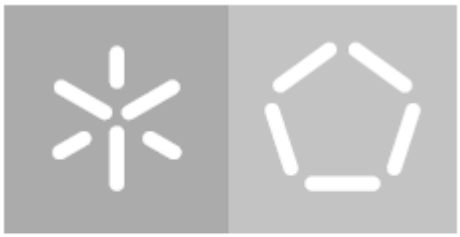
Universidade do Minho
Escola de Engenharia
Departamento de Informática

Joana Maria Martins Ribeiro

Transcriptional regulation of neurogenesis by the proneural factor Ascl1

June 2022



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Joana Maria Martins Ribeiro

Transcriptional regulation of neurogenesis by the proneural factor Ascl1

Master dissertation

Master Degree in Bioinformatics

Dissertation supervised by

Professor Doutor Miguel Francisco de Almeida Pereira da Rocha

Doutor Diogo Pinto da Cruz Sampaio e Castro

June 2022

ACKNOWLEDGEMENTS

I want to thank Professor Doutor Miguel Rocha for giving me support during this dissertation. I also want to thank Doutor Diogo S. Castro for allowing me to be part of his team at i3S and for giving all the support and orientation throughout the development of this work. I want to thank Lígia Tavares for sharing her knowledge, addressing my questions and unconditional help. My thanks to Rúben Rodrigues for always being available to my questions regarding coding and for the helpful discussions about R packages. I want to thank Abeer Heskol for providing me support with the chromatin state model. Also, I want to thank Inês Coutinho, Raquel Marques, Mário Soares, and Miguel Silva for all the support and fellowship during my stay in the laboratory.

Finally, I want to thank my partner João and my parents for their unlimited support and for making this journey possible.

This thesis is dedicated to my son Gil.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

This project aims to provide a better understanding of the transcriptional regulation of neurogenesis by the proneural factor Ascl1. The first genome-wide characterization of Ascl1 transcriptional program in the embryonic mouse brain was performed by ChIP-chip. However, the restriction to proximal promoter regions, excluding genes bound by Ascl1 to distal enhancers, and the need to validate the model with a more robust experimental approach, prompted the use of ChIP-seq. Genome-wide mapping of Ascl1 binding sites with higher resolution, reveals 3054 high confidence binding regions in ventral telencephalon. The chromatin states of genomic regions associated with Ascl1 recruitment were also characterised, concluding that these bear marks of distal enhancers, but also proximal promoter regions. Further integration of expression profiling data from Ascl1 LoF experiments identifies 643 target genes. Results from functional annotation of these targets corroborate previous findings, showing that Ascl1 coordinates neurogenesis by regulating a large number of target genes with a wide variety of biological functions, and associated with different stages of neurogenesis. Additional investigations should address how Ascl1 coordinates this complex transcriptional program along the neuronal lineage. This could explore a possible crosstalk with the Notch program, taking advantage of the 105 regulatory regions identified where Ascl1 is co-recruited by RBPJ, as assessed by ChIP-seq.

Key words: Neurogenesis, Ascl1, chromatin immunoprecipitation followed by sequencing (ChIP-seq), chromatins states, target genes, RBPJ

RESUMO

O objetivo principal deste projeto consiste em compreender melhor a regulação transcricional da neurogénese pelo fator proneural Ascl1. A primeira caracterização à escala do genoma do programa de transcrição do Ascl1 no cérebro de embriões de ratinho foi realizada pela técnica de ChIP-chip. No entanto, a restrição a regiões próximas do promotor, com exclusão de genes ligados pelo Ascl1 a *distal enhancers*, e a necessidade de validar o modelo com uma abordagem experimental mais robusta, motivou o recurso à técnica de ChIP-seq. A análise de localização, com alta resolução, ao longo de todo o genoma para sítios de ligação do Ascl1, revelou 3054 regiões de ligação de elevada confiança no telencéfalo do ratinho. De seguida, caracterizaram-se os *chromatin states* de regiões genómicas associadas com o recrutamento do Ascl1. Desta análise conclui-se que estas regiões possuem marcas de *distal enhancers*, mas também de regiões próximas do promotor. A posterior integração de perfis de expressão em experiências de perda-de-função para o Ascl1 identificou 643 genes alvo. Os resultados da anotação funcional desses alvos corroboram as conclusões anteriormente publicadas, mostrando que o Ascl1 coordena a neurogénese através da regulação de um grande número de genes alvo, com uma ampla diversidade de funções biológicas, associados a diferentes fases da neurogénese. Estudos futuros deem abordar de que forma o Ascl1 coordena este programa de transcrição complexo ao longo da linhagem neuronal. Tal poderia explorar um possível *crosstalk* com o programa Notch, tirando partido das 105 regiões regulatórias identificadas por ChIP-seq, onde o Ascl1 é co-recrutado pelo RBPJ.

Palavras-chave: Neurogénese, Ascl1, imunoprecipitação de cromatina seguida de sequenciação (ChIP-seq), *chromatin states*, genes alvo, RBPJ

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

B

Basic helix-loop-helix (bHLH)

C

Central nervous system (CNS)

Chromatin immunoprecipitation followed by sequencing (ChIP-seq)

Chromatin immunoprecipitation followed by DNA microarrays (ChIP-chip)

F

Forward (FW)

I

Intermediate progenitors (IP)

Irreproducible Discovery Rate (IDR)

N

Neural stem (NS)

Neural progenitor (NP)

P

Protein-of-interest (POI)

R

Radial glial (RG)

Reverse (RV)

T

Transcription start sites (TSS)

Transcription factor (TF)

1 INTRODUCTION

1.1 Motivation

Neurogenesis is a developmental process whereby fully functional neurons are generated from multipotent neural stem cells. It requires a fine balance between gene expression programs that regulate self-renewal and differentiation of neural progenitors and, is to large extent, regulated by proneural transcription factors such as Ascl1 (Vasconcelos & Castro, 2014).

Previously, the first genome-wide characterization of Ascl1 transcriptional program in the embryonic mouse brain was performed by combining gene expression profiling with chromatin immunoprecipitation, followed by hybridization to promoter oligonucleotide arrays (ChIP-chip) (Castro et al., 2011). This work revealed a set of Ascl1 target genes which regulates different steps in the neurogenic program. Additionally, it identified a novel and unexpected function of Ascl1 in sustaining progenitor proliferation. Together, these findings suggest that Ascl1 promotes sequentially the proliferation and differentiation of neural progenitors along the neuronal lineage, by activating distinct sets of target genes (Castro et al., 2011; Vasconcelos & Castro, 2014).

The mechanisms by which different Ascl1 target genes are activated differently in proliferating, versus differentiating progenitors, remain poorly understood. One possibility is that pathways co-regulating Ascl1 target genes may contribute to this process (Castro et al., 2011; Vasconcelos & Castro, 2014). Castro et al., 2011 found the consensus binding sequence for RBPJ (the effector of the Notch signalling pathway) to be highly enriched in the vicinity of Ascl1 binding events near genes that promote cell proliferation. RBPJ can mediate transcriptional activation or repression in cells with high or low Notch signalling, respectively. As result, such co-regulation may influence specifically the expression of a subset of Ascl1 targets, with different consequences in progenitors with distinct levels of Notch activity (Soares et al., 2022).

The restriction to proximal promoter regions, and the lack of chromatin states data, reveals the need to validate the model with a more robust experimental approach. Genome-wide mapping of Ascl1 binding sites with higher resolution can be achieved by coupling ChIP with Next Generation Sequencing (NGS) (ChIP-seq) (Nakato & Sakata, 2020; Nakato & Shirahige, 2017). Integration of ChIP-seq data characterizing the TF binding profiling, and the chromatin landscape, together with expression data of loss-of-function models, would result in a more comprehensive and mechanistic view of Ascl1 transcriptional program.

1.2 Objectives

The main goals of this work were to:

1. Provide a genome-wide characterization of Ascl1 transcriptional program in ventral telencephalon, and their associated target genes. This required the integration of Ascl1 genome-wide location data at E13.5, generated by chromatin immune-precipitation followed by Next Generation Sequencing (ChIP-seq), with expression data profiling Ascl1 loss-of-function (LoF) ventral telencephalon tissue, at the same developmental stage.

2. Characterize distinct chromatin states present along the genome. Data publicly available for histone modifications at the same developmental stage was analysed to map gene regulatory regions.

3. Identify changes in Ascl1 binding during development. Data from genome-wide mapping of Ascl1 binding sites across developmental stages was compared.

4. Find cis-regulatory regions that may respond transcriptionally to both Notch and Ascl1 pathways. For that, integration with genome-wide ChIP-seq data of RBPJ binding sites in ventral telencephalon at E11.5 was performed.

5. Determine how many binding sites were recovered with the Chip-seq approach from the list derived from the first genome-wide characterization of Ascl1 transcriptional program in the embryonic mouse brain performed by ChIP-chip (Castro et al., 2011).

1.3 Structure

This document is described by the following structure:

Chapter 2: State of the art

A brief introduction to neurogenesis, with a focus on the ventral telencephalon of the mouse brain. Importance and role of the transcription factor Ascl1 in the neurogenic process, including characterization of its transcription targets. Description of the main methodologies and bioinformatic tools necessary to answer the proposed biological questions. Explanation of the data analysis pipeline, including a description of the data integration steps used for downstream analysis.

Chapter 3: Methods

This chapter contains the description of relevant parameters for the different bioinformatic tools used for the data analysis pipeline during the development of this work, including explanations for their choice.

Chapter 4: Results and Discussion

The main results during this dissertation. How do they integrate and add knowledge to the state of the art.

Chapter 5: Conclusion and Future Work

A short analysis whether the proposed objectives were achieved and possible directions for future work.

2 STATE OF THE ART

2.1 Neurogenesis

Neurogenesis is a developmental process whereby neurons are generated from neural stem (NS) cells (Homem et al., 2015). Two cardinal properties of stem cells are shared by NS cells: multipotency and self-renewal. Most neurons in the mammalian brain are originated during embryonic development by NS cells that are called radial glia (RG) cells. By asymmetric division, RG cells can self-renew and have the potential to generate neurons (during early stages of development) and glial cells (at later stages) (Figure 1) (Kriegstein & Alvarez-Buylla, 2009). During the neurogenic phase, each RG cell divides into two different daughter cells: one RG cell and one immature neuron (direct neurogenesis), or one neuronal-committed intermediate progenitor (IP) cell (indirect neurogenesis). Different progenitors have been characterized in different germinal layers of the telencephalon, the rostral-most division of the forebrain. While RG divide in the ventricular zone (VZ), IPs migrate into the subventricular zone (SVZ) where they divide a few times and give rise to immature neurons by symmetric neurogenic divisions. The ventral part of the telencephalon is the region with the largest SVZ in the mouse brain during embryonic development (Turrero García & Harwell, 2017). As proliferation continues, immature neurons migrate into outer layers of the telencephalon (mantle zone) (Imayoshi & Kageyama, 2014; Kriegstein & Alvarez-Buylla, 2009; Sueda & Kageyama, 2020). Neurogenesis in the developing brain requires a fine balance between proliferation and differentiation events to allow for neuronal differentiation while maintaining the neural stem cell pool.

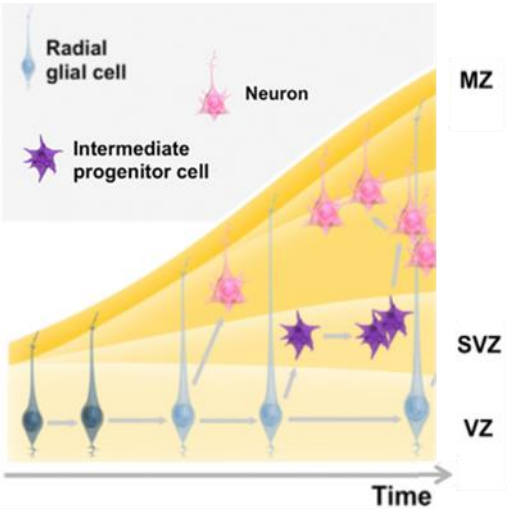


Figure 1: Neural progenitor (NP) cells in the developing telencephalon. MZ, Mantle zone; SVZ, subventricular zone; VZ, ventricular zone (adapted from (Imayoshi & Kageyama, 2014).

Neurons born in ventral telencephalon may become part of ventral structures (e.g. striatum), or migrate towards more distal locations, such as the cerebral cortex, or the olfactory bulb (Turrero García & Harwell, 2017).

2.2 Ascl1 in neurogenesis

Vertebrate neurogenesis is to large extent regulated by proneural TFs, such as Ascl1. Proneural factors are considered master regulators of neurogenesis, as these are both required and sufficient for inducing a full program of neuronal differentiation in the developing mammalian brain. Ascl1 function has been particularly well-described in the mouse ventral telencephalon, where its expression is confined to the germinal layers (VZ and SVZ), being expressed in both NS cells and neuronal committed intermediate progenitor cells (Castro et al., 2011). Ascl1 stops being expressed soon after cells exit the cell cycle, being absent from the MZ, where new-born neurons reside (Figure 2) (Castro et al., 2011).

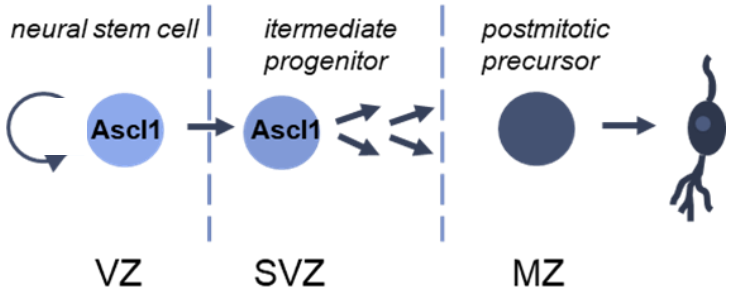


Figure 2: Ascl1 is expressed in both RG NS cells (VZ) and IPs (SVZ).

Genetic analysis in the mouse ventral telencephalon, where Ascl1 is the only known proneural gene to be expressed, has shown that this TF drives a GABAergic neurogenic program characterised by high lineage expansion (Casarosa et al., 1999; Castro et al., 2011; Turrero García & Harwell, 2017). Mice that carry a null mutation in Ascl1, present neural developmental defects associated with reduced generation of neurons (Casarosa et al., 1999). Conversely, overexpression of Ascl1 *in vivo* or in cultured progenitors results in a rapid cell cycle exit and full neuronal differentiation (Farah et al., 2000; Nakada et al., 2004). Ascl1 has also been used in protocols to convert various somatic cells into induced neurons (Vasconcelos & Castro, 2014) making it an interesting tool in the context of cell-replacement therapies for tackling neurodegenerative diseases (Vasan et al., 2021).

While promoting neuronal differentiation, Ascl1 induces Notch ligands, such as Dll1 (Delta1). Dll1 will bind to Notch receptors at the cell surface of neighbouring progenitors, triggering intracellular transmembrane proteases to cleave and release the Notch Intracellular Domain (NICD) (reviewed in (Sueda & Kageyama, 2020)). NICD is subsequently translocated into the nucleus where it is recruited by the transcription factor RBPJ. Upon Notch activation, RBPJ/NICD form a complex with coactivators that will induce the expression of Notch target genes, such as Hes1/5. These, in turn, repress the expression of proneural genes, including Ascl1. This process is called lateral inhibition, and functions to transiently counteract neuronal differentiation, contributing to the maintenance of a pool of neural stem/progenitor cells (Kageyama et al., 2019).

Ascl1 is a transcriptional activator, being a member of the basic helix-loop-helix (bHLH) family. The HLH domain is required for dimerization, while the basic domain is involved in sequence-specific DNA binding (Bertrand et al., 2002). Ascl1 forms heterodimers with another class of bHLH factors called E-proteins, which bind the E-box consensus sequence (CAGSTG) (Figure 3(A)). In line with its role in mediating direct DNA binding by Ascl1, this motif has been found at promoter regions of Ascl1 target genes (Figure 3 (B)) (Borromeo et al., 2014; Castro et al., 2011; Raposo et al., 2015).

