting mRNAs of combinations of these TFs into frog embryos.
Overall, the authors' statements are clear and the data presented appear reasonable. However, the computational analysis seems a bit poor and the subsequent experimental analyses may not be vigourous and significant enough to be considered for publication in the EMBO journal. In particular, the usage of the ATAC-seq data of whole embryos rather than those of specific regions for the integrated analysis with scRNA-seq would not be appropriate. The conclusion of the authors may

change if they used the region-specific ATAC-seq or even scATAC-seq data. If it were not possible to use the data of AC and DMZ with some reasons, the authors should include and discuss the data of epigenetic profiles such as H3K4me3 and H3K27me3, which were examined in Figure1 and S1, from a global view point.

I would also like to suggest the authors to consider a better representation of the data and their explanation in several parts in the manuscript. It was not easy to understand the figures and the way of their analyses.

1. The authors should present the statistics data regarding the quality of all sequencing data, especially for scRNA-seq.
2. In the Figure 1C, the authors mentioned that the genes associated with lower levels of ATAC-seq signal were upregulated. For RNA-seq data, which cell types did the authors examined?
3. In the result section, the authors explained that the visualization of scRNA-seq was performed using force-directed layout and the clustering using Louvain method. However, in the method section, it was stated that they were done using t-SNE and "FindClust" function in a package of R. The method for the visualization and clustering for Figure 3 did not seem to be appropriate.
4. Related to the point #2, it would be more informative to see the transcription levels of genes by cluster-wise manner in Figure 1B.
5. The title stated that "Combinatorial action of TFs in open chromatin contributes to early cellular heterogeneity", but I could not understand which data support the statement that the open chromatin contributed to cellular heterogeneity. The authors should clarify this point.

Response to reviewers

November 13, 2020 – Bright et al. EMBO J – Revision

We thank the reviewers and the editor for their constructive comments and support. We are happy to address the concerns. Below the reviewer comments are presented in regular font. Our response is indicated in **bold** after "> ".

Referee #3:

This paper integrates chromatin accessibility and gene expression datasets, including single cell analysis, across early development in Xenopus tropicalis. The authors discover that lineage-restricted genes, in this case those expressed in the organizer, are associated with accessible chromatin in the organizer, while those that are not expressed are not associated with accessible chromatin in the organizer, as might be expected. However organizer genes are also associated with accessible chromatin in animal cap cells, where they are not expressed, which raises the question of what activates their expression in the organiser region but not the animal cap, and how chromatin accessibility is regulated in these different regions. Using motif analysis on accessible regions around genes that are associated with specific cell groups in the embryo, the authors identify several candidate transcription factors that may be activating organiser genes, including Eomes and Foxb1, and validate experimentally that they act alone or in combination to up-regulate gene expression. In addition, the authors find that most of these transcription factors, which are zygotically expressed, act at promoters marked by maternal histone modifications.

This is an interesting and useful paper that probes how (combinatorial) transcription factor activity acting at open chromatin can regulate gene expression in a specific cell types. The results obtained and data within the manuscript also pave the way for more studies that will give insights into how genes are differentially regulated in different regions of the early embryo, leading to cell fate specification.

My main concern as a reader is that large amounts of the data that are discussed are not provided with the manuscript, meaning that readers will not be able to interrogate the data properly. This manuscript should be published with minor corrections (below) and the proviso that these missing data are made available.

> We thank the reviewer for the interest in our work. The NCBI GEO accession associated with our manuscript (link and reviewer token are in the manuscript) contains all raw sequencing data (FASTQ files). In addition, this GEO accession includes peak files (bed) for all ATAC samples, count tables for both the single cell and bulk RNA-seq data, and the fragments file of the scATAC-seq data. These processed data can be used directly for custom analyses. To facilitate exploration of the data without a need for deep analysis, we

are now also providing additional supplementary tables: ATAC-seq peak regions (ST1), scATAC-seq marker genes (ST2), Hypervariable genes, mean cluster expression, reanalysis scRNA-seq data St.8-12 (ST3), Count tables AC-DMZ scRNA-seq (ST4), Marker genes AC-DMZ clusters C0-C6 (ST5), Differentially expressed genes in AC upon factor overexpression (ST6), and Sequencing and mapping statistics (ST7).

