Developmental Dynamics of the Epigenome and Methods to Find Relevant Regulatory Motifs

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Dania Machlab

2021

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Filippo M. Rijli
PD Dr. Michael B. Stadler
Prof. Dr. Dirk Schübeler
Dr. Philipp Bucher

Basel, den 16.11.2021

Prof. Dr. Marcel Mayor
(Dean of Faculty)

To My Parents
Petra & Hassan

ABSTRACT

Studying the epigenome and which transcription factors interact with it gives us a better understanding of how developmental processes are regulated and harmoniously orchestrated. For sensory neurons, such signals correspond to environmental stimuli. A group of genes called immediate early genes (IEGs) are known to play important roles during development, and they are some of the first to respond to signals a cell receives. They tend to encode for transcription factors (TFs), are activated within minutes and regulate the activity of other genes. Studying the features of these genes, we found a new epigenetic signature that hints at why they can be induced so fast. They have the active H3K27ac mark on their promoters, and the repressive H3K27me3 mark on their gene bodies. We found a few hundred genes with this signature and called them 'bipartite' genes.

Bipartite genes are very lowly expressed, or not at all. They are, however, in a poised state that is even more ready to be quickly induced than the known bivalent genes. The needed transcriptional machinery is already sitting at the promoter. We used t-distributed stochastic neighbor embedding to jointly visualize chromatin accessibility and several histone marks on all genes in barrelette neurons of the somatosensory system in mice. Moreover, we used several developmental time points to visualize genome-wide changes in chromatin states across development. This allowed us to visualize the epigenetic dynamics that bipartite genes undergo by observing how they move from one developmental time point to another in these chromatin landscapes.

As mentioned, IEGs correspond to TFs that have important regulatory roles. Knowing which TFs play relevant or functional roles is key to understanding the underlying developmental processes. Motivated by the importance of finding relevant TFs, we developed computational methods that enable us to make predictions, in an unbiased way, about which TFs could explain an experimental measure of interest, typically coming from sequencing data. We created an R package called *monaLisa*, short for "**mo**tif a**na**lysis with **Lisa**", that allows for these methods to be used in a user-friendly manner.

The package offers two main ways of identifying regulatory motifs. In the first approach, we made use of an existing method of correcting for sequence composition differences to apply a binned motif enrichment analysis. This method links motif enrichment to an experimental value, for example changes in DNA methylation between two conditions. The second approach uses linear regression to select a set of TFs that are likely to explain the given observations. Specifically, we use randomized lasso stability selection to discover relevant motifs. The new epigenetic signature with the bipartite genes illustrates how the epigenome can control a timely transcriptional response during development, and the methods in *monaLisa* further enable us to decipher which TFs could be key players.

ACKNOWLEDGEMENTS

The resulting work in this thesis would not have been possible without the support and encouragement of many people I wish to thank here. Firstly, I am very grateful to my supervisors, Michael and Filippo. Thank you for your guidance, patience, kindness and support. I have learnt a great deal under your supervision, and our conversations always left me even more enthusiastic about the science we pursue. Thank you for always making time for me. My thanks also extends to Luca Giorgetti for providing valuable feedback during our thesis committee meetings. I am thankful to Taro Kitazawa for a fantastic and enriching collaboration. My gratitude also extends to all current and former members of the Rijli lab for their helpful comments and support. Upasana and Leslie, thank you for making me feel so welcome at the FMI. I am also thankful to Lukas Burger for the great collaboration and to all the members of the computational biology platform. Charlotte, Pan, Dimos and Hans-Rudolf, thank you for the discussions, the guidance, and for making the office a joyful place to be. I am also thankful to Harold, Vero and Milica for making my time in Basel a fun and memorable one. Thanks to Brilé, Maria and Martin. I am also indebted to my family — to my brothers Hadi and Karim, and my parents Petra and Hassan — for their constant love and support. Finally, I wish to thank Mathias for making my time in Basel a very special one. Thank you for your kindness, your encouragement and for making me laugh.

CONTENTS

ABBREVIATIONS

A	adenine
ATAC-seq	assay for transposase-accessible chromatin using sequencing
AUC	area under the curve
bp	base pairs
BS-seq	bisulfite sequencing
4C	circular chromosome conformation capture
C	cytosine
ChIP-seq	chromatin immunoprecipitation sequencing
DNA	deoxyribonucleic acid
ES cells	embryonic stem cells
FPR	false positive rate
G	guanine
$H_3K_4me_2$	di-methylation on histone 3 at lysine residue 4
$H_3K_{27}ac$	acetylation on histone 3 at lysine residue 27
$H_3K_{27}me_3$	tri-methylation on histone 3 at lysine residue 27
IEG	immediate early gene
LMR	lowly methylated region
mRNA	messenger RNA
NP	neural progenitors
PC	principal component
PCA	principal component analysis
PFER	per-family error rate
PFM	position frequency matrix
PPM	position probability matrix
PWM	position weight matrix
RNA	ribonucleic acid
RNA-seq	RNA sequencing
RSS	residual sum of squares
SNR	signal-to-noise ratio

T	thymine
TAD	topologically associated domain
TF	transcription factor
TFBS	TF binding sites
TPR	true positive rate
t-SNE	t-distributed stochastic neighbor embedding
TSS	transcription start site
ZOOPS	zero or one occurrence per sequence

1

INTRODUCTION

'Spirit' comes from the Latin word 'to breathe'. What we breathe is air, which is certainly matter, however thin. Despite usage to the contrary, there is no necessary implication in the word 'spiritual' that we are talking of anything other than matter (including the matter of which the brain is made), or anything outside the realm of science. On occasion, I will feel free to use the word. Science is not only compatible with spirituality; it is a profound source of spirituality. When we recognize our place in an immensity of light-years and in the passage of ages, when we grasp the intricacy, beauty, and subtlety of life, then that soaring feeling, that sense of elation and humility combined, is surely spiritual.

— Carl Sagan (1997)

Deoxyrobonucleic acid (DNA) is the material that gave and continues to give rise to all living and extinct creatures we know of today. Be it a human being, a cat, a bacterium in the gut, or a tree that lived hundreds of millions of years ago, the cells from all of those organisms contain a variation of this material. It can be thought of as the constructional blueprint that cells and organisms rely on to develop and function properly — a bundle of code to produce essential things needed by cells, along with its unique and remarkable ability to replicate. Studying specific properties of this molecule, how other factors in a cell interact with it, how it forms in three dimensions, how it changes over time, and what rules govern all of these facets, has captured the fascination of many scientists over the years. This dissertation presents work we have done studying the DNA, describing a particular epigenetic signature we found on a subset of genes that gives them a functional advantage, and providing a few computational tools that can help make predictions on relevant or regulatory factors that could be interacting with the DNA. This chapter introduces concepts in biology and statistics that serve as background information for chapters 2 and 3.

1.1 THE EUKARYOTIC GENOME

All cells in a single living organism contain the same DNA, a long sequence of combinations of four nucleotides or bases, namely adenine (A), cytosine (C), guanine (G), and thymine (T). The DNA molecule is a double helix consisting of two strands, a forward and a complementary reverse strand, where A and T nucleotides interact to form base pairs, and G and C nucleotides interact with each other as base pairs (bp) (Watson & Crick, 1953). The DNA is organized on several levels to compactly fit inside the nucleus of a cell and the final resulting structures are called chromosomes.

base pairs

genes

Specific regions along the DNA constitute genes, which are sequences that are transcribed into messenger ribonucleic acid (mRNA) and subsequently translated into proteins with the appropriate machinery in the cells. Genes, however, constitute a small fraction of the DNA, with most of the genome being non-protein-coding. In humans, for instance, less than 5 % of the genome consists of coding sequences (Lander, Linton, *et al.*, 2001). The region surrounding the transcription start site (TSS) of a gene is known as a promoter, and it is where additional proteins and factors are recruited to initiate transcription (Sandelin, Carninci, *et al.*, 2007).

high-throughput sequencing

RNA-seq

The genome is regulated in complex ways and there are various experiments that measure different properties of the DNA or RNA that can give deep insight into some of these mechanisms. The advancement of technologies, particularly next-generation sequencing (Metzker, 2010), has allowed the production of vast volumes of sequences, and the ability to do genome-wide measurements relatively inexpensively. We can for instance measure genome-wide mRNA levels for all genes, to study gene expression and activity with RNA sequencing (RNA-seq). In mapping the sequenced reads back to an annotated reference genome, these experiments allow us to quantitatively examine the measure of interest and how it changes between different conditions.

TF

enhancer

Gene activity is regulated by proteins called transcription factors (TFs), which bind near the TSSs of gene promoters, or at genomic regions that are farther away and typically a few hundred bp long, called enhancers (Spitz & Furlong, 2012). TF binding on an enhancer and the promoter of a gene can allow those two otherwise not-so-close regions to communicate by recruiting other transcriptional cofactors, to control and regulate the transcriptional activity of the gene (Zabidi & Stark, 2016). However, there is a lot unknown in terms of exactly how the rules of enhancer regulation are governed genome-wide. The binding of the right TFs, at the right locations, and at the right time is key during development (Spitz & Furlong, 2012).

DNA methylation

BS-seq

TF binding can also be influenced by DNA methylation. Modifications to the DNA beyond the sequence composition are termed as epigenetic, and they can also have major regulatory roles across the genome. DNA methylation is an example of an epigenetic feature, whereby a methyl group is added to a nucleotide, most commonly to the C base which is followed by a G (Bestor & Coxon, 1993). The binding of some TFs like NRF1 on the DNA can be methylation-sensitive, whereby in the presence of DNA methylation these proteins do not bind specific DNA regions, but then bind upon the removal of methylation (Domcke, Bardet, *et al.*, 2015). A common experimental approach to measure the methylation of Cs in this context is called bisulfite sequencing (BS-seq). Methylated Cs are converted to uracil and then to thymine after deep sequencing (Frommer, McDonald, *et al.*, 1992).

TAD

The long-range interactions between enhancers and genes typically occur within units called topologically associated domains (TADs) (Dixon, Selvaraj, *et al.*, 2012), although interactions outside of these boundaries are also possible. TADs are regions in the genome, typically a few hundred kilo bases or even a mega base long, where regions within a TAD interact more frequently with each other, and less so with regions outside of the TAD boundaries (Giorgetti, Servant & Heard, 2013). There is thus a structure and organization in how the genome interacts with itself in space. Chromosomes, for example, are known to occupy specific territories in the cell nucleus (T. Cremer & C. Cremer, 2001). There are various chromosome conformation capture techniques to measure three-dimensional chromosomal interactions. However, the emergence of Hi-C, which uses proximity-based ligation followed by deep sequencing, has allowed us to study chromosomal interactions in a genome-wide

fashion (Lieberman-Aiden, van Berkum, *et al.*, 2009). Circular chromosome conformation capture (4C) on the other hand, measures any chromosomal interactions with a pre-defined locus on the genome (Simonis, Klous, *et al.*, 2006), for example the promoter of a specific gene.

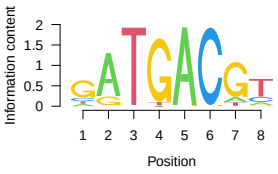
4C

1.1.1 *Transcription factor binding sites*

TFs bind at specific sequences on the genome that are around 6-12 bp long (Spitz & Furlong, 2012). These binding sites are referred to as transcription factor binding sites (TFBS). A common way to represent these TFBS and have a representative motif, is through a position weight matrix (PWM) (Stormo, T. D. Schneider, *et al.*, 1982). Starting from a set of sequences that contain the TFBS, we can first summarize them in a position count matrix or position frequency matrix (PFM), where the rows correspond to the four nucleotides A, C, G and T, and the columns represent the positions along the binding site. Each entry in this matrix shows the number of times a particular nucleotide was observed at a specific position. This matrix can be transformed into a position probability matrix (PPM) by dividing the entries by the column sums per position. The entries per position then sum to one. To get the PWM, the PPM is divided by an expected background and $\log_2$-transformed. A uniform background would be a probability of 0.25 for observing each of the four nucleotides. When scanning the genome for a hit for a particular motif using the PWM, the score of a hit is calculated by summing the PWM values of a specific sequence of nucleotides. For example, GATGACGT would take the PWM value of G at position 1, add it to that of A at position 2 and so on. If this score is greater than some defined threshold, the sequence is considered to be a TFBS for the TF of interest. Finally, we can visually represent the motifs using an information content matrix, where we can get an idea of how conserved the individual positions are in a motif. For each position, we take the total information content and subtract the uncertainty at that position to get the amount of information present at the position (T. D. Schneider & Stephens, 1990).

TFBS

PWM



Atf3 motif from the JASPAR2020 database with information content per position

1.1.2 *Chromatin accessibility*

Looking at the accessible regions of the DNA can tell us which regions are active. By using the appropriate computational tools, we can also analyze what TFs may be bound and playing regulatory roles in these accessible sites. There are several experimental methods that measure chromatin accessibility, one of them being the assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro, Giresi, *et al.*, 2013). The so-called Tn5 transposase enzyme cuts open DNA while simultaneously inserting the adapter sequence at the cut site. The resulting DNA fragments are then amplified and sequenced. The resulting reads are mapped back to the genome. Genomic regions that were very accessible will have a high number of reads mapped to them, whereas regions that were inaccessible due to for example compaction of the DNA or having other factors bound at the DNA, will have no or very few reads mapped to them. These experiments allow us to measure chromatin accessibility in a quantitative manner, and for example define regulatory regions in the genome, where the chromatin is open.

ATAC-seq

1.1.3 *Histone marks*

nucleosome

To enable compact storing of the long DNA molecule, DNA is wrapped around histone proteins forming structures called nucleosomes. Each nucleosome typically spans 146 bp of DNA wrapped around eight histone proteins (Li & Reinberg, 2011). An additional ninth histone protein, known as a linker histone, binds to the outside of the nucleosome core, stabilizing the whole structure (Lawrence, Daujat & R. Schneider, 2016). The DNA stretches connecting one nucleosome to another are called linker DNA.

histone

Modifications at specific positions in the histone proteins have been implicated with activity or repression, and can thus have an influence over gene and enhancer activity. For instance, a trimethylation on histone 3, at lysine residue 27, called H3K27me3 or the Polycomb mark, is associated with repression. Acetylation at the same position, called H3K27ac, on the other hand, is a mark associated with activity and transcription, as is the H3K4me2 modification (Dong & Weng, 2013).

These histone marks and their distributions can be experimentally quantified genome-wide using chromatin immunoprecipitation sequencing (ChIP-seq) (Johnson, Mortazavi, *et al.*, 2007).

ChIP-seq

The DNA is fragmented, and an antibody targeting the specific histone modification is used to pull down fragments of the DNA which are bound to this mark, thus enriching for the target of interest. These are then amplified and deeply sequenced. ChIP-seq can also be used to target specific proteins or transcription factors present on the DNA after first cross-linking, to find the binding sites or positions where a protein of interest was present or interacting with the DNA (Park, 2009).

Histone marks and combinations thereof have functional roles in gene regulation and development. The co-occurrence of the repressive H3K27me3 mark and the active H3K4me2 mark on the

bivalent genes

promoters of certain genes has been described and termed "bivalent". Such genes are in a poised state with little or no expression, and await to be induced given the right stimuli or cues (Bernstein, Mikkelsen, *et al.*, 2006). Cranial neural crest cells for instance, which migrate to form different parts of the craniofacial skeleton, have many TF genes that are in a poised bivalent state. Once the neural crest cells migrate to a specific region on the face, they acquire the appropriate regional identity by starting to express these TFs in response to the environmental cues. Moreover, different groups of bivalent genes from the neural crest cells become active in the different subsets of the craniofacial skeleton. The appropriate previously bivalent genes thus become active and the cells acquire a positional identity (Minoux, Holwerda, *et al.*, 2017).

1.1.4 *Single-cell measurements*

The sequencing experiments described so far allow for genome-wide studies of several aspects of the genome and its activity, in bulk samples. That means that these measurements reflect average readouts of a population of cells. However, single-cell sequencing experiments, like single-cell RNA-seq or single-cell ATAC-seq are also possible now. These data sets tend to be much more sparse than the bulk samples. Technologies that do single-cell experiments as well as the computational methods that can be used to analyze such complex data sets and the problems they come with is still an active field of research.

1.2 EPIGENETICS IN NEUROBIOLOGY

Epigenetic features like the described DNA methylation or histone modifications are key during neuronal development, and correct timing is especially important. Yet not much is known about how these changes and gene expression relate to neural activity in the brain. The onset of next-generation sequencing has allowed us to study the behavior of neurons on a genome-wide molecular level.

The mouse brain is often used as a model system to study neural development and activity. The so-called barrelette system is an example of a system that can be perturbed to understand neuronal development, how these neurons respond to the right stimulus, and how they form circuits in the brain. The barrelette map can be seen as being organized into blocks of cells or neurons, with each block receiving sensory input from a whisker on the face of the mouse. Moreover, the whisker positions on the face of the mouse are faithfully represented in the barrelette map, with whiskers that are close to each other on the face, also mapping to regions that are close to each other in the barrelette map in the brain. The barrelette map already starts forming in the embryo a few days before birth, but is further refined after birth as it receives sensory input when the mice start using their whiskers (Kitazawa & Rijli, 2018). Systems like this somatosensory system can be helpful to further our understanding of how neurons respond to environmental signals and how the epigenetic states of key genes could play important roles in such contexts.

A class of genes known as immediate early genes (IEGs) has been described as some of the first responders to stimuli in cells. The response is within minutes and they usually encode for TFs that can activate and regulate other genes (Fowler, Sen & Roy, 2011). IEGs are known to have roles in neuronal plasticity, a feature that is important to be able to react to environmental stimuli and adapt accordingly. They have been implicated in learning and memory and other cognitive functions (Okuno, 2011).

1.3 DIMENSIONALITY REDUCTION

The various sequencing experiments described provide us with reads that are typically mapped to a reference genome and quantified on genes or defined regions to get count matrices and follow up with quantitative analyses. With a large number of samples and many data types quantified across a large set of features, we end up with high-dimensional data sets. There are various dimensionality reduction methods that enable us to visualize these complex data sets in a lower-dimensional representation. Principal component analysis (PCA) is one such technique (Pearson, 1901). It is *PCA*
a linear dimensionality reduction method that produces a set of principal components (PCs), ordered by the amount of variance they explain in decreasing order. Each PC is an axis in a new coordinate system that better captures the variability of the data. An example use case of PCA is an expression count matrix, where the genes are represented as rows are the samples as columns. The high dimensionality is at the level of the thousands of genes quantified. In doing a PCA and plotting the samples in the new coordinate system of PC1 and PC2, we can see how the samples and the underlying conditions and batches group. Samples that are similar to each other will group closely, whereas the more dissimilar they are, the farther apart they will be.

A dimensionality reduction can also be done to reduce the number of features for further analysis, since having too many features can be a problem in terms of feasibility, complexity and run-time. Following up with the gene-by-sample matrix example, the first fifty PCs could be used for further downstream analyses. Another dimensionality reduction method commonly used is t-distributed stochastic neighbor embedding (t-SNE) (Maaten & Hinton, 2008). In contrast to the PCA, this is a stochastic and non-linear dimensionality reduction technique. t-SNE tries to produce a lower-dimensional representation of the data, maintaining as much as possible the distances between the points from the higher dimensions. That means that points that are close to each other in the high-dimensional space, will be close to each other in the lower representation of two or three dimensions. Such techniques allow us to visualize many data sets and features together. For example, t-SNE can be used to jointly visualize the quantifications of several histone marks and accessibility across genes (Kitazawa, Machlab, *et al.*, 2021). There, genes with similar chromatin states are grouped together, and it is possible to visualize how genes change chromatin state from one developmental time point to another by observing how they move on this t-SNE plot.

t-SNE

1.4 LINEAR REGRESSION

Linear regression is frequently used to make predictions or understand the relationships between variables of interest, and how they explain observations, like the observed counts or fold changes measured between different conditions. In a simple linear regression, the variable y depends on one explanatory variable x:

simple linear regression

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1.1}$$

where β_0 and β_1 are the intercept term and the coefficient for x, respectively. The error term ε captures the noise or lack of fit to the linear equation. When y depends on more than one explanatory variable, it can be modeled through a multiple linear regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{1.2}$$

multiple linear regression

for observation i, and where we now have p explanatory variables. This can be further formulated as a vector y of N observations, a vector β of p coefficients, and an N x p matrix X, of the values of the explanatory variables (columns) per observation (rows). This gives us $y = X\beta + \varepsilon$, where y and X are often called the response variable and the predictor matrix, respectively.

In a linear regression, the values of the coefficients ($\beta_1, \ldots, \beta_p$) are estimated such that the difference between the predicted and the actual y is minimized. We are thus trying to minimize the residual errors ε_i when we fit the coefficients. A common way to do this is by minimizing the residual sum of squares (RSS) (Hastie, Tibshirani & J. Friedman, 2009):

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \tag{1.3}$$

RSS

Equation (1.3) is taking the difference between the actual and the predicted y value per observation i, squaring this difference, and summing the squared differences for all N observations. The sizes of the estimated coefficients will indicate how much a variable or predictor contributes to explaining the response. Variables with high coefficients would contribute a lot, while a low coefficient would indicate little contribution.

1.4.1 *Penalized regression*

Multiple linear regression suffers from a few problems when it comes to estimating reasonable beta coefficients in practice. On the one hand, the estimated values can be problematic when predictors are highly correlated to each other (Hastie, Tibshirani & J. Friedman, 2009). For example, suppose that two predictors x_1 and x_2 are highly correlated with each other, but not with the response. If both predictors get equal values but opposite signs in their beta estimates, they cancel each other out in equation 1.3. In such a case the beta coefficients can get arbitrarily large without adding much to the RSS. Interpreting the values of the estimated beta coefficients can thus be problematic. On the other hand, the multiple linear regression suffers from over-fitting. Since the measurements in a response vector or predictor matrix come with some levels of noise in practice, the beta values estimated in one data set may vary quite substantially when estimated in a new data set of the same conditions but with new measurements. To overcome these problems and estimate beta coefficients that are a bit more robust and representative, shrinkage methods have been proposed that impose a penalty on the beta coefficients. This problem is known as the variance-bias tradeoff, since the beta estimates will vary less with such imposed penalties at the expense of being slightly more biased (James, Witten, *et al.*, 2013).

ridge regression

Ridge regression is an example of a linear regression that places a limit on the beta coefficients. It does so by adding to equation 1.3 a penalty term λ on the sum of the squared coefficients when estimating them:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{1.4}$$

where $\hat{\beta}$ is the vector of estimated beta coefficients such that the equation is minimized. When λ is zero, equation 1.4 is equivalent to the RSS in equation 1.3. The bigger the value of the penalty λ, the harder the limit on the sum of the coefficients is, and the more shrinkage there is.

lasso regression

Ridge regression shrinks the values of the estimated coefficients towards zero with the penalty term. However the coefficients are not equal to zero and every variable has an estimated beta coefficient. In order to select meaningful variables by setting the coefficients of unimportant variables to zero, the lasso regression imposes a penalty that allows for this. Lasso regression penalizes the sum of the absolute values of the coefficients, which allows for variable selection, and can be formulated in the following equivalent Lagrangian form (Hastie, Tibshirani & J. Friedman, 2009):

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{1.5}$$

elastic-net regression

Elastic-net regression offers a middle ground between the ridge and lasso regressions. It selects variables like the lasso, and shrinks the coefficients of correlated predictors like the ridge (Hastie, Tibshirani & J. Friedman, 2009), using the following penalty:

$$\lambda \sum_{j=1}^{p} \left(\alpha\beta_j^2 + (1-\alpha)|\beta_j|\right) \tag{1.6}$$

with the $\alpha \in [0, 1]$ dictating how close the regression is to a lasso ($\alpha = 0$) or ridge ($\alpha = 1$).

The various regression methods are useful in different ways, whether the aim is doing good predictions, interpreting the beta coefficients, or selecting meaningful variables. They are relatively simple yet powerful models due to how well understood and interpretable they are.

1.5 GOALS AND STRUCTURE OF THE THESIS

This dissertation is composed of two main projects presented in chapters 2 and 3. In the project discussed in chapter 2, the goal was biological discovery, where we performed computational analyses on several types of sequencing data to study the epigenome and gene expression during neuronal development. There, we found a new epigenetic signature, that several IEGs have, which leaves these genes in a poised state ready for activation upon the right signal, but not expressed due to the presence of the repressive H3K27me3. In the project presented in chapter 3, the goal was to develop computational tools that can be used to identify regulatory TFs in a given sequencing experiment. There, among other things, we made use of linear regression to find important motifs. Chapter 4 briefly discusses the findings from both projects and provides some final remarks and an outlook.

2

CHROMATIN IN THE CONTEXT OF NEURONAL DEVELOPMENT

Immediate early genes are a particular group of genes that are first to be expressed in response to the appropriate signal or stimulus. In this project, we investigated what could explain this rapid response on the epigenetic level by studying developing sensory neurons in the mouse somatosensory system. In examining the chromatin marks of these genes, we found the active H3K27ac mark embedded on the promoter region coexisting with the repressive H3K27me3 mark on the first 2,000 bp of the gene body. We called genes with such chromatin profiles bipartite genes. Bipartite genes are not or very lowly expressed due to the presence of the Polycomb mark. However, their promoters are already accessible and ready to be induced. Unlike bivalent genes, which are known to have the active H3K4me2 mark and the repressive H3K27me3 marks around their promoters, bipartite genes are more readily and quickly inducible. In fact, RNA polymerase II is already present at the promoter. We found bipartite genes in various tissue types and developmental time points, and estimated their numbers to be in the range of a few hundred genes. To visualize how bipartite genes change chromatin states from one developmental time point to another, we applied a non-linear dimensionality reduction method called t-distributed stochastic neighbor embedding. This method was used on quantifications of chromatin accessibility and several histone marks across all genes and at several developmental time points in barelette neurons. We could then see how the genes changed chromatin states between different time points, and observe that bipartite genes are the most dynamic of all (Kitazawa, Machlab, *et al.*, 2021).

This project was a collaboration with Taro Kitazawa as well as the other authors listed in the publication. My contributions included handling the project on the computational side, analyzing and integrating the data sets. I performed the computational analyses on the sequencing data sets with the exception of the 4C and single-cell RNA-seq analyses which were done by Michael Stadler and Charlotte Soneson, respectively.

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00789-z

Check for updates

A unique bipartite Polycomb signature regulates stimulus-response transcription during development

Taro Kitazawa[1,6], Dania Machlab[1,2,3,6], Onkar Joshi[1], Nicola Maiorano[1], Hubertus Kohler[1], Sebastien Ducret[1], Sandra Kessler[1], Henrik Gezelius[4,5], Charlotte Soneson[1,2], Panagiotis Papasaikas[1,2], Guillermina López-Bendito[4], Michael B. Stadler[1,2] and Filippo M. Rijli[1,3] ✉

Rapid cellular responses to environmental stimuli are fundamental for development and maturation. Immediate early genes can be transcriptionally induced within minutes in response to a variety of signals. How their induction levels are regulated and their untimely activation by spurious signals prevented during development is poorly understood. We found that in developing sensory neurons, before perinatal sensory-activity-dependent induction, immediate early genes are embedded into a unique bipartite Polycomb chromatin signature, carrying active H3K27ac on promoters but repressive Ezh2-dependent H3K27me3 on gene bodies. This bipartite signature is widely present in developing cell types, including embryonic stem cells. Polycomb marking of gene bodies inhibits mRNA elongation, dampening productive transcription, while still allowing for fast stimulus-dependent mark removal and bipartite gene induction. We reveal a developmental epigenetic mechanism regulating the rapidity and amplitude of the transcriptional response to relevant stimuli, while preventing inappropriate activation of stimulus-response genes.

During development, cells are exposed to a variety of distinct environmental signals to which they may need to rapidly respond in a spatiotemporally regulated manner, in order to keep their differentiation schedule. Stimulus-response genes are essential for rapid cellular responses to extracellular signals[1–3]. Among them, immediate early genes (IEGs) are induced in multiple cell types within minutes in a stimulus-dependent manner, often encoding transcription factors (for example, Fos and Egr1), which in turn regulate the expression of downstream late-response genes (LRGs) through activation of enhancers[1,4–6]. Before induction, IEGs share key regulatory properties, which poise them for rapid stimulus-dependent activation. In general, these include accessible promoters and enhancers bound by serum response factor, nuclear factor-κB, cyclic AMP response element-binding protein (CREB) and/or activator protein-1 transcription factors, which are posttranslationally modified upon stimulus response, as well as transcriptionally permissive histone modifications (H3K4me2/3) and paused RNA polymerase II (RNAPII)[2,7]. Despite their shared organization, differences in transcription initiation, elongation or mRNA processing and stability may result in IEG induction differences[2,7]. Moreover, IEGs are both general (that is, the same IEGs are induced in most cell types in response to different stimuli) and cell-type specific (responding to specific signals in different cell types)[3,6,8–12]. How spatiotemporal regulation and specificity of the IEG transcriptional response is achieved in developing cells, and how untimely induction of IEGs in response to spurious signals is prevented are poorly understood.

Here, we asked whether and how chromatin states might also contribute to stimulus-dependent transcriptional regulation of IEGs during development, choosing the mouse developing somatosensory neurons as a suitable model. We then further confirmed the general validity of our findings in developing neural crest, heart, liver and embryonic stem cells (ESCs). We discovered, and functionally investigated, a unique H3K27ac/H3K27me3 bipartite chromatin signature, which provides an epigenetic mechanism to modulate the rapidity and amplitude of the transcriptional response of inducible IEGs to distinct stimuli during development. Our findings support the involvement of Polycomb (Pc)-dependent H3K27me3 on the gene body in inhibiting the productive elongation of RNAPII on bipartite genes. While strong stimuli allow for the rapid removal of Pc marking of gene bodies and fast transcriptional induction, Pc marking of gene bodies of bipartite stimulus-response genes may establish a threshold to prevent rapid transcriptional induction of IEGs in response to suboptimal and/or nonphysiologically relevant levels of environmental stimuli.

Results

Transcriptional and chromatin profiling of activity-regulated genes in developing neurons. During early postnatal sensory neuron development, IEGs and LRGs are transcriptionally induced by sensory experience, which drives neuronal and circuit maturation[6,8]. In the mouse somatosensory system, topographic representations of the mystacial vibrissae (whiskers) on the face are generated at brainstem, thalamus and cortical levels[13,14]. In the brainstem, the whisker-related neuronal modules, or barrelettes, are generated in the ventral principal trigeminal sensory nucleus (vPrV), and sensory neuronal activity is required at perinatal/early postnatal stages for the maturation of barrelette neuron connectivity and map formation[13,14].

To characterize IEG and LRG activity-response genes (ARGs) in developing barrelette neurons, we set out a genetic strategy to isolate

[1]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. [2]Swiss Institute of Bioinformatics, Basel, Switzerland. [3]University of Basel, Basel, Switzerland. [4]Instituto de Neurociencias de Alicante, Universidad Miguel Hernández-Consejo Superior de Investigaciones Científicas (UMH-CSIC), Sant Joan d'Alacant, Spain. [5]Present address: Science for Life Laboratory, Solna, Sweden. [6]These authors contributed equally: Taro Kitazawa, Dania Machlab. ✉e-mail: filippo.rijli@fmi.ch

embryonic day (E) 10.5 mitotic progenitors and postmitotic barrelette neurons at E14.5 (early postmitotic), E18.5 (perinatal) and postnatal day (P) 4 (consolidated barrelette stage) by FACS, and we profiled them by mRNA sequencing (mRNA-seq; Smart-seq2) and chromatin immunoprecipitation followed by sequencing (ChIP–seq; ChIPmentation) of the Pc-dependent repressive H3K27me3 and active H3K4me2 and H3K27ac histone modifications, and chromatin accessibility by an assay for transposase-accessible chromatin followed by sequencing (ATAC–seq; Methods, Fig. 1a, Extended Data Fig. 1a–e, Supplementary Table 1, Supplementary Figs. 1 and 2 and Supplementary Note).

To identify ARGs induced in barrelette neurons at the beginning of the sensory-dependent maturation period (E18.5–P2/P3)[15], we collected E18.5 *Kir2.1*-overexpressing, activity-deprived, vPrV postmitotic barrelette neurons by FACS sorting (Extended Data Fig. 1e–p, Supplementary Table 1, Supplementary Fig. 2, Supplementary Note and Methods), profiled them by mRNA-seq (Smart-seq2), and compared to E14.5 and E18.5 vPrV wild-type barrelette neurons. Among the genes with undetectable or low basal expression level (reads per kilobase per million mapped reads (RPKM) < 3) in E14.5 barrelette neurons, we identified 56 genes, referred to as barrelette sensory ARGs (bsARGs; Supplementary Table 2), which were upregulated at E18.5 in a neuronal activity-dependent manner (Extended Data Fig. 1q–s, Supplementary Note and Methods). Barrelette sensory ARGs comprised 4 IEGs, namely *Fos*, *Egr1*, *Junb* and *Zfp36* (Fig. 1b and Supplementary Table 2), and at least 23 putative LRGs (for example, *Cd38* and *Osmr*; Supplementary Table 3). We next identified additional ARGs referred to as non-barrelette ARGs (nbARGs; $n = 83$; Methods and Supplementary Note), which included both LRGs and 12 IEGs and were transcriptionally induced by distinct activity-dependent stimuli in neuronal types other than barrelette neurons[16–18], but that displayed undetectable or low basal expression level (RPKM < 3) in E14.5, E18.5 and P4 barrelette neurons.

Pc group proteins regulate dynamics and plasticity of gene expression during development[19–23]. We found that in E14.5 barrelette neurons, 32/56 (57%) and 67/83 (84%) of bsARGs and nbARGs, respectively, were embedded in H3K27me3$^+$ domains of Pc-repressive chromatin (Methods) with, however, H3K4me2$^+$/ATAC$^+$ promoters (Fig. 1c and Extended Data Fig. 1t,u).

Immediate early genes carry a unique Polycomb bipartite signature during development. The bsARGs and nbARGs with a H3K27me3$^+$/H3K4me2$^+$/ATAC$^+$ Pc chromatin profile at E14.5 included all 16 IEGs: *Fos*, *Egr1*, *Egr3*, *Egr4*, *Fosb*, *Fosl2*, *Junb*, *Zfp36*, *Klf4*, *Maff*, *Npas4*, *Nr4a3*, *Apold1*, *Atf3*, *Dusp5* and *Arc*[16–18]. In analysis of their chromatin profile, only 4 of 16 (25%) IEGs (*Junb*, *Egr3*, *Egr4* and *Atf3*) displayed a conventional Pc bivalent[24–26] signature (Fig. 1d), that is, with promoters marked by both active H3K4me2 and repressive H3K27me3 histone modifications. Interestingly, 12 of 16 (75%) IEGs displayed a unique distinct 'bipartite' Pc signature (Fig. 1d; see genome browser snapshots at *Fos*, *Egr1*, *Fosb* and *Nr4a3* in Fig. 1e; Extended Data Fig. 2a). Namely, H3K27me3 deposition was restricted to their gene bodies, whereas the accessible H3K4me2$^+$ promoters were devoid of H3K27me3 and decorated instead with the active mark H3K27ac, notably with no or only low basal levels of detected mRNA. H3K27me3 on gene bodies did not stretch further than 2–3 kb downstream of the transcription start site (TSS), even when the gene was longer (for example, *Nr4a3*; Extended Data Fig. 2a). H3K27ac deposition at promoters of bipartite IEGs was not induced by the dissociation procedure (Extended Data Fig. 2b and Methods). Conversely, we found that among the remaining 83 of 99 H3K4me2$^+$/H3K27me3$^+$/ATAC$^+$ ARGs, which included putative barrelette neuron LRGs and non-barrelette neuron LRGs[16–18] (for example, *Osmr* and *Pdlim1*, respectively; Fig. 1e), 66 of 83 (80%) were in a bivalent state, whereas only 17 of 83 (20%) carried the bipartite Pc signature (Fig. 1d and Methods).

In summary, at prenatal stages, the rapidly inducible IEGs are preferentially in a Pc bipartite state, while the LRGs are preferentially enriched with a Pc bivalent signature (Fig. 1d,e). Similarly to developing barrelette neurons, the Pc bipartite signature was also present at IEGs in prenatal cortical progenitors and postmitotic neurons, although neither in adult excitatory neurons nor in 7-d cultured embryonic cortical neurons (Extended Data Fig. 2a). Thus, the bipartite chromatin organization is specifically established at IEGs during prenatal neuronal development.

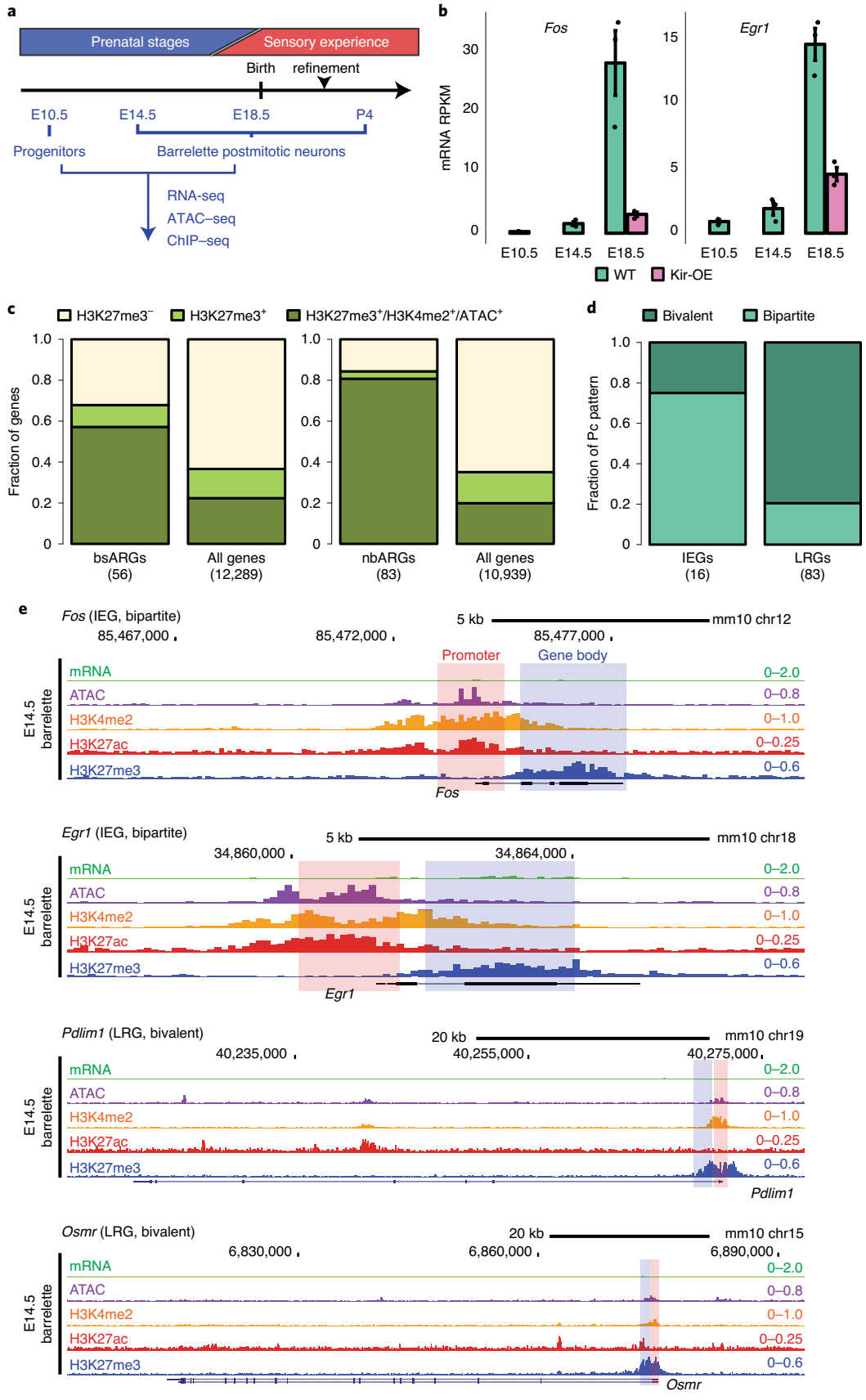
The bipartite signature is found on stimulus-response genes during development and is not restricted to neurons. We next investigated the genome-wide distribution of the Pc bipartite chromatin signature. We assigned each gene with a 'bipartiteness' score related to their promoter H3K27ac and gene body H3K27me3 levels and a 'bivalency' score related to H3K27me3 and H3K4me2 at promoters (Methods and Extended Data Fig. 3a,b). Considering the estimated false-positive rates of this scoring approach, we conservatively evaluated the total numbers of true bipartite genes from at least 140 at E10.5, to 177 at E14.5, to 219 at E18.5 and decreasing to 113 at P4 in barrelette neurons (Fig. 2a and Methods). At all stages, approximately 1,500 genes were instead in a bivalent state (Fig. 2a and Methods). Aggregate profile plots of chromatin marks of the top 100 E14.5 barrelette neuron genes scored as bipartite (E14.5Bip) or bivalent (E14.5Biv; Methods) further confirmed their clearly distinct chromatin signatures (Fig. 2b and Extended Data Fig. 3c).

In addition to IEGs, Gene Ontology analysis of E14.5Bip genes identified genes encoding transcriptional regulators and transmembrane domain receptors responding to distinct signaling pathways including bone morphogenetic protein and transforming growth factor-β signaling, voltage-gated ion channels, and dendritic, axonal and synaptic genes (Fig. 2c and Supplementary Table 4).

Furthermore, by our ranking method, we additionally found 124, 99, 185 and 107 genes carrying the bipartite chromatin signature in mouse E14.5 heart tissue, E14.5 liver tissue, E10.5 cranial neural-crest-derived cells (NCCs) and ESCs, respectively (Fig. 2a–c, Extended Data Fig. 3d and Supplementary Table 4). Bipartite genes are tissue- and stage-specific as only a few bipartite genes were shared among the different cell types (Fig. 2d), including typical IEGs (for example, *Fos*, *Jun*, *Fosl2*, *Myb*, *Egr2* and *Arc*; Fig. 2c). Nonetheless, bipartite genes appear to be consistently 5–15% of the bivalent genes at all times and in all the distinct cell types analyzed (Fig. 2a).

Fig. 1 | Immediate early- and late-response genes carry distinct Polycomb signatures during sensory neuron development. a, Cell isolation strategy diagram. E10.5 mitotic progenitors and E14.5 (early postmitotic), E18.5 (perinatal) and P4 (consolidated barrelette stage) barrelette neurons were isolated and collected as indicated. **b**, *Fos* (left) and *Egr1* (right) IEG mRNA RPKM expression bar plots in E10.5 progenitors (*K20$^{tdTomato/+}$*), E14.5 and E18.5 postmitotic *Drg11$^{vPrV\text{-}ZsGreen/+}$* barrelette (wild-type (WT); green bars) and activity-depleted Kir2.1-overexpressing E18.5 *Drg11$^{vPrV\text{-}Kir}$* barrelette neurons (Kir-OE; purple bars; $n = 3$ biologically independent littermates). Bar plots represent average values and error bars show the standard deviation. **c**, Bar plots showing the fraction of Pc target gene profiles in E14.5 *Drg11$^{vPrV\text{-}ZsGreen/+}$* barrelette neurons. Left: fraction of Pc target profiles at E14.5 among the 56 bsARGs compared to all other $n = 12{,}289$ genes (RPKM < 3 at E14.5); right: fraction of Pc target profiles at E14.5 among 83 nbARGs compared to all other $n = 10{,}939$ genes (RPKM < 3 at E14.5, E18.5 and P4) (Supplementary Note and Methods). **d**, Bar plots showing the fractions of Pc bivalent or bipartite chromatin signatures among the 16 IEGs (bsARGS + nbARGs; left) and 83 LRGs (right) in E14.5 *Drg11$^{vPrV\text{-}ZsGreen/+}$* barrelette neurons. **e**, Genome browser profiles of representative bipartite IEGs (*Fos* and *Egr1*) and bivalent LRGs (*Pdlim1* and *Osmr*), in E14.5 *Drg11$^{vPrV\text{-}ZsGreen/+}$* barrelette neurons.

Lastly, sequential ChIP–seq on E14.5 bulk hindbrain tissue and single-cell mRNA-seq (scRNA-seq; 10x Genomics) analysis of FACS-isolated E14.5 postmitotic barrelette neurons and E10.5 progenitors demonstrated that the H3K27ac and H3K27me3 histone marks coexist at the promoter and gene body of bipartite genes, correlating with low or undetectable mRNA



transcription (Supplementary Note, Fig. 2e,f and Extended Data Figs. 3e and 4a–e).

These results show that the bipartite signature is not an exclusive feature of developing neurons but is widely used during development, raising the intriguing possibility that it could regulate rapid IEG transcriptional inducibility.

The bipartite signature originates from bivalent chromatin and is dynamic during development. To investigate how the bipartite signature is established, maintained and resolved during development, we created a two-dimensional (2D) projection of autosomal genes according to chromatin accessibility, H3K27me3, H3K4me2 and H3K27ac levels at promoters and gene bodies (Extended Data Fig. 5a) using *t*-distributed stochastic neighbor embedding (*t*-SNE; Fig. 3a–d, Extended Data Fig. 5b–l, Supplementary Note and Methods). We generated a single map for E10.5 progenitors and a combined E14.5, E18.5 and P4 *t*-SNE map of chromatin states for postmitotic barrelette neurons (Fig. 3a,b, Extended Data Fig. 5b,c and Methods). Genes with similar chromatin patterns were grouped together, which also correlated with mRNA-seq data (Extended Data Fig. 5c–e).

Top-scoring bipartite and bivalent genes at E10.5 and postmitotic stages mapped to distinct, largely nonoverlapping, regions on the respective *t*-SNE maps (Fig. 3a–d, Extended Data Fig. 5f,g,i–l, Supplementary Note and Methods). Furthermore, genes mapping to the same region of the combined E14.5/E18.5/P4 *t*-SNE map revealed a stable chromatin state, unlike genes changing their localization between developmental stages (Fig. 3a–d, Extended Data Fig. 5h and Supplementary Note).

At P4, distinct fractions of the E14.5Bip genes had transitioned into productive transcription (Bip → Exp; RPKM > 3), bivalency (Bip → Biv) or remained bipartite (Bip → Bip; Extended Data Fig. 6a). As compared to E14.5, Bip → Exp genes displayed higher levels of H3K27ac, increased accessibility (ATAC–seq) and mRNA levels, and decreased H3K27me3, in contrast to genes that remained bipartite (Bip → Bip) or became bivalent (Bip → Biv; Extended Data Fig. 6a). The developmental progression through distinct bipartite patterns and into the active chromatin state of E14.5Bip genes could also be readily visualized as relocation of their position on the E10.5, E14.5 and P4 *t*-SNE plots (Fig. 3c); representative examples include *Fos*, *Egr1* and *Bcl6* (involved in postmitotic neuronal fate through repression of Wnt/Notch/Fgf/Shh[27]), *Nr3c1* (glucocorticoid receptor) and *Plekhh3* (signal transduction in axon growth; Fig. 2c), while Figure 3d shows the fraction of E14.5Bip genes that switched to bivalency at P4 (Bip → Biv). Genome browser views of *Fos* and *Egr1* (Bip → Exp) and *Gpr88* (Bip → Biv) confirmed the transcriptional and epigenetic changes (Fig. 3e,f and Extended Data Fig. 6b). Moreover, by using circular chromosome conformation capture, coupled to high-throughput sequencing (4C–seq), we found that the bipartite signature at the *Fos* locus allowed for reciprocal physical contacts between its active enhancers and promoter, irrespective of productive transcription (Supplementary Note, Fig. 3e and Extended Data Fig. 6c).

Next, we investigated the developmental origin of the bipartite signature. At E10.5, about 50% of E14.5Bip genes were already in a bipartite state, as they mapped within the green contour region of the E10.5 *t*-SNE plot; however, as much as 40% of E14.5Bip genes were in a bivalent state in E10.5 progenitors, as they were contained within the bivalent red contour region (Fig. 3b; see Fig. 3f for a representative example, *Gpr88*). While bipartite and bivalent genes had similar CpG content and distribution (Extended Data Fig. 3f), the E14.5Bip promoters were enriched in nuclear factor-κB-related and forkhead box-related factor binding motifs (Extended Data Fig. 3g and Methods).

Thus, the bipartite state originates from bivalent chromatin in early progenitors and, during postmitotic neuron development, displays bidirectional dynamics, reverting into a bivalent state for a subset of genes, or resolving into productive transcription.

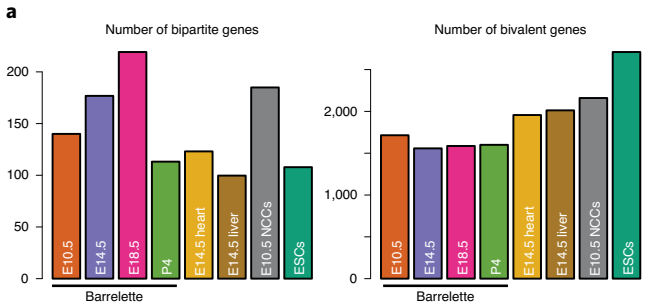
RNA polymerase II transcripts of bipartite genes are not efficiently processed to productive mRNA. We next investigated additional chromatin features of bipartite genes (Fig. 4a and Supplementary Note).

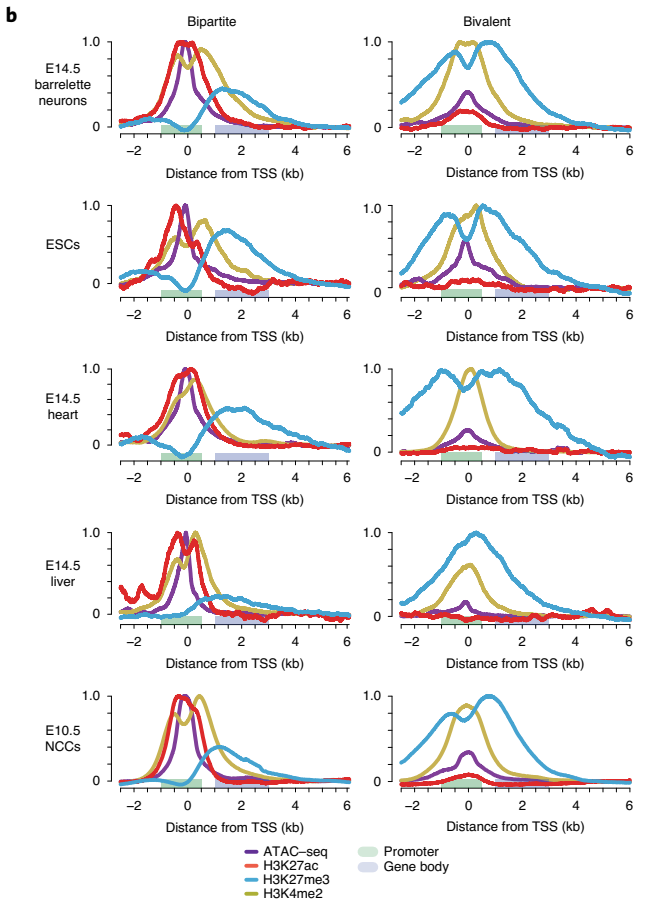
Moreover, E14.5Bip genes displayed dramatically lower productive mRNA levels than E14.5 non-bipartite genes with Bip-matching promoter H3K27ac levels (E14.5AcP; Fig. 4a, mRNA). To investigate why active bipartite promoters did not drive higher levels of productive transcription, we determined ChIP enrichment for distinct phosphorylated forms of the main subunit of RNAPII. The RNAPII C-terminal domain changes its serine phosphorylation pattern as RNAPII progresses from initiation (S5P) through productive transcription and elongation (S7P and S2P)[28–30]. Transcriptionally productive and elongating RNAPII-S5P^{+}S7P^{+}S2P^{+} is detected at active genes, whereas not productively elongating RNAPII-S5P^{+}S7P^{-}S2P^{-}, also little or not recognized by the 8WG16 antibody, is detected at Pc-repressed bivalent genes[29,31,32].

We found a unique pattern of RNAPII at E14.5Bip genes. Namely, 8WG16, RNAPII-S5P and RNAPII-S7P levels at E14.5Bip promoters were similar to those at E14.5AcP promoters, and higher than non-Bip genes with low, Bip-matching, levels of productive mRNA transcription (E14.5mRNALow; Methods) and E14.5Biv promoters (Fig. 4a,b). In contrast, around the E14.5Bip transcription end sites (TESs; Methods), the levels of RNAPII-S2P and H3K36me3, a mark of productive mRNA elongation into gene bodies[28], were significantly lower than those in E14.5AcP genes, although higher than those in E14.5Biv genes and comparable to those in E14.5mRNALow genes (Fig. 4b). Genome browser views of E14.5Bip *Fos* and *Egr1* loci confirmed that both RNAPII-S5P and RNAPII-S7P paused at the promoter–proximal first exon regions, while RNAPII-S7P and

Fig. 2 | Developmental cellular representation and genome-wide distribution of bipartite chromatin, and promoter H3K27ac and gene body H3K27me3 coexistence at bipartite genes. a, Estimated numbers of bipartite and bivalent genes in E10.5 *K20*$^{tdTomato/+}$ progenitors, E14.5, E18.5 and P4 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons, E14.5 mouse heart tissue, E14.5 mouse liver tissue, E10.5 NCCs and mouse ESCs. **b**, Aggregate plots of chromatin features (ATAC–seq and ChIP–seq, as indicated) around the TSS of bipartite and bivalent genes in the distinct developing cell types, as indicated. Promoters and gene bodies are highlighted; *y* axes of bipartite and bivalent plots were scaled to allow direct comparison of the same mark between plots (Methods). **c**, Gene Ontology of bipartite genes identified in the different developing cell types. MAPK, mitogen-activated protein kinase; TGF, transforming growth factor. **d**, UpSet plot showing intersections among bipartite genes in different developing cell types. Bipartite genes are mostly tissue- and stage-specific with only a few shared. In **b**–**d**, the 100 top-scoring genes for bipartiteness were used, as a conservative definition of bipartite genes. **e**, Genome browser of bipartite *Fos* displaying accessibility (ATAC–seq), 2–3-kb fragment H3K27ac and H3K27me3 single ChIP–seq, and H3K27me3/H3K27ac sequential ChIP–seq from E14.5 hindbrain. H3K27me3 and H3K27ac coexist on gene body and promoter, respectively (Supplementary Note). **f**, Violin plots displaying promoter H3K27ac, bulk mRNA-seq levels (Smart-seq2) and single-cell fractions with detected mRNA transcripts (10x Genomics) in E14.5 barrelette neurons, E14.5Bip genes ($n = 97$) and E14.5 non-bipartite genes with Bip-matching promoter H3K27ac levels (E14.5AcP genes; $n = 97$) are compared (Supplementary Note and Methods). Plots extend from the data minima to the maxima; the white dot indicates the median, the box shows the interquartile range and whiskers extend to the most extreme data point within 1.5 times the interquartile range. *P* values are from two-sided Wilcoxon's tests. NS, not significant ($P > 0.05$).

