A Thesis submitted for the degree of Master of Science

Inferring Gene Regulatory Networks from single cell RNA-seq Data for Extending Mechanistic Models in Embryonic Neurogenesis

---

Dana Barilan

Supervisor and Examiner: Prof. Dr. Jana Wolf

Co-examiner: Prof. Dr. Heike Siebert

Freie Univeristät Berlin
Department of Computer Science
March 31st, 2024

Declaration

I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.

Place, Date	Signature

Abstract

foo bar

Contents

1 Introduction

1.1 Motivation

Neurogenesis...

The relatively simple central nervous system (CNS) of the Drosophila embryo provides a useful model system for investigating the mechanisms that generate and pattern complex nervous systems (Segment polarity genes in neuroblast formation and identity specification during Drosophila neurogenesis) - motivation?

Modeling of GRN limitations and why SC-data can help (importance of data driven GRN inferring)....

Traditionally, GRNs are reconstructed gene by gene, by experimental data. ... High trhroughput sequencing opens the potential for large GRNs reconstruction.

This thesis offers an approach of complementing a Boolean model GRN with SC-rna seq data.

1.2 Research Structure

A gene regulatory network was inferred from single cell RNA sequencing data using SCENIC workflow. Then used to extend an existing Boolean model.

2 Related Work

Give a brief overview of the work relevant for your thesis.

3 Background

3.1 Central Nervous System Development in *Drosophila* melanogaster

The central nervous system (CNS) of the *Drosophila* embryo is a complex organ encompassing approximately 15,000 neurons and glial cells and develops over the course of roughly one day [1–5]. This system is composed of the brain and the ventral nerve cord (VNC) and is structured into segments known as neuromeres. In the VNC, Each neuromere is subdivided by a set of midline cells into two bilaterally symmetric hemineuromeres, or hemisegments [1, 6]. The VNC itself is comprised of a total of 28 hemineuromeres. The VNC is generated in the ventral-lateral region of the embryo called the neuroectoderm [1].

Neuroblasts Formation and Specification in the VNC During Neurogenesis

Neuroblasts are the progenitor cells that give rise to the diverse neuronal and glial populations within the Drosophila CNS. During neurogenesis NBs within the hemisegments of the VNC are emerging by delamination from a sheet of surrounding cells in five consecutive waves (S1 to S5) [7], and are arranged into a pattern of rows and columns. The formation and fate specification of neuroblasts are uniform across all hemisegments. Different NBs give rise to different types of neurons or glia cells later in development. These NBs' fates are determined by their relative position within the hemisegment and time of formation. That means that NBs in different hemisegments, that are developed in the same row and column within the hemisegment, acquire the same fate. Eventually each NB undergoes division to generate a distinct group of neurons or glial cells that make up a functioning CNS [8].

Neuroectoderm Patterning

Neuroblasts formation in the neuroectoderm takes place between stages 8-9 to late stage 11 of development, during approximately four hours. Each hemineuromer counts

about 10 NBs at the beginning of the NB formation process, and ends up with about 32 [1, 7, 9, 10].

NBs are formed within specialized groups of cells known as preneural clusters, where the specific pattern of gene expression within these clusters dictates the identity and eventual fate of each NB. In each hemisegment NBs are organized initially into a grid of four rows and three columns oriented along the anteroposterior (AP) and dorsoventral (DV) axes of the embryo, respectively [1, 8, 9].

The patterning of the neuroectoderm is regulated by the interplay between a number of signaling pathways. The activity of different pathways in different rows and columns of the hemisegment result in different gene expression profile of each NB depending on it's position, hence determining its identity and unique fate. Segmentation genes were found to be involved in NBs specification as well, some of them as part of known signalling pathways. However the full regulatory mechanism in which signalling pathways employ different genes is not yet fully understood [9]. This remains a pursuit in developmental biology.

Signaling Pathways

Several signaling pathways were found to play an important role in NBs formation and specification, such as Notch, Wingless (Wg/Wnt), Hedgehog (Hh) and Epidermal growth factor receptor (EGFr) signaling pathways.

Notch signaling pathway was found to affect the formation of all NBs [1].

Wg, Hh and EGFr pathways were shown to be involved in the separation of NBs to different rows and columns. More specifically, the interaction between Wg and Hh were found to pattern the embryo repetitively to different rows along the AP axis. For example, Wg pathway is active in rows 4 and 5 during neurogenesis. However it was shown that row 5 it unaffected by Wg activity when the *gooseberry* segment polarity gene, which is part of the Hh pathway, is expressed [9, 11].

The EGFr signaling pathway illustrates columnar patterning along the dorsoventral (DV) axis, playing a role in dividing the hemisegment into three columns that express different segmentation genes called the columnar genes, with EGFr signaling active in the medial and intermediate columns [12].

Segmentation Genes

Two key groups of segmentation genes were found to play a crucial role in the formation and specification of neuroblasts in the neuroectoderm, namely segment

polarity genes on the AP axis and columnar genes on the DV axis. The specific combination of segmentation and columnar genes expressed in both the preneural clusters and the NBs themselves establishes the unique identity of each NB, thereby determining its fate [8, 9]. Segmentation genes of these two groups serve as markers to identify specific NBs, as different NBs express unique combinations of these gene, resulting in distinct expression profile that define each NB's identity [13].

The activity of the segment polarity genes segments the neuroectoderm into four rows per hemisegment and can largely explain how NBs that develop in these different transverse rows acquire different fates [8, 9]. Segment polarity genes include signaling molecules (*wingless (wg)*, *hedgehog (Hh)*), transmembrane receptors (*patched (ptc)*), transcription factors (*gooseberry (gsb), engrailed (en))* that are part of the Wg and Hh signaling pathways [9].

Columnar gene activity was shown to subdivides the neuroectoderm into three longitudinal columns, ensuring that NBs that develop in different columns acquire different fates [8]. Columnar genes include the *EGF receptor (EGFr)* mentioned above, and the three transcription factors *ventral nerve cord defective (vnd)*, *intermediate nerve cord defective (ind)*, and *drop (Dr)* also known as *msh*. *Vnd*, *ind* and *Dr* are expressed in the medial, intermediate and lateral columns respectively [8, 12, 14–17].

Neuroblasts Nomenclature

As mentioned above, NBs within each hemisegments are initially organized into a grid of 4 rows and 3 columns. NBs continue to form in between this initial grid, and it becomes difficult to discern clear columns or rows of NBs, due to variability in NB positions [13]. NBs are given numerical names based on their final position in the pattern. They get two numbers: the first one indicates the AP position (rows) and the second one indicates the mediolateral position on the dorsoventral axis (columns). For example NB 1-1 is the most medial NB of row 1. The first appearing four NBs are numbered with odd numbers (1, 3, 5, 7) while the even numbers (2, 4, 6) are developed shortly later between them [8].

3.2 Gene Regulatory Networks

Gene regulatory networks (GRNs) are topological diagrams or graphs that model the regulatory interactions within a cell, where the nodes represent molecular entities such as genes, proteins, or RNA molecules, and the directed edges depict the regulatory relationships. Edges can also be signed to represent the nature of the regulation, for

example positive or negative regulation [18, 19]. Typically these networks include the regulatory connections between transcription factors (TFs) and the genes they regulate known as target genes (TGs) [18]. TFs are DNA binding proteins that either activate or repress the transcription of their TGs when binding to their gene regulatory regions [20]. More specifically, TFs bind to regulatory regions of the DNA called cis-regulatory modules (CRMs), such as enhancers, silencer and promoters. CRMs control the activity of genes and usually contain numerous binding sites of different TFs [21–23].

