



LIÈGE université
GIGA
Genomics

VARNAVEL

Development of methods to map within -and between-
individual variation in single cell RNA velocity-based
fate determination

Aguilar Ortmans Diego

Master of Biomedical Sciences, professional focus in biomedical data management

Faculty of Medicine

UNIVERSITY OF LIEGE

Academic Year 2020/2021

Supervisors: Prof. Georges Michel, Ph. D. Demeulenaere Loïc

Examining Board: Prof. Nguyen Laurent, Prof. Haesbroeck Gentiane, Ph. D. Lavergne Arnaud

Acknowledgements

The first person I wish to thank is Loïc Demeulenaere, the postdoctoral researcher in charge of the VARNAVEL project. First for answering my countless questions about RNA velocity and Markov chains with both precision and patience. Then for helping me on several occasions to find the errors in my programs and to correct them. Finally, for reviewing parts of this work to point me in the right direction. Working with him was a pleasure that had a very positive impact on the progress of my dissertation and I hope to have the opportunity to collaborate with him again in the future.

I also wish to express my gratitude to my supervisor, Professor Michel Georges, for giving me the opportunity to work on a subject I am passionate about. Despite his busy schedule, he took time to give me advice and encourage me towards undertaking a PhD.

I also thank Professor Gentiane Haesbroeck, Professor Laurent Nguyen and Ph. Doctor Arnaud Lavergne for taking the time to evaluate my work and be part of my jury.

Finally, I would like to thank my relatives for their presence and support, especially my father, Javier Aguilar Medina for proofreading the entire work and my sister Céline Aguilar Ortman for helping me with the layout.

Diego Aguilar Ortman

Table of Contents

Introduction	1
Monogenic diseases and complex diseases	1
Induced pluripotent stem cells	2
Organoids.....	3
Vertebrate retina	4
Laminar organization of the retina	5
Cell types in the retina	5
Retinal progenitor cells.....	5
Photoreceptor precursors	5
Photoreceptors	6
Bipolar cells.....	6
Horizontal cells	6
Amacrine cells.....	7
Retinal ganglion cells	7
Müller glia	7
Neuroepithelium	7
Retinal pigment epithelium	7
Cellular death	8
Differences between the human retina and the murine retina.....	8
Retinal organoids.....	9
Differences between retinas and retinal organoids.....	10
Enhanced S-cone syndrome	11
Single cell RNA sequencing.....	11
RNA velocity.....	13
Markov chains.....	14
Hypothesis, Objectives and Experimental Strategy	16
Hypothesis	16
Objectives	17
Experimental Strategy of this work	17
Experimental strategy of the VARNAVEL project	18
Materials and methods	19
Datasets	19
First Dataset	19
Second Dataset.....	19
Generation of the first database	19
Single-cell RNA sequencing.....	20

Bridge PCR	20
Data processing	21
Demultiplexing	22
Read alignment	22
Identification of spliced and unspliced mRNAs	22
Barcode and UMI filtering	22
Duplicate marking	23
Cell filtering	23
First gene filtering	23
Determination of cell types	24
Dimensionality reduction and visualization	25
Metacells	26
Second gene filtering	26
Loom files	28
Computation of RNA velocities	28
Computation of probability distributions using Markov chains	29
Prediction models	30
First model	31
Second model	31
Third model	31
Long term fate	32
Compute the transition matrix to a high power	32
Exponentiation by squaring	32
Simulations	32
Reproducibility of simulations	33
Simulation graphs	33
Results	34
First Model	35
Proportion graphs analysis	35
Mature cell types	35
Progenitor cell types	36
Conclusion	36
Second Model	37
Proportion graphs analysis	37
Mature cell types	37
Progenitor cell types	38
Conclusion	38
Third Model	39
2D UMAP analysis	39
Discussion	40

Similarity between first and third model	40
Model comparison.....	41
UMAP graph interpretation.....	42
Identification of spliced and unspliced mRNAs	42
Cellular Death	43
Lack of cells.....	43
Use of multiple databases	44
Calculation of RNA velocity.....	44
Low transcript capture per cell.....	45
Selected genes	45
Conclusion.....	46
Outlook.....	46
Bibliography	47

List of Acronyms

cDNA	Complementary DNA
DR	Dimensionality reduction
ERPC	Early retinal precursor cell
ESC	Embryonic stem cell
ESCS	Enhanced S-Cone Syndrome
GCL	Ganglion cell layer
INL	Inner nuclear layer
IPL	Inner plexiform layer
iPSC	Induced pluripotent stem cell
LRPC	Late retinal precursor cell
mRNA	Messenger RNA
NE	Neuroepithelium
ONL	Outer nuclear layer
OPL	Outer plexiform layer
PCA	Principal Component Analysis
RGC	Retinal ganglion cell
RPC	Retinal precursor cell
RPE	Retinal pigment epithelium
scRNA-seq	Single cell RNA-sequencing
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
UMI	Unique molecular identifier
VARNVEL	Variation of RNA velocity

Abstract

Background: RNA velocity is a new theoretical model whose objective is to predict the short-term future of a cell in terms of its transcriptome from single cell RNA sequencing data. In addition, the production of single cell data has drastically increased for several years.

Objectives: From single cell RNA sequencing databases produced to study development, the objective is to create a computer model recapitulating the differentiation trajectories by using the concepts of RNA velocity and Markov chains.

Methods: The database used comes from a study about retinal development (Georges et al., 2020) which contains murine retinal cells collected at 4 stages of development. By associating the transcriptomic profile of each of these cells to a state, the long-term evolution of these cells can be determined using Markov chains. Transition probabilities are defined from RNA velocities, providing a biological basis for predictions. These velocities are calculated with the steady state model (La Manno et al., 2018). Three models have been developed to calculate the transition probabilities. These take into account the angle between the RNA velocity vector and the vector connecting the two states involved in the transition. Moreover, the distance between these two states is also considered.

Results: Of the three models created, none was able to completely recapitulate the process of retinal development. This is partly due to the inability of photoreceptor precursors to differentiate. However, the results obtained do not depend only on the model used. Other factors can be responsible for the problems encountered, such as a lack of cells in the database, biases in the calculation of RNA velocity, the fact that cell death is not accounted for in our models, an incorrect gene filtering, the poor capture of the transcriptome with the 10X method and difficulties to determine whether an RNA molecule is spliced or not.

Conclusion: In order to obtain more biologically consistent results, the models must be optimized and the external factors mentioned above must be taken into account. Once this is done, the early genes responsible for the distinct differentiation pathways could then be identified by analyzing the regions where the main trajectories split into several different trajectories by using principal curves.

*A French version of the abstract is available on MatheO.

CHAPTER 1

Introduction

This introduction aims to cover the theoretical notions necessary for a good understanding of the work done in this project and the global objectives of the VARNAVEL project. In brief, the objective of the VARNAVEL project is to develop computer tools that identify genes involved in development in healthy and diseased individuals. In order to generate and test these tools, databases are required. These databases can be obtained from organoid cells or from cells taken directly from an animal model.

The introduction first discusses the difference between monogenic and complex diseases as it is known that they often have an impact on development and are therefore interesting to model to study a diseased development. The ability of induced pluripotent stem cells and organoids to be reliable models is then described. The monogenic disease to be studied is called “enhanced S-cone syndrome” and affects retinal development. Therefore, the second part of the introduction will describe the organization of the retina in vertebrates, the difference between human and mouse retinas, the ability of retinal organoids to accurately represent retinal development, the difference between a retina and a retinal organoid and a brief description of the disease in question. Finally, the techniques needed to study such development are described in the last part of the introduction, namely single cell RNA-sequencing, RNA velocity and Markov chains.

For the purpose of this work, not all of the elements described are necessary, especially the section describing the difference between mouse and human retina and the section describing the disease. However, I thought it was relevant to include them for a better understanding of the project as a whole and because they will be useful for the next steps, especially when comparing the results obtained in mice with those obtained in healthy and sick humans.

Monogenic diseases and complex diseases

A monogenic disease is a disease caused by mutations on a single gene. Its inheritance follows a Mendelian pattern and therefore can be recessive, dominant or X-linked. Even if they are rare diseases¹, more than 5,000 monogenic diseases have been described² such as cystic fibrosis, Huntington’s disease, hemophilia A³ and enhanced S-cone syndrome⁴. However, not all diseases with a genetic origin are monogenic. Indeed, a category of diseases called “complex diseases” are caused on the one hand by a set of mutations on different genes and on the other hand by environmental factors⁵ and lifestyle habits. Complex diseases include, for example, hypertension, diabetes, Alzheimer’s disease and Parkinson’s disease. The multiplicity of variables involved in these diseases makes their transmission unpredictable given that they do not follow Mendelian patterns¹.

The distinction between monogenic and complex diseases is, in fact, blurrier than what is explained in the previous paragraph. First because diseases qualified as complex diseases can be caused, in a small proportion of the affected population (between 1 and 7 percent), by a single mutant gene and are thus transmitted in the same way as monogenic diseases⁶. Another nuance to bring concerns the penetrance of monogenic diseases. Indeed, alleles responsible for monogenic diseases were characterized as sufficient and necessary to develop the disease⁷.

This implies that these alleles must be completely penetrant. However, studies showed that a small proportion of the individuals carrying mutations responsible for monogenic diseases do not develop the associated symptoms⁸. The reason of this incomplete penetrance is not clearly defined. Copy number variation of the mutated alleles and the influence of the rest of the genome are two potential factors responsible for this observation⁸.

Induced pluripotent stem cells

The fact that the DNA present in the nucleus of somatic cells contain the same genetic information as embryonic stem cells was proved in 1962 by John Gurdon⁹. Moreover, the concept of reprogramming environment necessary for somatic cells to become stem cells was discovered by Ian Wilmut et al. when they performed somatic cloning¹⁰. On the basis of these discoveries, Takahashi and Yamanaka achieved in 2006 the reprogramming of mouse fibroblasts into stem cells by using four different transcription factors: Oct3/4, Sox2, c-Myc, and Klf4¹¹. In 2007, the same achievement was successfully performed by using the same four transcription factors on human fibroblasts¹². Another combination of genes (OCT4, SOX2, NANOG, and LIN28) was found that year by another team¹³.

The cells produced by this process are called induced pluripotent stem cells (iPSCs) and as the name suggests, are pluripotent cells obtained from differentiated somatic cells. This technology is simple and reproducible¹⁴ although not very efficient. Indeed, the proportion of transfected fibroblasts that ultimately become induced pluripotent stem cells is inferior to one percent¹⁴. The expression of the four transcription factors used to reprogram somatic cells was initially achieved by retrovirus or lentivirus -mediated transfection¹¹. However, this method caused insertional mutagenesis and therefore has been replaced by the use of plasmids and Sendai viruses¹⁴.

Nowadays, the main application of induced pluripotent stem cells is disease modeling. Another promising application consists in deriving iPSCs from a patient and differentiate them into specific cells in order to re-inject them in the patient and change the course of the illness. This is called “cellular therapy“. In the context of this work, the differences between iPSCs and embryonic stem cells (ESCs) will only be described for disease modeling¹⁵.

When it comes to producing human disease models, iPSCs are preferred to ESCs. Indeed, in order to model genetic diseases, the responsible mutation has to be induced in ESCs by genome editing. Before the discovery of the CRISPR gene editing technology, techniques enabling the precise insertion of a specific mutation in a gene had very low yield. It was therefore more efficient to derive induced pluripotent stem cells from somatic cells of patients carrying the mutation of interest¹⁶. In addition, by using patient-derived iPSCs, we ensure that the genotype will cause the expected phenotype, thus avoiding the risk of having a protective genetic background¹⁷. Finally, iPSCs are not concerned by ethical issues in the same proportion as human ESCs¹⁸. Besides, the capacity of induced pluripotent stem cells to model diseases is not limited to monogenic diseases. Indeed, polygenic diseases can also be mimicked as demonstrated with Alzheimer’s disease¹⁹⁻²¹, schizophrenia²² and Parkinson’s disease²³.