(A) (B)

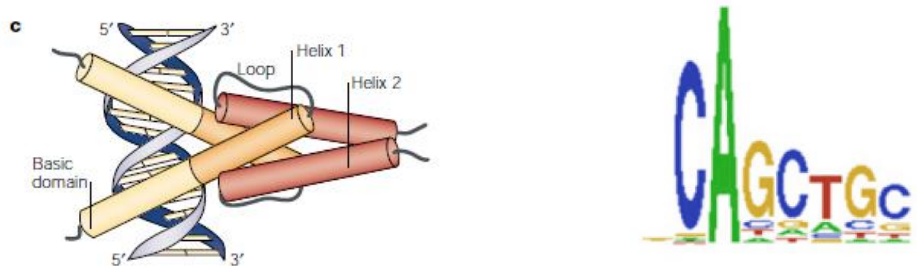


Figure 3: (A) Schematic representation of the structure of a bHLH dimer in a complex with DNA (adapted from (Bertrand et al., 2002)). (B) Motif enriched in Ascl1-bound promoters (Castro et al., 2011)

Genes directly bound and regulated by Ascl1 have been characterized *in vivo* both in ventral telencephalon and dorsal spinal cord of the developing mouse embryo (Borromeo et al., 2014; Castro et al., 2011). Both studies have shown Ascl1 functions as a transcriptional activator by directly inducing a large number of genes involved in several steps of neurogenesis. In line with this, Ascl1 target genes have

different onsets of expression along the neuronal lineage, from undifferentiated NS cells, to early born neurons.

In the ventral telencephalon, functional annotation analysis of the promoters bound by Ascl1 included biological processes associated with early steps of neurogenesis, such as "Notch signalling pathway", "cell fate commitment" and "regulation of cell cycle", and others related with later steps of neuronal differentiation, including "neurotransmitter biosynthetic process" and "cell projection morphogenesis" (Castro et al., 2011). The same study also showed a previously uncharacterized function of Ascl1 in maintaining cell proliferation, by directly regulating genes involved in cell cycle progression. These include E2f1, Cdk1, Cdk2, Skp2, Cdc25b and the oncogene FoxM1. The group of Ascl1 target genes involved in cell cycle arrest, comprised CcnG2, Btg2, Hipk2, Prmt2 and Gadd45g. Overall, these results demonstrate that Ascl1 regulates two groups of target genes with opposing roles in cell cycle control (Castro et al., 2011).

The mechanism that selects which genes are regulated by Ascl1 in proliferating versus differentiating progenitors, is poorly understood. In fact, multiple mechanisms are likely to operate. One of such mechanisms is a change in Ascl1 mode of expression, switching from oscillatory to sustained. Accordingly, Ascl1 promotes NS cell proliferation when it oscillates, while inducing neuronal differentiation and cell cycle exit, when its expression becomes sustained (Imayoshi et al., 2015; Vasconcelos & Castro, 2014). Ascl1 oscillations are driven by the oscillatory behaviour of Hes1, which can function as an intrinsic oscillator. Repression of Notch signalling is characteristic of the onset of differentiation, at which point Ascl1 expression becomes sustained.

Other possible mechanisms consist of cross-talks with other transcriptional pathways. In support of this, the enrichment of a consensus motif for RBPJ binding was found specifically at promoters of Ascl1 proliferation targets (Castro et al., 2011). RBPJ can mediate transcriptional activation or repression in cells with high or low Notch signalling, respectively. One possibility, is that co-binding of RBPJ and Ascl1 to gene regulatory regions at a subset of Ascl1 targets, will impact their expression in progenitors with distinct levels of Notch activity (Soares et al., 2022). Addressing this model will require mapping the binding events of Ascl1 and RBPJ at a genome-wide level and identify the genes they regulate.

2.3 ChIP-seq analysis

In vivo and *in vitro* genome-wide technologies are important to understand transcriptional regulation and gene expression. The large amount of data delivered by such high-throughput methods requires increasingly advanced statistical and computational analyses to integrate and extract biologically meaningful information. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is an application of such technologies. ChIP-seq data was extensively used in this thesis and its analysis will be described in detail.

2.3.1 ChIP-seq pipeline overview

ChIP-seq is a technique used for genome-wide profiling of proteins that come in close vicinity to DNA (e.g. DNA binding transcription factors, transcriptional co-factors or histones and their posttranslational modifications) (Nakato & Sakata, 2020; Steinhauser et al., 2016). ChIP-seq pipeline, overviewed in Figure 4, refers to cross-linked chromatin, fragmented and subjected to antibody-specific immunoprecipitation followed by DNA purification (Nakato & Shirahige, 2017). Chromatin fragments before immunoprecipitation are used as input DNA. After library preparation and DNA sequencing, reads are mapped onto a reference genome (Nakato & Sakata, 2020; Nakato & Shirahige, 2017). Genomic regions significantly enriched for immunoprecipitated reads, when compared with input reads, are detected as “peaks” (Nakato & Shirahige, 2017). The remaining genomic regions are considered background (Nakato & Shirahige, 2017). Called peaks represent regions for interaction of the protein under study. Several computational steps are required to determine the genomic coordinates of protein-DNA interactions from raw ChIP-seq data. When performing any ChIP-seq experiment, a single assay is often subject to a substantial amount of variation. To minimize this issue, the use of biological replicates is recommended. In addition, applying quality metrics during several steps of the ChIP-seq analysis protocol, in a combination with site-inspection-based evaluation, provides an overall assessment of the experimental outcome and data quality (Angarica & del Sol, 2017; Nakato & Sakata, 2020; Nakato & Shirahige, 2017; Steinhauser et al., 2016). In this section, I describe the main steps of a typical ChIP-seq data analysis together with strategies, challenges, and data QC metrics available for each step.

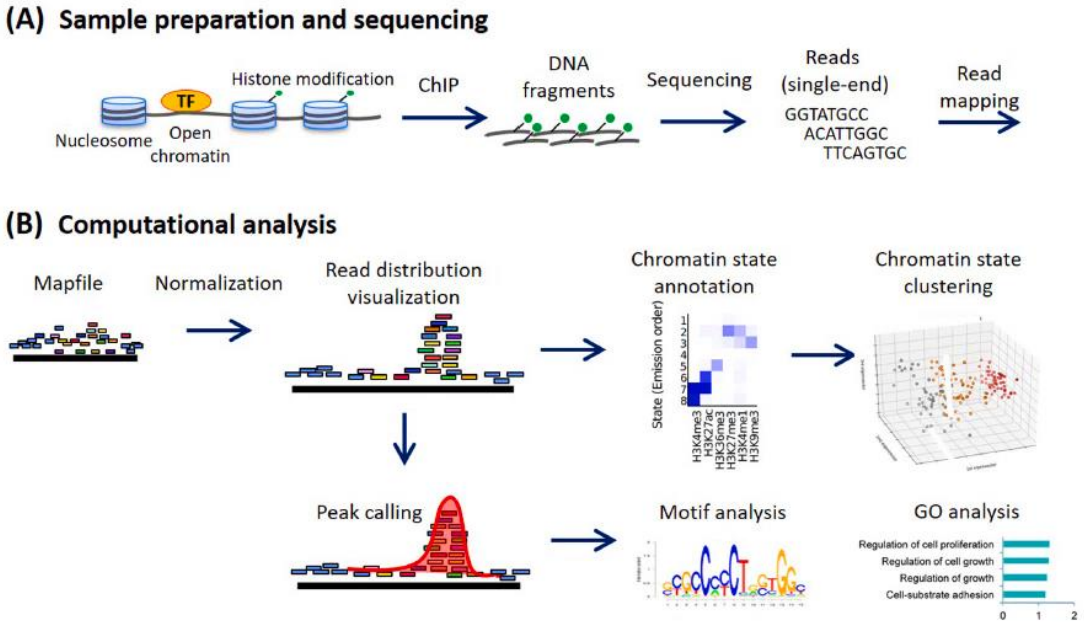


Figure 4: ChIP-seq analysis pipeline. (A) Sample preparation and sequencing. (B) Computational analysis in a basic ChIP-seq analysis (Nakato & Sakata, 2020).

2.3.2 Sequencing

Libraries can be sequenced using either single-end or paired-end, generating short sequence reads from one end or both ends of the DNA template, respectively (Kidder et al., 2011; Nakato & Shirahige, 2017). Although more expensive, benefits of paired end sequencing include improved identification of duplicated reads, increased mapping efficiency to repeat regions and better estimation of fragment sizes (Kidder et al., 2011).

2.3.3 Sequenced reads quality assessment

Sequencing generates a collection of large files, frequently in FASTQ format, containing the sequence data. For each nucleotide in the read there is an associated quality score which represents the probability that the corresponding nucleotide call is incorrect. Probability values are generated by base calling algorithms and depend on the signal captured during base incorporation (Illumina, n.d.).

At this stage, quality assessment of sequenced reads is recommended to determine the success of the sequencing step (Kidder et al., 2011; Nakato & Sakata, 2020). Quality of the data increases with the number of times a base is sequenced.

2.3.4 Alignment

Alignment or mapping of sequenced reads consists of finding the corresponding part of each sequenced read in its reference genome sequence. Over the last years, many alignment programs have been developed to efficiently process millions of short reads including, Bowtie2 (Langmead & Salzberg, 2012) and BWA (Chong et al., 2003).

2.3.5 Read filtering

Several filters may be applied to obtain an appropriate set of reads for further analysis.

2.3.5.1 Unmapped, multiple mapped and redundant reads

The first quantitative measure determined after the alignment step is the mapping ratio. This is the percentage of reads that successfully aligned to the reference genome and reflects the proportion of reads that derive from true genomic DNA (Nakato & Sakata, 2020). Aligned reads comprise uniquely mapped reads (reads mapped to a single genomic location) and multiple mapped reads (reads mapped to multiple loci on the reference genome) (Davis et al., 2018; Nakato & Shirahige, 2017). The inclusion of multiple mapped reads increases the number of usable reads, but the number of false positives might be higher, decreasing confidence in site discovery (Nakato & Shirahige, 2017). Therefore, it is standard practice to eliminate all multi-mapped reads from further analysis (Stark & Hadfi, 2016).

The next issue concerns PCR duplicates. Uniquely mapped reads comprise redundant reads (reads obtained multiple times that align to the same genomic location) and non-redundant reads (uniquely mapped reads that remain after PCR duplicates being removed) (Davis et al., 2018). Sonication of ChIP-seq samples produces a collection of overlapping reads. It is highly unlikely that two DNA fragments will be generated and map the same genomic location. Hence, reads with identical mapping positions likely correspond to PCR duplicates, arising from amplification of the same fragment during PCR. The default approach is to filter out all duplicated reads, leaving only a single exemplar of the read before further analysis.

The ENCODE project sets a minimum of at least 20 million distinct uniquely mapped reads (that remain after removing PCR duplicates) to analyse sharp-mode peaks, such as the ones obtained with TFs (Blanco & Abril, 2004).

2.3.5.2 Blacklisted regions

Blacklisted regions represent a set of regions that have anomalous, unstructured, or high signal in ChIP-seq experiments independent of cell line or experiment (Amemiya et al., 2019). Some regions overlap specific types of repeats such as centromeres, telomeres, and satellite repeats (Amemiya et al., 2019). Also, if some of these regions comprise uniquely mappable regions, simple filters are unable to remove them. Therefore, removal of blacklisted regions is a standard procedure. ENCODE consortia provides blacklists for various species and genome versions (Davis et al., 2018).

2.3.6 Relative Strand Correlation (RSC)

RSC quantifies fragment clustering prior to peak calling (Nakato & Shirahige, 2017). For high quality ChIP-seq experiments, it is expected significant clustering of enriched sequenced reads on the FW and RW strands at locations bound by the protein of interest (Landt et al., 2012). To calculate RSC, a measurement of agreement between the two strands (e.g., Pearson correlation) must be computed. Then, it should be re-computed after shifting one strand relative to the other (Landt et al., 2012; Nakato & Shirahige, 2017). The RSC is then calculated as the ratio of the fragment-length cross-correlation value minus the background cross-correlation value, divided by the read-length cross-correlation value minus the background cross-correlation value (Landt et al., 2012; Nakato & Shirahige, 2017). Encode quality standards recommend RSC values equal or greater than 0.8 for transcription factor analysis (Landt et al., 2012; Nakato & Shirahige, 2017).

2.3.7 Input samples

It is standard practice to generate both input and ChIP libraries at the same time (Stark & Hadfi, 2016). Input sample corresponds to the sonicated chromatin, without being subject to the immunoprecipitation step. Input samples provide a model of background distribution of genomic loci to separate truly enriched regions from those not associated with the protein of interest. Additionally, the input reveals the open chromatin "signature" of each cell type (Stark & Hadfi, 2016). As DNA fragmentation is not a uniform process, some regions can be preferentially represented because open chromatin regions tend to be more easily fragmented than closed regions (Feng et al., 2012; Stark & Hadfi, 2016).

2.3.8 Peak calling

Peak calling is used to identify areas in the genome that have been enriched with aligned reads, which represent the locations of protein-DNA interactions (e.g., transcription factor binding sites). To each genomic location, peak calling algorithms start assigning the number of reads that cover this position (Zhang et al., 2008). This number corresponds to the strength of the protein binding event. Then, these tools segment the signal into background regions and regions with potential peaks. Different methods can be applied for segmentation: window-based approaches or methods like hidden Markov models (HMMs) (Steinhauser et al., 2016). Finally, a statistical test is performed to check whether the potential peaks significantly differ from the background signal (input sample) to determine if the site of enrichment is likely to be a real binding site (Landt et al., 2012). As an output, a list of peaks is usually provided where each peak is assigned to a P-value and/ or FDR value. Significance values calculated from different peak callers are not directly comparable (Landt et al., 2012) as different statistical models are used.

MACS2 (Model-based Analysis of ChIP-seq) is one of the most used peak caller tools for identifying transcription factor binding sites (Feng et al., 2012). Its workflow is represented on Figure 5 (Feng et al., 2012). MACS2 assumes that DNA fragments in ChIP-Seq experiments are equally likely to be sequenced from both positive and negative ends (Feng et al., 2012). As a result, the read density around a protein-DNA interaction location is likely to show a bimodal enrichment pattern (or paired peaks). MACS2 uses this bimodal pattern to build a peak model. First, MACS2 considers window size (bandwidth) and fold-enrichment (mfold) parameters. Fold-enrichment consists of a threshold for the ratio of densities (read counts) of test over input within windows. The window size corresponds to the sonication size and MACS2 divides the genome into windows of the given size. Then, it slides a two bandwidth windows across the genome searching for windows with reads more than mfold enrichment relative to a random read distribution. By default, the bandwidth size is 300 bp and the fold-enrichment is the range of ratios from 10 to 30. After detecting all the regions that satisfy the given fold-enrichment, MACS2 randomly samples 1000 of those regions. For each peak, MACS2 separates their positive and negative stranded reads and aligns them by the midpoint between their centres. The distance between the midpoint of the peaks represents the estimated fragment length (d). Then, algorithm merges each pair of positive and negative peaks, which are located next to each other, in a single peak. This is done by shifting the reads of each peak of the pair toward the 3'ends by d/2. The model is overridden for paired-end reads, as there is no need for d estimation.

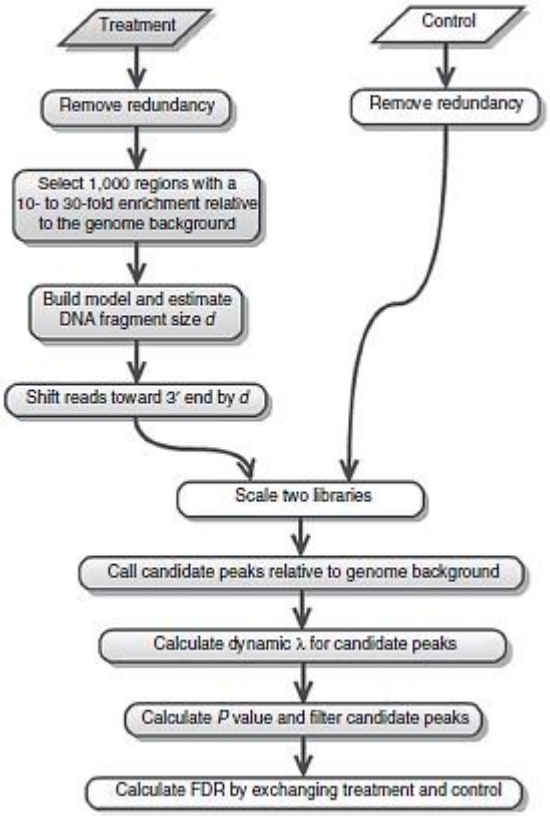


Figure 5: General MACS workflow (adapted from Feng et al., 2012).