Reconstruction of GRNs is crucial in systems biology for uncovering the regulatory mechanisms underlying biological processes. Essential cellular functions are driven by sophisticated interactions among genes, which regulate one another to ensure accurate gene product synthesis [24].

Developmental processes in particular are guided by regulation of genes responsible for producing transcription factors and cell signalling pathway components. The development of animal structures is governed by precise patterns of gene expression in the embryo, synchronized and location specific. Generally, complex patterns of gene expression require the interaction of multiple cis-regulatory modules [21, 25]. Given that multiple TFs can regulate multiple modules, and each module may be affected by multiple TFs, gene expression regulation during development forms a complicated network [21].

Computational tools are therefor essential to allow systemic analysis and understanding of complex and large scaled GRNs. This is particularly important given the increasing volume of data generated by advancements in experimental methods. These tools enable to investigate the behavior of GRNs under various conditions, assess their robustness, gain insight about their functionality in regards to their individual components and more [26].

Modeling: Mbodj et al. (2013) [27]

3.2.1 Logical Modeling of Gene Regulatory Networks

Several methods exist for modeling GRNs, notably logical and continuous modeling. Continuous modeling approaches, such as ordinary differential equations, use real-valued parameters over a continuous timescale to provide in-depth analyses of network behaviors. However, its use is limited to specific and usually smaller systems due to the need for extensive kinetic data and the computational constraints [26, 28]. In contrast, logical modeling, which is based on discrete functions, bypasses the requirement for kinetic parameters, allowing for qualitative, more global and computationally efficient

descriptions of system behaviors [28].

Logical models were first introduced by Kauffman in 1969, who used discrete logic to model the biological process of gene regulation [29]. It has become a key approach in systems biology, widely used for gene regulatory and signaling networks analysis. In multilevel logical modeling, the variables take a number of discrete states. However the most common form of this modeling uses Boolean networks, where variables represent the binary states of being inactive (0) or active (1). This discrete approach was shown to capture the essence of biological regulation, where interactions and regulatory effects typically take place only after the involved molecules reach a specific concentration threshold [30]. Thus, this modeling approach mirrors the natural "on" or "off" states of genes based on their regulatory conditions, without delving into the specific examples of regulator and target molecule dynamics [31, 32].

Formalism of GRNs logical modeling

Logical descriptions use variables with a discrete value, specifically 1 or 0 in the Boolean model case. The following formalism is based on the review done by Abou-Jaoudé et al. (2016) [33].

Logical Model Definition

A logical model (G, K) of a regulatory network is defined as follows:

- $G = \{g_1, g_2, ..., g_n\}$**:** A set of n regulatory components (e.g. genes), each g_i is associated with an integer variable with a value in the range $\{0, ..., max_i\}$, defining the level of activity. In Boolean models $max_i = 1$ for all i, hence variables are binary and equal either 0 or 1, representing "OFF" or "ON" activity of the component.

- **State space** $S = \prod_{i=1,...,n}\{0, ..., max_i\}$**:** The combination of all possible states of G, defined as the cartesian product of the ranges for each component g_i. The model state is a vector $g = (g_1, ..., g_n)$. The number of states in S is always finite.

- **Discrete transition function** $K : S \to S$**:** For each component g_i, a discrete function $K_i : S \to \{0, ..., max_i\}$ define its value depending on the model state. Therefor for a state vector g, $K(g) = (K_1(g), ..., k_n(g))$. K therefor defines the behaviour of the model. K_i functions are logical functions, function that use the logical operators AND, OR and NOT.

Regulatory Graph

The regulatory graph (G, R) representing a family of logical models. It consists of nodes (G) representing the regulatory components, and signed directed edges (R) representing activation or inhibition. The edges are defined by K, however it is worth mentioning that several logical rules sets can result in the same regulatory graph topology.

Logical Model Dynamics

A logical model defines discrete dynamics over its state space S. Given a state g, the transition function K specifies the possible changes of the model variables.

- **Initial state:** Represents the initial conditions of the system.
- **Attractor:** Maximal set of mutually reachable states with no transitions leaving the set.
  - **Stable State:** When an attractor consists of only one state, $K(g) = g$, meaning the next state is identical to the current state of the model. Each component value is maintained constant. Stable states can often represent a phenotype, like a cell differentiated state for example.
  - **Cyclic Attractor:** When an attractor consists of multiple states. May denote a biologically periodic behaviour (e.g cell cycle).
- **State Transition Graph**: A diagram of state vectors as nodes, and directed edges as transition to the next state according to the logical rules. It is usually practical to examine the whole state transition graph for smaller networks [34].

Synchronous vs Asynchronous Updating Schemes

Many times a state g has more than one variable to change its value. The order in which this happens is not trivial. The two most used updating schemes are *synchronous* and *asynchronous*. In synchronous update scheme all variables are updated simultaneously, hence there is an underlying assumption that all updates require the same time to take place, which is usually biologically unrealistic. Nonetheless it is widely used, due to its simplicity of application and interpretation, and low computational complexity. Alternatively, Thomas et al. ([31]) introduced an asynchronous updating scheme, in which each variable is updated independently, yielding a transition per update variable. This results with non deterministic dynamics. Its

advantage is it covers all possible transitions, including biologically meaningful ones. Its disadvantage is its more complex computationally, and harder to interpret, since not all states correspond to a biological meaning.

Related Work

Early *Drosophila* development has been the focus of a large number of dynamical modelling studies, however many times there are lacking kinetic parameters - "logical modeling of Drosophila signalling pathways" paper [27] .

3.2.2 Gene Regulatory Network Inference from Single-Cell RNA-seq Data

GRN inference is building a GRN network based on molecular data observations, under the assumption that the impacts of an actual underlying GRN are detectable and quantifiable within this data. [35]. GRNs are traditionally constructed based on experimentally confirmed regulatory relationships, often found in literature or compiled in databases. Since validating every part of a GRN is challenging, many networks remain small, presenting a compromise between accuracy and completeness [18, 21]. Beyond experimental validations, GRNs have also been inferred from gene co-expression in bulk transcriptomics data. This approach offers less general and more context-specific insights when sufficient data is available, however transcriptomics data alone is not always enough for accurate gene regulation inference, and sometimes is better when combined with other data types, like chromatin accessibility. Moreover, bulk profiling does not distinguish between regulatory processes across different cell types within the same tissue [35]. The latter challenge can be addressed using since-cell sequencing (SC-seq) technologies. SC-seq enables the inference of GRNs considering different cell types [36, 37]. This advancement has led to the development of numerous new methods for inferring GRNs [35].

Methods for inferring GRNs from Bulk Transcriptomics Data

Methods inferring regulatory relationships from transcriptomics data generally aim to explain the observed variability in the expression of each gene by considering the expression of other genes.

A popular and simple implementation of this approach is Weighted Gene Co-expression Network Analysis (WGCNA) [38], which detects groups of genes that are expressed together by performing pairwise correlations across all genes in the

transcriptome to form a co-expression network. However correlation alone produces an undirected network, which makes it hard to interpret as regulatory relationships between the genes and often results in many false positives [35].

GENIE3 [39] and its faster implementation GRNBoost2 [40] are co-expression based methods for inferring GRNs from gene expression data, that also incorporate prior knowledge from databases to distinguish TFs from target genes. This not only reduces the number of potential gene pairs to consider, but also yields directed results that are easier to interpret [35]. Moreover, both methods utilize ensemble-tree approaches, enabling them to capture complex regulatory relationships beyond linear correlations [41]. GENIE3 employs random forest models to predict the expression of each gene, using TF expressions as predictors. The models generate weights for the TFs, indicating their importance in regulating each target gene, with the highest weights suggesting regulatory links [42]. GRNBoost2 uses a gradient-boosting machines (GBM) algorithm instead of random forest, which, while similar in principle to GENIE3, offers the advantage of higher parallelism that significantly reduces runtime, making it more suitable for larger datasets [40].

