BMC Bioinformatics

RESEARCH

Open Access

Fusion of single-cell transcriptome and DNA-binding data, for genomic network inference in cortical development

Thomas Bartlett*

*Correspondence:
thomas.bartlett.10@ucl.ac.uk
University College London, Gower Street, London WC1E 6BT, UK

Abstract

Background: Network models are well-established as very useful computational-statistical tools in cell biology. However, a genomic network model based only on gene expression data can, by definition, only infer gene co-expression networks. Hence, in order to infer gene regulatory patterns, it is necessary to also include data related to binding of regulatory factors to DNA.

Results: We propose a new dynamic genomic network model, for inferring patterns of genomic regulatory influence in dynamic processes such as development. Our model fuses experiment-specific gene expression data with publicly available DNA-binding data. The method we propose is computationally efficient, and can be applied to genome-wide data with tens of thousands of transcripts. Thus, our method is well suited for use as an exploratory tool for genome-wide data. We apply our method to data from human fetal cortical development, and our findings confirm genomic regulatory patterns which are recognised as being fundamental to neuronal development.

Conclusions: Our method provides a mathematical/computational toolbox which, when coupled with targeted experiments, will reveal and confirm important new functional genomic regulatory processes in mammalian development.

Keywords: Gene regulatory networks, Single-cell RNA-seq, Cortical development

Background

Network models have become very popular in cell biology in recent years, proving their usefulness in many contexts. Example applications include gene regulatory, co-expression and protein signalling networks. Most applications in cell biology continue to use static network models, including in the context of single-cell RNA-seq data [1, 2]. However, processes such as development are inherently dynamic, and hence for such applications, a time-varying genomic network model would be more appropriate. Better inference of time-varying genomic networks will allow regulatory patterns to be inferred which better characterise dynamic biological processes such as development.

Developmental processes are characterised by transient expression of certain key genes at specific times. Morphogen gradients set up at particular developmental stages

© The Author(s), 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

provide specific information about location of cells, leading to appropriate patterns of gene expression in those cells. These gene expression patterns define cellular lineages, which can then be locked in place by persistent expression of, for example, homeobox and bHLH genes [3]. In the neural lineage, particular subtypes of fully differentiated neurons and glial cells are arrived at as a result of sequential expression of such positional markers and fate-determining genes [4–7].

In the cortex of mammalian embryos, early in neural development morphogen gradients driven by WNT and SHH signalling are set up defining location in the cortex. For example, SP8 and COUP-TFI are expressed most strongly at either end of a rostrodorsal to caudoventral gradient, and PAX6 and EMX2 expressed most strongly at either end of a rostroventral to caudodorsal gradient. The positional information from these morphogen gradients leads to specific developmental trajectories being followed in human embryos, involving the expression of numerous other master regulators and positional markers such as NKX2-1, SOX6, COUP-TFII, GSX2, DLX1/2, and OLIG2 [8]. For example, later in neural development, these trajectories can lead (depending on location) to the sequential expression RELN, TBR1, CTIP2, CUX1 and SATB2 [4], which determine the specification of excitatory neuron subtypes. For a more detailed background on the molecular mechanisms of neuronal specification, a thorough review is provided by Guillemot and colleagues [9].

Cells can be characterised in terms of their progression through a dynamic process such as neural development according to their gene expression patterns. In this context, cells with similar gene expression patterns are characterised as being at a similar point along the developmental process, or trajectory. The notion of 'developmental time' of cells can be used to quantify the progression of those cells through the developmental process, or trajectory. Hence, developmental time can be characterised in terms of the gene expression patterns of the cells. It is also recognised that genomic network inference in single-cell data should be carried out on cells of specific types [10]. A time-varying network model allows the dynamic genomic network structure to be inferred from relatively homogenous groups of cells, with each such group of cells corresponding to a different developmental time-point. In this model, each such group of cells represents a different time-step along the developmental process, or trajectory.

Any genomic network model based only on gene expression data can by definition only infer gene co-expression networks. In order to infer gene regulatory patterns, it is necessary to also include data which relates to the physical binding of the products of some genes to DNA of other genes. Previous work has been successful at inferring these more complicated genomic network structures, by incorporating data of several different modalities [11, 12]. However, those models are very computationally intensive, and are appropriate only for small networks involving the influence of few tens of gene regulators (such as transcription factors) on a few hundreds of genes. On the other hand, the method which we propose here is able to infer genomic regulatory patterns from genome-wide data, based on the expression of tens of thousands of genes and / or transcripts. However, we also note that to confirm any novel findings of marker genes or transcription factors important in neurogenesis, any analysis using this method will need to be coupled with experimental verification. This

would require a combined dry/wet lab setting, which is beyond the scope of this investigation.

In this work, we propose a method to infer dynamic genomic network structure, fusing single-cell RNA-seq data from specific experiments (i.e., gene expression data) with publicly-available DNase-seq data (i.e., DNA-binding data). This paper is organised as follows. In "Methods" section, we define our model, and describe our inference method. In "Results" section, we present the results of applying our model/method to data from human fetal cortical development. Then in "Discussion" section, we discuss our findings and their wider implications.

Methods

Model overview

We infer genomic network structure by fitting a sparse linear model locally around each 'target gene'. This sparse linear model has the log gene expression for target gene i at time t as the response, and the log gene expression for all genes $j \neq i$ genome-wide at time t as potential predictors; variables are standardised before model fitting. From this genome-wide choice of potential predictor genes, the sparse model fit chooses a small set of predictor genes which together are able to predict the expression level of the target gene i. This chosen set of genes are then used to infer the local network structure around the target gene i. To infer the global network structure, we infer the local network structure around each target gene i in turn.

As well as gene expression data, we also use DNA-binding data to inform the sparse model fits. We use this DNA-binding data to reduce the sparsity of the model fit for predictor genes for which there is evidence of a physical DNA interaction between the gene-product of predictor gene j with the DNA of target gene i. This means that the sparse model fit is more likely to infer genomic network interactions between predictor genes and a target gene whenever there is evidence of a physical interaction between the gene product of the predictor genes and the DNA of the target gene.

Time-varying network model

Following earlier work [13], denoting log(gene expression+1) at time t as y_t for the target gene i and $\mathbf{x}_t$ for the $p-1$ other genes, the model is defined as:

$$y_t = a + \mathbf{b}_{t,:}\mathbf{x}_t^\top + \epsilon_t. \tag{1}$$

The time-varying coefficient vector $\mathbf{b}_{t,:}$ encodes the time-varying local network structure at time t around the target gene. If there is a non-zero element of this vector at $b_{t,j}$ (after thresholding to remove trivially small values), then a network edge is inferred between genes j and i. The row-vector $\mathbf{b}_{t,:}$ is a row of the matrix $\mathbf{b}$, and hence $\mathbf{b}$ encodes the time-varying local network structure around the target gene i for all times $t \in \{1, \ldots, T\}$.

Assuming local decomposability of the global network structure as in [14] allows the local network structure to be inferred separately around each target gene i. Also assuming Gaussian distributed errors with constant variance leads to the log-likelihood

$$\ell = -\sum_{t=1}^{T}(y_t - \mathbf{b}_{t,:}\mathbf{x}_t^\top)^2 - \Psi, \tag{2}$$

where the subtraction of the Ψ term leads to 'regularisation', or 'penalisation' (of the model likelihood). A special case of the model, which assumes no DNA-binding data is available to adjust local sparsity, defines Ψ as

$$\Psi = \lambda\left(\sum_{t=1}^{T}\|\mathbf{b}_{t,:}\|_1 + \sum_{t=2}^{T}\|\mathbf{b}_{t,:} - \mathbf{b}_{t-1,:}\|_1\right). \tag{3}$$

The first term of Ψ, i.e., $\sum_{t=1}^{T}\|\mathbf{b}_{t,:}\|_1$, encourages choosing a smaller number of regulator genes, minimising the number of non-zero entries in $\mathbf{b}_{t,:}$, which is referred to as 'sparsity within time' [13]. The second term of Ψ, i.e., $\sum_{t=2}^{T}\|\mathbf{b}_{t,:} - \mathbf{b}_{t-1,:}\|_1$, encourages smooth time-variation of $\mathbf{b}_{t,:}$, which is referred to as 'sparsity across time' [13].