To establish the significance of peaks first MACS2 assumes that read counts have a Poisson distribution with a parameter lambda BG. Then, it performs a local checking for each peak by using Poisson with a dynamic parameter, lambda local. This is estimated from the candidate region under consideration in the input sample and is deduced by taking the maximum value across various window sizes (Feng et al., 2012). A p-value is calculated for each candidate peak. Finally, peaks with p-values under a user defined threshold are reported. P-values can be further corrected for multiple comparison using the Benjamini-Hochberg correction. When sequence depth differs between input and treatment samples, MACS2 linearly scales the total input read count to be the same as the total ChIP read count. By default, the larger sample is scaled towards the smaller sample (GitHub Macs3 call peaks).

MACS2 outputs several files. The narrowPeak file is a BED 6+4 format file which contains the peak locations together with peak summit, p-value and q-value and is used for downstream analysis. The output bed files contain the peak summit locations for every peak. There are two BedGraph format output files: one containing the pileup normalised signals from the ChIP sample, and the other containing the local

biases estimated for each genomic location from the input sample. These are recommended for purposes of visualization.

2.3.8.1 Faction of reads in peaks (FRiP)

FRiP is an indicator for the signal-to-noise ratio in the data. It is estimated by calculating the ratio of reads that uniquely map to a single location of the genome within significantly enriched peaks, and the overall number of uniquely mapped reads (Davis et al., 2018). The higher the FRiP, the better the signal-to-noise ratio. In fact, only a minority of reads in ChIP-seq experiments occur in significantly enriched genomic regions (i.e., peaks), the rest of the reads represents background. For a mammalian genome a minimum of 1% FRiP is expected (Landt et al., 2012). This metric is particularly useful for comparing results obtained with the same antibody across cell lines/ conditions when derived from the same peak caller (Landt et al., 2012). This metric is sensitive to the peak calling method, including the way the algorithm identifies regions of enrichment, the parameters and thresholds used (Landt et al., 2012).

2.3.9 Reproducibility

ChIP-seq experiments consist of samples groups and replicates. To take the analysis forward, intervals from peaks identified at each replicate must be combined. When the analysis aims to work directly with peaks identified by the peak caller (qualitative analysis), then a conservative method that minimizes false positive (like intersection) might be the choice (Stark & Hadfi, 2016). When performing a quantitative differential analysis, which is considered robust with respect to noise, a more relaxed approach may be used. This includes the generation of a consensus peak set resulting of the union of all (or most) of the identified peaks (Stark & Hadfi, 2016). The ENCODE project recommends that all ChIP experiments should be performed on two independent biological replicates and replicate agreement should be assessed by the irreproducible discovery rate (IDR) analysis pipeline (Landt et al., 2012).

2.3.9.1 Irreproducible Discovery Rate (IDR) algorithm

IDR algorithm is based on the concept that if two replicates measure the same underlying biology, the most significant peaks, which are likely to be genuine signals, are expected to have high consistency

between replicates (Q. Li et al., 2011). Advantages of this approach include not requiring peak calling initial cut-offs, which are not comparable for different peak callers (Q. Li et al., 2011). As IDR is based on ranks, only order of the signals is important. IRD thresholds are constrained by the quality and enrichment of the weakest replicate (Landt et al., 2012; Nakato & Shirahige, 2017). True peaks will be discarded by this analysis if they are not reproducible in the weak replicate (Landt et al., 2012; Nakato & Shirahige, 2017). IDR is an open-source Python tool used in the IDR pipeline. The first part of this pipeline aims to determine the replicate peak rank threshold that later, in the second part, is being used to truncate the ranked list of pooled-data peaks. Briefly, for each pair of replicates, peak calling is performed with MACS2 using a less stringent P or Q-value. After that, each peak list is ranked by a score that can be either heuristic based (e.g., fold enrichment) or probabilistic based (e.g. P-value) (Q. Li et al., 2011). IDR framework considers peaks that are present in both replicates to belong to one of two groups: a reproducible group and an irreproducible group (Landt et al., 2012; Q. Li et al., 2011). This tool uses a two-component probabilistic copula-mixture model to fit the peak rank distributions from the pairs of replicates (Q. Li et al., 2011). This way, the method learns the degree of peak-rank consistency in the signal scores and the proportion of peaks belonging to each group (Q. Li et al., 2011). As a result, an IDR score for each peak is assigned, which reflects the posterior probability that the peak belongs to the irreproducible group (Q. Li et al., 2011). Low IDR scores represent high-confidence peaks. A user defined threshold for the IDR score is used to obtain the peak rank thresholds (Kundaje, n.d.). When IDR determines the replicate peak rank threshold based on a pair of true replicates (Nt,Figure 6), the resulting output is a "conservative" peak set. Peak regions in this list can be interpreted as high confidence (Davis et al., 2018). On the other hand, if the replicate peak threshold is based on a pair of pseudo-replicates (Np,Figure 6), then the "optimal" peak set is determined. This sampling strategy consists on mapped reads being pooled across all replicates of a dataset, and then randomly sampled (without replacement) to generate two pseudo-replicates with equal numbers of reads (Kundaje, n.d.). The aim is to transfer signal from stronger replicates to the weaker replicates, thereby balancing data quality and sequencing depth across replicates (Davis et al., 2018). If a dataset has more than two replicates, all pairs of replicates are analysed using the IDR tool (Kundaje, n.d.). The maximum peak rank threshold across all pairwise analyses is used as the final replicate peak rank threshold. Finally, for each dataset, the best of the Nt and Np thresholds are used to obtain a final consolidated set of peaks.

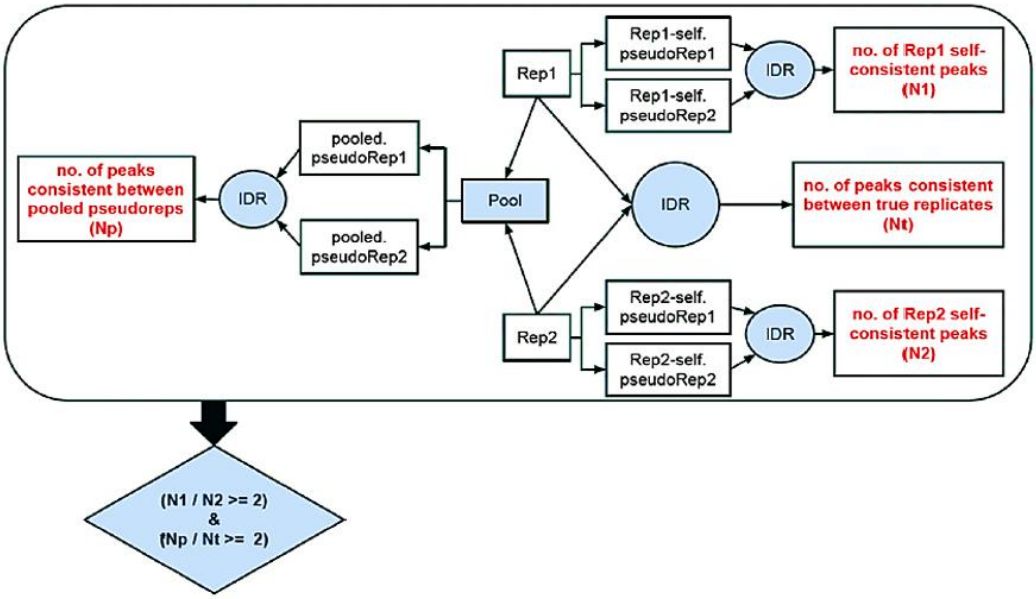


Figure 6: IDR pipeline for assessing ChIP-seq quality using replicates (adapted from (Landt et al., 2012)).

2.3.9.2 IDR QC metrics

The IDR pipeline also provides QC metrics to assess the reproducibility between replicates. The self-consistency ratio measures consistency within a single experiment and is used to ensure similar weighting of individual replicates when identifying binding regions (Landt et al., 2012). It is obtained by partitioning reads, from each replicate and use IDR comparison within each group (Figure 6Table 1). The self-consistency ratio should have a value less than 2 (max(N1/N2/min(N1,N2)) (Landt et al., 2012). When the replicates within a single experiment are not comparable, ENCODE developed the rescue ratio to measure consistency between datasets (Davis et al., 2018). It is the ratio between the thresholds Np and Nt (Figure 6) (Landt et al., 2012). ENCODE portal provides a scheme to classify ChIP-seq data into two different statuses (Table 1). An experiment passes if both self-consistency and rescue ratios are less than 2.

Table 1: ENCODE data quality classification based on IDR QC metrics (adapted from (Davis et al., 2018)

Self-consistency Ratio	Rescue Ratio	Resulting Data Status	Flag colors
Less than 2	Less than 2	Ideal	None
Less than 2	Greater than 2	Acceptable	Yellow
Greater than 2	Less than 2	Acceptable	Yellow
Greater than 2	Greater than 2	Concerning	Orange

2.3.10 Visualization of ChIP-seq data and analysis of known positive targets

Several visualization tools can be used for visual inspection of read distribution, allowing detailed analysis of peaks and biological interpretation (e.g. Integrated Genome Viewer (IGV)) (Nakato & Sakata, 2020). When using web servers, such as UCSC genome browser (Kent et al., 2002), it is easier to integrate the ChIP-seq results with other available data, such as histone modifications and gene expression in various tissues (Nakato & Sakata, 2020). bigWig is a standard file format commonly used for ChIP-seq data visualization (Nakato & Sakata, 2020).

2.4 Downstream analysis

In case of ChIP-seq for TFs, enriched regions represent likely locations of where a TF binds to the genome. After obtaining the lists of peak coordinates, the study of biological implications of protein-DNA bindings helps answering the underlying biological questions.

2.4.1 Qualitative Analysis

Qualitative analysis aims to identify peaks that are unique or common to sample groups. This analysis uses as input peaks directly identified by the peak caller (Stark & Hadfi, 2016). Once lists of highly reproducible peak intervals have been derived, lists from different groups are overlapped to isolate peak regions that are common or unique between them. Some false positives or negatives might arise from these binary comparisons due to noise intrinsic to peak calling and the imbalance in sample numbers between the groups (Nakato & Sakata, 2020; Stark & Hadfi, 2016).

2.4.2 Peak annotation with genes and genomic regions

Functional interpretation of genomic intervals (peaks) requires integration with known genomic annotations. The standard approach to identify which genes are associated with the binding sites consists of annotating peaks with the nearest gene and proximal genomic regions such as promoters, exons, introns, and distal regions where the peak is located (McLean et al., 2010; Yu et al., 2015).

2.4.3 Gene Ontology terms enrichment analysis

Gene Ontology (GO) enrichment analysis consists of identifying predominant biological functions, processes or pathways that are over-represented on a particular gene set (Nakato & Sakata, 2020). A typical analysis compares the total fraction of genes annotated for a given ontology term with the fraction of annotated genes identified by proximal binding events to obtain a gene-based P value for enrichment (McLean et al., 2010). This approach presents a disadvantage: restricting the analysis to proximal binding events (for example, under 2-5 kb from the TSS) will discard distal events (McLean et al., 2010). These contain cis-regulatory elements, such as enhancers, that can play crucial roles in controlling gene expression in specific cell types, conditions, and developmental stages. Moreover, binding sites might be located between two start sites of different genes or hit different genes, which have the same TSS location in the genome (Yu et al., 2015). To account for these issues, several tools provide parameters to specify distances from the TSS or define regulatory domains when performing GO analysis (McLean et al., 2010; Yu et al., 2015).

Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010) analyses the functional significance of cis-regulatory regions identified by ChIP-seq across an entire genome. First, every gene in the genome is assigned a regulatory domain and all noncoding sequences that lie within the regulatory domain are assumed to regulate that gene. GREAT supports three different parametrized association rules to define gene regulatory domains. Then, each genomic region (peak) is associated with the gene whose regulatory domain it overlaps. To determine functional enrichments, the number of peaks that land on each regulatory domain having associated a GO term is calculated. To ensure an accurate annotation of enrichments for genomic regions, GREAT performs both the binomial test over genomic regions and the hypergeometric test over genes. Each test has a bias that is compensated by the other test.

2.4.4 DNA motif analysis

There are many thousands of sequences in the non-coding portion of the genome whose role is to mediate the interaction of sequence-specific TFs and chromatin. For a particular TF many variations of binding sites may exist, differing by only some nucleotides from one another (Tran & Huang, 2014). Thus, identifying the conserved short DNA sequences shared across these specific regions reveals the consensus binding motif for that TF (Boeva, 2016; B. Liu et al., 2018; Rocha & Ferreira, 2018; Tran &

Huang, 2014). Often, genomic locations identified by ChIP-seq not only contain the binding motifs for the TF under study but are also enriched with binding motifs for other TFs (cofactors) (Sharov & Ko, 2009). This happens when the TF used for the immunoprecipitation binds to DNA indirectly, through binding to other TFs that directly binds to DNA (Sharov & Ko, 2009) or when they require co-factors to bind or regulate DNA. Discovered motifs are compared with databases of known motifs to identify potentially bound transcription factors and further elucidate mechanisms of transcription regulation.

The Position Weight Matrix (PWM) is the most frequently used mathematical model to represent binding motifs (Boeva, 2016; B. Liu et al., 2018). It contains information about the position-dependent frequency or probability of each nucleotide in the motif and can be visualized using sequence logos. The height of each nucleotide in the logo is proportional to its probability for each position and the four nucleotides are ordered by probability with the most likely nucleotides shown on top of the stack (Boeva, 2016).

Many bioinformatic tools have been developed for motif finding. CisFinder (Sharov & Ko, 2009) web-based software produces a list of motifs enriched in a set of DNA sequences and describes them with position frequency matrices (PFMs). This tool can process large sequences and discover multiple and weak motifs in a single run, even with a low level of enrichment (Sharov & Ko, 2009). For this end, the algorithm estimates PFMs directly from counts of n-mer words with and without gaps. After that, it extends PFMs over gaps and flanking regions and clusters them to generate non-redundant sets of motifs.

2.4.5 Identification of TF target genes

A binding event of a TF does not necessarily imply the existence of a gene regulatory function. To infer direct target genes, genomic location analysis needs to be integrated with further evidence for gene regulation, usually expression profiling experiments designed to perturb gene function. Typically, micro-array or RNA-seq and ChIP-seq assays are performed independently. Differentially expressed genes are then compared with peak annotated genes from the ChIP-seq using a simple overlap.

2.4.6 Mapping of gene regulatory regions

Chromatin encodes epigenetic information in the form of histone modifications, histone variants and regions of open chromatin (Ernst & Kellis, 2017; Gorkin et al., 2020). Genome-wide maps of these

chromatin marks provide important information for annotating the noncoding genome, including identifying regulatory elements. The four basic types of histones proteins H2A, H2B, H3, and H4 are subject to different modifications, including acetylation, methylation, and phosphorylation. For example, trimethylation of histone H3 lysine 4 (H3K4me3) marks active gene promoters, but long-term repression regions are marked by trimethylation of histone H3 lysine 9 (H3K9me3). In contrast, monomethylation on lysine 4 of histone 3 (H3K4me1) has been linked to distal enhancer regions (Andersson & Sandelin, 2020; Bernstein et al., 2006; Calo & Wysocka, 2013). Characteristic combinatorial and spatial patterns of these epigenetic marks, termed chromatin states, have been shown to correlate with various functional elements in the genome, such as active promoters, poised or strong enhancers, and transcribed, repressed, and repetitive regions (Baker, 2011).