RNAPII-S2P levels were barely detectable in the H3K27me3^{+} gene body regions (Extended Data Fig. 6d). This distribution is generally shared by E14.5Bip genes (Fig. 4). In addition, total RNA analysis (Ovation SoLo RNA-seq; Methods) showed that E14.5Bip nascent RNA transcripts were not efficiently processed to productive mRNA (Extended Data Fig. 7a,b).





c

Bipartite genes in E14.5 barrelette neurons (E14.5Bip genes)

Term	*P* value	Example of genes
Regulation of transcription, DNA-templated	$P = 1.81 \times 10^{-3}$	*Egr1, Zbtb16, Nr6a1, Nr3c1, Npas4, Nr4a3, Fos, Eya1, Rel, Bcl6, Myb, Klf4*
TGF-β receptor signaling pathway	$P = 4.36 \times 10^{-2}$	*Fos, Bambi*
Voltage-gated ion channel activity	$P = 2.68 \times 10^{-6}$	*Kcnc2, Kcnj6, Kcna1*
Synapse	$P = 6.37 \times 10^{-5}$	*Kcnc2, Myrip, Unc5c*
Signal transduction	$P = 4.96 \times 10^{-2}$	*Fzd8, Gpr158, Plekhh3, Camk4, Unc5c, Grem1, Gpr88*

Bipartite genes in ESCs

Term	*P* value	Example of genes
Positive regulation of transcription, DNA-templated	$P = 1.28 \times 10^{-3}$	*Fos, Fosl2, Jun, Myb, Zfhx3*
Response to cAMP	$P = 2.30 \times 10^{-2}$	*Fos, Jun, Vgf*
Negative regulation of Notch signaling pathway	$P = 8.88 \times 10^{-3}$	*Gdpd5, Bcl6*

Bipartite genes in E14.5 heart

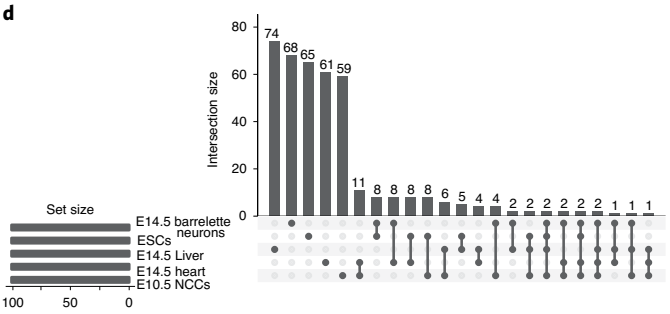
Term	*P* value	Example of genes
Positive regulation of transcription, DNA-templated	$P = 8.75 \times 10^{-5}$	*Egr2, Nr4a3, Rorb, Gli2*
Membrane	$P = 1.45 \times 10^{-4}$	*Arc, Pigz, Gli2, Btbd11*
Steroid hormone-mediated signaling pathway	$P = 2.63 \times 10^{-2}$	*Nr4a3, Rorb*

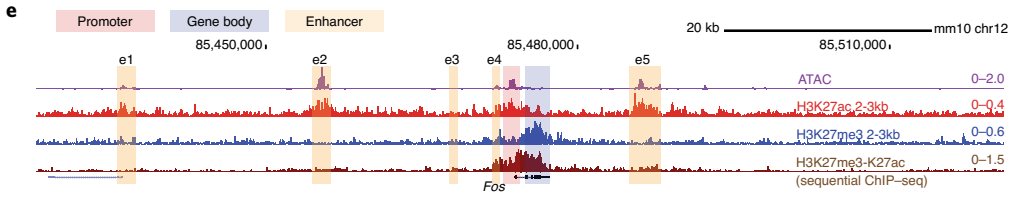
Bipartite genes in E14.5 liver

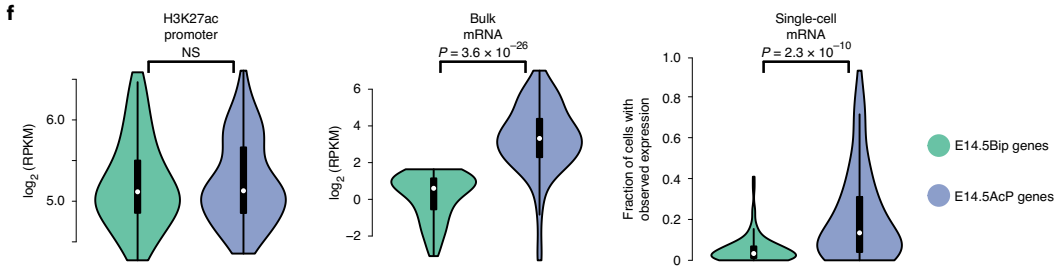
Term	*P* value	Example of genes
Regulation of transcription, DNA-templated	$P = 8.09 \times 10^{-4}$	*Fosb, Nr4a3, Zfp318, Fos*
cAMP signaling pathway	$P = 2.79 \times 10^{-2}$	*Fos, Camk2d, Pik3r3*

Bipartite genes in E10.5 NCCs

Term	*P* value	Example of genes
Positive regulation of transcription, DNA-templated	$P = 4.55 \times 10^{-3}$	*Egr2, Crebl2, Fos, Klf4, Arntl*
MAPK signaling pathway	$P = 1.22 \times 10^{-2}$	*Dusp10, Prkca, Mapt, Fos*







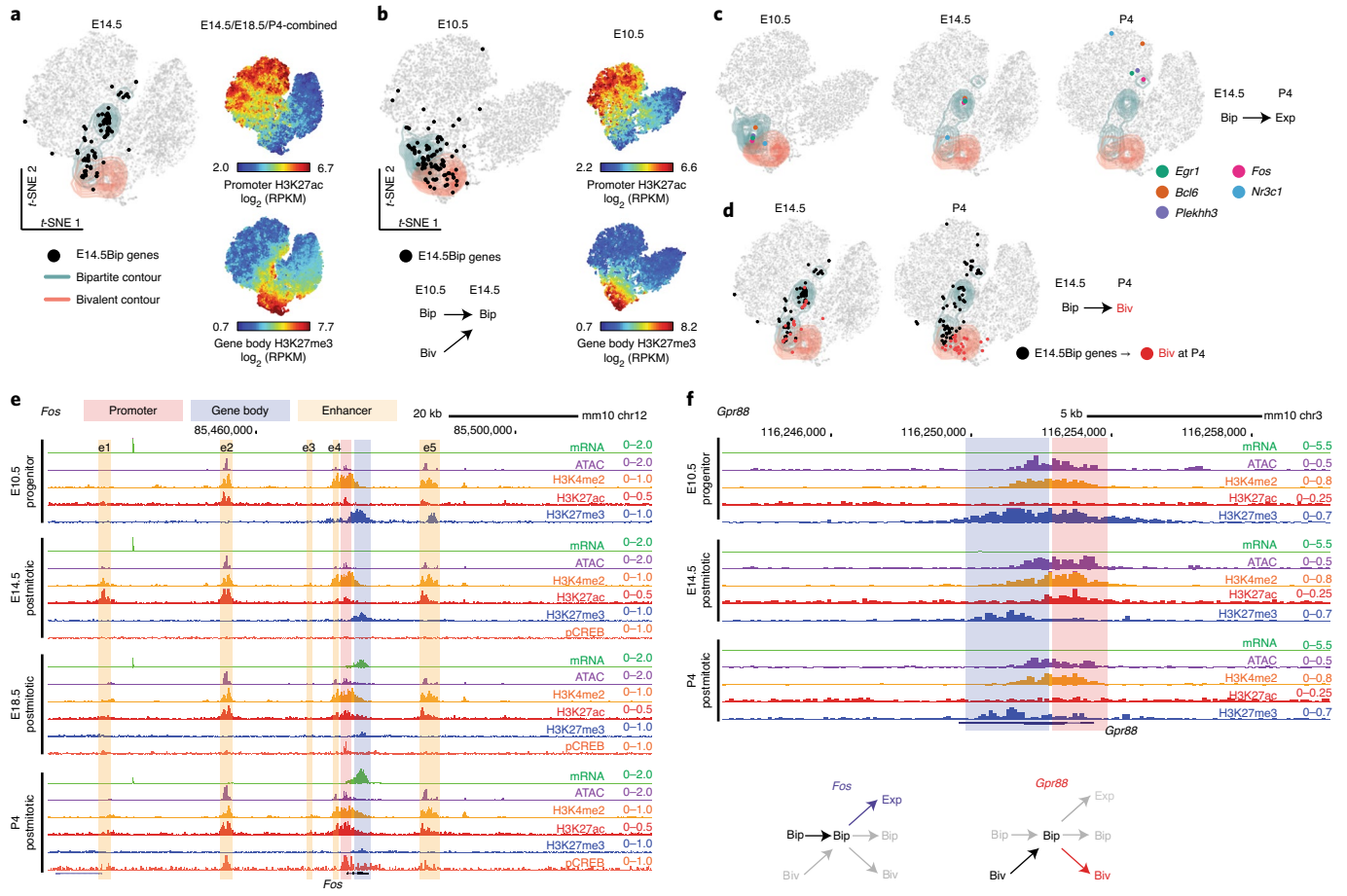


Fig. 3 | Bipartite chromatin dynamics during barrelette neuron development. a, A 2D projection of all autosomal genes (dots) visualized in E14.5/E18.5/P4-combined *t*-SNE maps, according to the chromatin profiles of *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons at E14.5 (Extended Data Fig. 5b,c, Methods and Supplementary Note). Contour lines depict *t*-SNE regions enriched for bipartite or bivalent genes. E14.5Bip genes are mapped as black dots. Color-coded *t*-SNE plots: each gene is labeled according to promoter H3K27ac and gene body H3K27me3 signal levels. **b**, *t*-SNE plot of E10.5 *K20*$^{tdTomato/+}$ progenitor gene chromatin profiles. Black dots: distribution at E10.5 of E14.5Bip genes. **c**,**d**, Stage-specific developmental dynamics of E14.5Bip gene chromatin profiles on single (E10.5) and E14.5/E18.5/P4-combined (E14.5 and P4) *t*-SNE maps (Extended Data Fig. 6a). In **c**, five representative E14.5Bip genes, *Egr1*, *Fos*, *Bcl6*, *Nr3c1* and *Plekhh3*, which became expressed (Exp; mRNA RPKM ≥ 3) at P4, show their dynamic 'movement' (change of chromatin state) between E10.5 progenitors, and E14.5 and P4 postmitotic neurons. In **d**, *t*-SNE map shows the change of spatial distribution between E14.5 and P4 of E14.5Bip genes (black dots) and E14.5 and P4 distributions of E14.5Bip genes switching to bivalent (Biv) at P4 (red dots). **e**,**f**, Genome browser views of chromatin and transcriptional states of *Fos* and *Gpr88* in E10.5 *K20*$^{tdTomato/+}$ progenitors and E14.5, E18.5 and P4 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons. Bottom right: summary diagram of chromatin state developmental transitions at *Fos* and *Gpr88*. e1-e5, activity-dependent *Fos* enhancers.

In summary, mRNA processivity of E14.5Bip genes is intermediate between bivalent (E14.5Biv) genes and genes with comparable H3K27ac promoter levels (E14.5AcP). We also demonstrate that mRNA elongation through the gene bodies of E14.5Bip genes is maintained at a low rate, in line with E14.5RNALow genes (Fig. 4a,b), despite having similar promoter H3K27ac and RNAPII-S5P levels to E14.5AcP genes.

Polycomb marking of bipartite genes on gene bodies inhibits productive mRNA processing. Little is known about a potential role of Pc on gene bodies[22,23,33,34]. We conditionally inactivated *Ezh2* (ref. [35]), which catalyzes H3K27me3 deposition, in mouse rhombomere 3 (r3) hindbrain derivatives, enriched in vPrV barrelette progenitors and postmitotic neurons (*Ezh2*-cKO$^{r3\text{-}RFP}$, Supplementary Table 1 and Supplementary Figs. 1 and 3). In control FACS-isolated E14.5 r3-derivatives, E14.5Bip barrelette neuron genes were in a bipartite state (Extended Data Fig. 3e and Supplementary Note), whereas in *Ezh2*-cKO$^{r3\text{-}RFP}$ homozygous mutant cells, the H3K27me3 mark on E14.5Bip gene bodies was strongly reduced and replaced by the H3K27ac mark (Fig. 5a; see below). Productive mRNA transcription of E14.5Bip genes was significantly increased in *Ezh2*-cKO$^{r3\text{-}RFP}$ mutant cells (Fig. 5b). Total RNA-seq analysis indicated that nascent E14.5Bip transcripts were more efficiently processed to productive spliced mRNA in *Ezh2*-cKO$^{r3\text{-}RFP}$ mutant cells than controls (Extended Data Fig. 7c). Moreover, accumulation of reads at gene TSS proximal regions was reduced in mutant compared to wild-type cells (Fig. 5c and Extended Data Fig. 7d). Moreover, likely as a direct result of ectopic Fos induction, 85 activity-regulated Fos-binding enhancers[4] (Methods), which normally became open only in postnatal barrelette neurons, gained precocious accessibility in E14.5 FACS-isolated *Ezh2* homozygous mutant neurons from bulk hindbrain (*Ezh2*-cKO$^{HB\text{-}RFP}$; Supplementary Fig. 3, Supplementary Table 1 and Methods), suggesting incorrect precocious activation of an early postnatal Fos-driven enhancer program (Fig. 5d, Extended Data Fig. 8a,b and Supplementary Note).

Next, we investigated the levels and distribution of elongation marks in *Ezh2* mutants. H3K36me3 levels were increased at E14.5Bip genes in *Ezh2*-cKO$^{HB\text{-}RFP}$ mutant cells, as compared to wild-type cells

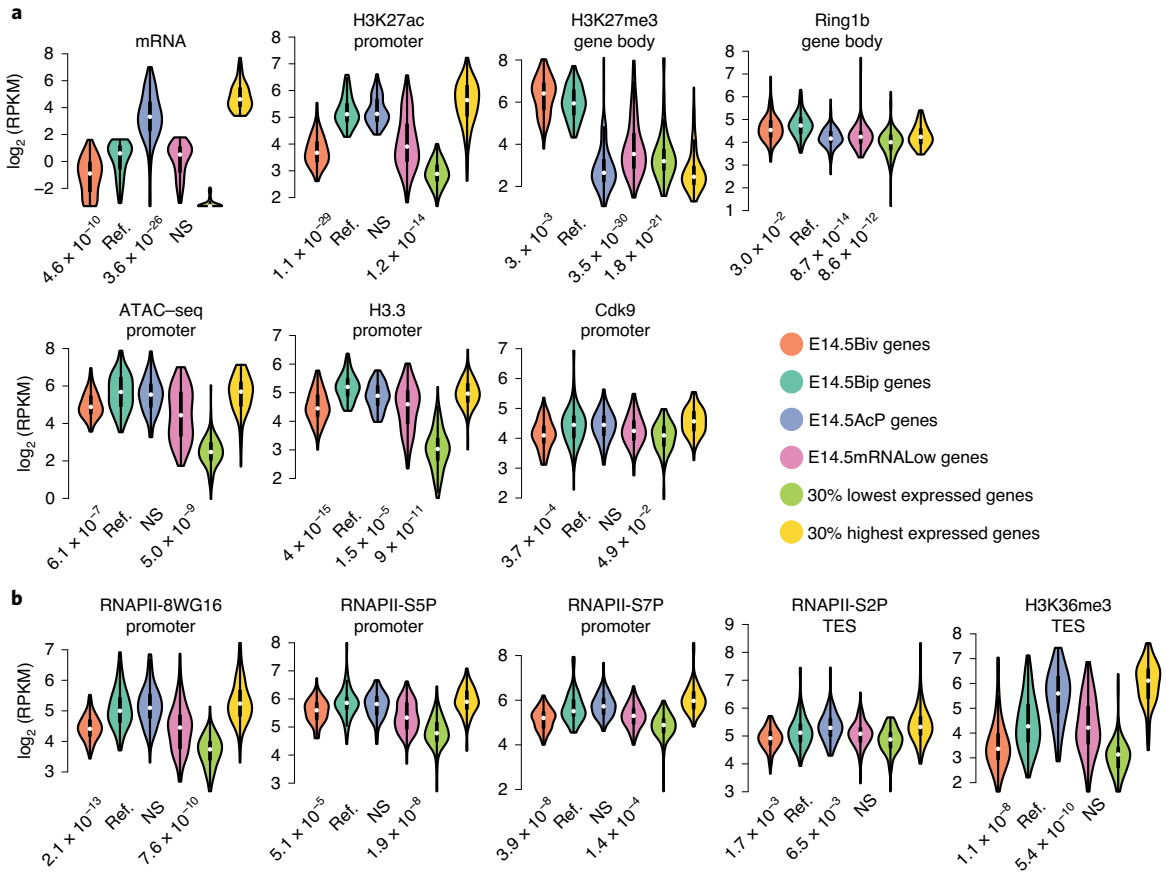


Fig. 4 | RNA polymerase II transcripts of bipartite genes are not efficiently processed to productive mRNA. a,**b**, Violin plots of RNA expression (RNA-seq) and chromatin features (ATAC–seq and ChIP–seq) of bivalent, bipartite and control gene sets. Comparison of *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neuron E14.5Biv genes, E14.5Bip genes, E14.5AcP genes, E14.5mRNALow genes, and the 30% lowest and 30% highest expressed genes illustrating the maximal signal range (Methods). Each gene set contains $n = 97$ genes (Methods). Plots extend from the data minima to the maxima, the white dot indicates the median, the box shows the interquartile range and whiskers extend to the most extreme data point within 1.5 times the interquartile range. *P* values are from two-sided Wilcoxon's tests between each gene set and the E14.5Bip gene set, labeled as 'ref' for reference.

(Extended Data Fig. 7e). To overcome the unfeasibility of obtaining large amounts of cells from *Ezh2*-cKO embryos, we used *Eed*^{KO} mouse ESCs in which the H3K27me3 mark is removed genome wide[36]. We carried out RNAPII-S2P ChIP–seq and mRNA-seq in wild-type and *Eed*^{KO} ESCs. For genes carrying H3K27me3 in gene bodies, upregulation of mRNA levels in *Eed*^{KO} cells correlated with a modest but significant increase of RNAPII-S2P signals in the TES region, compared with wild-type ESCs (Extended Data Fig. 7f,g and Supplementary Note). We then analyzed the transcriptional upregulation of bipartite genes in full *Ezh1*-KO;*Ezh2*-KO and Ezh2 catalytically inactive *Ezh1*-KO;*Ezh2*^{Y726D} mutant ESCs[37] and found that the H3K27me3 mark itself on the gene body, rather than recruitment of Pc proteins, was required for the inhibition of bipartite gene productive transcription (Supplementary Note and Extended Data Fig. 7h). Taken together, these results indicate that the Pc-dependent H3K27me3 marking of the gene bodies of bipartite genes inhibits productive mRNA elongation.

To further support these findings, we selectively depleted the H3K27me3 mark from specific bipartite gene bodies and analyzed its acute effect on productive mRNA transcription. We developed an ex vivo short-term culture of E12.5 neurons from bulk hindbrain tissue; in this system, we observed no H3K27me3 depletion from bipartite gene bodies normally observed in long-term (1 week) hindbrain and cortical neuron embryonic cultures[38] (Extended Data Figs. 2a and 9a). Overexpression of the catalytically 'dead' Cas9 (dCas9) fused to the H3K27me3 demethylase UTX (Kdm6a; dCas9–UTX) resulted in the selective decrease of H3K27me3 from the bipartite gene body (that is, *Fos*; Extended Data Fig. 9b–d). Quantification of mRNA levels confirmed that dCas9–UTX targeted to gene bodies of bipartite genes (*Fos* and *Egr1*) caused significant transcriptional upregulation of these genes (Fig. 5e), whereas dCas9–UTX targeted to non-bipartite gene bodies (*Actb* and *Gapdh*) did not affect gene expression (Extended Data Fig. 9e).

Together, these results indicate that the H3K27me3 histone mark on gene bodies of bipartite genes interferes with the production and accumulation of mature mRNA from the bipartite active promoters.

The bipartite signature regulates the rapidity and amplitude of transcriptional response to stimuli. Next, we asked whether the bipartite state might still allow rapid stimulus-dependent inducibility of IEGs, and whether bipartite or bivalent IEGs would display distinct transcriptional responses. We FACS-isolated cells from E14.5 hindbrain bulk tissue and treated them with 55 mM KCl for 8 or 30 min. KCl-mediated depolarization of cultured neurons results in an increase of intracellular calcium signaling and phosphorylation of CREB, a readout of stimulus-dependent transcription, at IEG promoters and is widely used to mimic the transcriptional response to a wide range of sensory stimuli[2,4,17,38]. While an 8-min KCl treatment caused rapid induction of the bipartite *Fos* and *Egr1* IEGs, the bivalent *Junb* IEG (Extended Data Fig. 10a) was not induced;

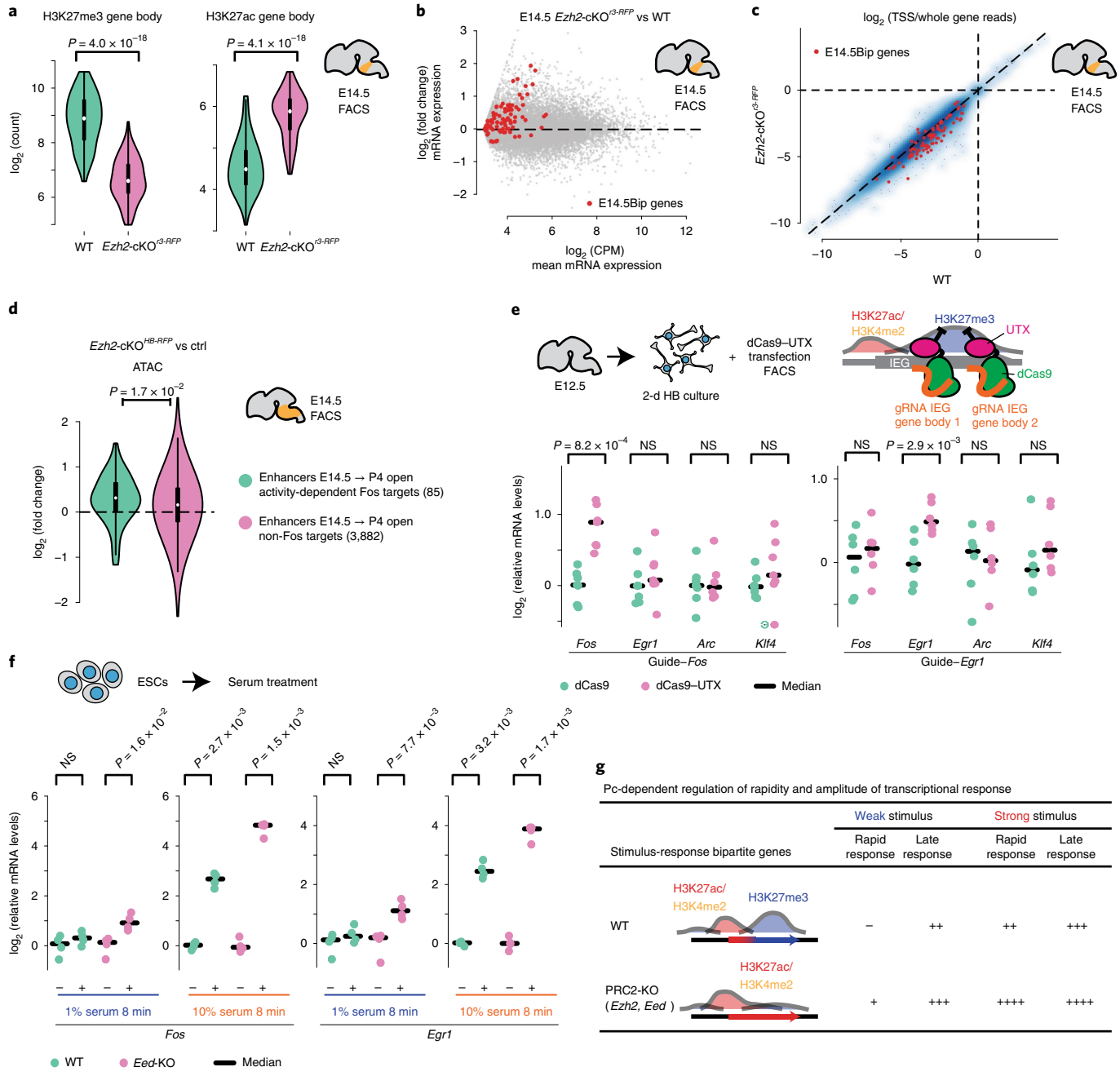


Fig. 5 | Polycomb marking of gene bodies inhibits productive mRNA processing and regulates rapidity and amplitude of transcriptional response to stimuli. a, Violin plots of gene body H3K27me3 and H3K27ac levels of bipartite genes (E14.5Bip; $n=100$; Methods) in control $K20^{tdTomato/+}$ (WT; green) and *Ezh2*-cKO$^{r3\text{-}RFP}$ (purple) conditional knockout (cKO) E14.5 hindbrain cells. **b**, MA plot comparing productive Smart-seq2 mRNA levels of E14.5Bip genes (red dots; $n=100$) between E14.5 *Ezh2*-cKO$^{r3\text{-}RFP}$ and WT hindbrain cells. CPM, counts per million. **c**, Fractions of total RNA-seq reads at gene TSS proximal regions (Methods) in E14.5 WT compared to *Ezh2*-cKO$^{r3\text{-}RFP}$ hindbrain cells. A blue smooth scatter represents the density of all genes. E14.5Bip genes ($n=82$ genes; genes with CPM ≤ 1 are not shown) show improved elongation in *Ezh2*-cKO$^{r3\text{-}RFP}$ versus WT (Extended Data Fig. 7d). **d**, Violin plots, show $\log_2$ fold changes of enhancer chromatin accessibilities in E14.5 *Ezh2*-cKO (*Ezh2*-cKO$^{HB\text{-}RFP}$) homozygous mutant neurons as compared to heterozygous control neurons; activity-dependent Fos-binding enhancers, normally open only at P4 ($n=85$), gain precocious accessibility when compared to all the remaining enhancers only open at P4 ($n=3{,}882$; Methods). **e**, *Fos*, *Egr1*, *Arc* and *Klf4* mRNA levels were measured by quantitative PCR with reverse transcription (RT-qPCR) in short-term cultured E12.5 hindbrain (HB) neurons overexpressing control dCas9 or dCas9-UTX targeted to *Fos* ($n=7$ biologically independent neuron cultures) or *Egr1* ($n=6$ biologically independent neuron cultures) gene bodies. gRNA, guide RNA. **f**, *Fos* and *Egr1* mRNA levels by RT-qPCR in serum-starved WT and *Eed*^{KO} mouse ESCs treated with a low (1%) or high (10%) concentration of FCS for 8 min ($n=4$ biologically independent cultured cells). Treatment with 1% FCS was not sufficient to induce rapid transcriptional responses in WT, unlike in *Eed*^{KO}, ESCs; while 10% FCS induced rapid transcriptional responses in both WT and mutant, although levels were higher in *Eed*^{KO} ESCs. **g**, Summary of Pc-dependent regulation of rapidity and amplitude of bipartite gene transcriptional response to environmental stimuli with distinct strengths. In **a** and **d**, plots extend from the data minima to the maxima, the white dot indicates the median, the box shows the interquartile range and whiskers extend to the most extreme data point within 1.5 times the interquartile range. *P* values are from paired two-sided Wilcoxon's tests. In **e** and **f**, the median expression is indicated by bars. *P* values are from Welch's two-sample two-sided *t*-tests.

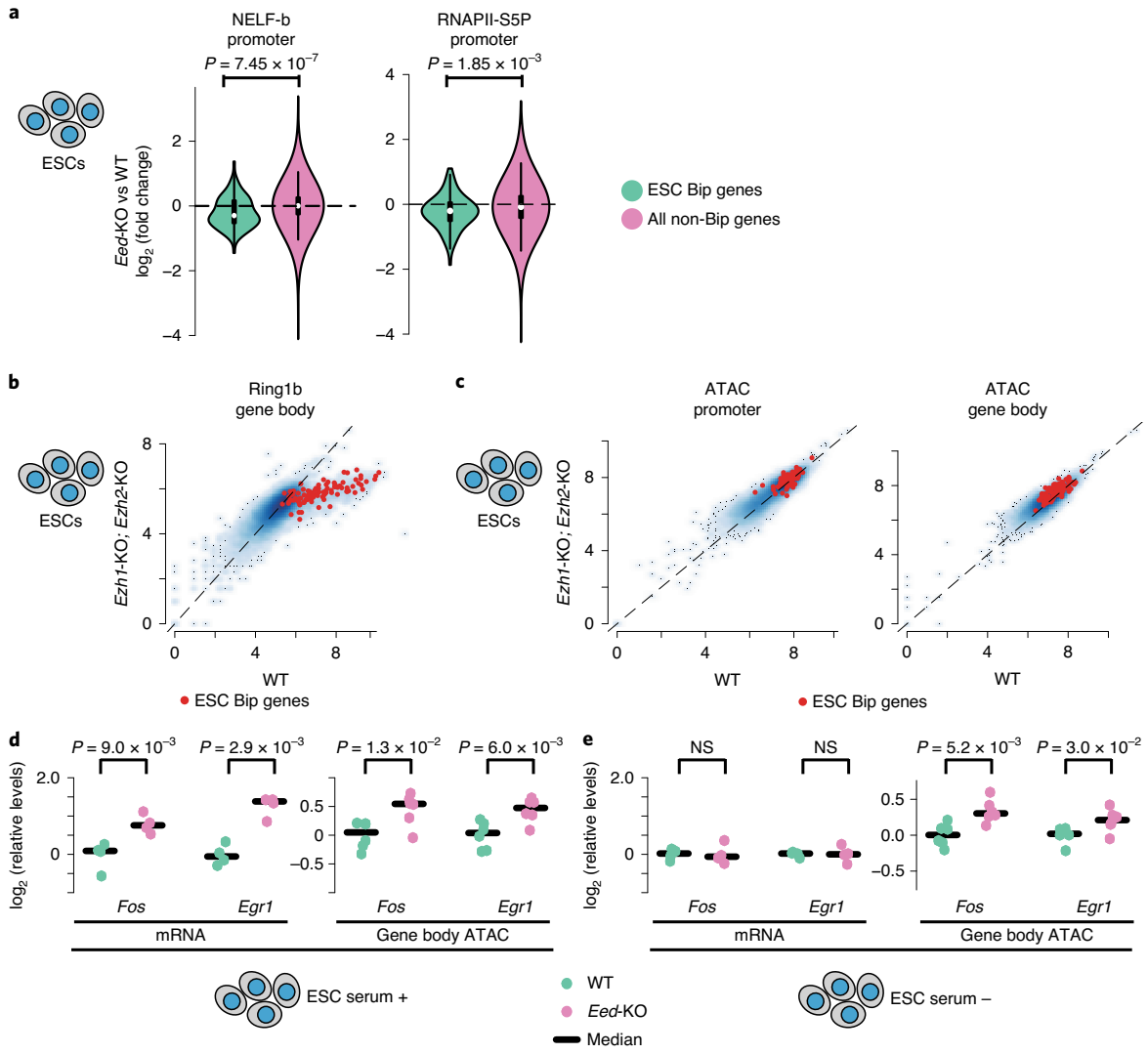


Fig. 6 | Polycomb marking of bipartite genes bodies hampers productive mRNA elongation by inhibition of stimulus-dependent NELF release and chromatin compaction. a, Violin plots visualizing $\log_2$ fold changes of NELF-b and RNAPII-S5P levels at promoters in Eed^{KO} versus WT ESCs; decreased NELF-b and RNAPII-S5P levels were selectively detected in ESC bipartite genes (100 top-scoring genes for bipartiteness in ESCs) as compared to non-Bip genes. Plots extend from the data minima to the maxima, the white dot indicates the median, the box shows the interquartile range and whiskers extend to the most extreme data point within 1.5 time the interquartile range. *P* values are from a two-sided Wilcoxon's test. **b**,**c**, Scatterplots comparing gene body Ring1b levels (**b**) and promoter and gene body accessibilities (**c**) between WT and *Ezh1*-KO;*Ezh2*-KO ESCs. Bipartite genes ($n = 100$) are highlighted in red. **d**,**e**, mRNA expression and gene body accessibilities of bipartite *Fos* and *Egr1* are visualized in Eed^{KO} as compared to WT ESCs in serum-containing (**d**) and serum-starved (**e**) conditions. In the serum-starved condition, although *Fos* and *Egr1* mRNA levels did not differ between WT and Eed^{KO} ESCs, gene body accessibilities (ATAC) were higher in Eed^{KO} as compared to WT ESCs. mRNA ($n = 4$ biologically independent cell cultures) and accessibility ($n = 6$ biologically independent cell cultures) levels at *Fos* and *Egr1* loci were quantified by qPCR and the median is indicated by bars. *P* values are from Welch's two-sample two-sided *t*-tests.

however, its transcripts could be detected after 30 min (Extended Data Fig. 10b). Thus, if developing neurons become exposed to a relevant signal, the bipartite signature at the *Fos* and *Egr1* loci might still allow for rapid inducibility, whereas the bivalent state constrains the *Junb* IEG to a slower response and only in the presence of prolonged stimulation.

We next evaluated the amplitude of the transcriptional response of bipartite IEGs to distinct strengths of the same signal. We used serum treatment after starvation in mouse ESCs, a well-known model to rapidly induce expression of IEGs[3]. *Fos* and *Egr1* carried the bipartite signature also in mouse ESCs (Extended Data Fig. 3d). We treated serum-starved wild-type and Eed^{KO} ESCs with a low (1%) or high (10%) concentration of fetal calf serum (FCS) for a short (8 min) or a longer (16 min) time of exposure and quantified *Fos* and *Egr1* transcriptional induction (Fig. 5f and Extended Data Fig. 10c). Treatment with 10% FCS could induce a rapid (within 8 min) *Fos* and *Egr1* transcriptional response in both wild-type and Eed^{KO} backgrounds; however, the amplitude of the *Fos* and *Egr1* transcriptional responses was higher in Eed^{KO} than wild-type ESCs (Fig. 5f). Furthermore, lowering the concentration of the stimulus by tenfold (that is, treating with 1% FCS) was not sufficient to elicit a transcriptional response after an 8-min treatment in wild-type ESCs but caused significant *Fos* and *Egr1* induction in the Eed^{KO} background (Fig. 5f). In wild-type ESCs, the bipartite *Fos* and *Egr1* IEGs could only be induced after prolonged exposure (that is, 16 min) to 1% FCS (Extended Data Fig. 10c).

In summary, H3K27me3 marking of bipartite IEG gene bodies, while still allowing for rapid induction, regulates the amplitude of the transcriptional response to relevant stimuli. Moreover, Pc marking of gene bodies of bipartite stimulus-response genes may establish a transcriptional threshold to prevent rapid productive induction of IEGs in response to suboptimal and/or nonphysiologically relevant levels of environmental stimuli (Fig. 5g).

Mechanism of stimulus-dependent transition of bipartite to active chromatin. Negative elongation factor (NELF) negatively regulates transcriptional elongation by pausing RNAPII at TSSs[28]. Stimulus-dependent NELF removal from IEG promoters causes release of paused RNAPII into elongation[39]. We found that H3K27me3 on gene bodies inhibits transcriptional elongation in bipartite genes in part by interfering with stimulus-dependent NELF release (Fig. 6a and Supplementary Note). Moreover, *Ezh1/Ezh2* removal caused a reduction in levels of Ring1b in the gene body of bipartite genes (Fig. 6b), correlating with a significant increase of bipartite gene body, although not promoter, accessibility in *Ezh1*-KO;*Ezh2*-KO mouse ESCs (Fig. 6c and Supplementary Note). Such decompaction of bipartite gene bodies was not only merely correlative with increased transcription, but was at least partially caused by the removal of H3K27me3 (Fig. 6d,e and Supplementary Note).

As for the transition from a bipartite to an active state, we reasoned that stimulus-dependent posttranslational modification of transcription factors prebound to promoters could be involved, and in turn induce an increase of H3K27ac, decrease of H3K27me3 and gain of productive transcription (Extended Data Fig. 6a). CREB phosphorylation is rapidly increased in response to neuronal activity and/or other environmental stimuli and induces CREB-binding protein-dependent H3K27ac increase and transcription of IEGs[2,9]. Indeed, phosphoCREB (pCREB) levels increased in the promoter regions of genes that were bipartite at E14.5 and became active at P4 (Fig. 7a; Bip → Exp), including neuronal activity-induced IEGs such as *Fos* and *Egr1* (Fig. 3e and Extended Data Fig. 6b). This correlated with the resolution of the bipartite signature and productive transcription (Fig. 7a and Extended Data Fig. 6a,d,e).

Are strong inducing stimuli (for example, neuronal activity) able to resolve the bipartite epigenetic state? By treating E12.5 short-term cultured hindbrain neurons with 55 mM KCl, after overnight incubation with a cocktail of neuronal activity blockers (TDN cocktail = tetrodotoxin + D-AP5 + NBQX; Methods), the H3K27me3 mark was removed from IEG gene bodies (Fig. 7b). Notably, the decrease of the H3K27me3 mark was detectable as early as 8 min after KCl treatment (Fig. 7b), showing that H3K27me3 removal starts very rapidly after exposure to the inducing stimuli. In addition, treatment of embryonic neurons with a TDN cocktail prevented the removal of H3K27me3 from IEG gene bodies in long-term hindbrain neuron culture (Fig. 7c; also see above and Extended Data Fig. 9a). This indicates that the removal of H3K27me3 from IEG gene bodies is rapid and stimulus dependent.

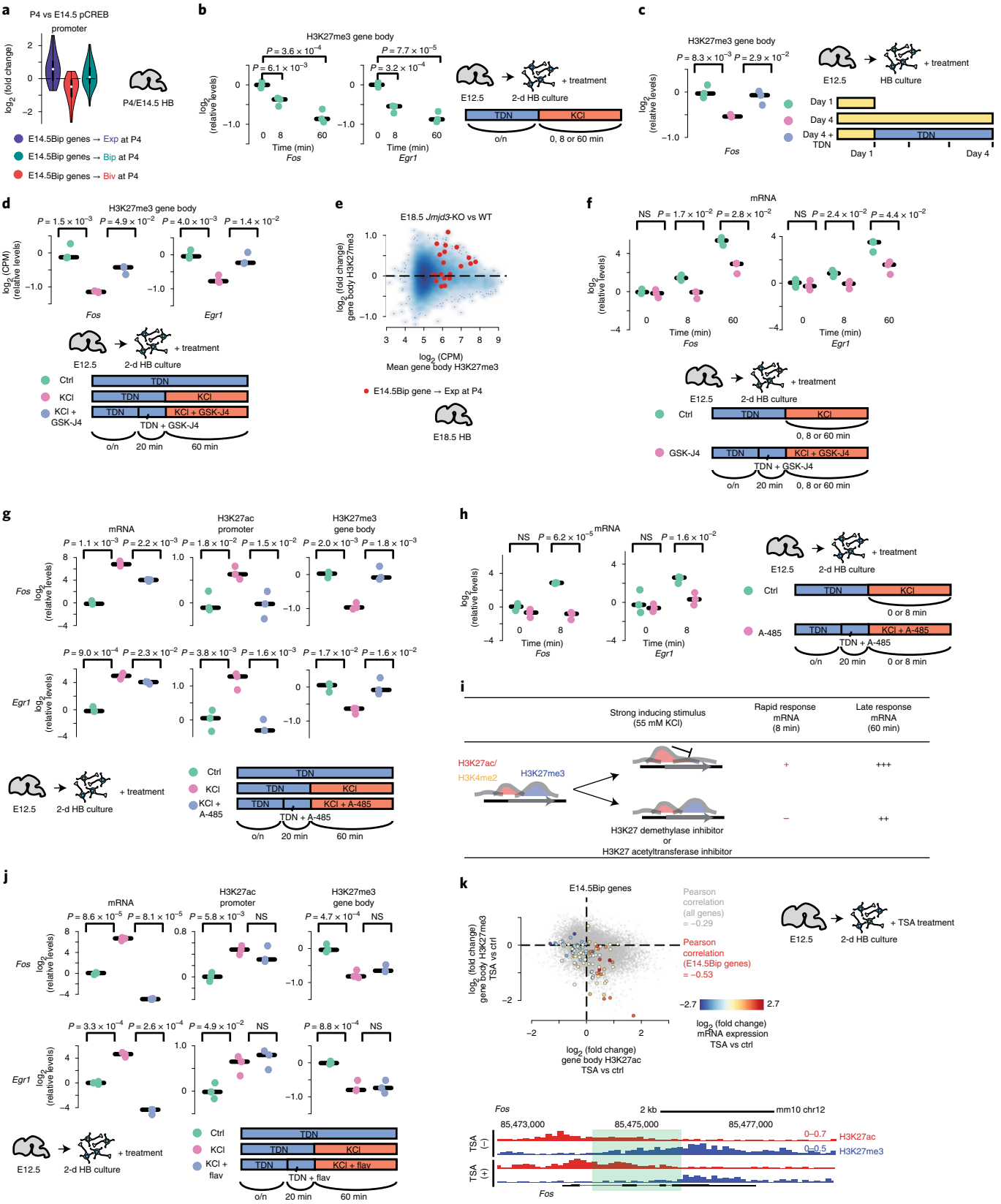
Furthermore, treatment with GSK-J4, an inhibitor of H3K27me3 demethylases (that is, UTX (Kdm6a), Jmjd3 (Kdm6b)) prevented neuronal activity-dependent gene body H3K27me3 removal (Fig. 7d). Similarly, inactivation of *Jmjd3* inhibited, at least partially, gene body H3K27me3 removal from the E14.5Bip genes that became active at perinatal/postnatal stages (Fig. 7e, Supplementary Table 1 and Supplementary Note). These results indicate that the stimulus-dependent removal of H3K27me3 from IEG gene bodies requires active demethylation. In addition, GSK-J4 treatment prevented the rapid transcriptional induction of bipartite IEGs after short (8 min) exposure to the inducing stimulus (Fig. 7f). Taken together with our previous observation that, in the absence of the H3K27me3 mark in Eed^{KO} ESCs, the amplitude of the rapid bipartite IEG transcriptional response upon short exposure (8 min) to inducing stimuli (that is, FCS) is enhanced as compared to that in wild-type control (Fig. 5f), these results indicate that stimulus-dependent gene body H3K27me3 mark removal is essential to achieve rapid and sizeable transcriptional induction of bipartite IEGs.

On the other hand, after prolonged exposure (that is, 60 min) to the inducing stimulus, GSK-J4-treated neurons showed transcriptional upregulation of bipartite IEGs, even though mRNA levels remained significantly lower than those in control neurons (Fig. 7f). Thus, in the event of incomplete H3K27me3 mark removal from the gene body, while rapid bipartite IEG mRNA induction is impaired, transcripts can nonetheless accumulate over time upon prolonged stimulation, albeit never reaching optimal levels.

We then tested the requirement of de novo promoter H3K27 acetylation in activity-dependent removal of the gene body H3K27me3. We treated E12.5 short-term cultured neurons with KCl in the presence of A-485, an inhibitor of H3K27 acetyltransferase p300/CREB-binding protein. A-485 inhibited KCl-dependent increase of promoter H3K27ac levels and prevented the removal of the

Fig. 7 | Mechanism of stimulus-dependent transition from bipartite to active chromatin. a, Violin plots visualizing $\log_2$ fold changes of promoter pCREB levels of E14.5Bip genes from E14.5 to P4; E14.5Bip genes that became expressed ($n = 20$ genes), bivalent ($n = 25$ genes) or remained bipartite ($n = 55$ genes) at P4 were compared. Plots extend from the data minima to the maxima, the white dot indicates the median, the box shows the interquartile range and whiskers extend to the most extreme data point within 1.5 times the interquartile range. **b**, After overnight (o/n) treatment with TDN cocktail (inhibitors of sodium channel, NMDA and AMPA receptors), E12.5 short-term cultured hindbrain neurons were treated with 55 mM KCl for short (8 min) and prolonged (60 min) time courses at day 2. H3K27me3 decrease was detectable as early as 8 min after KCl treatment. **b**–**d**,**f**–**h**,**j**, qPCR quantification of mRNA, H3K27ac and H3K27me3 levels from $n = 3$ biologically independent neuron cultures (bars indicate the median). **c**, Shorter (1 d) or longer (4 d) E12.5 hindbrain neuron cultures in the presence or absence of the TDN cocktail. **d**, E12.5 short-term cultured hindbrain neurons treated by KCl for 1 h in the absence or presence of GSK-J4 at day 2. KCl treatment caused H3K27me3 removal from bipartite IEGs through active demethylation. **e**, MA plot comparing gene body H3K27me3 levels of E14.5Bip genes that became expressed at P4 ($n = 20$ genes; **a**) between E18.5 *Jmjd3*-KO and WT hindbrain cells. **f**, Short (8 min) or prolonged (60 min) KCl treatment of E12.5 short-term cultured hindbrain neurons in the absence or presence of GSK-J4 at day 2, showing that inhibition of H3K27me3 removal prevented rapid induction of bipartite IEGs. **g**, KCl treatment (1 h) of E12.5 short-term cultured hindbrain neurons in the absence or presence of A-485 at day 2. **h**, Short (8 min) KCl treatment of E12.5 short-term cultured hindbrain neurons in the absence or presence of A-485 at day 2. Inhibition of de novo promoter H3K27ac prevented rapid KCl-dependent induction of bipartite IEGs ($n = 3$; *P* values are from a two-sided *t*-test). **i**, Diagram of H3K27 demethylase (GSK-J4) or acetyltransferase (A-485) inhibitor treatments on neuronal activity-dependent transcriptional induction of bipartite IEGs after short (8 min) or prolonged (60 min) exposure to strong inducing stimulus (55 mM KCl). **j**, KCl treatment (1 h) of E12.5 short-term cultured hindbrain neurons in the absence or presence of flavopiridol (Flav) at day 2. KCl-induced H3K27me3 removal of bipartite IEGs appears to be dependent on de novo promoter H3K27ac but not on transcriptional elongation. **k**, Scatter plot showing $\log_2$ fold changes of gene body H3K27ac (*x* axis) and H3K27me3 (*y* axis) levels upon overnight TSA treatment in short-term cultured E12.5 hindbrain neurons. E14.5Bip gene distribution is mapped. Colors indicate the $\log_2$ fold changes of mRNA levels of E14.5Bip genes. The genome browser view of *Fos* in TSA-treated short-term cultured hindbrain neurons shows H3K27me3 levels were reduced by expansion of H3K27ac in the coding region (green highlight). In **b**–**d**, **g** and **j**, *P* values are from analysis of variance followed by Tukey's HSD posthoc tests. In **f** and **h**, *P* values are from Welch's two-sample two-sided *t*-tests.

H3K27me3 mark from bipartite IEG gene bodies (Fig. 7g), indicating that gene body H3K27me3 removal requires stimulus-dependent de novo promoter H3K27 acetylation. Furthermore, A-485 treatment prevented rapid induction of bipartite IEGs after short-time (that is, 8 min) exposure to KCl (Fig. 7h), similarly to GSK-J4 treatment (Fig. 7f), indicating that fast bipartite IEG transcriptional induction requires de novo H3K27 acetylation and rapid removal of the gene body H3K27me3 mark through active demethylation (Fig. 7i and Supplementary Note). Moreover, the KCl-dependent gene body H3K27me3 removal is not merely the consequence



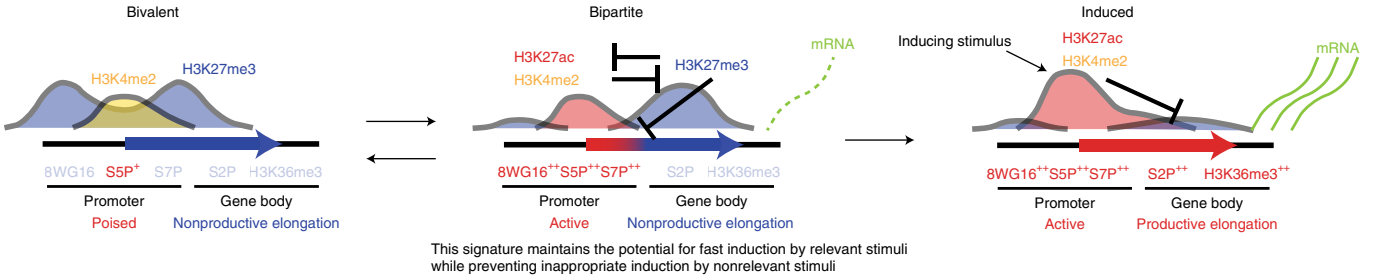


Fig. 8 | Polycomb-dependent regulation of stimulus-response genes during development. During development, a subset of stimulus-response genes displays a bipartite chromatin signature that carries active H3K27ac^{+}/H3K4me2^{+} on promoters but repressive H3K27me3^{+} on gene bodies. The bipartite signature originates from H3K27me3^{+}/H3K4me2^{+} bivalent chromatin maintaining a transcriptionally poised state and nonproductive transcription. Distinct transcription factors cause partial resolution of a subset of bivalent poised promoters of stimulus-response genes into the bipartite state. Active promoters of bipartite genes carry actively initiating RNAPII (high 8WG16, S5P and S7P); however, productive mRNA processing and elongation in gene bodies are maintained at a low rate (low RNAPII-S2P and H3K36me3) due to inhibition of stimulus-dependent NELF release and chromatin compaction by Pc (*Ezh1/2, Eed*)-dependent H3K27me3 (inhibition sign). Pc-dependent marking on gene bodies also inhibits spreading of H3K27ac and accessibility in bipartite gene body regions (mutual inhibition signs). The bipartite state is dynamic and could revert to bivalency. The bipartite signature maintains the potential for fast induction by relevant/strong stimuli while preventing inappropriate induction by nonrelevant/weak stimuli. Inducing stimuli cause de novo promoter H3K27 acetylation, which causes H3K27 demethylase (Kdm6; that is, UTX and Jmjd3)-dependent rapid removal of the Pc mark from bipartite gene bodies (inhibition sign) and fast transcriptional response by transcription factors.

of transcriptional elongation but it is at least partly driven by the de novo promoter acetylation per se (Fig. 7j and Supplementary Note).

Lastly, treatment of E12.5 short-term cultured neurons with the histone deacetylase inhibitor trichostatin A (TSA) resulted in the spreading of H3K27ac into the bipartite gene bodies and H3K27me3 removal, increase of mRNA levels and resolution of the bipartite signature into an active state (Fig. 7k). Together with the analysis of E14.5 *Ezh2*-cKO hindbrain cells (Fig. 5a), and the finding that bipartite genes can revert to bivalency during development (Fig. 3d and Extended Data Fig. 6a), we propose that a dynamic reciprocal balance between the H3K27ac and H3K27me3 marks maintains the bipartite signature.

In summary, stimulus-dependent increase of promoter H3K27ac causes active and rapid H3K27me3 removal from the gene body and release of the elongation barrier, switching from the bipartite to the productive active transcription state.

Discussion

During development, the response to environmental signals requires rapid, stimulus-dependent, transcriptional responses through the induction of IEGs, whose gene products in turn regulate the activation of specific LRGs, driving cell-type-specific differentiation schedules[6,8]. How chromatin states and epigenetic regulation contribute to the timely and rapid activation of stimulus-induced developmental transcriptional programs is poorly understood. Here, we discovered an unusual Pc-dependent bipartite chromatin signature at stimulus-response IEGs before their transcriptional induction in developing neurons, whereas LRGs were preferentially maintained in a bivalent chromatin state. Moreover, we found that the bipartite state is not an exclusive feature of developing neurons, but it is generally present in developing cell types and in ESCs. The bipartite state originates from the bivalent state and is dynamic during development, reverting to bivalency or resolving into rapid activation (Fig. 8). Bipartite genes carry an active promoter and the Pc-dependent H3K27me3 mark on the gene body, which inhibits RNAPII transcriptional elongation regulating the transition into stimulus-dependent productive transcription of bipartite genes (Fig. 8). We demonstrate that this unique chromatin signature provides a suitable epigenetic structure to modulate the rapidity and amplitude of the transcriptional response of inducible IEGs to distinct stimuli during development while inhibiting IEG productive transcription in response to suboptimal and/or nonphysiologically significant levels of environmental stimuli (Fig. 5g). Additional discussion can be found in the Supplementary Discussion.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00789-z.

Received: 5 October 2020; Accepted: 19 January 2021;
Published online: 18 February 2021

References

1. Fowler, T., Sen, R. & Roy, A. L. Regulation of primary response genes. *Mol. Cell* **44**, 348–360 (2011).
2. West, A. E. & Greenberg, M. E. Neuronal activity-regulated gene transcription in synapse development and cognitive function. *Cold Spring Harb. Perspect. Biol.* https://doi.org/10.1101/cshperspect.a005744 (2011).
3. Greenberg, M. E. & Ziff, E. B. Stimulation of 3T3 cells induces transcription of the c-*fos* proto-oncogene. *Nature* **311**, 433–438 (1984).
4. Malik, A. N. et al. Genome-wide identification and characterization of functional neuronal activity-dependent enhancers. *Nat. Neurosci.* **17**, 1330–1339 (2014).
5. Vierbuchen, T. et al. AP-1 transcription factors and the BAF complex mediate signal-dependent enhancer selection. *Mol. Cell* **68**, 1067–1082 (2017).
6. Stroud, H. et al. An activity-mediated transition in transcription in early postnatal neurons. *Neuron* **107**, 874–890 (2020).
7. Mayer, A., Landry, H. M. & Churchman, L. S. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Curr. Opin. Cell Biol.* **46**, 72–80 (2017).
8. Yap, E. L. & Greenberg, M. E. Activity-regulated transcription: bridging the gap between neural activity and behavior. *Neuron* **100**, 330–348 (2018).
9. Lonze, B. E. & Ginty, D. D. Function and regulation of CREB family transcription factors in the nervous system. *Neuron* **35**, 605–623 (2002).
10. Toth, A. B., Shum, A. K. & Prakriya, M. Regulation of neurogenesis by calcium signaling. *Cell Calcium* **59**, 124–134 (2016).
11. Ginty, D. D., Glowacka, D., Bader, D. S., Hidaka, H. & Wagner, J. A. Induction of immediate early genes by Ca^{2+} influx requires cAMP-dependent protein kinase in PC12 cells. *J. Biol. Chem.* **266**, 17454–17458 (1991).
12. Greenberg, M. E., Greene, L. A. & Ziff, E. B. Nerve growth factor and epidermal growth factor induce rapid transient changes in proto-oncogene transcription in PC12 cells. *J. Biol. Chem.* **260**, 14101–14110 (1985).
13. Erzurumlu, R. S., Murakami, Y. & Rijli, F. M. Mapping the face in the somatosensory brainstem. *Nat. Rev. Neurosci.* **11**, 252–263 (2010).

14. Kitazawa, T. & Rijli, F. M. Barrelette map formation in the prenatal mouse brainstem. *Curr. Opin. Neurobiol.* **53**, 210–219 (2018).
15. Erzurumlu, R. S. & Gaspar, P. Development and critical period plasticity of the barrel cortex. *Eur. J. Neurosci.* **35**, 1540–1553 (2012).
16. Hrvatin, S. et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21**, 120–129 (2018).
17. Tyssowski, K. M. et al. Different neuronal activity patterns induce different gene expression programs. *Neuron* **98**, 530–546 (2018).
18. Valles, A. et al. Genomewide analysis of rat barrel cortex reveals time- and layer-specific mRNA expression changes related to experience-dependent plasticity. *J. Neurosci.* **31**, 6140–6158 (2011).
19. Mohn, F. et al. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol. Cell* **30**, 755–766 (2008).
20. Ferrai, C. et al. RNA polymerase II primes Polycomb-repressed developmental genes throughout terminal neuronal differentiation. *Mol. Syst. Biol.* **13**, 946 (2017).
21. Hirabayashi, Y. et al. Polycomb limits the neurogenic competence of neural precursor cells to promote astrogenic fate transition. *Neuron* **63**, 600–613 (2009).
22. Aranda, S., Mas, G. & Di Croce, L. Regulation of gene transcription by Polycomb proteins. *Sci. Adv.* **1**, e1500737 (2015).
23. Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome regulation by Polycomb and Trithorax: 70 years and counting. *Cell* **171**, 34–57 (2017).
24. Bernstein, B. E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
25. Minoux, M. et al. Gene bivalency at Polycomb domains regulates cranial neural crest positional identity. *Science* https://doi.org/10.1126/science.aal2913 (2017).
26. Piunti, A. & Shilatifard, A. Epigenetic balance of gene expression by Polycomb and COMPASS families. *Science* **352**, aad9780 (2016).
27. Bonnefont, J. et al. Cortical neurogenesis requires Bcl6-mediated transcriptional repression of multiple self-renewal-promoting extrinsic pathways. *Neuron* **103**, 1096–1108 (2019).
28. Chen, F. X., Smith, E. R. & Shilatifard, A. Born to run: control of transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **19**, 464–478 (2018).
29. Brookes, E. & Pombo, A. Modifications of RNA polymerase II are pivotal in regulating gene expression states. *EMBO Rep.* **10**, 1213–1219 (2009).
30. Zaborowska, J., Egloff, S. & Murphy, S. The pol II CTD: new twists in the tail. *Nat. Struct. Mol. Biol.* **23**, 771–777 (2016).
31. Brookes, E. et al. Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. *Cell Stem Cell* **10**, 157–170 (2012).
32. Stock, J. K. et al. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nat. Cell Biol.* **9**, 1428–1435 (2007).
33. Simon, J. A. & Kingston, R. E. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. Cell Biol.* **10**, 697–708 (2009).
34. Blackledge, N. P., Rose, N. R. & Klose, R. J. Targeting Polycomb systems to regulate gene expression: modifications to a complex story. *Nat. Rev. Mol. Cell Biol.* **16**, 643–649 (2015).
35. Shen, X. et al. EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency. *Mol. Cell* **32**, 491–502 (2008).
36. Schoeftner, S. et al. Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. *EMBO J.* **25**, 3110–3122 (2006).
37. Lavarone, E., Barbieri, C. M. & Pasini, D. Dissecting the role of H3K27 acetylation and methylation in PRC2-mediated control of cellular identity. *Nat. Commun.* **10**, 1679 (2019).
38. Kim, T. K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
39. Schaukowitch, K. et al. Enhancer RNA facilitates NELF release from immediate early genes. *Mol. Cell* **56**, 29–42 (2014).