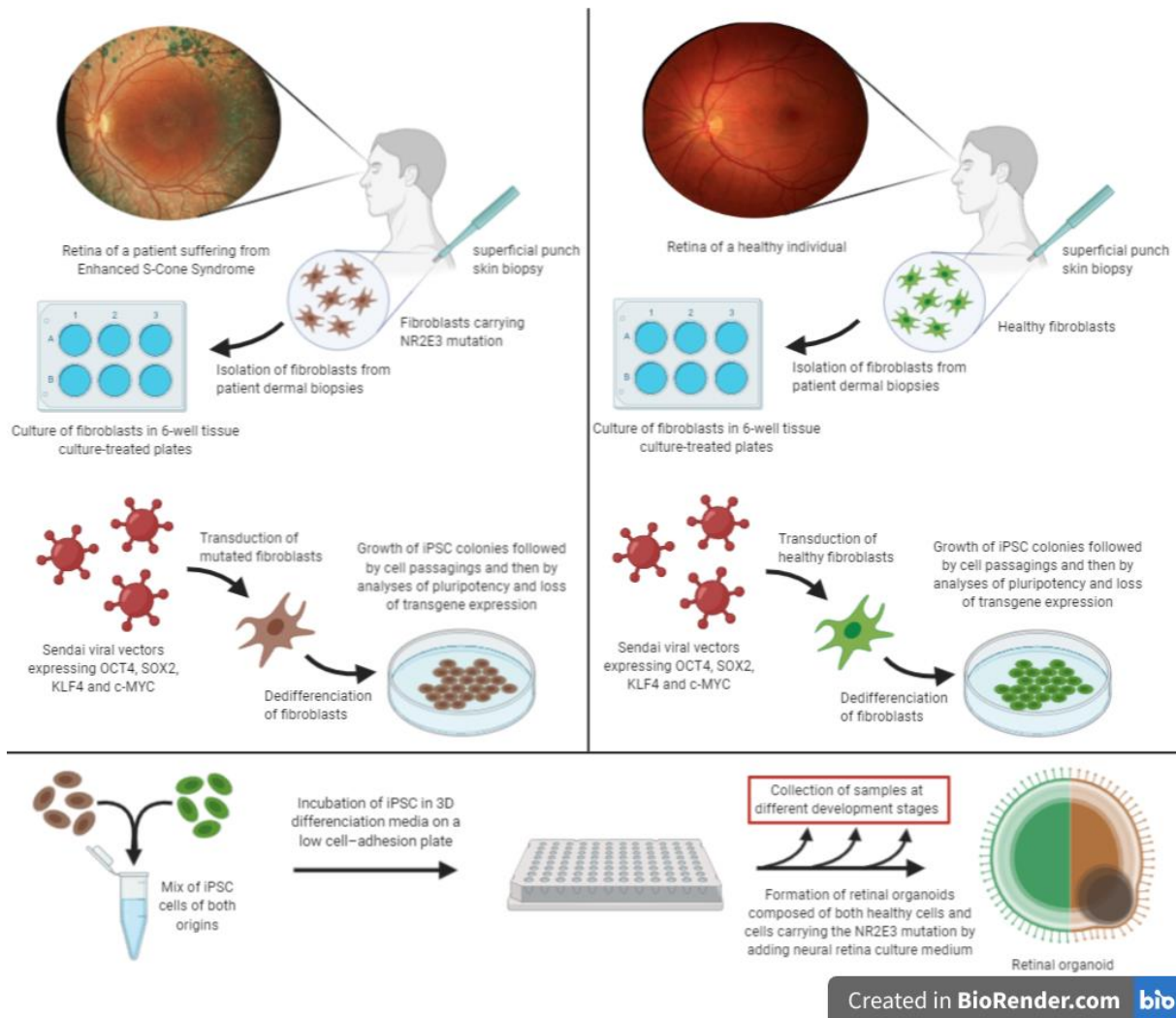


Figure 1.1: Summarized steps of the generation, from human skin biopsies, of retinal organoids composed of both healthy cells and cells carrying the NR2E3 mutation responsible for enhanced S-cone syndrome. These organoids may be used to model the differentiation trajectories of both healthy and diseased development of the retina. Figure created in BioRender.com

Nevertheless, in specific cases, the usage of iPSCs may not be adapted to model diseases. For instance, when the disease to model has a high early mortality rate. Indeed, by taking only patient-derived iPSCs, the model will be based only on genomic profiles of patients who survived. This introduces a bias that diminishes the quality of the model to represent the disease in its globality, including all the profiles that did not survive¹⁵. Another case is when the disease to model is caused by mutations situated in the genes responsible for the reprogramming pathway used to produce iPSCs. In this situation, iPSCs cannot be obtained as the reprogramming does not complete¹⁵. Similarly, some genes are resistant to the process of reprogramming and therefore are not expressed in the model. Thus, diseases affecting these genes cannot be modeled with iPSCs¹⁵.

Organoids

An organoid is a tridimensional multicellular structure used to model a given organ. This is made possible thanks to its similarity with the native form of that organ. Organoids are produced in vitro from specific cells, such as cells sampled from a tissue, induced pluripotent stem cells (iPSC) or embryonic stem cells, by following a specific cell culture protocol. They are very promising tools 1) to study the development of healthy tissues and organs, 2) to evaluate the impact of diseases on that developmental process, 3) to perform drug screening on cultured cells and 4) to generate cells suited for cell therapy. Even if they are used in fundamental research, there is still too much variability when generating organoids to use them in clinic²⁴.

Generating organoids consists in replicating in vitro the in vivo developmental conditions in order to incite stem cells to differentiate and organize themselves in the same way as native stem cells do. Indeed, physical and chemical signals are used to induce development in vitro which implies the renewing of stem cells and their differentiation in different cell types. These signals act on different pathways to stimulate the proliferation of cells, their differentiation, their migration, their selection and their organization in complex structures²⁴.

Compared with 2D cell cultures, organoids better mimic the architecture, the metabolic function and the protein expression of native organs. They are even able to reproduce the specific functions of some organs such as the production of mucus and the absorption and secretion of molecules that are performed by the intestinal organoids²⁵. Therefore, organoids are more physiologically relevant than 2D cell cultures²⁴.

Using pluripotent stem cells instead of cells originating from tissue samples in order to produce organoids presents the advantage to model organs whose tissue sampling is difficult. As described hereafter, iPSC-derived organoids can be used to model rare human embryonic tissues to study organogenesis. Moreover, if the induced pluripotent stem cells are produced from a sample of a patient suffering from a genetic disease, the generated iPSCs will carry the mutation and the effect of this disease on development can be studied (Figure 1.1).

However, more advances are required before organoids completely recapitulate native organs. Indeed, the lack of a mesenchymal compartment, vascularization and microbiome, could explain why most organoids do not produce every specialized cell type present in the native organ. Moreover, their effectiveness is also lower in vitro than in vivo. In addition to the fact that all the signals involved during development are not known yet, the absence of these cells may be the reason why none of the established organoids is as functional as its native version²⁴.

Another important difference is that organoids cannot be kept in culture indefinitely. This is particularly true with human iPSC-derived organoids that do not last enough to model organs beyond the fetal stage^{26,27}. In addition, given that nutrient supply and waste removal depend on diffusion, the effect of these two processes decrease as organoids get bigger²⁴.

Given that some processes such as cell fate and cellular self-organization are still difficult to control in vitro, organoid generation protocols lack of reproducibility. In order to reduce this variability and take full advantage of the properties of organoids, some improvements need to be implemented. Ideally, organoid generation would be fully automated but protocol's complexity impedes it for now. Moreover, matrices and media used in organoid culture should be defined and standardized. Indeed, the extracellular matrix plays an important role in stem cell self-renewal and differentiation by creating stimulating signals and providing a physical structure^{28,29}. The most widely used matrix is called Matrigel and is derived from mice. Its animal origin and its complexity (more than 2 000 different proteins^{28,29}) result in difficulties to define and standardize it²⁴.

In addition, some of the variability is produced by the insufficient precision of the techniques used to monitor the organoids. To address that issue, miniature biosensors could be placed in strategic places to produce more accurate information. For example, how functional the organoid is cannot be precisely assessed with traditional optical monitoring techniques. Finally, other factors bring variability such as the starting cell population, their positioning and aggregation. To address that issue, microwell structures and microfluidic devices are used to control cell aggregation²⁴.

Vertebrate retina

The global structure of the retina is common to all vertebrates and is constituted of six cell types. Among these cell types, five are neuronal cells (retinal ganglion cells, amacrine cells, photoreceptors, bipolar cells and horizontal cells) connected to each other. These connections form circuits regulating photoreceptor generated signals and transmitting them to the optic nerve and then to the brain^{30,31}. The last cell type composing the vertebrate retina is the Müller glia, whose function is to protect retinal neuronal cells, ensure their metabolic necessities and maintain their homeostasis³⁰.

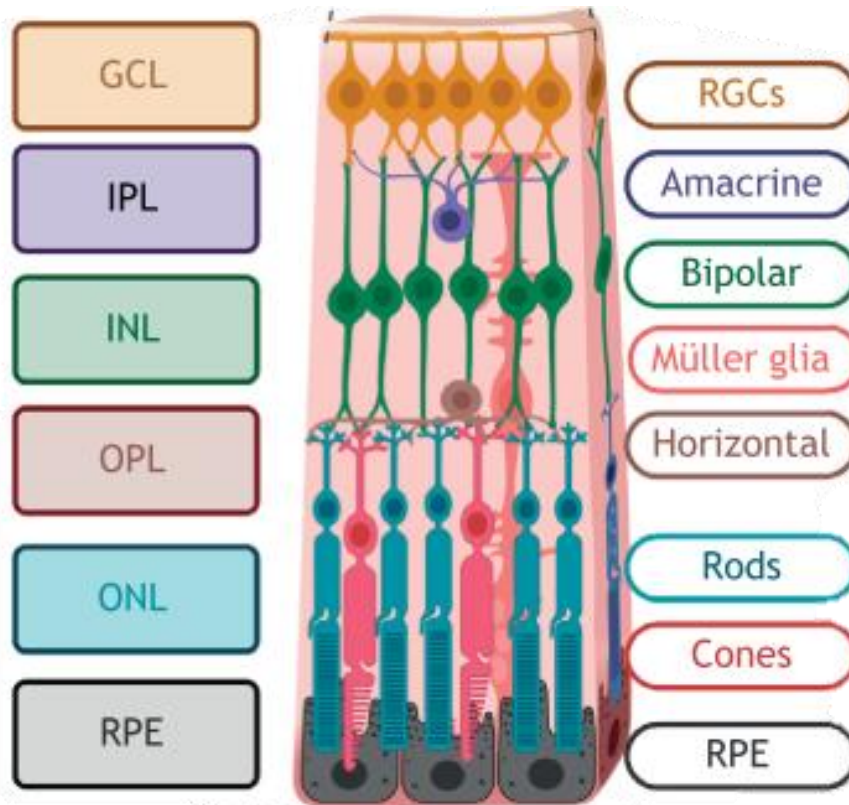


Figure 1.2: *In vivo*, the seven main neuroretinal cell types populate the layers of the retina with retinal pigmented epithelium (RPE) next to the outer nuclear layer (ONL). Interneurons synapse with photoreceptors in the outer plexiform layer (OPL) and retinal ganglion cells (RGCs) in the inner plexiform layer (IPL) to relay signals to the brain. Figure and explanations from “Retinal organoids: a window into human retinal development”, figure 4a by Michelle O’Hara-Wright and Anai Gonzalez-Cordero, 2020³². GCL = Ganglion cell layer, INL = Inner nuclear layer

Laminar organization of the retina

The mature vertebrate retina is organized in six different layers. The retinal pigment epithelium (RPE) is the most external layer of the retina that overlays the outer nuclear layer (ONL). This one contains the nuclear bodies of both rods and cones and is situated above the outer plexiform layer (OPL) which represents the plane zone where photoreceptors make synaptic connections with bipolar and horizontal cells. The inner nuclear layer (INL) is situated under the OPL and is made up of horizontal cells, bipolar cells, amacrine cells and Müller glia. Bipolar and amacrine cells make synaptic connections with retinal ganglion cells in the inner plexiform layer (IPL) localized between the INL and the ganglion cell layer (GCL) that contains the cell bodies of retinal ganglion cells^{32,33} (Figure 1.2).