Minor corrections:

Figure 1A: What do the numbers on the left of the peak plots refer to, please indicate in the legend? A scale bar should be included to allow the reader to judge the extent of the genomic region shown. This applies to all figures visualizing peak data. In addition, is the black line the sox2 gene, or is the gene located at the position of the sox2 label? This should be made clearer.

> We have made the appropriate changes to the figures and the legend. A proper scale bar has been added to all figures related to visualization of peak profiles. The number at left indicates the scale used to plot each track profile and the black line is indeed *sox2* gene. We have added necessary text in the figure legend to clarify this.

Pg 4: "... we analyzed the transcript levels of nearby genes". Make it clear how these nearby genes are annotated - it seems from the methods that this an intersect with GREAT analysis and each peak is linked to the nearest gene. Is this the case? If so, perhaps "...we analyzed the transcript levels of the nearest gene for each accessible chromatin region" would be better? In addition it is not clear what RNA-seq data were used for this analysis and should be made clear.

> We thank the reviewer for the textual suggestion. We indeed used GREAT regions (McLean et al. 2010, PMID 20436461), which represents a better way to associate genomic locations to genes than the closest gene. Please note that the GREAT regions of genes can overlap, so in some cases a peak overlaps with more than one GREAT region. We have made the necessary changes to the text (Page 4). We have used the RNA-seq data generated by the Gilchrist lab (Owens et al., 2016) for this analysis (reference provided in the manuscript).

Figure 2D: This is not clearly explained. How were regulatory regions for pluripotent genes defined? Is one line of the heat map one regulatory region? Within what distance of a gene was a regulatory region deemed to belong to that gene? What other pluripotent genes are represented in the heatmap but not labelled? A table of this information would be useful. Were the same number and size of random regions chosen? In short, how was this analysis done?

> We apologize for not explaining the analysis well. Shown are the ATAC signals of all the accessible regions overlapping with the GREAT regions of the four *ventx1 / ventx2* genes

(homologues of Nanog), the three *pou5f3* genes and *sox2*. These accessible regions overlap with p300 bound regions and thus we categorize them as enhancers. Every row here represents a regulatory region. In order to keep all the peaks the same size we used peaks summits, and extended them by 100 bp on both sides. We created random regions also with similar size in order to be compatible with our peaks sets. We have added necessary information to the legend.

Figure 5: Analysis was done to associate regulatory elements with hypervariable genes. Again, how were regulatory elements linked to a gene? Can a gene have more than one regulatory element?

> We have clarified this in the text. We used GREAT regions (see comment above). Yes, most developmental genes have more than one accessible region (promoters and enhancers).

Motif analysis (Figure S2 and Figure 5): The motif analysis is not well explained and points the reader to a preprint (Bruse and Heeringen, 2018, Biorxiv) which does not yet appear to have been peer-reviewed, and which needs to be read in some detail to pick out the information. For the benefit of the reader of this submitted manuscript, the method should be summarised here as well, and specifically the concept of 'motif activity' explained in more detail as this is not at all clear. For instance, the workflow in Figure 5 is not well explained and because of this, it was difficult for this reviewer to interpret the data shown. In particular negative motif activity is not explained or commented on.

> We are very sorry that this was unclear; we did not properly explain the relatively sophisticated motif analysis that we performed. We have now adjusted the text to explain better.

> We have made some modifications to this analysis and have now also added new panels to explain the method properly. The analysis is based on regressing peak signals (in Fig. 5A, S8A) or transcript signals (in Fig. 5B-C, ST8C-D) on the presence of motifs. This is not a standard motif analysis, but not conceptually novel (cf. the following references in the paper: Balwierz et al., 2014; Madsen et al., 2018; Suzuki et al., 2009). In the case of regressing peak signals, this means we use the motifs in underlying DNA sequence to predict (explain the variance of) peak strength. In the case of regressing transcript signals, we use motifs in the accessible regions of the GREAT regions of genes, to predict (explain the variance of) the transcript signals. A negative apparent 'motif activity' means that the presence of the motif anti-correlates with the signal to which the regression is performed. Simply put, a positive motif activity may be explained by an activator binding to the motif, while a negative motif activity may be explained by a repressor binding to the motif. This

method works surprisingly well. Of course these are inferences that need validation. Many of the inferred transcription factor activities confirm roles of these factors in early development that have been documented in the literature, testifying to the usefulness of the approach. In addition, we perform follow-up experiments to test the activities of factors for which much less was known (Fig. 6).