The type of likelihood penalisation specified by Ψ (Equation (3)) falls within the generalised lasso framework [15], meaning that Ψ can be written as

$$\Psi = \lambda\|\mathbf{D}\,\mathrm{vec}(\mathbf{b}^\top)\|_1, \tag{4}$$

where $\mathrm{vec}(\cdot)$ is the vectorisation operator (which vectorises the $(p-1)\times T$ matrix $\mathbf{b}^\top$ to a $(p-1)T\times 1$ vector), and λ controls how sparse the model is. The penalty matrix $\mathbf{D}\in\mathbb{R}^{m\times(p-1)T}$ controls which elements of $\mathbf{b}$ are sparse, as well as controlling which differences between elements of $\mathbf{b}$ are sparse. Each row of $\mathbf{D}$ defines a different component of the sparsity. If exactly one element of a row of $\mathbf{D}$ is non-zero, then this leads to a contribution to the sparsity within time, i.e., sparsity for the variable and time-point at the corresponding location in $\mathrm{vec}(\mathbf{b}^\top)$. If exactly two elements of a row of $\mathbf{D}$ are non-zero, are of equal magnitude but opposite sign, and correspond to locations in $\mathrm{vec}(\mathbf{b}^\top)$ for the same variable at adjacent time-points, then this leads to a contribution to the sparsity across time. These are the only scenarios for this model in which any element of $\mathbf{D}$ is non-zero.

To achieve sparsity within time, there must be a separate row in $\mathbf{D}$ for each gene j for each time-point, with one non-zero element in each of these rows. The level of sparsity can be varied for each gene j by varying the magnitude of these non-zero elements, as long as the magnitude of the non-zero elements is the same for all T rows of $\mathbf{D}$ which correspond to gene j. Similarly, to achieve sparsity across time, there must be a separate row in $\mathbf{D}$ for each gene j for each pair of adjacent time-points $t-1$ and t, with $t\in\{2,\ldots,T\}$. Each of these rows must have exactly two non-zero elements with the same magnitude and opposite signs, at locations corresponding to times $t-1$ and t for gene j. Again, the level of sparsity can be varied for each gene j by varying the magnitude of these non-zero elements, as long as the magnitude of the non-zero elements is the same for all $T-1$ rows of $\mathbf{D}$ which correspond to gene j. In practice, we set the magnitude of the non-zero elements to be the same for all $2T-1$ rows of $\mathbf{D}$ which correspond to gene j—this covers both sparsity within and across time.

Using DNA-binding data to adjust local sparsity

To set the magnitude of the non-zero elements of the rows of **D** which correspond to gene *j*, we use the evidence available in DNA-binding data for any interaction of the protein-product of gene *j* with the promoter DNA of the target gene *i*. We set these magnitudes from model probabilities which quantify the evidence in the DNA-binding data for this protein-DNA interaction. These fitted model probabilities can come from any model, meaning that the framework presented here is independent of the model of the DNA-binding data which is used. Other authors have previously used such a notion of model independence for data-fusion in genomics [12].

The magnitude of the non-zero elements of **D** defaults to 1 whenever there is no evidence for the binding of the protein-product of gene *j* to the promoter DNA of the target gene *i*. This is typically the case when, for example, gene *j* does not code for a transcription factor. When there is evidence of binding of the protein-product of gene *j* to the promoter DNA of the target gene *i*, the magnitude of the corresponding elements of **D** is decreased below 1. The amount by which this magnitude is decreased varies according to the strength of evidence of DNA binding, with a minimum magnitude of $1/\eta$ (for $\eta > 1$) when the binding evidence is strongest. This means that the sparsity is decreased for gene *j* according to the strength of evidence of an interaction between the protein product of gene *j* with the promoter (or other regulatory) DNA of the target gene *i*.

We assume that the model of the DNA-binding data gives binding probabilities $p_{ji} \in [0, 1]$ for the interaction of the protein-product of gene *j* with the promoter DNA of target gene *i*. Then, we want $p_{ji} = 1$ and $p_{ji} = 0$ to correspond to magnitudes of $1/\eta$ and 1 respectively, for the non-zero elements in the rows of **D** which correspond to gene *j*. For $0 < p_{ji} < 1$, we want these magnitudes to scale proportionally to $\log(p_{ji})$. The intuition behind this proportionality of scaling is simply that by referring to the log-likelihood in Eq. (2), we can observe that the sparsity term Ψ is on the scale of log-probability. Hence we want any additive components of Ψ (as in Eqs. (3–4)) to scale proportionally with log-probabilities.

To achieve this scaling between $1/\eta$ and 1 proportionally to $\log(p_{ji})$, we set the magnitude of the non-zero elements of the rows of **D** which correspond to gene *j* as

$$\frac{1}{\eta} + \frac{\eta - 1}{\eta} \cdot \frac{\log(p_{ji})}{\log(p_{\min})}, \tag{5}$$

for $j \in \{j : p_{ji} \geq p_{\min}, \}$, where $p_{\min} > 0$ is the minimum model probability of interest. We set the magnitude of the non-zero elements of **D** to be 1 otherwise. Thus, η represents the factor by which λ is scaled down from gene *j* to gene j', when $p_{ji} \leq p_{\min}$ and $p_{j'i} = 1$. Like λ, the scaling factor $\eta > 1$ is set by the user. We have found that $\eta = 4$ works well in practice, and we previously found that $\lambda = 20$ is optimal [13].

Modelling co-regulation to achieve consistency in sparse model fits

The lasso is a linear model with L_1 penalisation, such as in Equation (3) where Ψ is comprised of $\|\cdot\|_1$ terms. It is well known that lasso models may lead to inconsistent results, if the same model is fit several times to the same data under slight perturbations of the set of variables available to the model [16]. This inconsistency happens because there

may be many subsets of the available variables which produce model fits which are virtually as good. A novel way to overcome this inconsistency in the context of genomic network inference is to look for genes which are consistently inferred as regulators of many genes in a particular gene-set.

In the time-varying genomic network model described here, a gene is inferred as a regulator of a target gene i at time t if it is included in the genes j which are selected as predictors of target gene i according to the non-zero model coefficients $|b_{t,j}| > 0$. Genes j which are inferred as predictors of target gene i are then inferred as being connected to this target gene in the local network structure around this target gene. If a particular gene j is inferred in this way as being connected to many such target genes i, where those target genes make up a particular gene-set, then consistency is demonstrated in the regulation of the target genes i of this gene set by the gene j. Such a gene-set could correspond to, for example, the marker genes known to specify particular cell-types.

Results

Inference of developmental pseudo-time

The main neuro-developmental data-set analysed here [17] consists of transcriptome measurements from single cells from developing fetal brains. Some of these cells are neural stem-cells, some are fully differentiated neurons and other cell types, and there is a whole spectrum of cells in between. No information is available for each cell other than its gene expression measurements. We inferred a time-ordering of all the cells before model fitting: this time-ordering gives the relative position on a 'developmental trajectory' which goes from neural stem cell (inferred time $\hat{t} = 1$) to fully differentiated cell type ($\hat{t} = T$). These developmental trajectories also branch with lineage, as cells are specified and differentiate into various different cell types. A time ordering inferred like this is often referred to as 'pseudo-time'. Several methods have been published previously, to carry out this pseudo-time inference [18–21]. We have followed a common theme amongst these methods, summarised as follows: (1) Dimensionality reduction e.g., by t-SNE (t-distributed stochastic neighbour embedding) [22]. (2) Trajectory and branch inference (often after some clustering, to assign cells of the same phenotype to the same pseudo-time point). (3) Biological inference (using prior knowledge to relate trajectory extrema to known cell-types).

Figure 1 shows a developmental pseudo-time ordering according to the strategy described above, based on the gene-expression measurements for the 2136 cells of the main neuro-developmental data-set analysed here [17]. The trajectories in Figure 1 are inferred from a minimum spanning tree, where the root represents the location of the stem cells and the leaves represent the fully differentiated cell types. This minimum spanning tree is fitted to the medioids of clusters obtained by the 'partition around medioids (PAM)' method, resulting in $T = 8$. The visualisation is via a t-SNE projection into three dimensions. Figure 1 also shows the cells coloured according to mean expression of marker genes identified for different cell-types (Additional file 1: Table S1). N.B., these marker genes (provided by domain-experts) were not used in any way to infer the pseudo-time ordering or the developmental trajectories, other than to determine which end of the trajectory is the stem cell. Hence these colourings blindly verify the appropriateness of the inferred developmental pseudo-time orderings.