Cell types in the retina

Retinal progenitor cells

Retinal progenitor cells (RPCs) are multipotent precursor cells that produce all the cells composing the mature retina. Cellular types generated during the development of the retina depend on the stage of the retinogenesis. In fact, there are two kinds of retinal progenitor cells that participate in this process: early retinal progenitor cells (ERPCs) and late retinal progenitor cells (LRPCs). Early RPCs are produced at the beginning of retinogenesis while late RPCs appear during the late stage of development³⁴. Each type is responsible for the generation of a collection of cell types. Retinal ganglion cells, horizontal cells, cones and certain subtypes of amacrine cells are produced by differentiation of early RPCs. Late RPCs spawn the other subtypes of amacrine cells, bipolar cells, rods and Müller glia³⁵.

During development, RPCs replicate first by symmetrical mitosis to increase the size of the cell pool and then asymmetrically to generate differentiated cells while maintaining a constant pool. The new cell generated by asymmetric mitosis is qualified as neurogenic because it expresses the genes necessary for its differentiation into retinal neuronal cells. Nevertheless, during the last asymmetric division, the mitosis does not produce a neurogenic cell but a Müller cell. Both early and late RPCs produce neurogenic RPCs but these are different because they do not give rise to the same cell types. Indeed, early RPCs generate early neurogenic RPCs that differentiate into ganglion cells, horizontal cells, cones and amacrine cells while late RPCs produce late neurogenic RPCs that differentiate into amacrine cells, bipolar cells, rods and Müller glia as described above³⁶.

Photoreceptor precursors

Photoreceptor precursors originate from both early and late neurogenic RPCs and give rise to cones, rods and bipolar cells. The transcription factor Otx2 targets the promoters and the enhancers of genes specific of photoreceptors and bipolar cells³⁷⁻³⁹ and therefore is a marker of photoreceptor precursors. The distinction of fate between photoreceptor cells and bipolar cells is mediated by two transcription factors (Vsx2 and Prdm1) whose expression depends on the expression of Otx2⁴⁰. The expression of Vsx2 induces the differentiation of photoreceptor precursors into bipolar cells by repressing photoreceptor-specific genes⁴¹⁻⁴³. On the other hand, Prdm1 is necessary for photoreceptor precursors to become photoreceptors^{38,44,45}.

Finally, factors such as Ror β and Nrl induce the differentiation into rods instead of cones⁴⁶. Indeed, the combination of Otx2 and Ror β triggers the expression of Nrl that activates rod-specific genes such as the transcription factor Nr2e3 responsible for the induction of rod genes and the repression of cone genes^{47–52}. The specific factors promoting the differentiation into cones are still rather unknown. However, the thyroid hormone receptor Thr β 2 is a marker specific to cones during the development⁴⁰.

Photoreceptors

Retinal photoreceptors are photosensitive cells represented by two distinct cell types: cones and rods. Together, they account for more than 70 percent of cells in the mature retina^{35,53}. Their task is to convert light stimuli into electrochemical signals and ensure the transmission of this information to the rest of the retina^{31,35,54}. Cones and rods differ in the visual pigment they contain. Indeed, rods incorporate rhodopsin, a biological pigment extremely sensitive to light, hence enabling night vision. The biological pigments associated to cones are called opsins and define different kinds of cones, depending on their wavelength of maximal absorption. Human retina contains three different subtypes of cones identified by their initials: L-cones, M-cones and S-cones. These initials correspond respectively to “Long”, “Medium” and “Short”, which indicate the optimal range of wavelengths for each cone. L-cones are the most abundant subtype in the human retina and are sensitive to red light. The second most abundant subtype are M-cones which are specific to green light. S-cones are a minority representing around eight percents of the cones in the human retina and are sensitive to blue light⁵⁵. The specificity of each cone to a precise wavelength associated to a defined color enables color vision³¹. Nevertheless, cones are less sensitive to light than rods and are therefore more suited for bright light conditions^{30,56}.

Retinas are generally duplex which means that they contain simultaneously cones and rods⁵⁴. However, the proportion of rods is considerably superior to the cone proportion in the retina. This difference is estimated by calculating the ratio of rods to cones and may vary from one species to another. In fact, this ratio will depend on the circadian cycle of the animal. Given that cones are more adapted to daylight, their number is higher in diurnal than nocturnal species. For instance, the ratio of rods to cones in the human retina is around 20:1 while it is equal to 30:1 in the murine retina³⁵.

Bipolar cells

The shape of bipolar cells is very specific: two branches emerging from a central body but with opposite directions. The first branch synapses with photoreceptors while the opposite branch synapses with a ganglion cell³¹. This shape is associated with its function that is to ensure the communication of the signal from photoreceptors to ganglion cells^{35,54}. Moreover, the activity of bipolar cells can be regulated by amacrine cells that contact several bipolar cells³⁵.

Horizontal cells

Horizontal cells are inhibitory interneurons laterally connected to multiple photoreceptors that modulate the transmission of the signals emerging from rods and cones^{31,54}. Their assumed function is to increase contrast through lateral inhibition and adapt signal processing depending on the luminosity⁵⁷.

Amacrine cells

Amacrine cells are predominantly inhibitory interneurons that regulate the excitatory signals transmitted between bipolar cells and ganglion cells^{35,54}. They are connected laterally to a certain number of bipolar cells in the same way as horizontal cells are connected to photoreceptors⁵⁸. Given that ganglion cells receive direct inputs from bipolar and amacrine cells, the regulatory actions of amacrine cells on signals are both direct and indirect. Indeed, the direct effect is applied through the synapses between ganglion cells and amacrine cells and is called feedforward inhibition while the interactions between bipolar and amacrine cells carry out the indirect effect called feedback inhibition^{30,58}.

Retinal ganglion cells

The last cells that conduct the signal and whose function is to send it to the visual centers of the brain through the optic nerve are called retinal ganglion cells (RGCs). Each retinal ganglion cell receives different inputs from bipolar and amacrine cells which reflect the activity of photoreceptors. These inputs are processed in the ganglion cells and the final output is sent through their axons to the brain under the form of depolarizing spike trains^{35,54}.

In fact, there are over 40 different types of RGCs⁵⁹. Four of them express the transcription factor Tbr1⁵⁹ and thus form a subpopulation distinct from the rest of the RGCs. In this work, this subpopulation will be called RGCs 2 while the main population will be called RGCs 1.

Müller glia

Müller glia contribute to the maintenance of the retina. Indeed, these cells ensure different support tasks such as establishing the laminar structure of the retina, providing nutrients, releasing neurotrophic factors and interacting with the extracellular environment by capturing neurotransmitters⁶⁰. These functions are made possible by the processes originating from their cell bodies and reaching the different layers of the retina³⁵.

Neuroepithelium

During the formation of the eye, the optic cup is formed of two layers of different size. The inner layer is thicker than the outer layer and will give rise to the neural retina. The outer layer of the optic cup will become the retinal pigment epithelium (RPE) by the production of melanin granules in the cells composing it⁶¹.

The neuroepithelium (NE) is a pseudostratified epithelium composed of neuroepithelial cells connected by junctional complexes that compose these two layers⁶². These cells are neural stem cells that compose the inner layer of the optic cup and that will give rise, among other things, to all the neuronal cells of the retina and to the retinal pigment epithelium⁶³.

Retinal pigment epithelium

The retinal pigment epithelium is situated between the outer nuclear layer and the choriocapillaris. It ensures the role of outer blood-retinal barrier and therefore controls what substances coming from the blood can reach the retina. Retinal pigment epithelium is thus

responsible for the transport of nutrients, ions and water. Another function of the RPE is to reduce the oxidative stress existing in the retina⁶⁴. A part of this stress is caused by the photooxidation of lipids which is decreased by the presence of different pigments in the RPE that absorb specific wavelengths of the light⁶⁵⁻⁶⁷. Moreover, the RPE contains antioxidants to reduce the production of reactive oxygen species due to the high consumption of oxygen⁶⁴. The RPE plays another critical role because it produces the enzyme capable of isomerizing all-trans-retinal into 11-cis-retinal. Indeed, 11-cis-retinal is a chromophore present in photoreceptors that allow for the vision⁶⁸. Finally, RPE ensure other functions such as the renewal of the outer segment of photoreceptors⁶⁹⁻⁷¹, the control of the ion composition in the subretinal space responsible for the conservation of the excitability of photoreceptors⁷⁰ and the secretion of growth factors and factors that help maintain the structure of the retina and the survival of its cells⁶⁴.

Cellular Death

The elimination of a fraction of the cells by apoptosis is a physiological process that takes place during the development of the retina⁷². Its objective is to adapt the size of the retina population. Indeed, nearly 70% of the cells generated do not reach the mature stage of the retina and are eliminated by apoptosis⁷³.

During mouse retinal development, two waves of programmed cell death occur. The first one involves neuronal progenitor population and starts at embryonic day 12 and reaches its maximum between day 14 and 16. The second wave takes place essentially during the first two weeks after birth. 90% of the retinal ganglion cells are eliminated during the first 11 days, a part of the amacrine cells is also eliminated during these 11 days. A fraction of rods is eliminated from the fifth day until the 24th day. Finally, a part of bipolar cells and Müller cells is also eliminated from the fifth day but only until the 18th day⁷³.

Differences between the human retina and the murine retina

Human and mouse have a different circadian rhythm. Indeed, humans are diurnal animals while mice are nocturnal animals. For that reason, mice are dichromats which means that their retina contains two different types of cones. By contrast, humans have three subtypes of cones and are thus trichromats. Mice possess around five percent of S-cones and ninety-five percent of M-cones⁷⁴⁻⁷⁷ while human retina includes L-cones, M-cones and S-cones. Moreover, the proportion of rods is superior in mice while the proportion of cones is superior in humans^{78,79}. Besides, the size of photoreceptors also differs when nocturnal and diurnal species are compared⁸⁰. Moreover, there are two types of horizontal cells in the human retina^{81,82} and only one type in the murine retina⁸³.

Another difference between humans and mice is the presence of a macula in the human retina. Indeed, the macula is a small specialized region of the retina composed of three concentric zones: the outermost layer is the perifovea, the center is the fovea and the parafovea is the region in between⁸⁴. The vascularization rate and the ratio of rods to cones

are two parameters that vary from one zone to the other. Out of the macula, the ratio of rods to cones is high and the retina is highly vascularized. The perifovea shares these characteristics but also has a higher proportion of cones and ganglion cells^{55,85}. Vascularization and rods density decrease in the parafovea, while the amount of ganglion cells and cones increases further. Thus, the rods to cone ratio decreases to around four rods for one cone⁸⁴. This decrease of vascularization and cone density escalates in the fovea until the proportion of cones is superior to the proportion of rods. Since cones perform under bright light conditions, the macula is specialized in the high visual acuity of the photopic vision⁸⁴.

Retinal organoids

Historically, studies on the retinal development used animal models to observe and identify the processes responsible for the differentiation of precursors cells and their spatial organization. Although these models provided precious information of the development of retina in different species, it was still necessary to validate these concepts in the human retinogenesis. Indeed, the little anatomical information available was obtained from rare human fetal tissue. Advances in induced pluripotent stem cell and organoid technologies made it possible to study in vitro the development of the human retina without using human embryonic tissues.

Techniques to form retinal organoids evolved through time, starting with simple 2D adherent cultures. This method consists in inducing the differentiation of retinal stem cells by exposing these cells to Wnt or BMP inhibitors such as DKK1 and Noggin and to IGF1. Photoreceptors are generated but they are not organized in layers like in the native retina³². In 2011, 3D optic cups were generated from murine embryonic stem cells aggregated in embryoid bodies. To that end, a serum free suspension culture method coupled with Matrigel matrix was used⁶³. The same results were then obtained with human embryonic stem cells³². After that, the production of retinal organoids in classic 3D suspension culture was performed. Finally, 2D/3D approaches result from the combination of 2D and 3D techniques. In fact, pluripotent stem cells first grow on adherent cultures and form retinal vesicles once confluence is reached. These retinal vesicles are then mechanically excised and placed into a 3D suspension culture. These conditions enable the alignment of photoreceptors expressing rhodopsin in a layer similar to the ONL³².