Still, as summarized in the review of Badia-i-Mopel et al. (2023) [35], such unsupervised GRN inference methods based on transcriptomics data alone have only a moderate success in accurately inferring GRNs, primarily because they often overlook other regulatory mechanisms like chromatin accessibility, leading to numerous false positive results. Moreover, when these methods are based on bulk omics data, they are also not suitable for GRNs specifically for cell types or cell state.

Current Methods for inferring GRNs Single Cell Transcriptomics Data

Single-cell technologies and single-cell RNA sequencing (scRNA-seq) in particular enable to examine individual cell types within tissues and hence allows a more detailed analysis of the GRNs that guide different cellular processes, like differentiation and specification [43]. Suitable GRN reconstruction methods have been developed to infer cell type-specific TF–gene interactions [35].

Single-Cell Regulatory Network Inference and Clustering (SCENIC) [42] is one of the earliest GRN inference methods designed specifically for scRNA-seq data. SCENIC was implemented in R [42] and later also in Python [41].It consists of three main consecutive steps: Initial GRN inference, motif enrichment and cellular enrichment.

- **Co-expression modules constructions:** Co-expression modules of TFs and candidate target genes are inferred using GENIE3/GRNBoost mentioned above.

- **Motif enrichment:** Co-expression modules are refined into regulons. In essence, *i-cisTarget* method [44] is used to predict which genes are more likely to be directly regulated by the TF by using databases of genome-wide, cross-species motif rankings, to check if TF motifs are over-represented near the transcription start sites of genes in the inferred module.

- **Cellular enrichment:** *AUCell* method [42] evaluates the enrichment of a regulon in individual cells. It ranks each cell's genes based on expression levels, and measures for each regulons its relative activity compared to the other genes within the cell. This score can be used in various ways, for example for clustering the single cells, or for generating binary scores of activity with AUCell score threshold based on the scores distribution across the whole dataset.

Nguyen et al. (2020) [43] conducted a survey in which they compared by simulation 15 available methods of GRN inference methods that use scRNA-seq data, including SCENIC. The comparison is based on three metrics: accuracy in reconstruction reference networks, sensitivity to different levels of dropout/sparsity of the given data, and time complexity. SCENIC to outperformed the other methods regarding the accuracy, and came in the top three in terms of time complexity and sensitivity. It was relatively effective in analyzing datasets with a high degree of sparsity compared to most other methods, but with relatively more variability in results as sparsity rises.

4 Methodology

Single-cell RNA-sequencing Dataset

The scRNA-seq dataset used in this project was generated by Ana Veloso from the Zinzen lab, MDC-BIMSB, and was pre-processed by Yozlem Bahar from the Wolf lab in MDC. The dataset was generated using 10X Genomics approach, by a protocol developed by Veloso (2022) to specifically target NBs cells of the Drosophila embryos in the sequencing. The dataset contains samples from two partially overlapping time points that span developmental stages 8 to mid-stage 10, in which the NBs delamination waves S1-S3 take place.

The dataset underwent both cell-level and gene-level filtering by Bahar. Additionally, Bahar annotated the dataset for NBs cell types by clustering and identifying marker genes. For further details, see Bahar (2024).

4.1 GRN inference from scRNA-seq Dataset

SCENIC workflow

A GRN was infered from the dataset using pySCENIC (Van de Sande et al., 2020 [41]) v.0.12.1, a Python implementation of the SCENIC workflow, available as a Python package. A Python script was written to run the SCENIC workflow. The full script will be made publicly available in Github - add link here. The script utilizes the three main steps of SCENIC in a pipeline manner, where the output of one step is the input of the next.

Here, the SCENIC workflow used the *GRNBoost2* [40] method for the initial co-expression modules construction, followed by the motif enrichment (*i-cisTarget*) and cellular enrichment (*AUCell*) steps, called here also the first, second and third step respectively.

Additional input used in the SCENIC workflow

SCENIC workflow was run sequentially where one step's output is the input for the next one. The first step requires a gene expression matrix, which here is the counts matrix of the pre-processed dataset (without any annotations, like cell types). Additionally, SCENIC incorporates input from databases for the first and second step, as follows:

- **Step 1:** co-expression modules construction (GRNBoost2)
  - **Species specific list of candidate TFs** A list of all known TFs in Drosophila
- **Step 2:** Motif enrichment (i-cisTarget)
  - **Motif ranking database** Drosophila Whole genome rankings for regulatory motifs
  - **Motif Annotations database** Drosophila regulatory motifs annotations

The databases were sourced from the Stein Aerts Lab website (`https://resources.aertslab.org/cistarget`), which utilises data from FlyBase [45], a database for Drosophila genetics and molecular biology, along with databases related to other species. When not available, one can create custom databases suitable for SCENIC (see `https://scenic.aertslab.org`).

Multiple iterations

As described in the background section, GRNBoost2 is stochastic, which means that every time the workflow is run, the inferred GRN is somewhat different. To address this variability, the workflow was executed 50 times, following the recommendation of Van de Sande et al. (2020) [41]. As illustrated in fig. 1, regulons that appear in more than 80% of the runs ($>$ 40 runs) were considered for further analysis of the GRN. In more detail, a list of regulons is produced after the motif enrichment step in the SCENIC workflow, and each each regulon consists of a TF and an accompanying list of target genes. TFs (the head of regulons) were aggregated across runs, and those appearing in over 40 runs were saved as a subset for further analysis. For convenience purposes, the persistent regulons that were common for over 80% of the iterations will be referred to as *core regulons.*

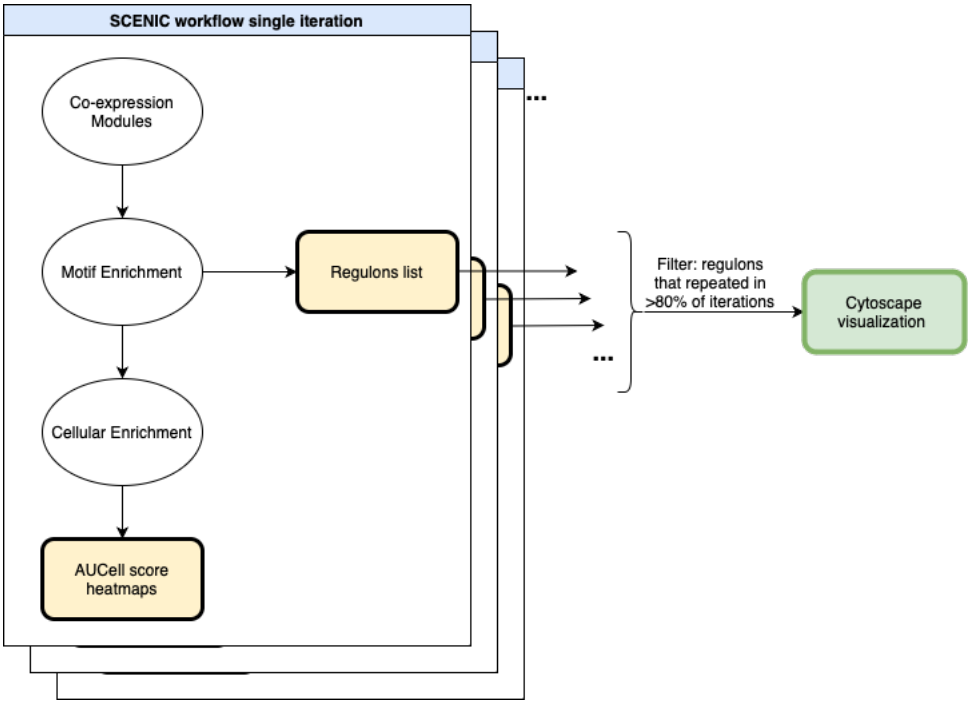

Figure 1: Illustration for multiple iterations of SCENIC workflow

Cellular enrichment clustering and visualisation in heatmaps

The third and last step of SCENIC is cellular enrichment, in which each regulon gets an AUCell score that symbolises its activity in each of the cells in the dataset. In each SCENIC iteration, a matrix of $cells \times regulons$ is produced, with the AUCell scores. This matrix was clustered and visualized in a heatmap using Python's *seaborn.clustermap* hierarchical clustering method. Additionally, cell types annotations of the dataset were added to the heatmap.