Different computationally approaches have been developed to predict and annotate chromatin states (Angarica & del Sol, 2017; Vu & Ernst, 2021). ChromHMM (Ernst & Kellis, 2017) is based on a multivariate hidden Markov model (HMM) that performs the probabilistic modelling of both the combinatorial presence/absence of multiple histone marks and the spatial constraints of how these mark combinations occur relative to each other across the genome (Ernst & Kellis, 2017). This tool accepts as input aligned reads for each chromatin mark, and chromatin states are then analysed at 200 bp intervals (approximately nucleosome size) across the genome. Advantages of this tool include speed, ease of use and being able to automatically compute state enrichments for external annotations (Ernst & Kellis, 2017). However, ChromHMM does not provide the biological significance of each chromatin state, a task that requires additional knowledge.

2.4.7 Environment setup

Analysis pipelines use different bioinformatic tools, written in different computational languages. Each language requires a different setup method, and several tools require a specific OS version. Moreover, analysis pipelines used by large-scale projects are difficult to modify and/or update (Nakato & Sakata, 2020). Depending on the analysis pipeline chosen, it is an advantage to use virtual environments like Docker (*Docker*, n.d.). This option overcomes the need to replace or update packages, providing bioinformatic tools released as pre-complied computational environments (Nakato & Sakata, 2020).

2.5 Overall tasks

In the previous section, I described the various approaches to improve the resolution of Ascl1 binding sites mapping and to solve the lack of chromatin states data integration, validating the model with a more robust experimental approach. The following list describes the main tasks achieved during the current work:

1. Analysis of ChIP-Seq data using a critical approach into existing bioinformatic tools (functionality, features, specific requirements, and limitations) to establish and refine the data analysis pipeline.

2. Generation of lists of highly reproducible peaks using the Irreproducibility Discovery Rate (IDR) framework.

3. Comparison between lists of Ascl1-bound genomic regions and Ascl1 LoF expression arrays data to identify Ascl1 target genes directly affected by Ascl1 loss.

4. Understand which chromatin states Ascl1 preferentially binds to.

5. Identify the co-occurrence of RBPJ and Ascl1 binding sites in the genome.

6. Determine how many binding sites were recovered with the ChIP-seq approach in comparison with the ChIP-chip methodology from Castro et al. (2011).

3 Methods

The bioinformatics pipeline for global characterization of Ascl1 transcriptional program is shown in Figure 7. Each step is explained in the following sections, including the bioinformatic tools and parameters used, and the reasons for their choice. The code for the data analysis is accessible in the GitHub repository online through the URL: https://github.com/rjoana1/Ascl1.

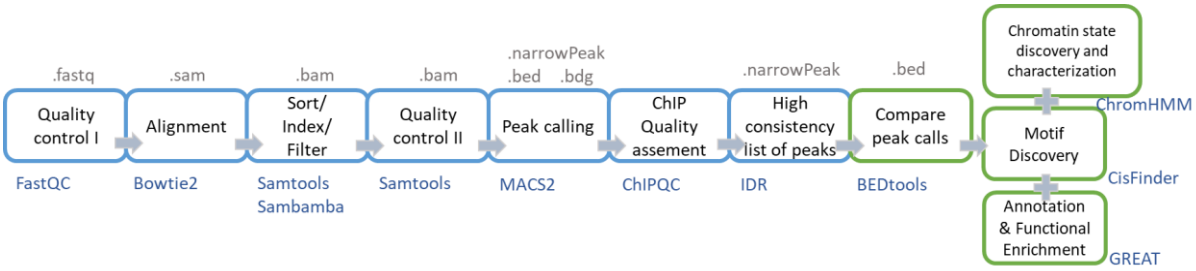


Figure 7: ChIP-seq analysis pipeline and downstream analysis. Blue boxes indicate the steps involved in ChIP-seq analysis while green boxes represent downstream analysis. When appropriate, the file format output (on top, grey) from each tool (bottom, in blue) is indicated.

3.1 Dataset

ChIP-sequencing data previously produced by the laboratory from mouse ventral telencephalon was available for two transcription factors at different time points: Ascl1 at E11.5, E13.5 and E14.5 and RBPJ at E11.5. Those developmental stages correspond to a more proliferative (E11.5) and a more differentiated (E14.5) population of progenitor cells. Datasets corresponding to input controls for E11.5 and E14.5 time points were also available. All ChIP-seq experiments were performed in duplicate. Additionally, microarray data (Affymetrix MOE430 2.0 arrays) from E13.5 Ascl1-null mutant embryos, also previously produced by the laboratory, was used for expression profiling analysis. Finally, a chromatin state model previously generated in the laboratory, based on publicly available histone (H3K27ac, H3K4me3, H3K4me1) ChIP-seq data (Lindtner et al., 2019), was used in this study.

3.2 Quality Control I – FastQC

The command line version of Fastqc (v0.11.9) (Andrews, 2010) was used to analyse the FASTQ files containing the reads. This Java tool performs quality control checks on raw sequence data coming from high throughput sequencing pipelines. The analysis outputs per sample summary reports including

per base sequence quality, per base sequence content, sequence duplication levels and overrepresented sequences.

3.3 Alignment – Bowtie2

Raw reads were aligned to the mouse (GRCm38/mm10) genome using Bowtie2 (version 2.4.1_cv1) (Langmead & Salzberg, 2012). This alignment tool is based on a Burrow-Wheeler transform algorithm and outputs unsorted Sequence Alignment Map (SAM) files. These files contain information for each individual read and its alignment to the genome (Langmead & Salzberg, 2012). This information includes the genomic position of the best mapping (chromosome, start position, strand) and quality metrics for the confidence that the read is correctly and uniquely aligned. The local alignment mode of Bowtie2 was chosen to omit some read characters at the end of the read (soft-trimming) aiming to remove poor quality bases or adapters from untrimmed reads (Langmead & Salzberg, n.d.). This option increases the alignment score by removing poor quality bases or adapters from untrimmed reads. Previously to the mapping process, the reference mm10 file was indexed to be used by Bowtie 2 using the "*bowtie2-build"* command. The aim was to construct a Bowtie index from the set of DNA sequences in the reference file.

3.4 Filtering uniquely mapped reads – Samtools and Sambamba

Before the filtering step, SAM files were converted to the Binary Alignment/Map (BAM) format, sorted by read coordinate locations and indexed for fast random access using the Samtools (H. Li et al., 2009). BAM is a compressed, binary representation of the SAM format which improves the performance despite keeping the original information. To keep only uniquely mapping reads for further analysis, BAM files were filtered by specifying a set of filters for the Sambamba command-line tool (Tarasov et al., 2015). Unmapped, duplicates and multimappers were filtered out by specifying in the filter "not unmapped", "not duplicate" and [XS] == null, respectively (Tarasov, 2016). Samtools flagstat command was used to calculate and print statistics on aligned files, before and after filtering (H. Li et al., 2009).

3.5 Peak calling – MACS2

Peaks of enriched binding regions against input DNA control were called using MACS2 (version 2.2.7) (Zhang et al., 2008). This process was performed for both individual replicates and merged

replicates (as part of the IDR pipeline, see 3.7). MACS2 function "*callpeak*" was used with default parameters for all samples except the p-value cut-off set to 0.01. For paired-end data, the "BAMPE" parameter was chosen to use the actual insert sizes of pairs of reads to build fragment pileup (Zhang et al., 2008).

3.6 Quality Control II – ChIPQC

The Bioconductor package ChIPQC (version 1.30.0) (Carroll et al., 2022; Thomas et al., 2014) was used to compute the following quality assessment metrics and plots: Reads in Blacklists, Reads in Peaks, Coverage Histogram and Relative Strand Correlation (RSC). These were generated using the sorted BAM files and the individual replicate called peaks files as inputs.

3.7 IDR pipeline

For Ascl1 transcription factor data, an IDR score threshold of 0.05 (5%) was used to obtain the conservative peak rank threshold (Nt) from individual true replicates, previously ranked by signal-score (–log10(p-value)) using IDR (version 2.0.4). The same IDR score threshold was applied for RBPJ E11.5 data but pseudo-replicates were used instead to calculate an optimal peak rank threshold (Np). These pseudo-replicates were generated by polling mapped reads across all replicates of the dataset and randomly sampling it (without replacement). Finally, reads from individual replicates from each TF were pooled and MACS2 was used to call peaks on the pooled data with a relaxed p-value of 0.01. Pooled-data peaks were ranked by signal-score and Nt or Np previously calculated were used to threshold these lists and obtain the final reproducible lists of peaks. The self-consistency ratio was obtained by partitioning reads, from each replicate, into two equal groups (Figure 6) and use IDR comparison within each group.

3.8 Filtering blacked regions – bedtools

All post-IDR lists of peaks were screened against a curated blacklist of regions in the mouse genome (Amemiya et al., 2019) and peaks overlapping the blacklisted regions were discarded. This step was performed by the "intersect" function of bedtools (Quinlan & Hall, 2010), using the parameter "-v". This way, only those entries in post-IDR lists that presented no overlap in the curated blacklist were reported.

This software suite, written in C++, is used to compare, manipulate and annotate genomic features bed and gtf file formats.

3.9 Visualization – bedGraphToBigWig

The software bedGraphToBigWig, available on ENCODE Portal (*ENCODE: Encyclopedia of DNA Elements*, n.d.) was used to convert bedGraph files generated by MACS2 (pileup normalised signals from each ChIP sample and corresponding inputs) to bigWig format.

3.10 Peak annotation with genomic features – ChIPseeker

The Bioconductor package ChIPseeker (v1.28.3) (Yu et al., 2015) was used to determine the genomic annotations associated with Ascl1 E13.5 peak regions. The analysis was performed on Ascl1 E13.5 high confidence peak list (post-IDR bed file) by defining a range of 1kb upstream and downstream from TSS as the promoter region. In case of genomic features overlap, the following priority was adopted for annotation: Promoter, Exon, Intron, Downstream (defined as the downstream of gene end) and Intergenic.

3.11 Peak annotation with genes and Gene Ontology analysis – GREAT

GREAT (v. 4.0.4) (McLean et al., 2010) was used for both peak-gene annotation and functional enrichment analysis of Ascl1 E13.5 post-IDR peak list (3 column bed files used as input). After parameter optimization, the "single nearest gene" approach was chosen to define the regulatory domains. This consisted of extending the regulatory domain in both directions to the midpoint between the gene's TSS and the nearest gene's TSS but no more than the maximum extension in one direction. Gene names contained in GREAT's output list were converted into Entrez IDs. For that an R script was created using the following R packages: org.Mm.eg.db (Carlson, 2019) and AnnotationDbi (Pagès et al., 2021).

3.12 Mapping of gene regulatory regions – ChromHMM

Using the default parameters, a 6-state HMM model had been previously generated in the laboratory via ChromHMM (Ernst & Kellis, 2017) (version 1.14). For that, WT data from three histone

modifications (H3K4me1, H3K4me3, and H3K27ac) from basal ganglia WT samples at E13.5 (Lindtner et al., 2019) was used and chromatin states were assigned based on the emission probabilities. The output segmentation .bed file (mm9 build) containing ChromHMM's genome annotation, including the coordinates of each segment and the corresponding state was first converted to the mm10 genome assembly with LiftOver (Hinrichs et al., 2006). Then, it was used to compute the neighbourhood enrichment heatmap (NeighborhoodEnrichment command) for each state for each 200-bp bin position within 2 kb around the peak summits of Ascl1 E13.5 binding sites. Candidate state descriptions based on Ernst & Kellis, 2017.

3.13 DNA motif analysis – CisFinder

CisFinder software (Sharov & Ko, 2009) was used to identify DNA over-represented motifs in the entire set of Ascl1 E13.5 bond segments. The analysis was performed within a 50-bp region centered at the peak summits, using default parameters (FDR cut-off set to 0.05) and clustering by similarity. Coordinates corresponding to the list of Ascl1 E13.5 bound segments systematically shifted by 3000 bp upstream from the summits were used as control. To identify motifs enriched for other TFs (cofactors), all sequences containing the consensus Ascl1-binding motif were masked. Briefly, getfasta command from bedtools (Quinlan & Hall, 2010) was used to extract the sequences from the reference mouse genome (mm10) FASTA file for each of the intervals defined in the subset of bound segments common to both Ascl1 E13.5 and RBPJ E11.5 peak BED file. Next, masking was performed by substituting the PWM GCAGCTG with N characters. CisFinder analysis was performed as before.

3.14 Ascl1 E13.5 LoF expression data integration

Starting from Ascl1 E13.5 LoF probe based analysed data, an R script (dplyr R package v.1.0.7) containing several filters was developed and applied to annotate peaks with expression values. First, probes with the same Entrez ID and highest fold change were filtered in. Entrez ID missing values were excluded. Then, ties between probes with the same Entrez ID were solved by filtering in probes annotated with the smallest p-value lower than 0.05. The log2 fold cut-off for deregulated probes was set to less than -0.3 or more than 0.3. Finally, the resulting list of deregulated genes was intersected with the Ascl1 E13.5 peak list, previously annotated with Entrez IDs using the GREAT strategy (3.11). Entrez Gene IDs annotation was chosen because these IDs are considered more stable and outdated IDs are easier to

map to current IDs than Ensembl Gene IDs. Additionally, Gene Symbol annotation was not a choice because each gene symbol might have more than one entry at Ensembl/NCBI.

3.15 Qualitative analysis – Bedtools

Post-IDR lists of peaks from Ascl1 E11.5, Ascl1 E13.5 and RBPJ E11.5 were intersected, in pairs, to generate lists of unique and common peaks. This step was performed by the "intersect" function of bedtools (Quinlan & Hall, 2010). When using the parameter "-v", only entries in the list used as a reference that showed no overlap in the other list were reported. To create lists with common peaks, the parameter "-wo" was used. This way, the original entries of both lists were written together with the number of base pairs of overlap between the two features. Only features with a minimum overlap of 1bp from the list used as a reference were reported. The same strategy was used to determine the number of peaks identified in this thesis that were common or unique when compared to the initial ChIP-chip binding sites identified by Castro et al., 2011.

3.16 Master table

During this work, a master table (Supplementary table S1, accessible in the GitHub repository online though the URL: https://github.com/rjoana1/Ascl1) was created, and layers of annotation retrieved by the different methods performed were added to annotate each of the Ascl1 E13.5 binding regions (peaks). The goal was to prepare a worksheet that could be processed on a user-friendly interface to extract data, display, and perform manual curation of results.

4 Results and discussion

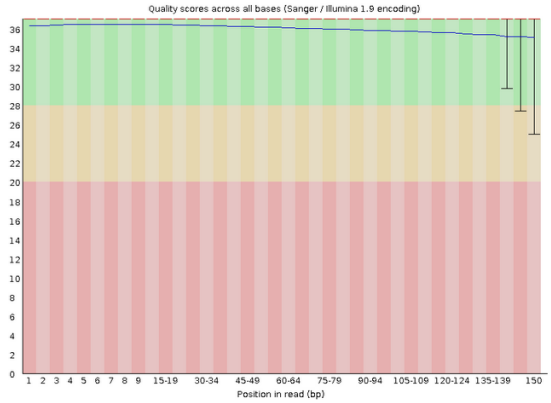
4.1 Quality control, peak calling and deriving lists of highly reproducible peaks

Sequenced reads quality assessment revealed no samples flagged as "poor quality" by fastqc (Table 2). When further inspecting the "Per base sequence quality" plots (Figure 8), showing the quality collectively across all reads within a sample, it is possible to observe that the median quality for any base in all analysed samples is higher than a Phred score of 30. This corresponds to a probability of less than 1 in 1000 that any base was called incorrectly. In some samples, the quality of reads slightly decreases toward the end of the reads, as seen by the whiskers dropping into the orange regions. This situation is often due to signal decay (degradation of the fluorophores) or phasing (loss of synchronicity by the cluster) during the sequencing run. As these are still scores of reasonable quality, the use of an alignment tool that omits ("soft clips") some bases at the ends of reads was the chosen strategy to achieve the greatest possible alignment scores.