Retinal organoids are new models of human retinogenesis that complement the pre-existing animal models. The advantage of human retinal organoids is that they have a morphology and a cellular composition similar to the human native retina. This is essential in order to clarify the knowledge acquired by means of the previous models and to discover new signaling mechanisms and cell interactions leading to retinal development. Retinal organoids have the potential to recapitulate the developmental trajectories of each cell type present in the retina. To achieve that goal, diversified differentiation protocols are established to help and/or accelerate the development of specific cell types. Nevertheless, the use of external factors in these protocols can modify the original progression of the developing retina. This could decrease the quality of the model due to unsynchronized developmental temporal timelines³². Lastly, retinal organoids can model the effects of genetic retinal diseases by using induced pluripotent stem cells obtained from patients carrying the mutation responsible for the disease.

Differences between retinas and retinal organoids

Firstly, retinal ganglion cells are produced at a lower level in retinal organoids. Indeed, two RGC specific markers (POU4F1 and NEFL) are less detected in organoids than in fetal conditions⁸⁶. This could be explained by the absence of other ocular structures and a shortage in neurotrophic factors. In addition, RGCs tend to disappear in long term cultures⁸⁷ which is confirmed by the difference of expression of RGC related genes measured in scRNA-seq studies^{86,88,89}. Moreover, retinal organoid differentiation protocols are usually designed to produce a maximum of photoreceptors. These conditions may not be optimal for the survival of RGCs³².

Secondly, retinal organoids generate an insufficient amount of bipolar, horizontal and amacrine cells. This production deficit could be attributed to the loss of synaptic connections with RGCs due to their progressive disappearance. These synaptic connections are necessary to keep bipolar, horizontal and amacrine cells alive³². Furthermore, retinal organoids lose the inner layer lamination containing these cells in long term cultures⁸⁶.

Thirdly, even if retinal organoids mimic natural photoreceptor developmental dynamics (cones containing S-opsin are produced before the formation of cones expressing L or M opsins)⁷⁸, photoreceptors maturation is not fulfilled with in vitro conditions⁹⁰. This maturation aims to form the outer segment of photoreceptors which is essential in the light detection process³². To address that issue, improved maturation conditions were developed and enable, inter alia, the formation of maturing outer segments^{91,92}. Fourthly, the retinal pigment epithelium is not juxtaposed to the outer nuclear layer in retinal organoids³². Indeed, in organoids, RPE differentiate from neuroepithelial cells at a molecular level but not at a structural level⁹³.

Fifthly, the macula is not produced in retinal organoids. Indeed, no fovea-like structure is detected by immunohistochemistry techniques nor is the expected evolution of rods to cones ratio present in the macula³². Even if promising molecules, such as triiodothyronine⁷⁸, retinoic acid⁸⁹ and fibroblast growth factor 8⁹⁴ were identified, more research is needed to obtain a macular region in retinal organoids.

Inversely, Müller glia are adapted to the culture conditions of organoids and therefore keep the natural morphology that they have in the native retina and are sufficiently produced³².

The study that produced the dataset used in this work⁹⁵ report that, in retinal organoids, the control of the transcriptome of cells was less tight in terms of space and time than in native retina⁹⁵. That means that the variations between the transcriptomic profile of cells of a same cell type are more important in retinal organoids than in native retina. Moreover, the cell types resulting from differentiation tend to appear earlier in the development in retinal organoids⁹⁵. Indeed, photoreceptor precursors differentiate more quickly but are not able to completely finish their differentiation process into mature photoreceptors or bipolar cells and therefore accumulate. The fact that photoreceptor precursors appear earlier negatively impacts the other cell type populations, specifically retinal progenitor cells⁹⁵.

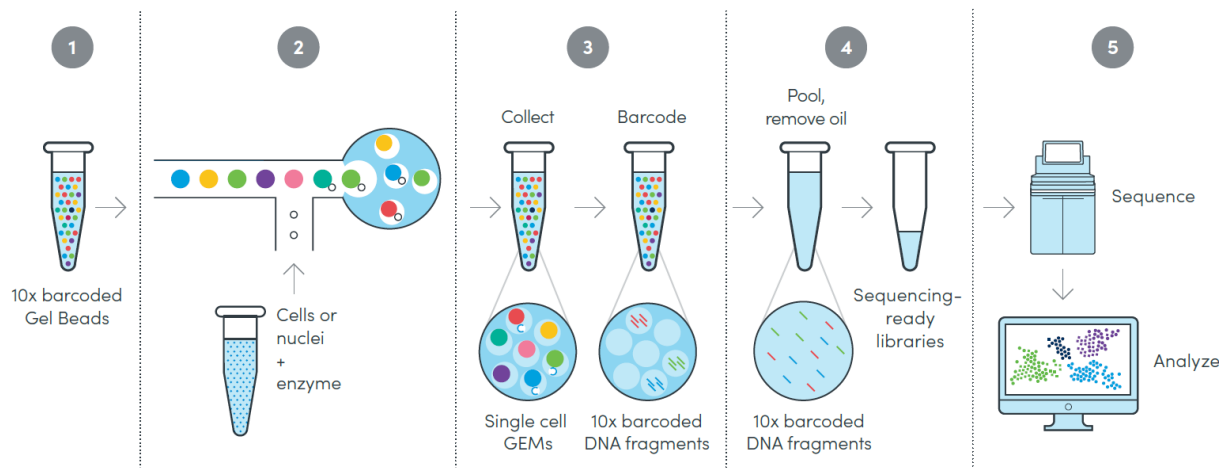


Figure 1.3: 10X Genomics Chromium workflow for single cell gene expression analysis. Figure from “The Power of Single Cell Partitioning”, page 2 by 10x Genomics, 2020⁹⁶.

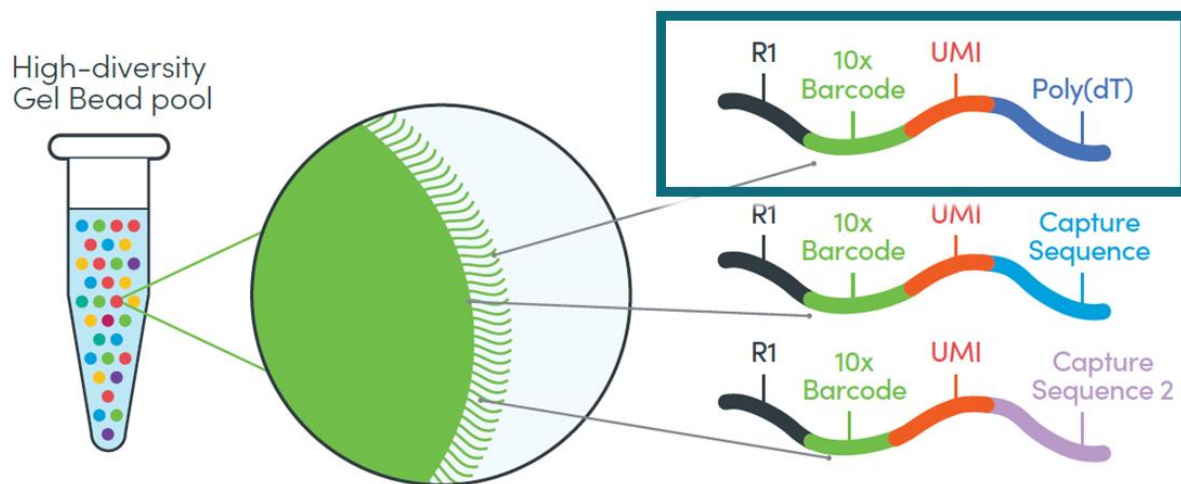


Figure 1.4: Representation of a gel bead (middle) with the structure of the primers (right) among a pool of gel beads (left). The type of primers used in this project is framed (Poly(dT)). Figure from “The Power of Single Cell Partitioning”, page 3 by 10x Genomics, 2020⁹⁶.

Finally, it is important to mention that protocols as much as cell lines used to produce retinal organoids yield organoids with different proportion of cells and timings^{89,97-100}.

To conclude this section, the following aspects need to be improved to get better organoids: 1) the precise apposition of the RPE with photoreceptors is necessary, 2) the survival of bipolar, horizontal, amacrine and retinal ganglion cells and 3) the formation of a macula-like region³².

Enhanced S-cone syndrome

Enhanced S-cone syndrome (ESCS) is a rare monogenic autosomal recessive retinal disease resulting from mutations in the NR2E3 gene. This gene encodes the NR2E3 (nuclear receptor subfamily 2, group E, member 3) protein which is exclusively expressed in the retina¹⁰¹. NR2E3 is a ligand-dependent transcription factor that interacts with other transcription factors such as neural retina leucine zipper (NRL) and cone-rod homeobox (CRX)^{51,102,103}. Its function is to induce the differentiation of photoreceptor precursors into rod photoreceptors⁴⁸ while repressing the genes responsible of the formation of short-wavelength sensitive opsin¹⁰⁴. Consequently, ESCS is defined by an abnormally high proportion of S-opsin-positive cones in the retina at the expense of M- and L-cone photoreceptors and rod photoreceptors¹⁰⁵. Therefore, patients suffering from ESCS are more sensitive to light with small wavelengths such as blue light. Sensitivity to light with medium and long wavelengths, such as green and red light, depends on the quantity of M and L cones present in the retina. The deficit of rod photoreceptors causes a deterioration of night vision¹⁰⁴. Moreover, loss of visual acuity and visual field are other symptoms of ESCS^{106,107}.

Among the thirty described mutations in NR2E3 responsible for diseases, most of them are located in ligand-binding domains or in DNA binding domains. These mutations not only cause ESCS but also Goldmann-Farve syndrome (GFS) or autosomal dominant retinitis pigmentosa (adRP)^{102,106,108,109}.

Single cell RNA sequencing

Single cell RNA sequencing (scRNA-Seq) is a technique that enables transcriptome analysis of isolated cells. This method consists in five main steps: (1) cell separation, (2) reverse transcription, (3) amplification, (4) library generation and (5) sequencing. In this project, 10X Genomics chromium workflow is used to perform the scRNA-Seq⁹⁶ (Figure 1.3).

Single cell partitioning is realized by a droplet-based method. In this method, gel beads are covered with oligonucleotide primers composed of four different parts (Figure 1.4). The first part (R1) is the Illumina TruSeq Read 1 which is a primer binding site used to initiate the sequencing of Read 1 (additional information in materials and methods). This allows the 10X barcode and the UMI to be sequenced. The second part is the 10X barcode which is identical in all the primers of a same bead but different from one bead to the other. It is used to regroup all the RNA sequences coming from the same cell. The third part is the unique molecular

identifier (UMI) that is different for each primer of the bead. Its function is to mark each RNA molecule that has been captured by the bead to avoid PCR quantitative bias during amplification. The last part is a poly(dT) tail that binds with the poly(A) tail of mRNAs. Capture sequences can replace the poly(dT) tail to capture specific RNAs¹¹⁰

The first step of this method is to associate a gel bead with exactly one cell inside a droplet. To that end, microfluidic chips coupled with partitioning oil are used. Once the bead, the enzymes and the cell are in the same droplet, the cell is lysed, mRNAs can bind the bead primers and reverse transcription is then performed. Given that the primers contain the four parts described above, these parts are integrated into the cDNA. Then the complementary strand is synthesized. Afterwards, the cDNA sequences obtained from each droplet are pooled and amplified by PCR¹¹⁰.

The following step is the preparation of the sequencing library which is required for next generation sequencing. It is composed of three successive operations. First, the strands of cDNA that will be sequenced must be fragmented. Indeed, short-read sequencing technologies (Illumina) are limited to sequence DNA fragments of small size. Only the fragments containing the barcode and the UMI are kept. Then, to enable the attachment of the adapters, the formation of an overhang is induced. An overhang is an ensemble of unpaired nucleotides situated at the extremity of a double stranded DNA molecule. In this situation, it is usually a unique adenine base that is added. Finally, the fragments and the adapters are brought together to allow their bonding. Indeed, adapters also possess an overhang which is composed of a single thymine. The two overhangs are therefore complementary which enables the hybridization. Then, a ligase completes the link. These adapters perform two main functions. First, they make the attachment to the flow cell possible. In addition, they can contain indexes to mark all the fragments of a same sample and therefore enable multiplexing¹¹¹. The composition and the functions of adapters is further described in the materials and methods section.