From this AUCell score heatmap .. more heatmaps were produced:

- **Binary heatmaps:** to binarize the AUCell score, a two-component Gaussian mixture model is fitted to define a threshold. This threshold labels cells as 'on' or 'off' for each regulon. When the AUCell distribution across cells is skewed, the threshold is set along the right tail of the distribution [41]. Thresholds can be tuned or set manually according to biological knowledge, however here the binarization was not manually changed, due to lack of biological knowledge.
- Reduced heatmap for the persistent regulons
- AUCell score heatmap grouped by average AUCell score per cell type.

4.2 GRN Visualization

The SCENIC-inferred GRN is inferred after the motif enrichment step of SCENIC. The subset of regulons that repeated across over 40 runs were extracted and exported as a CSV file containing TF-target genes pairs, and their occurrence count across runs. This CSV was visualized as a network using Cytoscape software (Shannon et al., 2003 [46]) v.3.10.0. The network was contains TFs and target genes as nodes, with directional edges from a TF to its target genes. Some TFs are also target genes of other TFs. TFs and target genes were styled differently for better readability by uploading an additional list of annotations of genes as TFs or target genes. Within Cytoscape, the count of reoccurring TF-target gene across iterations was captured as the thickness of the regulatory edges. Thicker edge shows higher occurrence across SCENIC iterations. The GRN was then filtered to show the known marker genes of NBs, and their nearest neighbors.

4.3 Boolean Model Analysis

Boolean model of signaling mechanism if early Drosophila nervous system development

A Boolean model of the signaling mechanism in Drosophila during early development, that was developed by Yozlem Bahar from the Wolf lab at MDC, was analysed here. Fig. 2 illustrates its topography. The model consists of 30 nodes and 57 directed edges. Nodes represent genes and edges represent a regulatory relationship of transcription activation (green edges) or suppression (red edges). This model is based on literature knowledge about signalling pathways, and marker genes that characterize different NBs types. The yellow nodes are input nodes that represent signaling pathways, namely Hh, Wg and Egfr. Gray nodes represent marker genes. Round nodes represent NB cell types. Six NBs types are included in this model, namely 5-6, 5-3, 5-2, 6-2, 7-4, 7-1. The model with the full logical rules is available in the supplementary materials.

4.3.1 Stable state analysis

Gene Interaction Network simulation (GINsim), Gonzalez et al. (2006) [47] is a software designed to edit, simulate and analyse logical models of regulatory networks. Here, GINsim v.3.0 [34] was used to calculate the stable states, or fixed points, of the Boolean model in both the graphic user interface and as a python package, available

via CoLoMoTo [48], a Python interface developed for various tools for the construction and analysis of logical models. Stable states were calculated in GINsim to reproduce to the stable states analysis done by Bahar et al.

Correlation of Variables

Stable states tables were further analysed by calculating the pairwise correlation between the different variables. Kendall rank correlation was used.

Perturbation Analysis

Knock out perturbation was done by iteratively setting one node at a time to 0, and calculating the stable states with that perturbation.

4.4 Boolean Model Extension with GRN Inferred from Single Cell Sequencing Data

The Boolean model was compared to the GRN filtered in different ways, as will be shown in the results (section 5). The GRN in Cytoscape filtered to contain the marker genes, which appear in the Boolean model, and their direct neighbors, was compared to the model. New regulatory relationships that include nodes or edges that did not appear in the model but are connected directly according to the SCENIC GRN, were searched for validation in literature and in FlyBase database. New nodes and edges with defined logical rules were added to the Boolean model based on literature. The extended model underwent stable state analysis and compared to the original model.

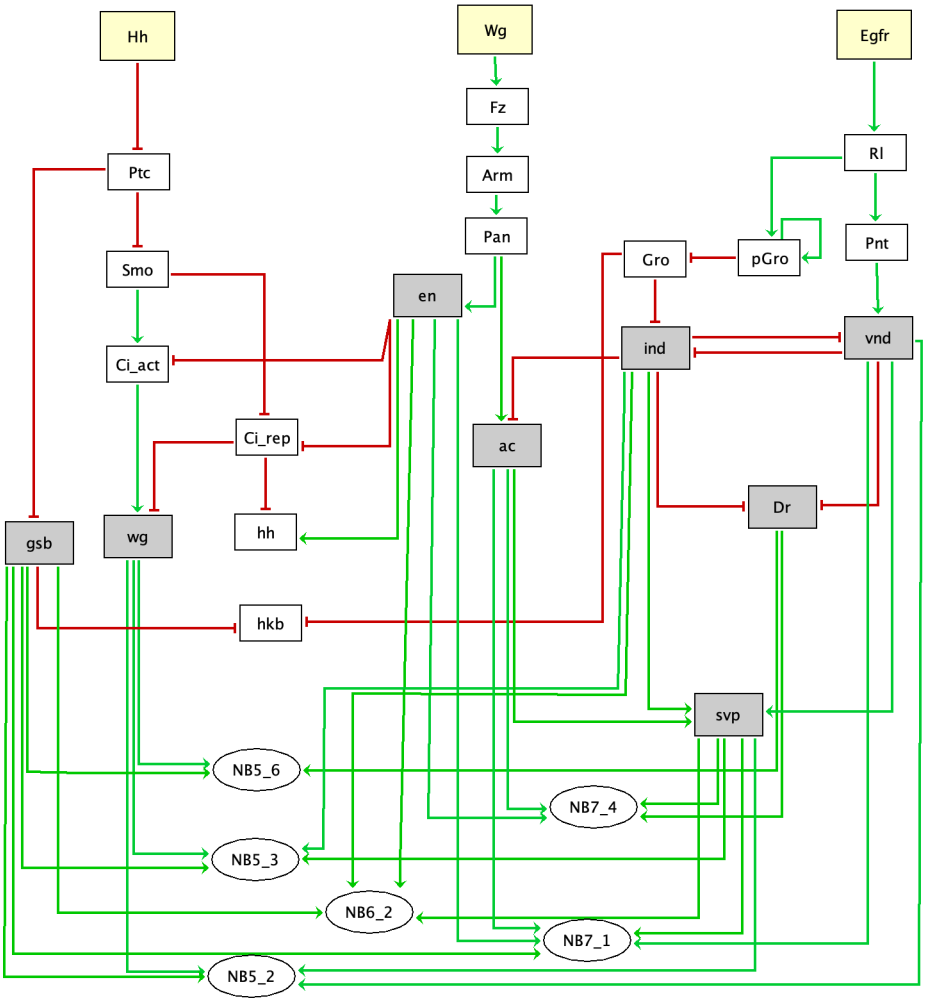

Figure 2: Boolean model of Drosophila GRN during neuroblasts specification. Developed by Yozlem Bahar, Wolf lab, MDC (2024)

5 Results

scRNA-seq dataset

The pre-processed dataset used here contains 3855 genes and 9751 cells. There are 15 different cell types. Among these, 14 types are NBs and one is Midline Glia (MG) which is a precursor for Glia cells. Fig. 3 displays the number of cells per type, arranged in ascending order from the fewest to the most. Numbers represent NBs (i.e 3-5 is NB3-5).