Methods

Mating scheme. To obtain E10.5 and E14.5 *Krox20::Cre;R26*^{tdTomato} (*K20*$^{tdTomato/+}$) embryos, the *Krox20::Cre* transgenic mouse line[40] was crossed with the *R26*^{tdTomato} reporter mouse line[41] (The Jackson Laboratory, 007905). To obtain E14.5, E18.5 and P4 *Drg11::Cre;R26R*^{ZsGreen};*r2::mCherry* (*Drg11*$^{ZsGreen/+}$;*r2*$^{mCherry/+}$) mice, the *Drg11::Cre* transgenic mouse line[42], the *R26R*^{ZsGreen} reporter mouse line[41] (The Jackson Laboratory, 007906) and the *r2::mCherry* (*r2*$^{mCherry/+}$) transgenic mouse line[42] were crossed (Extended Data Fig. 1c). To obtain E14.5, *Drg11::Cre;R26*^{tdTomato};*r2::EGFP* (*Drg11*$^{tdTomato/+}$;*r2*$^{EGFP/+}$) mice, the *Drg11::Cre* transgenic mouse line, the *R26*^{tdTomato} reporter mouse line and the *r2::EGFP* (*r2*$^{EGFP/+}$) transgenic mouse line (Supplementary Methods) were crossed (Extended Data Fig. 1d). To obtain E18.5 *Drg11::Cre;R26*$^{Kir\text{-}mCherry}$;*r2::EGFP* (*Drg11*$^{Kir/+}$;*r2*$^{EGFP/+}$) mice, the *Drg11::Cre* transgenic mouse line, the *R26*$^{Kir\text{-}mCherry}$[43] mouse line and the *r2*^{EGFP} transgenic mouse line were crossed (Extended Data Fig. 1e). To obtain E14.5 *Krox20::Cre;Ezh2*$^{flox/flox}$;*R26*^{RFP} (*Ezh2*-cKO$^{r3\text{-}RFP}$) mice, the *Krox20::Cre;Ezh2*$^{flox/+}$ mouse line was crossed with the *Ezh2*$^{flox/flox}$;*R26*^{RFP} mouse line. The *R26*^{RFP} mouse line was described before[44]. To obtain E14.5 *Hoxa2::Cre;R26*^{tdTomato} (*Hoxa2*$^{tdTomato/+}$) embryos, the *Hoxa2::Cre* transgenic mouse line[45] was crossed with the *R26*^{tdTomato} reporter mouse line. To obtain E14.5 *Hoxa2::Cre;Ezh2*$^{flox/flox}$;*R26*^{RFP} (*Ezh2*-cKO$^{HB\text{-}RFP}$) mouse, the *Hoxa2::Cre;Ezh2*$^{flox/+}$;*R26*^{RFP} mouse line was crossed with the *Ezh2*$^{flox/flox}$;*R26*^{RFP} mouse line. *Hoxa2::Cre* line, which labels from r2 to posterior hindbrain neurons, was used to collect a relatively large number of hindbrain neurons to enable the molecular analysis of *Ezh2*-null neurons (see below). To obtain P8 *Krox20::Cre;R26*$^{Kir\text{-}mCherry}$ (*K20*$^{Kir/+}$) mice, the *Krox20::Cre* transgenic mouse line was crossed with the *R26*$^{Kir\text{-}mCherry}$ mouse line. To obtain P8 *Krox20::Cre;R26*^{tdTomato};*r2::EGFP* (*K20*$^{tdTomao/+}$;*r2*$^{EGFP/+}$) mice, the *Krox20::Cre;r2:EGFP* mouse line was crossed with the *R26*^{tdTomato} reporter mouse line. To obtain P8 *Krox20::Cre;R26*$^{Kir\text{-}mCherry}$;*r2::EGFP* (*K20*$^{Kir/+}$;*r2*$^{EGFP/+}$) mice, the *Krox20::Cre;r2::EGFP* mouse line was crossed with the *R26*$^{Kir\text{-}mCherry}$ mouse line. *Jmjd3*$^{-/-}$ (*Jmjd3*-KO) mouse line was described previously[46]. To obtain P10 *Krox20::Cre;LSL-R26*$^{TVA\text{-}LaxZ}$ (*K20*$^{TVA/+}$) mice, the *Krox20::Cre* transgenic mouse line was crossed with *LAL-R26*$^{TVA\text{-}LacZ}$ transgenic mouse line[47]. To obtain P10 *Krox20::Cre;LSL-R26*$^{TVA\text{-}LacZ}$;*R26*$^{Kir\text{-}mCherry}$ (*K20*$^{TVA/Kir}$) mice, the *Krox20::Cre;LSL-R26*$^{TVA\text{-}LacZ}$ mouse line was crossed with the *R26*$^{Kir\text{-}mCherry}$ transgenic mouse line. Nomenclatures for mouse lines are summarized in Supplementary Table 1.

Dissociation of hindbrain tissue and isolation of cells by FACS. To collect r3-derived progenitors from E10.5 *K20*$^{tdTomato/+}$ mice, r2–r4 regions were microdissected. Dissected tissue was kept in 1× PBS on ice, then treated with papain digestion mix (10 mg ml^{-1} papain, 2.5 mM cysteine, 10 mM HEPES (pH 7.4), 0.5 mM EDTA and 0.9× DMEM) for 3 min at 37 °C and immediately put on ice. Tissue was rinsed by ice-cold 1× DMEM, dissociated by pipetting and filtered. FACS was used to collect r3-derived cells (Supplementary Fig. 1). Processing of these cells was adapted for further analyses (for example, RNA-seq, ATAC–seq and ChIP–seq). To collect postmitotic barrelette neurons of vPrV (*Drg11*$^{vPrV\text{-}ZsGreen/+}$, *Drg11*$^{vPrV\text{-}tdTomato/+}$and *Drg11*$^{vPrV\text{-}Kir/+}$; for nomenclatures, see Extended Data Fig. 1c–e, Supplementary Table 1 and Supplementary Note) from E14.5, E18.5 and P4 *Drg11*$^{ZsGreen/+}$;*r2*$^{mCherry/+}$, *Drg11*$^{tdTomato/+}$;*r2*$^{EGFP/+}$ or *Drg11*$^{Kir/+}$;*r2*$^{EGFP/+}$ mice, r2–r3-derived regions were microdissected. The boundary between r3 and r4 was identified by the position of the facial nerve. Dissected tissue was kept in 1× PBS on ice, then treated with papain digestion mix for 4 min at 37 °C and immediately put on ice. Tissue was rinsed with 1× ice-cold DMEM, dissociated by pipetting and filtered. Wild-type (*Drg11*$^{vPrV\text{-}ZsGreen/+}$ and *Drg11*$^{vPrV\text{-}tdTomato/+}$) barrelette neurons were FACS-sorted by selecting green single-positive cells from *Drg11*$^{ZsGreen/+}$;*r2*$^{mCherry/+}$ mice or red single-positive cells from *Drg11*$^{tdTomato/+}$;*r2*$^{EGFP/+}$ mice, while activity-deprived barrelette neurons (*Drg11*$^{vPrV\text{-}Kir/+}$) were sorted by collecting red single-positive cells from *Drg11*$^{Kir/+}$;*r2*$^{EGFP/+}$ mice (Extended Data Fig. 1 and Supplementary Fig. 2). Processing of these cells was adapted for further analyses (for example, RNA-seq, ATAC–seq and ChIP–seq). To collect r3-derived hindbrain cells from E14.5 *K20*$^{tdTomato/+}$ or *Ezh2*-cKO$^{r3\text{-}RFP}$ mice, r2–r4 regions were microdissected. Dissected tissue was kept in 1× PBS on ice, then treated with papain digestion mix for 3 min at 37 °C and immediately put on ice. Tissue was rinsed with 1× ice-cold DMEM, dissociated by pipetting and filtered. FACS was used to collect r3-derived cells (Supplementary Fig. 3a). Processing of these cells was adapted for further analyses (for example, RNA-seq and ChIP–seq). To collect hindbrain cells from E14.5 *Hoxa2*$^{tdTomato/+}$ or *Ezh2*-cKO$^{HB\text{-}RFP}$ mice, hindbrain regions (from the exit of the trigeminal nerve in the rostral hindbrain down to the beginning of the spinal cord) were microdissected. Dissected tissue was kept in 1× PBS on ice, then treated with papain digestion mix for 3 min at 37 °C and immediately put on ice. Tissue was rinsed with 1× ice-cold DMEM, dissociated by pipetting and filtered. Hindbrain-derived cells were collected by FACS (Supplementary Fig. 3b). Processing of these cells was adapted for further analyses (for example, RT–qPCR, ATAC–seq and ChIP–seq).

Overexpression of dCas9–UTX. Ex vivo cultured E12.5 hindbrain neurons (Supplementary Methods) were transfected with Lipofectamine 2000 (Thermo Fisher Scientific, 11668019) at culture day 1. Next, dCas9 or dCas9–UTX fusion protein overexpression vector was co-transfected with two gRNA/EGFP overexpression vectors (pGuide_EGFP) targeted to the mouse genes (that is *Fos*, *Egr1*, *Actb* and *Gapdh*; Supplementary Methods). After 24 h (day 2), neurons were dissociated with 0.05% trypsin/EDTA. About 1% GFP-positive neurons were collected by FACS (Supplementary Fig. 4). Processing of these cells was adapted for further analyses (RNA extraction and ChIP–seq of H3K27me3; see below and Supplementary Methods).

KCl, trichostatin A, TDN cocktail, GSK-J4, A-485 and flavopiridol treatment. Ex vivo cultured E12.5 hindbrain neurons were treated with 2 μM TSA (MBJ, JM-1606-1) at the culture day 1 and incubated for 16 h. For KCl treatment, cultured hindbrain neurons were treated with a cocktail of neuronal activity blockers (TDN cocktail = 1 μM tetrodotoxin (TOCRIS, 1069) + 100 μM D-AP5 (Sigma, A8054) + 20 μM NBQX (TOCRIS, 0373) at day 1 for an overnight incubation, and 55 mM KCl-containing medium was treated at day 2 in the presence or absence of 35 μM GSK-J4 (Sigma, SML0701), 50 μM A-485 (TOCRIS, 6387) or 10 μM flavopiridol (Sigma, F3055) after a rinse. For *Drg11*$^{tdTomato/+}$ cultured hindbrain neurons, tdTomato^{+} neurons were sorted immediately after the KCl treatment. Processing of these cells was adapted for further analyses (mRNA-seq, ATAC–seq and ChIP–seq; see below and Supplementary Methods).

Serum shock of mouse ESCs. Wild-type and *Eed*^{KO} mouse ESCs were cultured up to 80% confluence in normal culture medium (Supplementary Methods) and were subsequently serum starved overnight in the culture medium that did not contain FCS. Serum-starved ESCs were treated by a low (1%) or high (10%) concentration of FCS for a short (8 min) or a longer (16 min) exposure time. After reaction, total RNA was immediately extracted by RNeasy Mini Kit (QIAGEN, 74104) with genomic DNA digestion using RNase-Free DNase I Set (QIAGEN, 79254) according to the manufacturer's protocol, and RT–qPCR was conducted (Supplementary Methods).

Sample preparation and RNA isolation and sequencing. For RNA-seq experiments, total RNA was extracted by a Single Cell RNA Purification Kit (NORGEN, 51800) with genomic DNA digestion using an RNase-Free DNase I Kit (NORGEN, 25710) according to the manufacturers' protocols. Library preparation protocols for poly A^{+} mRNA (Smart-seq2 protocol[48]) and total RNA (Ovation SoLo RNA-seq System), as well as for for single-cell RNA-seq (10x Genomics), are described in Supplementary Methods.

Sample preparation and chromatin immunoprecipitation and sequencing. Cells were cross-linked with 1% formaldehyde for 10 min at and quenched with 125 mM glycine for 5 min at room temperature. To achieve the sequencing of chromatin immunoprecipitated from small amounts of cells, preparation of the ChIP–seq library was mostly done by ChIPmentation protocol[49]. Cells were lysed in sonication buffer (10 mM Tris-HCl (pH 8), 5 mM EDTA, 0.5% SDS, 0.1× PBS, 1× protease inhibitor cocktail (cOmplete EDTA-free; Roche, 04693132001)) on ice, and sonicated using the Covaris machine to obtain DNA fragments between 150 bp and 500 bp. The supernatant was transferred to a new tube, diluted with equilibration buffer (10 mM Tris-HCl (pH 8), 1 mM EDTA, 140 mM NaCl, 1% Triton X-100, 0.1% sodium deoxycholate and 1× protease inhibitor cocktail). Chromatin solutions were incubated overnight at 4 °C with antibodies. The next day, protein G magnetic beads (Dynabeads Protein G; Thermo Fisher, 10004D) were added and the incubation was continued for 2 h at 4 °C. The beads were then washed and resuspended in tagmentation buffer (10 mM Tris-HCl (pH 8) and 5 mM $MgCl_2$) containing Tagment DNA Enzyme from the Nextera DNA Sample Prep Kit (Illumina, FC-121-1030) and incubated at 37 °C for 10 min. The beads were washed, and DNA was eluted from the beads with elution buffer (10 mM Tris-HCl (pH 8), 5 mM EDTA, 300 mM NaCl, 0.5% SDS and proteinase K) at 65 °C. DNA was purified with SPRI AMPure XP beads (Beckman Coulter; sample to beads ratio of 1:2) and eluted in 10 mM Tris-HCl (pH 8). Libraries were prepared in a 50-μl reaction (1× KAPA HiFi Hot Start Ready Mix and 0.8 μM primers). Enriched libraries were purified with size selection using SPRI AMPure XP beads (sample to beads ratio of 1:0.6) to remove long fragments and recovering the remaining DNA (sample to beads ratio of 1:2). Sequencing was performed on an Illumina HiSeq 2500 (50-bp read length, single-end). The ChIP–seq protocol was optimized for different experiments (for example, FACS-sorted cells, bulk tissue, cultured cells, prefixed tissue and sequential ChIP–seq; Supplementary Methods).

Sample preparation and assay for transposase-accessible chromatin. ATAC–seq experiments were performed as described previously[50] with minor modifications. For each experiment, 50,000–70,000 cells were used. Two independent biological replicates were prepared. For the detailed protocol of ATAC–seq, see Supplementary Methods.

Reference genome and annotation. The mouse GRCm38/mm10 genome assembly was used as reference. The most variable TSS in promoter chromatin accessibility in our datasets was selected per gene. Promoter (P) regions were defined as 1,000 bp upstream and 500 bp downstream of that TSS. Gene body (GB) regions were defined from 1,001 bp to 3,000 bp downstream of the TSS (Extended Data Fig. 5a), and TESs as the last 2,000 bp of the most downstream transcript.

Spliced and unspliced counts for total RNA datasets (Ovation SoLo RNA-seq) were obtained for GENCODE transcripts. The unspliced transcriptome was created by including intronic regions in each transcript. For the sequential ChIP analysis, regions were defined as the start of P to the end of GB. In Fig. 5c, TSS proximal regions were defined from 100 bp upstream to 200 bp downstream of the TSS (exonic regions) and compared with all exonic regions in the gene.

Read alignment to the reference genome. For read alignment, mRNA and total RNA reads were aligned to the genome using STAR and converted to bam files with 'samtools'. In addition, for total RNA datasets, 'salmon' was used to estimate spliced and unspliced transcript abundances. For better comparability with 51-mer paired-end samples, reads in ATAC–seq samples that had been sequenced as 76-mer paired ends were trimmed to 51-mer paired-end samples using 'cutadapt', followed by adaptor sequence trimming. The trimmed ATAC–seq and ChIP–seq samples were aligned using 'bowtie2'. For genome browser views, the number of alignments per 100-bp window and per million alignments in each sample was calculated and stored in BigWig format with 'QuasR' using 'qExportWig'. When appropriate, counts were corrected to reduce between-sample nonlinearities using 'normalizeCyclicLoess' in limma[51]. Coordinates of expected 4C fragments were created by in silico digesting of the genome with DpnII. Valid fragments were defined as fragments containing an NlaIII site at least 30 bp away from the fragment start and end. Reads were aligned to the genome with QuasR using 'qAlign'.

Peaks were called on ATAC–seq samples per condition using MACS2 with parameters --f BED, --nomodel, --shift −100, --extsize 200 and --keep-dup all. For E14.5 barrelette H3K27me3, Pc peaks were defined using a hidden semi-Markov model with 'mhsmm' package[52] to detect Pc regions of varying sizes. Each gene in Fig. 1c was defined as Pc overlapping if any of its transcripts overlapped with the defined Pc regions. For other ChIP samples, positive regions were defined using a Gaussian mixture model as described by Minoux et al.[25].

To define barrelette enhancers, we used the union of the ATAC peaks from both E14.5 and P4 barrelette neurons, which were at least 1,000 bp away from any TSS with an ATAC $\log_2$ fold change greater than 1.5 times that from E14.5 to P4. Using neuronal activity-dependent Fos targets from work by Malik et al.[4], we divided our enhancers into Fos overlapping (85) and non-Fos overlapping (3,882).

Read quantification and abundance estimation. All analyses downstream of alignment steps were performed in R. RNA-seq, ATAC–seq and ChIP–seq samples were quantified with QuasR's 'qCount' on genes (exons) or specified genomic regions defined above. Salmon was used with 'tximport' to quantify spliced and unspliced transcripts per million for total RNA datasets. Single-cell RNA-seq data were quantified with CellRanger[53], followed by quality control and log transformation of UMI counts with 'scran' and 'scater'.

Raw counts were corrected for library size differences by multiplying by scaling factors, calculating counts per million or calculating RPKM values, followed by averaging across replicates and a $\log_2$ transformation. For samples with a strong GC bias or genome-wide signal changes (*Ezh2*-cKO), specific normalizations were applied (Supplementary Methods).

Activity-response genes. ARGs specific to barrelette neurons (bsARGs) were defined as genes present at low expression at E14.5 (RPKM < 3), upregulated from E14.5 to E18.5 and downregulated between E18.5 Kir-OE and E18.5 wild-type neurons (56 genes; Extended Data Fig. 1q–s). Differential expression analyses were performed with 'edgeR' using 'glmQLFit'. Non-barrelette ARGs (nbARGs) were defined based on the literature[16–18]. All activity-dependent genes (rapid and late induced) were used, and only genes not expressed (RPKM < 3) in all of E14.5, E18.5 and P4 barrelette neurons and not contained in the bsARGs were kept (83 genes). The bsARGs and nbARGs were grouped as IEGs or LRGs as obtained from the literature[16–18] (Supplementary Methods) and manually classified as bipartite or bivalent based on histone modifications at E14.5.

Bipartite and bivalent gene scores. To calculate a 'bipartiteness' score, we selected genes with low expression (RPKM < 3), with more H3K27ac in P than in GB and with more of H3K27me3 in GB than in P. We separately ranked H3K27ac in P and H3K27me3 in GB from low to high and summed the two ranks for each gene. A 'bivalency' score was calculated similarly for genes with low expression (RPKM < 3) followed by summing the ranks of H3K27me3 in P and H3K4me2 in P.

Visual inspection of individual gene loci on the genome browser and correlation between biological replicates were used to evaluate both scores, and confirmed strong correlation of bipartiteness and bivalency scores with true bipartite or bivalent chromatin signatures, respectively (Extended Data Fig. 3a,b). By calculating the fraction of true positives (Extended Data Fig. 3a,b) for different score values, we estimated the total number of bipartite and bivalent genes in each condition (Fig. 2a). We used conservative definition of bipartite genes and considered only the top 100 scoring genes (E14.5Bip genes). We confirmed that this threshold selects at least 75–80% true bipartite genes (Extended Data Fig. 3a), allowing for efficient detection of bipartite genes without manual classification. The chromatin mark distributions surrounding the TSSs for the top 100 bipartite and bivalent genes were obtained using QuasR's 'qProfile', normalized for library size, scaled between 0 and 1 and smoothed with 'runmean' (Fig. 2b and Supplementary Methods).

For the top 100 E14.5Bip and bivalent genes in barrelette neurons, the CpG observed-over-expected ratio was calculated in 100-bp bins around the TSS and averaged across genes. Motif enrichment analysis on promoters of bipartite and bivalent genes was carried out using 'monaLisa' and 'Homer' tools[54].

Visualizing combined chromatin states with *t*-distributed stochastic neighbor embedding. H3K27me3, H3K27ac and H3K4me2 and chromatin accessibility (ATAC) were quantified for each gene in P and GB regions as $\log_2$ RPKM values, and *t*-SNE[55] was used to create a 2D embedding, placing genes with similar chromatin landscapes close together. For the combined *t*-SNE (Fig. 3a), additional normalization steps were performed to reduce between-sample nonlinearities (Supplementary Methods).

Using the 100 top-scoring genes for the single time point *t*-SNE (E10.5 *t*-SNE) map and the top 300 genes for the combined *t*-SNE map (E14.5/E18.5/P4-combined *t*-SNE), 2D densities for bipartite and bivalent genes were estimated with 'kde2d' from MASS[56] and visualized as contour lines (Extended Data Fig. 5g,l). We calculated Euclidean distances of genes between P4 and E14.5 on the original eight-dimensional space consisting of normalized $\log_2$(RPKM) counts of ATAC, H3K27me3, H3K27ac and H3K4me2 on the P and GB regions and colored the E14.5 *t*-SNE by this distance (Extended Data Fig. 5h).

The 100 top-scoring genes for bipartiteness at E14.5 were divided into three groups: genes that became expressed at P4 (RPKM ≥ 3; 20 genes), genes that became bivalent (move into the bivalent contour at P4; Fig. 3d; 25 genes) and those that remain bipartite (55 genes; Extended Data Fig. 6a).

Gene sets for comparison to E14.5 bipartite genes. For Fig. 4, we first selected the 100 top-scoring bipartite and bivalent genes at E14.5 (E14.5Bip and E14.5Biv) and excluded three genes contained in both sets. Control sets of the same number of genes (97) were then created using 'swissknife': E14.5AcP genes were sampled from all genes except the top 400 E14.5Bip genes and E14.5Biv genes, to have similar H3K27ac distribution in P as E14.5Bip. E14.5mRNALow genes that matched E14.5Bip in $\log_2$ RPKM mRNA expression were sampled similarly from all genes excluding E14.5Biv, the top 400 E14.5Bip genes and E14.5AcP. Finally, two sets were sampled from the bottom and top 30% of genes ordered by mRNA expression, excluding any of the genes already contained in the previous sets. For the E10.5 samples, sets with a total of 99 genes were similarly created, excluding 1 gene that was common between the top 100 bipartite and bivalent genes (Extended Data Fig. 4e).

For Extended Data Fig. 7a,b, Entrez identifiers from the top 100 E14.5Bip genes were mapped to Ensemble IDs using biomaRt[57], resulting in 90 successfully mapped identifiers. A control set of the same size with matching spliced transcript abundance, excluding the top 400 E14.5Bip genes, was then randomly sampled.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All raw sequencing data and processed data used for this study are available through ArrayExpress and will be released to the public without restrictions: mRNA-seq (Smart-seq2), E-MTAB-8314; total RNA-seq (Solo RNA-seq), E-MTAB-8311; ChIP–seq (ChIPmentation), E-MTAB-8317; ATAC–seq, E-MTAB-8313; 4C–seq, E-MTAB-8295; and single-cell RNA-seq (10x Genomics), E-MTAB-8312. FACS gating strategies/source data are presented in Supplementary Figs. 1–4. Public sequencing datasets were obtained from the Gene Expression Omnibus and ENCODE as follows: mouse cortical culture (GSE21161 and GSE60192), mouse embryonic forebrain (GSE93011 and GSE52386), mouse adult cortical excitatory neuron (GSE63137), mouse ESCs (GSE36114 and GSE94250), mouse ESCs for *Ezh2*-KO experiments (GSE116603), mouse E14.5 heart tissues (GSE82764, GSE82637, GSE82640 and GSE78441; ENCSR068YGC), mouse E14.5 liver tissues (GSE78422, GSE82407, GSE82615 and GSE82620; ENCSR032HKE) and E10.5 mouse NCCs isolated from the frontal nasal process (GSE89437).

Code availability

Computational analyses were performed in R using the mentioned publicly available packages (Methods, Reporting Summary and Supplementary Methods). The custom tool monaLisa (v0.1.28), used for motif enrichment, can be found on GitHub at https://github.com/fmicompbio/monaLisa/. The custom tool swissknife (v0.10) is available on https://github.com/fmicompbio/swissknife/.

References

40. Voiculescu, O. et al. Hindbrain patterning: *Krox20* couples segmentation and specification of regional identity. *Development* **128**, 4967–4978 (2001).
41. Madisen, L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2010).

42. Bechara, A. et al. *Hoxa2* selects barrelette neuron identity and connectivity in the mouse somatosensory brainstem. *Cell Rep.* **13**, 783–797 (2015).
43. Moreno-Juan, V. et al. Prenatal thalamic waves regulate cortical area size prior to sensory processing. *Nat. Commun.* **8**, 14172 (2017).
44. Luche, H., Weber, O., Nageswara Rao, T., Blum, C. & Fehling, H. J. Faithful activation of an extra-bright red fluorescent protein in 'knock-in' Cre-reporter mice ideally suited for lineage tracing studies. *Eur. J. Immunol.* **37**, 43–53 (2007).
45. Di Meglio, T. et al. *Ezh2* orchestrates topographic migration and connectivity of mouse precerebellar neurons. *Science* **339**, 204–207 (2013).
46. Maheshwari, U. et al. Postmitotic Hoxa5 expression specifies pontine neuron positional identity and input connectivity of cortical afferent subsets. *Cell Rep.* **31**, 107767 (2020).
47. Seidler, B. et al. A Cre-loxP-based mouse model for conditional somatic gene expression and knockdown in vivo by using avian retroviral vectors. *Proc. Natl Acad. Sci. USA* **105**, 10137–10142 (2008).
48. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
49. Schmidl, C., Rendeiro, A. F., Sheffield, N. C. & Bock, C. ChIPmentation: fast, robust, low-input ChIP–seq for histones and transcription factors. *Nat. Methods* **12**, 963–965 (2015).
50. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
51. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
52. O'Connell, J. & Hojsgaard, S. Hidden semi-Markov models for multiple observation sequences: the mhsmm package for R. *J. Stat. Softw.* **39**, 4 (2011).
53. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
54. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
55. Maaten, L. J. P. V. D. & Hinton, G. E. Visualizing high-dimensional data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
56. Venables, W. N. & Ripley, B. D. Modern Applied Statistics with S 4th edn (Springer, 2002).
57. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

Acknowledgements

F.M.R. wishes to dedicate this paper to the memory of P. Sassone-Corsi (1956–2020), dear friend and eminent scientist who made seminal contributions to immediate early gene and activator protein-1 transcriptional regulation and with whom F.M.R. discussed this project at an early stage. We thank A. Pombo for very useful discussion. We thank N. Vilain, S. Smallwood, D. Gaidatzis and the members of the Rijli group and FMI facilities for excellent technical support and discussion. We thank S. H. Orkin (Harvard Medical School) and A. Wutz (ETH Zurich) for the kind gifts of the *Ezh2*^{flox} mouse line and *Eed*^{KO} ESCs, respectively. The *LAL-R26*$^{TVA\text{-}LacZ}$ line was a kind gift from D. Saur (Technische Universitat Munchen). T.K. was supported by a Japan Society for the Promotion of Science fellowship, and O.J. was supported by an EMBO Long-Term fellowship. F.M.R. was supported by the Swiss National Science Foundation (31003A_149573 and 31003A_175776). This project has also received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant no. 810111-EpiCrest2Reg). F.M.R. and M.B.S. were also supported by the Novartis Research Foundation.

Author contributions

T.K. and F.M.R. conceived the study, designed experiments and analyzed experimental data. T.K. performed most of the experiments. H.K. carried out cell sorting. D.M., T.K. and M.B.S. performed computational analysis. O.J. performed 4C–seq. S.K. carried out some ChIP–seq assays. S.D., H.G. and G.L.-B. contributed to Kir-OE mouse generation and characterization. N.M. analyzed the phenotype of Kir-OE mice. C.S. performed analysis of scRNA-seq. C.S. and P.P. contributed to the Solo RNA-seq analysis; T.K., D.M. and M.B.S. wrote the first draft. F.M.R. revised and wrote the final manuscript.

Competing interests

The authors declare no competing interests.

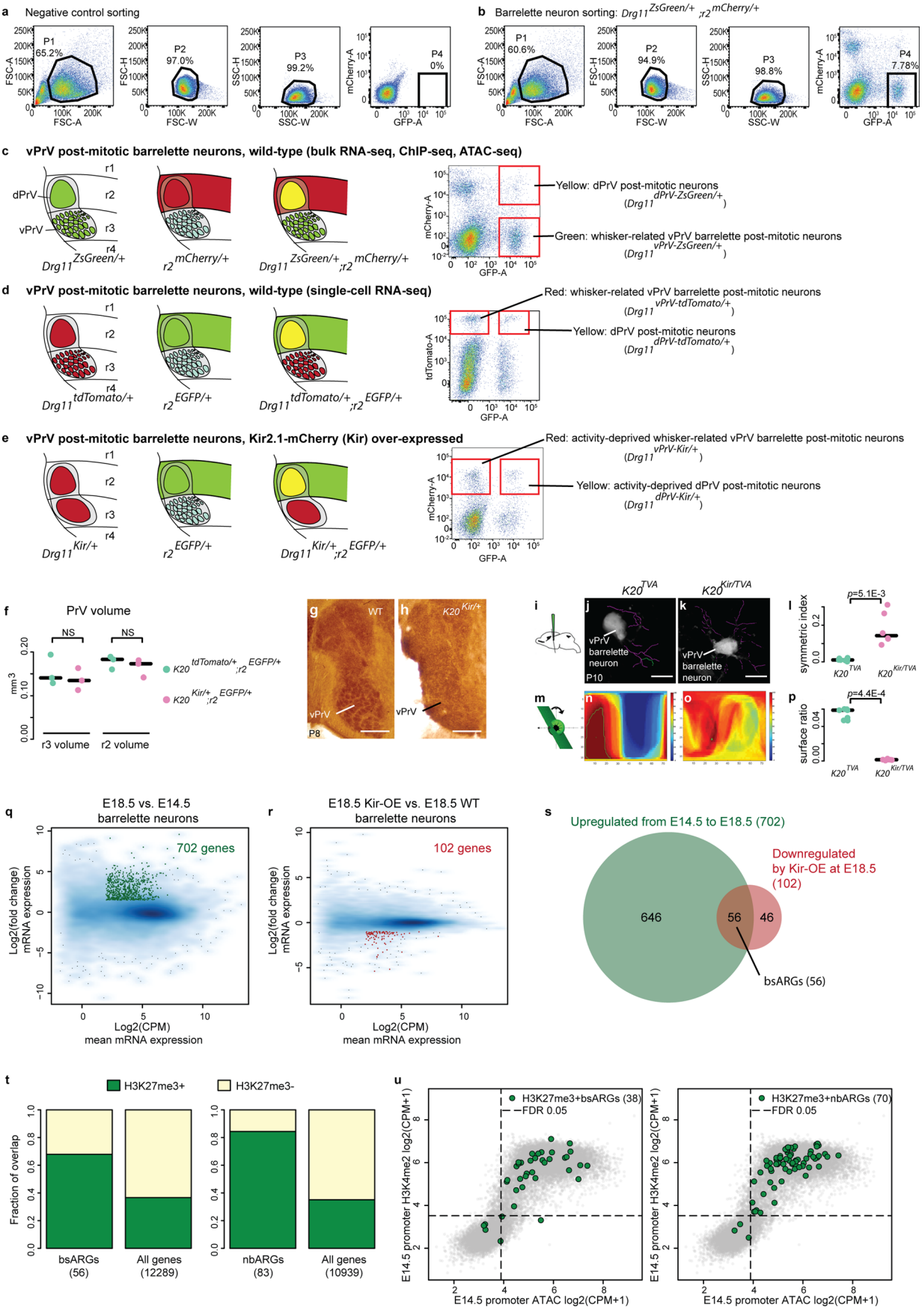
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00789-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00789-z.

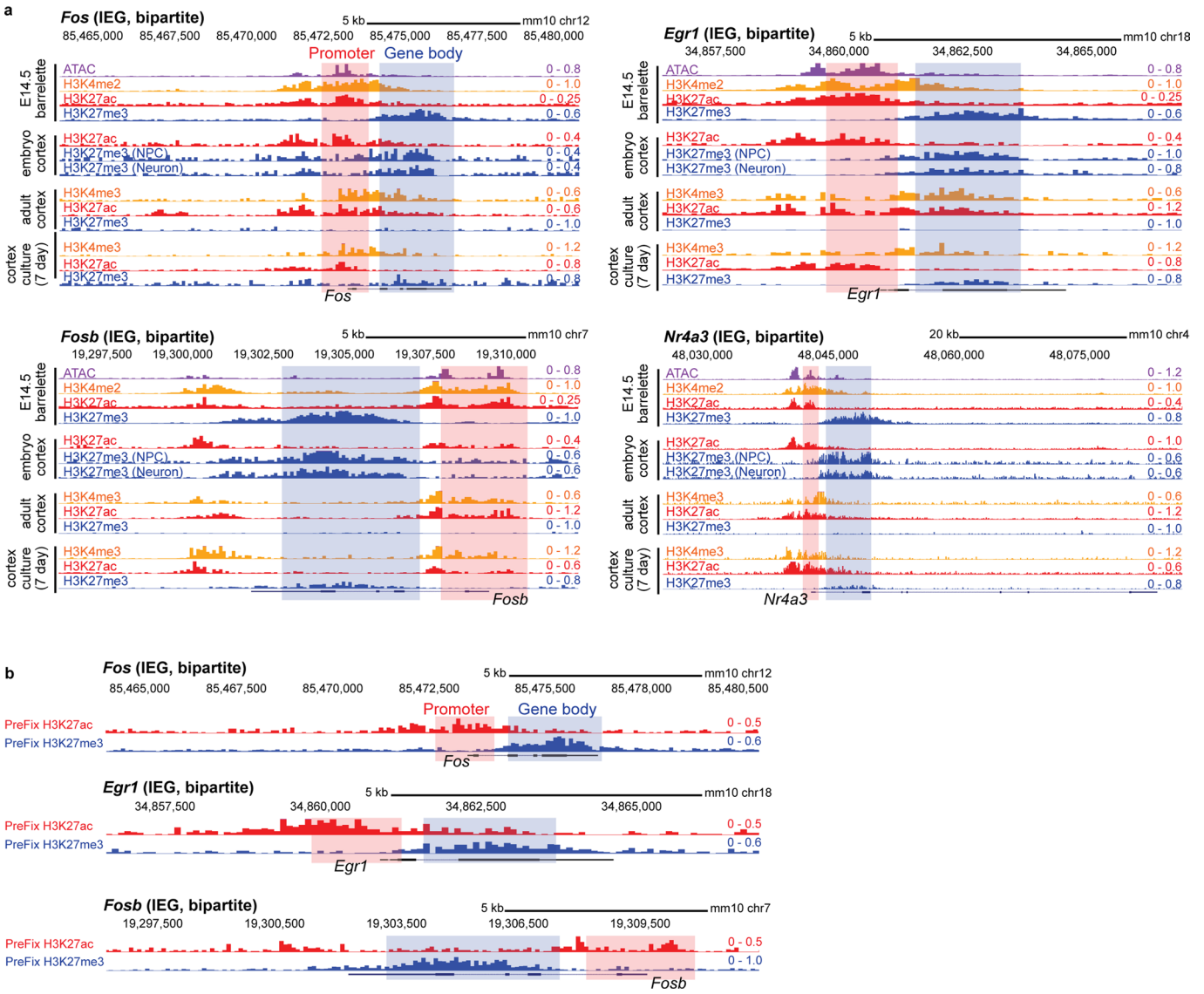
Correspondence and requests for materials should be addressed to F.M.R.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

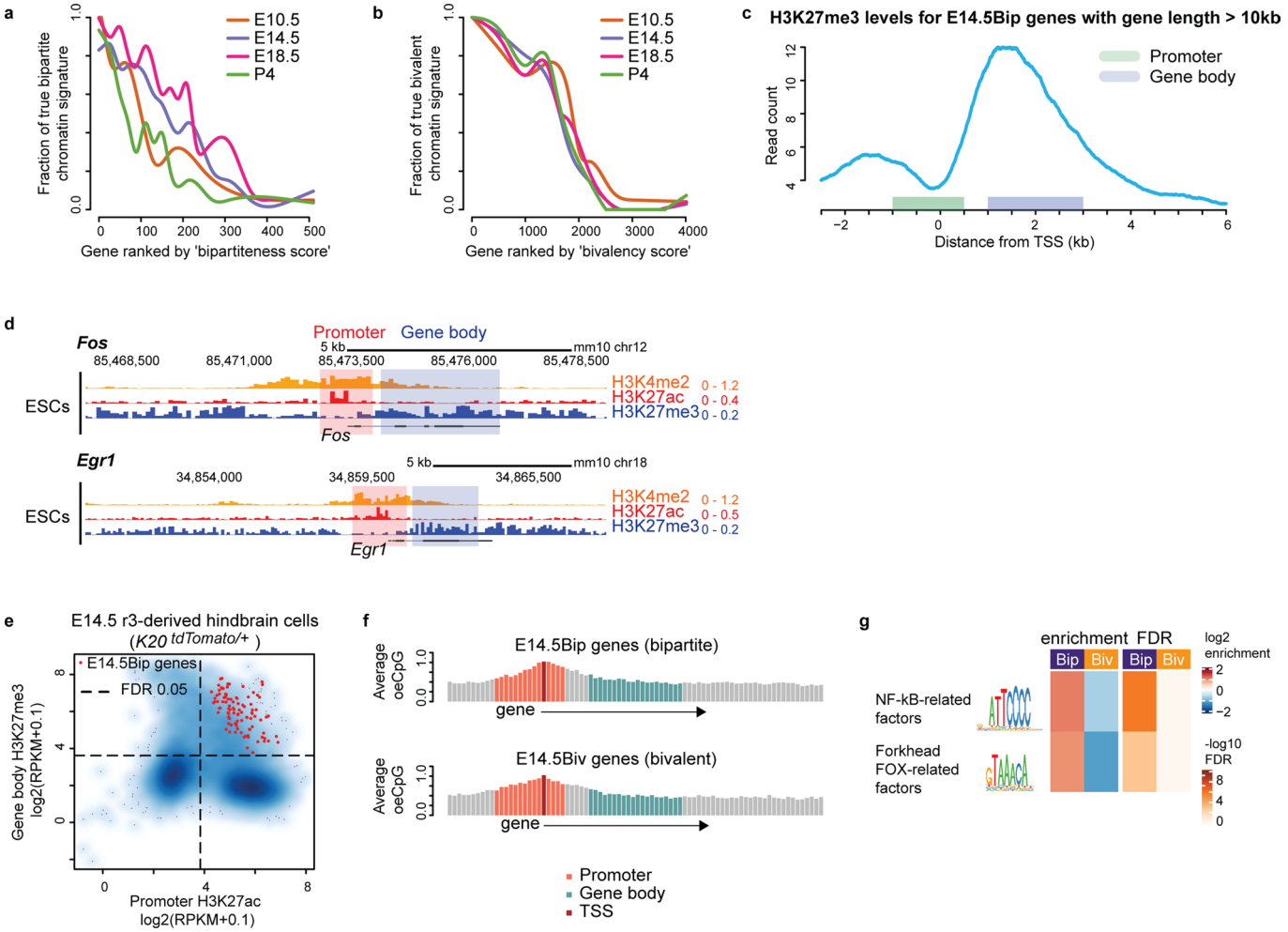


Extended Data Fig. 1 | See next page for caption.

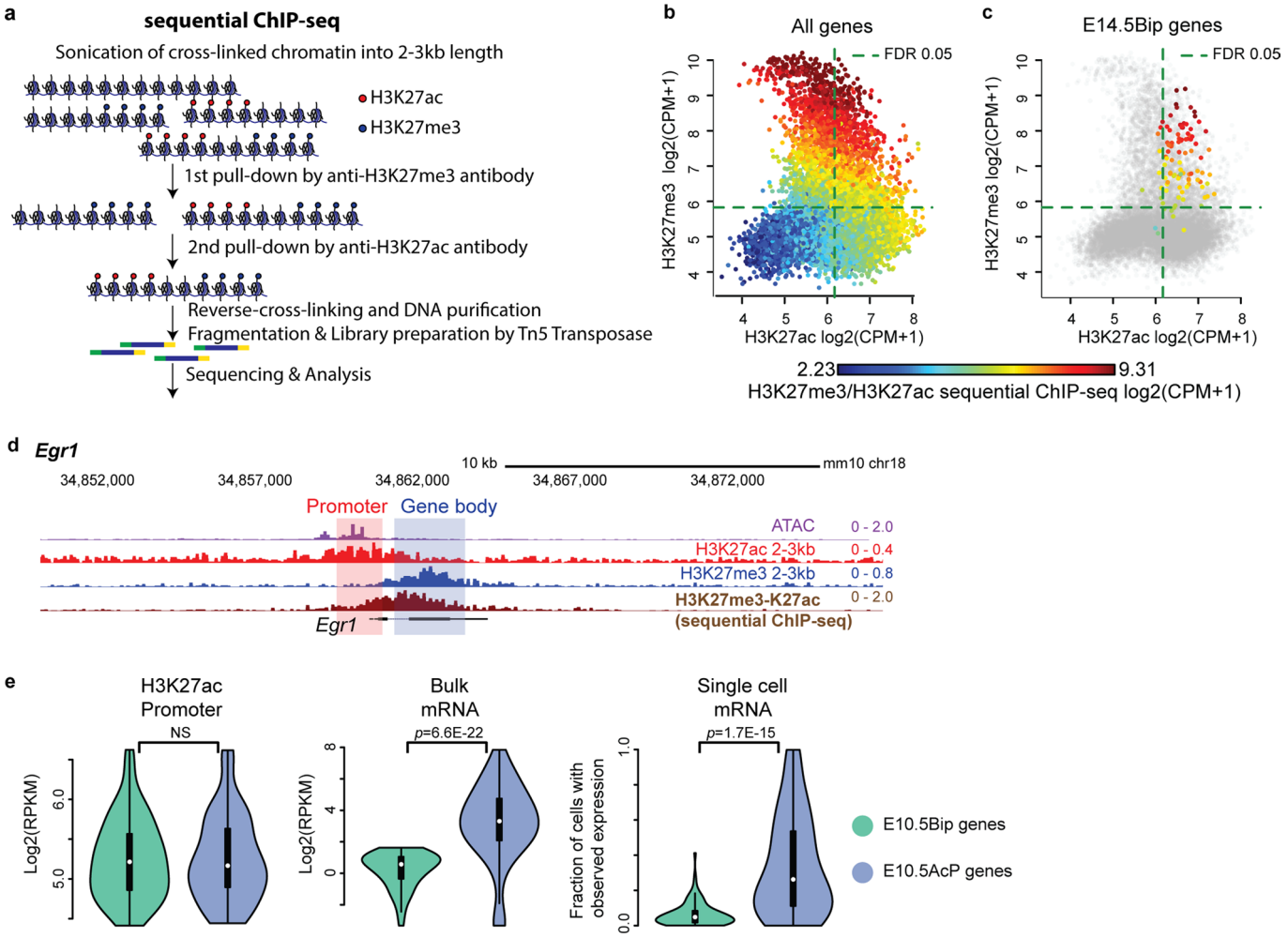
Extended Data Fig. 1 | Genetic strategy of barrelette neuron isolation and identification of activity response genes. a and **b**, Representative FACS gating for barrelette neurons (Supplementary Figs. 1–4). **c** and **d**, Intersectional strategies to FACS isolate E14.5, E18.5 or P4 postmitotic barrelette neurons (*Drg11*$^{vPrV\text{-}ZsGreen/+}$ (**c**), *Drg11*$^{vPrV\text{-}tdTomato/+}$ (**d**)) from ventral principal trigeminal nucleus (vPrV). (Supplementary Table 1, Supplementary Note). **e**, Intersectional strategy to FACS isolate Kir2.1(Kir)-mCherry overexpressing, neuronal activity-deprived, vPrV barrelette neurons (*Drg11*$^{vPrV\text{-}Kir/+}$; Supplementary Table 1, Supplementary Note). **f**, Volumes of PrV in control (*K20*$^{tdTomato/+}$;*r2*$^{EGFP/+}$) and Kir overexpressing (*K20*$^{Kir/+}$;*r2*$^{EGFP/+}$) mice ($n=3$, biologically independent animals). **g** and **h**, P8 cytochrome oxidase (CO) staining in wild-type (WT; **g**) and *K20*$^{Kir/+}$ (**h**) mice. Representative images of $n=3$ biologically independent animals. Scale bars: 200 μm. **i**–**p**, P10 barrelette neuron dendrite orientation by GFP expression after pseudotyped rabies virus (EnvA-SADΔG-GFP) injection in P3 thalamus in control *Krox20::Cre;LSL-R26*$^{TVA\text{-}LacZ}$ (*K20*$^{TVA/+}$) (**j**) and Kir overexpressing *Krox20::Cre;R26*$^{Kir\text{-}mCherry}$;*LSL-R26*$^{TVA\text{-}LacZ}$ (*K20*$^{TVA/Kir}$) (**k**) mice. Scale bars: 10 μm. Symmetry index (**l**) and surface ratio (**p**) are compared (Supplementary Methods). $n=8$ (*K20*$^{TVA/+}$) and $n=6$ (*K20*$^{TVA/Kir}$) biologically independent animals were used, and **j, k, n, o** are representative images. **q** and **r**, MA-plots comparing E18.5 and E14.5 mRNA levels in control *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons (**q**), and E18.5 *Drg11*$^{vPrV\text{-}Kir/+}$ (Kir-OE) and E18.5 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ wild-type (WT) barrelette neurons (**r**). 702 genes (green dots, **q**) increase their expression at E18.5 as compared to E14.5 (log2(fold change) > 1.5), while 102 genes (red dots, **r**) decrease their expression in E18.5 Kir-OE barrelette neurons (log2(fold change) < –1) (Methods). **s**, Identification of 56 bsARGs. **t**, Fractions of the 56 bsARGs (left) and 83 nbARGs (right) with H3K27me3- and H3K27me3+ profiles in E14.5 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons (see Fig. 1c). **u**, Scatterplots showing ATAC–seq (*x* axis) and H3K4me2 (*y* axis) signals on promoters (1 kb around TSS) in E14.5 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons. Dashed lines indicate thresholds corresponding to a 5% false discovery rate (FDR) (Methods) (see Fig. 1c). **f**, **l**, **p**, Bars indicate median and *P* values are from Welch's two-sample two-sided *t*-tests. NS: not significant ($P > 0.05$).



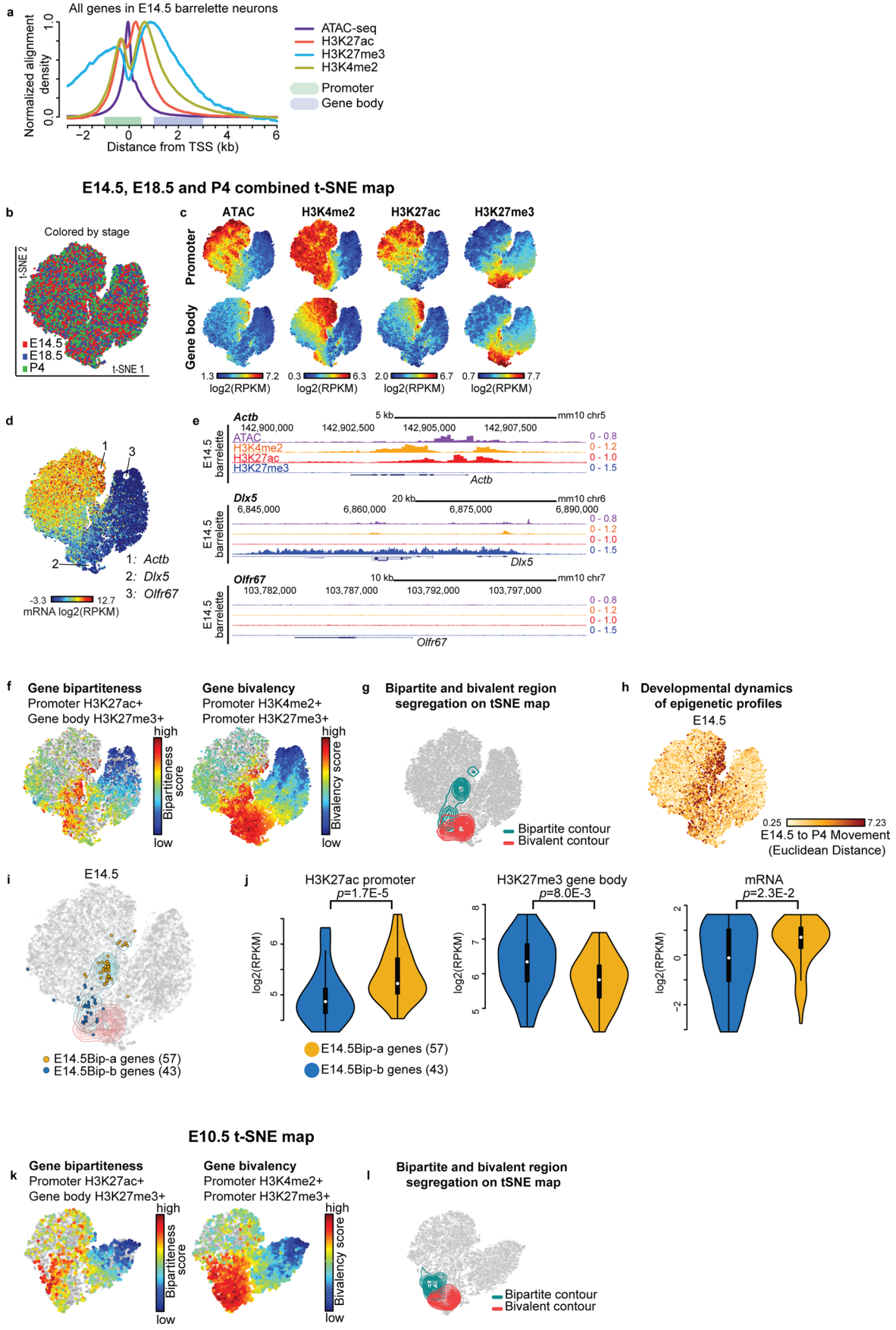
Extended Data Fig. 2 | Chromatin profiles of IEGs carrying the bipartite signature. a, *Fos, Egr1, Fosb* and *Nr4a3* genome browser views. ATAC (violet), H3K4me2 (yellow, E14.5 barrelette neurons), H3K4me3 (yellow, adult cortical neurons and cultured embryonic neurons), H3K27ac (red) and H3K27me3 (blue) are shown in E14.5 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons, E15.5 embryonic cortical neural progenitors (NPC) and postmitotic neurons (PMID: 28793256), adult cortical excitatory neurons (PMID: 26087164), and embryonic cortical 7 day *ex vivo* cultured neurons (PMID: 20393465). Shaded boxes highlight promoters (pink) and gene bodies (blue). IEGs display a bipartite chromatin signature characterized by promoter-H3K27ac and gene body-H3K27me3 in E14.5 barrelette neurons and E15.5 cortical progenitors and postmitotic neurons. In contrast, H3K27me3 is not present on IEG gene bodies in postnatal cortical excitatory neurons, similar to postnatal barrelette neurons (see Fig. 3e and Extended Data Fig. 6b). Also, culturing embryonic cortical neurons for one week results in depletion of the H3K27me3 mark, similar to embryonic hindbrain neuron culture (see Extended Data Fig. 9a). In long genes (for example *Nr4a3*), H3K27me3 deposition on gene body does not stretch throughout the gene body, but is restricted only to the proximal region downstream of the promoter (also see Extended Data Fig. 3c). **b**, IEG (*Fos*, *Egr1* and *Fosb*) genome browser views at E14.5, pre-fixed prior to the dissociation procedure. Shaded boxes highlight promoters (pink) and gene bodies (blue). The presence of the bipartite pattern (H3K27ac + promoter/H3K27me3+ gene body) in the pre-fixed tissue indicates that it is not induced by the dissociation procedure.



Extended Data Fig. 3 | Genome-wide characterization of the bipartite signature. a and **b**, Bipartite (**a**) and bivalent (**b**) gene rank (*x* axis) and corresponding rate of correct bipartite and bivalent classification obtained through genome browser visual inspection of individual loci (*y* axis, fraction of bipartite true-positive, Methods) in E10.5 *K20*$^{tdTomato/+}$ progenitors and E14.5, E18.5 and P4 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons. **c**, Aggregate plot showing profile of H3K27me3 around the transcription start site (TSS) in E14.5Bip genes (top 100 genes ranked by bipartiteness scores in E14.5 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons) with long gene length (>10 kb). Promoters (defined as 1 kb upstream to 500 bp downstream of TSS) and gene bodies (from 1 kb to 3 kb downstream of TSS) are highlighted. Note that H3K27me3 on gene bodies does not stretch further than 2-3 kb downstream of the TSS, even when genes are long. **d**, Genome browser profiles of representative bipartite IEGs (*Fos*, *Egr1*) in mouse ESCs. Note that the H3K27me3 mark is deposited not only on downstream (gene body) but also upstream regions of these H3K27ac promoters. **e**, Scatterplots showing promoter H3K27ac (*x* axis) and gene body H3K27me3 (*y* axis) signals in E14.5 rhombomere 3 (r3)-derived *K20*$^{tdTomato/+}$ hindbrain cells. E14.5Bip genes identified in *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons are mapped (red dots). Dashed lines indicate thresholds corresponding to a 5% false discovery rate (FDR) based on a gaussian mixture model with two components (for foreground and background, see Methods). Barrelette neuron E14.5Bip genes show high levels of promoter H3K27ac and gene body H3K27me3 indicating that they are bipartite also in r3-derived *K20*$^{tdTomato/+}$ hindbrain cells. **f**, CpG average observed/expected (o/e) ratios in a 100 bp window in E14.5Bip and E14.5Biv gene loci. Bins overlapping with the promoter, TSS, and gene body are indicated. **g**, Transcription factor binding motifs specifically enriched in E14.5Bip as compared to E14.5Biv promoters (Methods).