The sequencing of all these fragments is realized simultaneously thanks to the sequencing-by-synthesis method. Briefly, this method consists in sequencing a DNA fragment by synthesizing its complementary strand. The trick is that the nucleotides used are marked with fluorophores and their 3'-OH group is replaced by chemical groups blocking further synthesis. Thus, the identity of the synthesized nucleotide can be determined through the light emitted by the fluorophore. Once the information is gathered, the fluorophore and the 3' blocking group are chemically removed and the following nucleotide is synthesized in the same way. The sequencing is therefore performed step by step throughout the synthesis of the complementary strand.

The advantage of this method is that it enables the sequencing of a tremendous number of fragments in parallel. This is made possible by the hybridization of the fragments onto the flow cell. Indeed, flow cells are covered with primers that are complementary to the adapters added during the library preparation step. By an amplification process called bridge PCR (see methods), each initial fragment will form a cluster containing a high number of strands identical to the initial fragment. The light information can then be studied simultaneously by distinguishing each cluster¹¹².

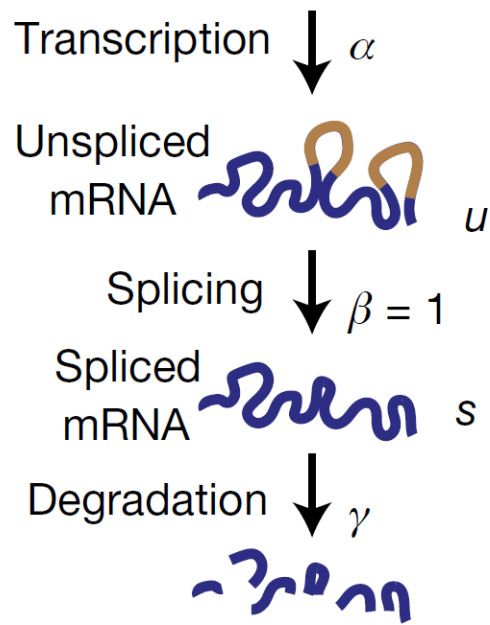


Figure 1.5: schematic representing the processes that define the amount of spliced and unspliced mRNA. α corresponds to the rate of transcription, the splicing rate is represented by β and considered as a constant and the degradation rate is represented by γ . Figure from “RNA velocity of single cells”, figure 1b by La Manno et al., 2018¹¹³.

The last step is the data processing step and is developed in the materials and methods section. For each cell analyzed by scRNA-seq, the number of mRNAs detected for each gene is obtained after the data processing step. This data can be presented as a matrix where cells are represented by rows while each gene is represented by a column.

This information enables the clustering of cells that share a similar transcriptomic profile. Some of these clusters can be matched to known cell types based on the expression of cell type specific marker genes while others clusters could help identifying new cell types. Once clusters are associated with cell types, they can be used to perform differential expression analysis to find genes with particular high or low expression in a given cell type compared with the others. It is also possible to determine if the mRNAs sequenced are spliced or unspliced by searching for intronic regions¹¹³.

RNA velocity

RNA velocity is a new concept that allows to study the process of cell differentiation from single cell RNA sequencing data. Indeed, scRNA-seq captures transcriptome information of a precise instant. This is incompatible with the study of dynamic processes that evolve with time, such as development. In order to get insights into the “direction” of differentiation of cells, the concept of RNA velocity is applied¹¹³.

In fact, the RNA velocity of a specific gene is the quantification of the expected variation in the expression of that gene in a specific cell at the time of sequencing. In other words, the RNA velocity of a gene is a prediction of the level of expression of that gene in the short-term future. The prediction of the evolution of the expression of a gene is based on the proportion of spliced and unspliced mRNAs sequenced for that gene¹¹⁴. Thus, the aim of RNA velocity is (i) to determine if a gene is induced, transcribed at a constant level or repressed and (ii) to quantify this process.

During gene upregulation, the transcription rate increases and the amount of unspliced mRNA grows. Then, these new unspliced mRNAs are progressively spliced which increases the quantity of spliced mRNAs and offset the increase of unspliced mRNA. The quantity of unspliced mRNAs spliced per time unit is proportional to the amount of unspliced mRNAs. In the same way, the quantity of spliced mRNAs degraded per time unit is proportional to the amount of spliced mRNAs. Thus, the intensity of degradation increases with the increase in the amount of spliced mRNAs until a steady state is reached. The steady state represents the state of a cell that expresses a constant number of spliced and unspliced mRNAs through time. Indeed, in this situation, the amount of new unspliced mRNAs generated is equal to the amount of unspliced mRNAs that undergo splicing and equal to the number of spliced mRNAs that are degraded. When a gene is downregulated, the rate of transcription decreases causing a sharp reduction of the production of unspliced mRNA. This downregulation leads to a decrease of unspliced mRNA caused by splicing followed by a decline in spliced mRNAs due to degradation (Figure 1.5).

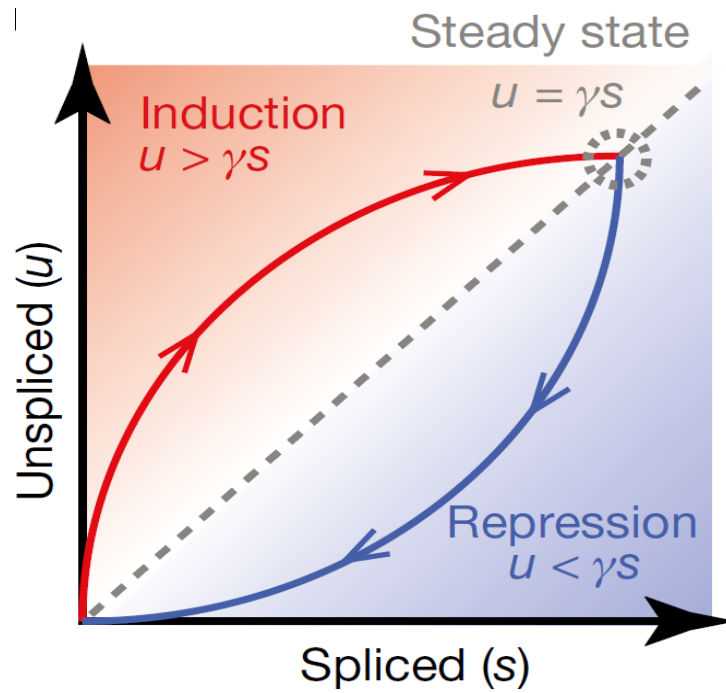


Figure 1.6: Phase portrait showing the transcriptional dynamics (solid curves). *Steady states for different values of transcription rates α fall on the diagonal given by slope γ (dashed line). Levels of unspliced mRNA above or below this proportion indicate increasing (red shading) or decreasing (blue shading) expression of a gene, respectively.* Figure and explanations from "RNA velocity of single cells", figure 1d by La Manno et al., 2018¹¹³.

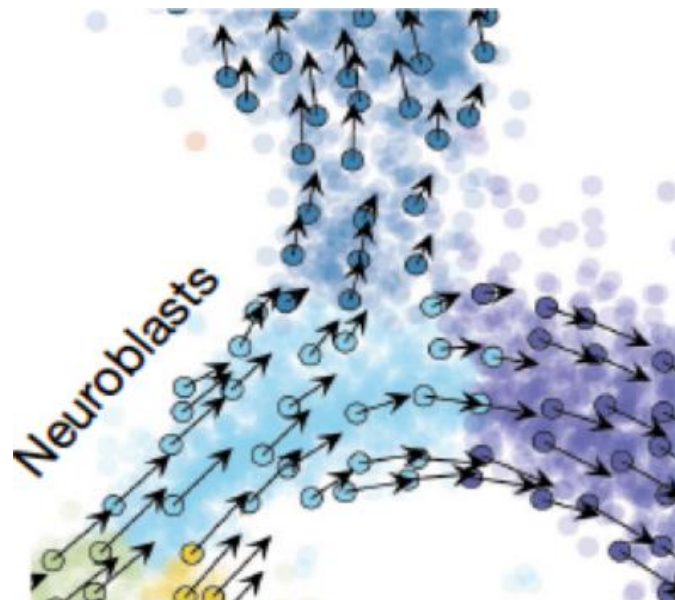


Figure 1.7: Representation of a set of cells with their associated RNA velocities. Each velocity points towards the fate of the associated cell in the short term. Figure modified from "RNA velocity of single cells", figure 3e by La Manno et al., 2018¹¹³.

For a given gene, each cell of the dataset can be plotted on a graph with its values of spliced and unspliced mRNAs as coordinates. This graph is called a “Phase portrait graph” (Figure 1.6). The degradation coefficient γ is determined by performing linear regression on the extreme expression quantiles (see methods). The cells at steady state will always be on the grey dashed line, depending on the transcription rate α . For the cells located above the steady state line, the gene is upregulated (red arrow). Indeed, for these cells the velocity is positive and therefore the amount of spliced and unspliced mRNAs is increasing. Conversely, cells located under the steady state line undergo transcription downregulation of this gene (blue arrow) because their RNA velocity is negative.

Mathematically, RNA velocity is the time derivative of the spliced mRNA amount. It is calculated from the amount of unspliced RNA, the splicing rate, the amount of spliced mRNA and the degradation rate (see methods and Figure 1.5). It can also be obtained as the deviation from the steady state on a phase portrait graph¹¹³.

The RNA velocity of a cell corresponds to a vector the components of which represent the RNA velocity of the different genes (Figure 1.7). This vector enables the prediction of the state of a cell in a very close future based on the genes that are activated and repressed^{113,114}. Therefore, by comparing their vectors, it is possible to find fate differences between neighboring cells. Long term predictions are also possible by combining RNA velocity with Markov chains. Together, these predictions can be used to represent the global movements of cells during differentiation on a UMAP visualization graph (Figure 1.8).

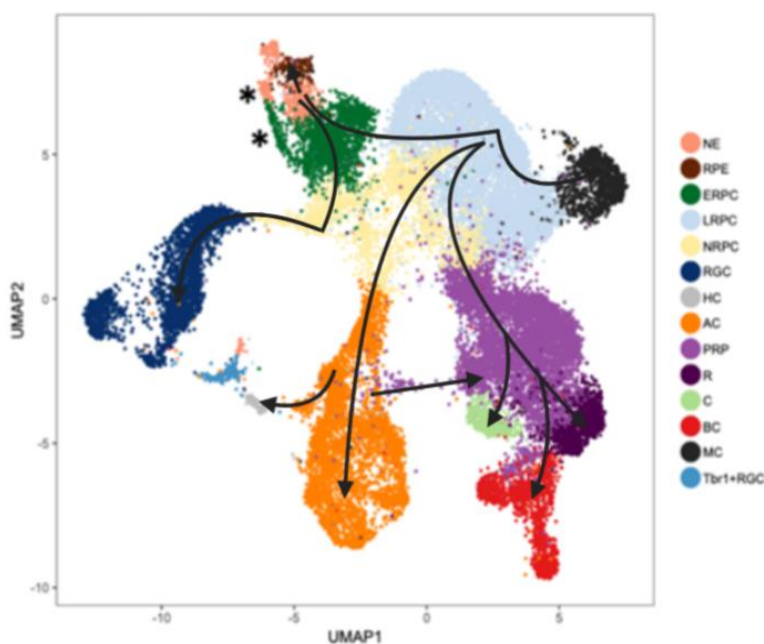


Figure 1.8: 2-Dimensional UMAP representation of the dataset used in this work. Cell types are indicated by the color legend. The black arrows represent differentiation trajectories. These arrows are not based on the data, they are purely illustrative.