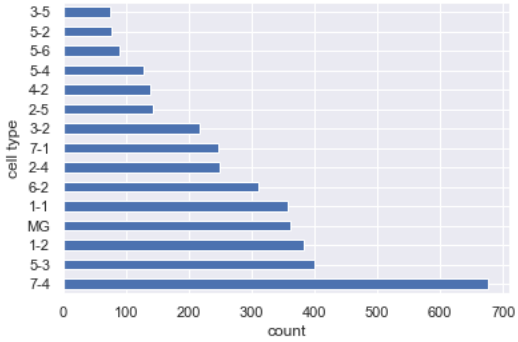

Figure 3: Pre-processed scRNA-seq dataset cell types count

5.1 SCENIC Workflow

Inferring a GRN

As explained in Methodology (section 4), the SCENIC workflow was executed repeatedly 50 times and each run produced a list of predicted regulons, and an AUCell cell enrichment score for each individual cell per regulon.

For each SCENIC iteration, after motif enrichment, a list of regulons was saved in a CSV file, with each regulon composed of one TF and its target genes. Regulons are named after the TF at their head. The occurrences of TFs were aggregated across all runs, identifying 352 unique TFs. The histogram in fig. 4 shoes how frequently each TF was repeated across different runs. The x-axis represents the number of

run in which a TF appears, ranging from 1 to 50 (with bin size of five). The y-axis indicates the number of TFs that appears in that specific number of runs. The red line indicates the threshold set at 80% of runs (>40 iterations); 38 TFs repeated in over 40 iterations in regulons and were selected for further analysis, with 3330 unique genes. It is important to note that between different iterations, regulons with the same TF still differ in target genes composition.

To check the variability of the genes within the 38 regulons of the most occurring TFs across iterations, each regulon was spread to a list of TF-target couples, and the occurrences of these couples across iterations was counted. Similarly to fig, 4, the histogram in fig. 5 shows the frequency of number of repetitions among these TF-target couples.

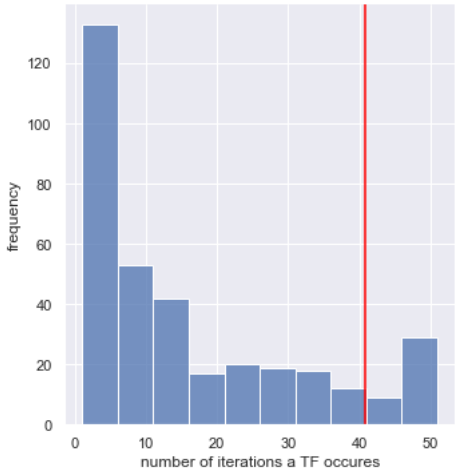

Figure 4: Frequency of repeating TFs across SCENIC iterations

In both figure it is clear that the variability between different iterations was high, since it is more frequent for TFs and TFs-target couples are less repetitive between runs.

5.1.1 Cellular Enrichment Score Clustering

AUCell score clustering

Each iteration produces a different AUCell score that is clustered and visualized in a heatmap. Fig. 6 is an example for such heatmap for one iteration. Columns are individual cell and rows are regulons. Note that regulons are named after their TF. The (+) sign attached to the regulons names symbolises activation regulatory relationship between the TF and its targets. The cells types annotation are attached

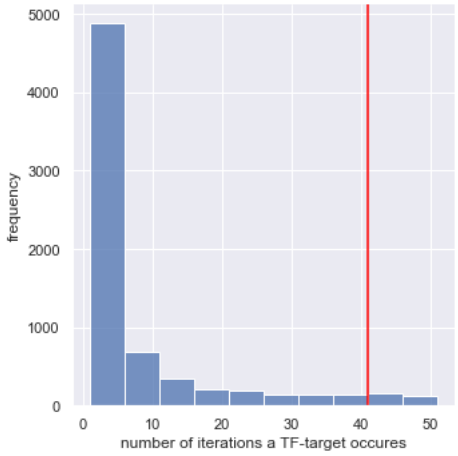

Figure 5: Frequency of repeating TF-target gene across SCENIC iterations

as the top row. In this specific iteration, there were 100 predicted regulons in total. In fig. 6 the cells are generally not clustered clearly by cell type, nonetheless some MG cells appear to cluster together, as well as some 2-5 NB cells. This was repetitive throughout the vast majority of iterations (the rest of the heatmaps are available in supplementary materials). Among the regulons that have relatively high AUCell scores in MG cells in fig. 6 are transcription factors from the enhancer of split gene complex ($E(spl)mbeta, m3, m5, m8 - HLH$). According to FlyBase [45], these TFs are known to take part in neuronal differentiation. Additionally, Myc, aop and $Mef2$ regulons were clustered with repeatedly high AUCell scores for MG cells in most iterations.

AUCell score clustering of most repetitive TFs

AUCell score clustering with binarization

AUCell score clustering grouped by cell type

The heatmap in fig. 9 was based on the same AUCell scores as in fig. 7, and was grouped by cell type. The value for each cell type is the average of AUCell scores for all of the cells of that same cell type. To improve the readability of the heatmap and to enhance the visual contrasts, centering around zero and z-score scaling per row was applied.