Table 2: Quality assessment of sequenced reads files (fastq) of ChIP-seq data for Ascl1 and RBPJ transcription factors using fastqc version 0.11.9. Summary statistics on specific modules for each sample. Dev. Stage, developmental stage; SE, single-end sequencing; PE, paired-end sequencing.

	Dev. stage	Sample	Type of sequencing	Poor quality reads	Sequence Length (shortest-longest)
Ascl1	E11.5	Ascl1_11_SE	SE	0	39-76
		Ascl1_11_PE1	PE	0	150
		Ascl1_11_PE2	PE	0	150
	E13.5	Ascl1_13_PE1_1	PE	0	150
		Ascl1_13_PE1_2	PE	0	150
		Ascl1_13_PE2_1	PE	0	150
		Ascl1_13_PE1_2	PE	0	150
	E14.5	Ascl1_14_SE	SE	0	35-76
		Ascl1_14_PE1	PE	0	150
		Ascl1_14_PE2	PE	0	150
RBPJ	E11.5	RBPJ_11_PE1_1	PE	0	150
		RBPJ_11_PE1_2	PE	0	150
		RBPJ_11_PE2_1	PE	0	150
		RBPJ_11_PE2_2	PE	0	150

(a)



(b)

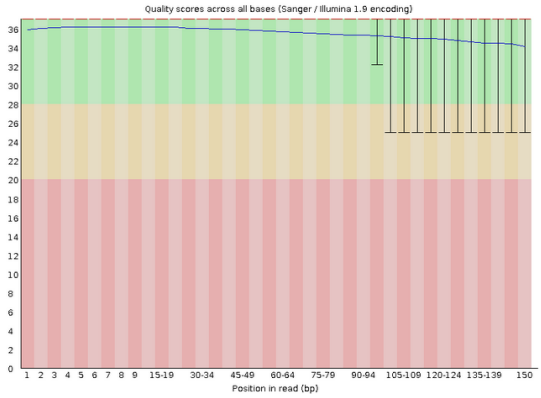
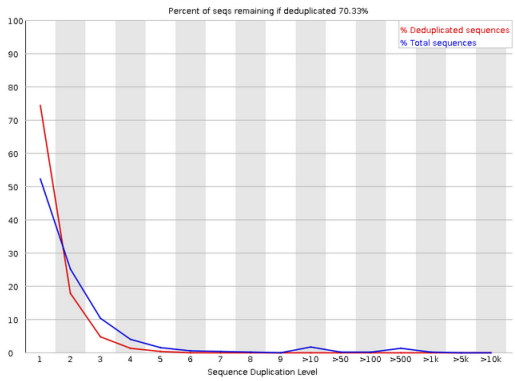


Figure 8: FASTQC "Per base sequence quality" plots. For each position, a boxplot is drawn with the median value, represented by the central red line, the inter-quartile range (25-75%), represented by the yellow box, the 10% and 90% values in the upper and lower whiskers and the mean quality, represented by the blue line. The y-axis shows the quality scores. The background of the graph divides the y-axis into very good quality scores (green), scores of reasonable quality (orange) and reads of poor quality (red). Plots for sample replicates Ascl1_13_PE1_1 (a) and Ascl1_13_PE1_2 (b) are shown.

The Sequence duplication level plots (Figure 9) indicate which proportion of the library corresponds to duplicates. Although RBPJ_11_PE2_1 fails this fastqc module (Figure 9b), the duplication levels observed are not concerning, suggesting the number of reads available for mapping and peak calling will not be reduced. Moreover, duplicates are removed in the next step of the ChIP-seq workflow.

(a)



(b)

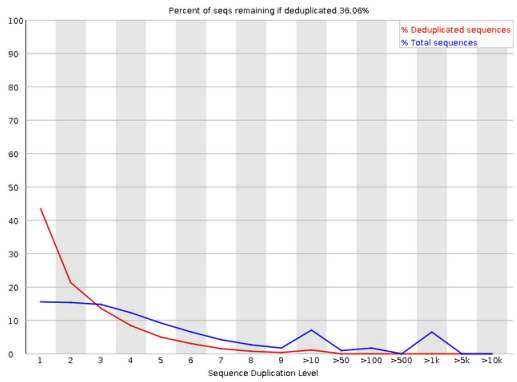


Figure 9: FastQC Sequence Duplication Levels plots. The plot (a) corresponds to the Ascl1_13_PE1_2 which passed this assessment. RBPJ_11_PE2_1 fails this module as shown on plot (b).

The next step of the pipeline consisted of filtering aligned reads to keep only uniquely mapped ones. Except for Ascl1_13_PE1 and Ascl1_14_SE, all samples show more than 20 million non-redundant

reads (Table 3), which is the minimum value recommended by the ENCODE project to analyse transcription factor ChIP-seq data.

Table 3: Summary statistics for each replicate before and after alignment and filtering steps. All metrics were obtained using Samtools flagstat 1.1.0.

	Dev. stage	Sample	Replicate	Total number of sequenced reads	Total number of mapped reads	Mapping ratio	% of reads filtered out	Non-redundant reads
Ascl1	E11.5	Ascl1_11_SE	1	40 330 317	38 197 497	94.71	38.1	23 641 577
		Ascl1_11_PE	2	59 914 236	47 535 439	79.34	40.9	28 096 169
	E13.5	Ascl1_13_PE1	1	40 475 710	27 951 967	69.06	40.7	16 572 290
		Ascl1_13_PE2	2	47 864 402	38 721 153	80.90	40.6	22 984 081
	E14.5	Ascl1_14_SE	1	30 792 551	29 758 028	96.64	39.5	17 998 996
		Ascl1_14_PE	2	45 569 050	37 914 917	83.20	40.6	22 539 807
RBPJ	E11.5	RBPJ_11_PE1	1	61 788 868	38 768 065	62.74	41.6	22 633 802
		RBPJ_11_PE2	2	50 701 338	41 763 381	82.37	41.0	24 627 663

The following checks set by ENCODE aimed primarily to verify appropriate enrichment of reads in peaks. All replicates show FRiP values greater than 1%, which is the minimum percentage of reads in peaks expected for a mammalian genome (Table 4). RSC values greater than 0.8 are observed for all replicates except for Ascl1_E11.5_2 and Ascl1_E13.2_1 (Table 4). This might suggest that there is lower enrichment in these two replicates.

Table 4: Summary statistics for each replicate after alignment and filtering steps. Relative Strand Correlation (RSC), Percentage of reads in peaks (FRiP) and within Blacklist regions were computed using the ChIPQC R package (v. 1.30.0).

	Dev. stage	Replicate	RSC	FRiP	Percentage of reads within Blacklist regions
Ascl1	E11.5	1	1.77	2.53	1.31
		2	0.734	2.91	1.52
	E13.5	1	0.631	6.13	3.37
		2	1.06	5.58	1.71
	E14.5	1	0.969	2.29	1.44
		2	0.912	5.24	1.48
RBPJ	E11.5	1	1.12	1.87	1.8
		2	1.32	2.13	1.49

Evidence of enrichment can also be seen by inspecting the coverage histograms (Figure 10). In these plots, all ChIP replicates present reasonable enrichment seen by existence of more positions (higher values on the y-axis) having higher sequencing depth, compared with the input samples.

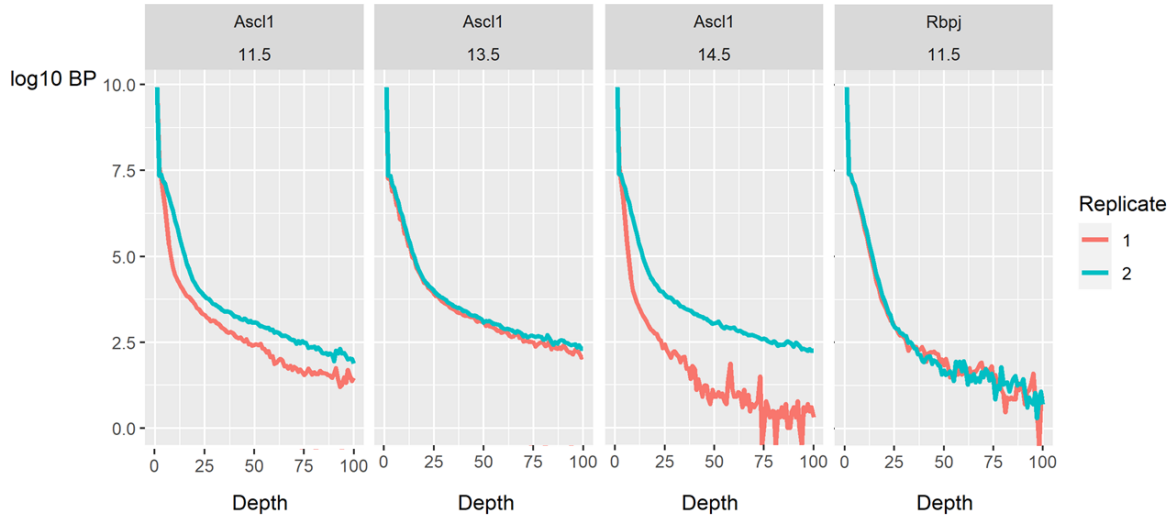


Figure 10: Coverage histograms generated using the ChIPQC package in Bioconductor. The X-axis represents the number of reads overlapping a single base in the reference genome (coverage), while the Y-axis represents the number (on a log scale) of base positions in the genome with exactly that level of coverage. For visualization purposes, a cut-off is applied at 100bp.

As part of the IDR pipeline, Ascl1 and RBPJ TFs bam files from each replicate were merged before calling peaks with MACS2 (p-value = 0,01). MACS2 successfully estimated the fragment size d to be 205 bp and 194 bp for Ascl1 E11.5 and Ascl1 E14.5 merged replicates, respectively (Figure 11). As expected, the model was not generated for Ascl1 E13.5 or RBPJ E11.5 paired-end samples.

(a) (b)

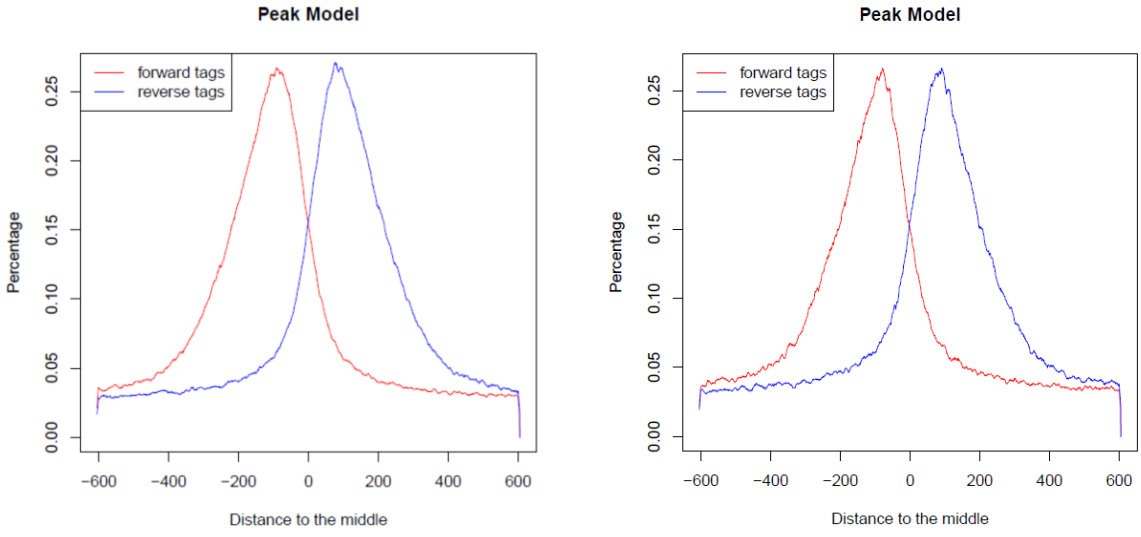


Figure 11: Peak model built by MACS2 using Ascl1 E11.5 (a) and Ascl1 E14.5 (b) data sets. The red curve represents the percentage of positive strand reads at each base pair, and the blue curve models reads on the negative strand.

To evaluate the consistency of peaks between two replicates, the IDR algorithm (Figure 6) was applied to all datasets under study. The number of peaks that passed the 5% IDR threshold (values shown in Figure 12) were used to truncate the sorted merged replicate peak files. Except for RBPJ, where an optimal peak rank threshold was used, all Ascl1 peak rank thresholds originated conservative lists (see 2.3.9.1).

Blacklisted regions were then filtered out from this peak lists. Results show that genome-wide mapping of Ascl1 binding profile by ChIP-seq identified 2337 high-confident binding events at E11.5, 3054 at E13.5 and 6218 at E14.5. RBPJ E11.5 binding profile retrieved 1431 high confident peaks.

The IDR pipeline also computes IDR QC scores to check if the data meets the ENCODE standards. Results show that both the self-consistency ratio (N1/N2) and rescue ratio (Np/Nt) for RBPJ E11.5 data have values less than 2, indicating a concerning data status (Table 5).

There are several batch effects that could have affected data quality of Ascl1 and RBPJ datasets during the sample preparation step: sample replicates performed in different days, library preparation and sequencing steps carried out in different days, at different settings and using different sequencing methods.

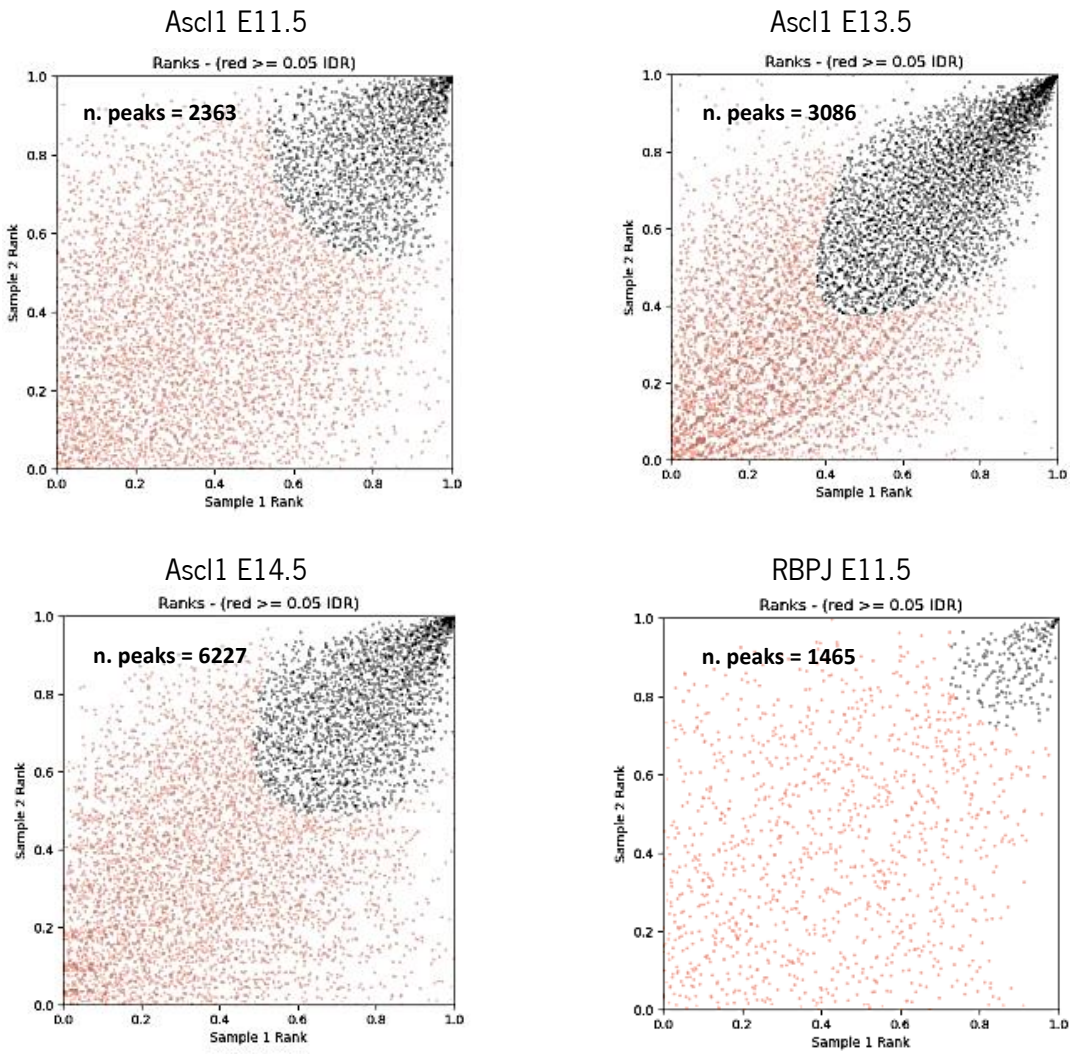


Figure 12: IDR output plots for each sample. Replicate 1 peak ranks versus Replicate 2 peak ranks. Peaks that do not pass an IDR threshold of 0.05 are coloured red.