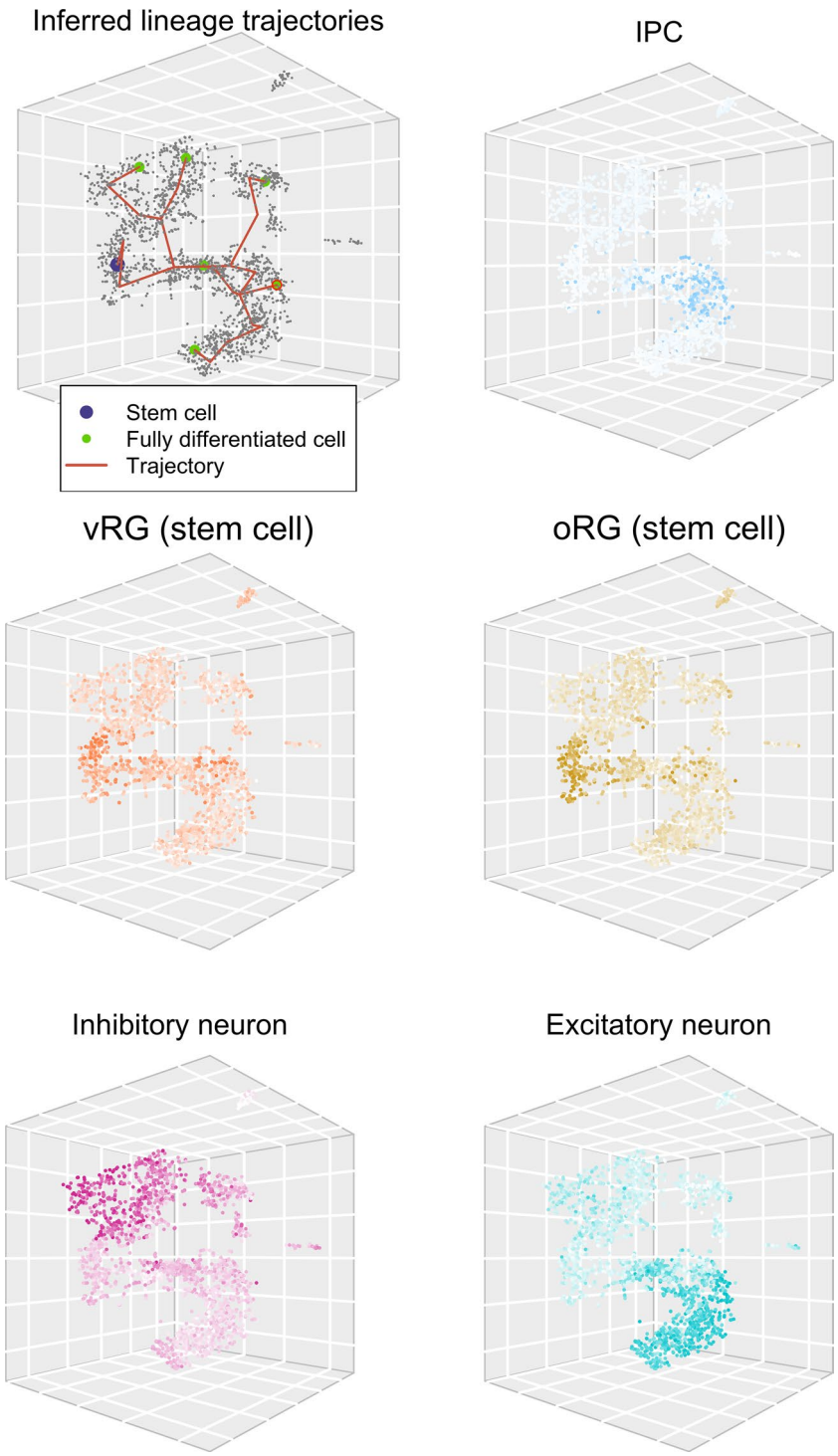

Fig. 1 Inferred lineage trajectories, together with cellular identities for verification of trajectories. The six detected terminally-differentiated neuronal subpopulations are shown in green, with a red outline for the one analysed further in "Inferring local structure with the proposed time-varying network model" section. vRG and oRG refer to 'ventricular radial glia' and 'outer radial glia' respectively

Inference of promoter-DNA binding

In order to fit the time-varying genomic network model of "Methods" section, we need to obtain probabilities quantifying the evidence for the interaction of the protein-product of gene j with the promoter DNA of target gene i. We denote these probabilities $p_{ji} \in [0, 1]$, and we estimate them as the posterior probabilities of this protein-DNA interaction using the *CENTIPEDE* model [23]. We use the *CENTIPEDE* model with 14 DNAse-seq data-sets from human fetal brain tissue, downloaded from *ENCODE* (www.encodeproject.org). This provides us with the posterior probabilities of each of 415 transcription factors j interacting with the DNA of 13907 target genes i. To fit the *CENTIPEDE* model, we specify that binding should be within 5000 base-pairs (5kbp) upstream of the transcriptional start site, with a 90% minimum probability weight matrix match score. These probability weight matrices were downloaded from the *JASPAR* database (jaspar.genereg.net).

Inferring local structure with the proposed time-varying network model

The developmental trajectories inferred according to "Inference of developmental pseudo-time" section were used to obtain a time-stamp for each cell. Using these time-stamps, together with with the promoter-DNA binding inference of "Inference of promoter-DNA binding" section, the time-varying genomic network model of "Methods" section was fit to the main neuro-developmental single-cell RNA-seq data-set [17]. This model fit infers the local network structure around a target gene, by choosing which genes best predict the expression of the target gene. The model makes this choice from all other 10774 genes, genome-wide, which are present in both the main neuro-developmental data-set and the DNA-binding data-set. This model fitting procedure takes 35 minutes on one processor core (MacBook Pro, 2019, 2.6 GHz). The model can be fitted to each target gene in turn from a panel of genes of interest, or genome-wide, to give the local network structure around each target gene. We note that this procedure can easily be run in parallel on multiple cores for a large panel of target genes of interest. Figure 2 shows the inferred time-varying network structure, which resulted from fitting a single model around the target-gene SATB2.

The gene SATB2 is well known for defining the identity of certain types of neuron [4, 24]. As would be expected for such a gene, network edges (representing direct or indirect genomic regulatory effects) appear at later times (Fig. 2), as the cells take on their neuronal identities. These network edges connect SATB2 to several genes, which are all relevant to neuronal development, as follows. At the final time-point (Fig. 2), NEUROD2 is recognised as the prominent neurogenesis gene 'neurogenic differentiation factor 2' [25]. The transcription factor NFIX is essential for neural development [26], and has a role in both embryonic and adult neurogenesis [27]. The transcription factor TFAP2C is part of the core cortical development programme [28]. The transcription factor LHX2 is known to promote neuronal as opposed to glial fate [29], as well as regulating the timing of cortical neurogenesis [30]. MEIS2 is known as a co-factor of the ventral neural fate marker PAX6 in neurogenesis [31] (PAX6 being one of the most important genes in neurogenesis [32]). The gene RBFOX1 is well known for regulating alternative splicing in neuronal development [33, 34]. The gene DCC encodes an axon guidance receptor

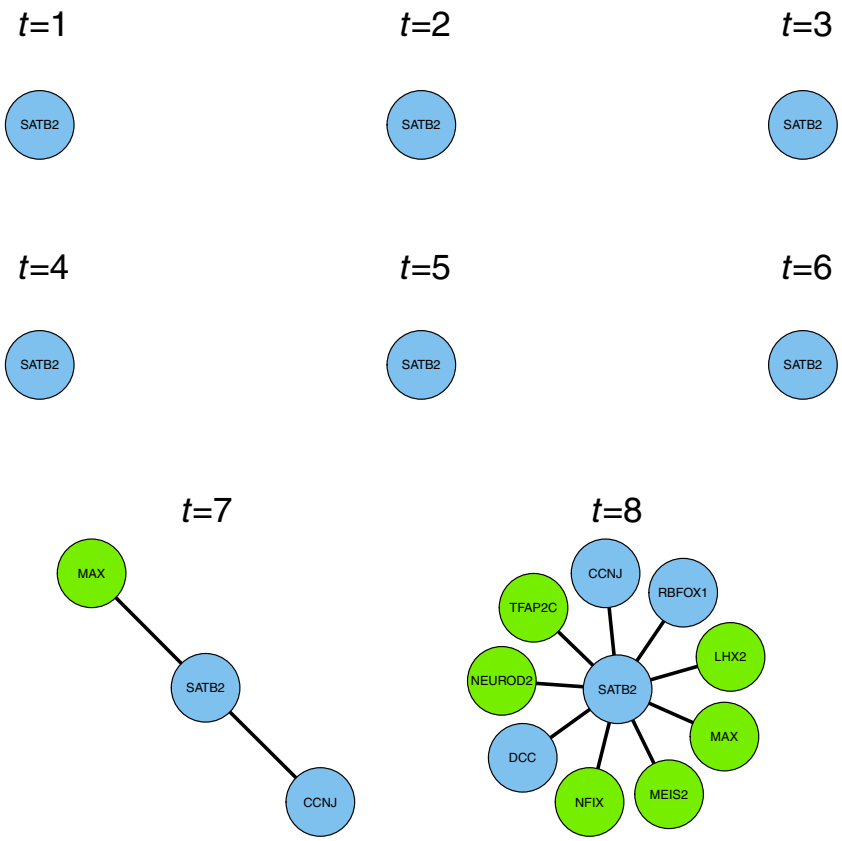

Fig. 2 Time-varying local network structure inferred around the target gene SATB2 with the model of "Methods" section. Genes with transcription-factor binding data are shown in green, and other genes are shown in blue