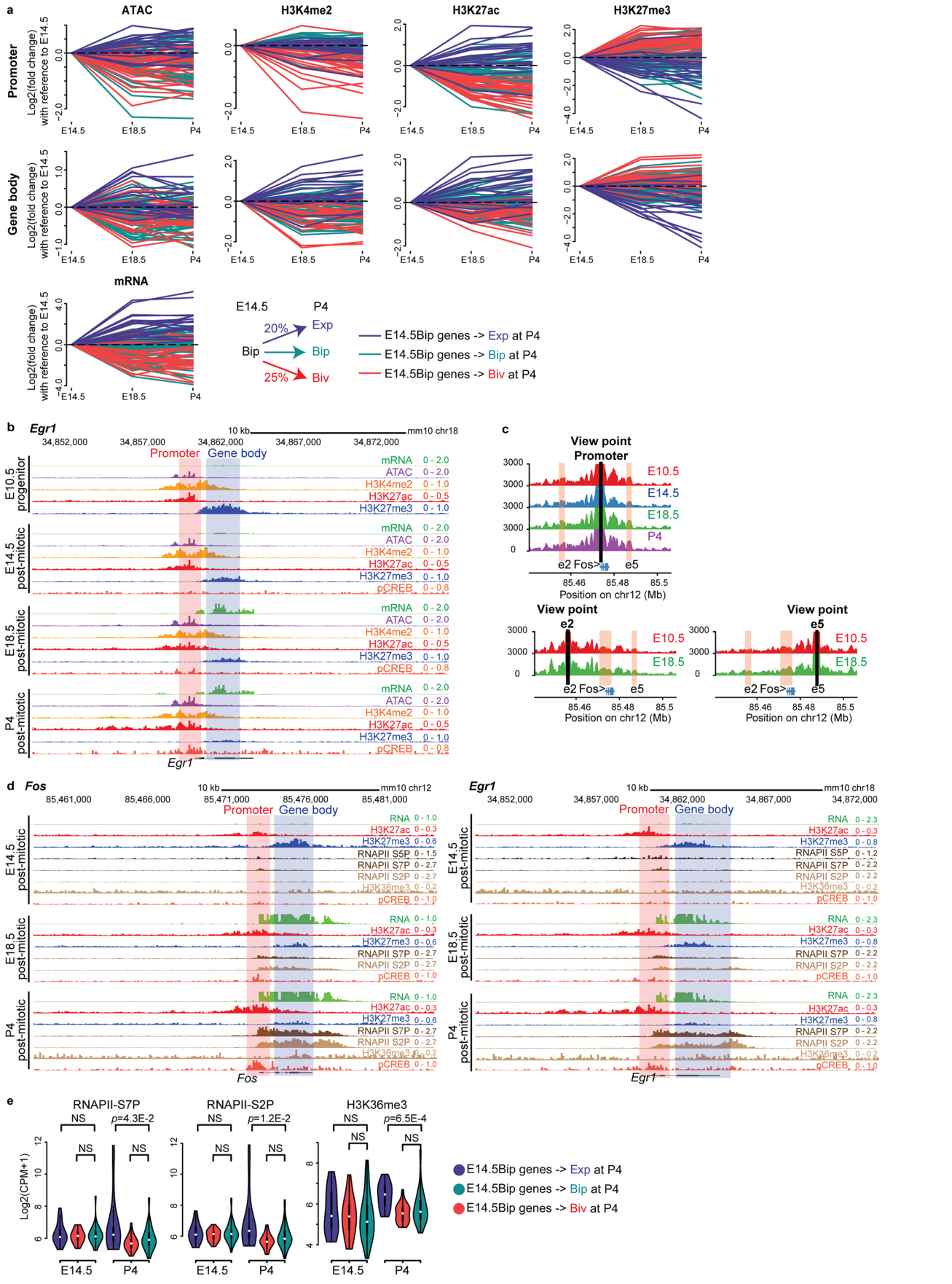


Extended Data Fig. 4 | Coexistence of H3K27ac and H3K27me3 at promoter and gene body of bipartite genes. a, Scheme of sequential ChIP-seq protocol (Supplementary Methods). **b** and **c**, Scatterplots comparing H3K27ac (*x* axis) and H3K27me3 (*y* axis) signals detected in regions from −1 kb to +3 kb around the transcription start site (TSS) by single ChIP-seq performed with large (2-3 kb) chromatin fragments in E14.5 hindbrain tissue. The colors indicate the corresponding H3K27me3/H3K27ac sequential-ChIP-seq signals, either for all autosomal genes (**b**), or only for E14.5Bip genes (**c**). Dashed lines indicate thresholds corresponding to a 5% FDR based on a gaussian mixture model with two components (for foreground and background, see Methods). Stratified by one single ChIP signal (for example H3K27ac), the sequential ChIP signal still correlates with the other (for example H3K27me3), which indicates that single chromatin fragments have been double-marked and thus have been enriched at both steps of the sequential ChIP experiments. **d**, Genome browser view of bipartite *Egr1* gene locus displaying chromatin accessibility (ATAC-seq), 2-3kb-fragment H3K27ac, H3K27me3 and H3K27me3/H3K27ac sequential-ChIP-seq in E14.5 hindbrain tissue. H3K27me3 and H3K27ac coexist on gene body and promoter, respectively. **e**, Violin plots displaying promoter H3K27ac (left), bulk mRNA-seq (middle, Smart-seq2), and single cell fraction with detected mRNA-seq (right, 10X genomics) of E10.5 *K20*$^{tdTomato/+}$ progenitors. E10.5 bipartite (E10.5Bip) genes (green, $n = 99$ genes, see Methods) and E10.5 non-bipartite genes with Bip-matching promoter H3K27ac levels (blue, $n = 99$ genes) are compared. E10.5Bip gene transcripts are only detected in as little as 6% of single cells on average. Plots extend from the data minima to the maxima with the white dot indicating median, the box showing the interquartile range and whiskers extending to the most extreme data point within 1.5X the interquartile range. *P* values are from two-sided Wilcoxon's tests. NS: not significant ($P > 0.05$).



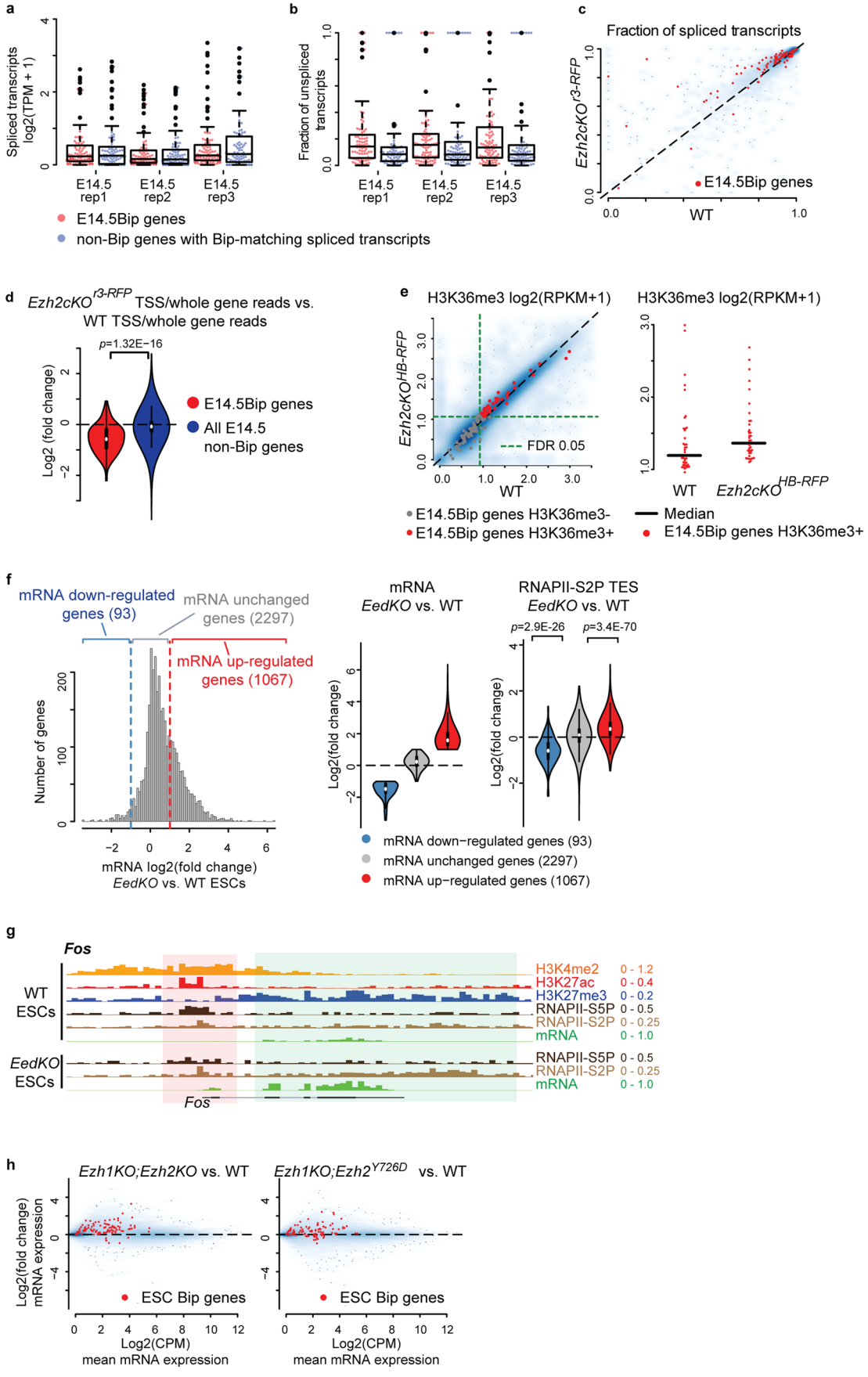
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | t-SNE visualization of mouse genes according to chromatin pattern. a, Aggregate plot of chromatin profiles (ATAC-seq, ChIP-seq) around transcript start sites (TSSs) of all autosomal genes in E14.5 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons. Promoter and gene body regions are highlighted (Methods). **b**-**d**, Two-dimensional (2D) projection on a E14.5, E18.5 and P4-combined t-SNE map of autosomal genes according to chromatin accessibility, H3K27me3, H3K4me2, and H3K27ac levels at promoters and gene bodies in *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons (Methods). **b**, E14.5 (red), E18.5 (blue) and P4 (green) genes are indicated (Supplementary Note). **c**, Color-coded t-SNE gene maps according to promoter (top row) and gene body (bottom row) chromatin profiles (columns). **d**, Color-coded t-SNE gene maps indicating mRNA levels. Numbers 1–3: example genes in E14.5 barrelette neurons. **e**, Chromatin profiles of the example genes in **d**, namely H3K4me2+/H3K27ac+/H3K27me3-/ATAC+(active, *Actb*), Polycomb-dependent H3K4me2+/H3K27ac-/H3K27me3+/ATAC+(permissive, *Dlx5*) and H3K4me2-/H3K27ac-/H3K27me3-/ATAC- (repressed, *Olfr67*). **f**, color-coded E14.5, E18.5 and P4-combined t-SNE gene maps displaying bipartiteness (left) or bivalency (right) scores in E14.5, E18.5 and P4 barrelette neurons (Methods). **g**, E14.5, E18.5 and P4-combined t-SNE map with contour lines indicating regions enriched with bipartite (green) and bivalent (red) genes (Methods). **h**, E14.5, E18.5 and P4-combined t-SNE map in which all E14.5 genes are colored according to their developmental change of chromatin state from E14.5 to P4 (Supplementary Note and Methods). **i**, E14.5Bip genes are subdivided into two subgroups according to their localization on the t-SNE plot ($n=57$, E14.5Bip-a genes, orange dots; $n=43$, E14.5Bip-b genes, blue dots). **j**, Violin plots showing promoter H3K27ac (left), gene body H3K27me3 (middle) and mRNA (right) levels in E14.5Bip-a (orange) and E14.5Bip-b genes (blue) in E14.5 barrelette neurons (Supplementary Note and Methods). Plots extend from the data minima to the maxima with the white dot indicating median, the box showing the interquartile range and whiskers extending to the most extreme data point within 1.5X the interquartile range. *P* values are from two-sided Wilcoxon's tests. **k** and **l**, E10.5 *K20*$^{tdTomato/+}$ progenitor t-SNE maps with bipartiteness or bivalency scores (**k**) and contour lines (**l**) at E10.5 (Methods).



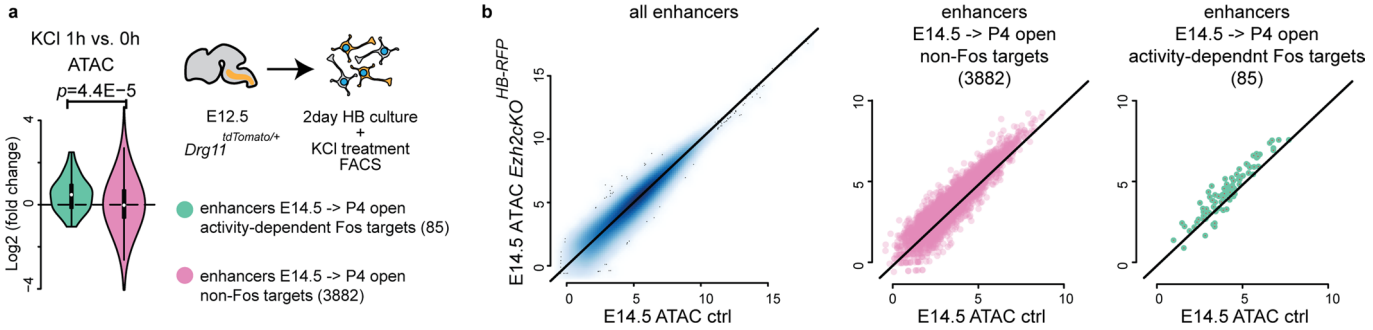
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Developmental dynamics of the bipartite chromatin signature. a, Developmental dynamics of chromatin profiles (ATAC-seq and ChIP-seq signals in promoter and gene body regions) and mRNA levels of E14.5Bip genes in E14.5, E18.5 and P4 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons. Log2 fold changes are calculated with reference to E14.5. At P4, 20% of E14.5Bip genes become expressed (Exp, that is RPKM >= 3 at P4), 25% become bivalent (Biv, red dots in Fig. 3d), and the rest (55%) remain bipartite (Bip), (blue, red and green lines, respectively). Bottom: summary diagram. **b**, Genome browser view of the *Egr1* locus at the E10.5, E14.5, E18.5 and P4 stages. **c**, 3D interaction map (4C-seq) using the *Fos* promoter (top left), enhancer 2 (e2, bottom left) and enhancer 5 (e5, bottom right) as viewpoints in E10.5, E14.5, E18.5 and P4 hindbrain tissue. Normalized read per 4 C fragment is visualized. **d**, Genome browser views of *Fos* and *Egr1* at the E14.5, E18.5 and P4 stages. **e**, Violin plots showing transcription end site (TES, Methods) RNAPII-S7P (left), RNAPII-S2P (middle), and H3K36me3 (right) levels of E14.5Bip genes at E14.5 and P4 (see **a**). E14.5Bip genes that become expressed (Exp, blue, $n = 25$) at P4 displayed significantly higher levels of RNAPII-S7P, -S2P and H3K36me3 marks as compared to E14.5Bip genes that become bivalent (Biv, red, $n = 20$) or remain bipartite (Bip, green, $n = 55$) at P4: violin deviations between groups are to be compared within the same time point since the time points reflect different batches. Plots extend from the data minima to the maxima with the white dot indicating median, the box showing the interquartile range and whiskers extending to the most extreme data point within 1.5X the interquartile range. *P* values are from two-sided Wilcoxon's tests. NS: not significant ($P > 0.05$).

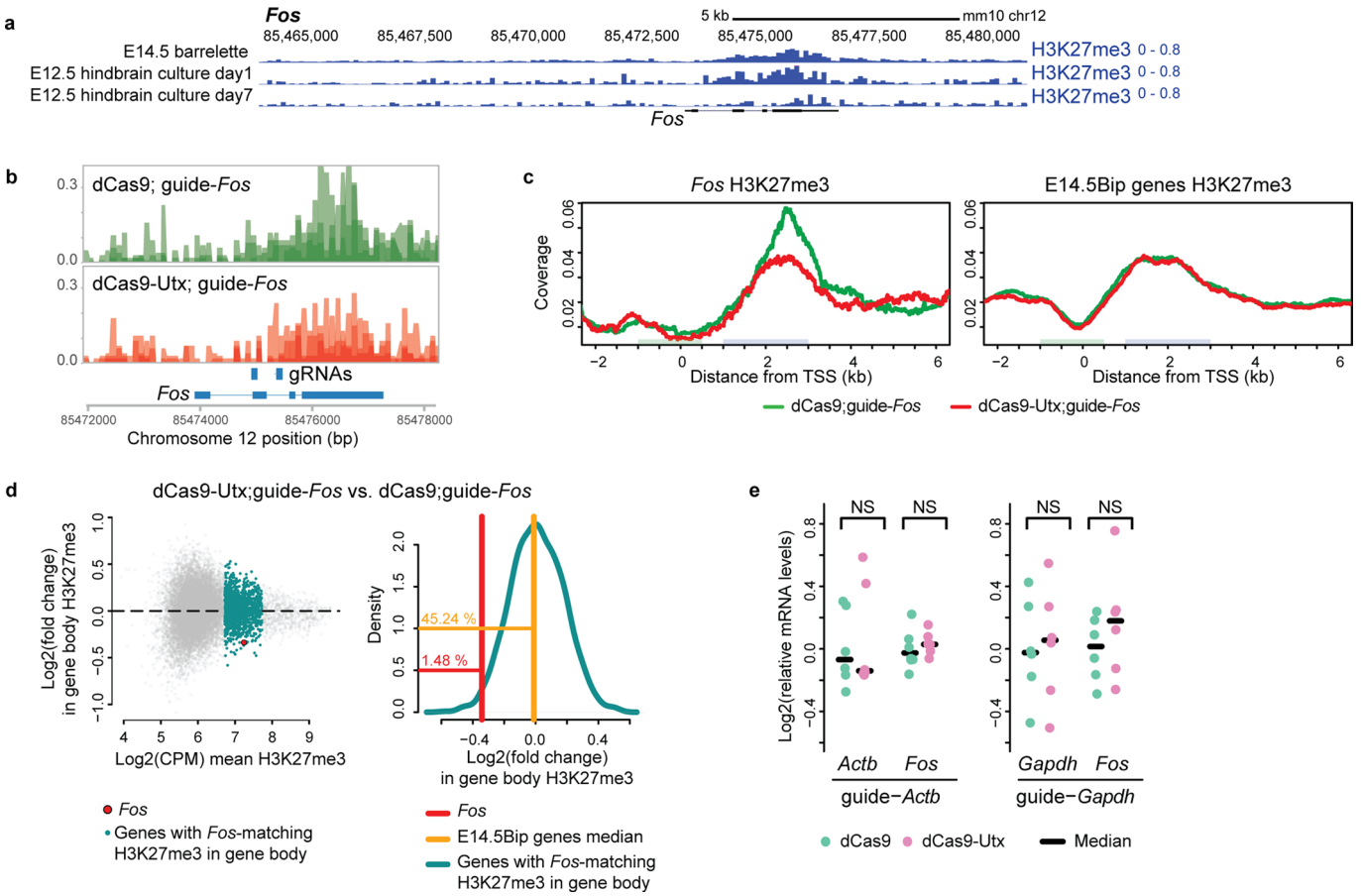


Extended Data Fig. 7 | See next page for caption.

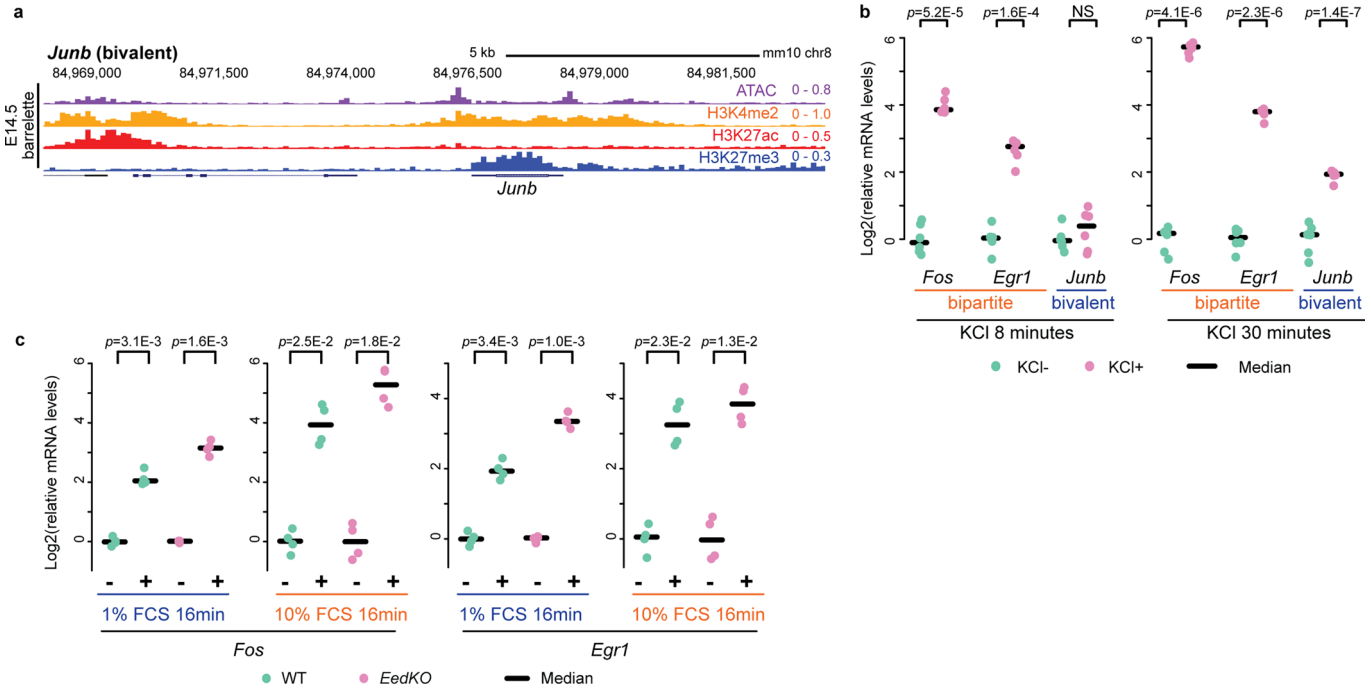
Extended Data Fig. 7 | Polycomb marking of gene bodies inhibits productive mRNA elongation. a, Spliced transcript expression (transcripts per million, TPM, Methods) of E14.5Bip genes (red) and control genes (blue) with matching distributions of spliced transcripts (Methods) ($n = 3$ biologically independent replicates). **b**, Unspliced over total transcript fractions for each gene set in **a** (Methods). Note the larger fraction of unspliced transcripts of E14.5Bip genes compared to control genes. **c**, Fractions of spliced over total transcripts of E14.5Bip genes (red) in E14.5 control $K20^{tdTomato/+}$ (WT) and $Ezh2cKO^{r3\text{-}RFP}$ hindbrain cells ($n = 3$, biologically independent littermates). **d**, Violin plots comparing the TSS/whole gene ratios of total RNA-seq reads between $Ezh2cKO^{r3\text{-}RFP}$ and WT hindbrain cells (see Fig. 5c). **e**, Scatter plot (left) comparing H3K36me3 levels on the coding region between E14.5 wild-type $Hoxa2^{tdTomato/+}$ (WT) and $Ezh2cKO^{HB\text{-}RFP}$ hindbrain cells, indicating H3K36me3-positive and negative E14.5Bip genes in red and gray, respectively (Methods). Increased H3K36me3 levels of H3K36me3-positive E14.5Bip genes in Ezh2-depleted cells are further illustrated (right panel). Bars indicate the median. **f**, (left, middle) Genes with non-zero expression in *EedKO* and wild-type (WT), and carrying H3K27me3 on gene bodies in mouse ESCs ($n = 3457$) were subdivided into genes that show up-regulated (red, $n = 1067$), down-regulated (blue, $n = 93$) and unchanged (gray, $n = 2297$) levels of mRNA in *EedKO* compared with WT ESCs (Supplementary Note, Methods). (right) Violin plots showing log2 fold changes of transcription end site (TES) RNAPII-S2P levels in *EedKO* mutant compared with WT ESCs. **g**, Chromatin profiles of *Fos* in WT and *EedKO* ESCs. While stalled RNAPII-S5P showed decrease in the promoter (pink highlight), elongating RNAPII-S2P was increased (green highlight) in *EedKO* ESCs. **h**, MA-plots comparing mRNA levels of bipartite (Bip) genes ($n = 100$, red) between full *Ezh1KO;Ezh2KO* (left) or Ezh2 catalytically inactive $Ezh1KO;Ezh2^{Y726D}$ (right) with WT ESCs. **d** and **f**, Plots extend from the data minima to the maxima with the white dot or middle bar indicating median, the box showing the interquartile range and whiskers extending to the most extreme data point within 1.5X the interquartile range. *P* value is from a two-sided Wilcoxon's test between the two groups.



Extended Data Fig. 8 | IEG bipartite chromatin is necessary to prevent precocious activity-dependent neuronal maturation. a, Short time (two days) E12.5 *ex vivo*-cultured *Drg11*$^{tdTomato/+}$ hindbrains. After over-night (o/n) treatment with a cocktail of neuronal activity blockers (TDN cocktail = TTX + D-AP5 + NBQX, inhibitors of sodium channel, NMDAR and AMPAR), cultured neurons were treated by 55 mM KCl for 1 hour. Drg11-positive immature trigeminal neurons were FACS-isolated for ATAC-seq analysis. Violin plots visualize log2 fold changes of enhancer chromatin accessibilities in 1 hour KCl-treated neurons as compared to non-treated control neurons. Increased accessibility is selectively detected in KCl-treated neurons at activity-dependent Fos-binding enhancers that normally become open only at P4 (green, $n = 85$ enhancers) (purple, all non-Fos-binding enhancers that gain accessibilities only at P4, $n = 3882$ enhancers). Plots extend from the data minima to the maxima with the white dot indicating median, the box showing the interquartile range and whiskers extending to the most extreme data point within 1.5X the interquartile range from the box. *P* value is from a two-sided Wilcoxon's test. **b**, Scatterplots comparing enhancer accessibilities (ATAC) in E14.5 *Ezh2* heterozygous control (ctrl) and homozygous mutant (*Ezh2cKO*$^{HB\text{-}RFP}$) hindbrain cells. All the barrelette enhancers in *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons (left), non-Fos-binding enhancers that gain accessibilities at P4 as compared with E14.5 in *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons ($n = 3882$ enhancers, middle, purple), neuronal activity-dependent Fos-binding enhancers that gain accessibilities at P4 as compared with E14.5 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons ($n = 85$ enhancers, left, green) are shown (Methods). 85 activity-dependent Fos-binding enhancers show precocious opening upon H3K27me3 removal at E14.5. Also see Fig. 5d.



Extended Data Fig. 9 | dCas9-UTX overexpression in E12.5 short-term ex vivo cultured neurons. a, H3K27me3 profiles of the *Fos* locus in E14.5 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons and E12.5 cultured hindbrain neurons at day 1 and day 7 of culture. One week hindbrain neuron culture results in the loss of the H3K27me3 mark from the *Fos* gene body, similarly to one-week embryonic cortical neuron culture (Extended Data Fig. 2a); in contrast, short-term (day 1) cultured hindbrain neurons still maintain H3K27me3 levels comparable to E14.5 barrelette neurons. **b**, H3K27me3 levels at the *Fos* locus in short-term cultured E12.5 hindbrain neurons overexpressing control dCas9 (green) or dCas9-UTX (red) targeted to *Fos* gene body: three biological replicates overlaid. Genomic regions targeted by guide-RNAs (gRNAs) are indicated. **c**, Averaged H3K27me3 profiles of *Fos* (left) and the rest of the E14.5Bip genes (right). Overexpression of control dCas9 (green) or dCas9-UTX (red) are compared. **d**, (left) MA-plot comparing H3K27me3 levels of dCas9-UTX against dCas9 targeting to *Fos* locus. *Fos* (red dot) shows a loss of gene body H3K27me3 compared to control genes carrying similar levels of H3K27me3 (green dots). (right) Density plot (green line) shows the distribution of the logFC values of the selected control genes (greed dots), highlighting the *Fos* logFC (red line) in the 1.48 % lower tail of the density plot and logFC of E14.5Bip genes (yellow line), indicating slight but significant decrease in *Fos* H3K27me3. $n = 3$ biologically independent neuron cultures. **e**, mRNA levels of *Actb*, *Gapdh* and *Fos* determined by RT-qPCR in short-term cultured E12.5 hindbrain neurons overexpressing control dCas9 (green) or dCas9-UTX (purple) targeted to *Actb* (left) or *Gapdh* (right) loci ($n = 6$ biologically independent neuron cultures). The median expression is indicated by bars. *P* values are from Welch's two-sample two-sided *t*-tests. NS: not significant ($P > 0.05$).



Extended Data Fig. 10 | Polycomb marking of bipartite gene bodies regulates the rapidity and amplitude of transcriptional response to relevant stimuli. **a**, Genome browser view of bivalent *Junb* in E14.5 *Drg11*$^{vPrV\text{-}ZsGreen/+}$ barrelette neurons. Chromatin accessibility (ATAC), H3K4me2, H3K27ac and H3K27me3 are shown. **b**, mRNA levels of bipartite (*Fos, Egr1*) and bivalent (*Junb*) ARGs, determined by RT-qPCR in E14.5 hindbrain cells treated by KCl for 8 (left) or 30 min (right) ($n = 4$, biologically independent embryos). The median expression is indicated by bars. *P* values are from Welch's two-sample two-sided *t*-tests. NS: not significant ($p > 0.05$). **c**, mRNA levels of *Fos* (left) and *Egr1* (right), determined by RT-qPCR in serum-starved WT (green) and *EedKO* (purple) mouse ESCs ($n = 4$, biologically independent cultured cells) treated with a low (1%) or high (10%) concentration of Fetal Calf Serum (FCS) for 16 minutes. WT ESCs *Fos* and *Egr1* could be induced only after prolonged exposure (that is 16 minutes) to 1% FCS: also see the effects of shorter (that is 8 minutes) time exposure to 1% FCS in Fig. 5f and g. The median expression is indicated by bars. *P* values are from Welch's two-sample two-sided *t*-tests.

3

COMPUTATIONAL METHODS TO IDENTIFY REGULATORY MOTIFS

This part of the thesis focuses on the computational approaches we developed that help us identify functional motifs in a given data set coming from an experiment, typically a sequencing experiment. We implemented and combined these methods as a software package in R. This chapter focuses on the package that we developed, called *monaLisa*, short for "**mo**tif a**na**lysis with **Lisa**", inspired by *Homer*, a well-known tool used for motif analyses. *monaLisa* allows the user to identify potentially important motifs that explain certain experimental observations either in a binned or a regression-based approach.

This project was a collaboration with Lukas Burger, Charlotte Soneson and Michael Stadler. My contributions include implementing the regression-based approach using stability selection to select meaningful motifs, as well as re-implementing, together with Michael, the specific *Homer* functionalities monaLisa uses in R. The package is publicly available on GitHub at https://github.com/fmicompbio/monaLisa.

3.1 BACKGROUND AND OVERVIEW

There are several tools and methods that do various types of motif analyses. When it comes to identifying motifs that are enriched in a particular set of sequences, *Homer* (Heinz, Benner, *et al.*, 2010) and *MEME* (Bailey, Boden, *et al.*, 2009) are some of the most commonly used tools. Another way to find important motifs is by making use of linear regression. Tools like *REDUCE* (Roven & Bussemaker, 2003) and *ISMARA* (Balwierz, Pachkov, *et al.*, 2014) have used such approaches to identify regulatory motifs that are likely to explain a given a set of observations. *monaLisa* makes use of such types of approaches, applying a binned way to calculate motif enrichments similarly to *Homer* or using an alternative to lasso regression, namely randomized lasso using stability selection (Meinshausen & Bühlmann, 2010) to select meaningful motifs.

Homer has many functions that can do various types of analyses on sequencing data. Here, we give an overview of how it finds significantly enriched motifs using a know set of motifs. *Homer* can identify motifs that are significantly enriched in a given set of sequences we call foreground sequences, compared to a set of background sequences. *Homer* will adjust for differences in GC content and sequence-composition between foreground and background, before doing enrichment tests. Specifically, it will first calculate weights for the background sequences to adjust for differences in GC content. These weights will then be iteratively updated to also adjust for sequence-composition differences up to a certain k-mer size, where a k-mer is a subsequence of size k and k is 3 by default. The result is that each foreground sequence will have a weight of one, and each background sequence will have a weight that adjusts for the sequence-composition differences compared to foreground. These weights are used when counting the number of sequences. Next, for each motif, *Homer* will

scan all sequences for hits in zero or one occurrence per sequence (ZOOPS) mode. That means that a given sequence will have a value of one if it has at least one TFBS for the motif, or a value of zero otherwise. For each motif, and using the previously calculated weights each sequence has, the number of foreground sequences with a motif hit, and the number of background sequences with a hit is calculated. Statistical hypothesis testing indicates how significantly enriched each motif is. The p-values are calculated with a binomial test, using the number of foreground sequences with a motif hit, the total number of foreground sequences, and the fraction of background sequences with a motif hit as the probability of success of a trial. The p-values are log-transformed, and the motifs are ordered by significance. *Homer* can thus for a given set of foreground and background sequences calculate p-values indicating significance of enrichment for a given set of known motifs.

Homer will test for the enrichment of each motif independently. On the other hand, a regression-based approach would jointly look for important motifs, with them competing against each other for importance. Regression-based approaches have been implemented in other tools. With *monaLisa*, however, we implement a regression method using the concept of stability selection proposed by Meinshausen & Bühlmann (2010). Specifically, we implement randomized lasso stability selection. Stability selection can be applied to any regression method, and particularly has advantages when the number of predictors exceeds the number of observations. Subsets of the response vector and corresponding rows of the predictor matrix are randomly selected multiple times, and a regression is done on each of those subsets. When using lasso regression for example, where predictors are selected by setting non-selected predictors to zero, stability selection helps select predictors that are consistently chosen and robustly explain the observations we have. The result is that each predictor gets a selection probability value, which is estimated by the fraction of times it was selected. The randomized lasso is a generalization of the lasso. Instead of penalizing each $|\beta_k|$ with the same penalty λ, the penalty term is predictor-specific and is a randomly chosen value in $[\lambda, \lambda/\alpha]$ where α is called a weakness term and $\alpha \in (0, 1]$ (Meinshausen & Bühlmann, 2010). The estimator is

$$\hat{\beta}^{\lambda,W} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k} \right) \tag{3.1}$$

where W_k are independent and identically distributed random variables in $[\alpha, 1]$ for $k = 1, \ldots, p$. When $\alpha = 1$, the equation above is equivalent to the lasso case. Meinshausen & Bühlmann (2010) demonstrate advantages to using randomized lasso stability selection for consistency in selection, better performance in noisy data sets and tighter error control. To some extent, it also performs better than other regression methods when predictors are correlated to each other.

In *monaLisa*, we provide two main methods to find relevant motifs. The first option is to find enriched motifs in a binned approach as first demonstrated by Ginno, Burger, *et al.* (2018), applying sequence-composition corrections similarly to how *Homer* does it. Applying the motif enrichment in a binned manner, where each bin associates a particular set of sequences to an experimental value like changes in DNA methylation, links the motif enrichment to the measured values. The second method uses randomized lasso stability selection with the improved error bounds proposed by Shah & Samworth (2013) to find relevant motifs. The regression-based approach does not by default correct for sequence-composition differences. However, differences that need to be adjusted for, like GC content differences between sequences, can be directly added as another column in the

predictor matrix, along with any artifact that needs to be accounted for, which makes this approach quite flexible as well.

3.2 DATA SETS IN *MONALISA*

In *monaLisa*, we provide two data sets that are used as examples in the vignette of the package, to illustrate how the user can identify relevant regulatory motifs using the approaches described so far. The first data set comes from DNA methylation experiments, where bisulfite sequencing data was produced for mouse pluripotent embryonic stem (ES) cells and neuronal progenitors (NP), and the data was generated as described by Stadler, Murr, *et al.* (2011). Briefly, lowly methylated regions (LMRs) were identified and for each LMR, the difference between NP and ES cells in terms of the average fraction of methylated CpGs was calculated.

The second data set comes from ATAC-seq experiments that are publicly available on *Encode* under accession numbers *ENCFF146ZCO*, *ENCFF109LQF*, *ENCFF203DOC*, and *ENCFF823PTD*. This data set consists of two replicates of mouse lung tissue and two replicates of mouse liver tissue, all at postnatal day zero. We called accessibility peaks genome-wide per condition using *MACS2* (Zhang, Liu, *et al.*, 2008). Only autosomal and distal peaks that were at least 1,000 bp away from any TSS were kept, followed by merging the peaks from both conditions and quantifying the ATAC-seq reads per sample across these peaks using *QuasR* (Gaidatzis, Lerch, *et al.*, 2015). Counts per million were calculated and averaged per condition, followed by getting $\log_2$ fold changes for liver vs lung tissue and randomly selecting 10,000 peaks to keep. We finally saved the genomic ranges of the kept peaks along with their computed $\log_2$ fold changes in accessibility.

3.3 BINNED APPROACH

monaLisa provides functions that can perform a motif enrichment analysis in a binned manner. That means that motif enrichments are calculated for a set of sequences belonging to the same bin vs a defined background set. The background sequences used in these calculations can be one of:

- *otherBins*: using all sequences belonging to the other bins (excluding the current one) as background. This is the default option.
- *allBins*: using the sequences from all bins (including the current bin).
- *zeroBin*: using sequences belonging to the zero bin, if a zero bin has bin defined.
- *genome*: using randomly sampled regions from the genome, that are similar in size and GC composition to the foreground set.

monaLisa does the binned enrichment analysis in R with the *calcBinnedMotifEnrR* function using PWMs from a reference database like *JASPAR* (Tan, 2021). The sequence-composition corrections and enrichment p-value calculations that take place per bin are similar to how *Homer* does this. Instead of calculating enrichment p-values per motif using the binomial test, however, *monaLisa* uses Fisher's exact test by default, but allows for the possibility of a binomial test should the user choose to do so. In addition, other assays like the negative $\log_{10}$ p-values, the negative $\log_{10}$ adjusted p-values correcting for multiple testing, the motif $\log_2$ enrichment values, the motif

enrichments as Pearson residuals, the expected number of foreground sequences with a motif hit, the actual number of foreground sequences with a motif hit, and the number of background sequences with a motif hit are also calculated. All of these results are conveniently stored as a *SummarizedExperiment* object in R (Morgan, Obenchain, *et al.*, 2021). In the end, all of these values can be associated to the bin they came from to see if there are patterns. For example, we can take the LMRs described in section 3.2, bin the regions according to levels of methylation differences, and use the *calcBinnedMotifEnrR* function in *monaLisa* to calculate motif enrichments per bin. In this manner we can link different levels of the DNA methylation differences to the extent of motif enrichment as depicted in figure 3.1.

3.3.1 *Re-implementing* Homer*'s sequence-composition correction in R*

In order to avoid depending on *Homer* being pre-installed on a user's machine, we re-implemented the functionalities that we need for *monaLisa* in R. Specifically, we implemented the correction for sequence-composition differences between foreground and background that is done in *Homer* when using the *findMotifsGenome.pl* function with options *-size given*, *-bg backgroundSeqs.bed*, *-nomotif* and *-mknown*.

To validate that we were successfully re-implementing *Homer*, we ran both our implementation and *Homer*'s on several data sets. Figure 3.2 shows the runs on the data sets that are in the *monaLisa* package. The figure shows the negative $\log_{10}$ p-values each motif gets to indicate significance of enrichment, as well as the sum of the weights of the sequences in fore- and background which are used to calculate these p-values. We attribute the small fluctuations to rounding errors and conclude that we can reproduce the results reasonably well.

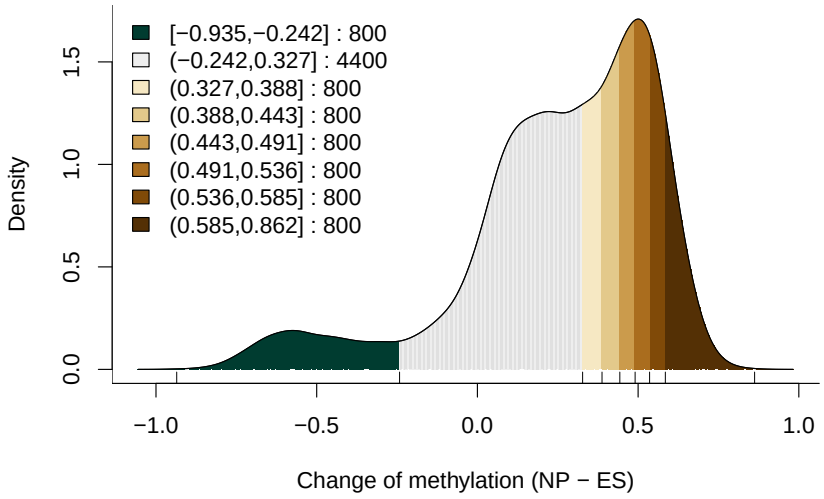
Aside from removing the *Homer* software dependency, we were able to efficiently implement the binned enrichment analysis in *monaLisa*, and were thus able to considerably decrease the run-time of this function. The run-times in figure 3.2 were around 11- to 17-fold faster with the *monaLisa* implementation. The current implementation allows the user to parallelize the process across the bins, by specifying the number of cores available for use.

Re-implementing the *Homer* sequence-composition correction in R inside *monaLisa* allowed us to better understand how the tool is doing this adjustment for sequence-composition differences between foreground and background sequences, as well as how it calculates enrichment p-values per motif. This gave us the opportunity to also create a more efficient implementation with improved parallelization. It also saves the user of *monaLisa* from needing to have *Homer* pre-installed in order to be able make use of the binned approach in the package to calculate motif enrichments.

3.4 REGRESSION APPROACH

With the binned approach, motifs are tested for enrichment significance independently of each other. However, in order to select for regulatory motifs in a manner that allows these motifs to compete against each other for selection, a regression-based approach is fitting. We implemented the randomized lasso stability selection method from Meinshausen & Bühlmann (2010) with the improved error bounds proposed by Shah & Samworth (2013). The *stabs* package (Hofner &

(a)



(b)

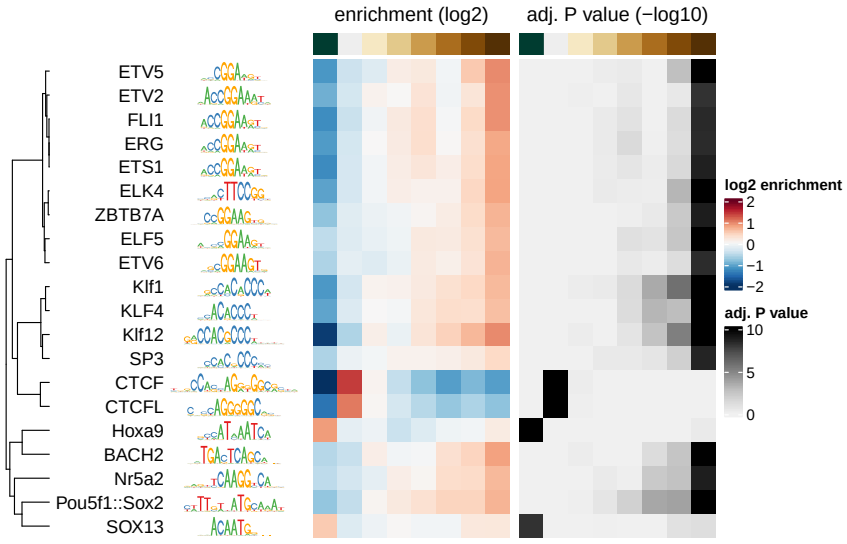


FIGURE 3.1: Binned motif enrichment analysis with *monaLisa* using the vertebrate motifs from the *JASPAR2018* database and the DNA methylation data set present in *monaLisa*. (a) Histogram of difference in methylation across genomic regions (the LMRs) between NP and ES cells, binned according to the levels of the difference. (b) Result of a binned motif enrichment analysis with *monaLisa*. The motifs with highest $\log_2$ enrichment values were selected and their enrichment and significance values as $-\log_{10}$(adjusted p-values) are displayed in the heat maps for each motif per bin.

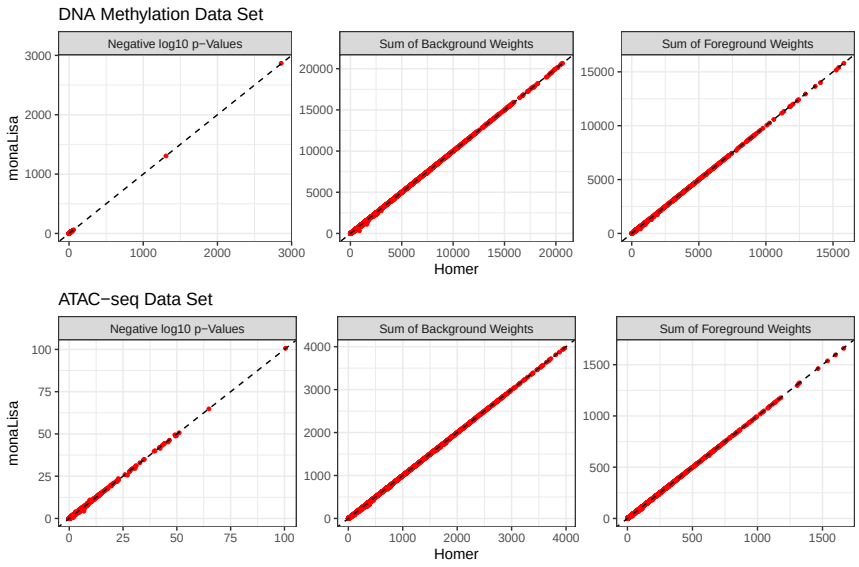


FIGURE 3.2: Reproducing *Homer*'s sequence-composition correction and p-value calculation in R. We implemented the way *Homer* calculates p-values per motif to indicate significance of enrichment. This includes filtering problematic sequences, correcting for GC content and sequence-composition differences, and calculating p-values per motif to test for enrichment significance using a binomial test. The x-axis shows the results from the *Homer* run and the y-axis the results from the implementation in *monaLisa*. Each red point is a motif in a specific bin and the black dashed line is the $y = x$ line. We use the data sets described in section 3.2 to illustrate the successful implementation, showing the negative $\log_{10}$ p-values (left), number of background sequences with a motif hit (center) and number of foreground sequences with a motif hit (right).

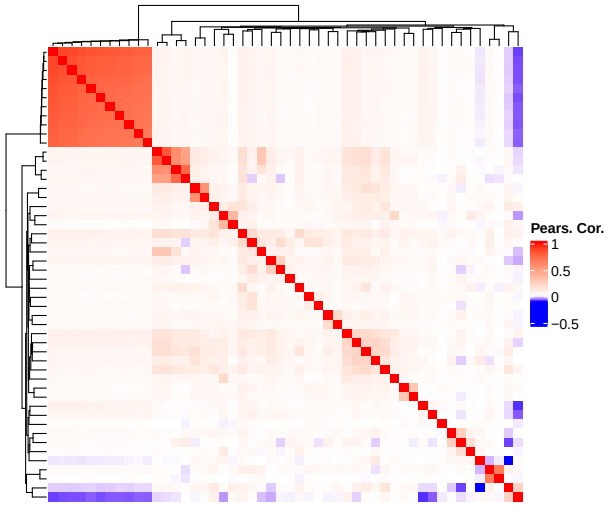


FIGURE 3.3: Pearson correlation between the predictors chosen as our true signal in the simulated predictor matrix. There is a wide range of correlation with some being even relatively high.

Hothorn, 2021) has implemented this for lasso stability selection, and so we adjusted that function to perform the randomized lasso. Specifically, we adjusted the *glmnet.lasso* function that is used by the *stabsel* function. *glmnet.lasso* uses the *glmnet* package (J. H. Friedman, Hastie & Tibshirani, 2010) to do the individual regressions. To motivate our choice for the randomized lasso stability selection in *monaLisa*, we compared it to a few other regression methods that are typically used for variable selection, namely lasso and elastic-net regression. We created a simulated data set to compare these approaches, with the predictor matrix containing a correlation structure we typically see in real settings.

3.4.1 *Simulated data set*

As mentioned in section 3.1, Meinshausen & Bühlmann (2010) have shown randomized lasso stability selection to perform better than lasso stability selection in noisy settings, that the performance is better in settings of highly correlated predictors, and that lasso stability selection does better than lasso with cross validation in terms of error control. To re-illustrate these claims and evaluate the methods on a correlation structure for the predictor matrix we usually see in real data sets, we created a synthetic example. We used the DNA methylation data set in *monaLisa*, and created the real predictor matrix from this data set with regions as rows and motifs as columns. Each entry in this matrix is the predicted number of binding sites a motif has in a specific region. To create the simulated predictor matrix, we used a negative binomial distribution per motif, with mean and

variance values coming from the real predictor matrix, to generate simulated values for binding sites. The real correlation structure between predictors was then introduced into the simulated one using Cholesky decomposition. We additionally introduced even higher correlation between some predictors since we have also observed such instances in other real data sets, and regression methods tend to generally suffer more with highly correlated variables. We then randomly selected 50 out of 500 predictors as our true signal, and summed them to create a response variable. Figure 3.3 shows the correlation between the chosen signal predictors. We tested the regression methods on several response variables with varying signal to noise ratios (SNR), as well as several parameters specific to the regression methods used. We chose to run lasso stability selection, randomized lasso stability selection, lasso with cross validation and elastic net with cross validation. On the one hand we wanted to see if stability selection outperforms cross validation methods. On the other hand, we also wanted to see if randomized lasso does better than the lasso in stability selection. We also wanted to see how it compares to elastic-net regression with cross validation since this would be the common method of choice if highly correlated variables are a concern. For more details on how these data sets were generated and how the regressions were done see appendix A.2.

Figure 3.4 depicts the results of these runs. Since stability selection and cross validation are stochastic, we ran a regression with the same set of parameters five times and averaged the true positive rate (TPR) and false positive rate (FPR) results to get a robust estimate. Generally, per set of parameters, all five runs were very similar to each other in TPR and FPR. From figure 3.4, we can appreciate that regression with stability selection tends to be on the conservative side with tight error control, as all runs had an FPR below 5 percent at all times. However, it would also rather select nothing than something incorrect, which is why it can have low TPRs. Increasing the tolerance for error however still shows stability selection to act conservatively, maintaining an FPR below 5 percent. Within stability selection, the randomized lasso controls for errors slightly better than the lasso. The regressions with cross validation, on the other hand, show less control of FPR. A reason for this is that these methods tend to co-select the signal predictors as well as non-signal predictors that happen to be correlated to the true ones. With elastic-net regression, there is additionally the *alpha* parameter that needs to be varied to make reasonable selections and it is not clear how that value is best chosen in practice.

3.4.2 *Implementation in* monaLisa

To illustrate how the described regression-based approach can be used to select meaningful motifs in *monaLisa* using the *randLassoStabSel* function, we used the ATAC-seq data set found in the package and described in section 3.2. It consists of ATAC-seq peaks, where the $\log_2$ fold change in accessibility between liver and lung was quantified. In the regression-based approach, this fold change is our response variable. The predictor matrix is the peaks by motifs matrix, with entries corresponding to the predicted number of binding sites a motif has. We used PWMs from the set of vertebrate motifs from the *JASPAR* database using the *JASPAR2018 bioconductor* package (Tan, 2021). Using randomized lasso stability selection, we can identify what motifs are likely to explain the observed changes in accessibility. In this example, we used a weakness value of 0.8, a probability cut-off of 0.8 (predictors with selection probabilities greater than this value will be selected), and a per-family error rate (PFER) value of 2 which indicates the number of falsely selected variables we

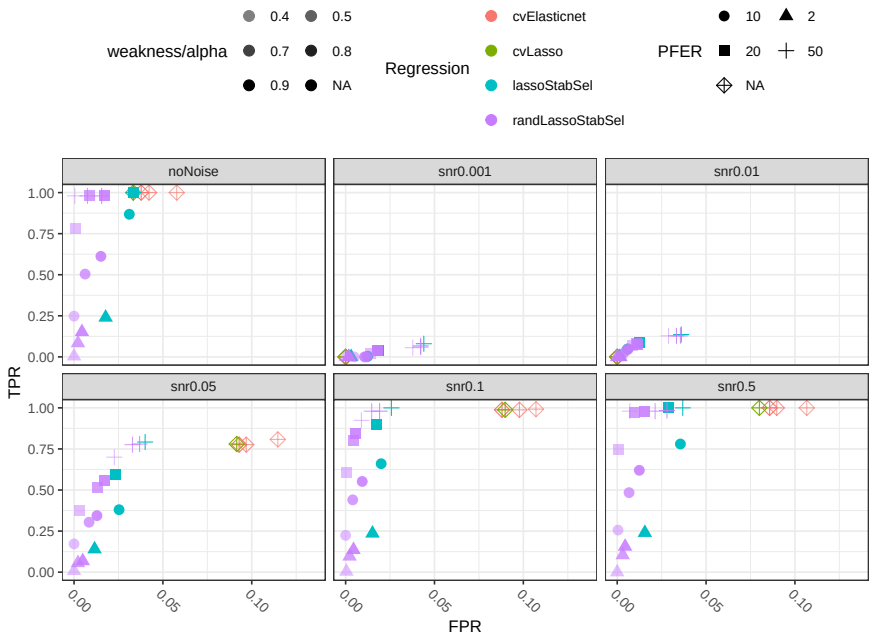


FIGURE 3.4: Comparing regression methods on the simulated data set. Each point is the average value of five runs with the same parameters. Stability selection methods generally better control for FPR than the cross validation methods. They also tend to rather choose nothing than something incorrect, as some runs have lower TPR values. Increasing the PFER value, which is the number of false positives allowed among the selected variables, increases the TPR while still controlling the FPR. Within the stability selection methods, randomized lasso has tighter error control than lasso. In the very noisy examples, it is generally difficult to select meaningful predictors, though error control is still maintained. The cross validation methods fare less well with the FPR, as they tend to co-select the true signal as well as non-signal predictors that are correlated to the true ones. The points are varyingly transparent depending on the *weakness* parameter (values of 0.4, 0.7, or 0.8) in the case of randomized lasso stability selection, or on the *alpha* parameter (values of 0.5, 0.7 or 0.9) in the case of elastic-net regression with cross validation.

allow for. Figure 3.5 shows the motifs that were selected with this approach and figure 3.6 shows the Pearson correlation between the selected motifs.

3.5 OTHER FUNCTIONALITIES

As described in detail in sections 3.3 and 3.4, *monaLisa* offers the *calcBinnedMotifEnrR* and *randLassoStabSel* functions to do a binned motif enrichment analysis, or regression-based selection of motifs with randomized lasso stability selection, respectively. The package also provides some plotting functions for each of those approaches to visualize the results (see figures 3.1 and 3.5), as well as the possibility to visualize GC content or other dinucleotide composition differences across pre-defined bins.

The package also offers the possibility to do a k-mer based binned enrichment analysis. This approach does not require pre-defined motifs, and instead looks for enriched sub-sequences of size k across the bins. This allows the user to potentially identify enriched k-mers that were not present in a motif database.

monaLisa can also be used to scan for motif hits across the genome or a given set of sequences relatively quickly. There, we have more efficiently implemented the *matchPWM* function from the *Biostrings* package (Pagès, Aboyoun, *et al.*, 2021). One can scan the provided sequences for hits using a PWM for each motif, and a provided threshold score to be considered a hit.

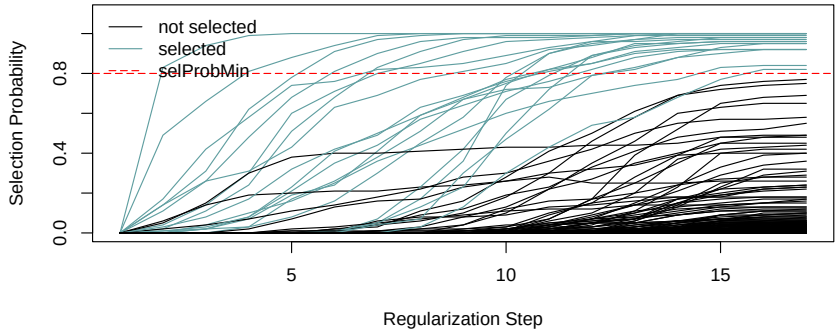
We also still provide the possibility for the user to use a pre-installed version of *Homer* to do the binned analysis with the *calcBinnedMotifEnrHomer* function. Since *Homer* requires a motif text file containing the PPMs of the motifs, whereas our implementation in *calcBinnedMotifEnrR* makes use of PWMs as a *PWMatrixList* object in R (Tan & Lenhard, 2016), we have also provided functions that can convert between these two formats.

3.6 DISCUSSION AND OUTLOOK

With *monaLisa*, we have made an R package that provides useful tools to identify potentially regulatory motifs. The package contains DNA methylation and ATAC-seq data sets to demonstrate how this can be achieved. However, any type of sequencing data could be used: ChIP-seq, HiC-seq, RNA-seq, single cell RNA-seq, single cell ATAC-seq etc. — any measure that can be quantified on a given set of regions or sequences. It makes use of existing *bioconductor* (Gentleman, Carey, *et al.*, 2004) objects and packages which makes it easy to integrate into other analyses that use *bioconductor* packages.

Implementing the way *Homer* adjusts for sequence-composition differences between fore- and background sequences allowed us to speed up and parallelize certain calculations in the binned approach. It also allowed us not to require the user of *monaLisa* to have *Homer* pre-installed on their machine in order to be able to do a binned motif enrichment analysis. As mentioned, this approach does a reasonable job, correcting for sequence-composition differences. However, when there are extreme differences in sequence-composition between the different bins, it is good to be aware of such differences and if they affect which motifs are enriched in the specific bins.