Missing data:

The current Supplementary Tables (ST1 and ST2) should include the a gene ID as well as a gene symbol, to allow genes to be more easily found if names become defunct (e.g. LOC100493862-like in cluster 0 (ST1) does not appear in a search of Xenbase, JGI, UCSC or NCBI).

> We share the reviewer's frustration about changing and disappearing gene identifiers for this model organism. Generally this involves genes without proper gene names or gene symbols (for example LOC#) which also tend to be absent in external data bases such as Ensembl. Where possible, we have added both Ensembl gene ID and gene symbols to the supplementary files. They are labelled in the current version as: Supplementary table ST5 and Supplementary table ST6.

A supplementary table showing metrics (e.g. total reads, % reads that map, etc...) for the all the -seq libraries generated in the study would be of benefit to the reader, so the quality of the data can more easily be judged.

> We have included an additional file (Supplementary Table ST7) that shows the quality metrics of the sequenced data.

Tabulated data for the results presented in Figs 1 and 2 should be given, e.g a table of the ATAC-seq peaks with the nearest gene (if this is what was annotated) and distance to that gene, a table of intersected ATAC-seq (whole embryo and AC/DMZ), H3K4me3 and p300 peaks, a table of differential gene and linked ATAC-seq peaks.

> See comment above. We have added a new supplemental file (Supplementary Table ST1) that includes the different ATAC-seq peak sets (whole embryo and AC/DMZ), in addition to the files uploaded to NCBI GEO. We have added the tabular data for Figs. 1B, including the p300 signals, in Supplemental Table ST1. The files are very big, and adding the comparable files from Figs. S1B-C crashed Excel. Direct intersections are problematic for this purpose, as it requires arbitrary cutoffs regarding noise versus signal. For casual exploration of the data we would advise to upload the peak sets in the UCSC genome browser.

Tabulated data for the results presented in Fig 3 for the whole embryo should also be given, i.e. not just for the AC and DMZ.
Tabulated data for the results presented in Fig 5 should be given, e.g. a table of ATAC-seq peaks associated with the hypervariable genes, a table of regulated genes associated with MaD or ZyD promoters.

> We are not entirely sure what the reviewer means. The figure cannot be represented as tabular data, at least not in a way that would be useful for exploration or custom analyses. However we have now provided a supplemental file (Supplementary Table ST3) with the the hypervariable genes, all genes, and their average expression in cell clusters.
We have now added additional figure panels and also made textual changes regarding the analysis of Figure 5.

Referee #4:

The manuscript entitled "Combinatorial action of transcription factors in open chromatin contributes to early cellular heterogeneity and organizer mesendoderm specification" by Brihgt et al. reports transcriptome and open-chromatin profiles of Xenopus embryo during peri-gastrula stage by scRNA-seq and ATAC-seq. The authors analyzed the chromatin accessibility of the animal cap (AC) where the pluripotent ectoderm occurs and the dorsal marginal zone (DMZ) where the cells with organizer activity occur, and found that the pluripotent cells in AC had a more open chromatin state than the lineage-restricted cells in (DMZ), a finding consistent with the case in mammalian embryonic stem cells (ESCs) and the differentiated cells. The authors performed an integrated analysis of the scRNA-seq and ATAC-seq data, and found several transcription factor (TF)-binding motifs in open chromatin regions, and the functions of key TFs were evaluated by experiments injecting mRNAs of combinations of these TFs into frog embryos.
Overall, the authors' statements are clear and the data presented appear reasonable. However, the computational analysis seems a bit poor and the subsequent experimental analyses may not be vigourous and significant enough to be considered for publication in the EMBO journal. In particular, the usage of the ATAC-seq data of whole embryos rather than those of specific regions for the integrated analysis with scRNA-seq would not be appropriate. The conclusion of the authors may change if they used the region-specific ATAC-seq or even scATAC-seq data. If it were not possible to use the data of AC and DMZ with some reasons, the authors should include and discuss the data of epigenetic profiles such as H3K4me3 and H3K27me3, which were examined in Figure1 and S1, from a global view point.
I would also like to suggest the authors to consider a better representation of the data and their explanation in several parts in the manuscript. It was not easy to understand the figures and the way of their analyses.