Table 5: Ascl1 and RBPJ datasets classification according to ENCODE standards.

	Self-consistency Ratio	Rescue Ratio	ENCODE Data Status
Ascl1 E11.5	Greater than 2	Less than 2	Acceptable
Ascl1 E13.5	Less than 2	Less than 2	Ideal
Ascl1 E14.5	Greater than 2	Less than 2	Acceptable
RBPJ E11.5	Greater than 2	Greater than 2	Concerning

4.2 Understanding the Ascl1 and RBPJ transcriptional network

4.2.1 Genomic features at Ascl1 target sites

To understand which genomic regions were associated with Ascl1 E13.5 peaks, binding events were first annotated with the respective genomic feature using the nearest gene method. When a range of 1kb upstream and downstream from TSS was defined as the promoter region, most of Ascl1 E13.5 binding events occur within intron regions and distal intragenic elements (62.4%) (Figure 13 (A)). A similar result was found when binding site locations were analysed relatively to the TSS of the associated genes using GREAT. According to the association rule established (see Methods), most of the input regions are at 50 kb to 500 kb upstream or downstream of their putatively regulated genes (Figure 13 (B)). This binding pattern is consistent with Ascl1 binding to distal enhancers, but also via proximal promoter regions (Raposo et al., 2015).

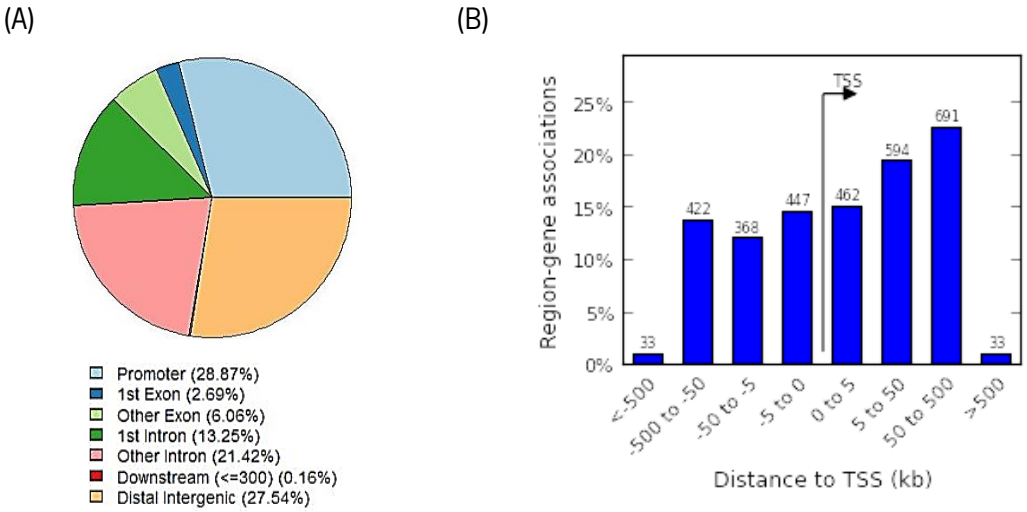


Figure 13: (A) Location of Ascl1 E13.5 binding events respective to genomic features using the nearest gene method for annotation (ChIPseeker R package v1.28.3). (B) Distance between Ascl1 E13.5 binding regions and their putatively regulated TSS genes using GREAT v4.0.4.

Next, a Hidden-Markov model was used to characterize the chromatin states at E13.5 Ascl1-bound regions (Figure 14). CpG islands, Exons, TSS and regions within 2kb of the TSS are shown to be associated with E6 or Active TSS, as expected, which corresponds to the co-occurrence of H3k27ac and H3K4me3. Ascl1 E13.5 binding sites fall mostly within three different chromatin states. One highly co-enriched for H3K27ac and H3K4me3, characteristic of active TSS (state 6). A second one, highly enriched for the three marks used, H3k27ac, H3K4me1 and H3K4me3, is characteristic of flanking regions

upstream TSSs (state 2). The third region of chromatin is a co-enrichment for H3K4me1 and H3K27ac of less intensity, located further away from each side of the summit point, and characteristic of active enhancers. Overall, results are in line with Ascl1 regulating gene expression via distal enhancers, but also by binding to proximal promoter regions.

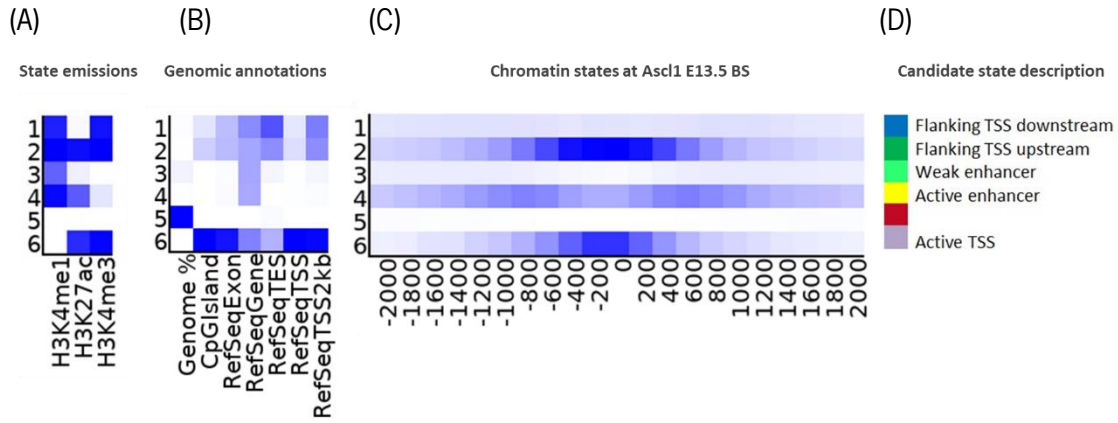


Figure 14: Chromatin state model and chromatin states at Ascl1 E13.5 binding regions. (A) Heatmap of the emission parameters in which each row corresponds to a different state, and each column corresponds to a different mark for the model defined based on the data for three histone modifications (H3K4me1, H3K4me3, and H3K27ac) from Lindtner et al., 2017. The darker blue colour corresponds to a greater probability of observing the mark in the state. (B) The heatmap displays the overlap fold enrichment for various external genomic annotations. A darker blue colour corresponds to a greater fold enrichment for a column-specific colouring scale. (C) The heatmap shows the fold enrichment for each state for each 200-bp bin position within 2 kb around the peak summits of Ascl1 E13.5 binding sites (BS). A darker blue colour corresponds to a greater fold enrichment, and there is one colour scale for the entire heatmap. (D) Candidate-state descriptions for each state.

4.2.2 Comparison across different developmental stages

Next, to identify binding events associated with neuronal differentiation or proliferation, binding coordinates identified by peak calling from genome-wide mapping of Ascl1 across developmental stages (E11.5 corresponding to a more proliferative and E13.5 to a more differentiated population of progenitor cells) were compared. Intersection of Ascl1 E11.5 and E13.5 lists identified 1905 common binding sites (Figure 15). Approximately half of E13.5 binding sites are specific for this developmental state.

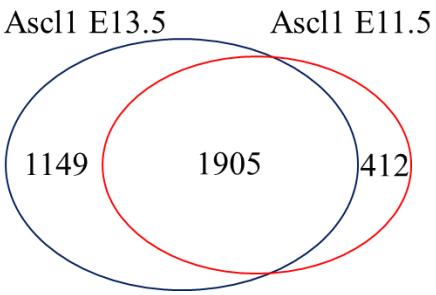


Figure 15: Venn diagram showing the overlap between binding events associated with Ascl1 E13.5 (blue) and binding events associated with Ascl1 E11.5 (red). The number of peaks in each section of the diagram is indicated.

Evaluation of read density across summits from the total number of peaks identified on each stage shows that the read enrichment signal from Ascl1 binding at E13.5 peak summits is stronger (more than 2.6 times) than the signal at E11.5 (Figure 16).

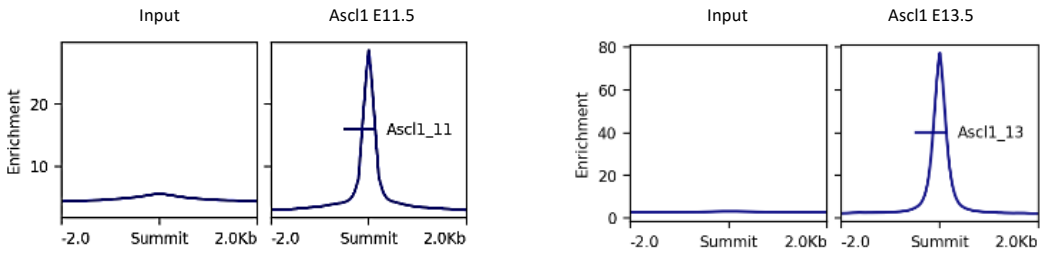


Figure 16: Profile plots of Ascl1 and corresponding input ChIP-seq read signal from merged replicates within ±2 kb of peak summits. Signal intensity represents average peak coverage.

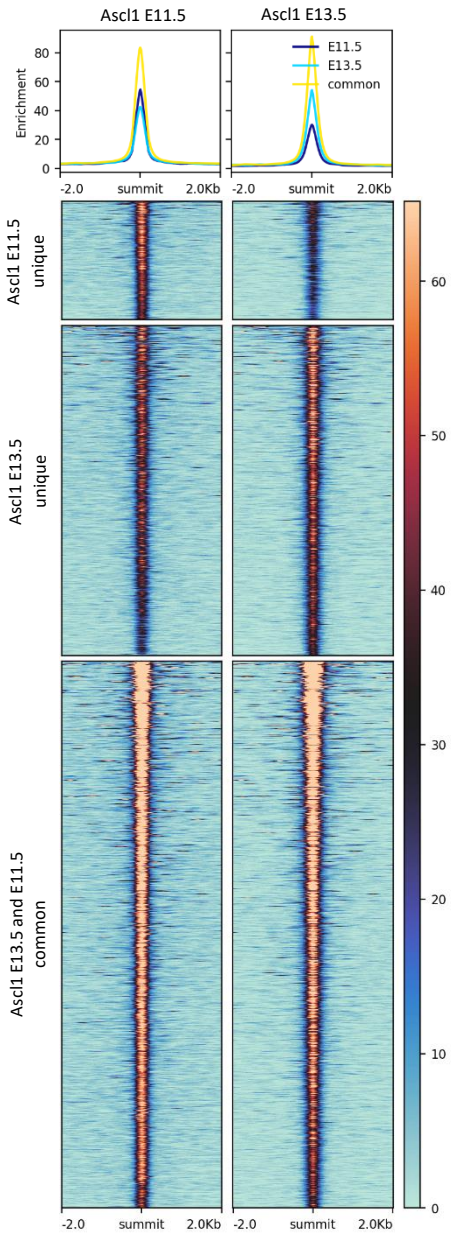


Figure 17: Density plots and corresponding heatmaps from Ascl1 at E11.5 (left panel) and E13.5 (right panel) peak summits. Read enrichment for specific peaks identified for Ascl1 at E11.5 (unique), Ascl1 at E13.5 (unique) and common peaks shared by both developmental stages (common) are shown.

When focusing on stage-specific peaks, the density of sequenced reads at peak summits on each stage shows a strong signal, but a clear enrichment in the other stage is still observed despite no peaks were identified by peak calling (Figure 17). It might be that peaks are, in fact, present, only just below the peak calling threshold. This suggests that the peak calling algorithm threshold used for peak detection may be too conservative. Indeed, when visually comparing the two stages, it is possible to observe that peaks associated with DLX1 (Figure 18 (A)) and Map3k1 (Figure 18 (B)) identified by the peak calling algorithm only at E13.5 are also visible at E11.5. Moreover, peaks associated with Insm1 (Figure 18 (C))

and Fbxw7 (Figure 18 (D)) genes identified by MACS2 in both stages, present very different enrichments across stages, suggesting quantitative differences. These may correspond to real differences between stages. To identify them, an alternative approach would be to follow a more quantitative comparison, such as the R package DiifBind (Stark & Brown, 2021) to identify genomic intervals where confidence statistics could be computed to characterise the likelihood of a difference in enrichment between developmental stages at each binding site. Nevertheless, the analysis performed did not uncover evidence of stage specific peaks across samples.

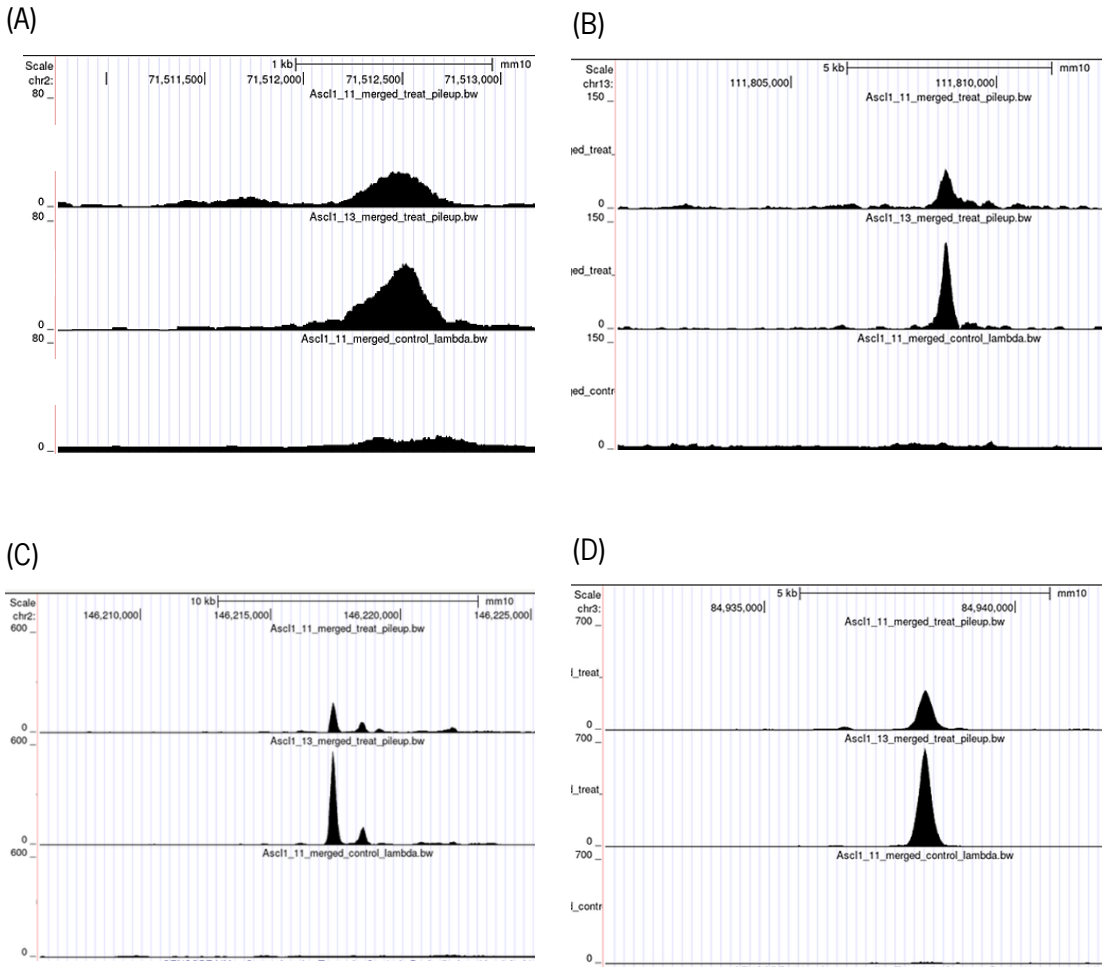


Figure 18: Ascl1 ChIP-seq enrichment profiles across developmental stages in the vicinity of selected Ascl1 bound genes. (A) Dlx1, (B) Map3k1, (C) Insm1 and (D) Fbxw7. First and second rows in each figure correspond to E11.5 and E13.5 developmental stages, respectively. The last row corresponds to the input sample. Image adapted from the UCSC genome browser.