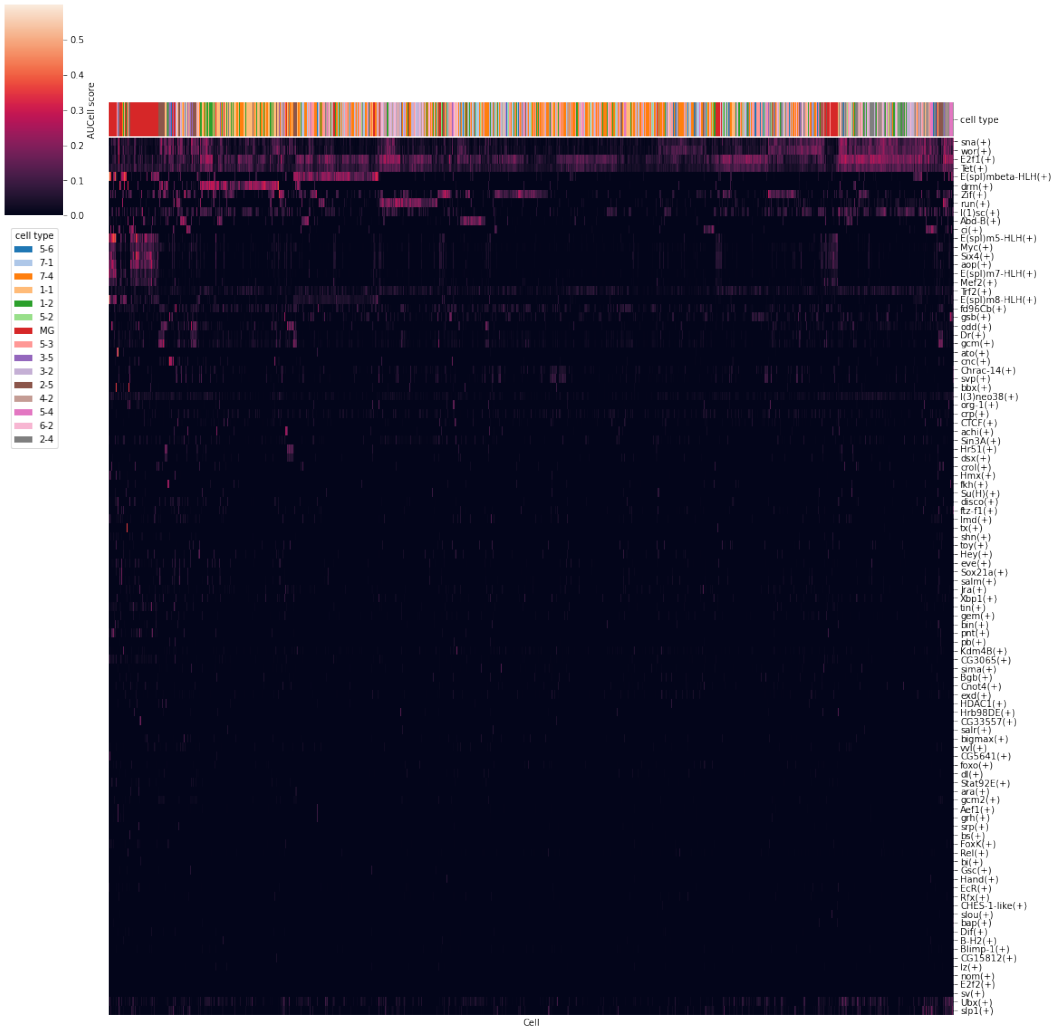

Figure 6: AUCell scores of all regulons

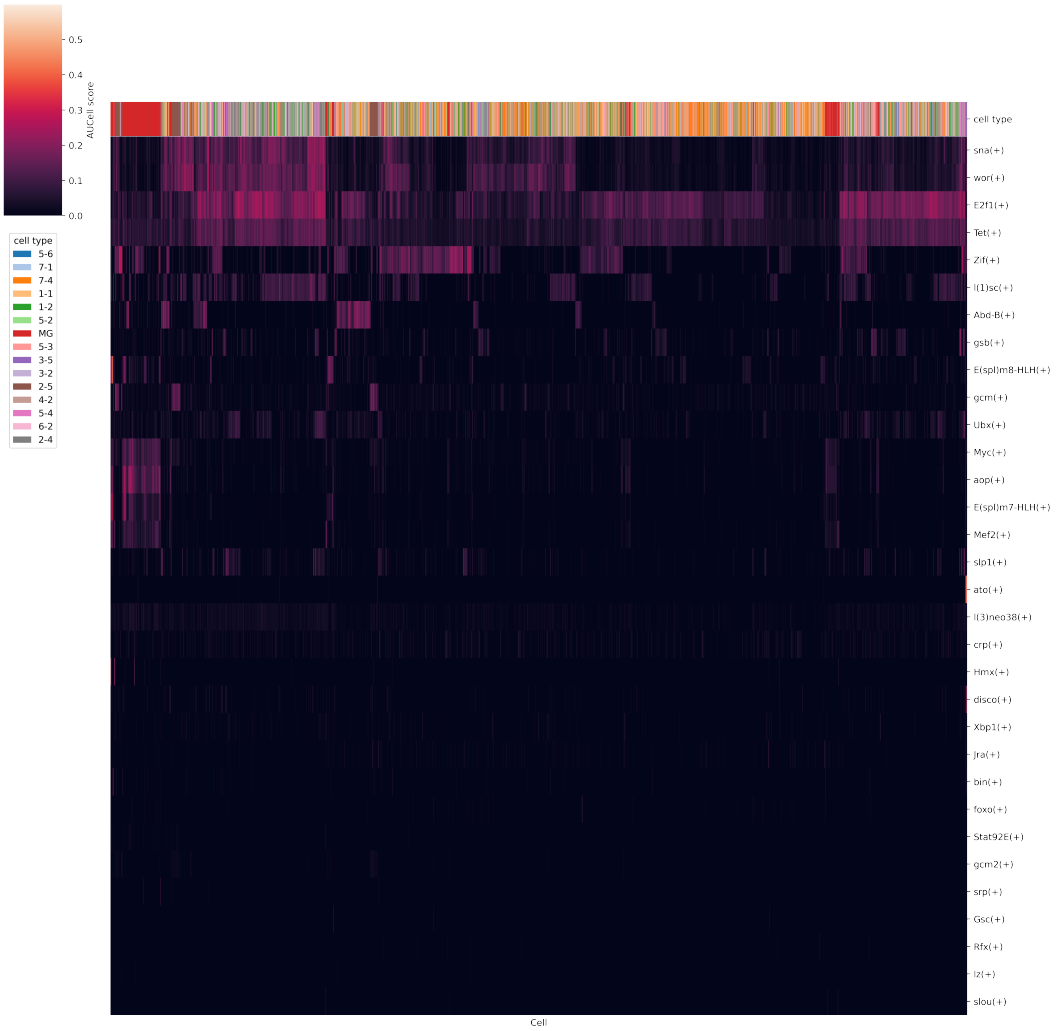

Figure 7: AUCell scores of core regulons

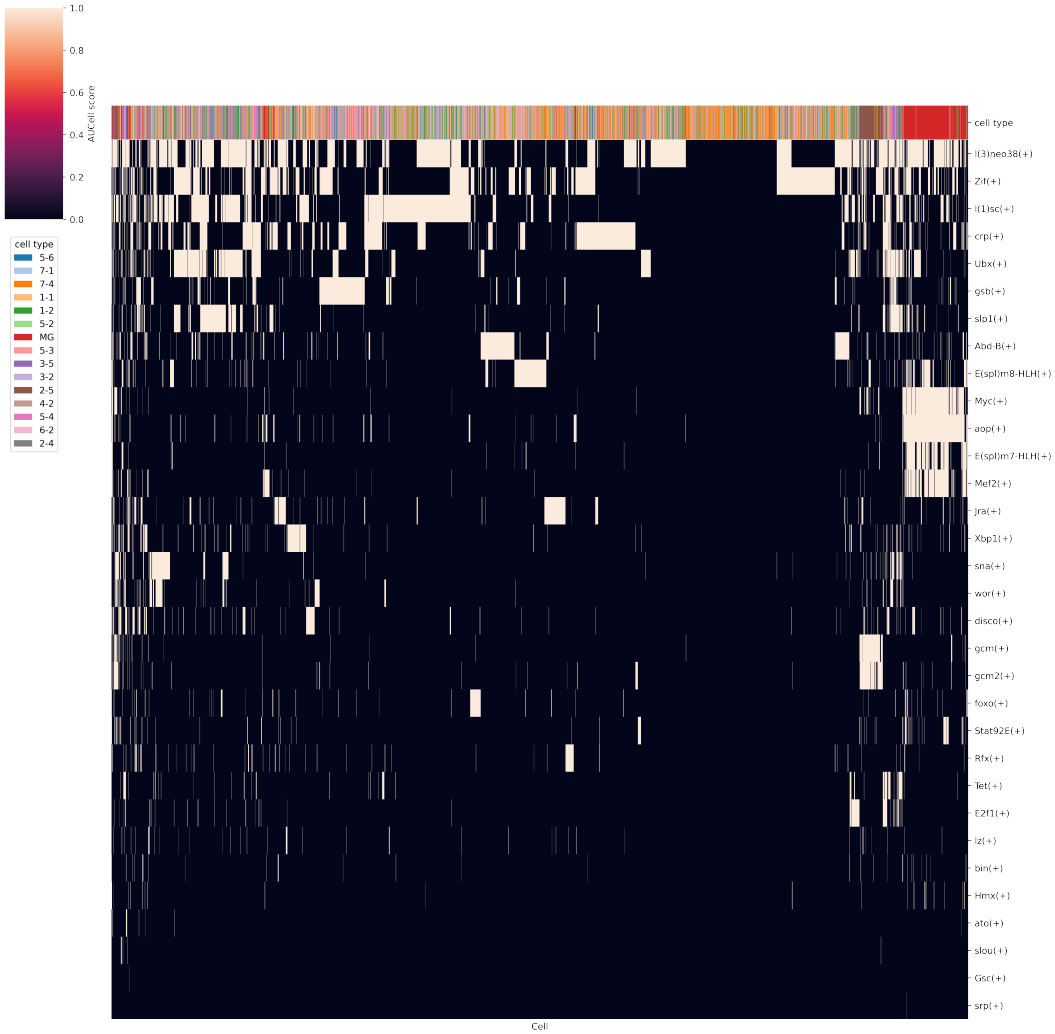

Figure 8: Binary AUCell scores of core regulons

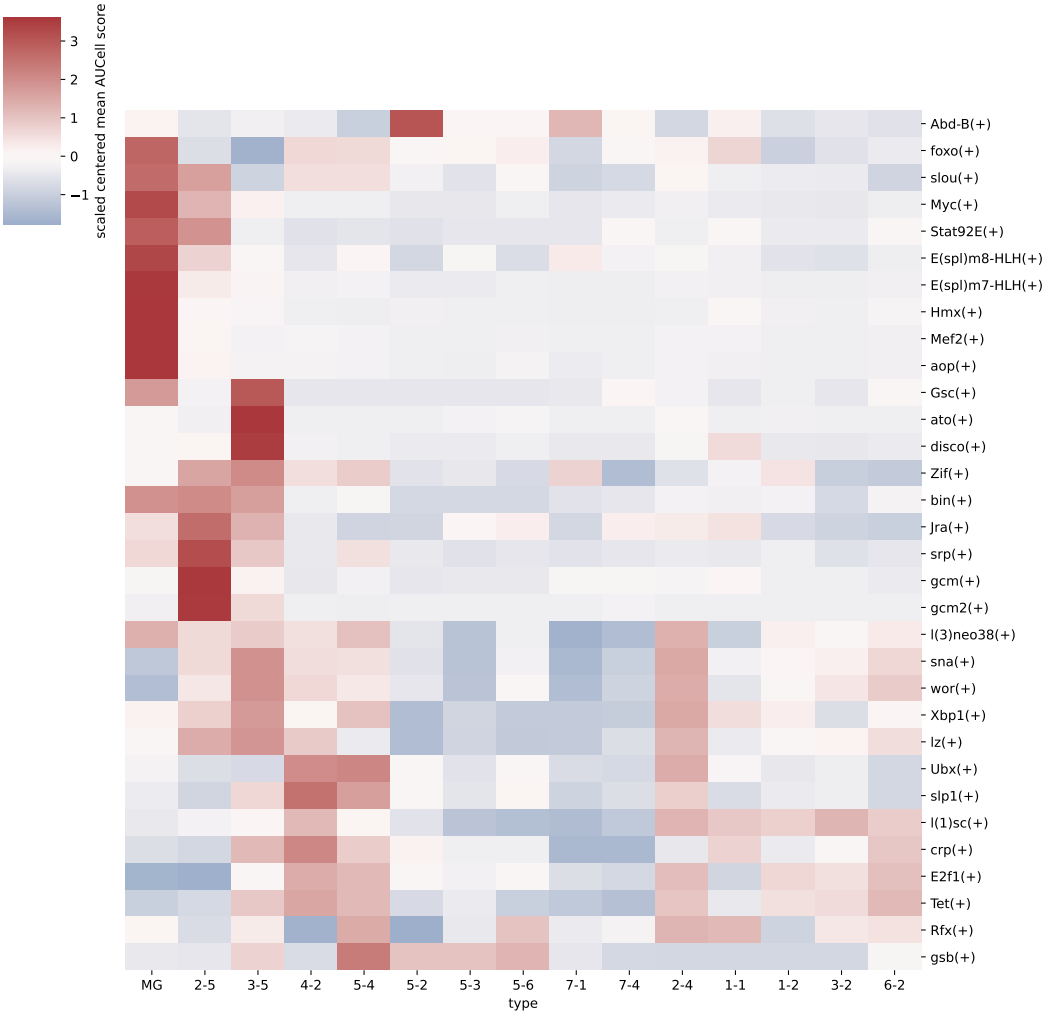

Figure 9: Average AUCell scores of core regulons for different cell types

5.1.2 Cytoscape GRN Visualization

Core TF-Target couples GRN

Most Repetitive TFs, with all targets GRN

Here the thickness of the edges represents the repetitiveness of the prediction throughout SCENIC iterations.

5.2 Boolean Model Analysis

5.3 Boolean Model Extension with GRN Inferred from Single Cell Sequencing Data

6 Conclusion

7 Acknowledgments

First and foremost, I would like to thank...

- advisers
- examiner
- person1 for the dataset
- person2 for the great suggestion
- proofreaders

Bibliography

[1] S. T. Crews, "Drosophila embryonic cns development: Neurogenesis, gliogenesis, cell fate, and differentiation," *Genetics*, vol. 213, pp. 1111–1144, 2019.

[2] K. Ito, J. Urban, and G. M. Technau, "Distribution, classification, and development of drosophila glial cells in the late embryonic and early larval ventral nerve cord," *Roux's archives of developmental biology*, vol. 204, pp. 284–307, 1995.

[3] E. S. Heckscher, F. Long, M. J. Layden, C.-H. Chuang, L. Manning, J. Richart, J. C. Pearson, S. T. Crews, H. Peng, E. Myers, *et al.*, "Atlas-builder software and the eneuro atlas: resources for developmental biology and neuroscience," *Development*, vol. 141, no. 12, pp. 2524–2532, 2014.

[4] I. M. Cobeta, B. Y. Salmani, and S. Thor, "Anterior-posterior gradient in neural stem and daughter cell proliferation governed by spatial and temporal hox control," *Current Biology*, vol. 27, no. 8, pp. 1161–1172, 2017.