> As mentioned above, we regret that we insufficiently explained the computational motif analysis of Fig. 5. Integration of ATAC and single cell RNA data, is, like many tools in the rapidly developing single cell field is, underdeveloped, and we are happy that our co-author and colleague Dr. Simon van Heeringen is developing and implementing tools for this purpose and is at the forefront in this area. See comments above (reviewer #3) for additional explanation on the computational approach we used.

> In the revised version of the manuscript, we have explained how we performed the analysis. We have also extended our analysis using region-specific ATAC-seq (AC-DMZ). In addition, we have also done a similar analysis using the scATAC-seq data. This is now clearly explained in the manuscript. We hope that this will provide a richer biological context for the interpretation and will also directly address the reviewer's concern.

1. The authors should present the statistics data regarding the quality of all sequencing data, especially for scRNA-seq.

> We have included an additional file (Supplementary Table ST7) that shows the quality metrics of the sequencing data.

2. In the Figure 1C, the authors mentioned that the genes associated with lower levels of ATAC-seq signal were upregulated. For RNA-seq data, which cell types did the authors examined?

> We used the time series RNA-seq data (Owens et al., 2016). This was not clearly stated before, we have now modified the legend accordingly. The stages used are indicated in the labels on top of Fig. 1C. Please note that we analyzed the correspondence of changes in transcript levels and chromatin accessibility between two stages (so "stage 9-10" in Fig. 1C analyzes the changes between stage 9 and stage 10). We have modified the legend in this respect. As an additional analysis, we calculated the median expression profiles for genes associated with the clusters from Fig. 1B, based RNA-seq data (Owens et al. 2016). This is now provided as Supplemental Figure S1C (see also point 4 below).

3. In the result section, the authors explained that the visualization of scRNA-seq was performed using force-directed layout and the clustering using Louvain method. However, in the method section, it was stated that they were done using t-SNE and "FindClust" function in a package of R. The method for the visualization and clustering for Figure 3 did not seem to be appropriate.

> We thank the reviewer for notifying us for an omission in the methods. The t-SNE in the methods describes the dimensionality reduction of Fig. 4. For the whole embryo data, we used a force-directed layout (this only affects the visualization). In both cases this was

clearly stated in the legend, but the methods were not complete in this respect. We have now fixed this. Also, the "FindClust" function in Seurat software uses Louvain clustering, so the text in the Results and the Methods is consistent. Also, for consistency we have now changed to UMAP for dimensionality reduction in all analyses. This has changed the looks of the figures (how the cells are visualized in two dimensions), but interpretation of the results remains the same.

4. Related to the point #2, it would be more informative to see the transcription levels of genes by cluster-wise manner in Figure 1B.

> We agree with the reviewer and have included an additional figure (Supplemental Fig. S1C). The graph shows median expression profiles for genes associated with clusters from Fig. 1B. Additional information has been added to Figure legends.

5. The title stated that "Combinatorial action of TFs in open chromatin contributes to early cellular heterogeneity", but I could not understand which data support the statement that the open chromatin contributed to cellular heterogeneity. The authors should clarify this point.

> We meant: The combinatorial action of transcription factors contributes to early cellular heterogeneity. This is what we show. We also show that this happens in largely open chromatin. However, the title may have been ambiguous and caused some confusion. We therefore changed it to "Embryonic heterogeneity arises by combinatorial action of transcription factors in open chromatin".

1st Revision - Editorial Decision 10th Dec 2020

Thank you for submitting your revised manuscript, we have now received the report from one of the initial referees (see comments below). I am pleased to say that the comments have overall been satisfactorily addressed and the referee now also supports publication. Therefore I would like to ask you to address a number of editorial issues that are listed in detail below in a final revised version. Please make any changes to the manuscript text in the attached document only using the "track changes" option. Once these remaining issues are resolved, we will be happy to formally accept the manuscript for publication.

Referee #4:

In the revised manuscript, the authors provided clear answers to my comments. I belive this article should be published in the EMBO journal.

Thank you again for submitting the final revised version of your manuscript. You will hear from me again regarding textual changes in the final files for transfer to our publisher, but for now I am pleased to inform you that we have formally accepted the manuscript for publication in The EMBO Journal.