Figure modified from "Combined analysis of single cell RNA-Seq and ATAC-Seq data reveals putative regulatory toggles operating in native and iPSC-derived retina" by Georges A, 2020.⁹⁵

Markov chains

As mentioned in the paragraph above, the use of Markov chains enables to make long-term predictions of the state of the cells whose velocities have been calculated. Markov chains enable the description of the evolution of a system through time, based on probabilistic predictions. In this project, Markov chains produce long-term predictions about the transcriptomic profile of cells.

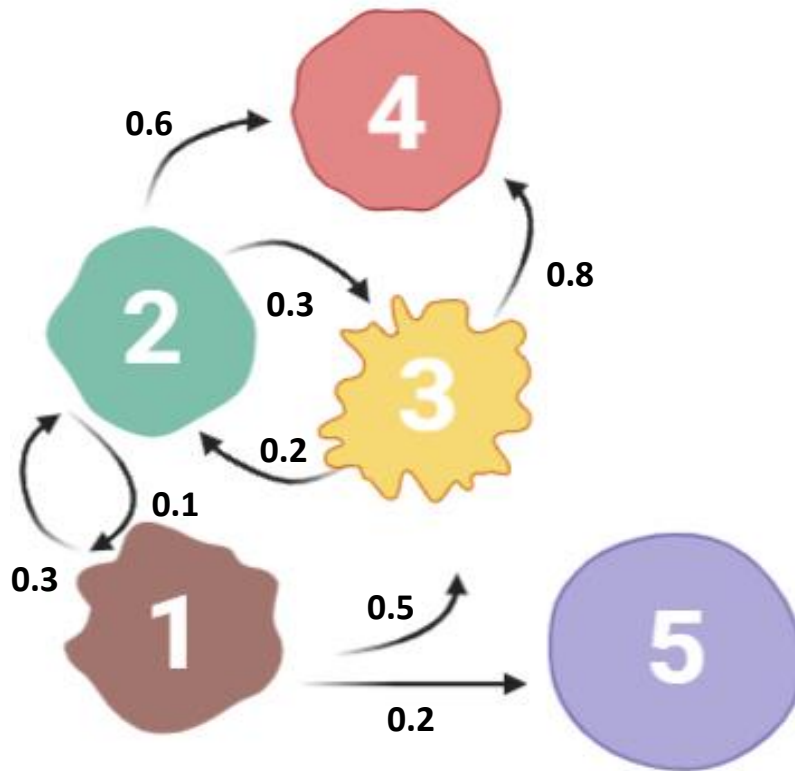


Figure 1.9: Weighted graph of transition probabilities of a hypothetical differentiation process. The probability of transition from a state to itself is not indicated. The sum of the transition probabilities of a state must always be equal to 1.

States	1	2	3	4	5
1	0	0.3	0.5	0	0.2
2	0.1	0	0.3	0.6	0
3	0	0.2	0	0.8	0
4	0	0	0	1	0
5	0	0	0	0	1

Figure 1.10: Transition matrix obtained from the weighted graph of transition probabilities (Figure 1.9). Each cell of the matrix indicates the transition probability from the state corresponding to the row to the state corresponding to the column.

The Markov chain method is a technique that is used in specific situations. Indeed, it is necessary to 1) have a list of states and 2) know the probability to go from each state to any other in one unit of time, called “transition” (Figure 1.9). Based on this information, the aim of the Markov chain is to predict with probabilities, from a starting state, the state of the system after a given number of transitions.

Moreover, for a process to be Markovian, the Markov property must be validated. This property states that the stochastic process used to determine a future state does not take into account the information about the previous states. It depends only on the current state to produce the predictions. Given that the initial state of each cell is known, the state of each cell after “n” transitions can be calculated thanks to the transition matrix by the probability distribution formula (see methods).

In order to apply the Markov chains method to single cell RNA sequencing data, it is required to define the list of states and the transitions probabilities. In our case, the transcriptomic profile of each cell is considered as a state of the system. Concerning the transitions probabilities, it is necessary to create a transition matrix (Figure 1.10). The transition matrix is a square matrix whose size is equal to the number of states existing in the system. It gives the probabilities of transition for all the possible state changes. For example, the transition probability of cell 1 becoming cell 2 in (Figure 1.9) is equal to 0,3.

As mentioned in the previous paragraph, each state corresponds to the transcriptomic profile of a cell of the dataset. In fact, a transcriptomic profile is a list indicating for each gene, the number of spliced mRNAs that were sequenced. This list of values can be seen as the coordinates of the state in a high dimensional space where each gene represents a dimension.

Actually, this high dimensional space is the same space where velocities are defined. Therefore, by plotting all the transcriptomic profiles in this space and by placing the origin of the velocity vectors on the corresponding points, the probabilities of each state to be the future state of a given state can be estimated. These probabilities are calculated using different models detailed in the materials and methods section. In summary, these models give higher transition probabilities to transcriptomic states that are aligned with the RNA velocity of the transcriptomic state whose transition probabilities are computed. The distance between the states is also considered in the majority of the models.

The idea behind Markov chains based long-term predictions is to recapitulate the route followed by each cell during the differentiation process. This route can be obtained by putting end to end a list of little “jumps” in the high dimensional space. Each jump corresponds to a transition of a cell in a given transcriptomic state towards another transcriptomic state. The RNA velocity of a cell only influences its first transition. Then, at each transition, it is the RNA velocity of the reached state that is considered. Given that the orientation of the jump is influenced by the RNA velocity vector and that this velocity is an instantaneous prediction concept, it can only be used for a single jump.

CHAPTER 2

Hypothesis, Objectives and Experimental Strategy

Hypothesis

The VARNAVEL project relies on recent technologies such as organoid formation and single cell RNA sequencing and new theoretical principles such as RNA velocity. Indeed, the improvement of organoid culture techniques from induced pluripotent stem cells now allows to model, in vitro, the development of rare human tissues such as tissues carrying important mutations. Furthermore, single cell RNA sequencing is a technique that allows the sequencing of the transcriptome of individual cells. RNA velocity is a new concept whose objective is to predict the short-term future of a cell in terms of its transcriptome from single cell RNA sequencing data.

By taking samples of organoids at different stages of development, cells in a transitional state between progenitor cells and mature cells are captured. The degree of differentiation of the collected cells depends on the stage at which the sampling was performed. Moreover, cells from the same sample are not exactly at the same level of differentiation. Indeed, some are more differentiated than others. By pooling all the samples, a set of cells is obtained including progenitor cells, mature cells and intermediate cells in a more or less advanced state of differentiation. Using single cell RNA sequencing, the transcriptomic profile of each of these cells can be obtained. Then, these transcriptomic profiles can each be represented by a point in a multidimensional space where each dimension corresponds to a gene. From the sequencing data, the RNA velocity of each cell can also be calculated and plotted in the multidimensional space as a vector. This vector points to the area of the multidimensional space corresponding to the transcriptomic state the cell would have had in the near future if it had continued its differentiation (Figure 1.7).

Since the multidimensional space describes the transcriptomic profile of cells, a change in the degree of differentiation can be seen as a shift in this space. Therefore, each differentiation trajectory corresponds to a global movement starting at the stem cells and ending at the mature cells. In order to recapitulate this whole movement and therefore characterize a differentiation trajectory, a set of small movements will be put end to end.

To do this, a set of states can be defined in which each state corresponds to the transcriptomic profile of a cell in the database. Moreover, each state is associated with its RNA velocity. Only movements from one state to another are possible. In addition, depending on the model used, each movement is associated with a probability of being realized. The probability is proportional to the cosine of the angle between the RNA velocity vector and the vector connecting the 2 states involved in the movement (Figure 2.1). The more the movement and the velocity vector are aligned, the smaller the angle will be and consequently the higher the cosine will be. The movements depend on RNA velocity and are therefore biologically based, which is not the case with other trajectory inference methods that are mainly based on the distances between points¹¹⁵. By letting the cells in the database make a large number of movements, they will move from state to state in the multidimensional space according to the RNA velocity of each state, which is supposed to lead them to states corresponding to more differentiated cells. Markov chains will allow to compute, for each cell, the states in which the cell has the highest probability to be found after a given number of movements.

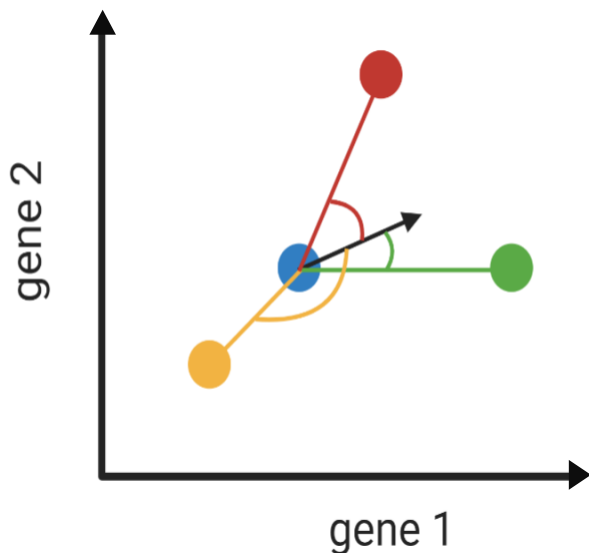


Figure 2.1: Two-dimensional representation of the angle between the RNA velocity vector and the vector connecting the 2 states involved in a movement. Given that each dimension is a gene, the position of the points on the graph depends on their level of spliced mRNAs for these two genes. The blue point corresponds to the starting state, the green, red and yellow points correspond to the neighboring states accessible in one transition and the arrow corresponds to the velocity vector of the starting state. The colored lines correspond to vectors that connect the starting state to the corresponding neighboring states. The angle between the velocity vector and the lines can then be calculated. Figure created in BioRender.com

Objectives

Since this work consists in a participation to the VARNAVEL project, the objective of this work is to program and test different models aiming at reconstructing the differentiation trajectories present in single cell RNA sequencing data from RNA velocity and long-term predictions calculated with Markov chains. The models must be tested to ensure that RNA velocity and Markov chains produce realistic long-term predictions of the fate of cells.

In the VARNAVEL project, the global objective is to that create computer tools that are able to find, by comparing cells from the same individual, the determinants responsible for commitment to one cellular pathway rather than another. Moreover, by comparing cells carrying a specific mutation and cells without that mutation, these tools should also identify the determinants causing the distinct differentiation trajectories taken by these cells. As an example, a mutation in the NR2E3 gene responsible for enhanced S-cone syndrome will be studied in the VARNAVEL project. The output of this tools should indicate which genes are involved in each developmental trajectory, when these genes operate and at what intensity.

Experimental strategy of this work

A dataset containing single cell RNA sequencing data of murine cells from iPSC-derived retinal organoids and native retina at four matched developmental stages will be used. A first step of data processing has already been done by the team that generated the data (Georges et al. 2020⁹⁵). This includes: a demultiplexing step, a read alignment step, the identification of spliced and unspliced mRNAs, a barcode and UMI filtering step, a duplicate marking step, a cell filtering step, a first gene filtering step, the determination of cell types and the generation of UMAP graphs. These different parts are developed in the materials and methods section.

From this data, the RNA velocities of all cells are calculated, metacells are then generated, the new velocities associated with these metacells are computed and a second gene filtering is applied.

These different steps are also developed in the materials and methods section. A new database is created from these metacells and the genes conserved. For the rest of the analysis, metacells will be treated as normal cells. Indeed, they have a transcriptomic profile, an RNA velocity and a cell type, in the same way as the initial cells of the database. Therefore, the different mathematical models for calculating transition probabilities work on both cells and metacells. These models were developed by Loïc Demeulenaere and are described in the material and methods section. These models are tested with different parameters in order to determine under which conditions the differentiation trajectories can be recapitulated in the most biologically accurate way. These tests and their analysis are detailed in the results and discussion sections of this work.

Experimental strategy of the VARNAVEL project

Once the models are able to efficiently recapitulate the different differentiation trajectories, these trajectories can be expressed mathematically thanks to a method called principal curve analysis. Regions where an initial trajectory splits into two distinct trajectories can be identified and analyzed to determine which genes are responsible for this split.

Since monogenic and complex diseases often have an impact on development, the determinants responsible for developmental impairment could be identified using the tools developed. By grouping cells from healthy retinas and cells from diseased retinas in a database, healthy and pathological differentiation trajectories could be identified and compared.