[5] B. Yaghmaeian Salmani, I. Monedero Cobeta, J. Rakar, S. Bauer, J. R. Curt, A. Starkenberg, and S. Thor, "Evolutionarily conserved anterior expansion of the central nervous system promoted by a common pcg-hox program," *Development*, vol. 145, no. 7, p. dev160747, 2018.

[6] R. Urbach, D. Jussen, and G. M. Technau, "Gene expression profiles uncover individual identities of gnathal neuroblasts and serial homologies in the embryonic cns of drosophila," *Development*, vol. 143, no. 8, pp. 1290–1301, 2016.

[7] T. Bossing, G. Udolph, C. Q. Doe, and G. M. Technau, "The embryonic central nervous system lineages of drosophila melanogaster: I. neuroblast lineages derived from the ventral half of the neuroectoderm," *Developmental biology*, vol. 179, no. 1, pp. 41–64, 1996.

[8] J. B. Skeath, "At the nexus between pattern formation and cell-type specification: the generation of individual neuroblast fates in the drosophila embryonic central nervous system," *BioEssays*, vol. 21, pp. 922–931, 10 1999.

[9] K. M. Bhat, "Segment polarity genes in neuroblast formation and identity specification during drosophila neurogenesis," *BioEssays*, vol. 21, pp. 472–485, 6 1999.

[10] H. Schmidt, C. Rickert, T. Bossing, O. Vef, J. Urban, and G. M. Technau, "The embryonic central nervous system lineages of drosophila melanogaster," *Developmental biology*, vol. 189, no. 2, pp. 186–204, 1997.

[11] J. A. McDonald and C. Q. Doe, "Establishing neuroblast-specific gene expression in the drosophila cns: huckebein is activated by wingless and hedgehog and repressed by engrailed and gooseberry," *Development*, vol. 124, no. 5, pp. 1079–1087, 1997.