EMBO PRESS

YOU MUST COMPLETE ALL CELLS WITH A PINK BACKGROUND ↓

PLEASE NOTE THAT THIS CHECKLIST WILL BE PUBLISHED ALONGSIDE YOUR PAPER

Corresponding Author Name: Gert Jan C. Veenstra
Journal Submitted to: The EMBO Journal
Manuscript Number: EMBOJ-2020-104913

Reporting Checklist For Life Sciences Articles (Rev. June 2017)

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. These guidelines are consistent with the Principles and Guidelines for Reporting Preclinical Research issued by the NIH in 2014. Please follow the journal's authorship guidelines in preparing your manuscript.

A- Figures

1. Data

The data shown in figures should satisfy the following conditions:

- ➔ the data were obtained and processed according to the field's best practice and are presented to reflect the results of the experiments in an accurate and unbiased manner.
- ➔ figure panels include only data points, measurements or observations that can be compared to each other in a scientifically meaningful way.
- ➔ graphs include clearly labeled error bars for independent experiments and sample sizes. Unless justified, error bars should not be shown for technical replicates.
- ➔ if n< 5, the individual data points from each experiment should be plotted and any statistical test employed should be justified
- ➔ Source Data should be included to report the data underlying graphs. Please follow the guidelines set out in the author ship guidelines on Data Presentation.

2. Captions

Each figure caption should contain the following information, for each panel where they are relevant:

- ➔ a specification of the experimental system investigated (eg cell line, species name).
- ➔ the assay(s) and method(s) used to carry out the reported observations and measurements
- ➔ an explicit mention of the biological and chemical entity(ies) that are being measured.
- ➔ an explicit mention of the biological and chemical entity(ies) that are altered/varied/perturbed in a controlled manner.

- ➔ the exact sample size (n) for each experimental group/condition, given as a number, not a range;
- ➔ a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates (including how many animals, litters, cultures, etc.).
- ➔ a statement of how many times the experiment shown was independently replicated in the laboratory.
- ➔ definitions of statistical methods and measures:
  - common tests, such as t-test (please specify whether paired vs. unpaired), simple $\chi 2$ tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
  - are tests one-sided or two-sided?
  - are there adjustments for multiple comparisons?
  - exact statistical test results, e.g., P values = x but not P values < x;
  - definition of 'center values' as median or average;
  - definition of error bars as s.d. or s.e.m.

Any descriptions too long for the figure legend should be included in the methods section and/or with the source data.

In the pink boxes below, please ensure that the answers to the following questions are reported in the manuscript itself. Every question should be answered. If the question is not relevant to your research, please write NA (non applicable). We encourage you to include a specific subsection in the methods section for statistics, reagents, animal models and human subjects.

USEFUL LINKS FOR COMPLETING THIS FORM

http://www.antibodypedia.com
http://1degreebio.org
http://www.equator-network.org/reporting-guidelines/improving-bioscience-research-repor

http://grants.nih.gov/grants/olaw/olaw.htm
http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Useofanimals/index.htm
http://ClinicalTrials.gov
http://www.consort-statement.org
http://www.consort-statement.org/checklists/view/32-consort/66-title

http://www.equator-network.org/reporting-guidelines/reporting-recommendations-for-tumo

http://datadryad.org

http://figshare.com

http://www.ncbi.nlm.nih.gov/gap

http://www.ebi.ac.uk/ega

http://biomodels.net/

http://biomodels.net/miriam/
http://jjj.biochem.sun.ac.za
http://oba.od.nih.gov/biosecurity/biosecurity_documents.html
http://www.selectagents.gov/

B- Statistics and general methods

Please fill out these boxes ↓ (Do not worry if you cannot see all your text once you press return)