In addition, RNA velocity short-range predictions and Markov chains long term predictions are new variables that can be used to compare neighboring cells. RNA velocities are obtained by considering both spliced and unspliced mRNAs while for cell coordinates, only spliced mRNAs are considered. This is because the RNA velocity only predicts the evolution of spliced mRNAs. Neighboring cells will therefore have a similar profile in terms of spliced RNAs but may have a different profile in terms of unspliced mRNAs and therefore have different velocities and fates. The aim of this comparison is to find regions where neighboring cells share a similar transcriptomic profile but not a similar cell fate. This indicates that the early determinants of these distinct fates are being regulated in these regions. It could then be possible to identify these determinants by analyzing the RNA velocity vectors of these cells.

CHAPTER 3

Materials and methods

Datasets

First Dataset

In the project of Georges et al. 2020⁹⁵, murine cells from iPSC-derived retinal organoids and native retina are sequenced with the scRNA-seq method to form a dataset (Figure 3.1). These cells were harvested at four equivalent developmental stages. The first sampling took place at embryonic day (E) 13 for native retinas which corresponds to the differentiation day (DD) 13 for retinal organoids. The second sampling took place at postnatal day (P) 0 for native retinas and at DD21 for retinal organoids. The third sampling was done at P5 which corresponds to DD25 and finally the last sampling was done at P9 which corresponds to DD29. In the end, data from 38,091 cells were obtained. Of these cells, 21,249 were collected from native retina and 16,842 were collected from iPSC-derived retinal organoids. On average, each cell possesses 5,940 unique molecular identifiers (UMIs, see methods) and 2,471 genes⁹⁵. In this project, this database is used to develop and test computer models that predict cell fate. The protocol used to obtain these cells is developed in the following section.


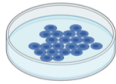
				
	Native retina		iPSC-derived retina	
	Biological replicate 1	Biological replicate 2	Biological replicate 1	Biological replicate 2
E13/DD13	1641	1571	1567	2291
P0/DD21	1456	1761	1498	1895
P5/DD25	4513	6212	3543	3890
P9/DD29	4468		3993	

Figure 3.1: Overview table showing the number of cells per sample at the four matched development stages for each origin of the retina and biological replicate. The values in this table are those obtained before a first cell filter. Therefore, the total amount of cells is slightly higher than the amount of cells described above.

Second Dataset

A second dataset will be produced by performing scRNA-seq on cells obtained from samples of human iPSC-derived retinal organoids at different developmental stages. These retinal organoids are derived from a mix of two types of iPSC. The first type is collected from healthy donors while the second is collected from patients carrying the NR2E3 mutation responsible for Enhanced S-cone syndrome. Therefore, the retinal organoids obtained contain cells with the mutation in the NR2E3 gene and cells without this mutation. The goal of generating hybrid organoids is for cells to share the same culture conditions regardless of the fact that they carry the mutation or not. This dataset will enable the comparison of the differentiation process of healthy cells with the differentiation process of mutated cells and thus provides insights into the effects of the NR2E3 mutation during retinal development (Figure 1.1).

Generation of the first database

The steps performed to obtain the database (maintenance of iPSCs, generation of iPSC-derived retinal aggregates, dissociation of native retinal tissue and 3D-culture retinal aggregates and single cell RNA sequencing) are described in the materials and methods section of the original article from Georges et al. 2020⁹⁵. For the purpose of this work, only the fact that Chromium Single Cell 3' reagent kits v2.0 were used to produce the sequencing libraries will be mentioned⁹⁵.

Single-cell RNA sequencing

The 10X Genomics methodology of single cell RNA sequencing is developed in the introduction. The purpose of this section is 1) to explain the bridge PCR amplification required to perform sequencing and 2) the post-sequencing data analysis step.

Bridge PCR

In order to perform the sequencing-by-synthesis method, a step of amplification of the fragments composing the sequencing library is required to make the light signal strong enough to be detected. As mentioned in the introduction, each cDNA fragment in the library is contained between two adapters. These two adapters are different from each other. The adapter next to the 10X barcode is composed of one part (P5) while the adapter next to the cDNA fragment is composed of three parts (TruSeq Read2, Sample Index, P7) (Figure 3.2). As a reminder, the TruSeq Read 1, the 10X barcode and the UMI are contained in the primer attached to the bead. TruSeq Read 1 and 2 are sequencing primer binding sites because it is where the sequencing primer hybridizes in order to initiate sequencing. TruSeq Read 1 will enable the sequencing of read 1 which contains the sequence corresponding to the 10X barcode and the UMI. TruSeq Read 2 is used to sequence read 2 containing the sequence corresponding to the cDNA fragment. Sample Index is the index sequence which identifies samples and therefore allow for multiplexing. Finally, P5 and P7 are regions that are complementary to the oligonucleotides that are attached to the flow cell. In fact, there are two different sequences of oligonucleotides that cover the flow cell. Thus, P5 is complementary to one of the two sequences and P7 is complementary to the other sequence. To be precise, since the two strands of cDNA are complementary, only the region at the 3' end is directly complementary to the oligonucleotides. Therefore, one strand will connect to the flow cell thanks to the P5 region while the other strand will connect to the flow cell thanks to the P7 region.

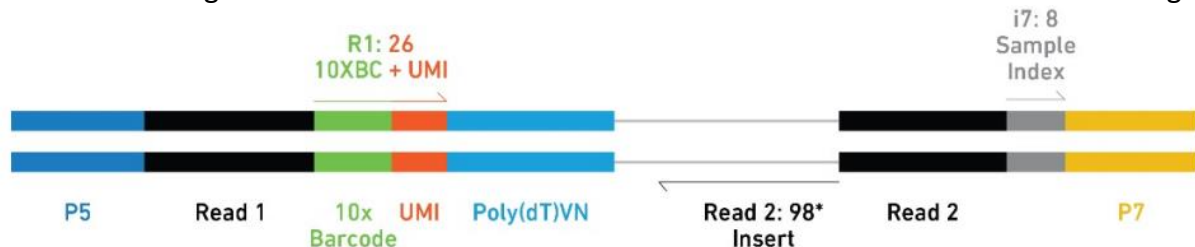


Figure 3.2: Illustration of a fragment of DNA composing a Chromium Single Cell 3' gene expression library. Figure from Chromium Single Cell 3' Reagent Kits v2 User Guide (2019)¹¹⁶.

Once the adapters are fixed to the DNA fragments, the two strands of each cDNA molecule are separated and the whole library is injected onto the flow cell. The strands of DNA hybridize to the complementary oligonucleotide and the complementary strands are synthesized. After a step of denaturation, the original strands are washed away while the new strands stay on the flow cell because they are attached to it. By complementarity, the new strands contain either the P5 or the P7 part at their 3' end and can thus hybridize with the other type of oligonucleotide. Given that these strands are attached to the flow cell, the hybridization of the extremity of these strands with the other type of oligonucleotides induces the folding of these strand which then takes the form of a bridge. The complementary strands are synthesized and are also attached to the flow cell through the oligonucleotide. These strands are then separated by a step of denaturation.

There are now two versions of the strands attached to the flow cell: a strand produced during the first step and complementary to the original strand and a strand produced during the second step and identical to the original strand. The process of bridge formation by hybridization of the 3' end is repeated a number of times to create a cluster of strands originating from the same molecule and attached to the plate. At the end of this cycle, each cluster contains as many complementary molecules as identical molecules.

In order to sequence Read 1, the strands linked to the flow cell via the P5 sequence are cleaved and washed away. To avoid unwanted RNA synthesis at the 3' ends of the fragments and oligonucleotides, their 3' end are blocked. Finally, the correct sequencing primers are added and hybridize to the sequencing primer binding site (TruSeq Read 1) in order to prepare the sequencing. Following the sequencing of Read 1, the index can be sequenced from the same strand using a different type of primer. The sequencing of Read 2 is done on the strands linked to the flow cell via the P5 sequence that were previously cleaved. These strands are generated again through a bridge formation step followed by a complementary strand synthesis step. This time, it is the strands attached to the plate via the P7 sequence that are cleaved. The primers specific to the sequencing primer binding site (TruSeq Read 2) are added in order to perform the sequencing (Figure 3.3). A read is the sequence data obtained by analyzing the light signals emitted by a cluster. Illumina sequencers produce reads ranging from 75 to 150 base pairs in length¹¹⁷.

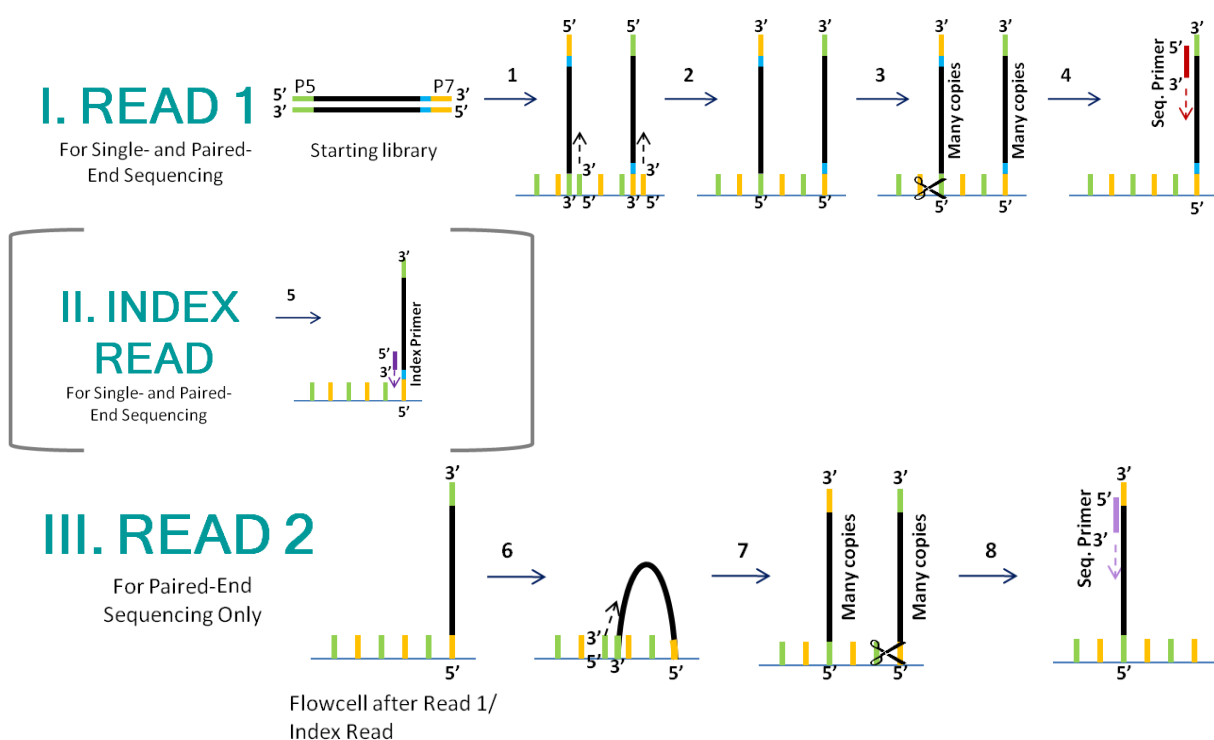


Figure 3.3: Illustration of the different steps required to obtain the 3 reads corresponding to a library fragment. Figure from the MGH NextGen Sequencing Core website¹¹⁸.

Data processing

A first step of data processing has already been done by the team that generated the data (Georges et al. 2020⁹⁵). This includes: the demultiplexing step, the read alignment step, the identification of spliced and unspliced mRNAs, the barcode and UMI filtering step, the duplicate marking step, the

cell filtering step, the first gene filtering step, the determination of cell types and the generation of UMAP graphs.

Demultiplexing

Next generation sequencing techniques enable the simultaneous sequencing of different samples. The demultiplexing step consists in grouping the reads coming from the same sample and saving their information in FASTQ files. This is made possible by the use of sample indexes which are sequences of 8 pair bases that mark all the sequences of a sample. Each sample is marked by a different sample index which allows to identify them¹¹⁹. This step was performed on the dataset by using Cell Ranger v2.1.1 (10X Genomics, CA)⁹⁵.