4.2.3 Ascl1 crosstalk with Notch pathway

To validate the possibility that the Notch pathway might be co-regulating Ascl1 putative target genes and its importance in neuronal development, RBPJ ChIP-Seq from mouse embryonic telencephalon at E11.5 data (Figure 19) was analysed. As demonstrated before, the quality of RBPJ ChIP-Seq dataset is lower but allow us to identify binding regions bound by both TFs by intersecting both lists. This approach generated a list of 105 common binding sites, corresponding to putative cis-regulatory regions (Figure 20). At this point, identification of those co-bound regions was a valuable result, however information about RBPJ and Ascl1 co-regulation data is still lacking.

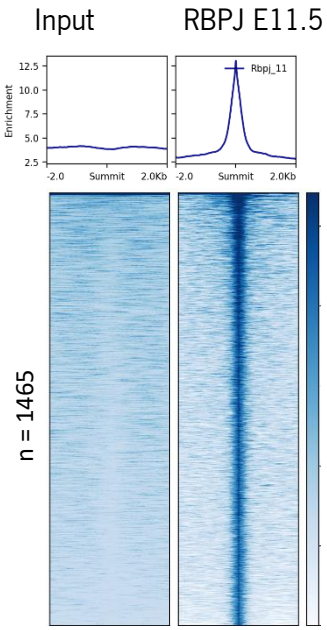


Figure 19: Density plots of RBPJ and corresponding input ChIP-seq reads from merged replicates within ±2 kb of peak summits. Signal intensity represents average peak coverage. Number on the side of heatmap represents the number of peaks called after IDR analysis.

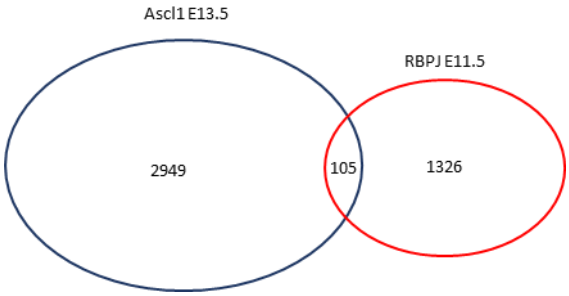


Figure 20: Venn diagram showing the overlap between binding events associated with Ascl1 E13.5 (red) and binding events associated with RBPJ E11.5 (blue). The number of peaks in each section of the diagram is indicated.

4.3 Finding over-represented DNA motifs on E13.5 binding regions

To determine the DNA sequence mediating Ascl1 binding, motifs enriched within 50-bp up- and downstream of Ascl1 peak summits were searched. Results show the hexamer sequence CAGCTG, corresponding to the E-box sequence previously associated with Ascl1 binding (Castro et al., 2011) and in agreement with its role in mediating direct DNA binding (Figure 21). Another PWM, GTGGGAACC, matching the consensus binding sequence for RBPJ was overrepresented in binding regions common to both Ascl1 and RBPJ.

(A)

Freq	Score	FDR
47553	1154.3.4	<1E(-4)

(B)

12	4.5	0.04

Figure 21: Enriched DNA motifs associated with E13.5 Ascl1 binding events (A) and with E13.5 Ascl1 binding events shared with RBPJ binding events (B). Motifs correspond to extended consensus binding sequences.

4.4 Gene annotation of E13.5 Ascl1 binding events

Biological interpretation of binding events requires prior knowledge of which genes are associated with the binding sites. According to the association rule established for peak-gene annotation using the GREAT tool, this strategy successfully assigned 99.4% of Ascl1 genomic regions to one gene. Ascl1 putatively regulates 3036 genes.

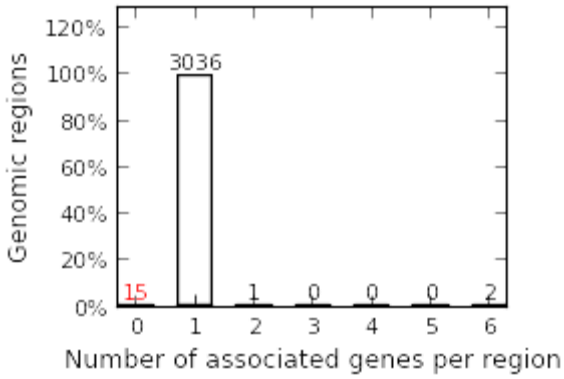


Figure 22:Number of genes associated with E13.5 Ascl1 binding sites using GREAT.

4.5 Integration of Ascl1 E13.5 loss-of-function (LoF) expression data

The binding of a transcription factor does not necessarily mean a regulatory function. To assess the impact of Ascl1 binding on the regulation of its putative target genes, ChIP-seq data was integrated with expression profiling data characterising transcriptional changes in ventral telencephalon of the Ascl1 null mutant at E13.5 stage of development.

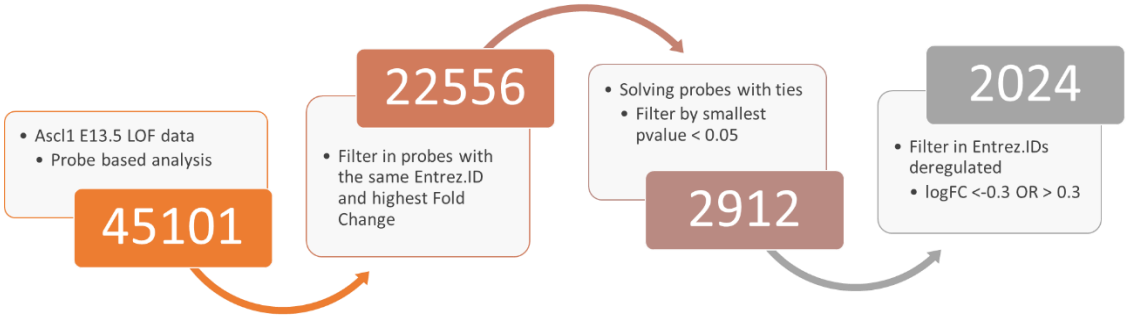


Figure 23: Annotation pipeline of Ascl1 E13.5 binding sites with LOF expression data. Results from each filtering step are shown across the pipeline.

The annotation strategy to assign Ascl1 E13.5 peaks with expression values resulted in a list of 2024 Ascl1 Entrez genes corresponding to deregulated genes in Ascl1 LoF experiments. A significant number of probes did not pass the filters developed in this strategy. Reasons included probes lacking an Entrez ID annotation and multiple probes interrogating the same gene (each interrogating a different mRNA transcript). Intersection of the lists of Ascl1-bound genes and of genes deregulated in Ascl1 LoF experiments identified 643 likely direct targets of Ascl1 at E13.5 (Figure 24 (A)). Of these, 62% (n = 398) were positively regulated and 38% (n = 245) negatively regulated by Ascl1 (Figure 24 (B)). The majority of Ascl1 bound and regulated genes are downregulated in the Ascl1 null embryo, in line with its described activity as a transcriptional activator (Raposo et al., 2015). In such case, upregulation of Ascl1 bound genes may be a consequence of compensatory mechanisms operating in Ascl1 null mutant embryos, where redundancy with other differentiation TFs, together with loss of Notch signalling, results in premature neuronal differentiation in the VZ (Casarosa et al., 1999; Castro et al., 2011). Another possibility, is that in many cases the "nearest gene" annotation does not properly associate an Ascl1 binding event with its target gene. One possible way to improve such association, is to consider the location of so-called topologically associated domains (TADs) (Bonev et al., 2017). TADs define the boundaries inside which long range interaction between promoters and cis-regulatory enhancers can take place. Since neural differentiation has been associated with highly cell type-specific 3D remodelling (Rajarajan et al., 2018), binding of Ascl1 to distal enhancers might occur by chromatin loop formation of

enhancer–promoter interactions that is not explained by the nearest-gene annotation used in this study (Kuang & Wang, 2021).

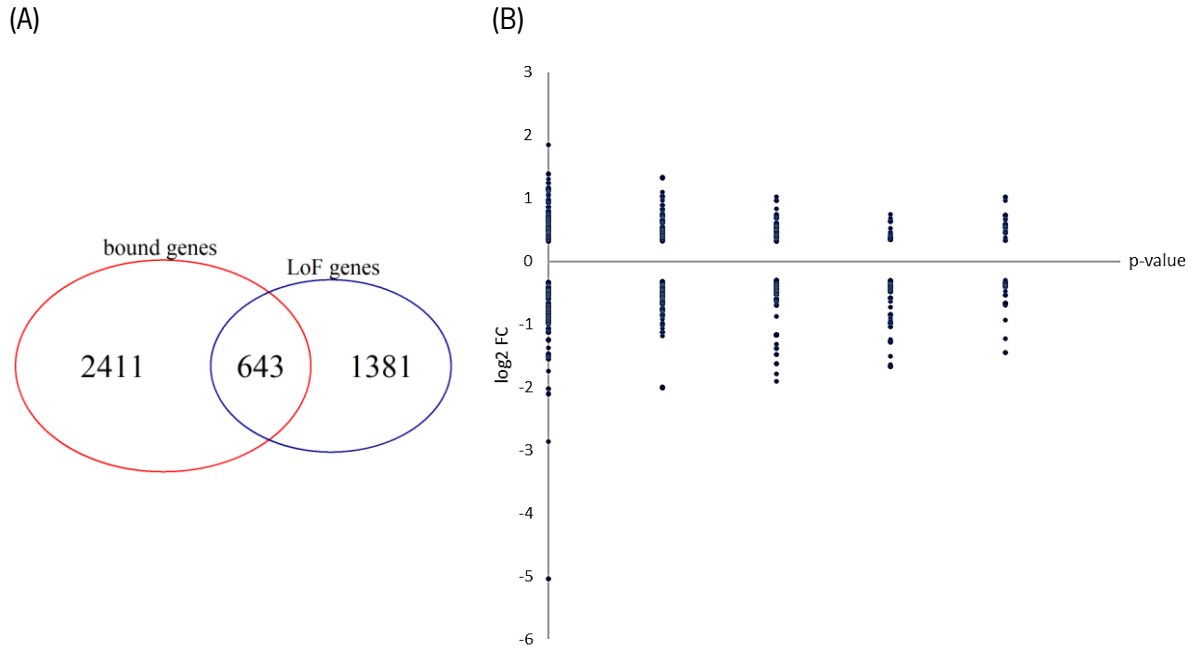


Figure 24: Comparison of Ascl1 E13.5 target genes with expression data derived from Ascl1 mutant embryonic ventral telencephalon. (A) Venn diagram showing the overlap between genes associated with Ascl1-binding events (red) and genes deregulated in Ascl1 LoF experiments (blue). (B) Graphical representation of Ascl1 E13.5 bound and deregulated genes in loss-of-function (LoF) in embryonic telencephalon. Log2 fold changes (cut-off <-0.3 or >0.3) of genes regulated are plotted against the associated p-value (cut-off <0.05).

4.6 Functional annotation of E13.5 Ascl1 target genes by Gene Ontology

Functional meaning of the target genes (i.e., bound and regulated) to detect statistically significant signalling pathways regulated by Ascl1 at E13.5 was achieved using GREAT (Figure 24). In line with previous studies (Castro et al., 2011; Raposo et al., 2015), significantly enriched GO terms (FDR < 0.05) were found to correspond to different states of neurogenesis. “Neural precursor cell proliferation” and “stem cell population maintenance” are terms related with Ascl1 role in proliferation neural precursor cells, resulting in the expansion of a cell population; “stem cell differentiation” and “neuron fate commitment” terms are consistent with Ascl1 role in the switch from proliferation to specialization of NPs, when these become restricted to develop into neurons. The “regulation of cell cycle arrest” prompted the inspection of the final list of Ascl1 E13.5 target genes (Supplementary Table S1) for known regulators of the cell cycle. These included the Gadd45g and Fbxw7 genes, known to promote cell cycle arrest, but also positive cell cycle regulators such as E2f1, Cdk6 and Ccnd1, essential for G1/S transition and Cdca7 and Cdc25b, necessary for entry in mitosis (Castro et al., 2011). These targets were shown to be downregulated when Ascl1 is absent. As previously described, results provide evidence of a functional

control of Ascl1 in G1-S and G2-M transitions (Castro et al., 2011). Ascl1 targets were also associated with "GABAergic neuron differentiation". Gad2 gene, specific of this type of neurons (Castro & Guillemot, 2011), is among Ascl1 E13.5 targets. As expected, the "regulation of Notch signalling pathway" term has also found in this GO analysis. Ascl1 regulates Notch/Hes pathway at different levels (Castro & Guillemot, 2011). The list of bound and deregulated genes included Notch1 receptor, the Notch ligands Dll1 and Dll3, the downstream Notch targets Hes5 and Hes6. Interestingly, Ascl1 targets are involved in "ephrin receptor signalling pathway" in the telencephalon, which is concordant with its direct role in regulating neuron migration (Y. Liu et al., 2017). When scanning the list of target genes, Ascl1 regulates various members of this pathway: Ephb2, Ephb3 and Epha3.

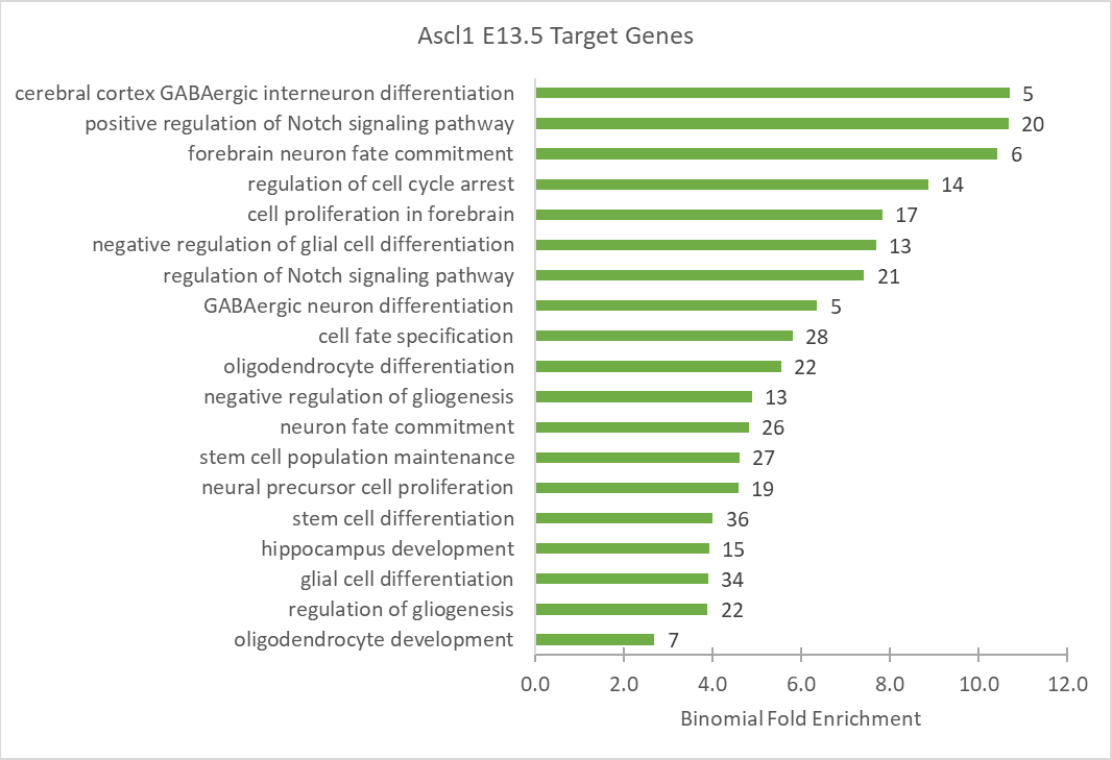


Figure 25: Selected Enrichment of Gene Ontology biological process terms among E13.5 Ascl1 target genes in ventral telencephalon. Total number of genomic regions associated with each term are shown at the end of each bar.