1.a. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?	The experiments on whole embryo ATAC-seq, dissected embryo ATAC-seq, single cell RNA-seq on dissected embryos, and bulk RNA-sequencing were performed in duplicate. The scATAC-seq experiment represents nuclei from 100 embryos. These numbers were deemed sufficient for stringent outlier statistics.
1.b. For animal studies, include a statement about sample size estimate even if no statistical methods were used.	N/A
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?	N/A
3. Were any steps taken to minimize the effects of subjective bias when allocating animals/samples to treatment (e.g. randomization procedure)? If yes, please describe.	N/A
For animal studies, include a statement about randomization even if no randomization was used.	N/A
4.a. Were any steps taken to minimize the effects of subjective bias during group allocation or/and when assessing results (e.g. blinding of the investigator)? If yes please describe.	N/A
4.b. For animal studies, include a statement about blinding even if no blinding was done	N/A
5. For every figure, are statistical tests justified as appropriate?	We use non-parametric tests as the data are not normally distributed
Do the data meet the assumptions of the tests (e.g., normal distribution)? Describe any methods used to assess it.	We use non-parametric tests as the data are not normally distributed
Is there an estimate of variation within each group of data?	We use mainly outlier statistics (corrected for multiple testing).

Is the variance similar between the groups that are being statistically compared?	N/A

C- Reagents

6. To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile. e.g., Antibodypedia (see link list at top right), 1DegreeBio (see link list at top right).	N/A
7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination.	N/A

* for all hyperlinks, please see the table at the top right of the document

D- Animal Models

8. Report species, strain, gender, age of animals and genetic modification status where applicable. Please detail housing and husbandry conditions and the source of animals.	Xenopus tropicalis, Nigerian, wild type, outbred. The animal were raised in-house for the specific purpose of the experiments. The housing and husbandry conditions were as mandated by European and Dutch law.
9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.	De central committee animal experiments (Centrale Commissie Dierproeven) has issued permission (application number AVD1030020171826) for the animal experiments, in accordance with the recommendations of the Dierexperimentencommissie RU DEC.
10. We recommend consulting the ARRIVE guidelines (see link list at top right) (PLoS Biol. 8(6), e1000412, 2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under 'Reporting Guidelines'. See also: NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance.	Yes

E- Human Subjects

11. Identify the committee(s) approving the study protocol.	N/A
12. Include a statement confirming that informed consent was obtained from all subjects and that the experiments conformed to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report.	N/A
13. For publication of patient photos, include a statement confirming that consent to publish was obtained.	N/A
14. Report any restrictions on the availability (and/or on the use) of human data or samples.	N/A
15. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable.	N/A
16. For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under 'Reporting Guidelines'. Please confirm you have submitted this list.	N/A
17. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under 'Reporting Guidelines'. Please confirm you have followed these guidelines.	N/A

F- Data Accessibility

18: Provide a "Data Availability" section at the end of the Materials & Methods, listing the accession codes for data generated in this study and deposited in a public database (e.g. RNA-Seq data: Gene Expression Omnibus GSE39462, Proteomics data: PRIDE PXD000208 etc.) Please refer to our author guidelines for 'Data Deposition'. Data deposition in a public repository is mandatory for: a. Protein, DNA and RNA sequences b. Macromolecular structures c. Crystallographic data for small molecules d. Functional genomics data e. Proteomics and molecular interactions	ATAC-seq, scATAC-seq, scRNA-seq and bulk RNA-seq data have been deposited at NCBI GEO, accession number GSE145619 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145619)
19. Deposition is strongly recommended for any datasets that are central and integral to the study; please consider the journal's data policy. If no structured public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under 'Expanded View' or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right).	We have provided Supplementary Tables to provide access to processed data.
20. Access to human clinical and genomic datasets should be provided with as few restrictions as possible while respecting ethical obligations to the patients and relevant medical and legal issues. If practically possible and compatible with the individual consent agreement used in the study, such data should be deposited in one of the major public access-controlled repositories such as dbGAP (see link list at top right) or EGA (see link list at top right).	N/A
21. Computational models that are central and integral to a study should be shared without restrictions and provided in a machine-readable form. The relevant accession numbers or links should be provided. When possible, standardized format (SBML, CellML) should be used instead of scripts (e.g. MATLAB). Authors are strongly encouraged to follow the MIRIAM guidelines (see link list at top right) and deposit their model in a public database such as Biomodels (see link list at top right) or JWS Online (see link list at top right). If computer source code is provided with the paper, it should be deposited in a public repository or included in supplementary information.	N/A

G- Dual use research of concern

22. Could your study fall under dual use research restrictions? Please check biosecurity documents (see link list at top right) and list of select agents and toxins (APHIS/CDC) (see link list at top right). According to our biosecurity guidelines, provide a statement only if it could.	N/a