which is important for the migration of developing neurons [35]. Then at the penultimate time-point (when the cells may still be proliferating), the transcription factor MAX is recognised as being involved with cellular proliferation [36]. Also, the gene CCNJ is a cyclin, and thus it is involved in cellular proliferation via its role in the cell-cycle.

Inferring network structure using the thresholded correlation matrix: a comparison

A very popular way to infer gene co-expression network structure is by thresholding the gene expression correlation matrix, e.g. at $|\rho| \geq 0.5$, where ρ is the Pearson or Spearman correlation coefficient. This method of inferring genomic networks is often used in the most high-profile studies [37]. We compare this thresholded correlation matrix method of inferring genomic networks, with the proposed time-varying genomic network inference method. To make this comparison, we have inferred the local network structure around the target gene SATB2 by thresholding the correlation matrix of the cells assigned to the final time-point in the developmental trajectory (i.e., $T = 8$ in Fig. 2). If we choose to threshold at $|\rho| \geq 0.5$, we do not find any edges in the local network structure around SATB2 from this method. So instead, we threshold at $|\rho| \geq 0.4$. The local network structure inferred in this way is shown in Fig. 3. Notably, there are no well-known neuro-developmental transcription factors amongst the genes in this local

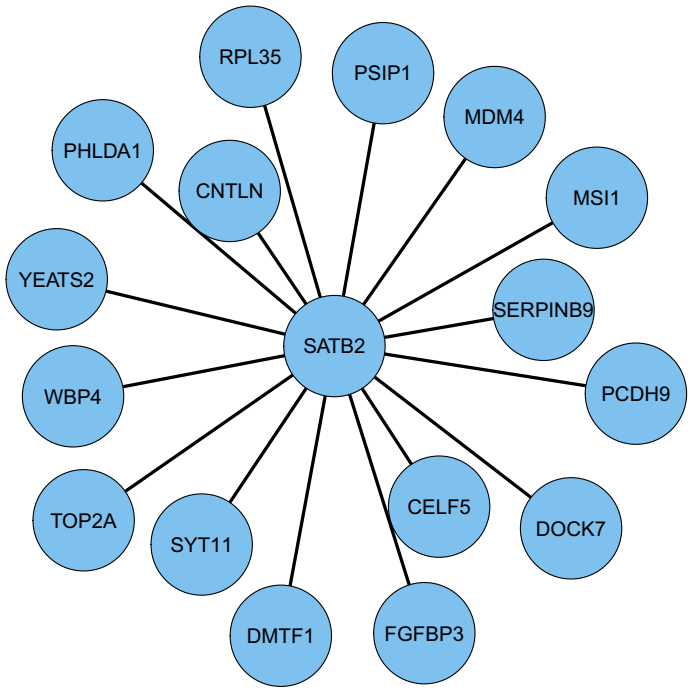

Fig. 3 Local network structure inferred around the target gene SATB2 from the correlation matrix of cells assigned to the final time-point

network structure. Many (but not all) of the genes in this local network structure have been associated previously with neural development and brain function, although the functional role is often less clear than those shown in Fig. 2. The role of the genes of Fig. 3 is summarised as follows. DOCK7 is thought to help regulate radial glial proliferation [38] (radial glia are an important type of cortical stem cell). MSI1 is known to be expressed in the sub-ventricular zone of neural stem cells [39]. RPL35, SYT11 and DMTF1 have been associated in a previous bioinformatic analysis with neurogenesis [40]. CNTLN has been correlated with RB-related protection from cell division during neurogenesis [41]. FGFBP3 has been previously associated with radial precursor cells [28]. TOP2A has previously been found to be expressed in the fetal telencephalon [42]. CELF5 is known to be expressed in the brain [43]. PCDH9 is a procadherin which may be expressed in the embryonic central nervous system [44]. PSIP1 has previously been associated with hereditary hearing loss [45]. Also, expression of PHLDA1 has been correlated with intractable epilepsy [46]. We also note that the inference of these gene co-expression network patterns is static, i.e., they do not vary with time; in other words the co-expression network is constrained so that no variation with time is possible. By using a softer constraint over time, smooth variation with time is possible.

Consistent regulation across excitatory neuron markers

The model used to infer these results (Eqs. (2)-(3)) is based on the lasso [47], which selects a smaller set of predictor variables (i.e., genes in this case) for the fitted model, from the full set of variables available (i.e., genome-wide in this case). A well known property of lasso-based models is that there may be several possible sets of predictor

variables which can be chosen by the model, each of which fit the data virtually equally well. Recent work has tried to overcome this issue, by looking for consistency amongst the chosen sets of predictor variables across many model fits [16]. In that work, the authors fit the model to several slight variations of the sets of predictor variables which are available to choose from, to find predictor variables which are chosen consistently across these model fits. An alternative approach which we use here is to fit the model to several different target genes, looking for predictor genes which are chosen consistently across these model fits. Importantly, these target genes are all taken from a particular gene-set of interest. This means that consistency across several model fits informs us about genomic regulatory processes which are fundamental to that gene-set of interest. Such a gene-set, comprising a highly curated selection of genes which is known to be important for excitatory neuronal identity, is given in Additional file 1: Table S2. Figure 4 shows the numbers of genes from this excitatory neuron gene-set which are found to be regulated by different transcription factors at the final time-point in the inferred dynamic network structure (the list of genes inferred as regulated by each TF is shown within the relevant bar of Fig. 4).

The transcription factors which appear in Fig. 4 are found to consistently regulate the genes of the excitatory neuron gene-set (Additional file 1: Table S2) and are known as being important for neural development, as follows. NEUROD2, LHX2, NFIX, MEIS2 and TFAP2C have been discussed already ("Inferring local structure with the proposed time-varying network model" section). Then, TCF4 is thought to be important in cortical and hippocampal neurogenesis [48]. STAT3 is thought to be important for neuronal

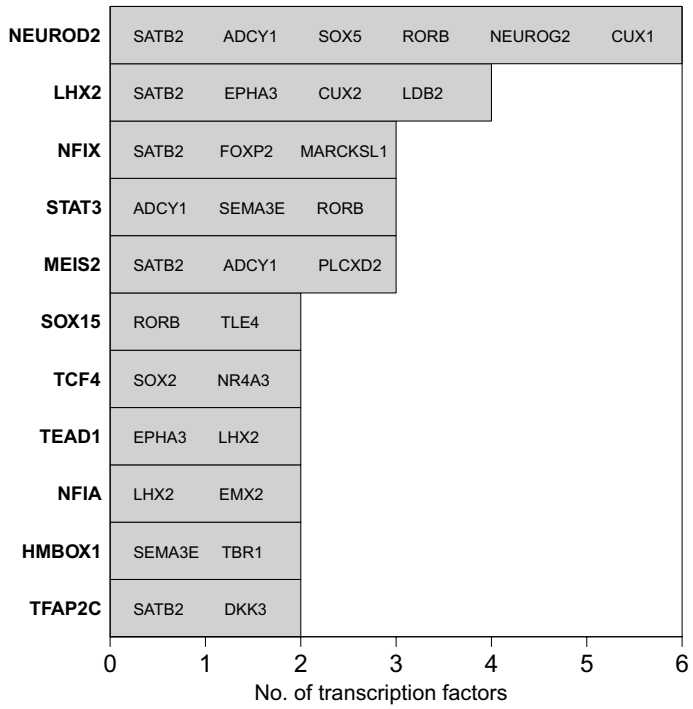

Fig. 4 Numbers of excitatory neuron genes from the selection in Additional file 1: Table S2, regulated by different transcription-factors