4.7 Comparison between ChIP-seq and ChIP-chip Ascl1 binding sites

The first genome-wide characterization of Ascl1 transcriptional program in the embryonic mouse brain, performed by ChIP-chip (hybridization of immunoprecipitated DNA to promoter oligonucleotide arrays) (Castro et al., 2011), identified 1265 promoters significantly bound by Ascl1 at E12.5. The restriction to proximal promoter regions, excluding genes bound by Ascl1 to distal enhancers, and the need to validate the model with a more robust experimental approach, prompted the use of ChIP-seq to

obtain a map of binding sites at genome-wide scale and with higher resolution. After successfully converting all mm9 genomic coordinates to the mm10 genome assembly, the number of peaks derived from the two techniques was compared. Using the ChIP-chip list as a reference, 835 out of 1265 (66.0%) binding sites do not overlap with any of the peak coordinates identified by ChIP-seq. When assessing how many peaks are in common, 430 out of 1265 (34,0%) peaks are shared between both lists. In theory, it was expected that nearly all ChIP-chip peaks had been identified by ChIP-seq. This difference might be explained, in part, by the fact that ChIP-chip peaks identified and validated at the promotors of Castro et al. (2011) study tended to present small enrichments, which were nevertheless highly reproducible by ChIP-PCR. Most likely, a large number of these binding events could be captured if the significance threshold used to call the ChIP-seq peaks was decreased. Another difference to take into account was the use of different peak calling algorithms. Nevertheless, when comparing the number of peak coordinates retrieved by both methods, as expected, ChIP-seq identified nearly the double of high-confident binding events.

5 Conclusion and future work

This work contributes to further elucidate the transcriptional regulation of neurogenesis by the proneural factor Ascl1. First, high confidence Ascl1 bound regions were mapped in the ventral telencephalon at a genome-wide scale. Then, the chromatin states of genomic regions associated with Ascl1 recruitment were characterized, concluding that these bear marks of distal enhancers, but also proximal promoter regions. This work corroborates previous findings, showing that Ascl1 coordinates neurogenesis by regulating a large number of target genes with a wide variety of biological functions and associated with different stages of neurogenesis.

Future work should address how Ascl1 coordinates this complex transcriptional program along the neuronal lineage. This could explore a possible crosstalk with the Notch program, taking advantage of the regulatory regions identified where Ascl1 is co-recruited by RBPJ, as assessed by ChIP-seq.

References

Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, *9*(1), 1–5. https://doi.org/10.1038/s41598-019-45839-z

Andersson, R., & Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, *21*(2), 71–87. https://doi.org/10.1038/s41576-019-0173-8

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Angarica, V. E., & del Sol, A. (2017). Bioinformatics tools for genome-wide epigenetic research. *Advances in Experimental Medicine and Biology*, *978*, 489–512. https://doi.org/10.1007/978-3-319-53889-1_25

Baker, M. (2011). Making sense of chromatin states. *Nature Methods*, *8*(9), 717–722. https://doi.org/10.1038/nmeth.1673

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., & Lander, E. S. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, *125*(2), 315–326. https://doi.org/10.1016/j.cell.2006.02.041

Bertrand, N., Castro, D. S., & Guillemot, F. (2002). Proneural genes and the specification of neural cell types. *Nature Reviews Neuroscience*, *3*(7), 517–530. https://doi.org/10.1038/nrn874

Blanco, E., & Abril, J. F. (2004). *ENCODE (Encyclopedia of DNA Elements)*. Dictionary of Bioinformatics and Computational Biology. https://doi.org/10.1002/9780471650126.dob0886

Boeva, V. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells. *Frontiers in Genetics*, *7*(FEB). https://doi.org/10.3389/fgene.2016.00024

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J. P., Tanay, A., & Cavalli, G. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, *171*(3), 557-572.e24. https://doi.org/10.1016/j.cell.2017.09.043

Borromeo, M. D., Meredith, D. M., Castro, D. S., Chang, J. C., Tung, K., Guillemot, F., Johnson, J. E., Borromeo, M. D., Meredith, D. M., Castro, D. S., Chang, J. C., Tung, K., Guillemot, F., & Johnson, J. E. (2014). A transcription factor network specifying inhibitory versus excitatory neurons in the dorsal spinal cord. *Development (Cambridge)*, *141*(14), 2803. https://doi.org/10.1242/dev.105866

Calo, E., & Wysocka, J. (2013). Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell*, *49*(5), 825–837. https://doi.org/10.1016/j.molcel.2013.01.038

Carlson, M. (2019). *org.Mm.eg.db: Genome wide annotation for Mouse. R package version 3.8.2.* https://doi.org/10.18129/B9.bioc.org.Mm.eg.db

Casarosa, S., Fode, C., & Guillemot, F. (1999). Mash1 regulates neurogenesis in the ventral telencephalon. *Development*, *126*(3), 525–534. https://doi.org/10.1242/dev.126.3.525

Castro, D. S., & Guillemot, F. (2011). Old and new functions of proneural factors revealed by the genome-wide characterization of their transcriptional targets. *Cell Cycle*, *10*(23), 4026–4031.

https://doi.org/10.4161/cc.10.23.18578

Castro, D. S., Martynoga, B., Parras, C., Ramesh, V., Pacary, E., Johnston, C., Drechsel, D., Lebel-Potter, M., Garcia, L. G., Hunt, C., Dolle, D., Bithell, A., Ettwiller, L., Buckley, N., & Guillemot, F. (2011). A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. *Genes and Development*, *25*(9), 930–945. https://doi.org/10.1101/gad.627811

Chong, C. F., Li, Y. C., Wang, T. L., & Chang, H. (2003). Stratification of adverse outcomes by preoperative risk factors in coronary artery bypass graft patients: an artificial neural network prediction model. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *June 1997*, 160–164.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801. https://doi.org/10.1093/nar/gkx1081

Docker. (n.d.). https://www.docker.com/

ENCODE: Encyclopedia of DNA Elements. (n.d.). https://www.encodeproject.org/

Ernst, J., & Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols*, *12*(12), 2478–2492. https://doi.org/10.1038/nprot.2017.124

Farah, M. H., Olson, J. M., Sucic, H. B., Hume, R. I., Tapscott, S. J., & Turner, D. L. (2000). *Farah MH, 2000.pdf*. *702*, 693–702.

Feng, J., Liu, T., Qin, B., Zhang, Y., & Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, *7*(9), 1728–1740. https://doi.org/10.1038/nprot.2012.101

Gorkin, D. U., Barozzi, I., Zhao, Y., Zhang, Y., Huang, H., Lee, A. Y., Li, B., Chiou, J., Wildberg, A., Ding, B., Zhang, B., Wang, M., Strattan, J. S., Davidson, J. M., Qiu, Y., Afzal, V., Akiyama, J. A., Plajzer-Frick, I., Novak, C. S., ... Ren, B. (2020). An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature*, *583*(7818), 744–751. https://doi.org/10.1038/s41586-020-2093-3

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., ... Kent, W. J. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, *34*(Database issue), 590–598. https://doi.org/10.1093/nar/gkj144

Homem, C. C. F., Repic, M., & Knoblich, J. A. (2015). Proliferation control in neural stem and progenitor cells. *Nature Reviews Neuroscience*, *16*(11), 647–659. https://doi.org/10.1038/nrn4021

Illumina, I. (n.d.). *FASTQ files explained*. https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html

Imayoshi, I., Ishidate, F., & Kageyama, R. (2015). Real-time imaging of bHLH transcription factors reveals their dynamic control in the multipotency and fate choice of neural stem cells. *Frontiers in Cellular Neuroscience*, *9*(AUGUST), 1–6. https://doi.org/10.3389/fncel.2015.00288

Imayoshi, I., & Kageyama, R. (2014). bHLH factors in self-renewal, multipotency, and fate choice of neural progenitor cells. *Neuron*, *82*(1), 9–23. https://doi.org/10.1016/j.neuron.2014.03.018

Kageyama, R., Shimojo, H., & Ohtsuka, T. (2019). Dynamic control of neural stem cells by bHLH factors.

Neuroscience Research, *138*, 12–18. https://doi.org/10.1016/j.neures.2018.09.005

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, and D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

Kidder, B. L., Hu, G., & Zhao, K. (2011). ChIP-Seq: Technical considerations for obtaining high-quality data. *Nature Immunology*, *12*(10), 918–922. https://doi.org/10.1038/ni.2117

Kriegstein, A., & Alvarez-Buylla, A. (2009). The glial nature of embryonic and adult neural stem cells. *Annual Review of Neuroscience*, *32*, 149–184. https://doi.org/10.1146/annurev.neuro.051508.135600

Kuang, S., & Wang, L. (2021). Deep Learning of Sequence Patterns for CCCTC-Binding Factor-Mediated Chromatin Loop Formation. *Journal of Computational Biology*, *28*(2), 133–145. https://doi.org/10.1089/cmb.2020.0225

Kundaje, A. (n.d.). *Schema for Uniform TFBS - Transcription Factor ChIP-seq Uniform Peaks from ENCODE/Analysis*. Retrieved May 5, 2021, from http://genome.ucsc.edu/cgi-bin/hgTables?db=hg19&hgta_group=regulation&hgta_track=wgEncodeAwgTfbsUniform&hgta_table=wgEncodeAwgTfbsHaibK562Mef2aV0416101UniPk&hgta_doSchema=describe+table+schema

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., … Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, *22*(9), 1813–1831. https://doi.org/10.1101/gr.136184.111

Langmead, B., & Salzberg, S. L. (n.d.). *Bowtie2*. http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#local-alignment-example

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Q., Brown, J. B., Huang, H., & Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, *5*(3), 1752–1779. https://doi.org/10.1214/11-AOAS466

Lindtner, S., Catta-Preta, R., Tian, H., Su-Feher, L., Price, J. D., Dickel, D. E., Greiner, V., Silberberg, S. N., McKinsey, G. L., McManus, M. T., Pennacchio, L. A., Visel, A., Nord, A. S., & Rubenstein, J. L. R. (2019). Genomic Resolution of DLX-Orchestrated Transcriptional Circuits Driving Development of Forebrain GABAergic Neurons. *Cell Reports*, *28*(8), 2048-2063.e8. https://doi.org/10.1016/j.celrep.2019.07.022

Liu, B., Yang, J., Li, Y., McDermaid, A., & Ma, Q. (2018). An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data. *Briefings in Bioinformatics*, *19*(5), 1069–1081. https://doi.org/10.1093/bib/bbx026

Liu, Y., Tsai, J.-W., Chen, J.-L., Yang, W., Chang, P.-C., Cheng, P.-L., Turner, D. L., Yanagawa, Y., Wang, T.-W., & Yu, J.-Y. (2017). Ascl1 promotes tangential migration and confines migratory routes by induction of Ephb2 in the telencephalon. *Scientific Reports*, *7*(March), 1–17. https://doi.org/10.1038/srep42895

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, *28*(5), 495–501. https://doi.org/10.1038/nbt.1630

Nakada, Y., Hunsaker, T. L., Henke, R. M., & Johnson, J. E. (2004). Distinct domains within Mash1 and Math1 are required for function in neuronal differentiation versus neuronal cell-type specification. *Development*, *131*(6), 1319–1330. https://doi.org/10.1242/dev.01008

Nakato, R., & Sakata, T. (2020). Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods*, *February*, 1–10. https://doi.org/10.1016/j.ymeth.2020.03.005

Nakato, R., & Shirahige, K. (2017). Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Briefings in Bioinformatics*, *18*(2), 279–290. https://doi.org/10.1093/bib/bbw023

Pagès, H., Carlson, M., Falcon, S., & Li, N. (2021). *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. R package version 1.56.2*. https://bioconductor.org/packages/AnnotationDbi

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rajarajan, P., Borrman, T., Liao, W., Schrode, N., Flaherty, E., Casiño, C., Powell, S., Yashaswini, C., LaMarca, E. A., Kassim, B., Javidfar, B., Espeso-Gil, S., Li, A., Won, H., Geschwind, D. H., Ho, S. M., MacDonald, M., Hoffman, G. E., Roussos, P., … Akbarian, S. (2018). Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science*, *362*(6420). https://doi.org/10.1126/science.aat4311

Raposo, A. A. S. F., Vasconcelos, F. F., Drechsel, D., Marie, C., Johnston, C., Dolle, D., Bithell, A., Gillotin, S., van den Berg, D. L. C., Ettwiller, L., Flicek, P., Crawford, G. E., Parras, C. M., Berninger, B., Buckley, N. J., Guillemot, F., & Castro, D. S. (2015). Ascl1 coordinately regulates gene expression and the chromatin landscape during neurogenesis. *Cell Reports*, *10*(9), 1544–1556. https://doi.org/10.1016/j.celrep.2015.02.025

Rocha, M., & Ferreira, P. G. (2018). Bioinformatics Algorithms. In *Bioinformatics Algorithms*. Elsevier. https://doi.org/10.1016/B978-0-12-812520-5.00019-5

Sharov, A. A., & Ko, M. S. H. (2009). Exhaustive search for over-represented DNA sequence motifs with cisfinder. *DNA Research*, *16*(5), 261–273. https://doi.org/10.1093/dnares/dsp014

Soares, D. S., Homem, C. C. F., & Castro, D. S. (2022). Function of Proneural Genes Ascl1 and Asense in Neurogenesis: How Similar Are They? *Frontiers in Cell and Developmental Biology*, *10*(February), 1–9. https://doi.org/10.3389/fcell.2022.838431

Stark, R., & Hadfi, J. (2016). Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing. In *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*. https://doi.org/10.1007/978-3-319-31350-4

Steinhauser, S., Kurzawa, N., Eils, R., & Herrmann, C. (2016). A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in Bioinformatics*, *17*(6), 953–966. https://doi.org/10.1093/bib/bbv110

Sueda, R., & Kageyama, R. (2020). Regulation of active and quiescent somatic stem cells by Notch signaling. *Development Growth and Differentiation*, *62*(1), 59–66. https://doi.org/10.1111/dgd.12626

Tarasov, A. (2016). *[sambamba view] Filter expression syntax*. https://github.com/biod/sambamba/wiki/%5Bsambamba-view%5D-Filter-expression-syntax

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*, *31*(12), 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Tran, N. T. L., & Huang, C. H. (2014). A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biology Direct*, *9*(1). https://doi.org/10.1186/1745-6150-9-4

Turrero García, M., & Harwell, C. C. (2017). Radial glia in the ventral telencephalon. *FEBS Letters*, *591*(24), 3942–3959. https://doi.org/10.1002/1873-3468.12829

Vasan, L., Park, E., David, L. A., Fleming, T., & Schuurmans, C. (2021). Direct Neuronal Reprogramming: Bridging the Gap Between Basic Science and Clinical Application. *Frontiers in Cell and Developmental Biology*, *9*(July), 1–29. https://doi.org/10.3389/fcell.2021.681087

Vasconcelos, F. F., & Castro, D. S. (2014). Transcriptional control of vertebrate neurogenesis by the proneural factor ascll. *Frontiers in Cellular Neuroscience*, *8*(DEC), 1–6. https://doi.org/10.3389/fncel.2014.00412

Vu, H., & Ernst, J. (2021). Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *BioRxiv*, *7*(1), 1–37. https://doi.org/10.1186/s13059-021-02572-z

Yu, G., Wang, L. G., & He, Q. Y. (2015). ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, *31*(14), 2382–2383. https://doi.org/10.1093/bioinformatics/btv145

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., & Shirley, X. S. (2008). *MACS – Model-based Analysis of ChIP-Seq*. Genome Biology. https://doi.org/10.1186/gb-2008-9-9-r137