(a)



(b)

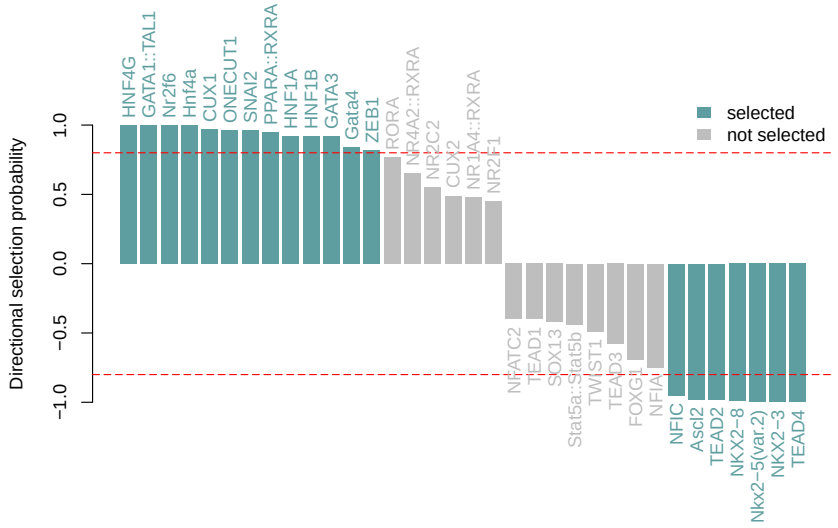


FIGURE 3.5: Randomized lasso stability selection in *monaLisa* using the ATAC-seq dataset. Selected motifs are highlighted in green and the probability cut-off as a dashed red line. (a) Stability paths per motif. We see the selection probability as a function of the regularization step. The area under the curve (AUC) of these paths can further indicate how important a motif is. (b) Bar plot showing the selection probabilities of individual motifs, multiplied by the sign of the Pearson correlation between a motif from the predictor matrix and the response variable. Positive and negative correlations can indicate a directionality in the sense that the motif explains positive (liver) or negative (lung) $\log_2$ fold changes, respectively.

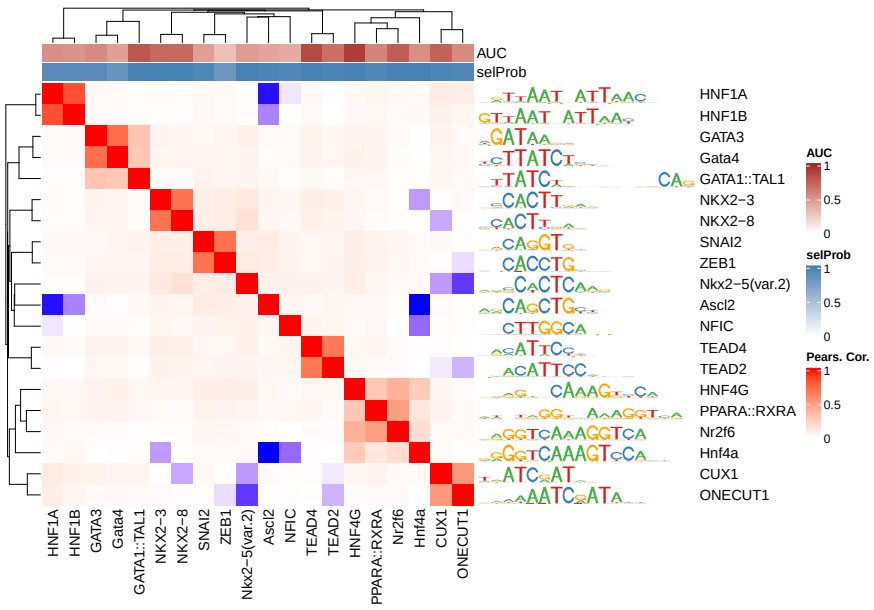


FIGURE 3.6: Heat map of the Pearson correlation in the predictor matrix, for the motifs that were selected with the regression. We can appreciate that some motifs with high correlations were co-selected, like HNF1A and HNF1B. The AUC of the stability path of a motif can give further indication of how important a motif is. A motif that is selected even under strict regularization with a high penalty term, and then remains selected as the penalty is relaxed can indicate a motif that is very likely to explain the response variable at hand (observed changes in accessibility).

The regression-based approach with randomized lasso stability selection allows motifs and any other feature of interest to compete for selection, to explain the observed experimental values. The predictor matrix consists of the number of TFBS across a set of genomic regions or sequences. TF binding could also be quantified in zoops mode in this matrix. TFs that are consistently selected end up with high selection probabilities. This approach is particularly useful when there are too few observations, or when the signal to noise ratio is very low. In practice, we see the method also co-select TFs that are highly correlated to each other. However, extremely correlated predictors are still problematic in any regression-based approach. Regression with stability selection, however, seems to have a much better error control than other regression-based approaches with cross validation. If a lot of predictors are very highly correlated to each other, the user may want to consider pre-grouping them, to have a representative predictor per group, before using the regression-based approach. This is something that may also need to be further investigated in the future, on how to handle such cases or best do such a grouping, for example by following the ideas of Bühlmann, Rütimann, *et al.* (2013).

Both described approaches offer an unbiased way to select meaningful TFs, using a public database of the TF motifs. The binned k-mer enrichment analysis further offers the opportunity to find enriched k-mers, and detect those that belong to motifs that are not present in the public database. There is room for further development with this approach in terms of grouping the enriched k-mers and trying to construct the motifs they may have come from.

We think this will be a tool that users, especially those already familiar with *bioconductor* packages, will be able to use fairly easily. It will allow them to discover regulatory motifs that could explain their data or experiments, and provide a list of potential candidates for follow-up experiments or validations.

4

DISCUSSION

Studying epigenetic changes in developing neurons led us to the discovery of bipartite genes, which we also find in many other tissues. The promoters of these genes are accessible and active, ready for fast induction. The presence of the repressive H3K27me3 on the first 2,000 bp of the gene body hinders any productive expression of mRNA. Once the right developmental cues come, the repressive mark gets removed and the genes can be expressed. We found bipartite genes to be some of the most dynamic genes when looking at the genome-wide changes in chromatin states from one developmental time point to another in the barrelette neurons in mice. Furthering our understanding of what regulators are responsible for these dynamics, how this is controlled, and how the chromatin of bipartite genes interacts with other regions in three dimensions requires further study and exploration.

On the other hand, being able to computationally identify relevant TFs that could for example explain the changes in epigenetic features between different conditions has been the aim with *monaLisa*. The binned motif enrichment analysis tests for enriched motifs per bin. *Homer* corrects for GC content and general sequence composition differences between fore- and background sequences before testing for significance in enrichment. Reproducing these processes in R allowed us to implement the binned enrichment analyses in *monaLisa*, without depending on the *Homer* tool. It also offered us the flexibility to do a more efficient and faster implementation of the underlying processes, with the option to parallelize along the bins. The binned approach tests for the enrichment of each motif independently. If we want to rather select a set of motifs and have them compete against each other for this selection, the available regression-based approach is fitting. This framework offers the flexibility to add any feature of interest on the given regions, like GC content, to correct for. Regression using stability selection is particularly useful when there is not much signal in the data, or there are too few observations. In simulated data sets, we have seen the method select variables rather conservatively, keeping a tight error control.

monaLisa integrates well with other *bioconductor* packages. We are currently in the process of submitting it to *bioconductor* to have it be available for others through that community of open-source software for computational biology. The two approaches described use a database of known motifs to find regulatory TFs. For de-novo motif discovery, there is still room for development in *monaLisa*. Open questions remain, for example, on how enriched k-mers can be used to reconstruct the motif or motifs they came from. There is also room for exploring how multiple conditions could be jointly used when finding regulatory TFs. When we have two conditions, we can calculate log fold changes in a measure of interest and ask what TFs explain these fold changes. With more than two condition, one may need to calculate fold-changes with respect to the average of all conditions. This would still require multiple runs of a binned enrichment analysis, or of the regression-based approach. In the regression setting, it may be worth exploring multivariate regression, where more than one response variable is modeled in a single regression.

Using the appropriate computational tools in chapter 2 has allowed us to study epigenetic signatures and their dynamics genome-wide. Specifically, using t-SNE to integrate quantifications of different histone modifications and chromatin accessibility, across all genes and in several developmental time points, enabled the joint visualization of changes in chromatin states. Identifying TFs that could potentially be responsible for such changes is possible through tools like *monaLisa*, and provides further hypotheses to test or TFs to experimentally follow up with. Having many different data types and measurements on the same set of features, like genomic regions, poses challenges in integrating them all in meaningful and informative ways. However, having so much information also presents opportunities for new discovery and understanding and for utilizing methods in computational statistics to challenge or further our knowledge in biology.

1. Sagan, C. *The demon-haunted world : science as a candle in the dark* (Ballantine Books, New York, 1997).
2. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737 (1953).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860 (2001).
4. Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. & Hume, D. A. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics* **8**, 424 (2007).
5. Metzker, M. L. Sequencing technologies — the next generation. *Nature Reviews Genetics* **11**, 31 (2010).
6. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* **13**, 613 (2012).
7. Zabidi, M. A. & Stark, A. Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. *Trends in Genetics* **32**, 801 (2016).
8. Bestor, T. H. & Coxon, A. Cytosine methylation The pros and cons of DNA methylation. *Current Biology* **3**, 384 (1993).
9. Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L. & Schübeler, D. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575 (2015).
10. Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. & Paul, C. L. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences* **89**, 1827 (1992).
11. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376 (2012).
12. Giorgetti, L., Servant, N. & Heard, E. Changes in the organization of the genome during the mammalian cell cycle. *Genome Biology* **14**, 142 (2013).
13. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics* **2**, 292 (2001).

14. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. & Dekker, J. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289 (2009).

15. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. & de Laat, W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature Genetics* **38**, 1348 (2006).

16. Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research* **10**, 2997 (1982).

17. Schneider, T. D. & Stephens, R. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **18**, 6097 (1990).

18. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10**, 1213 (2013).

19. Li, G. & Reinberg, D. Chromatin higher-order structures and gene regulation. *Current Opinion in Genetics & Development*. *Chromosomes and expression mechanisms* **21**, 175 (2011).

20. Lawrence, M., Daujat, S. & Schneider, R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics* **32**, 42 (2016).

21. Dong, X. & Weng, Z. The correlation between histone modifications and gene expression. *Epigenomics* **5**, 113 (2013).

22. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497 (2007).

23. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**, 669 (2009).

24. Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L. & Lander, E. S. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315 (2006).

25. Minoux, M., Holwerda, S., Vitobello, A., Kitazawa, T., Kohler, H., Stadler, M. B. & Rijli, F. M. Gene bivalency at Polycomb domains regulates cranial neural crest positional identity. *Science* **355**, eaal2913 (2017).

26. Kitazawa, T. & Rijli, F. M. Barrelette map formation in the prenatal mouse brainstem. *Current Opinion in Neurobiology. Developmental Neuroscience* **53**, 210 (2018).

27. Fowler, T., Sen, R. & Roy, A. L. Regulation of Primary Response Genes. *Molecular Cell* **44**, 348 (2011).

28. Okuno, H. Regulation and function of immediate-early genes in the brain: Beyond neuronal activity markers. *Neuroscience Research* **69**, 175 (2011).

29. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559 (1901).

30. Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579 (2008).

31. Kitazawa, T., Machlab, D., Joshi, O., Maiorano, N., Kohler, H., Ducret, S., Kessler, S., Gezelius, H., Soneson, C., Papasaikas, P., López-Bendito, G., Stadler, M. B. & Rijli, F. M. A unique bipartite Polycomb signature regulates stimulus-response transcription during development. *Nature Genetics* **53**, 379 (2021).

32. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer New York, New York, NY, 2009).

33. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning* (Springer New York, New York, NY, 2013).

34. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38**, 576 (2010).

35. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202 (suppl_2 2009).

36. Roven, C. & Bussemaker, H. J. REDUCE: an online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic Acids Research* **31**, 3487 (2003).

37. Balwierz, P. J., Pachkov, M., Arnold, P., Gruber, A. J., Zavolan, M. & Nimwegen, E. v. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research* (2014).

38. Meinshausen, N. & Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417 (2010).

39. Ginno, P. A., Burger, L., Seebacher, J., Iesmantavicius, V. & Schübeler, D. Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape. *Nature Communications* **9**, 4048 (2018).

40. Shah, R. D. & Samworth, R. J. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 55 (2013).

41. Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Nimwegen, E. v., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K. & Schübeler, D. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490 (2011).

42. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

43. Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M. B. QuasR: quantification and annotation of short reads in R. *Bioinformatics* **31**, 1130 (2015).

44. Tan, G. *JASPAR2018* Bioconductor. http : / / bioconductor . org / packages / JASPAR2018/ (2021).

45. Morgan, M., Obenchain, V., Hester, J. & Pagès, H. *SummarizedExperiment: SummarizedExperiment container* version 1.22.0. 2021.

46. Hofner, B. & Hothorn, T. *stabs: Stability Selection with Error Control* version 0.6-4. 2021.

47. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1 (2010).

48. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. *Biostrings: Efficient manipulation of biological strings* version 2.60.2. 2021.

49. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555 (2016).

50. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. & Zhang, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80 (2004).

51. Bühlmann, P., Rütimann, P., van de Geer, S. & Zhang, C.-H. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference* **143**, 1835 (2013).

A

APPENDIX

nature portfolio

https://doi.org/10.1038/s41588-021-00789-z

Supplementary information

A unique bipartite Polycomb signature regulates stimulus-response transcription during development

In the format provided by the authors and unedited

Supplementary Information for

A Unique Bipartite Polycomb Signature Regulates Stimulus-Response Transcription during Development

Taro Kitazawa,[1,5] Dania Machlab,[1,2,3,5] Onkar Joshi,[1] Nicola Maiorano,[1] Hubertus Kohler,[1] Sebastien Ducret,[1] Sandra Kessler,[1] Henrik Gezelius,[4,6] Charlotte Soneson,[1,2] Panagiotis Papasaikas,[1,2] Guillermina López-Bendito,[4] Michael B. Stadler,[1,2] and Filippo M. Rijli[1,3*]

[*] Correspondence to: filippo.rijli@fmi.ch

[5]These authors contributed equally to this work

This PDF file includes:

Supplementary Note

Identification of activity-regulated genes in barrelette neurons

To identify activity-regulated IEGs and LRGs in developing barrelette neurons, we set out a genetic strategy to isolate E10.5 mitotic progenitors and postmitotic barrelette neurons at E14.5 (early postmitotic), E18.5 (perinatal) and P4 (consolidated barrelette stage) by fluorescence activated cell sorting (FACS) (Fig. 1a, Extended Data Fig. 1a-e, Supplementary Figs 1 and 2, Methods). For easier reference, we have identified each mouse genotype with specific abbreviations and summarized all relevant details in Supplementary Table 1. Briefly, E10.5 tdTomato^{+} mitotic progenitors were obtained by crossing the *Krox20::Cre* mouse line, which labels rhombomere 3 (r3) derivatives including ventral principal trigeminal nucleus (vPrV) barrelette progenitors[1], with the Cre-dependent $R26^{tdTomato}$ floxed reporter line ($K20^{tdTomato/+}$) (Supplementary Table 1, Supplementary Fig. 1). Postmitotic vPrV barrelette neurons were FACS-isolated by an intersectional genetic strategy. Briefly, the postmitotic *Drg11::Cre*[2] line crossed with the Cre-dependent $R26R^{ZsGreen}$ line to generate $Drg11^{ZsGreen/+}$ line (Supplementary Table 1), labeling both the lower jaw-innervated dorsal PrV (dPrV) and whisker-associated vPrV barrelette neurons[2]. $Drg11^{ZsGreen/+}$ mice were further crossed with the *r2::mCherry* ($r2^{mCherry/+}$) transgenic line[2], expressing mCherry in r2 derivatives including dPrV neurons (yellow). We then selectively sorted ZsGreen^{+} (green) vPrV barrelette neurons ($Drg11^{vPrV\text{-}ZsGreen+}$) (Supplementary Table 1; Extended Data Fig. 1c; Methods; Supplementary Fig. 2).

To achieve genetic silencing of neuronal activity in vPrV barrelette neurons, we crossed the Cre-inducible $R26^{Kir\text{-}mCherry}$ line (*Kir*)[3], conditionally overexpressing the Kir2.1 inward rectifier potassium channel fused to mCherry, with either *Krox20::Cre* ($K20^{Kir/+}$) or *Drg11::Cre* ($Drg11^{Kir/+}$) lines (Supplementary Table 1). Genetic silencing of barrelette neuron activity resulted in vPrV of normal size (Extended Data Fig. 1f) but lacking whisker-related barrelette neuron map representation by CO staining (Extended Data Fig. 1g,h), due to asymmetric dendritic tree organization of vPrV barrelette neurons (Extended Data Fig.1i-p, see Supplementary Methods), resembling the phenotype induced by whisker sensory deprivation[4-7]. To obtain mRNA-seq (Smart-seq2) profile of perinatal (E18.5) activity-deprived barrelette neurons, we collected E18.5 Kir-overexpressing postmitotic barrelette neurons ($Drg11^{vPrV\text{-}Kir/+}$), by mating $Drg11^{Kir/+}$ mice and

the *r2::EGFP* (*r2*$^{EGFP/+}$) line and selectively FACS sorting mCherry^{+} neurons (red) (Extended Data Fig. 1e; Supplementary Table 1; Methods; Supplementary Fig. 2).

Next, to identify the activity response genes (ARGs) induced in barrelette neurons at the beginning of the sensory-dependent maturation period (that spans E18.5-P2/3)[8], we compared mRNA-seq profile of FACS-sorted E18.5 activity-deprived *Drg11*$^{vPrV\text{-}Kir/+}$ barrelette neurons to those of E14.5 and E18.5 wild-type barrelette neurons. Among the genes with undetectable or low basal expression level (reads per kilobase per million mapped reads, RPKM < 3) in E14.5 barrelette neurons, we identified those genes (*n*=56) that were up-regulated at E18.5 compared with E14.5, and down-regulated in E18.5 Kir-overexpressing, activity-silenced, barrelette neurons as compared to E18.5 wild-type neurons (Extended Data Fig. 1q-s; Methods). We referred to these genes as barrelette sensory ARGs (bsARGs) (Supplementary Table 2). bsARGs comprised 4 IEGs, namely *Fos, Egr1, Junb,* and *Zfp36* (Fig. 1b and Supplementary Table 2), as well as at least 23 putative LRGs involved in calcium signaling and late aspects of activity-dependent neuronal development (e.g. *Cd38, Osmr*) (Supplementary Table 3), thus validating our strategy.

We then identified additional activity regulated genes (ARGs) (*n*=83) that were transcriptionally induced by distinct activity-dependent stimuli in neuronal types other than barrelette neurons[9-11] but that displayed undetectable or low basal expression level (RPKM < 3) in E14.5, E18.5 and P4 barrelette neurons. We referred to these genes as non-barrelette ARGs (nbARGs, see Methods). nbARGs included both LRGs and 12 IEGs that were induced, respectively, in: i) KCl-treated mouse cultured cortical neurons[10], ii) primary somatosensory (S1) barrel cortex neurons after environmental enrichment[11], and iii) light-stimulated primary visual cortex (V1) neurons[9].

The H3K27ac and H3K27me3 histone marks coexist at the promoter and gene body of bipartite genes

To establish whether the H3K27ac and H3K27me3 marks indeed coexist at the promoter and gene body of bipartite genes, respectively, we next carried out sequential ChIP-seq. Since collecting sufficient numbers of E14.5 barrelette neurons for sequential ChIP-seq and other specific chromatin experiments requiring higher cell numbers was technically challenging, we have used developmentally matched bulk hindbrain tissue (Methods); indeed, in E14.5 hindbrain tissue,

barrelette neuron E14.5Bip genes were also maintained in a bipartite state (Extended Data Figs. 3e and 4a-c).

Since the two marks are expected to be non-overlapping though covering adjacent promoter and gene body regions, we applied sequential H3K27me3 and H3K27ac ChIP-seq on 2-3kb large chromatin fragments firstly immunoprecipitated by anti-H3K27me3 and subsequently by anti-H3K27ac antibodies (Extended Data Fig. 4a; Methods). E14.5Bip genes were selectively enriched among the H3K27ac+/H3K27me3+ co-immunoprecipitated fragments (Fig. 2e, Extended Data Fig. 4b-d), demonstrating that the bipartite signature exists *in vivo* and it is not the result of allelic or cellular heterogeneity in bulk ChIP-seq data.

To further investigate the issue of cell heterogeneity and support these findings, we next carried out single-cell mRNA-seq (scRNA-seq) analysis of FACS isolated E14.5 postmitotic *Drg11*$^{vPrV\text{-}tdTomato/+}$ barrelette neurons (Supplementary Table 1, Extended Data Fig. 1d, Supplementary Fig. 2, Methods) by droplet-based encapsulation using the 10X Genomics Chromium System. We reasoned that if the H3K27ac/H3K27me3 chromatin pattern of bipartite genes is merely the result of H3K27ac or H3K27me3 heterogeneity within the bulk cell population, then, in single cells, bipartite genes (E14.5Bip) carrying only H3K27ac should express comparable levels of mRNA expression as non-bipartite (non-Bip) genes with matching H3K27ac promoter levels (E14.5AcP; Fig. 2f, left and middle, Methods). However, while the E14.5AcP genes showed detectable mRNA expression in at least 21% (mean fraction) of single cells, the E14.5Bip gene transcripts were only detected in as few as 5% of single cells (Fig. 2f, right).

In addition, scRNA-seq analysis of E10.5 *K20*$^{tdTomato/+}$ mitotic progenitors confirmed the same finding: namely, transcripts of E10.5 top 100 bipartiteness scored (E10.5Bip) genes (Supplementary Table 4) were only detected in as few as 6% of single cells, while non-Bip genes with Bip-matching H3K27ac promoter level (E10.5AcP) showed detectable expression in 30% of single cells (Extended Data Fig. 4e).

In summary, the scRNA-seq analysis together with the bulk sequential ChIP-seq data strongly support that H3K27ac/H3K27me3 coexist at the promoter/gene body of bipartite genes, correlating with low or undetectable mRNA transcription.

Visualization of mouse genes by t-SNE according to their chromatin organization in barrelette neurons

To investigate how the bipartite signature is established, maintained, and resolved during development, we created a two-dimensional (2D) projection of autosomal genes according to chromatin accessibility, H3K27me3, H3K4me2, and H3K27ac levels at promoters and gene bodies (Extended Data Fig. 5a, Methods) using t-distributed Stochastic Neighbor Embedding (t-SNE) (Fig. 3a-d, Extended Data Fig. 5b-l). We generated a single map for E10.5 progenitors (Fig. 3b) and a combined E14.5, E18.5 and P4 t-SNE map of chromatin states for postmitotic barrelette neurons (Fig. 3a, Extended Data Fig. 5b, c; Methods). Genes (i.e. dots on t-SNE plots) with similar chromatin patterns were grouped together, as illustrated by the regions of homogeneous chromatin features at promoters and gene bodies (Fig. 3a, b, Extended Data Fig. 5b, c). Gene grouping also correlated with mRNA-seq data (compare Extended Data Fig. 5c-e).

Top-scoring bipartite and bivalent genes at E10.5 and postmitotic stages mapped to distinct, largely non-overlapping, regions on the respective t-SNE maps (visualized by green and red contour lines, respectively, depicting gene densities; Fig. 3a-d; Extended Data Fig. 5f, g, k, l; Methods). At postmitotic stages, while bivalent genes were grouped, bipartite genes appeared more spatially distributed on the t-SNE, indicating varying levels of H3K27ac and H3K27me3 at the promoter and gene body. To assess this point, we focused on the E14.5Bip genes (black dots, Fig. 3a) and subdivided them into two subgroups, E14.5Bip-a (n=57) and E14.5Bip-b (n=43) genes, according to their localization on the t-SNE plot (Extended Data Fig. 5i, orange and blue dots). E14.5Bip-a genes (orange dots) carried relatively higher H3K27ac on promoter and lower H3K27me3 on gene body (Extended Data Fig. 5j, left and middle); conversely, E14.5Bip-b genes (blue dots) carried relatively lower H3K27ac on promoter and higher H3K27me3 on gene body (Extended Data Fig. 5j, left and middle). To understand whether these differences might be biologically meaningful and correlate them with distinct transcript production, we analyzed mRNA levels. Indeed, E14.5Bip-b genes (blue cluster), located on t-SNE close to the bivalent region (Extended Data Fig. 5i), produced less transcripts than Bip-a genes (Extended Data Fig. 5j, right), located farther away from E14.5Bip-b genes on the t-SNE map and closer to active genes (Extended Data Fig. 5i, orange cluster).

Furthermore, the combined E14.5-E18.5-P4 t-SNE plot permitted to visualize chromatin state changes of each bipartite gene during neuron development (Fig. 3a-d). Namely, genes mapping to the same region of the combined t-SNE map at different developmental stages would reveal a stable chromatin state, whereas genes changing their localization between developmental

stages change their chromatin state (see Fig. 3c,d). Note that the three stages are homogenously mixed on the combined t-SNE (Extended Data Fig. 5b), indicating that gene clustering was conducted without potential stage-specific bias. To illustrate the dynamic nature of bipartite genes, we plotted on a t-SNE map all genes colored according to their overall change of chromatin state between E14.5 and P4 (Extended Data Fig. 5h, Methods). This analysis confirmed that bivalent genes tend to be more stable compared with bipartite genes and revealed that bipartite genes are the most dynamic of all genes, in terms of chromatin state change, during development.

The bipartite signature at the *Fos* locus allows for active enhancer-promoter contacts irrespective of productive transcription

Developmental analysis of the *Fos* locus revealed additional features of the bipartite chromatin organization (Fig. 3e). The *Fos* locus contains five activity-regulated enhancers (e1-e5) combinatorially activated in a stimulus-dependent manner in cultured and *in vivo* neurons[12-14]. In E10.5 progenitors and E14.5 early postmitotic neurons, even though *Fos* mRNA was barely detected, its promoter, e2, e5 and also e1 displayed active histone marks. At E18.5, phosphorylated CREB (pCREB), a readout of stimulus-dependent transcription[15], increased at the promoter and *Fos* was transcriptionally induced; the gene body was almost devoid of H3K27me3 and partially switched to H3K27ac. At P4, productive transcription and pCREB levels at the *Fos* promoter and e2 and e5 enhancers were further increased (Fig. 3e).

Circularized Chromosome Conformation Capture sequencing (4C-seq) of E10.5, E14.5, E18.5, and P4 bulk hindbrain tissue (Methods) using the *Fos* promoter, e2, and e5 as viewpoints revealed reciprocal physical interactions during development (Extended Data Fig. 6c), showing that the active promoter interacts with active enhancers, irrespective of net productive transcriptional output. On the other hand, enhancer acetylation levels change during development (Fig. 3e) as well as the frequencies of specific enhancer-promoter interactions in response to different stimuli[12].

Additional chromatin features of bipartite genes

E14.5Bip promoters were more accessible than E14.5Biv promoters, and displayed similar accessibility as E14.5AcP genes (Fig. 4a, ATAC-seq and H3K27ac). Moreover, non-Bip genes with low, Bip-matching, levels of productive mRNA transcription (Fig. 4a, E14.5mRNALow;

Methods) displayed much lower promoter accessibility and H3K27ac levels than E14.5Bip promoters (Fig. 4a). Unlike E14.5Bip and E14.5Biv, E14.5mRNALow and E14.5AcP genes were not marked by H3K27me3 on gene body (Fig. 4a). Moreover, E14.5Bip promoters carried higher levels of the histone variant H3.3, involved in histone turn-over at transcriptionally active genes in postmitotic neurons[16], than E14.5Biv promoters (Fig. 4a). Furthermore, promoter Cdk9 levels of E14.5Bip genes were higher than E14.5Biv genes, and comparable with E14.5AcP genes (Fig. 4a). Cdk9 is a core component of pTEFb that catalyzes phosphorylation of RNAPII-S2 to regulate the transition from paused RNAPII (S5P form) to elongation (S2P form)[17] and its recruitment to enhancers and promoters requires H3K27 acetylation[18,19].

Furthermore, we then correlated the dynamics of elongation marks (RNAPII-S7P, RNAPII-S2P, H3K36me3) and other chromatin features (ATAC-seq, H3K4me2, H3K27ac, H3K27me3) with the transcriptional state of E14.5Bip genes through E18.5 and P4 development. Genes that become activated (E14.5Bip->P4Exp) displayed specifically increased levels of RNAPII-S7P, -S2P and H3K36me3 marks at P4 compared to E14.5Bip->P4Bip and E14.5Bip->P4Biv genes (Extended Data Fig. 6a, e). These results further support that the bipartite chromatin signature is consistently inversely correlated with the productivity of mRNA elongation of bipartite genes.

IEG bipartite chromatin is necessary to prevent precocious activity-dependent neuronal maturation during development

We asked whether preventing the establishment of the Pc bipartite signature on inducible IEGs *in vivo* during development may have a profound impact on the cell maturation program. Activity-regulated AP-1 family TFs (e.g. Fos, Jun) mediate the maturation process of early postnatal neurons through the *de novo* activation of AP-1-specific enhancers in a neuronal subtype-specific manner[20]. We hypothesized that the precocious activation of bipartite IEG-TFs, including Fos, in immature neurons by the removal of the gene body H3K27me3 mark may lead to precocious opening and activation of the early postnatal Fos-specific enhancer maturation program. Among the enhancers that normally become open only in postnatal barrelette neurons (3967 enhancers), we identified 85 neuronal activity-regulated Fos-binding enhancers[14] (see Methods). Interestingly, these 85 Fos-binding enhancers gained precocious accessibility in E14.5 *Ezh2cKO*$^{HB\text{-}RFP}$ homozygous mutant neurons (Fig. 5d, Extended Data Fig. 8b). In a complementary approach, we

next assessed that these 85 Fos-binding enhancers also gained accessibility in response to neuronal stimulation of E12.5 short-term cultured hindbrain trigeminal sensory neurons (Extended Data Fig. 8a; Methods), thus likely as a direct result of Fos induction. Together with the untimely expression of bipartite IEGs in *Ezh2cKO* embryonic sensory neurons at early stages (Fig. 5b), these results strongly suggests that functional inactivation of *Ezh2* could have an impact on prenatal neuronal development and maturation *in vivo*, at least in part, through ectopic up-regulation of activity-dependent bipartite IEGs and precocious activation of the early postnatal Fos-driven enhancer program.

Additional analysis of bipartite gene transcriptional regulation in PRC2 KO ESCs

To overcome the unfeasibility of obtaining large amounts of cells from *Ezh2cKO* embryos, we utilized *EedKO* mouse ESCs in which the H3K27me3 mark is removed genome-wide[21]. We carried out RNAPII-S2P ChIP-seq and mRNA-seq in wild-type and *EedKO* ESCs. RNAPII-S2P ChIP-seq signals have a narrower dynamic range (bottom 30% and top 30% expressed genes, Fig. 4a, b) compared to other RNAPII phosphorylated forms, likely because elongating RNAPII-S2P is a small fraction of total RNAPII. To improve the robustness of our analysis and to extend the general value of its conclusions, we considered all genes (n=3457) carrying H3K27me3 on their gene bodies rather than focusing only on 100-200 bipartite genes. For all these Pc targets, up-regulation of mRNA levels correlated with modest but significant increase of RNAPII-S2P signals in the TES region, in *EedKO* compared with wild-type ESCs (Extended Data Fig. 7f, g).

Next, we addressed whether it is H3K27 methylation and/or recruitment of Pc to the gene body that is essential for inhibition of bipartite mRNA transcription. We mined published datasets and analyzed the transcriptional changes of bipartite genes in full *Ezh1KO;Ezh2KO* and Ezh2 catalytically inactive *Ezh1KO;Ezh2*^{Y726D} mutant ESCs[22]. In *Ezh1KO;Ezh2KO* double mutant ESCs, PRC2-targeted genes lack both PRC2 binding and the H3K27me3 mark. In contrast, in *Ezh1KO;Ezh2*^{Y726D} ESCs, while the H3K27me3 mark was depleted PRC2 recruitment was rescued[22]. We found that bipartite genes are up-regulated in both full *Ezh1KO;Ezh2KO* and Ezh2 catalytically inactive *Ezh1KO;Ezh2*^{Y726D} ESCs (Extended Data Fig. 7h), indicating that the H3K27me3 mark itself on the gene body is fundamental for bipartite gene transcriptional regulation, rather than recruitment of PRC2 proteins.

Taken together, these results (and those presented in the main manuscript) strongly associate the Ezh2-dependent H3K27me3 marking of the gene bodies of bipartite genes to the inhibition of productive mRNA elongation. Furthermore, these data suggest a general involvement of the H3K27me3 mark on gene bodies in the regulation of productive mRNA elongation, beyond just bipartite genes.

Polycomb marking of bipartite genes bodies hampers productive elongation through inhibition of stimulus-dependent NELF release and chromatin compaction

The negative elongation factor (NELF) negatively regulates transcriptional elongation by pausing RNAPII at TSSs[23]. Stimulus-dependent NELF removal from IEG promoters causes release of paused RNAPII into elongation[24]. We hypothesized that H3K27me3 on gene body may inhibit transcriptional elongation in bipartite genes by interfering with stimulus-dependent NELF release. Indeed, the levels of NELF-b, a core component of the NELF complex, in bipartite gene promoters were decreased in *EedKO* compared to wild-type ESCs (Fig. 6a, left). Importantly, this was accompanied by the release of paused RNAPII-S5P from promoter regions (Fig. 6a, right), phenocopying the consequences of NELF knock-down experiments[25,26]. Together with the finding that the Pc marking of gene bodies of bipartite IEGs limits the extent of rapid stimulus-induced transcription (Fig. 5f,g), these results strongly indicate that H3K27me3 in bipartite gene bodies interferes with RNAPII release and elongation in part by inhibiting stimulus-dependent NELF removal.

There is evidence that PRC1 interferes with transcriptional elongation through chromatin compaction[27,28]. One unique feature of the bipartite signature is a high-level deposition of Ring1b, a core component of PRC1, in bipartite gene bodies (Fig. 4a). To assess the involvement of H3K27me3 in Ring1b bipartite gene body deposition and in gene body compaction (i.e. accessibility) we analyzed ChIP-seq of Ring1b and ATAC-seq in wild-type and *Ezh1KO;Ezh2KO* mouse ESCs from published datasets[22]. We found that *Ezh1/Ezh2* removal caused a reduction of gene body Ring1b levels in bipartite genes (Fig. 6b), indicating that the gene body Ring1b deposition on bipartite genes is H3K27me3-dependent. This correlated with significant increase of bipartite gene body, though not promoter, accessibility in *Ezh1KO;Ezh2KO* mouse ESCs (Fig. 6c).

We next asked if the de-compaction (increased accessibility) of bipartite genes in PRC2 KO was primarily caused by the removal of H3K27me3 per se or whether it was merely a consequence of increased transcription. We carried out experiments to quantify chromatin compaction of bipartite IEGs (i.e. *Fos*, *Egr1*) using wild-type and *EedKO* ESCs in the serum-starved condition (Fig. 6d, e). While these IEGs showed a modest increase of expression in *EedKO* ESCs, as compared to wild-type, in the serum-containing medium (Fig. 6d, mRNA), in the serum-starved condition they were basically not expressed (Fig. 6e, mRNA; also see Fig. 5f), likely due to lack of inducing stimulus. We found that even in the serum-starved condition bipartite IEG gene bodies showed increased accessibilities in *EedKO* ESCs (Fig. 6e, ATAC), indicating that the de-compaction of bipartite gene bodies was not only merely correlative with increased transcription, but was also at least partially caused by PRC2 deletion and the removal of H3K27me3.

Stimulus-dependent H3K27me3 removal from IEG gene bodies requires active de-methylation

GSK-J4, an inhibitor of H3K27me3 demethylases (i.e. UTX (Kdm6a), Jmjd3 (Kdm6b)) prevented neuronal activity-dependent gene body H3K27me3 removal (Fig. 7d), indicating that H3K27me3 is removed through active demethylation. ChIP-seq of H3K27me3 in E18.5 wild-type and *Jmjd3KO* (Supplementary Table 1, Methods) hindbrain confirmed that inactivation of *Jmjd3* inhibited, at least partially, removal of the gene body H3K27me3 mark from the E14.5 bipartite genes that become active at peri/postnatal (P4) stages (Fig. 7e). These results strongly indicate that the stimulus-dependent removal of H3K27me3 from IEG gene bodies requires active de-methylation.

Effect of A-485 treatment on transcriptional induction of bipartite IEGs

We assessed the effect of A-485 treatment on neuronal activity-dependent transcriptional induction of bipartite IEGs. A-485 treatment prevents the rapid induction of bipartite IEGs after short-time (i.e. 8 minutes) exposure to KCl (Fig. 7h). Taken together with the effects of the H3K27me3 demethylase inhibitor (i.e. Gsk-J4) treatment on the rapid induction of bipartite IEGs (Fig. 7f), this result strongly indicates that fast bipartite IEG transcriptional induction requires *de novo* H3K27acetylation and rapid removal of the gene body H3K27me3 mark through active de-methylation (Fig. 7i, scheme).

On the other hand, after prolonged exposure (i.e. 60 minutes) to the KCl stimulus even A-485 treated neurons showed transcriptional up-regulation of bipartite IEGs. However, as is the case with the H3K27me3 demethylase inhibitor (Fig. 7f), mRNA levels remained significantly lower as compared to control neurons (Fig. 7g). This indicates that, even in the event of lack of H3K27ac increase at the bipartite IEG promoters, the existing H3K27ac levels are sufficient to allow the mRNA of bipartite IEGs to accumulate over time upon prolonged stimulation, albeit their transcripts do not reach optimal levels under such a condition (Fig. 7i, scheme).

KCl-dependent gene body H3K27me3 removal is driven by de novo promoter acetylation per se irrespective of transcriptional elongation

We asked whether the KCl-dependent gene body H3K27me3 removal is the consequence of transcriptional elongation or rather it is driven by promoter acetylation per se. We treated E12.5 short-term cultured neurons with KCl in the presence of flavopiridol, a Cdk9 inhibitor. Flavopiridol treatment caused a complete block of the KCl-dependent transcription of IEGs, while *de novo* promoter H3K27ac was not prevented (Fig. 7j). Interestingly, we found that flavopiridol did not prevent the KCl-dependent removal of H3K27me3 from bipartite IEG gene bodies (Fig. 7j), indicating that the gene body H3K27me3 mark is removed by KCl-induced neuronal stimulation regardless of mRNA transcriptional elongation, provided that *de novo* promoter H3K27 acetylation occurs.

Supplementary Discussion

We first investigated how the bipartite signature is established, maintained, and resolved during development. Bipartite genes are established in a cell type- and developmental stage-specific manner and constitute a small but consistent fraction (5-15%) of the bivalent genes (Figs. 2a, d and 3a-d), suggesting that the transition of a subset of bivalent genes into a bipartite state may depend on the distinct environmental conditions to which developing cells in different tissues are exposed. Indeed, in addition to typical stimulus response IEGs, bipartite genes encode for cell-type specific transcriptional regulators, receptors, and molecules responding to distinct signaling pathways (Fig. 2c).

We show that, unlike the more stable bivalent chromatin, the bipartite state allows for dynamic and reversible (i.e. from bivalent to bipartite to bivalent, or to active, state) chromatin changes through developmental stages and that this correlates with transcriptional output changes (Fig. 3, Extended Data Fig. 6a). The initial transition from bivalent to bipartite chromatin in different cell types and/or developmental stages might be regulated by specific sets of transcription factors. For instance, we show that in barrelette neurons at E14.5, the E14.5Bip gene promoters are enriched for NF-kB-related and forkhead FOX-related factor binding motifs (Extended Data Fig. 3g), suggesting that they might be involved in the transition leading to partial resolution of bivalency into a bipartite state. Once established, the bipartite signature is maintained through a reciprocal competitive balance between H3K27ac at promoters and H3K27me3 in gene bodies (Figs. 3d, 5a and 7a-k). In this respect, it is noteworthy that HDAC levels shuttling between the nucleus and cytoplasm can be signal-regulated[29].

As for the transition from the bipartite to active state, we found that pCREB binding levels increased in promoter (and enhancer, e.g. at *Fos* locus, Fig. 3e) regions of stimulus response genes that were bipartite at E14.5 and became active at E18.5/P4 (Fig. 3e, 7a), including neuronal activity-induced IEGs (Fig. 3e, Extended Data Fig. 6b,d). Increase of stimulus-dependent phosphorylation of CREB and subsequent increase of acetylation of H3K27 causes active H3K27me3 demethylation from bipartite gene bodies and elongation barrier release, correlating with bipartite chromatin resolution into an active state and productive transcription (Fig. 7a-k, summary diagram Fig. 8). Conversely, decrease of pCREB levels on E14.5Bip genes correlated with reversion into bivalency at P4 (Fig. 7a). Thus, while NF-kB and/or FOX factors might initially

increase the accessibility at E14.5Bip promoters contributing to partial resolution of bivalency and transition into the bipartite state, stimulus-dependent CREB phosphorylation binding levels might serve as an environmental sensor and a rheostat to bi-directionally regulate the balance between H3K27ac and H3K27me3 at bipartite target gene loci.

The bipartite state can be resolved into further repression or activation, a feature also shared with the bivalent state. However, unlike the bivalent state, we demonstrate that the bipartite chromatin still enables very rapid transcriptional inducibility of stimulus response genes, while integrating both levels and duration of the signal (Fig. 5f,g). How fast, when, or whether at all, a stimulus response gene might be transcriptionally induced by any given environmental signal that cells may experience during development has obviously an important impact on cell fate decisions and downstream transcriptional programs of maturation. We speculate that the bipartite chromatin provides a unique epigenetic mechanism regulating transcriptional sensitivity of target genes to different signals, ultimately resulting in context-dependent interpretation of signal relevance by the developing cell. In addition to molecular mechanisms regulating the competence of a cell to receiving the signal[30], we reveal a mechanism acting directly at stimulus response genes and providing competence to respond to the signal through their active promoter state, while maintaining epigenetic control on the timing and level of transcriptional response through the repressive Pc marking of the gene body (summary diagram, Fig. 5g), thus providing a way to evaluate signal relevance during development. During development, this chromatin structure at immediate early stimulus response genes might be most effective in preventing inappropriate induction by acute exposure to weak or non-physiologically relevant signals. In fact, based on our results (Fig. 7b-i), it is likely that short exposure to an inducing signal could achieve fast bipartite IEG transcriptional induction only if signal levels are sufficiently high to induce fast increase of promoter H3K27 acetylation and fast gene body H3K27me3 removal.

As for the molecular mechanism underlying the transcriptional regulation of bipartite genes, analysis of the distribution of RNAPII phosphorylated forms on bipartite genes revealed that at the stages when H3K27me3 levels on gene bodies are relatively high and bipartite genes are not or barely productively transcribed, RNAPII-S5P and RNAPII-S7P appear to be pausing at the promoter-proximal first exon regions, while both RNAPII-S7P and RNAPII-S2P signals are negligible in the gene body regions (Fig. 4a,b, Extended Data Fig. 6d). Indeed, some IEGs, including *Fos* and *Egr1*, are known to be regulated by RNAPII pausing in early elongation, and

by controlled, stimulus-dependent, release of paused RNAPII into productive RNA synthesis[31]. In particular, at the human *Fos* locus, RNAPII pauses within the first 300 nucleotides of the first exon[31]. This is compatible with the localization of RNAPII-S5P and -S7P peaks at the mouse *Fos* locus in developing E14.5 hindbrain tissue and ESCs (Extended Data Figs. 6d and 7g). On the other hand, the finding of RNAPII pausing on bipartite genes was intriguing; in fact, genes in a bipartite state are not or very lowly productively expressed, yet RNAPII pausing has been mainly observed on expressed genes[32]. Interestingly, this reminds of a particular class of 'paused and inactive' genes representing only < 2% of all genes (class III in Min et al., 2011[32]). This number is roughly compatible with the low number of bipartite genes we observe during development. It is therefore tempting to speculate that, during development, this class might be represented by the bipartite genes.

At the bipartite stage, RNAPII does not efficiently transit into productive elongation. At later stages, correlating with significant reduction or loss of H3K27me3 on gene bodies, we observed increasing accumulation of elongation marks (RNAPII-S7P, RNAPII-S2P, and H3K36me3) at the bipartite loci resolving into transcriptional activation. In developing neurons, this is driven by stimulus-dependent accumulation of pCREB and H3K27ac signals at promoters and enhancers (see above).

Functional analysis in developing hindbrain neurons and ESCs strongly supported a role of PRC2-dependent H3K27me3 marking of gene bodies in inhibiting RNAPII-dependent transcript elongation (Figs. 5 and 6, Extended Data Fig. 7, summary diagram Fig. 8). This is directly supported by the increase of elongation mark levels in PRC2 knockout mutants both in hindbrain neurons and ESCs (Extended Data Fig. 7e-g). Pc-dependent inhibition of productive mRNA elongation in bipartite IEGs may be achieved by multiple mechanisms including inhibition of stimulus-induced release of the NELF complex and chromatin compaction limiting RNAPII elongation. Indeed, we found that the H3K27me3 in bipartite gene bodies of bipartite IEGs limits the extent of rapid stimulus-induced transcription (Fig. 5f, g), as well as the release of both the NELF complex and of transcriptionally initiating RNAPII (Fig. 6a) at promoters. The Cdk9 binding profile (Fig. 4a), the increased gene body accessibility correlating with developmental removal of H3K27me3 (Extended Data Fig. 6a), the finding that the H3K27me3 mark in gene bodies of bipartite genes serves as a platform to recruit high levels of PRC1 (i.e. Ring1b, Figs. 4a and 6b), that elongation is not required for Eed deletion to increase chromatin accessibility and

that H3K27me3 (or Eed) is required for full compaction (Fig. 6c-e), further suggested that the poor mRNA elongation of bipartite genes may be additionally due to impairment of RNAPII to physically go through a Pc-compacted gene body chromatin.

Lastly, in prenatal postmitotic barrelette neurons, the bipartite state might critically prepare IEGs to rapidly respond to sensory (e.g. whisker) stimuli at the beginning of the critical period by activity-dependent increase of productive mRNA elongation. Active promoters and enhancers of bipartite genes, and high levels of pre-loaded RNAPII at bipartite gene promoters, might enable synchronous patterns of stimulus-dependent IEG transcription in response to whisker stimulus-specific correlated activity, which is critical to precisely refine the whisker-related barrelette map[8]. Moreover, we speculate that Pc-dependent epigenetic regulation at gene bodies of stimulus-inducible genes at prenatal stages might 'pre-label' gene bodies for further epigenetic marking in postnatal mature neurons replacing the developmental Pc marking pattern, such as e.g. gene body DNA methylation and recruitment of Mecp2[33], that is known to also cause chromatin compaction of targeted genomic regions[34].

Supplementary Methods

Animals

All animal experimental procedures were performed in accordance with Guide for Care and Use of Laboratory Animals, and were approved by the Veterinary Department of the Canton of Basel-Stadt.

Generation of the r2::EGFP mouse line

To generate the *r2::EGFP* (*r2*$^{EGFP/+}$) mouse line, an *EGFP* cassette (Clontech) was used to generate a pKS-β-globin-EGFP plasmid. The BamH1 2.5 kb rhombomere 2-specific enhancer of *Hoxa2*[35] was then cloned in reverse orientation 5′ of the β-globin promoter, thus generating a final construct consisting of the r2-specific enhancer, a β-globin minimal promoter, and *EGFP* encoding sequence. The construct was linearized, purified, and microinjected into the pronuclei of mouse zygotes. Founders were identified by PCR.

Mouse embryonic hindbrain neuron cultures

E12.5 CD1 wild-type or *Drg11*$^{tdTomato/+}$ mouse embryo hindbrains were dissected, and treated with 0.05% Trypsin/EDTA (Thermo Fisher Scientific, 25300054) for 3 minutes at 37°C, rinsed by ice-cold DMEM 1X, and dissociated by pipetting. Cells were seeded on dishes (Day0). Dishes were pre-coated over-night by collagen I (Thermo Fisher Scientific, A1048301) in PBS 1X. Neurons were maintained in Neurobasal Medium (Thermo Fisher Scientific, 21103049) containing 2% B27 Supplement (Thermo Fisher Scientific, A3582801), 2% GlutaMAX Supplement (Thermo Fisher Scientific, 35050061) and penicillin-streptomycin. For a long culture (upto 1 week), medium was changed every two day.

Mouse embryonic stem cells (ESCs) cultures

Mouse embryonic stem cells (ESCs) were cultured in the culture medium containing DMEM (Thermo Fisher Scientific, 41965), 55μM 2-Mercaptoethanol (Thermo Fisher Scientific, 21985023), 2% GlutaMAX Supplement (Thermo Fisher Scientific, 35050061), 1X MEM Non-Essential Amino Acids solution (Thermo Fisher Scientific, 11140050), 1000 unit/ml ESGRO

Recombinant Mouse LIF Protein (Merck, ESG1106), penicillin-streptomycin and 15% FCS. Dishes were pre-coated by 0.1% gelatin (Sigma, G2500).

Sample preparation, chromatin immunoprecipitation (ChIP) and sequencing (ChIP-seq)

ChIPmentation of the FACS-sorted cells

For ChIP-seq experiments using FACS-sorted cells, dissociated tissue was cross-linked with 1% formaldehyde in DMEM 1X/ FCS 10% for 10 minutes at room temperature (RT) and quenched with 125mM glycine for 5 minutes at RT. Cells were pelleted by centrifugation (1500 rcf, 5 minutes, 4°C) and rinsed twice with PBS 1X/ FCS 4%. Cells were filtered and collected by FACS. For each experiment, two to three independent biological replicates were used. To achieve the sequencing of chromatin immunoprecipitated from small amount of cells, preparation of ChIP-seq library was done by ChIPmentation protocol [36]. 50,000 cells (H3K4me2, H3.3) and 100,000 – 200,000 cells (H3K27ac, H3K27me3) were used. Cells were lysed in 20 – 40µl of Sonication Buffer (10mM Tris HCl pH8, 5mM EDTA, 0.5% SDS, 0.1X PBS, 1X Protease Inhibitor Cocktail (PIC – cOmplete – EDTA free, Roche, 04693132001)) for 15 min on ice, and sonicated using the Covaris machine to obtain DNA fragment the size of which distributes between 150bp and 500bp. The supernatant was transferred to a new tube, diluted five times with Equilibration Buffer (10mM Tris HCl pH8, 1mM EDTA, 140mM NaCl, 1% Triton X-100, 0.1% Sodium deoxycholate, 1X Protease Inhibitor Cocktail). Chromatin solutions were incubated over-night at 4°C with 1µg of anti-H3K4me2 (Millipore, 07-030, PRID: AB_10099880), 1µg of anti-H3K27ac (abcam, ab4729, PRID: AB_2118291), 1µg of anti-H3K27me3 (Millipore, 07-449, PRID: AB_310624) or 1µg anti-H3.3 (Millipore, 09-838, PRID: AB_10845793) antibodies. The next day, 20µl of protein G coupled to magnetic beads (Dynabeads Protein G, Thermo Fisher, 10004D) were added and the incubation was continued for 2 hours at 4°C. The beads were then washed five times with RIPA Buffer (10mM Tris HCl pH8, 1mM EDTA, 140mM NaCl, 1% Triton X-100, 0.1% SDS, 0.1% Sodium deoxycholate, 1X Protease Inhibitor Cocktail), twice with High-Salt RIPA Buffer (10mM Tris HCl pH8, 1mM EDTA, 500mM NaCl, 1% Triton X-100, 0.1% SDS, 0.1% Sodium deoxycholate, 1X Protease Inhibitor Cocktail), twice with LiCl Buffer (10mM Tris HCl pH8, 1mM EDTA, 250mM LiCl, 0.5% NP40, 0.5% Sodium deoxycholate, 1X Protease Inhibitor Cocktail), and twice with 10mM Tris HCl pH8. Beads were resuspended in 30µl Tagmentation Buffer (10mM Tris HCl pH8, 5mM $MgCl_2$) containing 1µl Tagment DNA Enzyme from the Nextera

DNA Sample Prep Kit (Illumina, FC-121-1030) and incubated at 37°C for 10min. The beads were washed twice with RIPA Buffer and twice with TE Buffer (10mM Tris HCl pH8, 1mM EDTA). DNA was eluted from the beads with 60µl Elution Buffer (10mM Tris HCl pH8, 5mM EDTA, 300mM NaCl, 0.5% SDS, proteinase K) at 65°C for 5hours. DNA was purified with SPRI AMPure XP beads (Beckman, sample to beads ratio 1:2) and eluted in 25µl 10mM Tris HCl pH8. 2µl of each library was amplified in 10µl qPCR reaction (1X Sybr Green (Thermo Fisher), 0.8µM primers, 1X KAPA HiFi Hot Start Ready Mix (Kapa Biosystems): StepOnePlus Real-Time PCR Systems (Thermo Fisher), 72°C for 5min; 98°C for 1min; 25 cycles of 98°C for 15sec, 63°C for 30sec, 72°C for 1min) to estimate the optimum number of enrichment cycles. Final enrichment of the libraries was performed in 50µl reaction (1X KAPA HiFi Hot Start Ready Mix and 0.8µM primers). Enriched libraries were purified with size selection using SPRI AMPure XP beads (sample to beads ratio 1:0.6) to remove long fragments, recovering the remaining DNA (sample to beads ratio 1:2). Sequencing was performed on an Illumina HiSeq 2500 machine (50bp read length, single-end).

ChIPmentation of bulk tissue

For ChIP-seq experiments using wild-type brainstem tissue and cultured embryonic hindbrain neurons dissociated tissue was cross-linked with 1% formaldehyde in PBS 1X for 10 minutes at room temperature (RT) and quenched with 125mM glycine for 5 minutes at RT. Cells were pelleted by centrifugation (1500 rcf, 5 minutes, 4°C) and rinsed twice with PBS 1X/ FCS 4%, and cells were filtered. Normally we prepare at least two biological replicate, but for some experiments of bulk ChIP-seq, independent biological replicates were not prepared, but analysis were supported by the use of internal controls and the sample was normalized and compared with samples from related condition (see below). Preparation of ChIP-seq library was done by ChIPmentation protocol, except for ChIP-seq of the H3K27ac marks of control and TSA-treated cultured E12.5 hindbrain neurons that were prepared according to manufacturer's protocol of NEBNext Ultra DNA Library Prep Kit for Illumina (NEB, E7370) 1,000,000 cells (H3K36me3), 5,000,000 cells (phospho-CREB (pCREB), non-phosphorylated RNAPII 8WG16, Ring1b) or 30,000,000 cells (Ser5P RNAPII, Ser7P RNAPII, Ser2P RNAPII, Cdk9) were used. Cells were lysed in 200 - 800 µl of Sonication Buffer (10mM Tris HCl pH8, 5mM EDTA, 0.15% SDS, 0.1X PBS, 1X Protease Inhibitor Cocktail) for 15 min on ice, and sonicated using the Covaris machine or the Bioruptor Pico machine to obtain DNA fragment the size of which distributes between 150bp and 500bp.

After centrifugation (1 min, 10,000rpm, 4°C), the supernatant was transferred to new tubes, diluted five times with Equilibration Buffer. Chromatin solutions were incubated over-night at 4°C with 1µg of anti-H3K36me3 (abcam, ab9050, PRID: AB_306966), 10µg of anti-pCREB (Millipore, 17-10131, PRID: AB_10807817), 10µg of anti-Ring1b (Cell Signaling, 5694, PRID: AB_10705604), 10µg of anti-non-phosphorylated RNAPII 8WG16 (Covance, MMS-126R, PRID: AB_10013665), 30µg of anti-Ser5P RNAPII 4H8 (Covance, MMS-128P, PRID: AB_10013820), 30µg of anti-Ser7P RNAPII 4E12 (Millipore, 04-1570, PRID: AB_10618152), 30µg of anti-Ser2P RNAPII 3E10 (Active Motif, 61083, PRID: AB_2687450) and 30µg of anti-Cdk9 (abcam, ab239364) antibodies. The next day, 40 - 200 µl of protein G coupled to magnetic beads were added and the incubation was continued for 2 hours at 4°C (as for anti-Ser7P RNAPII 4E12 and Ser2P RNAPII 3E10 rat-derived IgG_1, 200µl beads were pre-incubated with 60µl rabbit-derived anti-rat IgG antibody (abcam, ab6703, PRID: AB_956015) over-night to bridge primary antibodies and protein G). The beads were then washed twice with RIPA Buffer, once with High-Salt RIPA Buffer, once with LiCl Buffer, and twice with 10mM Tris HCl pH8. Beads were resuspended in 30-90µl Tagmentation Buffer containing 1-3µl Tagment DNA Enzyme and incubated at 37°C for 10min. The beads were washed twice with RIPA Buffer and twice with TE Buffer. DNA was eluted from the beads with 60µl Elution Buffer at 65°C for 5hours. DNA was purified with SPRI AMPure XP beads (sample to beads ratio 1:2) and eluted in 25µl 10mM Tris HCl pH8. 2µl of each library was amplified in 10µl qPCR reaction to estimate the optimum number of enrichment cycles. Final enrichment of the libraries was performed in 50µl reaction. Enriched libraries were purified with size selection using SPRI AMPure XP beads (sample to beads ratio 1:0.6) to remove long fragments, recovering the remaining DNA (sample to beads ratio 1:2). Sequencing was performed on an Illumina HiSeq 2500 machine (50bp read length, single-end).