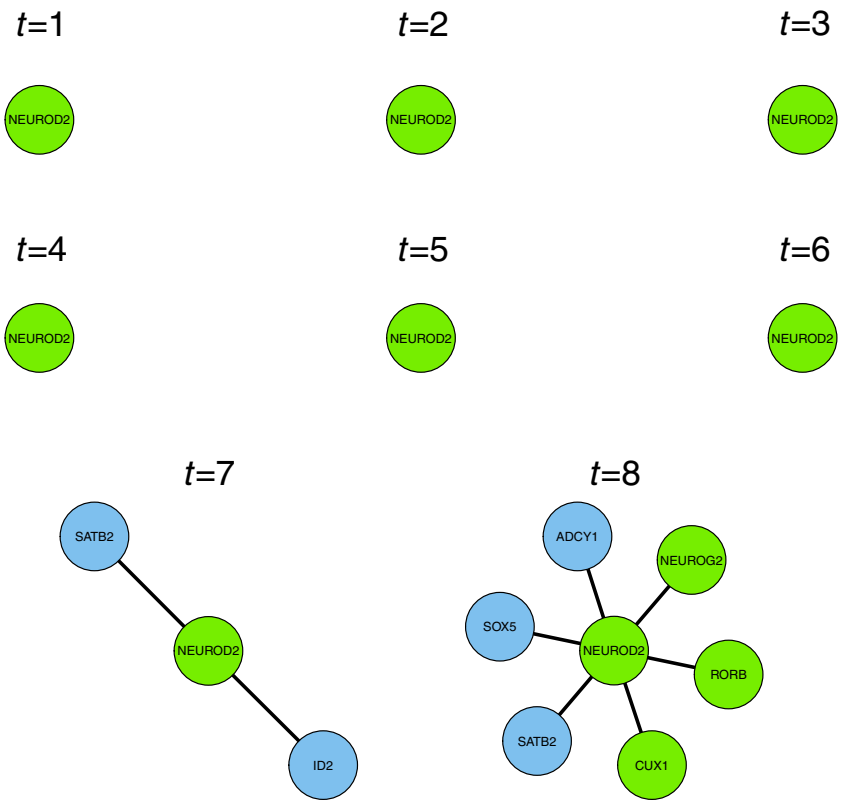

Fig. 5 Inferred time-varying genomic regulatory influence of NEUROD2, on the specified set of excitatory neuronal identity genes

differentiation [49], and more generally the JAK/STAT pathway is known to be important in the transition from neurogenesis to gliogenesis [50]. NFIA is best known as a transcription factor involved in the onset of gliogenesis [51] (i.e., following neurogenesis). Hence, the role of NFIA during neurogenesis is likely to be repressive. The TEAD transcription factors are known to have a role in neural progenitor specification [52], although a specific role for TEAD1 in neurogenesis has not yet been widely reported. Interestingly however, recent work has shown this gene has an important role in cell migration in the aggressive brain cancer glioblastoma [53]. Furthermore, TEAD1 has also been shown to be part of a neuronal transcriptional network which is fundamental to the progression of the pediatric brain cancer medulloblastoma [54]. SOX15 is not yet well known in neural development, although the SOX family of transcription factors are well known as co-factors in lineage specification throughout development [55]. HMBOX1, also known as HOT1, is a gene coding for a protein which binds to telomeres [56]. It's unclear what its role could be in neurogenesis, although it may be protective as newly generated neurons are known to be hypersensitive to telomere damage [57].

Inferring regulation by NEUROD2

The transcription factor identified in Fig. 4 as regulating the largest number of excitatory neuron genes is NEUROD2. Figure 5 shows the local network structure inferred, defined as the genes potentially regulated by NEUROD2. The network structure shown

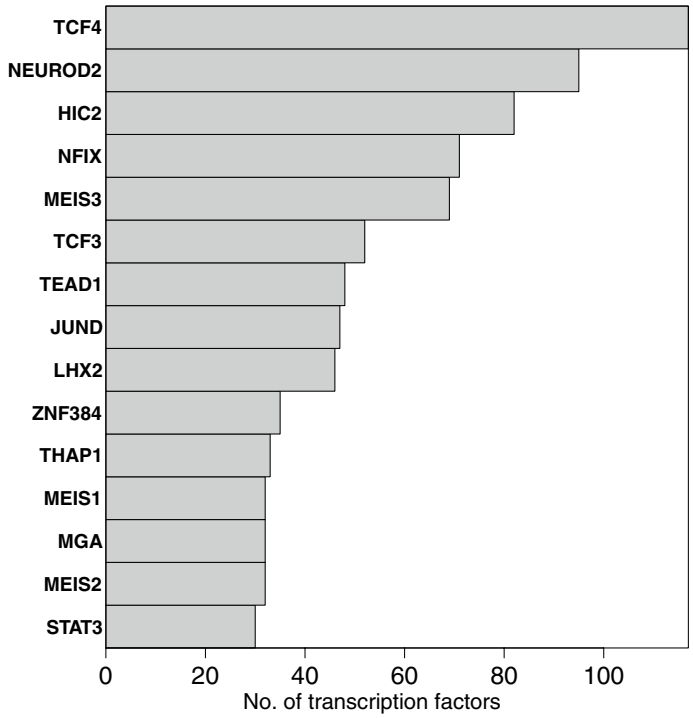

Fig. 6 Numbers of neuronal subtype genes of interest, regulated by different transcription-factors

in Fig. 5 is obtained from several model fits, each for a different target gene potentially regulated by NEUROD2. This is in contrast to the results for SATB2 in "2.3" section, which is obtained from just one model fit around the target gene SATB2.

The genes potentially regulated by NEUROD2 (Fig. 5) are all of interest to neural development. CUX1, like SATB2, is an important marker of specific neuronal subtypes [4]. NEUROG2 is important for cortical laminar fate specification [58]. RORB is involved in a mutually-repressive interaction with BRN1/2 to specify cortical laminar fate [59]. SOX5 has an important role in neuronal migration and differentiation [60]. ID2 is required for specification of certain types of neuron [61]. Finally, ADCY1 is a neuronal protein thought to have an important role in neuronal signal transduction and synaptic plasticity [62].

Consistent regulation across genes significant in a neuronal subpopulation

Similarly to the results of "2.5" section (shown in Fig. 4), we can also look for consistency of potential transcriptional regulators across target genes taken from a much larger gene-set. To identify such a gene-set, we used LIMMA and edgeR [63, 64] to identify genes which are significantly differentially expressed in a group of cells of interest, when compared to all the other cells in the data-set. We defined this group of cells of interest as those cells assigned to the final time-point in the developmental trajectory. This group of cells is expected to represent a particular excitatory neuronal subtype of interest. Significant genes were defined here as those genes with false discover rate-adjusted $p < 0.05$ in the differential expression analysis, which also increased their expression level in the cells of interest compared to all the other cells. Figure 6 shows the top

transcription factors in terms of the numbers of significant genes they are inferred to potentially regulate (the list of genes inferred as regulated by each TF is given in the Additional file 1).

The transcription factors which appear in Fig. 6 are all of interest to neural development. Several have been discussed already, and the rest are now discussed as follows. At the top of the list is TCF4. The role of TCF4 in neural development is not currently well understood, although loss-of-function mutations of this gene have been shown to be responsible for severe neurodevelopmental disorders [65], as well as conferring risk of schizophrenia [66]. However, recent work points to an important role for TCF4 in cortical and hippocampal neurogenesis [48]. Interestingly, HIC2 has no currently reported role in neurogenesis, although parallels have recently been drawn between the role of HIC2 in cardiovascular development, with the role of BRN3A in neural development [67]. MEIS3 is thought to mediate WNT-driven organisation of the neural plate in embryogenesis [68]. TCF3 is known as an inhibitor of neurogenesis, although interestingly this was reported previously in the spinal-cord [69]. JUND is known to have a role in the brain [70], although it may not be involved in development. The zinc-finger gene ZNF384 has previously been reported as having a role in neurogenesis, as part of the gene regulatory circuitry of the ventral neural fate marker PAX6 [71]. THAP1 has previously been reported to have a role in neural development [72]. MEIS1 is known to be an important developmental gene [73]. MGA is 'MAX gene-associated protein'; MAX is a transcription factor involved with cellular proliferation [36].