[12] J. B. Skeath and S. Thor, "Genetic control of drosophila nerve cord development," *Current Opinion in Neurobiology*, vol. 13, pp. 8–15, 2 2003.

[13] C. Q. Doe, "Molecular markers for identified neuroblasts and ganglion mother cells in the drosophila central nervous system," *Development*, vol. 116, no. 4, pp. 855–863, 1992.

[14] D. M. Mellerick and M. Nirenberg, "Dorsal-ventral patterning genes restrict nk-2 homeobox gene expression to the ventral half of the central nervous system of drosophila embryos," *Developmental biology*, vol. 171, no. 2, pp. 306–316, 1995.

[15] F. Jimenez, L. E. Martin-Morris, L. Velasco, H. Chu, J. Sierra, D. Rosen, and K. White, "vnd, a gene required for early neurogenesis of drosophila, encodes a homeodomain protein.," *The EMBO journal*, vol. 14, no. 14, pp. 3487–3495, 1995.

[16] M. D'Alessio and M. Frasch, "Msh may play a conserved role in dorsoventral patterning of the neuroectoderm and mesoderm," *Mechanisms of development*, vol. 58, no. 1-2, pp. 217–231, 1996.

[17] J. B. Weiss, T. Von Ohlen, D. M. Mellerick, G. Dressler, C. Q. Doe, and M. P. Scott, "Dorsoventral patterning in the drosophila central nervous system: the intermediate neuroblasts defective homeobox gene specifies intermediate column identity," *Genes & development*, vol. 12, no. 22, pp. 3591–3602, 1998.

[18] B. Wilczynski and E. E. Furlong, "Challenges for modeling global gene regulatory networks during development: Insights from drosophila," 2010.

[19] D. Mercatelli, L. Scalambra, L. Triboli, F. Ray, and F. M. Giorgi, "Gene regulatory network inference resources: A practical overview," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1863, no. 6, p. 194430, 2020.

[20] D. S. Latchman, "Transcription factors: an overview.," *International journal of experimental pathology*, vol. 74, no. 5, p. 417, 1993.

[21] M. Levine and E. H. Davidson, "Gene regulatory networks for development," *Proceedings of the National Academy of Sciences*, vol. 102, no. 14, pp. 4936–4942, 2005.

[22] S. Small, A. Blair, and M. Levine, "Regulation of even-skipped stripe 2 in the drosophila embryo.," *The EMBO journal*, vol. 11, no. 11, pp. 4047–4057, 1992.

[23] E. H. Davidson, *Genomic regulatory systems: in development and evolution.* Elsevier, 2001.

[24] L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris, and Z. Zakaria, "A review on the computational approaches for gene regulatory network construction," *Computers in biology and medicine*, vol. 48, pp. 55–65, 2014.

[25] S. Gray, P. Szymanski, and M. Levine, "Short-range repression permits multiple enhancers to function autonomously within a complex promoter.," *Genes & development*, vol. 8, no. 15, pp. 1829–1838, 1994.

[26] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature reviews Molecular cell biology*, vol. 9, no. 10, pp. 770–780, 2008.

[27] A. Mbodj, G. Junion, C. Brun, E. E. Furlong, and D. Thieffry, "Logical modelling of drosophila signalling pathways," *Molecular BioSystems*, vol. 9, no. 9, pp. 2248–2258, 2013.

[28] R.-S. Wang, A. Saadatpour, and R. Albert, "Boolean modeling in systems biology: an overview of methodology and applications," *Physical biology*, vol. 9, no. 5, p. 055001, 2012.

[29] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of theoretical biology*, vol. 22, no. 3, pp. 437–467, 1969.

[30] L. Glass and S. A. Kauffman, "Co-operative components, spatial localization and oscillatory cellular dynamics," *Journal of theoretical biology*, vol. 34, no. 2, pp. 219–237, 1972.

[31] R. Thomas, "Regulatory networks seen as asynchronous automata: a logical description," *Journal of theoretical biology*, vol. 153, no. 1, pp. 1–23, 1991.

[32] R. Samaga and S. Klamt, "Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks," *Cell communication and signaling*, vol. 11, pp. 1–19, 2013.

[33] W. Abou-Jaoudé, P. Traynard, P. T. Monteiro, J. Saez-Rodriguez, T. Helikar, D. Thieffry, and C. Chaouiya, "Logical modeling and dynamical analysis of cellular networks," *Frontiers in genetics*, vol. 7, p. 188073, 2016.

[34] A. Naldi, C. Hernandez, W. Abou-Jaoudé, P. T. Monteiro, C. Chaouiya, and D. Thieffry, "Logical modeling and analysis of cellular regulatory networks with ginsim 3.0," *Frontiers in physiology*, vol. 9, p. 646, 2018.

[35] P. Badia-i Mompel, L. Wessels, S. Müller-Dott, R. Trimbour, R. O. Ramirez Flores, R. Argelaguet, and J. Saez-Rodriguez, "Gene regulatory network inference in the era of single-cell multi-omics," *Nature Reviews Genetics*, vol. 24, no. 11, pp. 739–754, 2023.

[36] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, no. 5, pp. 1187–1201, 2015.

[37] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, *et al.*, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, 2015.

[38] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, pp. 1–13, 2008.

[39] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PloS one*, vol. 5, no. 9, p. e12776, 2010.

[40] T. Moerman, S. Aibar Santos, C. Bravo González-Blas, J. Simm, Y. Moreau, J. Aerts, and S. Aerts, "Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks," *Bioinformatics*, vol. 35, no. 12, pp. 2159–2161, 2019.

[41] B. Van de Sande, C. Flerin, K. Davie, M. De Waegeneer, G. Hulselmans, S. Aibar, R. Seurinck, W. Saelens, R. Cannoodt, Q. Rouchon, *et al.*, "A scalable scenic workflow for single-cell gene regulatory network analysis," *Nature protocols*, vol. 15, no. 7, pp. 2247–2276, 2020.

[42] S. Aibar, C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, *et al.*, "Scenic: single-cell regulatory network inference and clustering," *Nature methods*, vol. 14, no. 11, pp. 1083–1086, 2017.

[43] H. Nguyen, D. Tran, B. Tran, B. Pehlivan, and T. Nguyen, "A comprehensive survey of regulatory network inference methods using single cell rna sequencing data," *Briefings in bioinformatics*, vol. 22, no. 3, p. bbaa190, 2021.

[44] C. Herrmann, B. Van de Sande, D. Potier, and S. Aerts, "i-cistarget: an i ntegrative genomics method for the prediction of regulatory features and cis-regulatory modules," *Nucleic acids research*, vol. 40, no. 15, pp. e114–e114, 2012.

[45] J. Thurmond, J. L. Goodman, V. B. Strelets, H. Attrill, L. S. Gramates, S. J. Marygold, B. B. Matthews, G. Millburn, G. Antonazzo, V. Trovisco, *et al.*, "Flybase 2.0: the next generation," *Nucleic acids research*, vol. 47, no. D1, pp. D759–D765, 2019.

[46] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[47] A. G. Gonzalez, A. Naldi, L. Sanchez, D. Thieffry, and C. Chaouiya, "Ginsim: a software suite for the qualitative modelling, simulation and analysis of regulatory networks," *Biosystems*, vol. 84, no. 2, pp. 91–100, 2006.

[48] A. Naldi, C. Hernandez, N. Levy, G. Stoll, P. T. Monteiro, C. Chaouiya, T. Helikar, A. Zinovyev, L. Calzone, S. Cohen-Boulakia, *et al.*, "The colomoto interactive notebook: accessible and reproducible computational analyses for qualitative biological networks," *Frontiers in physiology*, vol. 9, p. 680, 2018.