ChIPmentation of tissues pre-fixed prior to dissociation

To avoid any effects of dissociation on chromatin state, we performed ChIP-seq against H3K27ac and H3K27me3 using E14.5 PrV tissue that was pre-fixed prior to dissociation. For this, micro-dissected PrV region was directly cross-linked with 1% formaldehyde in DMEM 1X/ FCS 10% for 12 minutes at room temperature (RT) and quenched with 125mM glycine for 5 minutes at RT. The tissue was rinsed twice with DMEM 1X/ FCS 4% and once with PBS, then treated with papain digestion mix for 15 minutes at 37°C and immediately put on ice. Tissue was rinsed by ice-cold DMEM 1X, and dissociated by pipetting and filtered. 500,000 cells were used per ChIP, and

reparation of library was done by ChIPmentation protocol as described above. We prepared only one biological replicate for each mark, but they were enough to confirm existence of the bipartite signature in IEGs.

Sequential (Double) ChIP-seq (H3K27me3/H3K27ac) of large chromatin fragments (Extended Data Fig. 4a)

For single ChIP-seq against H3K27ac or H3K27me3, or for sequential H3K27me3/H3K27ac ChIP-seq experiments with large (2-3kb) chromatin fragments, E14.5 wild-type brainstem tissue was dissociated and cross-linked with 1% formaldehyde in PBS 1X for 10 minutes at room temperature (RT) and quenched with 125mM glycine for 5 minutes at RT. Cells were pelleted by centrifugation (1500 rcf, 5 minutes, 4°C) and rinsed twice with PBS 1X/ FCS 4%, and cells were filtered. Six million cells were used. Cells were lysed in 200 µl of Sonication Buffer (10mM Tris HCl pH8, 5mM EDTA, 0.5% SDS, 0.1X PBS, 1X Protease Inhibitor Cocktail (PIC – cOmplete – EDTA free, Roche)) for 15 min on ice, and sonicated using the Bioruptor Pico machine to obtain DNA fragment the size of which distributes around 2kb. After centrifugation (1 min, 10,000rpm, 4°C), the supernatant was transferred to new tubes, diluted ten times with Equilibration Buffer. Samples were pre-cleaned by protein G coupled to magnetic beads at 4°C for one hour. As for single ChIP-seq, 10% of chromatin was incubated over-night at 4°C with 1µg of anti-H3K27ac or anti-H3K27me3 antibodies. The next day, 20µl of protein G coupled to magnetic beads were added and the incubation was continued for 2 hours at 4°C. The beads were then washed twice times with RIPA Buffer, once with High-Salt RIPA Buffer, once with LiCl Buffer, and once with TE Buffer. DNA was eluted from the beads with 100µl Elution Buffer at 65°C for over-night. DNA was purified with MinElute PCR Purification Kit (QIAGEN, 28004). As for sequential ChIP-seq, remaining chromatin was incubated over-night at 4°C with 5µg of anti-H3K27me3 antibody. The next day, 100µl of protein G coupled to magnetic beads were added and the incubation was continued for 2 hours at 4°C. The beads were then washed twice with RIPA Buffer, once with High-Salt RIPA Buffer, once with LiCl Buffer, and once with TE Buffer. To release chromatin from beads, DTT, high salt and detergent were used. Firstly beads were incubated with 50µl DTT 100mM at room temperature for 10 minutes, then 50µl 2X Chromatin Release Buffer (500mM NaCl, 2% SDS, 2% sodium deoxychorate, 2X Protease Inhibitor Cocktail) were added and mixed thoroughly and incubate at 37°C for 50 minutes. Magnetic beads were removed and samples were diluted four times by Equilibration Buffer, and concentrated using a 50Kda cutoff Amicon Ultra-

0.5 mL (Merck, FC505024). The collected samples were then further diluted ten times by Equilibration Buffer, and pre-cleaned by protein G coupled to magnetic beads at 4°C for one hour twice. Then samples were incubated over-night at 4°C with 1µg of anti-H3K27ac antibody. The next day, 20µl of protein G coupled to magnetic beads were added and the incubation was continued for 2 hours at 4°C. The beads were then washed twice with RIPA Buffer, once with High-Salt RIPA Buffer, once with LiCl Buffer, and once with TE Buffer. DNA was eluted from the beads with 100µl Elution Buffer at 65°C for over-night. DNA was purified with MinElute PCR Purification Kit. For both the single and sequential ChIP-seq, purified large fragments (2kb) of DNA were fragmented by 1µl Tagment DNA Enzyme from the Nextera DNA Sample Prep Kit (Illumina) at 55°C for 5 minutes. After tagmentation, DNA was purified with MinElute PCR Purification Kit by 25µl 10mM Tris HCl pH8. 2µl of each library was amplified in 10µl qPCR reaction to estimate the optimum number of enrichment cycles. Final enrichment of the libraries was performed in 50µl reaction. Enriched libraries were purified with size selection using SPRI AMPure XP beads (sample to beads ratio 1:0.6) to remove long fragments, recovering the remaining DNA (sample to beads ratio 1:2). Sequencing was performed on an Illumina HiSeq 2500 machine (50bp read length, single-end). Sequential ChIP-seq were prepared with two independent biological replicates. Long fragment single-mark ChIP-seq were prepared without replicates, but a bipartite chromatin pattern was clearly visible in these samples.

ChIPmentation of mouse ESCs

Mouse ESCs were cross-linked with 1% formaldehyde in DMEM 1X for 10 minutes at room temperature (RT) and quenched with 125mM glycine for 5 minutes at RT. Cells were rinsed twice with PBS 1X/ FCS 4% and collected by scraping. Normally we prepare at least two biological replicate, but for some experiments of bulk ChIP-seq, independent biological replicates were not prepared, but analysis were supported by the use of internal controls and the sample was normalized and compared with samples from related condition (see below). Preparation of ChIP-seq library was done by ChIPmentation protocol. 1,000,000 cells (H3K4me2, H3K27ac, H3K27me3), or 30,000,000 cells (Ser5P RNAPII-S5P, Ser2P RNAPII, NELF-b) were used. Cells were lysed in 200 - 800 µl of Sonication Buffer (10mM Tris HCl pH8, 5mM EDTA, 0.15% SDS, 0.1X PBS, 1X Protease Inhibitor Cocktail) for 15 min on ice, and sonicated using the Covaris machine or the Bioruptor Pico machine to obtain DNA fragment the size of which distributes between 150bp and 500bp. After centrifugation (1 min, 10,000rpm, 4°C), the supernatant was

transferred to new tubes, diluted five times with Equilibration Buffer. Chromatin solutions were incubated over-night at 4°C with 5µg of anti-H3K4me2 (Millipore, 07-030, PRID: AB_10099880), 5µg of anti-H3K27ac (abcam, ab4729, PRID: AB_2118291), 6µg of anti-H3K27me3 (Cell Signaling, 9733, PRID: AB_2616029), 30µg of anti-Ser5P RNAPII 4H8 (Covance, MMS-128P, PRID: AB_10013820), 30µg of anti-Ser2P RNAPII 3E10 (Active Motif, 61083, PRID: AB_2687450) and 30 µg of anti-NELF-b (abcam, ab237027) antibodies. The next day, 40 - 200 µl of protein G coupled to magnetic beads were added and the incubation was continued for 2 hours at 4°C (as for anti Ser2P RNAPII 3E10 rat-derived IgG_1, 200µl beads were pre-incubated with 60µl rabbit-derived anti-rat IgG antibody over-night to bridge primary antibodies and protein G). The beads were then washed twice with RIPA Buffer, and twice with 10mM Tris HCl pH8. Beads were resuspended in 30-90µl Tagmentation Buffer containing 1-3µl Tagment DNA Enzyme and incubated at 37°C for 10min. The beads were washed twice with RIPA Buffer and twice with TE Buffer. DNA was eluted from the beads with 60µl Elution Buffer at 65°C for 5hours. DNA was purified with SPRI AMPure XP beads (sample to beads ratio 1:2) and eluted in 25µl 10mM Tris HCl pH8. 2µl of each library was amplified in 10µl qPCR reaction to estimate the optimum number of enrichment cycles. Final enrichment of the libraries was performed in 50µl reaction. Enriched libraries were purified with size selection using SPRI AMPure XP beads (sample to beads ratio 1:0.6) to remove long fragments, recovering the remaining DNA (sample to beads ratio 1:2). Sequencing was performed on an Illumina HiSeq 2500 machine (50bp read length, single-end).

Sample preparation, RNA isolation and sequencing (RNA-seq)

For RNA-seq experiments, dissociated tissue was filtered and kept on ice until sorting. Sorted cells were collected into ice cold PBS 1X, and immediately pelleted by centrifugation (500 rcf, 5 minutes, 4°C). Total RNA was extracted by NORGEN Single Cell RNA Purification Kit (NORGEN, 51800) with genomic DNA digestion using RNase-Free DNase I Kit (NORGEN, 25710) according to manufacturer's protocol. For each experiment, three independent biological replicates were used.

Sequencing of poly A+ mRNA was done by Smart-seq2 protocol[37]. 1ng of total RNA was used as an input. Reverse transcription was conducted using 100U SuperScript II reverse transcriptase (Thermo Fisher, 18064014), 10U RNase inhibitor (Clontech, 2313A), 1X Superscript

II first-strand buffer, 5mM DTT (contained in SuperScript II reverse transcriptase), 1M Betaine (Sigma, B0300-1VL), 6mM MgCl2, 1µM template-switching oligos (TSOs) (exiqon), 1µM oligo-dT primer (Mycrosynth) and 1mM dNTP mix (Thermo Fisher, R0191) in total volume 10µl. Then PCR pre-amplification was done using 1X KAPA HiFi HotStart Ready Mix (KAPA Biosystems, KK2602), 0.1µM IS PCR primers (Mycrosynth) in total volume 25µl using 13 cycles of PCR. DNA was purified with SPRI AMPure XP beads (Beckman, sample to beads ratio 0.8:1) and eluted in 20µl 10mM Tris HCl pH8. Tagmentation reaction was done by Illumina Nextera XT DNA Library Prep Kit (Illumina, FC-131-1024) in total volume 5µl using 0.2ng DNA as an input, then amplification of adapter-ligated fragment was done using Illumina XT kits in a total volume 12.5µl with 12 cycles of PCR. Library was purified by SPRI AMPure XP beads (sample to beads ratio 0.6:1) and eluted in 12µl 10mM Tris HCl pH8. Sequencing was performed on an Illumina HiSeq 2500 machine (50bp read length, single-end).

Sequencing of total RNA was done by Ovation SoLo RNA-seq System (NuGEN, 0501-32), using 2.5 – 5ng of RNA as an input according to manufacturer's protocol. Sequencing was performed on an Illumina HiSeq 2500 machine (50bp read length, single-end).

Sample preparation and single-cell RNA sequencing (scRNA-seq)

To collect rhombomere 3 (r3)-derived progenitors from E10.5 *K20*$^{tdTomato/+}$ mouse, r2 – r4 regions were micro dissected. To collect E14.5 *Drg11*$^{vPrV\text{-}tdTomato/+}$ post-mitotic barrelette neurons from *Drg11*$^{tdTomato/+}$*;r2*$^{EGFP/+}$ mice, r2 – r3 derived regions were micro dissected. The boundary between r3 and r4 was identified by the position of the facial nerve. Dissected tissue was kept in PBS 1X containing 2µM actinomycin D (Sigma, A1410) for 10 minutes on ice, then treated with papain digestion mix (papain 10mg/ml/ cysteine 2.5mM/ HEPES pH7.4 10mM/ EDTA 0.5mM/ DMEM 0.9X/ 40µM actinomycin D) for 3 minutes at 37°C and immediately put on ice. Tissue was rinsed by ice-cold DMEM 1X containing 2µM actinomycin D, and dissociated by pipetting and filtered. r3-derived cells were collected by FACS (Extended Data Fig. 1d, Supplementary Figs. 1 and 2b).

Cells were sorted directly into 50µl of PBS-0.04%BSA. Cell concentration was determined using a TC20 automated cell counter (BioRad). As for E10.5 progenitors, 4,500 and 2,250 cells were respectively loaded into two different channels of a Chromium Single Cell A Chip (10X Genomics). As for E14.5 barrelette neurons, 8,000 cells were loaded into one channel of a Chromium Single Cell A Chip. Reverse-transcription, cDNA pre-amplification (13 cycles) and

library preparation were performed according to the manufacturer instructions (Chromium Single Cell 3′ Library & Gel Bead Kit v2, 10X Genomics). Libraries were sequenced on a Illumina NextSeq 500 platform (R1: 26bp, R2: 56bp, I1: 8bp). Demultiplexing and fastq files generation were performed using the CellRanger pipeline (10X Genomics).

Sample preparation and assay for transposase accessible chromatin (ATAC-seq)

Dissociated tissue was filtered and kept on ice until sorting. Two independent biological replicates were prepared. Sorted cells were collected into ice cold PBS 1X, and immediately pelleted by centrifugation (500 rcf, 5 minutes, 4°C). 50,000 – 70,000 cells were used for each experiment. Cells were washed once with 50µl PBS 1X and pelleted (500 rcf, 5 minutes, 4°C), and gently resuspended by pipetting in ice cold 50µl cold lysis buffer (10mM Tris HCl pH7.4, 10mM NaCl, 3mM $MgCl_2$, 0.1% IGEPAL CA-630) to extract nuclei and pelleted (500 rcf, 10 minutes, 4°C). Nuclei were washed once with ice cold 50µl Tagmentation Buffer (10mM Tris HCl pH8, 5mM $MgCl_2$) and pelleted (500 rcf, 10 minutes, 4°C). Then nuclei were tagmented in transposition reaction mix (2.5µl Tn5 Transposase (Illumina, FC-121-1030), 25µl 2X TD Buffer (Illumina, FC-121-1030), and 22.5µl nuclease free water) at 37°C for 30 minutes. DNA was purified with MinElute PCR Purification Kit (Qiagen, 28004) by 10µl 10mM Tris HCl pH8. Library amplification was started with 1X KAPA HiFi HotStart Ready Mix, 1.25µl primers and 0.6X Sybr Green (Thermo Fisher) in total volume 50µl: 72°C for 5min; 98°C for 1min; 5 cycles of 98°C for 15sec, 63°C for 30sec, 72°C for 1min. 5µl of 5 cycled DNA was taken and quantitative PCR was conducted to optimize PCR cycle (StepOnePlus Real-Time PCR Systems (Thermo Fisher)), and then remaining 45µl was further PCR amplified according to the necessity. DNA was purified with SPRI AMPure XP beads (Beckman, sample to beads ratio 1:2) and eluted in 20µl 10mM Tris HCl pH8. Sequencing was performed on an Illumina HiSeq 2500 machine (50bp read length, paired-end) or on an Illumina NextSeq 500 (75bp or 150bp read length, paired-end).

Sample preparation and 4C-Seq

4C assays using wild-type brainstem tissue were performed as described previously[38] with minor modifications. Two million of dissociated cells were cross-linked for 10 minutes with 2% paraformaldehyde, quenched with glycine and lysed in 50 ml lysis buffer (10mM Tris pH 7.5, 10mM NaCl, 2% NP-40, 1X protease inhibitors) for 30 minutes. Nuclei were then digested with

800U NlaIII enzyme (NEB, R0125) followed by 8 hours 'in nuclei' ligation at 16˚ C with 2000U T4 DNA Ligase (NEB, M0202) [39]. Reverse crosslinked and purified DNA was further digested with 50U DpnII enzyme (NEB, R0543), followed by circularization. 3200ng of 4C library was amplified with bait-specific inverse primers (read and amplification primers), pooled and purified. Amplified library was adaptor ligated, PCR amplified (8 cycles) and paired-end sequenced on the Illumina NextSeq 500 to obtain 75bp long reads. Read and amplification primers are listed in Supplementary Table 5.

Reverse transcription-PCR (RT-PCR) and quantitative-PCR (qPCR)

RT-PCR was done by SuperScript III Reverse Transcriptase (Thermo Fisher, 18080093) according to manufacturer's protocol. Oligo $d(T)_{18}$ mRNA primer (Thermo Fisher, SO131), 10mM dNTP Mix (Thermo Fisher, R0191) and RNase inhibitor (Promega, N2111) were used. qPCR was done using StepOnePlus Real-Time PCR System (Thermo Fisher) with SYBR Green PCR Master Mix (Thermo Fisher, 4309155) according to manufacturer's protocol using primers listed in Supplementary Table 5.

Quantification of *Fos, Egr1, Arc, Klf4, Junb* and *Gapdh* was done by ΔΔCt using *Actb* as an internal control, while the quantification of *Actb* (Extended Data Fig. 9e) was done using *Gapdh* as an internal control. A statistical analysis was performed by Welch's two-sample two-sided *t*-tests or by one-way analysis of variance (ANOVA) followed by Tukey's honest significant difference (HSD) post-hoc tests.

ChIP-qPCR

Embryonic hindbrain neurons were cross-linked with 1% formaldehyde in PBS 1X for 10 minutes at room temperature (RT) and quenched with 125mM glycine for 5 minutes at RT. Cells were rinsed twice with PBS 1X/ FCS 4%. Normally we prepared at least three biological replicates with 1,000,000 cultured neurons. Cells were lysed in 100 µl of Sonication Buffer (10mM Tris HCl pH8, 5mM EDTA, 0.15% SDS, 0.1X PBS, 1X Protease Inhibitor Cocktail) for 15 min on ice, and sonicated using the Covaris machine or the Bioruptor Pico machine to obtain DNA fragment the size of which distributes between 150bp and 500bp. After centrifugation (1 min, 10,000rpm, 4°C), the supernatant was transferred to new tubes, diluted four times with Equilibration Buffer. Chromatin solutions were incubated over-night at 4°C with 1µg of anti-H3K27me3 or anti-

H3K27ac antibodies. The next day, 40 µl of protein G coupled to magnetic beads were added and the incubation was continued for 2 hours at 4°C. The beads were then washed twice with RIPA Buffer and once with TE Buffer. DNA was eluted from the beads with 100 µl Elution Buffer at 65°C for 5hours. DNA was purified with SPRI AMPure XP beads (sample to beads ratio 1:2) and eluted in 60µl 10mM Tris HCl pH8. 4 µl of each sample was used for qPCR reaction. qPCR was done using StepOnePlus Real-Time PCR System with SYBR Green PCR Master Mix according to manufacturer's protocol using primers listed in Supplementary Table 5.
Quantification of *Fos* and *Egr1* was done by ΔΔCt using *Actb* or *Oct4* as an internal control. A statistical analysis was performed by Welch's two-sample two-sided *t*-tests or by one-way ANOVA followed by Tukey's HSD post-hoc tests.

ATAC-qPCR

ATAC-qPCR[40] experiments were performed as described previously with minor modifications. For each experiment, six independent biological replicates were used. Dissociated ESCs were collected into ice cold PBS 1X, and immediately pelleted by centrifugation (500 rcf, 5 minutes, 4°C). 20,000 cells were used for each experiment. Cells were washed once with 50µl PBS 1X and pelleted (500 rcf, 5 minutes, 4°C), and gently resuspended by pipetting in ice cold 50µl cold lysis buffer to extract nuclei and pelleted (500 rcf, 10 minutes, 4°C). Then nuclei were tagmented in transposition reaction mix (1 µl Tn5 Transposase (Illumina, FC-121-1030), 15µl 2X TD Buffer (Illumina, FC-121-1030), and 14µl nuclease free water) at 37°C for 30 minutes. DNA was purified with MinElute PCR Purification Kit by 10µl 10mM Tris HCl pH8. Library amplification was started with 1X KAPA HiFi HotStart Ready Mix, 1.25µl primers in total volume 50µl: 72°C for 5min; 98°C for 1min;11 cycles of 98°C for 15sec, 63°C for 30sec, 72°C for 1min. DNA was purified with SPRI AMPure XP beads (Beckman, sample to beads ratio 1:2) and eluted in 60µl 10mM Tris HCl pH8. 4 µl of each sample was used for qPCR reaction. qPCR was done using StepOnePlus Real-Time PCR System with SYBR Green PCR Master Mix according to manufacturer's protocol using primers listed in Supplementary Table 5. Quantification of *Fos* and *Egr1* was done by ΔΔCt using *Actb* as an internal control. A statistical analysis was performed by Welch's two-sample two-sided *t*-tests.

Plasmid construction

dCas9 over-expression vector (control vector, pEF1a_dCas9) was constructed from dCAS9-VP64_GFP plasmid by removing the *VP64* sequence. dCAS9-VP64_GFP was a gift from Feng Zhang (Addgene Plasmid #61422)[41]. Subsequently, dCas9-UTX fusion protein over-expression vector (pEF1a_dCas9-UTX) was generated by inserting the CDS of mouse *UTX* to the 3' of *dCas9*. Guide RNA (gRNA) over-expression vector (pGuide_EGFP) was constructed by replacing the *hSpCas9* sequence of pX330-U6-Chimeric_BB-CBh-hSpCas9 plasmid with the *EGFP* sequence. pX330-U6-Chimeric_BB-CBh-hSpCas9 plasmid was a gift from Feng Zhang (Addgene Plasmid #42230)[42]. Two gRNAs targeting the gene body regions of mouse *Fos*, *Egr1*, *Gapdh*, and *Actb* gene loci are designed and inserted into the BpiI (Thermo Fisher, ER1011) cut site of pGuide_EGFP. Inserted sequences are following. Targeted sequences are listed in Supplementary Table 5.

TVA/EnvA trans-synaptic tracing and PrV neuron dendrite analysis

EnvA-pseudotyped G-deleted rabies viruses (*EnvA-SAD-ΔG-Rabies:EGFP* virus[43]) were stereotaxically injected (Kopf Instruments) into P3 wild-type $K20^{TVA/+}$ or barrelette neuron activity-deprived $K20^{TVA/Kir}$ VPM (from Bregma: 1.2 mm posterior, 1.0 mm lateral, and 2.4 mm ventral). A small craniotomy was performed and a pulled-glass pipettes were used for local infusion of the virus by multiple short pulses using a picospritzer (Parker). To trace vPrV barrelette and for the transsynaptic experiments neurons mice were sacrificed 7 days later (P10), perfused with 4% PFA, and 60-µm vibratome brain sections collected. To trace vPrV barrelette at least 5 neurons were analysed for each mouse (for $K20^{TVA/+}$ and $K20^{TVA/Kir}$ conditions). Confocal imaging was performed with 40× objective (Zeiss LSM700 microscope). Arbors were traced using ImageJ neurite tracer software.

Matlab ad-hoc code was created in order to calculate the symmetric index and the surface ratio. Neurons were divided in the two half spheres by a rotating plane (5 degrees steps). Dendrites density was calculated in both half spheres and the ratio among the two densities (numerator as lower term and denominator as the highest) was calculated for each step of the plane. In the Extended Data Fig. 1n, o, example of the ratio on each plane calculated for a $K20^{TVA/+}$ (n) and $K20^{TVA/Kir}$ (o) barrelette neuron is shown by the color codes (percentage of the lower dendritic density over the higher dendritic density). The lower point (zero in case of total asymmetry) was defined as the Symmetric Index. Surface Ratio describes the probability that the plane will be

settled in a point of the function having a Symmetric index equal to zero. Statistical analysis was performed by Welch's two-sample two-sided *t*-tests on the averages of the Symmetry indexes and Surface Ratios calculated from different cells in the same animals.

3D reconstruction and nuclei volume calculation

Volumes of PrV were calculated upon a 2.5D reconstruction. Coronal vibratome sections (60 µm) of the PrV from P8 pups $K20^{tdTomato/+};r2^{EGFP/+}$ or $K20^{Kir/+};r2^{EGFP/+}$ animals were collected. Imaging was performed using a confocal microscope (Zeiss LSM700). 2.5 D reconstructions were made using Fiji/ImageJ and volumes were calculated upon processing with Imaris® software selecting the corresponding color. A threshold was imposed in order to eliminate small dots and background. A statistical analysis was performed by Welch's two-sample two-sided *t*-tests.

Histological analysis

Postnatal brains were perfused, dissected and also left overnight in 4%PFA. Sections were obtained upon vibratome cut (60 µm). Sections were incubated overnight at 4°C with primary antibodies anti-GFP (1:500, Abcam, ab290, PRID: AB_303395) or anti-RFP (1:1,000, chromotek, 5f8-20), rinsed three times in PBS and incubated for 60–90 min at room temperature with Alexa Fluor 488- and/or 546- and/or 647 conjugated secondary antibodies (1:1,000, Invitrogen, A11034, PRID: AB_2576217, A21245, PRID: AB_2535813, A11077, PRID: AB_2534121). Nuclei were stained with DAPI (ThermoFisher, D1306). Cytocrome oxydase staining was performed as previously described[1].

Reference genome and annotation

The mouse GRCm38/mm10 genome assembly was used as a reference and the analyses were done using R (https://www.r-project.org, versions 3.5.1 and 3.5.3).

TSS selection:

Transcription start sites (TSS) were obtained from the *TxDb.Mmusculus.UCSC.mm10.knownGene* Bioconductor package (https://doi.org/doi:10.18129/B9.bioc.TxDb.Mmusculus.UCSC.mm10.knownGene, version 3.4.0), and in the case of multiple TSSs per gene the one with the highest variance in chromatin accessibility (ATAC-seq, window of 2kb centered on TSS) was chosen. For this, ATAC-seq reads

were counted using *qCount* from *QuasR* [44](version 1.22.1) for the E10.5 progenitor, and E14.5, E18.5, and P4 barrelette (vPrV) samples. The raw counts were normalized using calculated scaling factors as follows: For each sample, ATAC peaks were called using MACS2 [45](version 2.1.1.20160309) with options *–f BED*, *--nomodel*, *--shift -100*, *--extsize 200*, and *--keep-dup all*. For each sample, the number of reads per base (rpb_{sample}) was calculated by dividing the sum of reads outside called peaks by the non-peak genome (autosomes only). The scaling factor (sf) was calculated as sf_{sample} = min(rpb of all samples)/rpb_{sample}. The corrected counts were log2-transformed with a pseudo-count of 8. For each gene, the TSS with the highest variance in accessibility across all conditions was selected. If multiple TSSs shared equally high variance, a random TSS was chosen. Only autosomal genes were considered. Choosing the TSS in other ways gave similar results. Namely, selecting the TSS that has the maximum ATAC signal per gene resulted in a set of TSSs similar to the previous one. Excluding genes where the TSS was randomly chosen, the TSS overlap between the two methods was 94.6%.

Promoter and gene body regions:

Promoter (P) regions were defined as 1000bp upstream and 500bp downstream of the chosen TSS per gene. Gene Body (GB) regions were defined as the region between 1001 bp and 3000 bp downstream the chosen TSS, giving GBs a maximum width of 2 kb (green and blue rectangles in Extended Data Fig. 5a). This definition also left a gap of 500 bp between promoter and gene body regions in order to prevent any overlap of signals between them. Genes with overlapping promoter or gene body regions, as well as genes that were too short were excluded.

Transcription end site:

The transcription end sites (TES) were defined for genes in the P and GB sets as the last 2kb of the transcript with the most downstream transcript end site from all transcripts of each gene.

Salmon spliced and unspliced transcript indexing:

Spliced transcripts were obtained from the *GENCODE*[46] vM19 transcriptome fasta file (ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M19/gencode.vM19.transcripts.fa.gz), and for each transcript, the corresponding unspliced version was created as the full sequence from the first exon to the last (including all introns and exons in between). Sequence-duplicated spliced transcripts and short spliced transcripts (less than 31 bp) along with their corresponding unspliced forms were removed. *Salmon* [47](version 0.12.0) was used to build a quasi mapping index on the fasta file with *–k 31*.

Read alignment to the reference genome

mRNA-seq (SmartSeq2):

Reads were aligned to the reference genome with *STAR* [48](version 2.5.2b) using parameter values similar to the ones suggested by *ENCODE* with options *--outFilterType BySJout*, *--outFilterMultimapNmax 20*, *--outMultimapperOrder Random*, *--alignSJoverhangMin 8*, *--alignSJDBoverhangMin 1*, *--outFilterMismatchNmax 999*, *--alignIntronMin 20*, *--alignIntronMax 1000000*, and *--alignMatesGapMax 1000000,* and converted to bam files with *samtools* [49](version 1.2).

Total RNA-seq (Ovation SoLo RNA-Seq):

Reads were trimmed for the GATCGGAAGAGCACACGTCTGAACTCCAGTCAC adapter from the 3' end with *cutadapt*[50] (version 1.15) using *overlap=1*. Trimmed reads were mapped to the reference genome using *STAR* (version 2.5.2b, using identical parameters as above for the SmartSeq2 data) and *samtools* (version 1.2). In addition, for the E14.5 *Drg11*$^{vPrV\text{-}ZsGreen/+}$, E14.5 *K20*$^{tdTomato/+}$ and E14.5 *Ezh2cKO*$^{r3\text{-}RFP}$ samples, reads were used to estimate spliced and unspliced transcript abundances per gene with *salmon* (version 0.12.0) with the parameters *--validateMappings*, *--fldMean 350, --seqBias*, and *--gcBias*. For downstream analyses, transcript abundances were combined to obtain gene abundances (see below).

ATAC-seq:

For better comparability with 51-mer paired-end (PE) samples, the samples that had been sequenced as 76-mer PE reads were trimmed to 51-mer PE reads by removing the last 25 bp using *cutadapt* (version 1.12 and 1.18). In addition, the adapter sequence CTGTCTCTTATACACA was trimmed from the 3' end of all samples with *overlap=1*. The trimmed reads were aligned with *bowtie2*[51] (version 2.3.0 and 2.3.4.2) with the options *--fr*, *--minins 0*, *--maxins 1000*, *--no-discordant* and *--dovetail*, and converted to bam files using *samtools* (version 1.2 and 1.9).

ChIP-seq:

The ChIP-seq samples that were prepared by the ChIPmentation protocol were trimmed for the adapter sequence CTGTCTCTTATACACA from the 3' end using *cutadapt* (versions 1.15 and 1.18) with *overlap=1*. The ChIP-seq samples prepared by NEBNext Ultra DNA Library Prep Kit (i.e. E12.5 short-term cultured hindbrain neurons that were treated by DMSO (control) or TSA, see above) were trimmed using *cutadapt* (version 1.15) for the

GATCGGAAGAGCACACGTCTGAACTCCAGTCAC adapter from the 5' and then the 3' ends since the adapter was seen on both ends during visual inspection of the fastq files. The trimmed reads were mapped to the reference genome using *bowtie2* (versions 2.3.3.1 and 2.3.4.2) and converted to bam files using *samtools* (versions 1.6 and 1.9).

The ChIP-seq samples from the public datasets were processed in the same way, but differed at the adapter trimming step. For the mouse embryonic and adult cortex ChIP-seq (GSE93011 and GSE52386, GSE63137), reads were trimmed for the GATCGGAAGAGCACACGTCTGAACTCCAGTCAC adapter from the 5' end. The cultured embryonic cortical neuron ChIP-seq (GSE21161) was mapped to the reference genome using *bowtie*[52] (version 1.0.0) with the *–C* option to allow for colorspace alignment. For the public ESC dataset (GSE36114 and GSE94250), the adapter (GATCGGAAGAGCACACGTCTGAACTCCAGTCAC) was trimmed on both the 5' and 3' ends, while it was trimmed from the 5' end for the dataset from Lavarone et al.,[22] (GSE116603).

4C-seq:

Coordinates of expected 4C fragments were created by *in silico* digesting the reference genome with restriction enzyme one (DpnII). Valid fragments were defined as non-blind fragments (containing a site for restriction enzyme two (NlaIII)) where the distance between the fragment start/end to the first/last site for restriction enzyme two was greater than 30bp. Reads were aligned to the genome with *QuasR* (version 1.22.1) using *qAlign* with default parameters. Sample quality was assessed using the criteria (percent of covered fragment ends 0.2 Mb around viewpoint and percent of reads in cis) described in [53]. Reads were counted per fragment using *qCount* from the *QuasR* package, normalized by dividing by the number of aligned reads in a sample on the cis chromosome (the chromosome containing the viewpoint) and multiplying with 1e6 (counts per million cis-chromosome alignments) and exported as a bedGraph file for visualization using *rtracklayer*[54] (version 1.42.2). For visualization of combined replicates, normalized fragment counts were averaged across replicates and smoothed using a running mean over three adjacent fragments.

BigWig Files:

For genome browser views, the number of alignments per 100 bp window in the genome and per million alignments in each sample were calculated and stored in BigWig format with *QuasR* (version 1.22.1) using the *qExportWig* function with *binsize=100* and *createBigWig=TRUE.* The

autosomal library size was used to normalize for library size differences across samples, since each sample was a mixture of mouse embryos of different sexes (this was achieved by setting *scaling* = autosomal library / full library size * 1e6, which will cancel out the full library size normalization the function does internally). Counts per bin were averaged across replicates to create BigWig files that represented a single condition. For the ATAC-seq data, *pairedAsSingle=TRUE* option was used in order to create a wig file of fragment cut sites, instead of fragment mid points.

For a subset of samples, normalized BigWig files were created in order to reduce between-sample non-linearities as follows: For the ATAC-seq, H3K27me3, H3K4me2, and H3k27ac of the samples from E10.5 progenitors, E14.5vPrV, E18.5vPrV, and P4vPrV, initial BigWig files were created setting *scaling* equal to the total mapped reads (so that it cancels out the normalization in the function). The same kind of normalization was done for the RNAPII-S2P, RNAPII-S7P, and H3K36me3 samples at E14.5, E18.5 and P4. For each chromatin mark, counts from the resulting BigWig files were read into *R* and *limma's normalizeCyclicLoess*[55] (version 3.38.3) function was used on the log2-transformed counts per bin with *method="fast"*. This normalization corrected for differences in library size and other non-linearities between different samples. The corrected counts were transformed back from the log2 to the raw count space and exported as normalized BigWig files. These normalized BigWig files were scaled to the counts per million (CPM) level afterwards. Counts per bin were averaged across replicates to create BigWig files that represented a single condition.

To visualize the BigWig files for replicates of H3K27me3 ChIP-seq of the dCas9 or dCas9-UTX over-expressed cultured hindbrain neurons on top of each other (see Extended Data Fig. 9b), the *plotCoverage* function from the *wiggleplotr*[56] package (version 1.6.1) was used with the options *mean_only = FALSE*, *alpha=0.5*, *flanking_length = c(200,200)* and *rescale_introns = FALSE*.

Read quantification and abundance estimation

RNA-seq:

Reads were quantified for genes from the *TxDb.Mmusculus.UCSC.mm10.knownGene* Bioconductor package (https://doi.org/doi:10.18129/B9.bioc.TxDb.Mmusculus.UCSC.mm10.knownGene, version 3.4.0) and *QuasR*'s *qCount* function with *mapqMin=11* and *mapqMax=255*. Only autosomal

genes were kept since the samples consisted of mixtures of embryos of different sexes. Reads per kilobase million (RPKM) were then calculated, averaged across replicates, and log2-transformed with a pseudo-count of 0.1.

For samples whose transcript abundance was quantified with *salmon*, the *tximport* function from the *tximport*[57] (version 1.10.1) package was used to get the abundance matrix which contains the spliced and unspliced transcript per million (TPM) counts for each gene and sample. The TPMs were log2-transformed with a pseudo-count of 1.

Single Cell RNA-seq:

The single-cell RNA-seq data was quantified with *CellRanger*[58] (version 3.0.2), using a reference transcriptome generated from the mouse mm10 genome and the Gencode (version M20) genome annotation. Quantifications were imported into R (version 3.6.1) using the *DropletUtils* package[59] (version 1.4.3), quality control was performed using *scater*[60] (version 1.12.2), and normalization and log-transformation of UMI counts with *scran*[61] (version 1.12.1) and *scater*. Cells with total UMI count below 6,500 (E10.5) or 3,500 (E14.5) were excluded from further analyses. The remaining 2,646 cells from E10.5 had a median UMI count of 12,419 (range 6,502-47,172), and the median number of detected genes was 4,070 (range 2,454-7,739). The 2,513 retained cells from E14.5 had a median UMI count of 6,339 (range 3,505-21,288), and the median number of detected genes was 2,922 (range 1,836-5,847). A gene was considered 'observed' in a cell if the UMI count was greater than 0.

ATAC-seq and ChIP-seq:

Reads were quantified on the defined regions using *qCount* from *QuasR* (version 1.22.1). For H3.3, Cdk9, Ring1b and 8WG16 RNAPII, RPKMs were calculated using the read count sum on the P and GB as the library size per sample, averaged across replicates and finally log2-transformed with a pseudo-count of 0.1. In calculating RPKMs on the P region for S5P RNAPII and S7P RNAPII the library size was the read count sum on the P regions. The RPKMs for S2P RNAPII and H3K36me3 on the TES were calculated using the read count sum on the TES regions as the library size (Fig. 4a, b).

For the H3K27me3 ChIP-seq from the dCas9 experiments, the counts per million (CPM) on the GB regions were calculated, averaged across replicates, and log2-transformed with a pseudo-count of 8. From the MA plot shown in Extended Data Fig. 9d, a set of genes (green) that have similar average log2-transformed CPM (logCPM) as the Fos gene (red) were selected by

taking all genes +/- 0.5 logCPM of the Fos gene logCPM (1081 genes selected), in order to compare the Fos gene against a group of genes that have similar levels of H3K27me3. The median of the E14.5Bip genes' log2-fold change (logFC) that fall in this region of the MA plot was calculated as well (Extended Data Fig. 9d). The percentages of genes within the selected groups, that have a logFC value less than or equal to that of Fos (1.48%) or the median E14.5Bip genes (which fall in the green region) (45.24%) were calculated.

The Jmjd3 ChIPs were analyzed by quantifying reads on the defined P and GB regions. The sum of reads on P and GB was used as library size per sample, and used when calculating counts per million (CPM) across the genes. The CPMs on the GB regions were averaged across replicates, log2-transformed with a pseudo-count of 8, and used to create the MA plot in Fig. 7e, highlighting E14Bip genes that gain expression at P4 (RPKM > 3 at P4).

The pCREB ChIP-seq reads for E14.5 and P4 (see Fig. 7a) were quantified on the promoters, corrected for library size differences using the total mapped reads, and log2-transformed with a pseudo-count of 1. These were then used to calculate log-fold changes (logFC) between the P4 and E14.5 samples (logFC P4/E14.5). More detailed analyses of other ChIP-seq datasets that required specific normalizations are described below.

Visualizing combined chromatin states with t-SNE

The H3K27me3, H3K27ac, and H3K4me2 marks as well as chromatin accessibility (ATAC) were quantified for each gene in P and GB regions using the *qCount* function from *QuasR* (version 1.22.1). RPKMs were calculated using the sum of the total P and GB counts as the library size per sample. RPKMs were averaged across replicates and log2-transformed with a pseudo-count of 0.1 to obtain a matrix with genes as rows and the log2-RPKMs in P and GB regions of the mentioned chromatin marks as columns. t-Distributed stochastic neighbor embedding (t-SNE)[62] was used to create a 2-dimensional embedding of this matrix, placing genes with similar chromatin landscapes close together.

For the E14.5, E18.5 and P4 samples, an additional normalization step was performed to reduce non-linearities between samples and therefore improve comparability, using *limma's* (version 3.38.3) *normalizeCyclicLoess* with *method="fast"*. A t-SNE embedding was then calculated on the corrected counts using *Rtsne* (version 0.13).

Bipartite and bivalent scores:

To calculate a 'bipartiteness' score for genes, we selected genes that had low expression (RPKM < 3), greater levels of H3K27ac in P than in GB, and greater levels of H3K27me3 in GB than in P. The selected genes' H3K27ac in P and H3K27me3 in GB were ranked separately from low to high, and the two ranks for each gene were summed up.

A 'bivalency' score was calculated for genes that also had low expression values (RPKM < 3). H3K27me3 in P and H3K4me2 in P were ranked from low to high, and the two ranks for each gene were and summed up. Genes that were high in both marks thus got high 'bivalency' scores.

Both scores were evaluated separately for each time point by selecting the top 100 high-scoring genes and about 300 additional genes selected from a broad range of scores, and inspecting the distribution and levels of histone marks in a genome browser (e.g. to evaluate the E14.5 barrelette neuron profile, the first top 100 bipartiteness scoring genes were all visually inspected one by one whereas from 101-500 we randomly chose sets of 20 genes every 40-100 in the list for visual inspection). Inspected genes were manually classified into bipartite (high H3K27ac in P, high H3K27me3 in GB) or bivalent (both high H3K27me3 and H3K4me2 in P) in order to calculate the fraction of true positives at any given score (Extended Data Fig. 3a, b). For every developmental time point, a function was fitted on the true positive fraction versus the top scoring genes using the *interpSpline* function with *bSpline=TRUE* from the *splines*[63] R package. Producing the bipartite and bivalent scores using 2 biological replicates of the ChIP-seqs gave highly similar scores across genes between the 2 replicates and a pearson correlation of 0.91 and 0.93 for the bipartite and bivalent scores, respectively, between the two replicate scores.

The total number of bipartite and bivalent genes in each time point was then estimated as the area under the curve (AUC) of the fitted splines. The same steps were done to obtain bipartiteness and bivalency scores for genes using the public datasets for mouse ESCs, our own ESC datasets, E10.5 NCCs and E14.5 liver and heart tissues (Fig. 2a). To inspect the intersection of the top 100 bipartite genes from the different tissues, the *upset* function from the *UpSetR* package[64] (version 1.3.3) was used with *order.by="freq"* (see Fig. 2d).

Bipartite and bivalent regions on t-SNE:

Using the 100 top-scoring genes for the single time point t-SNE (E10.5 t-SNE) map and the top 300 genes for the combined t-SNE map (E14.5/E18.5/P4 combined t-SNE), two-dimensional densities for bipartite and bivalent genes were estimated with the *kde2d* function from the *MASS*[65]

(7.3.51.1) package using a grid size of 25, and shown as the contour lines on the t-SNE plots (see Extended Data Fig. 5g, l).

Chromatin profiles of top 100 bipartite and bivalent genes:

The chromatin mark distribution surrounding the selected TSSs (2.5 kb upstream and 6 kb downstream) for the top 100 bipartite and top 100 bivalent genes in E14.5 barrelette, ESCs, E10.5 NCCs, E14.5 mouse liver and E14.5 mouse heart tissues were obtained using *QuasR*'s *qProfile* (version 1.22.1) function (see Fig. 2b). The counts were corrected for library size differences by multiplying the counts by scaling factors (sf) that were obtained using the total number of reads mapped on the autosomal genome as the library size per sample: sf_{sample} = min(lib size) / sample-specific lib size. The corrected counts were averaged per condition and chromatin mark and smoothed with the *runmean* function with *endrule="constant"* and using *k=601* for the E10.5 NCC and *k=801* for the rest.

For visualization, chromatin mark profiles were scaled to the interval [0,1] for easier comparison as follows: For each mark, we chose a minimum (min) value that was the median of the counts on the first and last 500 bp (-2.5kb to -2kb, and 5.5kb to 6kb relative to the TSS) and a maximum (max) value that was the maximum count in both the bipartite and bivalent profiles to also make the two profiles comparable to each other after the scaling. For each mark, the min value was then subtracted from the profile and the result was divided by (max-min) resulting in a scaled count between 0 and 1 (scaled count = (raw count - min) / (max - min)).

Chromatin dynamics in the t-SNE:

We calculated Euclidean distances of genes between P4 and E14.5 on the original 8-dimentional space consisting of normalized log2(RPKM) counts of ATAC, H3K27me3, H3K27ac, and H3K4me2 on the P and GB regions, and colored the E14.5 t-SNE by this distance (see Extended Data Fig. 5h). To further explore the properties of the E14.5Bip genes across development, they were divided into 3 groups: E14.5Bip genes that become expressed at P4 (RPKM >=3 at P4), that become bivalent (selected by taking the genes that move into the red bivalent contour on the t-SNE at P4, see Fig. 3d), and that remain bipartite (the remaining E14.5Bip genes) (see Extended Data Fig. 6a).

Selection of genes to compare to E14.5Bip Genes

Groups in Fig. 4:

We first selected the top 100 bipartite genes at E14.5 (E14.5Bip genes, using the 'bipartiteness' score) and the top 100 bivalent genes at E14.5 (E14.5Biv genes, using the 'bivalency' score) and excluded 3 genes that were found in both sets, leaving two distinct sets of 97 E14.5Bip and 97 E14.5Biv genes. Several control sets of the same number of genes (97) were then created as follows: E14.5AcP genes were sampled from all genes except for the top 400 E14.5 bipartite genes and E14.5Biv genes to have similar H3K27ac distribution in P as the E14.5Bip genes, using the *sampleControlElements* function from the *swissknife* package (https://github.com/fmicompbio/swissknife, version 0.10). E14.5mRNALow genes that match the E14.5Bip genes in log2-RPKM mRNA expression (smartSeq2) were sampled similarly from all genes except E14.5Biv genes, the top 400 E14.5 bipartite genes and E14.5AcP genes. Finally, two gene sets were sampled from the bottom and top 30% of genes ordered by mRNA expression, excluding any of the genes already contained in the previously defined groups.

Genes in Extended Data Fig. 7a, b:

Entrez identifiers from the top 100 E14.5Bip genes were first mapped to their corresponding Ensemble IDs using the *useMart* function from *biomaRt*[66] (version 2.38.0) with the options *biomart="ensembl"* and *dataset="mmusculus_gene_ensembl"*, resulting in a set of 90 successfully mapped E14.5Bip genes. A control set of 90 genes matching the E14.5Bip genes in spliced transcript abundance (in the log2(average TPM + 1)) and not in the top 400 E14.5 bipartite genes was then randomly sampled as described above. Genes with zero total counts were excluded (between 0 and 14 genes per replicate and gene set).

Ezh2cKO analyses (E14.5)

H3K27ac and H3K27me3 ChIP-seq (Fig. 5a):

The knock-out of *Ezh2* caused genome-wide changes in histone modification levels and thus the total number of alignments cannot be used as a library size proxy for normalization. To normalize these samples for differences in library size, we therefore selected a group of genes that were constantly expressed, assuming that the promoter regions of these genes are likely to not exhibit changes in chromatin marks between the two conditions. We selected genes that had an absolute logFC of less than 0.2 between the wt and *Ezh2cKO* and an average logCPM >= 5 based on differential expression analysis with *edgeR*[67] (version 3.24.3). We further excluded from our selection of control genes the genes that overlapped with any of the Polycomb regions on our t-

SNE maps (including the bipartite and bivalent regions), resulting in a total of 1,166 genes. Promoter (as defined in the previous sections: -1kb and +500bp relative to the selected TSS per gene) reads of the selected genes were summed for each sample *i* (N_i) and used to calculate scaling factors for sample *i* = min(N_i)/ N_i .These scaling factors were then multiplied by the raw counts in each sample, and corrected counts were then averaged across replicates and log2-transformed with a pseudo-count of 1.

SmartSeq2 MA Plot (Fig. 5b):

The CPMs of the autosomal genes were calculated, averaged across replicates, and log2-transformed with a pseudo-count of 8. The logFC (M values) and average logCPM (A values) were calculated and E14.5Bip genes were shown in red.

E14.5 vPrV fraction of TSS proximal reads (Fig. 5c):

To calculate the fraction of reads at the beginning of each gene, we used the selected TSS for each gene as described above. The beginning of the gene was defined as 200 bp downstream and 100 bp upstream of the TSS, including only exonic regions. Reads were then counted in this region for each gene and normalized by calculating CPMs using the same library size as for the normalization of the whole gene counts. Replicate CPMs were averaged, and only genes with a CPM of at least 1 in both the wt and *Ezh2cKO* were kept (of the E14.5Bip genes, 82 out of 100 remained). For visualization, the log2 ratio compared to the whole genes was calculated as $\log2((CPM_{beginning} + 1)/(CPM_{whole} + 1))$.

Total RNA (SoloSeq) Spliced and Unspliced Transcript Abundance:

We used the total RNA-seq dataset to look at differences in spliced and unspliced transcript abundances between the wt and *Ezh2cKO* on genes that had a TPM of at least 0.05 in both their spliced and unspliced forms in at least 1 sample. For each of the two conditions, the fraction of transcripts in the spliced form was calculated as (TPM spliced) / (TPM spliced + TPM unspliced) (Extended Data Fig. 7c).

ATAC Ezh2cKO vs Ctrl (Fig. 5d and Extended Data Fig. 8b):

E14.5 and P4 barrelette enhancers were defined using the peaks called on each condition with the ATAC-seq. The defined peaks were combined for both conditions, and only distal (at least 1kb away from any TSS) and autosomal peaks were kept and used as our set of enhancers. ATAC-seq for the *Ezh2cKO* and ctrl samples, as well as the E14.5 and P4 samples were quantified on the defined enhancers. The counts were log2-transformed with a pseudo-count of 1, underwent

normalization with the *normalizeCyclicLoess* function from *limma*, and were averaged across replicates. Using the bed files from Malik et al., 2014[14], we used the union of the two KCl-treated Fos ChIP bed files to define neuronal activity-dependent Fos targets, and divided our enhancers into Fos-overlapping (enhancer overlapping Fos target) and non-Fos-overlapping (enhancers not overlapping Fos targerts). We focused our analysis on enhancers that show an ATAC log-fold change greater than 1.5 from E14.5 to P4: 85 Fos-overlapping enhancers, and 3,882 non-Fos-overlapping enhancers, and looking at their *Ezh2cKO* vs ctrl ATAC logFC.

H3K36me3 Ezh2cKO vs WT (Extended Data Fig. 7e):

Reads were quantified on the full selected transcript per gene, and RPKMs were calculated followed by a log2-tranformation with a pseudo-count of 1. In each condition, we used the *Mclust* function from the *mclust* package[68] (verison 5.4.3) with *G=2* and *modelNames="V"* to fit a two-component Gaussian-mixture model to the log2-RPKMs, and used the fitted distributions to define a threshold value for a specific mark corresponding to an FDR of 0.05 (dashed green lines in Extended Data Fig. 7e) as described in Minoux et al., 2017[69].

Sequential ChIP analysis

Reads were quantified on joint P and GB regions, defined as the start of P to the end of GB. CPMs were calculated and averaged across replicates. Genes with low mappability (less than 80% mappable positions for 50-mer reads in joint P and GB regions) were removed. H3K27me3 was plotted against H3K27ac and colored by the sequential ChIP (Extended Data Fig. 4b). The sequential ChIP signal (color gradient) increases both in the directions of the x and y axes, indicating that the sequential ChIP was successful and that the two histone modifications co-exist on individual chromatin fragments.

We used the *Mclust* function from the *mclust* package[68] (verison 5.4.3) with *G=2* and *modelNames="V"* to fit a two-component Gaussian-mixture model to the log2-counts of each ChIP mark, and used the fitted distributions to define a threshold value for a specific mark corresponding to an FDR of 0.05 (dashed green lines in Extended Data Fig. 4b) as described in Minoux et al., 2017[69].

Defining activity response genes (ARGs)

bsARGs:

barrelette sensory ARGs (bsARGs) were defined using the mRNA samples of E14.5 vPrV, E18.5 vPrV, and E18.5 vPrV Kir-OE barrelette neurons. Only autosomal genes were considered and differential expression analyses were done with *edgeR* (version 3.24.3). The model was fit with *glmQLFit* using all the mRNA samples from E10 to P4 WT and Kir-OE on genes that have a CPM > 1 in at least 2 samples. Genes that were not or only lowly expressed at E14.5 (RPKM < 3) and that were upregulated from E14.5 to E18.5 WT were selected using a logFC >= 1.5 and a logCPM >= 2 (n = 702) (Extended Data Fig. 1q). Genes that were not or only lowly expressed at E14.5 (RPKM < 3) and that were downregulated between E18.5 vPrvKir-OE and E18.5 vPrV WT were selected as having a logFC <= -1 and a logCPM >= 2 (n = 102) (Extended Data Fig. 1r). The intersection of these two sets of genes (n = 56 genes) was taken as the bsARGs (Extended Data Fig. 1q-s). The barplot in Fig. 1c compared the bsARGs to all other genes that were also not expressed at E14.5. A gene was Polycomb-overlapping if any of its transcripts overlapped with the defined Polycomb regions using *findOverlaps* from GenomicRanges (version 1.34.0).

nbARGs:

Non-barrelette ARGs (nbARGs) were defined based on the literature[9-11] using the ARGs as described in the public datasets section. All activity-dependent genes (rapid and late induced genes) were used and only the genes that are not expressed (RPKM < 3) in all of E14.5, E18.5 and P4 in the barrelette neurons and do not overlap with the bsARGs were kept (83 genes). In the barplot (Fig. 1c) they were compared to all other genes that are not ARGs from the 3 papers, nor bsARGs, and also not expressed at any of the three developmental time points (RPKM < 3).

ATAC and H3K4me2 of bsARGs and nbARGs (Extended Data Fig. 1u):

All autosomal TSSs were obtained using the *transcripts* function with the *TxDb.Mmusculus.UCSC.mm10.knownGene* Bioconductor package (https://doi.org/doi:10.18129/B9.bioc.TxDb.Mmusculus.UCSC.mm10.knownGene, version 3.4.0) and the E14.5 vPrV ATAC-seq was quantified on all TSSs +/- 1kb. The TSS with the highest ATAC signal was chosen per gene, and if multiple TSSs had equally high values the selection was done randomly with the *sample* function. A 2 kb region (1 kb upstream and 1 kb downstream the selected TSS) was defined as a promoter region and the ATAC-seq and H3K4me2 were quantified on these regions with *qCount* from *QuasR* (version 1.22.1). CPMs per sample were calculated, averaged across replicates, and log2-transformed with a pseudo-count of 1.

To define the cutoff for being positive or negative for ATAC and H3K4me2, a two component Gaussian-mixture model was fit per mark on the log2-transformed counts using the *mclust* package[68] (version 5.4.1) and a threshold corresponding to an FDR of 0.05 was calculated as described in Minoux et al., 2017[69].

IEGs and LRGs (Fig. 1d):

A set of immediate early genes (IEGs) and a set of late response genes (LRGs) were obtained from the literature[9-11]. All bsARGs and nbARGs were considered and only the genes that are Polycomb-overlapping (a gene is considered to be Polycomb-overlapping if any one of its transcripts overlaps with the defined Polycomb regions as described in the section below), ATAC+ and H3K4me2+ (a gene is positive for ATAC or H3K4me2 if its count on the defined promoters described above was greater than the threshold derived with the set Gaussian-mixture model and FDR), and are rapid response genes in any of the papers were defined as IEGs (IEGs with RPKM < 3 at E14.5). The LRGs were the rest of the ARGs that are also polyc-overlapping, ATAC+, H3K4me2+, and non-IEGs (also with RPKM < 3 at E14.5). The IEGs and LRGs were viewed in the genome browser to manually classify them as bipartite or bivalent (in E14.5 vPrV).

Identifying E14.5 vPrV Polycomb domains

A hidden semi-marov model (HSMM) was used to identify Polycomb domains (chromatin that is positive for the H3K27me3 histone modification) genome-wide. This approach was used instead of typical peak callers since the size of Polycomb regions varies a lot across the genome, from short peaks up to large domains. A hidden markov model (HMM) or peak callers like *MACS2* will be unable to find optimal parameters that model both short and long regions and typically do not account for this variation. We therefore tiled the genome into 500 bp regions and the H3K27me3 ChIP-seq read count per tile was log2-transformed with a pseudo-count of 1. In an HMM setting these are the observations, and the hidden states that are to be inferred are “Polycomb” and “non-Polycomb”. The HSMM in addition models the duration of a state and thus was better suited to call Polycomb domains.

The *mhsmm*[70] (version 0.4.16) R package was used. The *hsmmspec* function was used to estimate the model parameters assuming a Gaussian distribution on the emissions (the log2-count data of the two states) and a gamma distribution on the sojourn (duration of a state). The *hsmmfit*

function then used the model's estimated parameters with *mstep=mstep.norm* and *maxit=100*, to call Polycomb and non-Polycomb regions.