Coregulation by NEUROD2 and TCF4

Figure 7 shows a particular co-regulated subnetwork structure of interest, at the final time-point in the inferred dynamic network. This subnetwork structure comprises the neuronal identity genes which are inferred as being potentially regulated by NEUROD2 and TCF4. It is clear from Fig. 7 that many of these genes are co-regulated by both these transcription factors.

There are 29 genes connected to both NEUROD2 and TCF4 in the subnetwork of Fig. 7, indicating potential co-regulation by this pair of transcription factors. These 29 genes are: CALCOCO1, CHRDL1, CPSF4, DOCK4, FAM110A, FAM126A, FBLN1, GNG3, HDAC2, HECTD4, IGDCC3, ITPR2, KIAA1324, LRP8, MEX3A, NCS1, NFASC, PGAP1, RBCK1, RBFOX2, RPL37A, SCAF1, SEZ6, SIDT2, SNAP25, TMEM86A, TRIOBP, YWHAG, and ZDHHC20. Several of these are known to be of particular interest, as follows. The gene CHRDL1 promotes neuronal differentiation by inhibiting the important neural development gene BMP4 [74]: the 'bone morphogenetic proteins' (BMPs) have a fundamental, but complex, role in cellular specification throughout neural development [75]. The chromatin remodelling gene HDAC2 is 'histone deacetylase 2', which has been shown to control the progression of neural precursors to neurons during neural development [76]. The gene KIAA1324 encodes a transmembrane protein, and has been shown to be differentially expressed between the ventricular and subventricular cortical zones [77]. RBFOX2 regulates the alternative splicing of many important neuronal transcripts [78]. DOCK4 is thought to play a fundamental role in formation of neurites and dendrites [79, 80]. NFASC is thought to be involved in neuronal projection

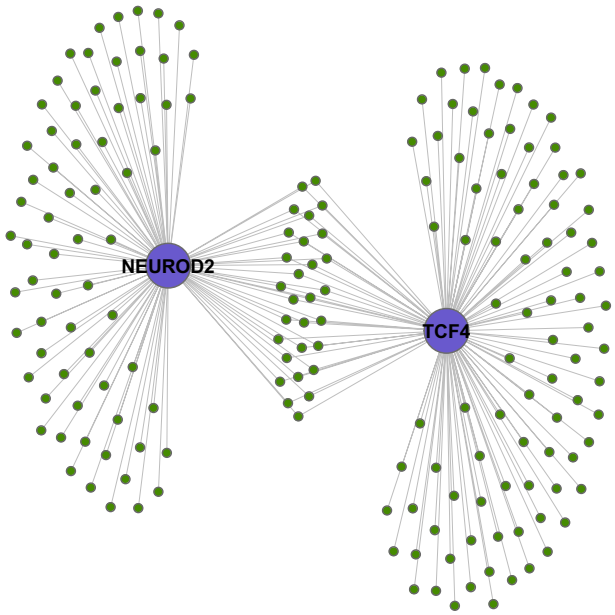

Fig. 7 Co-regulation of neuronal subtype genes (shown in green) by the transcription factors TCF4 and NEUROD2 (shown in purple). There are 29 genes connected to both NEUROD2 and TCF4, as well as 66 genes connected to only NEUROD2, and 87 genes connected to only TCF4

morphogenesis [81]. LRP8 is thought to be involved in neuronal migration via its interaction with RELN [82]. GNG3 is expressed in proliferating neural progenitors and immature neurons [83]. IGDCC3 is associated with a committed neuron phenotype [84]. FBLN1 is required for morphogenesis in neural crest-derived structures [85]. SNAP25 is involved in vesicular fusion and neurotransmitter exocytosis, with different isoforms in developing and adult tissue [86].

Discussion

In this paper, we have presented a new dynamic genomic network model, for inferring patterns of genomic regulatory influence in dynamic cell-biological processes such as development. We have applied this method to genome-wide data from human fetal cortical tissue, finding genomic interactions which are known to be fundamental to excitatory neuron specification. Our method compares very favourably with equivalent findings which we obtain from the same data using a popular method for network inference based on the data correlation matrix.

Our proposed method uses a large repository of publicly-available chromatin accessibility (DNAse-Seq) data, to identify transcription-factor (TF) bindings events that are possible in the neural lineage. It then uses expression data to infer potential regulatory relationships occurring at different times, guided by the possibilities identified in the chromatin accessibility data. We note that a potential regulatory relationship can still be identified if the evidence in the expression data is strong enough, even without corresponding evidence in the chromatin accessibility data. The evidence in the chromatin accessibility data of the possibility of binding of the TF to the regulatory DNA of the

target gene effectively reduces the threshold required in the evidence from the expression data, in order to infer a regulatory relationship between these genes.

We have applied our method to data from human tissue, and have interpreted our findings based on knowledge available from the wider literature. Much of what is known about neural development and neurogenesis in mammals is the result of rodent studies. While the neurodevelopmental principles in humans are likely to be similar to rodents in many ways, there must also be key differences due to the much greater size of the human and more generally the primate cortex. Hence, the findings from earlier studies in rodents which we have sometimes cited can only be taken as an indication of genomic regulatory interactions which may take place in human neural development, and particularly cortical neurogenesis.

Recent advances in single-cell genomic profiling include single-cell chromatin accessibility data, including where these measurements are obtained from the same cells as the single-cell RNA-seq measurements. In principle, the methods presented in this manuscript should be directly applicable to such data-sets. However, we note that as single-cell RNA-seq data-sets become larger in size, some trade-off is necessary between run-time and the size of the data-set.

Our method is computationally efficient, and can be applied to genome-wide data with tens of thousands of transcripts. However, we note that in order to define and describe genomic interactions which are more specifically mechanistic, a finer-grained model will be needed. This finer-grained model will necessarily be more complex, and thus would not be feasible to run at this genome-wide scale. Hence, the method we propose is most appropriate for a course-grained genome-wide discovery or exploration stage. This discovery stage can then be followed by the finer-grained stage of mechanistic modelling, which should also incorporate experimental validation.

Conclusions

The method we propose here provides a new mathematical and computational tool, which could be used together with targeted experiments in order to reveal important new functional genomic regulatory processes in mammalian development.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04201-9.

Additional file 1. Supplementary Tables and Supplementary Information.

Acknowledgements
Not applicable.

Authors' contributions
TB conceived and designed the study, carried out all analyses, and wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work was funded by the MRC Grant MR/P014070/1. The funding body had no role in the design of the study, collection, analysis, and interpretation of data, or writing of the manuscript.

Availability of data and materials
The data-sets analysed during the current study are available in the NCBI database of genotypes and phenotypes (***dbGaP***) under accession number ***phs000989.v3*** with URL https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000989.v3.p1 and from ***ENCODE*** under accession numbers ***ENCFF196CDZ.bam ENCFF436NOR.bam ENCFF807SCV.bam ENCFF808BII.bam ENCFF537DKA.bam ENCFF108NTA.bam ENCFF431RLT.bam ENCFF542AQF.bam***

ENCFF542NTO.bam ENCFF602FMG.bam ENCFF846NFV.bam ENCFF848SVJ.bam ENCFF948DOQ.bam and *ENCFF952QLZ.bam* with URLs https://www.encodeproject.org/files/ENCFF196CDZ/, https://www.encodeproject.org/files/ENCFF436NOR/, https://www.encodeproject.org/files/ENCFF807SCV/, https://www.encodeproject.org/files/ENCFF808BII/, https://www.encodeproject.org/files/ENCFF537DKA/, https://www.encodeproject.org/files/ENCFF108NTA/, https://www.encodeproject.org/files/ENCFF431RLT/, https://www.encodeproject.org/files/ENCFF542AQF/, https://www.encodeproject.org/files/ENCFF542NTO/, https://www.encodeproject.org/files/ENCFF602FMG/, https://www.encodeproject.org/files/ENCFF846NFV/, https://www.encodeproject.org/files/ENCFF848SVJ/, https://www.encodeproject.org/files/ENCFF948DOQ/, and https://www.encodeproject.org/files/ENCFF952QLZ/

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The author declares that there are no competing interests.

Received: 5 November 2020 Accepted: 12 May 2021
Published online: 04 June 2021

References

1. van Dijk D, Nainys J, Sharma R, Kathail P, Carr AJ, Moon KR, et al. MAGIC: a diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. BioRxiv. p. 111591. 2017.
2. Jackson CA, Castro DM, Saldi GA, Bonneau R, Gresham D. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. Elife. 2020;9:e51254.
3. Alberts B. Molecular biology of the cell. 4th ed. New York: Garland Science; 2002.
4. Hansen DV, Rubenstein JL, Kriegstein AR. Deriving excitatory neurons of the neocortex from pluripotent stem cells. Neuron. 2011;70(4):645–60.
5. Zhou Q, Anderson DJ. The bHLH transcription factors OLIG2 and OLIG1 couple neuronal and glial subtype specification. Cell. 2002;109(1):61–73.
6. Muroyama Y, Fujiwara Y, Orkin SH, Rowitch DH. Specification of astrocytes by bHLH protein SCL in a restricted region of the neural tube. Nature. 2005;438(7066):360–3.
7. Hochstim C, Deneen B, Lukaszewicz A, Zhou Q, Anderson DJ. Identification of positionally distinct astrocyte subtypes whose identities are specified by a homeodomain code. Cell. 2008;133(3):510–22.
8. Ma T, Wang C, Wang L, Zhou X, Tian M, Zhang Q, et al. Subcortical origins of human and monkey neocortical interneurons. Nat Neurosci. 2013;16(11):1588–97.
9. Martynoga B, Drechsel D, Guillemot F. Molecular control of neurogenesis: a view from the mammalian cerebral cortex. Cold Spring Harbor Perspect Biol. 2012;4(10):a008359.
10. Bartlett TE, Müller S, Diaz A. Single-cell co-expression subnetwork analysis. Sci Rep. 2017;7(1):15066.
11. Novershtern N, Regev A, Friedman N. Physical module networks: an integrative approach for reconstructing transcription regulation. Bioinformatics. 2011;27(13):i177–85.
12. Franks A, Markowetz F, Airoldi E. Estimating cellular pathways from an ensemble of heterogeneous data sources. 2014. arXiv preprint arXiv:14065799.
13. Bartlett TE, Kosmidis I, Silva R. Two-way sparsity for time-varying networks, with applications in genomics. Ann Appl Stat. 2020. https://doi.org/10.1214/20-AOAS1416 (**in press**).
14. Kolar M, Song L, Ahmed A, Xing EP. Estimating time-varying networks. Ann Appl Stat. 2010;4(1):94–123.
15. Tibshirani RJ, Taylor JE, Candes EJ, Hastie T. The solution path of the generalized lasso. Stanford: Stanford University; 2011.
16. Cox D, Battey H. Large numbers of explanatory variables, a semi-descriptive analysis. Proc Natl Acad Sci. 2017;114(32):8592–5.
17. Nowakowski TJ, Bhaduri A, Pollen AA, Alvarado B, Mostajo-Radji MA, Di Lullo E, et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science. 2017;358(6368):1318–23.
18. Qiu P, Simonds EF, Bendall SC, Gibbs KD Jr, Bruggner RV, Linderman MD, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol. 2011;29(10):886–91.
19. Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, et al. Bifurcation analysis of singlecell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci. 2014;111(52):E5643–50.
20. Bendall SC, Davis KL, Amir EAD, Tadmor MD, Simonds EF, Chen TJ, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell. 2014;157(3):714–25.
21. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381–6.
22. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(2579–2605):85.
23. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011;21(3):447–55.
24. Alcamo EA, Chirivella L, Dautzenberg M, Dobreva G, Fariñas I, Grosschedl R, et al. Satb2 regulates callosal projection neuron identity in the developing cerebral cortex. Neuron. 2008;57(3):364–77.

25. Fong AP, Yao Z, Zhong JW, Cao Y, Ruzzo WL, Gentleman RC, et al. Genetic and epigenetic determinants of neurogenesis and myogenesis. Dev Cell. 2012;22(4):721–35.
26. Campbell CE, Piper M, Plachez C, Yeh YT, Baizer JS, Osinski JM, et al. The transcription factor Nfix is essential for normal brain development. BMC Dev Biol. 2008;8(1):52.
27. Urbán N, Guillemot F. Neurogenesis in the embryonic and adult brain: same regulators, different roles. Front Cell Neurosci. 2014;8:396.
28. Yuzwa SA, Borrett MJ, Innes BT, Voronova A, Ketela T, Kaplan DR, et al. Developmental emergence of adult neural stem cells as revealed by single-cell transcriptional profiling. Cell Rep. 2017;21(13):3970–86.
29. Subramanian L, Sarkar A, Shetty AS, Muralidharan B, Padmanabhan H, Piper M, et al. Transcription factor Lhx2 is necessary and sufficient to suppress astrogliogenesis and promote neurogenesis in the developing hippocampus. Proc Natl Acad Sci. 2011;108(27):E265–74.
30. Hsu LCL, Nam S, Cui Y, Chang CP, Wang CF, Kuo HC, et al. Lhx2 regulates the timing of-atenin-dependent cortical neurogenesis. Proc Natl Acad Sci. 2015;112(39):12199–204.
31. Agoston Z, Heine P, Brill MS, Grebbin BM, Hau AC, Kallenborn-Gerhardt W, et al. Meis2 is a Pax6 co-factor in neurogenesis and dopaminergic periglomerular fate specification in the adult olfactory bulb. Development. 2014;141(1):28–38.
32. Heins N, Malatesta P, Cecconi F, Nakafuku M, Tucker KL, Hack MA, et al. Glial cells generate neurons: the role of the transcription factor Pax6. Nat Neurosci. 2002;5(4):308.
33. Lee JA, Tang ZZ, Black DL. An inducible change in Fox-1/A2BP1 splicing modulates the alternative splicing of downstream neuronal target exons. Genes Dev. 2009;23(19):2284–93.
34. Gehman LT, Stoilov P, Maguire J, Damianov A, Lin CH, Shiue L, et al. The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. Nat Genet. 2011;43(7):706.
35. Duman-Scheel M. Netrin and DCC: axon guidance regulators at the intersection of nervous system development and cancer. Curr Drug Targets. 2009;10(7):602–10.
36. Amati B, Land H. MycMaxMad: a transcription factor network controlling cell cycle progression, differentiation and death. Curr Opin Genet Dev. 1994;4(1):102–8.
37. Treutlein B, Lee QY, Camp JG, Mall M, Koh W, Shariati SAM, et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. Nature. 2016;534(7607):391–5.
38. Yang YT, Wang CL, Van Aelst L. DOCK7 interacts with TACC3 to regulate interkinetic nuclear migration and cortical neurogenesis. Nat Neurosci. 2012;15(9):1201.
39. Yagita Y, Kitagawa K, Sasaki T, Miyata T, Okano H, Hori M, et al. Differential expression of Musashi1 and nestin in the adult rat hippocampus after ischemia. J Neurosci Res. 2002;69(6):750–6.
40. Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, et al. Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell. 2015;17(3):360–72.
41. Oshikawa M, Okada K, Nakajima K, Ajioka I. Cortical excitatory neurons become protected from cell division during neurogenesis in an Rb family-dependent manner. Development. 2013;140(11):2310–20.
42. Harkin LF, Gerrelli D, Gold Diaz DC, Santos C, Alzu'bi A, Austin CA, et al. Distinct expression patterns for type II topoisomerases IIA and IIB in the early foetal human telencephalon. J Anat. 2016;228(3):452–63.
43. Ladd AN, Charlet-B N, Cooper TA. The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. Mol Cell Biol. 2001;21(4):1285–96.
44. Liu Q, Chen Y, Pan JJ, Murakami T. Expression of protocadherin-9 and protocadherin-17 in the nervous system of the embryonic zebrafish. Gene Expr Patterns. 2009;9(7):490–6.
45. Girotto G, Scheffer DI, Morgan A, Vozzi D, Rubinato E, Di Stazio M, et al. PSIP1/LEDGF: a new gene likely involved in sensorineural progressive hearing loss. Sci Rep. 2015;5:18568.
46. Xi ZQ, Wang LY, Sun JJ, Liu XZ, Zhu X, Xiao F, et al. TDAG51 in the anterior temporal neocortex of patients with intractable epilepsy. Neurosci Lett. 2007;425(1):53–8.
47. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol). 1996;58:267–88.
48. Jung M, Häberle BM, Tschaikowsky T, Wittmann MT, Balta EA, Stadler VC, et al. Analysis of the expression pattern of the schizophrenia-risk and intellectual disability gene TCF4 in the developing and adult brain suggests a role in development and plasticity of cortical and hippocampal neurons. Mol Autism. 2018;9(1):20.
49. Snyder M, Huang XY, Zhang JJ. Stat3 is essential for neuronal differentiation through direct transcriptional regulation of the Sox6 gene. FEBS Lett. 2011;585(1):148–52.
50. He F, Ge W, Martinowich K, Becker-Catania S, Coskun V, Zhu W, et al. A positive autoregulatory loop of Jak-STAT signaling controls the onset of astrogliogenesis. Nat Neurosci. 2005;8(5):616.
51. Deneen B, Ho R, Lukaszewicz A, Hochstim CJ, Gronostajski RM, Anderson DJ. The transcription factor NFIA controls the onset of gliogenesis in the developing spinal cord. Neuron. 2006;52(6):953–68.
52. Cao X, Pfaff SL, Gage FH. YAP regulates neural progenitor cell number via the TEA domain transcription factor. Genes Dev. 2008;22(23):3320–34.
53. Tome-Garcia J, Erfani P, Nudelman G, Tsankov AM, Katsyv I, Tejero R, et al. Analysis of chromatin accessibility uncovers TEAD1 as a regulator of migration in human glioblastoma. Nat Commun. 2018;9(1):4020.
54. Łastowska M, Al-Afghani H, Al-Balool HH, Sheth H, Mercer E, Coxhead JM, et al. Identification of a neuronal transcription factor network involved in medulloblastoma development. Acta Neuropathol Commun. 2013;1(1):35.
55. Kamachi Y, Kondoh H. Sox proteins: regulators of cell fate specification and differentiation. Development. 2013;140(20):4129–44.
56. Kappei D, Butter F, Benda C, Scheibe M, Draˇskoviˇc I, Stevense M, et al. HOT1 is a mammalian direct telomere repeat-binding protein contributing to telomerase recruitment. EMBO J. 2013;32(12):1681–701.
57. Cheng A, Shin-ya K, Wan R, Tang SC, Miura T, Tang H, et al. Telomere protection mechanisms change during neurogenesis and neuronal maturation: newly generated neurons are hypersensitive to telomere and DNA damage. J Neurosci. 2007;27(14):3722–33.