TSA-treated cultured hindbrain neuron data

Quantifying and normalizing ChIP-seq counts on the gene body:

The control (DMSO-treated) and TSA-treated ChIP-seq (H3K27me3 and H3K27ac) samples had different log(raw cnt) vs percent GC trends (having quantified on 2kb tiles genome-wide). Left uncorrected, the logFC values between Ctrl and TSA-treated can be driven by differences in GC content. To correct for this, we first set out to define a reasonable set of background tiles (that are not positive for the marks). *MACS2* (version 2.1.2) was used to call peaks for each sample with the following parameters: *--nomodel*, *--extsize 100*, *--shift 0* and *--keep-dup all*. The background (non-signal) set of tiles were defined as those that do not overlap with any of the called peaks in any sample. The log2(raw cnt) vs percent GC trend was estimated with a linear fit for each sample on the defined background tiles (composing around 85% of all original tiles) using the *lm* function in R. These fits were used to correct the gene body counts (on the log2(raw counts + 8)). For each mark, the expected count for a given percent GC was calculated using the fits, and counts were corrected by setting one sample as the reference to which the trend of the other is pulled, by subtracting the difference in the expected values of the reference sample and the one being adjusted (delta), from the original count value: $\log2(\text{raw count adjusted})_{\text{sample}} = \log2(\text{raw count})_{\text{sample}} + \text{delta}_{\text{sample}}$, where $\text{delta}_{\text{sample}} = (\text{predicted reference sample count}) - (\text{predicted count})_{\text{sample}}$. In this manner, we corrected simultaneously for non-linearities in GC trends and library sizes using the fits on the background tiles. Where there were multiple replicates per condition, the corrected counts (which are in log2-space) were averaged per condition.

Fig. 7k:

The logFC values between TSA-treated and ctrl conditions on the gene body were calculated for H3K27me3 and H3K27ac using the corrected counts described above. For the mRNA, counts per million (CPM) were calculated on the autosomal genes, averaged across replicates, and log2-transformed with a pseudo-count of 1 to subsequently calculate logFC values. The E14.5Bip genes (top 100 bipartiteness ranked genes at E14.5 in barrelette neurons) are highlighted in Fig. 7k and colored by the mRNA (TSA/ctrl) logFC.

ESC *EedKO* analyses

Quantifying WT ChIP-seq:

The WT ESC levels of H3K27ac, H3K27me3 and H3K4me2 on the defined P and GB regions were quantified in R using *QuasR's* (version 1.22.10) *qCount* function. RPKM counts were calculated for the promoter and gene body count tables using the sum of the total reads on the promoter and gene body level as the library size for each sample. RPKMs were averaged per condition and log2-transformed with a pseudo-count of 0.1.

Quantifying mRNA (from Smart-seq2):

Counts on autosomal genes were quantified as described above. The counts per million (CPM) were quantified, averaged per condition, and log2-transformed with a pseudo-count of 1. The log-fold changes between *EedKO* and WT were calculated as the log2(*EedKO* CPM / WT CPM) for each gene. RPKM expression values of the autosomal genes were also calculated for the WT ESC condition, and averaged across replicates, to use in the definition of bipartite and bivalent genes. Bipartiteness and bivalency scores were calculated the same way as described above.

S5P logFC (Fig. 6a):

WT and *EedKO* counts for S5P and Nelf-b were quantified on the P, corrected for library size difference, and log2-transformed with a pseudo-count of 1. The logFC (*EedKO*/WT) violin plots of the ESCBip genes (top 100 bipartiteness scoring genes) were compared to the rest of the genes using the *vioplot* package[71] (version 0.3.0). The p-values were calculated using a two-sided Wilcoxon with the *wilcox.test* function from the *stats*[63] package in R with *alternative="two.sided"* and *paired = FALSE*.

Selecting genes positive for H3K27me3 on the gene body of ESC in WT:

We used the *Mclust* function from the *mclust* package[68] (verison 5.4.3) with *G=2* and *modelNames="V"* to fit a two-component Gaussian-mixture model on the gene body log2(RPKM+0.1) counts of the H3K27me3 WT mark, and used the fitted distributions to define a threshold value corresponding to an FDR of 0.05 as described in Minoux et al., 2019 [69]. Genes with H3K27me3 on the gene body higher than this threshold were selected as positive for the mark, and of those, genes that had a zero mRNA count in all samples were excluded (WT and *EedKO*), leaving us with a total of 3,457 genes that are positive for H3K27me3 on the gene body. These were further divided into 3 groups based on mRNA logFC between *EedKO* and WT: genes that

have a logFC < -1 (93), genes that have a logFC > 1 (1,067), and the remaining genes that have a -1 <= logFC <= 1 (2,297) (see Extended Data Fig. 7f).

S2P logFC (Extended Data Fig. 7f):

The S2P ChIP samples were quantified on the GB regions, corrected for library size differences, averaged per condition, and log2-transformed with a pseudo-count of 1. The logFC values were calculated by subtracting the WT log2-values from the *EedKO* log2-values, and displayed as violin plots for each of the 3 groups of genes (See Extended Data Fig. 7f). The p-values are the result of a two-sided Wilcoxon test using the *wilcox.test* function in R with *alternative="two.sided"* and *paired = FALSE*.

Sequence-specific characteristics of bipartite genes

Observed/expected CpG (oeCpG) ratio:

The regions 2kb upstream and 6kb downstream the TSS of the E14.5Bip and E14.5Biv genes were binned into consecutive 100bp bins, and the average oeCpG ratio was quantified in each set of genes. The oeCpG of a bin in a specific gene is calculated as the CG di-nucleotide frequency (with pseudo-count of 0.01) divided by the product of the C and G mono-nucleotide frequencies (with pseudo-count of 0.01) (See Extended Data Fig. 3f), using the *oligonucleotideFrequency* function from the *Biostrings*[72] package (version 2.50.2).

Motifs enriched in bipartite gene promoters:

The *monaLisa* package (version 0.1.28 and with R version 3.6.3, https://github.com/fmicompbio/monaLisa) was used to identify what motifs are enriched in E14.5Bip vs E14.5Biv promoters using the vertebrate list of motifs present in *JASPAR* (JASPAR2018)[73] and *Homer*[74] (version 4.10.4). Motifs that had an absolute log2-enrichment greater than 1 and were positively enriched in bipartite promoters were selected (RELB and FOXD1) (See Extended Data Fig. 3g).

Lavarone et al. Datasets (public):

mRNA MA plots

MA plots displaying logFC vs average logCPM values for each gene were made, using the public datasets from Lavarone et al[22]. A bipartiteness score was calculated using the same criteria as

described in previous sections, but using the ChIP-seq and mRNA-seq public datasets from Lavarone et al.[22] The top 100 bipartite genes specific to this dataset are highlighted in red in the MA plots (Extended Data Fig. 7h).

ATAC-seq and Ring1b (Fig. 6b,c):

The samples were quantified on the defined promoter and gene body regions, corrected for library size differences, and log2-transformed with a pseudo-count of 1. The top 100 bipartite genes of this ESC dataset are highlighted in red.

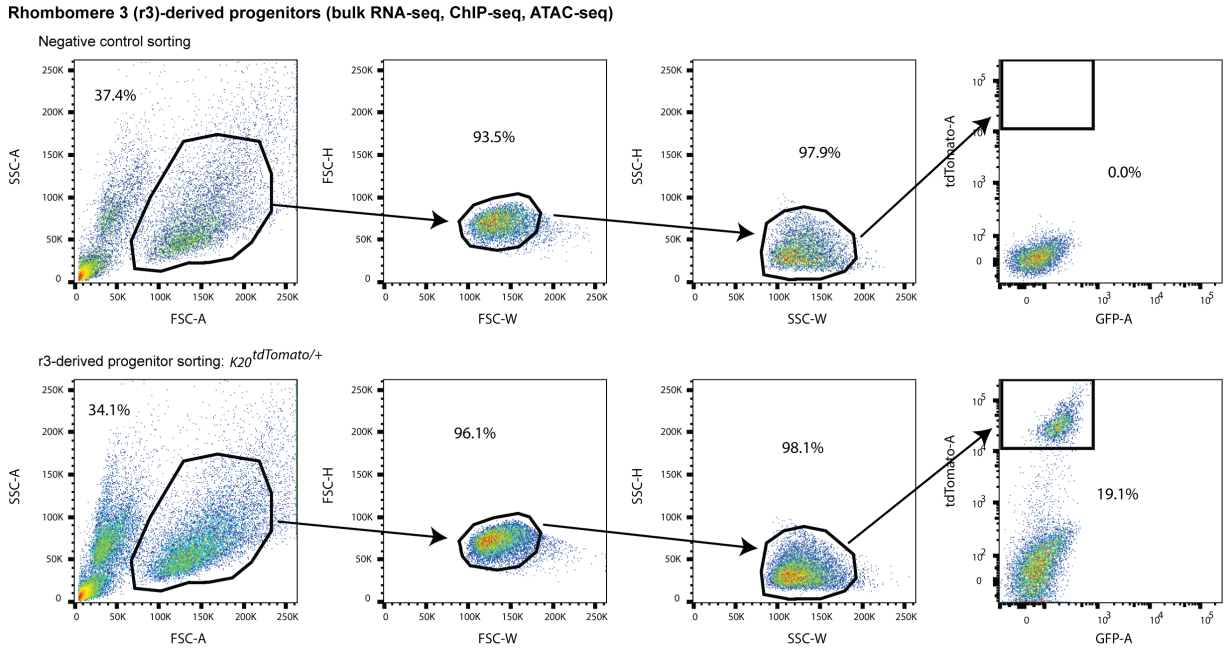
Public data sets used for analysis

A list of activity response genes (ARGs) (rapid primary response genes, delayed primary response genes and secondary response genes) from cultured cortical neurons treated by KCl were retrieved from Fig. 1 of Tyssowski et al., 2018[10]. ARGs from barrel cortex stimulated by environmental enrichment were retrived from Table 3 of Valles et al., 2011[11]. In Valles et al.,[11], 'early' ARGs (i.e. with a response comparable with IEGs) were defined as genes that are strongly up-regulated immediately after the completion of one hour of environmental enrichment (more than two folds) and show decreased expression after 4 hours (i.e. decreased expression at the time point of 4 hours compared with the time point of the completion of environmental enrichment). Eighteen genes were annotated in the mouse genome. 'Late' ARGs were defined as genes that are not strongly induced immediately after the completion of environmental enrichment, and show increased expression after 4 hours (at least 1.3 folds at the time point of 4 hours)[11]. Seven genes were annotated in the mouse genome. ARGs from visual cortex stimulated by light exposure were obtained from Table S3 of Hravatin et al., 2018[9]. 'Early' ARGs (i.e. with a response comparable with IEGs) were defined as genes that show increased expression in V1 excitatory neurons at the time point of 1 hour but not at the time point of 4 hours (at least three "a" and no "c" in excitatory neuron subpopulations). Twenty six genes were defined. 'Late'ARGs were defined as genes that are not induced at the time point of 1 hour but induced after 4 hours (at least two "c" and no "a" in excitatory neuron subpopulations). Nineteen genes were identified.

Public sequencing data sets were obtained as follows. Mouse cortical culture (GSE21161, GSE60192), mouse embryonic forebrain (GSE93011, GSE52386), mouse adult cortical excitatory neuron (GSE63137), mouse embryonic stem cells (GSE36114, GSE94250), mouse embryonic stem cells for PRC2-KO experiments in Lavarone et al., (GSE116603), mouse E14.5 heart tissues

(GSE82764, GSE82637, GSE82640, GSE78441, ENCSR068YGC), mouse E14.5 liver tissues (GSE78422, GSE82407, GSE82615, GSE82620, ENCSR032HKE) and E10.5 mouse neural crest cells isolated from the frontal nasal process (FNP) (GSE89437).

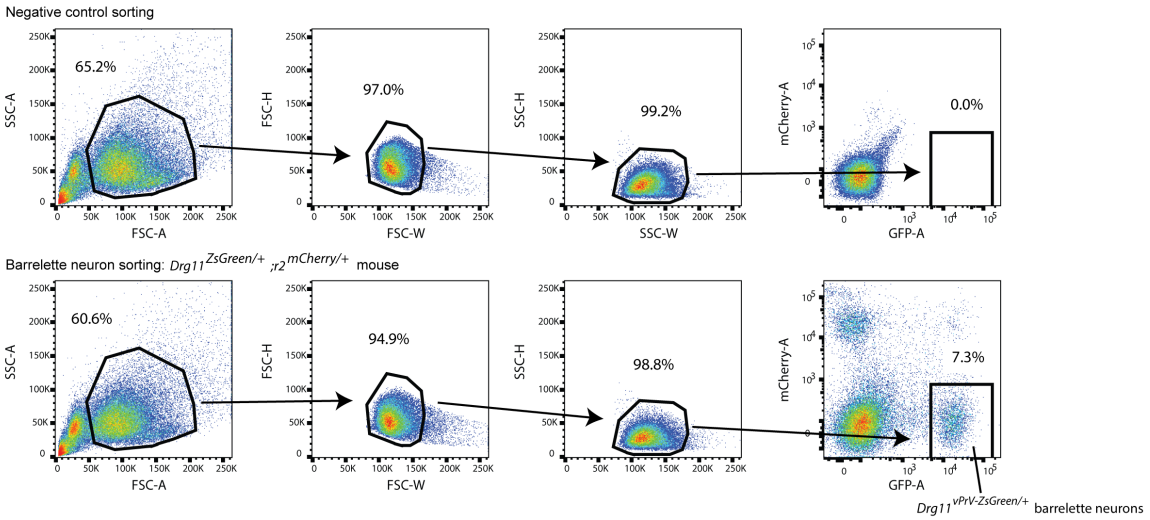
Supplementary Figures



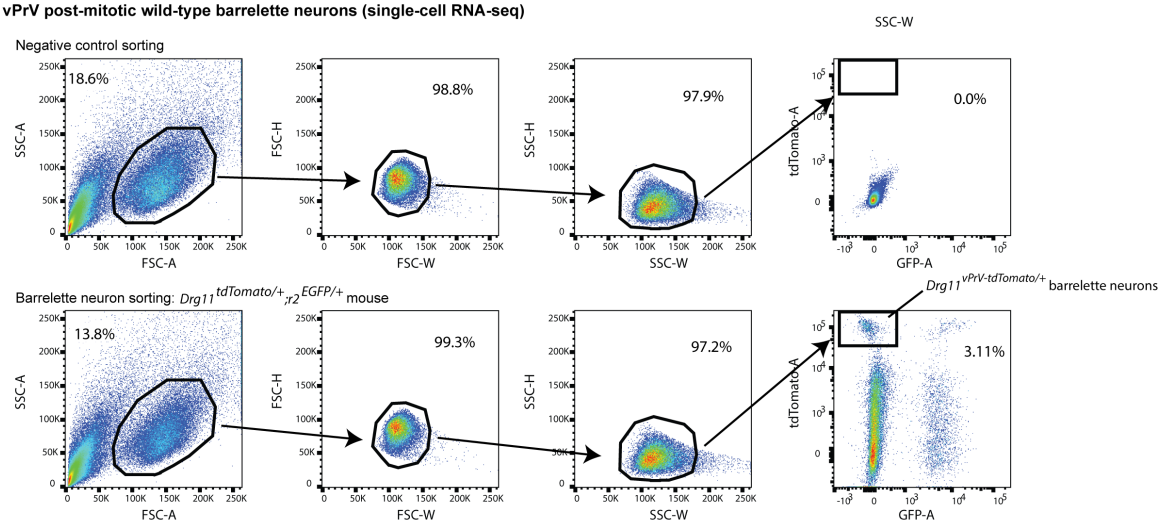
Supplementary Fig. 1 | FACS sequential gating strategy for E10.5 rhombomere 3 (r3)-derived progenitors.

Gating layouts of negative control and *K20*$^{tdTomato/+}$ mice are compared. Related to Fig. 1a, Extended Data Fig. 1. Nomenclatures for the mouse lines are summarized in Supplementary Table 1.

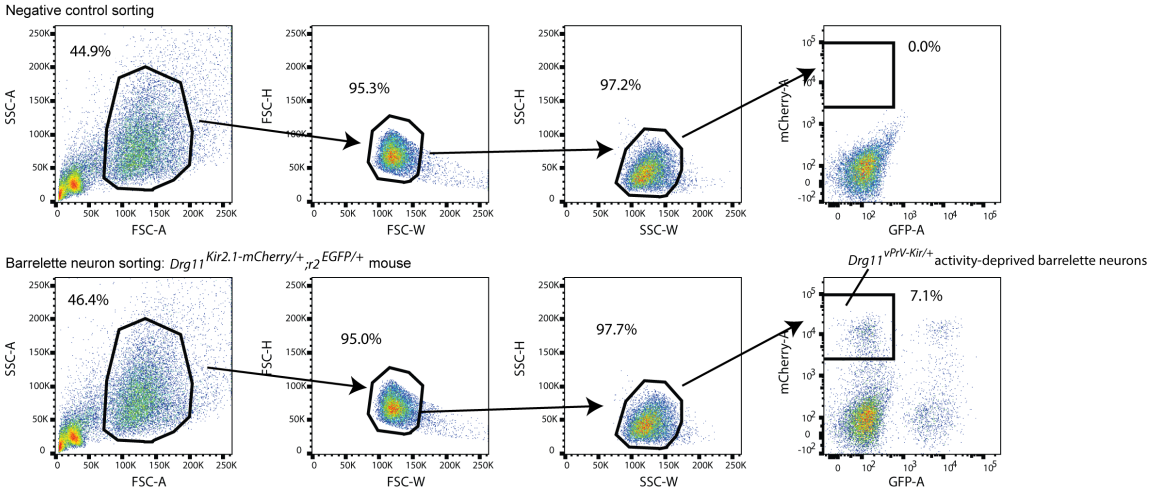
a **vPrV post-mitotic wild-type barrelette neurons (bulk RNA-seq, ChIP-seq, ATAC-seq)**



b **vPrV post-mitotic wild-type barrelette neurons (single-cell RNA-seq)**



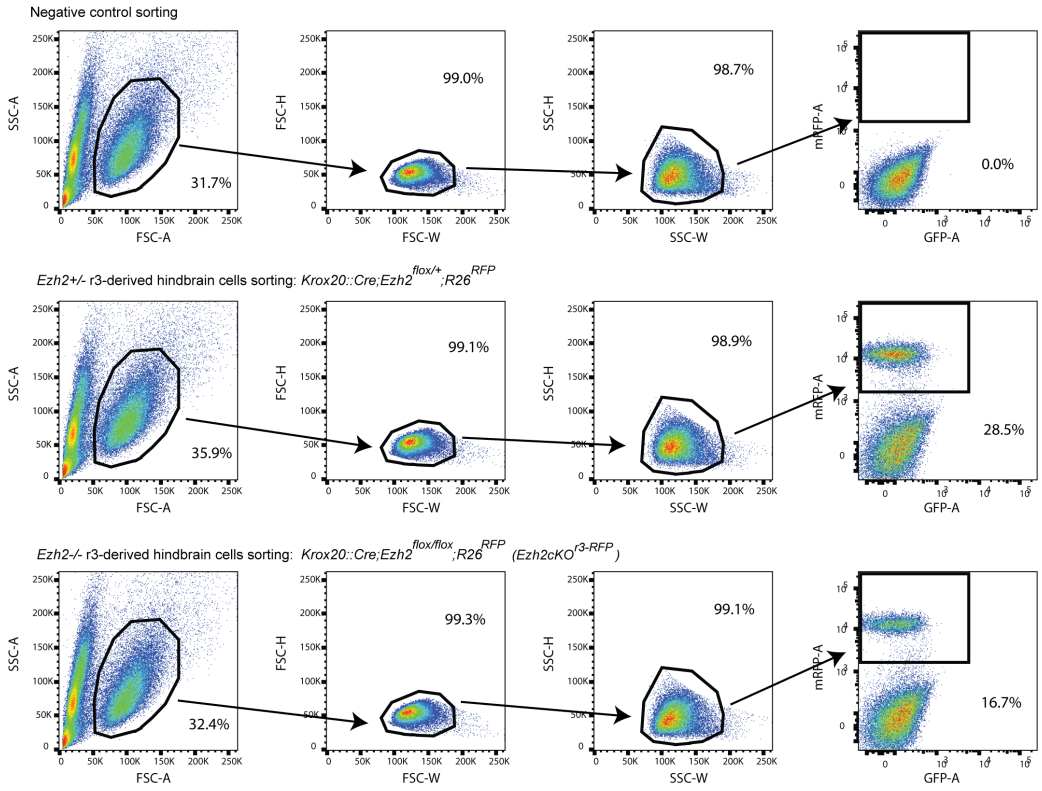
c **vPrV Kir2.1 over-expressing post-mitotic barrelette neurons**



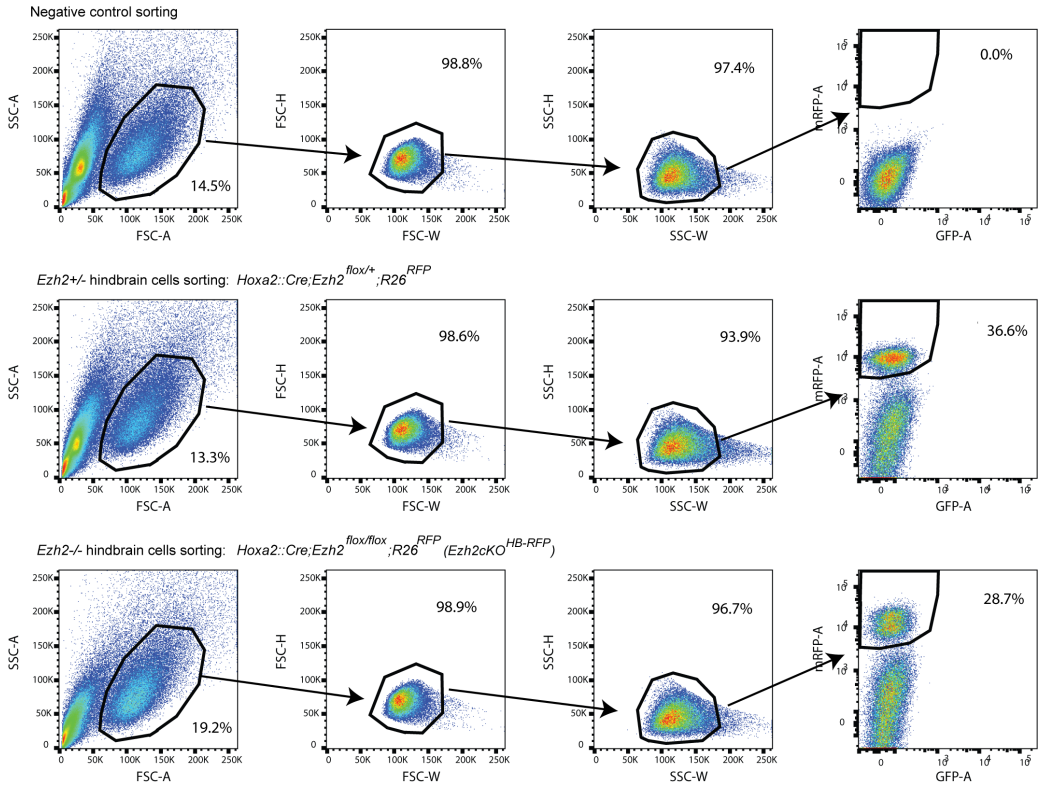
Supplementary Fig. 2 | FACS sequential gating strategy for vPrV post-mitotic barrelette neurons.

a, Gating layouts to sort wild-type *Drg11*$^{vPrV\text{-}ZsGreen/+}$ post-mitotic barrellette neurons. Negative control and *Drg11*$^{ZsGreen/+}$;*r2*$^{mCherry/+}$ mice are compared. Related to Fig. 1a, Extended Data Fig. 1a-c. **b**, Gating layouts to sort wild-type *Drg11*$^{vPrV\text{-}tdTomato/+}$ post-mitotic barrellette neurons. Negative control and *Drg11*$^{tdTomato/+}$;*r2*$^{EGFP/+}$ mice are compared. Related to Fig. 1a, Extended Data Fig. 1d. **c**, Gating layouts to sort neuronal activity-deprived Kir2.1 over-expressing *Drg11*$^{vPrV\text{-}Kir/+}$ post-mitotic barrellette neurons. Negative control and *Drg11*$^{Kir/+}$;*r2*$^{EGFP/+}$ mice are compared. Related to Fig. 1a, Extended Data Fig. 1e. Nomenclatures for the mouse lines and barrelette neurons are summarized in Supplementary Table 1.

a ***Ezh2cKO*** **r3-derived hindbrain cells (RNA-seq, ChIP-seq)**

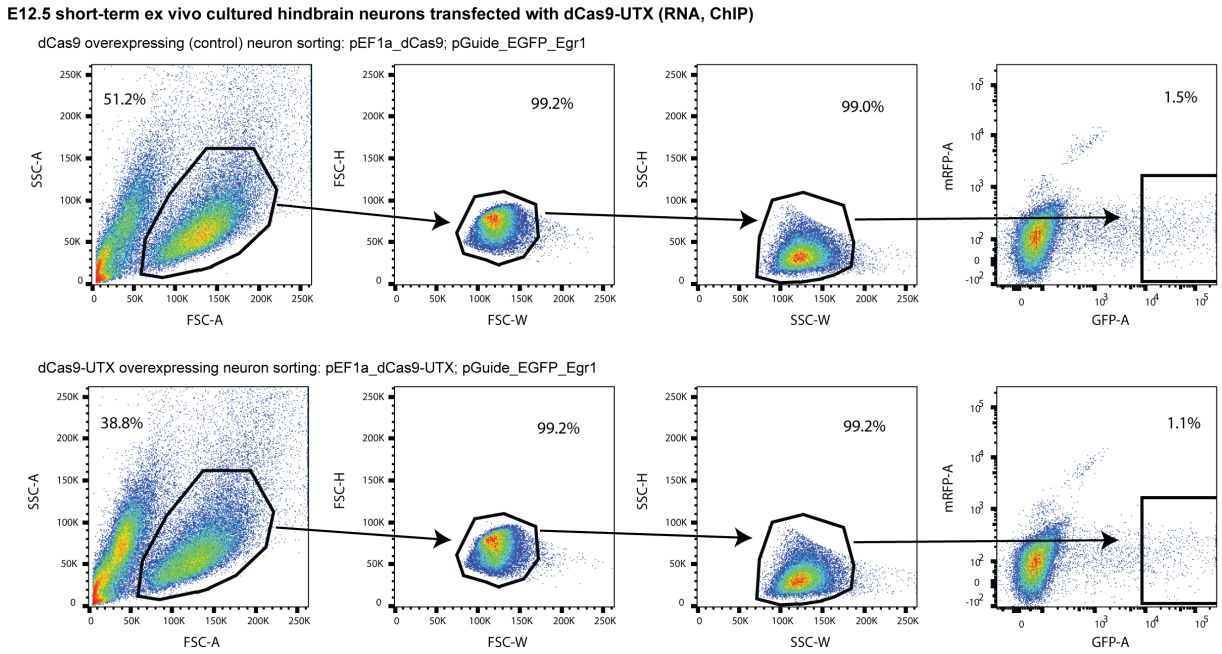


b ***Ezh2cKO*** **hindbrain cells (ATAC-seq, ChIP-seq)**



Supplementary Fig. 3 | FACS sequential gating strategy for E14.5 *Ezh2* conditionally knocked-out hindbrain cells.

a, Gating layouts to sort *Ezh2cKO* RFP-positive r3-derived hindbrain cells. Negative control, *Ezh2* heterozygous *Krox20::Cre;Ezh2*$^{flox/+}$*;R26*^{RFP} and *Ezh2* knocked-out *Krox20::Cre;Ezh2*$^{flox/flox}$*;R26*^{RFP} (*Ezh2cKO*$^{r3\text{-}RFP}$) mice are compared. Related to Fig. 5a-c, Extended Data Fig. 7c,d. **b**, Gating layouts to sort *Ezh2cKO* hindbrain cells. Negative control, *Ezh2* heterozygous *Hoxa2::Cre;Ezh2*$^{flox/+}$*;R26*^{RFP} and *Ezh2* knocked-out *Hoxa2::Cre;Ezh2*$^{flox/flox}$*;R26*^{RFP} (*Ezh2cKO*$^{HB\text{-}RFP}$) mice are compared. *Hoxa2::Cre* line, that labels from r2 to posterior hindbrain neurons, was utilized to collect relatively large number of hindbrain neurons to enable the molecular analysis of *Ezh2*-null neurons. Related to Fig. 5g, Extended Data Figs. 7e and 8i,j. Nomenclatures for the mouse lines are summarized in Supplementary Table 1.



Supplementary Fig. 4 | FACS sequential gating strategy for E12.5 short-term ex vivo cultured hindbrain neurons overexpressing dCas9-Utx.

Representative gating layouts of dCas9 or dCas9-UTX overexpressing E12.5 short-term cultured hindbrain neurons. pGuide_EGFP_Egr1 plasmid which overexpresses EGFP and *Egr1* targeting guide RNA was co-transfected with pEF1a_dCas9 or pEF1a_dCas9-UTX plasmids, and EGFP-positive neurons were sorted. Related to Fig. 5d, Extended Data Fig. 8a-e.

Supplementary References

1 Oury, F. *et al.* Hoxa2- and rhombomere-dependent development of the mouse facial somatosensory map. *Science* **313**, 1408-1413, doi:10.1126/science.1130042 (2006).

2 Bechara, A. *et al.* Hoxa2 Selects Barrelette Neuron Identity and Connectivity in the Mouse Somatosensory Brainstem. *Cell Rep* **13**, 783-797, doi:10.1016/j.celrep.2015.09.031 (2015).

3 Moreno-Juan, V. *et al.* Prenatal thalamic waves regulate cortical area size prior to sensory processing. *Nat Commun* **8**, 14172, doi:10.1038/ncomms14172 (2017).

4 Li, Y., Erzurumlu, R. S., Chen, C., Jhaveri, S. & Tonegawa, S. Whisker-related neuronal patterns fail to develop in the trigeminal brainstem nuclei of NMDAR1 knockout mice. *Cell* **76**, 427-437 (1994).

5 Erzurumlu, R. S., Murakami, Y. & Rijli, F. M. Mapping the face in the somatosensory brainstem. *Nat Rev Neurosci* **11**, 252-263, doi:10.1038/nrn2804 (2010).

6 Kitazawa, T. & Rijli, F. M. Barrelette map formation in the prenatal mouse brainstem. *Curr Opin Neurobiol* **53**, 210-219, doi:10.1016/j.conb.2018.09.008 (2018).

7 Lo, F. S. & Erzurumlu, R. S. Neonatal sensory nerve injury-induced synaptic plasticity in the trigeminal principal sensory nucleus. *Exp Neurol* **275 Pt 2**, 245-252, doi:10.1016/j.expneurol.2015.04.022 (2016).

8 Erzurumlu, R. S. & Gaspar, P. Development and critical period plasticity of the barrel cortex. *Eur J Neurosci* **35**, 1540-1553, doi:10.1111/j.1460-9568.2012.08075.x (2012).

9 Hrvatin, S. *et al.* Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat Neurosci* **21**, 120-129, doi:10.1038/s41593-017-0029-5 (2018).

10 Tyssowski, K. M. *et al.* Different Neuronal Activity Patterns Induce Different Gene Expression Programs. *Neuron* **98**, 530-546 e511, doi:10.1016/j.neuron.2018.04.001 (2018).

11 Valles, A. *et al.* Genomewide analysis of rat barrel cortex reveals time- and layer-specific mRNA expression changes related to experience-dependent plasticity. *J Neurosci* **31**, 6140-6158, doi:10.1523/JNEUROSCI.6514-10.2011 (2011).

12 Joo, J. Y., Schaukowitch, K., Farbiak, L., Kilaru, G. & Kim, T. K. Stimulus-specific combinatorial functionality of neuronal c-fos enhancers. *Nat Neurosci* **19**, 75-83, doi:10.1038/nn.4170 (2016).

13 Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187, doi:10.1038/nature09033 (2010).

14 Malik, A. N. *et al.* Genome-wide identification and characterization of functional neuronal activity-dependent enhancers. *Nat Neurosci* **17**, 1330-1339, doi:10.1038/nn.3808 (2014).

15 West, A. E. & Greenberg, M. E. Neuronal activity-regulated gene transcription in synapse development and cognitive function. *Cold Spring Harb Perspect Biol* **3**, doi:10.1101/cshperspect.a005744 (2011).

16 Maze, I. *et al.* Critical Role of Histone Turnover in Neuronal Transcription and Plasticity. *Neuron* **87**, 77-94, doi:10.1016/j.neuron.2015.06.014 (2015).

17 Zaborowska, J., Egloff, S. & Murphy, S. The pol II CTD: new twists in the tail. *Nat Struct Mol Biol* **23**, 771-777, doi:10.1038/nsmb.3285 (2016).

18 Greer, C. B. *et al.* Histone Deacetylases Positively Regulate Transcription through the Elongation Machinery. *Cell Rep* **13**, 1444-1455, doi:10.1016/j.celrep.2015.10.013 (2015).

19 Jang, M. K. *et al.* The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol Cell* **19**, 523-534, doi:10.1016/j.molcel.2005.06.027 (2005).

20 Stroud, H. *et al.* An Activity-Mediated Transition in Transcription in Early Postnatal Neurons. *Neuron* **107**, 874-890 e878, doi:10.1016/j.neuron.2020.06.008 (2020).

21 Schoeftner, S. *et al.* Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. *EMBO J* **25**, 3110-3122, doi:10.1038/sj.emboj.7601187 (2006).

22 Lavarone, E., Barbieri, C. M. & Pasini, D. Dissecting the role of H3K27 acetylation and methylation in PRC2 mediated control of cellular identity. *Nat Commun* **10**, 1679, doi:10.1038/s41467-019-09624-w (2019).

23 Chen, F. X., Smith, E. R. & Shilatifard, A. Born to run: control of transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* **19**, 464-478, doi:10.1038/s41580-018-0010-5 (2018).

24 Schaukowitch, K. *et al.* Enhancer RNA facilitates NELF release from immediate early genes. *Mol Cell* **56**, 29-42, doi:10.1016/j.molcel.2014.08.023 (2014).

25 Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nat Genet* **39**, 1507-1511, doi:10.1038/ng.2007.21 (2007).

26 Saha, R. N. *et al.* Rapid activity-induced transcription of Arc and other IEGs relies on poised RNA polymerase II. *Nat Neurosci* **14**, 848-856, doi:10.1038/nn.2839 (2011).

27 Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell* **171**, 34-57, doi:10.1016/j.cell.2017.08.002 (2017).

28 Simon, J. A. & Kingston, R. E. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat Rev Mol Cell Biol* **10**, 697-708, doi:10.1038/nrm2763 (2009).

29 Schlumm, F., Mauceri, D., Freitag, H. E. & Bading, H. Nuclear calcium signaling regulates nuclear export of a subset of class IIa histone deacetylases following synaptic activity. *J Biol Chem* **288**, 8074-8084, doi:10.1074/jbc.M112.432773 (2013).

30 Sagner, A. & Briscoe, J. Morphogen interpretation: concentration, time, competence, and signaling dynamics. *Wiley Interdiscip Rev Dev Biol* **6**, doi:10.1002/wdev.271 (2017).

31 Mayer, A., Landry, H. M. & Churchman, L. S. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Curr Opin Cell Biol* **46**, 72-80, doi:10.1016/j.ceb.2017.03.002 (2017).

32 Min, I. M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev* **25**, 742-754, doi:10.1101/gad.2005511 (2011).

33 Stroud, H. *et al.* Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. *Cell* **171**, 1151-1164 e1116, doi:10.1016/j.cell.2017.09.047 (2017).

34 Baker, S. A. *et al.* An AT-hook domain in MeCP2 determines the clinical course of Rett syndrome and related disorders. *Cell* **152**, 984-996, doi:10.1016/j.cell.2013.01.038 (2013).

35 Frasch, M., Chen, X. & Lufkin, T. Evolutionary-conserved enhancers direct region-specific expression of the murine Hoxa-1 and Hoxa-2 loci in both mice and Drosophila. *Development* **121**, 957-974 (1995).

36 Schmidl, C., Rendeiro, A. F., Sheffield, N. C. & Bock, C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods* **12**, 963-965, doi:10.1038/nmeth.3542 (2015).

37 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-181, doi:10.1038/nprot.2014.006 (2014).

38 Splinter, E., de Wit, E., van de Werken, H. J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* **58**, 221-230, doi:10.1016/j.ymeth.2012.04.009 (2012).

39 Nagano, T. *et al.* Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat Protoc* **10**, 1986-2003, doi:10.1038/nprot.2015.127 (2015).

40 Yost, K. E., Carter, A. C., Xu, J., Litzenburger, U. & Chang, H. Y. ATAC Primer Tool for targeted analysis of accessible chromatin. *Nat Methods* **15**, 304-305, doi:10.1038/nmeth.4663 (2018).

41 Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583-588, doi:10.1038/nature14136 (2015).

42 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).

43 Osakada, F. & Callaway, E. M. Design and generation of recombinant rabies virus vectors. *Nat Protoc* **8**, 1583-1601, doi:10.1038/nprot.2013.094 (2013).

44 Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M. B. QuasR: quantification and annotation of short reads in R. *Bioinformatics* **31**, 1130-1132, doi:10.1093/bioinformatics/btu781 (2015).

45 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

46 Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773, doi:10.1093/nar/gky955 (2019).

47 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).

48 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

49 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

50 Martin, M. Cutadapt removesadapter sequences from high-throughput sequencing reads. EMBnet.journal 17.1.

51 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

52 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

53 van de Werken, H. J. *et al.* 4C technology: protocols and data analysis. *Methods Enzymol* **513**, 89-112, doi:10.1016/B978-0-12-391938-0.00004-5 (2012).

54 Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841-1842, doi:10.1093/bioinformatics/btp328 (2009).

55 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).

56 Alasoo, K. Wiggleplotr: Make read coverage plots from bigwig files. (2019).

57 Soneson, C., Love, MI. & Robinson, MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved] F1000Research **4**, 1521 (2016).

58 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049, doi:10.1038/ncomms14049 (2017).

59 Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol* **20**, 63, doi:10.1186/s13059-019-1662-y (2019).

60 McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179-1186, doi:10.1093/bioinformatics/btw777 (2017).

61 Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122, doi:10.12688/f1000research.9501.2 (2016).

62 Maaten, L.J.P.V.D. & Hinton, GE. Visualizing High-Dimensional Data using t-SNE. Journal of Machine Learning Research **9**, 2579-2605 (2008).

63 R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, VA, 2020).

64 Gehlenborg, N. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Vi sualizing Intersecting Sets. R package version 1.4.0. https://CRAN.R-project.org/package =UpSetR (2019).

65 Venables, W.N. & Ripley, B.D. Modern Applied Statistics with S. (Fourth Edition. Springer, NY, 2002).

66 Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184-1191, doi:10.1038/nprot.2009.97 (2009).

67 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

68 Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* **8**, 289-317 (2016).

69 Minoux, M. *et al.* Gene bivalency at Polycomb domains regulates cranial neural crest positional identity. *Science* **355**, doi:10.1126/science.aal2913 (2017).

70 O'Connell, J. & Hojsgaard, S. Hidden Semi Markov Models for Multiple Observation Sequences: The mhsmm Package for R. Journal of Statistical Software **39**, 4 (2011).

71 Adler, D. & S. Kelly, S. K. vioplot: violin plot. R package version 0.3.4 https://github.com/TomKellyGenetics/vioplot (2019).

72 Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. R package version 2.56.0. (2020).

73 Tan, Ge. JASPAR2018: Data package for JASPAR 2018. R package version 1.1.1. http://jaspar.genereg.net/ (2017).

74 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

A.2 COMPARING REGRESSION METHODS USING SIMULATED DATA

This section of the appendix details how the simulated data sets in section 3.4.1 were generated and how the different regression methods were run and compared. It includes all the R code that was used to do this analysis to produce figure 3.4, as well as some additional figures.

TPR vs FPR in Selected Regression Methods

Dania Machlab

August 21, 2021

Contents

We use the DNA methylation data set from the `monaLisa` package to create a synthetic predictor matrix with a similar correlation structure. We run lasso stability selection, randomized lasso stability selection, a lasso regression and an elastic net regression, and compare the different approaches. Based on the publication from Meinshausen and Bühlmann (2010), we expect stability selection methods to outperform regular regression methods, and the randomized lasso stability selection to do better in noisy data sets.

We start by loading the packages we need.

```
setwd("/tungstenfs/groups/gbioinfo/machdani/thesis_plots/")

suppressPackageStartupMessages({
    library(monaLisa, lib.loc = "/tungstenfs/groups/gbioinfo/machdani/thesis_plots/libs/")
    library(GenomicRanges)
    library(SummarizedExperiment)
    library(JASPAR2018)
    library(TFBSTools)
    library(BSgenome.Mmusculus.UCSC.mm10)
    library(Biostrings)
    library(ggplot2)
    library(tidyverse)
    library(patchwork)
    library(ComplexHeatmap)
    library(circlize)
    library(Matrix)
    library(ROCR)
    library(limma)
    library(glmnet)
})
```

Create Synthetic Predictor Matrix

We use the DNA methylation data set present in `monaLisa` and first create a predictor matrix based on this real data set. The predictor matrix will consist of the sequences as rows and the motifs as columns. Each entry will be the number of binding sites a motif has at a specific sequence, which we will get by scanning the position weight matrix (PWM) of a motif across the sequences. We can do this using the `findMotifHits` function in `monaLisa`, which is a faster implementation of the `matchPWM` function from the `Biostrings` package. We will use the `JASPAR2018` database to get all the vertebrate PWMs.

We create the transcription factor binding site (TFBS) matrix based on the real data set. We will call this matrix `XObs`.

```
# data set
gr_path <- system.file("extdata", "LMRsESNPmerged.gr.rds", package = "monaLisa")
gr <- readRDS(gr_path)

# load PWMs
pfms <- getMatrixSet(JASPAR2018, list(matrixtype = "PFM", tax_group = "vertebrates"))
pwms <- toPWM(pfms)

# scan for motif hits
peakSeq <- getSeq(BSgenome.Mmusculus.UCSC.mm10, gr)
hits <- findMotifHits(query = pwms, subject = peakSeq, min.score = 10.0,
                      BPPARAM = BiocParallel::MulticoreParam(20))

# create TFBS matrix
TFBSmatrix <- unclass(table(factor(seqnames(hits), levels = seqlevels(hits)),
                            factor(hits$pwmname, levels = name(pwms))))
TFBSmatrix[1:6, 1:6]
```

```
##
##      Arnt Ahr::Arnt Ddit3::Cebpa NFIL3 Mecom FOXF2
##   s1    0         0            1     0     0     0
##   s2    0         0            0     0     2     0
##   s3    0         0            0     0     2     1
##   s4    2         0            0     0     0     0
##   s5    0         0            0     0     1     2
##   s6    0         1            1     0     0     0
```

```
# remove motifs with no hits at all
zero_TF <- colSums(TFBSmatrix) == 0
sum(zero_TF)
```

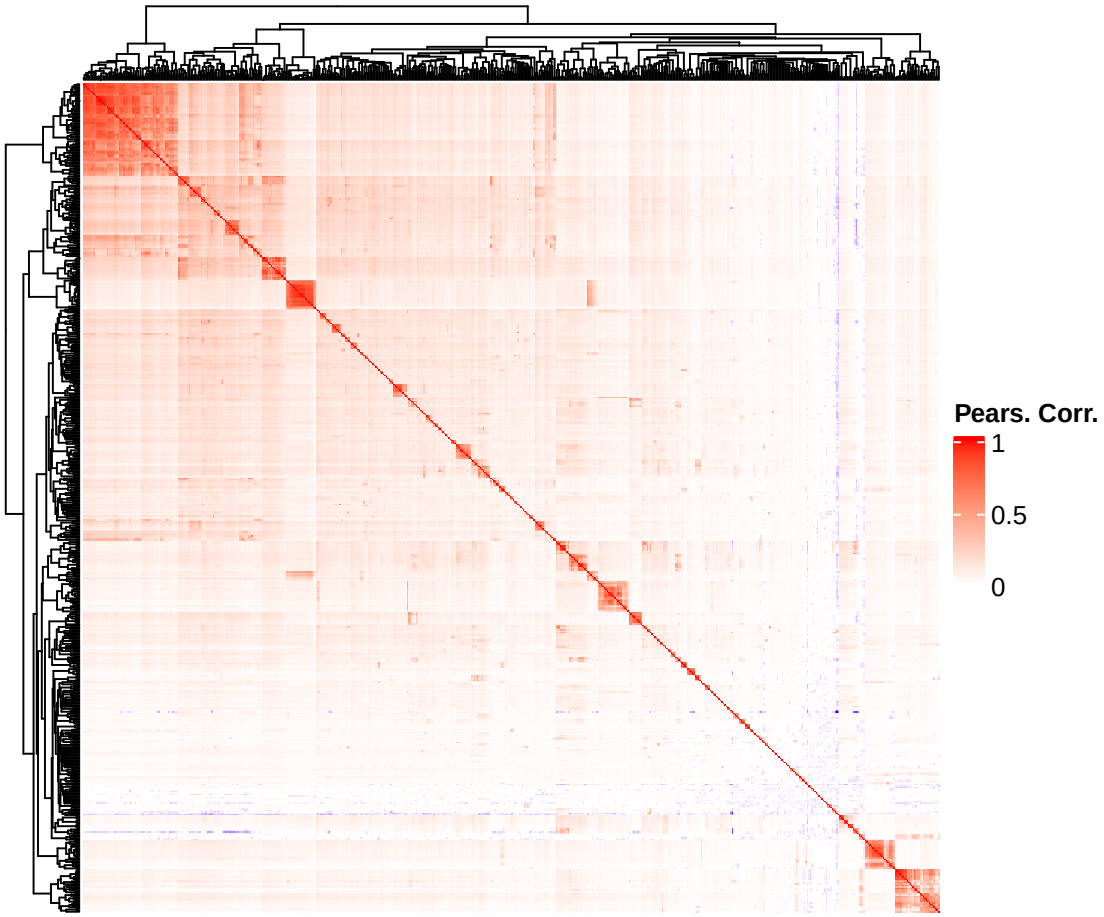
```
## [1] 5
```

```
XObs <- TFBSmatrix[, !zero_TF]

dim(XObs)
```

```
## [1] 45414   574
```

We visualize the correlation structure in `XObs`.

```
cor_XObs <- cor(XObs, method = "pearson")
col_fun = colorRamp2(c(range(cor_XObs)[1], 0, range(cor_XObs)[2]),
                     c("blue", "white", "red"))
Heatmap(cor_XObs, col = col_fun, show_row_names = FALSE, show_column_names = FALSE,
        name = "Pears. Corr.")
```



We create the synthetic predictor matrix. Each predictor in this matrix will be generated using a negative binomial distribution with the mean and variance parameters present in `XObs`.

```
# dim options for X (the synthetic predictor matrix)
n <- length(gr) # number of observations
p <- 500 # number of predictors or TFs

# randomly subsample p predictors from XObs
set.seed(123)
rand <- sample(1:ncol(XObs), p, replace = FALSE)
XObs <- XObs[, rand]
dim(XObs)
```

```
## [1] 45414   500
```

```
# we use the mean and variance each predictor has in XObs to create the synthetic X
Xmean <- colMeans(XObs)
Xvar  <- apply(XObs, 2, var)
```

```r
X <- do.call(cbind, lapply(1:p, function(i) {
  mu <- Xmean[i]
  size <- (Xmean[i])^2/abs(Xvar[i]-Xmean[i])
  rnbinom(n=n, size=size, mu=mu)
}))
dim(X)
```

```
## [1] 45414   500
```

```r
# plot correlation structure of X
cor <- cor(X)
col_fun = colorRamp2(c(range(cor)[1], 0, range(cor)[2]),
                     c("blue", "white", "red"))
H1 <- Heatmap(cor, name="Pears. Cor. on X", show_column_names = FALSE, col = col_fun,
              show_row_names = FALSE, column_title = "X before introducing correlation")

# introduce correlation structure from XObs using Cholesky decomposition.
cor_XObs <- cor(XObs, method = "pearson")
cor_XObs_pd <- nearPD(cor_XObs, corr=TRUE)$mat
cholesky <- chol(cor_XObs_pd)
X <- X %*% cholesky

# set values to integers and replace negative values with 0
X <- ceiling(X)
X <- as.matrix(X)
X[X < 0] <- 0

# ranges
range(X)
```

```
## [1]  0 79
```

```r
range(XObs)
```

```
## [1]   0 147
```

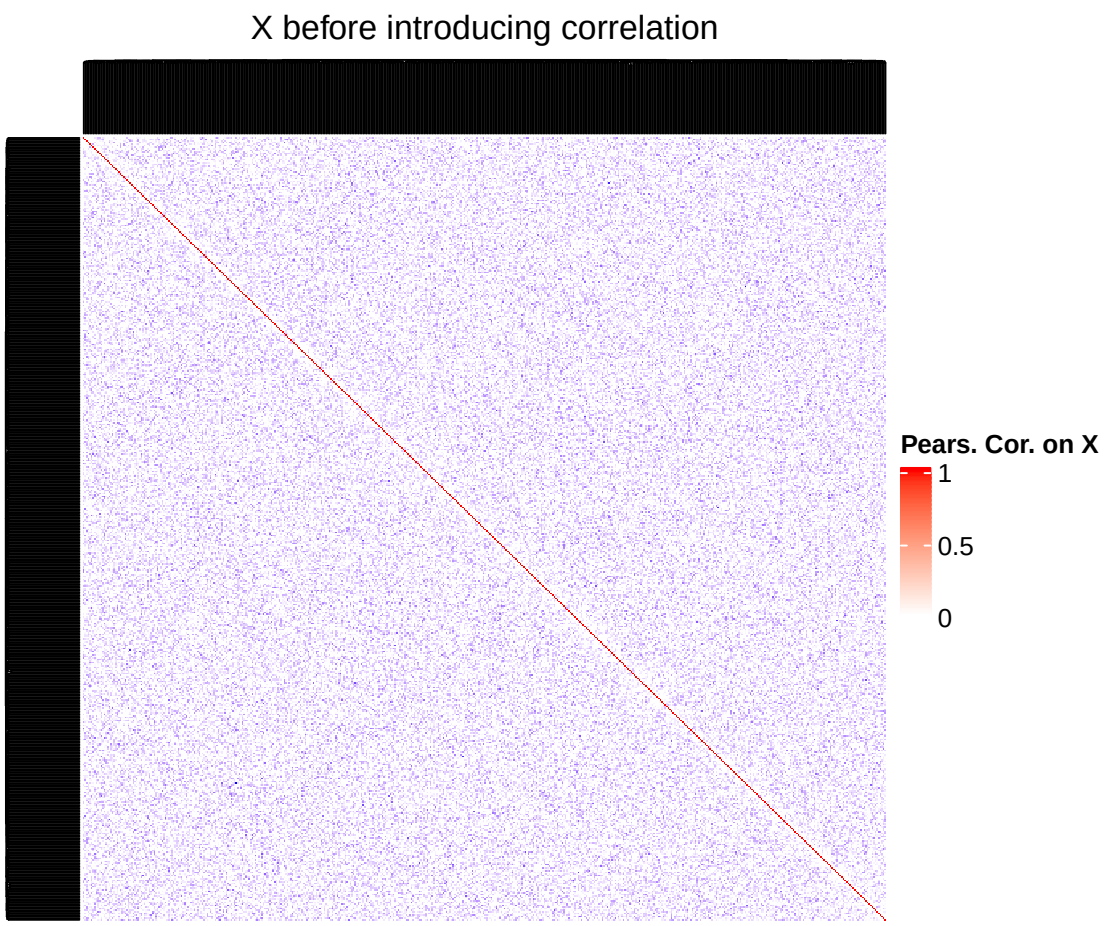
```r
# correlation structure in X
cor <- cor(X)
summary(as.vector(cor))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.44996  0.01646  0.03755  0.06094  0.07290  1.00000
```

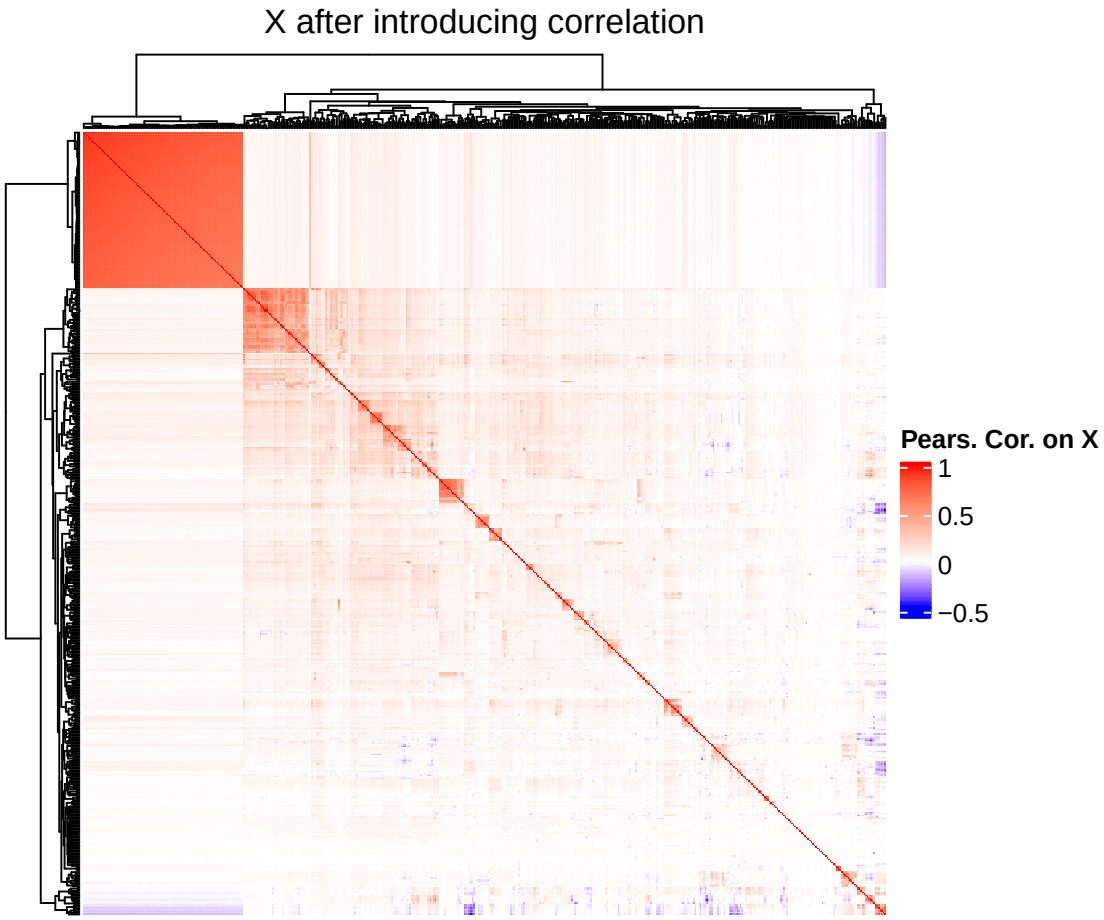
```r
# introduce a bit more correlation between 100 predictors
set.seed(123)
rand <- sample(1:ncol(X), 100, replace = FALSE)
# ... we take the first vector and vary it for the rest by adding a binomial distr
binom_prob <- seq(from = 0.01, to = 0.06, length.out = 99)
sel <- X[, rand[1]]
new_X_rand <- do.call(cbind, lapply(binom_prob, function(x){
    sel + rbinom(n = n, size = 1, prob = x)
}))
new_X_rand <- cbind(sel, new_X_rand)
X[, rand] <- new_X_rand

# plot correlation structure of X
cor <- cor(X)
```

```
col_fun = colorRamp2(c(range(cor)[1], 0, range(cor)[2]),
                     c("blue", "white", "red"))
H2 <- Heatmap(cor, name="Pears. Cor. on X", show_column_names = FALSE, col = col_fun,
              show_row_names = FALSE, column_title = "X after introducing correlation")
```

```
H1
```



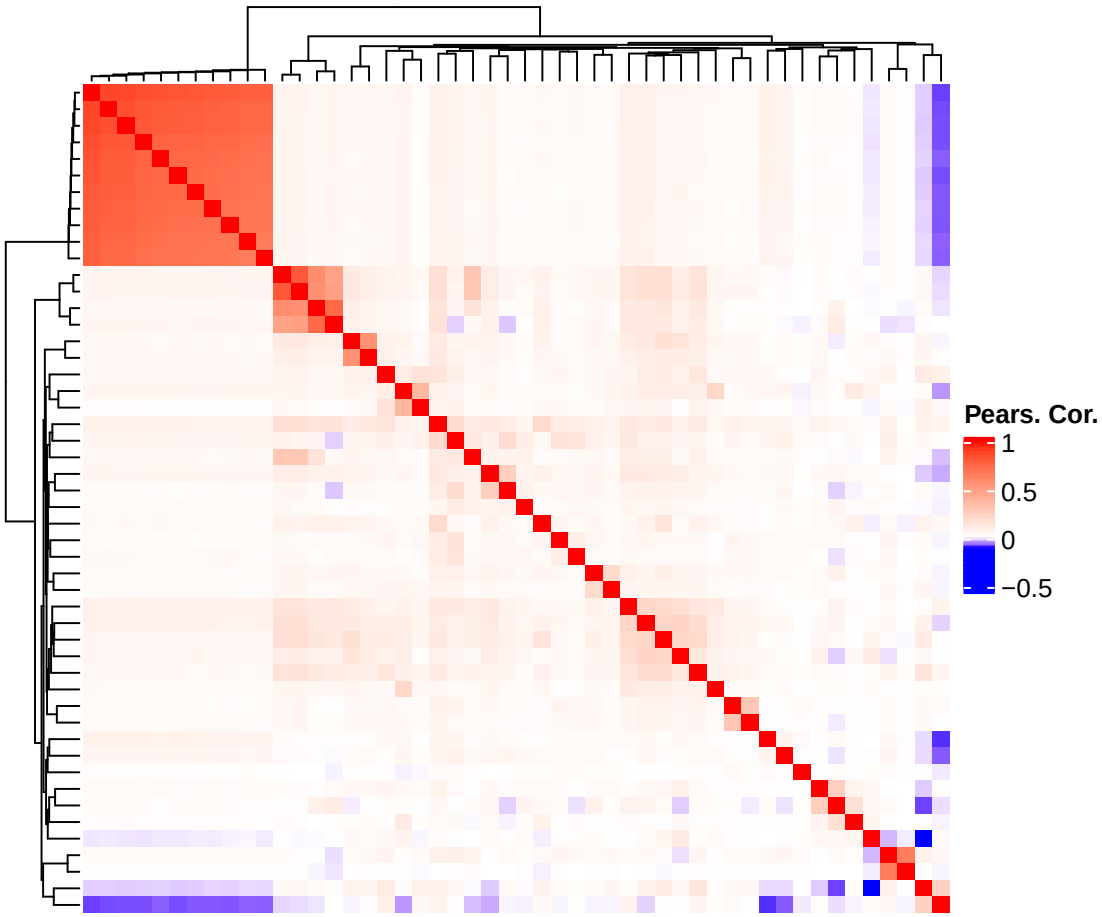

```
H2
```



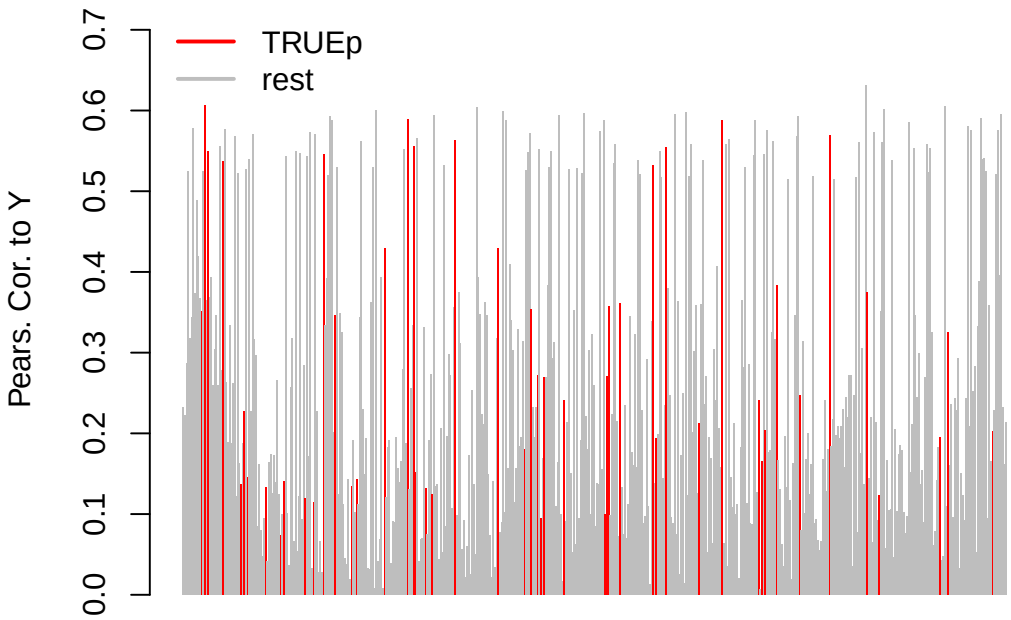
Choose Set of TRUE Predictors and Create Synthetic Response Vectors

We randomly select 50 predictors to be our TRUE predictors. The response vector is then the sum of these selected variables.

```
# choose our signal
set.seed(12345)
TRUEp <- sample(x = 1:ncol(X), size = 50, replace = FALSE)

# response vector
Y <- rowSums(X[, TRUEp])

# plot correlation structure of TRUEp
cor <- cor(X[, TRUEp])
col_fun = colorRamp2(c(range(cor)[1], 0, range(cor)[2]),
                     c("blue", "white", "red"))
Heatmap(cor, name="Pears. Cor.", show_column_names = FALSE, col = col_fun,
        show_row_names = FALSE, column_title = "")
```




```
# barplot correlation of predictors to the response
corr_to_y <- cor(Y, X)
cols <- rep("grey", ncol(X))
cols[TRUEp] <- "red"
barplot(corr_to_y[1, ], col=cols, border=NA, ylab="Pears. Cor. to Y",
        ylim = c(0, (range(corr_to_y)[2] + 0.1)))
legend("topleft", bty = "n", legend = c("TRUEp", "rest"), col = c("red", "grey"),
       lwd = 2)
```



create noisy versions of the response vector We add varying levels of noise to the response vector, in order to be able to evaluate which regression methods do a reasonable job selecting the correct predictors in such settings. If we do a linear regression on the response vector, without any added noise, the `TRUE` predictors all get coefficients of 1 estimated, whereas the rest of the predictors have coefficient values close to zero. We will thus take the signal to noise ratio to be the effect size in the un-noised setting, 1, divided by the standard deviation.

```
# linear fit
fit <- lm(Y ~ X)
range(coef(fit)[-1][TRUEp])
```

```
## [1] 1 1
```

```
range(coef(fit)[-1][-TRUEp])
```

```
## [1] -1.671253e-13  2.345629e-14
```

```
# create response vectors with varying levels of SNR
Y_snr0.5 <- Y + rnorm(n = length(Y), mean = 0, sd = 2)
Y_snr0.1 <- Y + rnorm(n = length(Y), mean = 0, sd = 10)
Y_snr0.05 <- Y + rnorm(n = length(Y), mean = 0, sd = 20)
Y_snr0.01 <- Y + rnorm(n = length(Y), mean = 0, sd = 100)
Y_snr0.001 <- Y + rnorm(n = length(Y), mean = 0, sd = 1000)
```

```
Y_list <- list(noNoise = Y, snr0.5 = Y_snr0.5, snr0.1 = Y_snr0.1, snr0.05 = Y_snr0.05,
               snr0.01 = Y_snr0.01, snr0.001 = Y_snr0.001)
```

Randomized Lasso Stability Selection

We run the randomized lasso stability selection with a few choices of the `weakness` and `PFER` parameters. We do this a few times since the process is stochastic, to also get an idea of reproducibility.