58. Dennis DJ, Wilkinson G, Li S, Dixit R, Adnani L, Balakrishnan A, et al. Neurog2 and Ascl1 together regulate a postmitotic derepression circuit to govern laminar fate specification in the murine neocortex. Proc Natl Acad Sci. 2017;14(25):E4934–43.
59. Oishi K, Aramaki M, Nakajima K. Mutually repressive interaction between Brn1/2 and Rorb contributes to the establishment of neocortical layer 2/3 and layer 4. Proc Natl Acad Sci. 2016;113(12):3371–6.
60. Kwan KY, Lam MM, Krsnik Ž, Kawasawa YI, Lefebvre V, Šestan N. SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. Proc Natl Acad Sci. 2008;105(41):16021–6.
61. Havrda MC, Harris BT, Mantani A, Ward NM, Paolella BR, Cuzon VC, et al. Id2 is required for specification of dopaminergic neurons during adult olfactory neurogenesis. J Neurosci. 2008;28(52):14074–87.
62. Wang H, Zhang M. The role of Ca2+-stimulated adenylyl cyclases in bidirectional synaptic plasticity and brain function. Rev Neurosci. 2012;23(1):67–78.
63. Smyth GK, et al. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):3.
64. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
65. Amiel J, Rio M, de Pontual L, Redon R, Malan V, Boddaert N, et al. Mutations in TCF4, encoding a class I basic helix-loop-helix transcription factor, are responsible for Pitt-Hopkins syndrome, a severe epileptic encephalopathy associated with autonomic dysfunction. Am J Hum Genet. 2007;80(5):988–93.
66. Steinberg S, de Jong S, Andreassen OA, Werge T, Børglum AD, Mors O, et al. Common variants at VRK2 and TCF4 conferring risk of schizophrenia. Hum Mol Genet. 2011;20(20):4076–81.
67. Dykes IM, van Bueren KL, Scambler PJ. HIC2 regulates isoform switching during maturation of the cardiovascular system. J Mol Cell Cardiol. 2018;114:29–37.
68. Elkouby YM, Elias S, Casey ES, Blythe SA, Tsabar N, Klein PS, et al. Mesodermal Wnt signalling organizes the neural plate via Meis3. Development. 2010;137(9):1531–41.
69. Gribble SL, Kim HS, Bonner J, Wang X, Dorsky RI. Tcf3 inhibits spinal cord neurogenesis by regulating sox4a expression. Development. 2009;136(5):781–9.
70. Herdegen T, Kovary K, Buhl A, Bravo R, Zimmermann M, Gass P. Basal expression of the inducible transcription factors c-Jun, JunB, JunD, c-Fos, FosB, and Krox-24 in the adult rat brain. J Comp Neurol. 1995;354(1):39–56.
71. Thakurela S, Tiwari N, Schick S, Garding A, Ivanek R, Berninger B, et al. Mapping gene regulatory circuitry of Pax6 during neurogenesis. Cell Discov. 2016;2:15045.
72. Zhao Y, Xiao J, Gong S, Clara JA, LeDoux MS. Neural expression of the transcription factor THAP1 during development in rat. Neuroscience. 2013;231:282–95.
73. Moens CB, Selleri L. Hox cofactors in vertebrate development. Dev Biol. 2006;291(2):193–206.
74. Gao WL, Zhang SQ, Zhang H, Wan B, Yin ZS. Chordin-like protein 1 promotes neuronal differentiation by inhibiting bone morphogenetic protein-4 in neural stem cells. Mol Med Rep. 2013;7(4):1143–8.
75. Bond AM, Bhalala OG, Kessler JA. The dynamic role of bone morphogenetic proteins in neural stem cell fate and maturation. Dev Neurobiol. 2012;72(7):1068–84.
76. Montgomery RL, Hsieh J, Barbosa AC, Richardson JA, Olson EN. Histone deacetylases 1 and 2 control the progression of neural precursors to neurons during brain development. Proc Natl Acad Sci. 2009;106(19):7876–81.
77. Fietz SA, Lachmann R, Brandl H, Kircher M, Samusik N, Schröder R, et al. Transcriptomes of germinal zones of human and mouse fetal neocortex suggest a role of extracellular matrix in progenitor self-renewal. Proc Natl Acad Sci. 2012;109(29):11836–41.
78. Gehman LT, Meera P, Stoilov P, Shiue L, O'Brien JE, Meisler MH, et al. The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function. Genes Dev. 2012;26(5):445–60.
79. Xiao Y, Peng Y, Wan J, Tang G, Chen Y, Tang J, et al. The atypical guanine nucleotide exchange factor Dock4 regulates neurite differentiation through modulation of Rac1 and actin dynamics. J Biol Chem. 2013;288(27):20034–45.
80. Ueda S, Fujimoto S, Hiramoto K, Negishi M, Katoh H. Dock4 regulates dendritic development in hippocampal neurons. J Neurosci Res. 2008;86(14):3052–61.
81. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell. 2014;159(7):1511–23.
82. Chai X, Frotscher M. How does Reelin signaling regulate the neuronal cytoskeleton during migration? Neurogenesis. 2016;3(1):e1242455.
83. Chen YJJ, Friedman BA, Ha C, Durinck S, Liu J, Rubenstein JL, et al. Single-cell RNA sequencing identifies distinct mouse medial ganglionic eminence cell types. Sci Rep. 2017;7:45656.
84. Butts JC, McCreedy DA, Martinez-Vargas JA, Mendoza-Camacho FN, Hookway TA, Gifford CA, et al. Differentiation of V2a interneurons from human pluripotent stem cells. Proc Natl Acad Sci. 2017;114(19):4969–74.
85. Cooley MA, Kern CB, Fresco VM, Wessels A, Thompson RP, McQuinn TC, et al. Fibulin-1 is required for morphogenesis of neural crest-derived structures. Dev Biol. 2008;319(2):336–45.
86. Bark IC, Hahn KM, Ryabinin AE, Wilson MC. Differential expression of SNAP-25 protein isoforms during divergent vesicle fusion events of neural development. Proc Natl Acad Sci. 1995;92(5):1510–4.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.