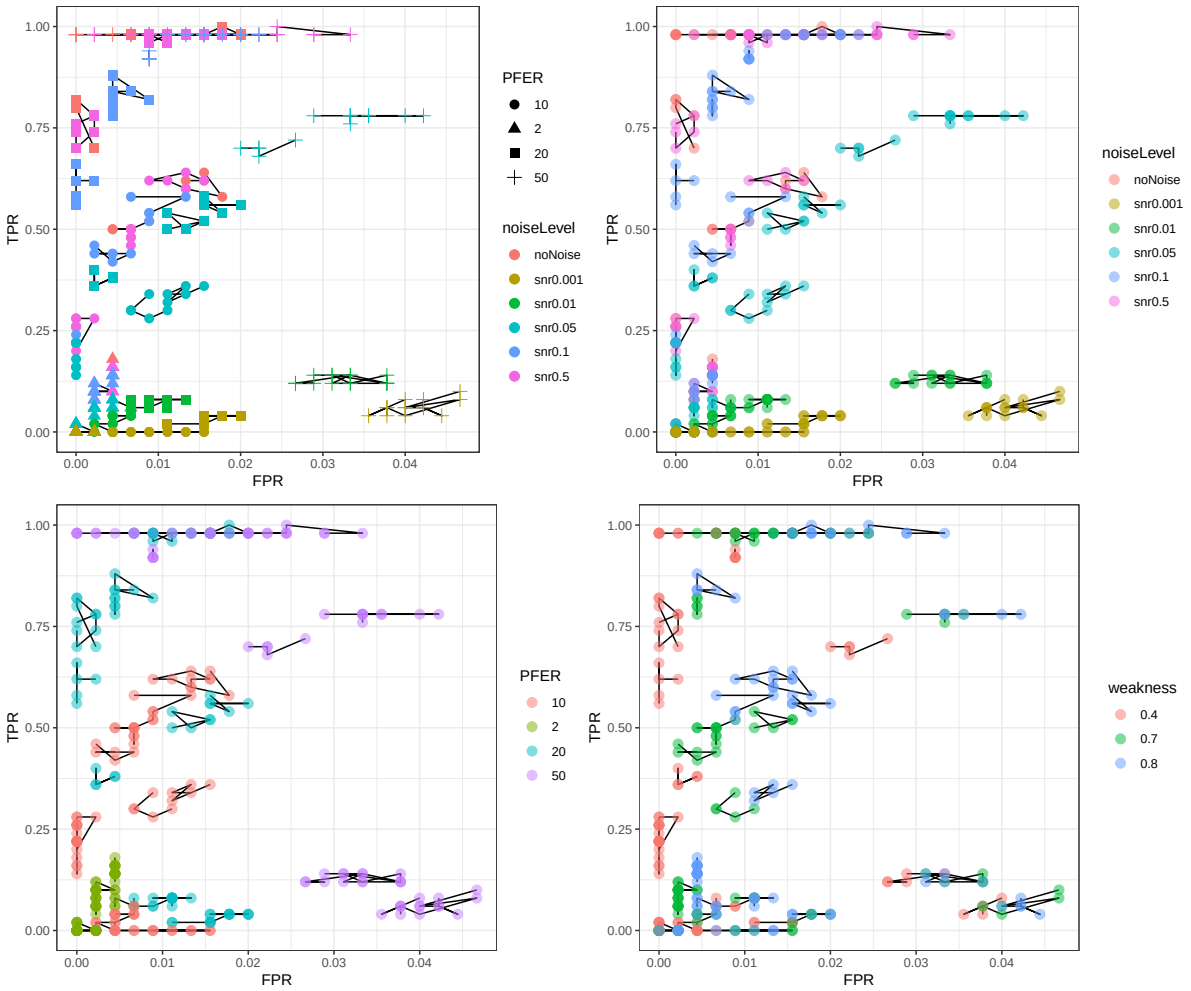
```
weakness <- c(0.4, 0.7, 0.8)
pfer <- c(2, 10, 20, 50)
prob_thresh <- c(0.7)
cores <- 20

randLassoStabSel_res <- list()
ind <- 1
# ... loop over datasets
for(i in 1:length(Y_list)) {
    response <- Y_list[[i]]
    nm_i <- names(Y_list)[i]
    # ... loop over weakness parameter
    for(j in 1:length(weakness)){
        w <- weakness[j]
        nm_j <- as.character(weakness[j])
        # ... loop over PFER parameter
        for(k in 1:length(pfer)){
            pf <- pfer[k]
            nm_k <- as.character(pfer[k])
            # ... loop over probability threshold
            for(l in 1:length(prob_thresh)){
                prob <- prob_thresh[l]
                nm_l <- as.character(prob_thresh[l])
                # ... run 5 times
                for(m in 1:5){
                    ss <- randLassoStabSel(x = X, y = response, weakness = w,
                                           cutoff = prob, PFER = pf, mc.cores = cores)
                    randLassoStabSel_res[[ind]] <- ss
                    names(randLassoStabSel_res)[ind] <- paste0(nm_i, "_W", nm_j, "_PFER",
                                                               nm_k, "_probCutoff",
                                                               nm_l, "_", m)
                    ind <- ind + 1
                }
            }
        }
    }
}
```

```
FALSEp <- 1:ncol(X)
FALSEp <- FALSEp[! FALSEp %in% TRUEp]

# true positive rate (TPR)
tpr <- sapply(randLassoStabSel_res, function(se){
    x <- colData(se)$selected
```

```
    TP <- sum(which(x) %in% TRUEp)
    FN <- sum(which(!x) %in% TRUEp)
    TP / (TP + FN)
})

# false positive rate (FPR)
fpr <- sapply(randLassoStabSel_res, function(se){
    x <- colData(se)$selected
    FP <- sum(which(x) %in% FALSEp)
    TN <- sum(which(!x) %in% FALSEp)
    FP / (FP + TN)
})
```

```
# prepare dataframe to plot
df_randLassoStabSel <-
    as.data.frame(do.call(cbind, list(TPR = tpr, FPR = fpr, uniqueNm = names(fpr),
                                      noiseLevel = strsplit2(names(fpr), "_")[, 1],
                                      weakness = strsplit2(names(fpr), "_|W")[, 3],
                                      PFER = strsplit2(names(fpr), "_|PFER")[, 4],
                                      probCutoff = strsplit2(names(fpr), "_|off")[, 5],
                                      run = gsub(".{2}$", "", names(fpr)))
    ))
df_randLassoStabSel$TPR <- as.numeric(df_randLassoStabSel$TPR)
df_randLassoStabSel$FPR <- as.numeric(df_randLassoStabSel$FPR)
```

```
transp <- 0.5
p <- ggplot(df_randLassoStabSel) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR, color = noiseLevel, shape = PFER),
               size = 3) +
    theme_bw()
p1 <- ggplot(df_randLassoStabSel) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR, color = noiseLevel), size = 3, alpha = transp) +
    theme_bw()
p2 <- ggplot(df_randLassoStabSel) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR, color = PFER), size = 3, alpha = transp) +
    theme_bw()
p3 <- ggplot(df_randLassoStabSel) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR, color = weakness), size = 3, alpha = transp) +
    theme_bw()

(p + p1) / (p2 + p3)
```



For all, FPR control is tight, below 5%. Randomized lasso stability selection is overall very careful not choose wrong predictors, even at the cost of not choosing anything at all, as TPR can be low even in the data sets where no noise has been added to the response. As tolerance for error increases, by increasing the PFER, the FPR is still well controlled, and TPR increases. For the very noisy data sets like the SNR 0.001 data set, it is difficult to improve TPR by much. Varying the weakness parameter seems to control more for the FPR, with lower weakness values showing a lower FPR. The points are connected by group, our version of "replicates" where the method was run again with the same parameters, to see how much the results can vary. Points belonging to the same group are fairly close to each other in the plots.

Lasso Stability Selection

We run lasso stability selection also 5 times per chosen group of parameters.

```
weakness <- 1
pfer <- c(2, 10, 20, 50)
prob_thresh <- c(0.7)
cores <- 20
```

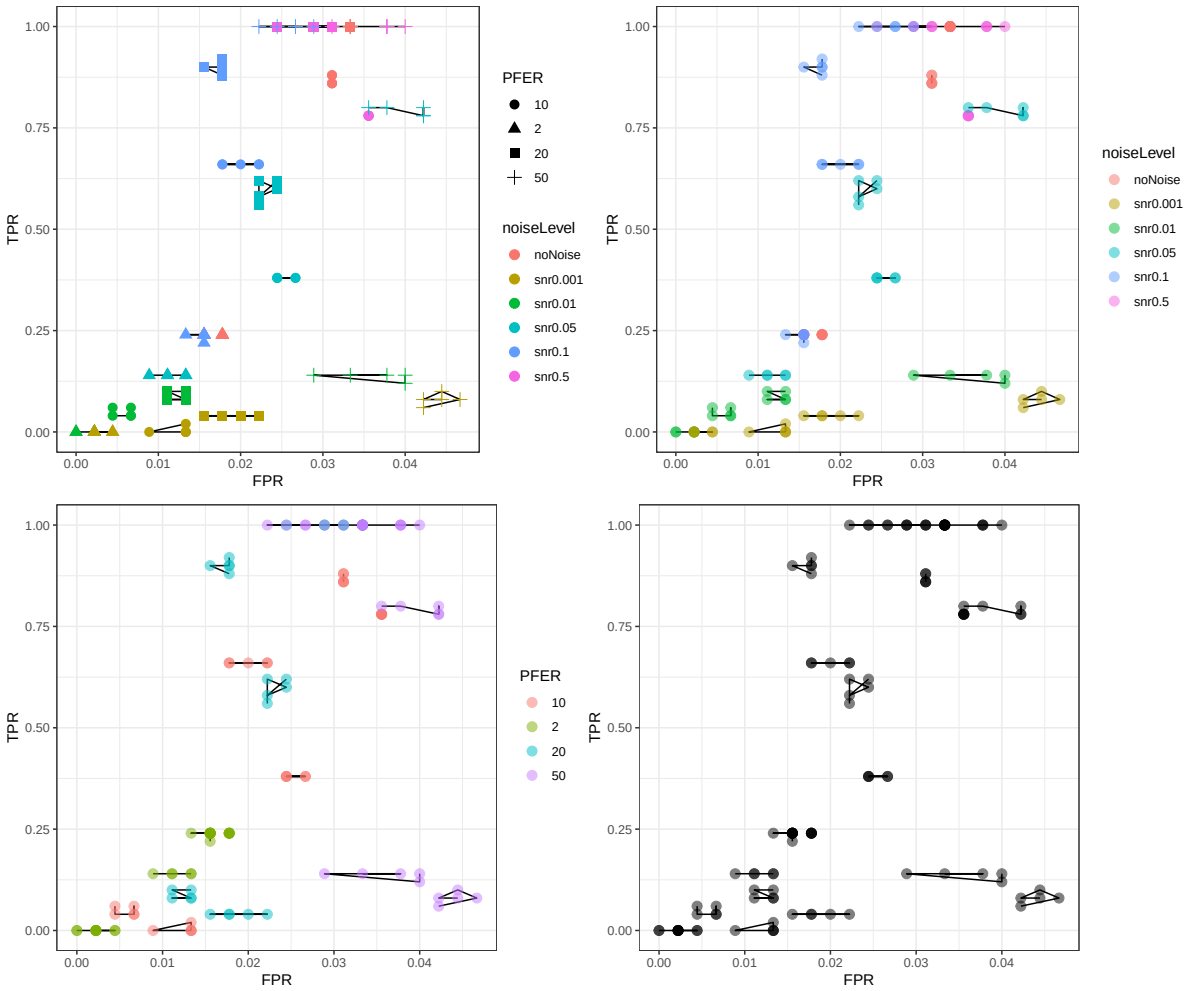
```
lassoStabSel_res <- list()
ind <- 1
# ... loop over datasets
for(i in 1:length(Y_list)) {
    response <- Y_list[[i]]
    nm_i <- names(Y_list)[i]
    # ... loop over PFER parameter
    for(k in 1:length(pfer)){
        pf <- pfer[k]
        nm_k <- as.character(pfer[k])
        # ... loop over probability threshold
        for(l in 1:length(prob_thresh)){
            prob <- prob_thresh[l]
            nm_l <- as.character(prob_thresh[l])
            # ... run 5 times
            for(m in 1:5){
                ss <- randLassoStabSel(x = X, y = response, weakness = weakness,
                                       cutoff = prob, PFER = pf, mc.cores = cores)
                lassoStabSel_res[[ind]] <- ss
                names(lassoStabSel_res)[ind] <- paste0(nm_i, "_PFER",
                                                       nm_k, "_probCutoff",
                                                       nm_l, "_", m)
                ind <- ind + 1
            }
        }
    }
}
```

```
FALSEp <- 1:ncol(X)
FALSEp <- FALSEp[! FALSEp %in% TRUEp]

# true positive rate (TPR)
tpr <- sapply(lassoStabSel_res, function(se){
    x <- colData(se)$selected
    TP <- sum(which(x) %in% TRUEp)
    FN <- sum(which(!x) %in% TRUEp)
    TP / (TP + FN)
})

# false positive rate (FPR)
fpr <- sapply(lassoStabSel_res, function(se){
    x <- colData(se)$selected
    FP <- sum(which(x) %in% FALSEp)
    TN <- sum(which(!x) %in% FALSEp)
    FP / (FP + TN)
})
```

```
# prepare dataframe to plot
df_lassoStabSel <-
    as.data.frame(do.call(cbind, list(TPR = tpr, FPR = fpr, uniqueNm = names(fpr),
                                      noiseLevel = strsplit2(names(fpr), "_")[, 1],
                                      PFER = strsplit2(names(fpr), "_|PFER")[, 3],
                                      probCutoff = strsplit2(names(fpr), "_|off")[, 4],
                                      run = gsub(".{2}$", "", names(fpr)))
```

```
    ))
df_lassoStabSel$TPR <- as.numeric(df_lassoStabSel$TPR)
df_lassoStabSel$FPR <- as.numeric(df_lassoStabSel$FPR)
```

```
transp <- 0.5
p <- ggplot(df_lassoStabSel) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR, color = noiseLevel, shape = PFER),
               size = 3) +
    theme_bw()
p1 <- ggplot(df_lassoStabSel) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR, color = noiseLevel), size = 3, alpha = transp) +
    theme_bw()
p2 <- ggplot(df_lassoStabSel) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR, color = PFER), size = 3, alpha = transp) +
    theme_bw()
p3 <- ggplot(df_lassoStabSel) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR), size = 3, alpha = transp) +
    theme_bw()

(p + p1) / (p2 + p3)
```



Lasso with Cross Validation

We run a regular lasso regression, setting the penalty term lambda to lambda 1se using cross validation. We do this five times.

```
require(doMC, lib.loc = "/tungstenfs/groups/gbioinfo/machdani/thesis_plots/libs")
```

```
## Loading required package: doMC
```

```
## Loading required package: foreach
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
## Loading required package: iterators
```

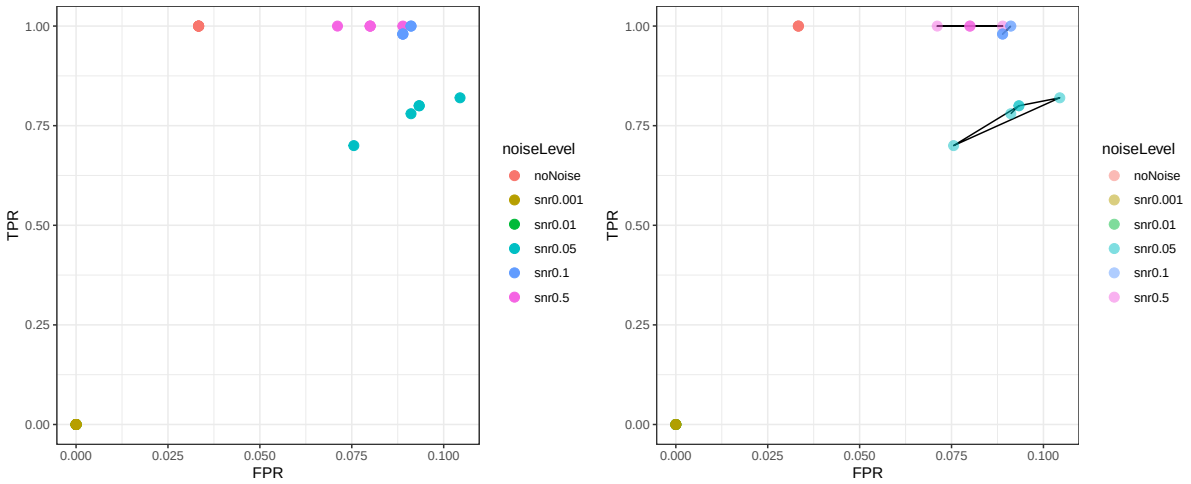
```r
registerDoMC(cores = 10)

cvLasso_res <- list()
ind <- 1
# ... loop over datasets
for(i in 1:length(Y_list)) {
    response <- Y_list[[i]]
    nm_i <- names(Y_list)[i]
    # ... run 5 times
    for(m in 1:5){
        # ... cross validation
        cv <- cv.glmnet(x = X, y = response, parallel = TRUE)
        cvLasso_res[[ind]] <- cv
        names(cvLasso_res)[ind] <- paste0(nm_i, "_", m)
        ind <- ind + 1
    }
}
```

```r
FALSEp <- 1:ncol(X)
FALSEp <- FALSEp[! FALSEp %in% TRUEp]

# true positive rate (TPR)
tpr <- sapply(cvLasso_res, function(fit){
    ind_lambda <- which(fit$lambda == fit$lambda.1se)
    beta <- fit$glmnet.fit$beta[, ind_lambda]
    x <- beta > 0
    TP <- sum(which(x) %in% TRUEp)
    FN <- sum(which(!x) %in% TRUEp)
    TP / (TP + FN)
})

# false positive rate (FPR)
fpr <- sapply(cvLasso_res, function(fit){
    ind_lambda <- which(fit$lambda == fit$lambda.1se)
    beta <- fit$glmnet.fit$beta[, ind_lambda]
    x <- beta > 0
    FP <- sum(which(x) %in% FALSEp)
    TN <- sum(which(!x) %in% FALSEp)
    FP / (FP + TN)
})
```

```r
# prepare dataframe to plot
df_cvLasso <-
    as.data.frame(do.call(cbind, list(TPR = tpr, FPR = fpr, uniqueNm = names(fpr),
                                      noiseLevel = strsplit2(names(fpr), "_")[, 1],
                                      run = gsub(".{2}$", "", names(fpr)))
    ))
df_cvLasso$TPR <- as.numeric(df_cvLasso$TPR)
df_cvLasso$FPR <- as.numeric(df_cvLasso$FPR)
```

```r
transp <- 0.5
p <- ggplot(df_cvLasso) +
    geom_point(aes(x = FPR, y = TPR, color = noiseLevel),
               size = 3) +
```

```
    theme_bw()
p1 <- ggplot(df_cvLasso) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR, color = noiseLevel), size = 3, alpha = transp) +
    theme_bw()

(p + p1)
```



Elastic Net with Cross Validation

We run an elastic net regression, setting the penalty term lambda to lambda 1se using cross validation. We do this five times using a few alpha values.

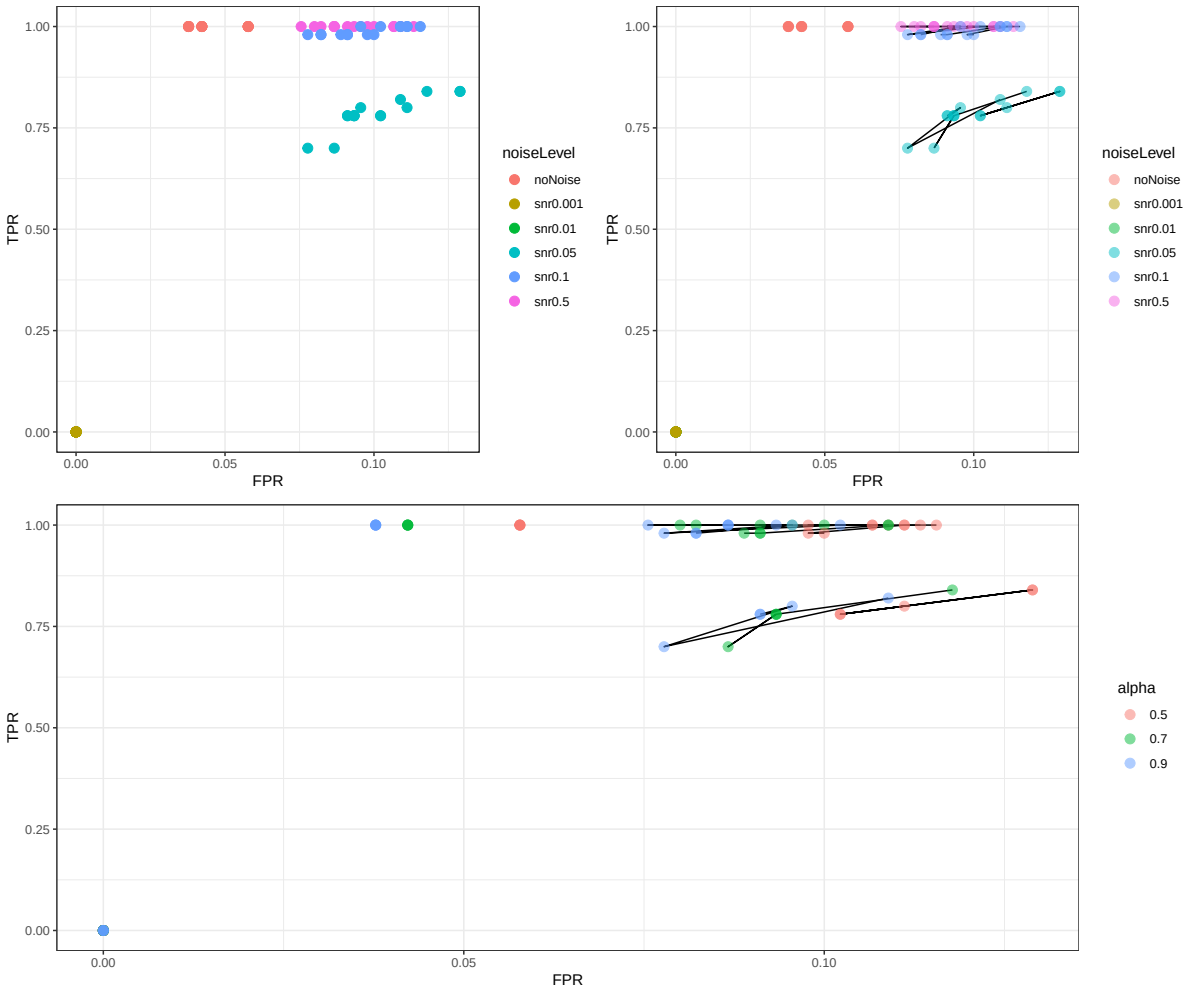
```
require(doMC, lib.loc = "/tungstenfs/groups/gbioinfo/machdani/thesis_plots/libs")
registerDoMC(cores = 10)
alpha <- c(0.5, 0.7, 0.9)

cvElasticnet_res <- list()
ind <- 1
# ... loop over datasets
for(i in 1:length(Y_list)){
    response <- Y_list[[i]]
    nm_i <- names(Y_list)[i]
    # ... loop over alpha
    for(j in 1:length(alpha)){
        al <- alpha[j]
        nm_j <- as.character(alpha[j])
        # ... run 5 times
        for(m in 1:5){
            # ... cross validation
            cv <- cv.glmnet(x = X, y = response, parallel = TRUE, alpha = al)
            cvElasticnet_res[[ind]] <- cv
            names(cvElasticnet_res)[ind] <- paste0(nm_i, "_alpha", nm_j, "_", m)
            ind <- ind + 1
```

```
        }
    }
}
```

```
FALSEp <- 1:ncol(X)
FALSEp <- FALSEp[! FALSEp %in% TRUEp]

# true positive rate (TPR)
tpr <- sapply(cvElasticnet_res, function(fit){
    ind_lambda <- which(fit$lambda == fit$lambda.1se)
    beta <- fit$glmnet.fit$beta[, ind_lambda]
    x <- beta > 0
    TP <- sum(which(x) %in% TRUEp)
    FN <- sum(which(!x) %in% TRUEp)
    TP / (TP + FN)
})

# false positive rate (FPR)
fpr <- sapply(cvElasticnet_res, function(fit){
    ind_lambda <- which(fit$lambda == fit$lambda.1se)
    beta <- fit$glmnet.fit$beta[, ind_lambda]
    x <- beta > 0
    FP <- sum(which(x) %in% FALSEp)
    TN <- sum(which(!x) %in% FALSEp)
    FP / (FP + TN)
})
```

```
# prepare dataframe to plot
df_cvElasticnet <-
    as.data.frame(do.call(cbind, list(TPR = tpr, FPR = fpr, uniqueNm = names(fpr),
                                      noiseLevel = strsplit2(names(fpr), "_")[, 1],
                                      alpha = strsplit2(names(fpr), "_|alpha")[, 3],
                                      run = gsub(".{2}$", "", names(fpr)))
    ))
df_cvElasticnet$TPR <- as.numeric(df_cvElasticnet$TPR)
df_cvElasticnet$FPR <- as.numeric(df_cvElasticnet$FPR)
```

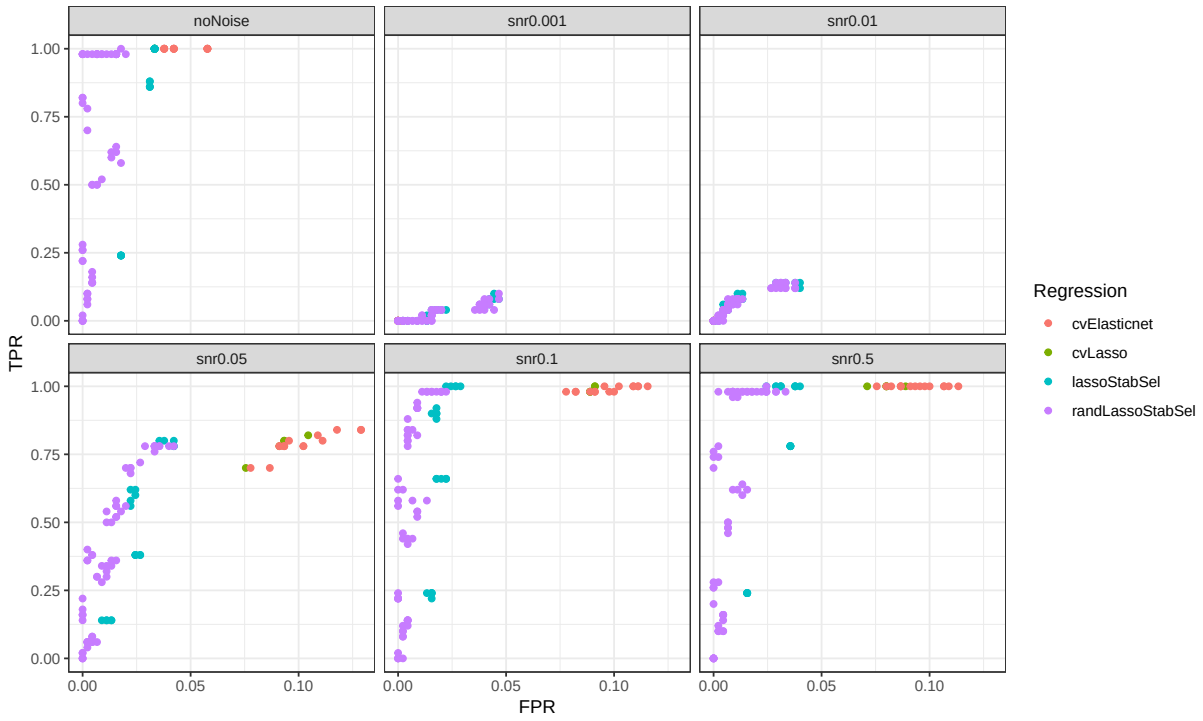
```
transp <- 0.5
p <- ggplot(df_cvElasticnet) +
    geom_point(aes(x = FPR, y = TPR, color = noiseLevel),
               size = 3) +
    theme_bw()
p1 <- ggplot(df_cvElasticnet) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR, color = noiseLevel), size = 3, alpha = transp) +
    theme_bw()
p2 <- ggplot(df_cvElasticnet) +
    geom_path(aes(x = FPR, y = TPR, group = run)) +
    geom_point(aes(x = FPR, y = TPR, color = alpha), size = 3, alpha = transp) +
    theme_bw()

(p + p1) / p2
```



Combine All

```
df_list <- list(cvLasso = df_cvLasso,
                cvElasticnet = df_cvElasticnet,
                lassoStabSel = df_lassoStabSel,
                randLassoStabSel = df_randLassoStabSel)
```

```
tbl <- df_list %>%
    bind_rows(.id = "Regression")

ggplot(tbl) +
    geom_point(aes(x = FPR, y = TPR, color = Regression)) +
    facet_wrap(~ noiseLevel) +
    theme_bw()
```

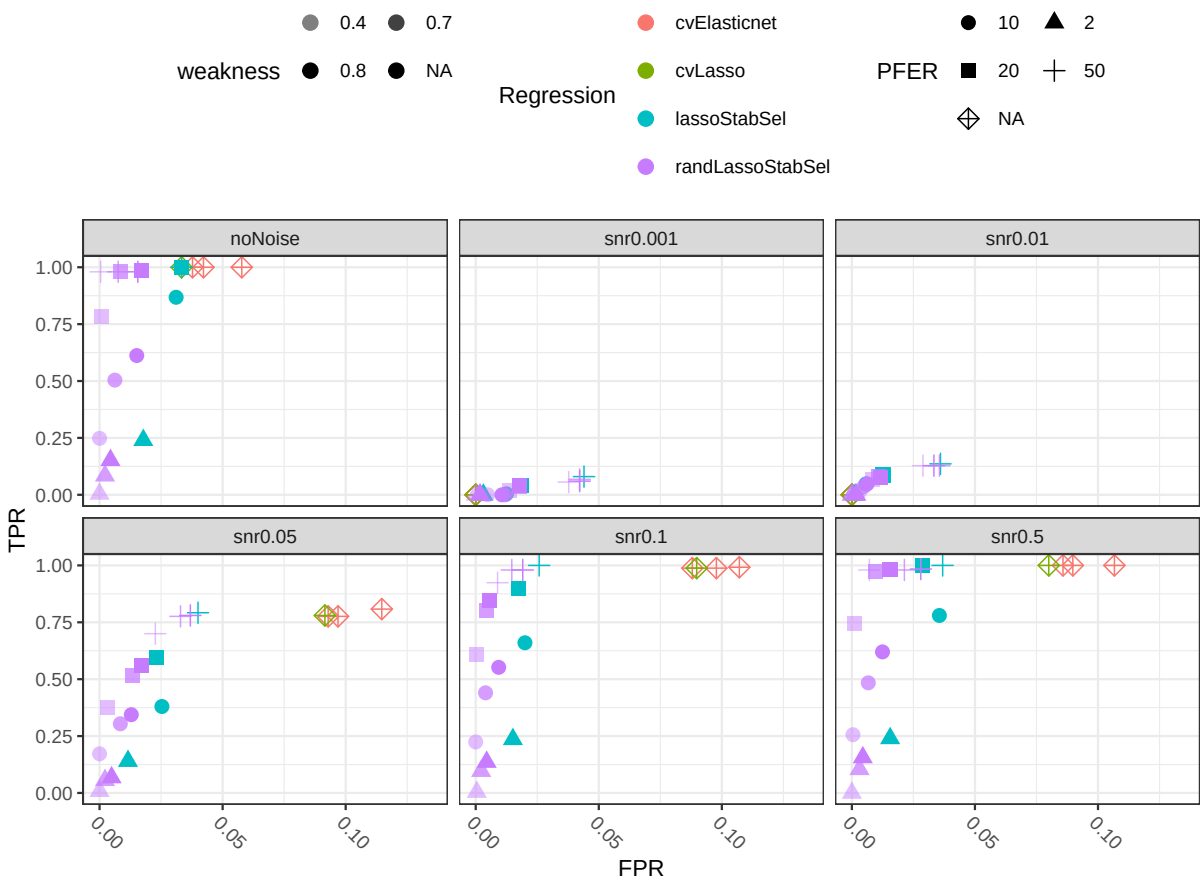


We calculate average values per group and plot.

```
# average over 'replicates'
tbl_avr <- tbl %>%
    as_tibble() %>%
    mutate(runParams = paste0(Regression, "_", run)) %>%
    group_by(runParams) %>%
    summarise(meanTPR = mean(TPR), meanFPR = mean(FPR)) %>%
    mutate(Regression = strsplit2(runParams, "_")[, 1],
           noiseLevel = strsplit2(runParams, "_")[, 2])
nm <- strsplit2(tbl_avr$runParams, "_alpha")[, 2]
nm[nm==""] <- NA
tbl_avr <- tbl_avr %>%
    add_column(alpha = nm)
nm <- strsplit2(tbl_avr$runParams, "_W")[, 2]
nm <- strsplit2(nm, "_")[, 1]
nm[nm==""] <- NA
tbl_avr <- tbl_avr %>%
    add_column(weakness = nm)
nm <- strsplit2(tbl_avr$runParams, "_PFER")[, 2]
nm <- strsplit2(nm, "_")[, 1]
nm[nm==""] <- NA
tbl_avr <- tbl_avr %>%
    add_column(PFER = nm)
tbl_avr <- tbl_avr %>%
    rename(TPR = meanTPR, FPR = meanFPR)
tbl_avr <- tbl_avr %>%
    mutate("weakness/alpha" = ifelse(Regression=="cvElasticnet", alpha, weakness))
```

```
tbl_avr %>%
    ggplot() +
    geom_point(aes(x = FPR, y = TPR, color = Regression, shape = PFER,
                   alpha = weakness), size = 3) +
    facet_wrap(~ noiseLevel) +
    scale_shape(na.value = 9) +
    scale_alpha_discrete(range = c(0.5,1)) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 315, hjust = 0)) +
    theme(legend.position = "top") +
    guides(shape = guide_legend(nrow = 3, byrow = TRUE),
           alpha = guide_legend(nrow = 3, byrow = TRUE),
           color = guide_legend(nrow = 4, ncol = 1, byrow = TRUE)) +
    coord_cartesian(xlim = c(0, range(tbl_avr$FPR)[2] + 0.02))
```

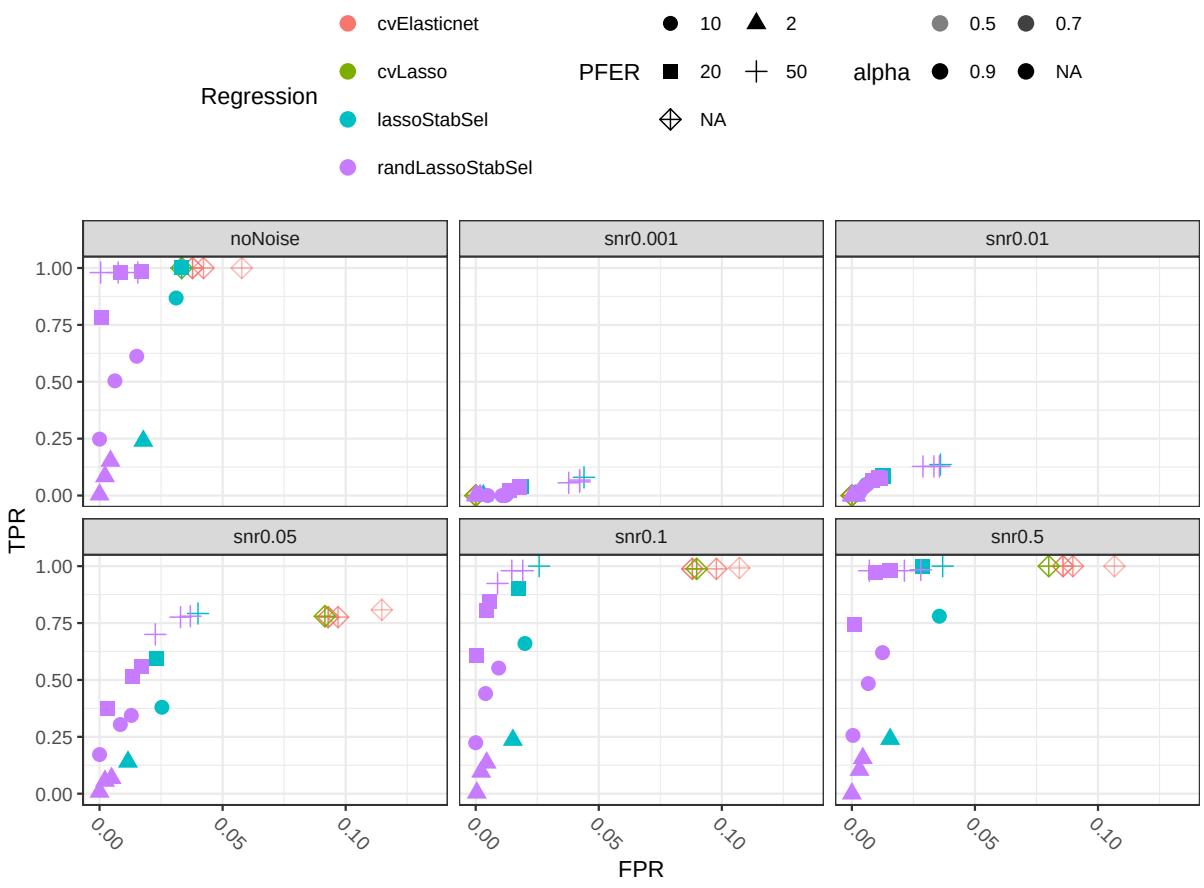
```
## Warning: Using alpha for a discrete variable is not advised.
```




```
tbl_avr %>%
    ggplot() +
    geom_point(aes(x = FPR, y = TPR, color = Regression, shape = PFER,
                   alpha = alpha), size = 3) +
    facet_wrap(~ noiseLevel) +
    scale_shape(na.value = 9) +
    scale_alpha_discrete(range = c(0.5,1)) +
```

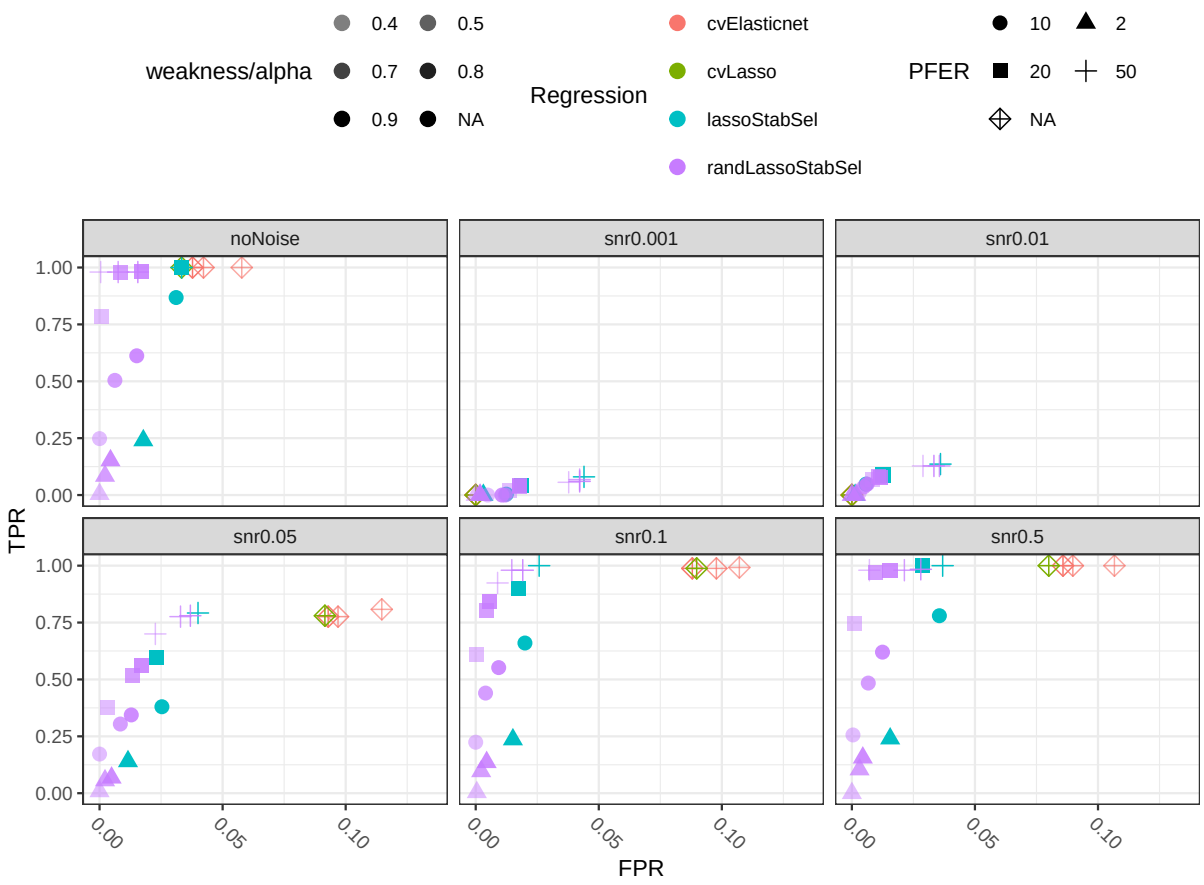
```
    theme_bw() +
    theme(axis.text.x = element_text(angle = 315, hjust = 0)) +
    theme(legend.position = "top") +
    guides(shape = guide_legend(nrow = 3, byrow = TRUE),
           alpha = guide_legend(nrow = 3, byrow = TRUE),
           color = guide_legend(nrow = 4, ncol = 1, byrow = TRUE)) +
    coord_cartesian(xlim = c(0, range(tbl_avr$FPR)[2] + 0.02))
```

```
## Warning: Using alpha for a discrete variable is not advised.
```




```
tbl_avr %>%
    ggplot() +
    geom_point(aes(x = FPR, y = TPR, color = Regression, shape = PFER,
                   alpha = `weakness/alpha`), size = 3) +
    facet_wrap(~ noiseLevel) +
    scale_shape(na.value = 9) +
    scale_alpha_discrete(range = c(0.5,1)) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 315, hjust = 0)) +
    theme(legend.position = "top") +
    guides(shape = guide_legend(nrow = 3, byrow = TRUE),
           alpha = guide_legend(nrow = 3, byrow = TRUE),
           color = guide_legend(nrow = 4, ncol = 1, byrow = TRUE)) +
    coord_cartesian(xlim = c(0, range(tbl_avr$FPR)[2] + 0.02))
```

```
## Warning: Using alpha for a discrete variable is not advised.
```



Conclusions

We can see stability selection better controlling for FPR, and the randomized lasso even more. It would, however select nothing rather than something wrong, which is why we see some low TPR values too. The fact that we had to increase our tolerance for false positives by increasing the PFER values also shows how conservative the methods are. For the very noisy examples, it's generally difficult to select meaningful predictors. However, the stability selection methods are able to pick up very few things rather than nothing compared to the lasso and elastic net with cross validation, while still controlling for FPR. Lasso or elastic net regressions will include predictors correlated to the true signal in their selections. That is why these methods select many predictors and control less well for false positives. As mentioned, here we have used stability selection proposed by Meinshausen and Bühlmann (2010), with improved error control proposed by Shah and Samworth (2013). For a different balance of TPR and FPR in stability selection, one can also vary the `assumption` parameter, based on the findings from Shah and Samworth (2013), which can be found in `stabs::stabsel`, which `monaLisa::randLassoStabSel` is built on.

We save the objects.

```
saveRDS(df_list, file = "files/04_regression_TPR_vs_FPR.rds")
saveRDS(X, file = "files/04_X.rds")
```

Session

```
date()
```

```
## [1] "Sun Aug 29 21:24:30 2021"
```

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: CentOS Linux 7 (Core)
##
## Matrix products: default
## BLAS/LAPACK: /tungstenfs/groups/gbioinfo/Appz/easybuild/software/OpenBLAS/0.3.12-GCC-10.2.0/lib/libop
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
##  [1] grid      parallel  stats4    stats     graphics  grDevices utils
##  [8] datasets  methods   base
##
## other attached packages:
##  [1] doMC_1.3.7                        iterators_1.0.13
##  [3] foreach_1.5.1                     glmnet_4.1-2
##  [5] limma_3.48.3                      ROCR_1.0-11
##  [7] Matrix_1.3-4                      circlize_0.4.13
##  [9] ComplexHeatmap_2.8.0              patchwork_1.1.1
## [11] forcats_0.5.1                     stringr_1.4.0
## [13] dplyr_1.0.7                       purrr_0.3.4
## [15] readr_2.0.1                       tidyr_1.1.3
## [17] tibble_3.1.3                      tidyverse_1.3.1
## [19] ggplot2_3.3.5                     BSgenome.Mmusculus.UCSC.mm10_1.4.0
## [21] BSgenome_1.60.0                   rtracklayer_1.52.1
## [23] Biostrings_2.60.2                 XVector_0.32.0
## [25] TFBSTools_1.30.0                  JASPAR2018_1.1.1
## [27] SummarizedExperiment_1.22.0       Biobase_2.52.0
## [29] MatrixGenerics_1.4.2              matrixStats_0.60.1
## [31] GenomicRanges_1.44.0              GenomeInfoDb_1.28.1
## [33] IRanges_2.26.0                    S4Vectors_0.30.0
## [35] BiocGenerics_0.38.0               monaLisa_0.2.0
##
## loaded via a namespace (and not attached):
##   [1] readxl_1.3.1               backports_1.2.1
##   [3] plyr_1.8.6                 igraph_1.2.6
##   [5] splines_4.1.1              BiocParallel_1.26.2
##   [7] digest_0.6.27              htmltools_0.5.1.1
```

```
##   [9] magick_2.7.3              GO.db_3.13.0
##  [11] fansi_0.5.0               magrittr_2.0.1
##  [13] memoise_2.0.0             cluster_2.1.2
##  [15] doParallel_1.0.16         tzdb_0.1.2
##  [17] annotate_1.70.0           modelr_0.1.8
##  [19] R.utils_2.10.1            colorspace_2.0-2
##  [21] blob_1.2.2                rvest_1.0.1
##  [23] haven_2.4.3               xfun_0.25
##  [25] crayon_1.4.1              RCurl_1.98-1.4
##  [27] jsonlite_1.7.2            TFMPvalue_0.0.8
##  [29] survival_3.2-11           glue_1.4.2
##  [31] gtable_0.3.0              zlibbioc_1.38.0
##  [33] GetoptLong_1.0.5          DelayedArray_0.18.0
##  [35] shape_1.4.6               scales_1.1.1
##  [37] DBI_1.1.1                 Rcpp_1.0.7
##  [39] xtable_1.8-4              clue_0.3-59
##  [41] bit_4.0.4                 stabs_0.6-4
##  [43] httr_1.4.2                RColorBrewer_1.1-2
##  [45] ellipsis_0.3.2            farver_2.1.0
##  [47] pkgconfig_2.0.3           XML_3.99-0.7
##  [49] R.methodsS3_1.8.1         dbplyr_2.1.1
##  [51] utf8_1.2.2                labeling_0.4.2
##  [53] tidyselect_1.1.1          rlang_0.4.11
##  [55] reshape2_1.4.4            AnnotationDbi_1.54.1
##  [57] munsell_0.5.0             cellranger_1.1.0
##  [59] tools_4.1.1               cachem_1.0.6
##  [61] cli_3.0.1                 DirichletMultinomial_1.34.0
##  [63] generics_0.1.0            RSQLite_2.2.8
##  [65] broom_0.7.9               evaluate_0.14
##  [67] fastmap_1.1.0             yaml_2.2.1
##  [69] knitr_1.33                bit64_4.0.5
##  [71] fs_1.5.0                  caTools_1.18.2
##  [73] KEGGREST_1.32.0           R.oo_1.24.0
##  [75] poweRlaw_0.70.6           pracma_2.3.3
##  [77] xml2_1.3.2                compiler_4.1.1
##  [79] rstudioapi_0.13           png_0.1-7
##  [81] reprex_2.0.1              stringi_1.7.3
##  [83] highr_0.9                 lattice_0.20-44
##  [85] CNEr_1.28.0               vctrs_0.3.8
##  [87] stringdist_0.9.7          pillar_1.6.2
##  [89] lifecycle_1.0.0           GlobalOptions_0.1.2
##  [91] bitops_1.0-7              R6_2.5.1
##  [93] BiocIO_1.2.0              codetools_0.2-18
##  [95] gtools_3.9.2              assertthat_0.2.1
##  [97] seqLogo_1.58.0            rjson_0.2.20
##  [99] withr_2.4.2               GenomicAlignments_1.28.0
## [101] Rsamtools_2.8.0           GenomeInfoDbData_1.2.6
## [103] hms_1.1.0                 rmarkdown_2.10
## [105] Cairo_1.5-12.2            lubridate_1.7.10
## [107] restfulr_0.0.13
```

References

Meinshausen, Nicolai, and Peter Bühlmann. 2010. "Stability Selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (4): 417–73. https://doi.org/10.1111/j.1467-9868.2010.00740.x.

Shah, Rajen D., and Richard J. Samworth. 2013. "Variable Selection with Error Control: Another Look at Stability Selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (1): 55–80. https://doi.org/10.1111/j.1467-9868.2011.01034.x.