Read alignment

The objective of this step is to determine, for each read, the locus that most likely produced the sequence corresponding to the read. The alignment can be performed on a reference genome and on a reference transcriptome¹¹⁹. This technique enables the interpretation of the data produced by single-cell RNA sequencing. Given that some reads can map several loci in the genome, Cell Ranger uses only confidently mapped reads that align to a single gene for UMI counting. This step was performed on the dataset by using Cell Ranger v2.1.1 (10X Genomics, CA)⁹⁵.

Identification of spliced and unspliced mRNAs

In order to compute RNA velocities, it is necessary to determine if the transcripts associated to each UMI are spliced or unspliced. Since the 10X method only retains the reads containing the barcode and the UMI after fragmentation, only the 3' fragment of the initial transcripts is analyzed. With this read, it is possible to try to determine if the initial transcript was spliced or unspliced. Indeed, either the read contains an intronic region and is considered unspliced, or it is composed of at least 2 different exonic regions without containing an intronic region and is considered spliced, or it aligns only on one exon and therefore cannot be classified with certainty. This step was performed on the dataset by using Cell Ranger v2.1.1 (10X Genomics, CA)⁹⁵.

Barcode and UMI filtering

10X cell barcodes are used to identify the cell from which the read comes. This is possible because each Gel Bead (see introduction) possesses a unique barcode of 16 base pairs. All the barcodes used are known before the sequencing. This is useful to remove reads whose barcode has been subject to mismatches during the sequencing. However, it is still possible to retrieve the information in the case only one mismatch happened and in a low-quality position. The barcode is corrected and the read is associated to the right cell¹¹⁹.

UMIs is the abbreviation for unique molecular identifiers and refers to short different sequences that tag uniquely each molecule of the sequencing library. Singular UMIs must also be removed. This includes homopolymer UMIs (such as AAAAAAAAAA) and the presence in the UMI of a base whose quality is not sufficient to ensure its identity. In a similar way to barcodes, UMIs that present one nucleotide mismatch compared with a higher-count UMI can be corrected to that UMI if they share a common cell barcode and correspond to the same gene¹¹⁹.

These two steps were also performed on the dataset by using Cell Ranger v2.1.1 (10X Genomics, CA)⁹⁵.

Duplicate marking

Thanks to UMIs, it is possible to identify reads that are duplicates of the same RNA molecule. By doing so, each different UMI detected corresponds to a unique RNA molecule. At this point, the unfiltered gene-barcode matrix can be obtained and indicates for each barcode, the number of UMIs detected for each gene¹¹⁹. This step was also performed on the dataset by using Cell Ranger v2.1.1 (10X Genomics, CA)⁹⁵.

Cell filtering

A first filter is applied to the dataset and consists in eliminating the barcodes whose total quantity of UMIs is lower than a given threshold. This threshold is computed as one tenth of the 99th percentile of the UMI counts among the expected recovered cells¹¹⁹. These cells are the N cells that have the highest UMI counts, N being a parameter specified by the user and which is by default equal to 3000¹²⁰. The objective of this filter is to remove the barcodes associated to Gel Beads that have not been in contact with a cell and which, therefore, are considered as noise. Once this step is completed, the filtered gene-barcode matrix can be obtained. This matrix is different from the unfiltered gene-barcode matrix in that each barcode is now supposed to correspond to a cell.

The second filter aims to remove aberrant data originating from several adverse events, namely the sequencing of multiplets, the sequencing of apoptotic cells and the presence of poorly informative cells¹¹⁹. Two variables are considered in order to perform this filter: the number of detected genes and the fraction of reads mapping to mitochondria-encoded genes. Because of inefficient mRNA capture, some cells do not contain enough information to be correctly interpreted. These cells are not stopped by the first filter because the number of UMIs is still above the threshold. In order to eliminate these cells, a minimum number of different genes detected is imposed. Any cell with a number of genes lower than this value is removed. In the same way, a maximum number of different genes detected is set to remove data coming from multiplets. Indeed, when more than one cell is captured with a Gel Bead, the data from these cells share the same barcode and are therefore no longer differentiable. Finally, a high proportion of reads mapping to mitochondria-encoded genes suggests that the cytoplasmic mRNAs have leaked out of the cell due to membrane damage and therefore only the mitochondrial mRNAs are still present. These damaged cells are removed by setting a maximum threshold (around 5 to 7 percent) for the fraction of reads mapping to mitochondrial genes¹¹⁹. These steps were also performed on the dataset by using Cell Ranger v2.1.1 (10X Genomics, CA)⁹⁵.

First gene filtering

The first filter of genes applied to the dataset is the one used in the article of La Manno et al. of 2018¹¹³. The average numbers of spliced and unspliced mRNA molecules are calculated for every cell type and for each gene. For a gene to be selected, these average numbers must be superior to a specific threshold in at least on cell type. The threshold is not the same for spliced and unspliced RNA molecules. The thresholds applied to our dataset are the same as those applied by La Manno et al.¹¹³ and their values are 0.5 for spliced molecules and 0.1 for unspliced molecules. In addition, the correlation between spliced and unspliced mRNAs must be greater than 0.05 for a gene to be selected.

Once the genes are selected, the number of molecules of spliced mRNAs and the number of molecules of unspliced mRNAs for a gene are respectively divided by the total number of spliced and unspliced molecules in the cell. This process is performed for every gene selected and for each

cell in the dataset. Finally, the spliced and unspliced values of each gene are standardized by dividing by the standard variation. This step was also performed on the dataset by using Cell Ranger v2.1.1 (10X Genomics, CA)⁹⁵.

Determination of cell types

In order to assign a cell type to the cells in the database, a clustering step must first be performed to group cells sharing a similar transcriptomic profile. The clustering method used on the dataset is the k-means clustering available in the Seurat software suite. The number of clusters is a parameter chosen by the user. In the case of our database, the number chosen is equal to 70. Thus, 70 clusters regrouping the 38,091 cells of the database were obtained (Figure 3.4). A cell type is then associated to each cluster by measuring the expression of specific marker genes of cell types of the retina. These gene expression signatures of retinal cell types come from the article by Clark et al. of 2018³⁶. Table 1 shows the list of marker genes for each cell type. Once each cluster has been assigned its cell type, clusters of the same cell type can be merged to obtain groups, each corresponding to a cell type (Figure 3.5). From our database, 14 groups were created corresponding to: neuroepithelium, retinal pigment epithelium, early retinal progenitor cells, late retinal progenitor cells, neurogenic retinal progenitor cells, retinal ganglionic cells, Tbr1 positive retinal ganglionic cells, horizontal cell, amacrine cells, photoreceptor precursor cells, cones, rods, bipolar cells and Müller cells.

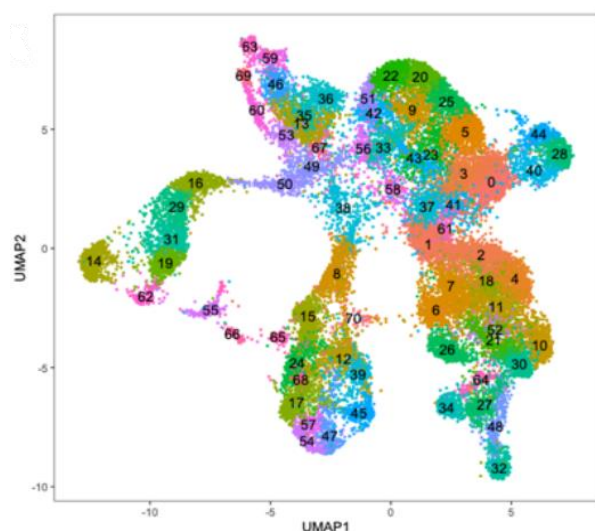


Figure 3.4: 2D UMAP manifold showing NaR and 3D-RA cells jointly and their assignment to 70 clusters by k-means clustering. Figure and explanations from "Combined analysis of single cell RNA-Seq and ATAC-Seq data reveals putative regulatory toggles operating in native and iPSC-derived retina" by Georges A, 2020⁹⁵. NaR = native retina, 3D-RA = 3D-retinal aggregates (organoids).

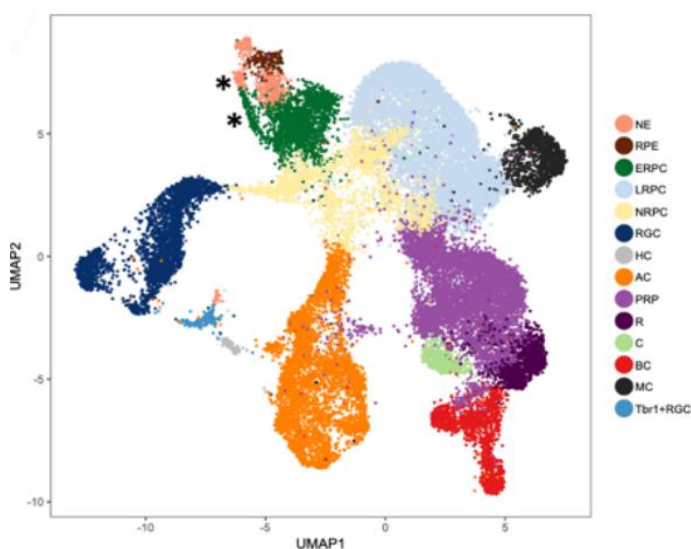


Figure 3.5: Merging of the clusters in 14 cliques corresponding to neuroepithelium (NE), retinal pigmented epithelium (RPE), early (ERPC), late (LRPC), neurogenic retinal progenitor cells (NRPC), retinal ganglionic cells (RGC), Tbr1 positive retinal ganglionic cells (Tbr1+RGC), horizontal cells (HC), amacrine cells (AC), photoreceptor precursor cells (PRP), cones (C), rods (R), bipolar cells (BC), and Müller cells (MC), on the basis of the expression of known marker genes. Cluster 69 (NE) and 60 (ERPC) (marked by asterisks) correspond to the ciliary marginal zone (CMZ) which forms a branch that clearly separates from the rest of NE and ERPC. Figure and explanations from "Combined analysis of single cell RNA-Seq and ATAC-Seq data reveals putative regulatory toggles operating in native and iPSC-derived retina" by Georges A, 2020.⁹⁵

NE	RPE	ERPC	LRPC	NRPC	RGC	HC	AC	PRP	R	C	BC	MC
Wfdc1	Foxg1	Sfrp2	Lhx2	Olig2	Pou4f2	Lhx1	Pax6	Crx	Nrl	Thrb	Vsx2	Sox9
Mitf	Mitf	Fgf15	Ascl1	Neurog2	Nefl	Onecut1	Barhl2	Otx2	Rho	Opn1sw	Grm6	Rlpb1
Ccnd2	Best1	Cdk4	Hes5	Hes6	Thy1	Onecut2	Nhlh2	Neurod1	Rcvrn	Gngt2	Vsx1	Slc1a3
H19	Mertk	Six6	Hes1	Atoh7	Rbpms	Calb1	Bhlhe22	Prdm1	Prdm1	Prdm1	Isl1	Aqp4
Mest	Otx1		Notch1	Ptf1a	Calb2		Tfap2a	Btg2		Crx		Clu
Msx1	Rpe65		Nfix	Neurod1	Atoh7		Tfap2b	Rax				Nfia
			Six3	Hes2			Ccnd1	Nfib				Glul
			Id1				Six3					Crym
			Id2				Calb2					

Table 3.1: Marker genes used for cell type discrimination between 13 different cell types (NE = neuroepithelium, RPE = retinal pigment epithelium, ERPC = early retinal progenitor cell, LRPC = late retinal progenitor cell, NRPC = neurogenic retinal progenitor cell, RGC = retinal ganglion cell, HC = horizontal cell, AC = amacrine cell, PRP = photoreceptor precursor cell, R = rod, C = cone, BC = bipolar cell and MC = Müller cell). Adapted from Georges et al. 2020⁹⁵, Supplementary Table 3.

Dimensionality reduction and visualization

The aim of dimensionality reduction (DR) is to decrease the number of variables describing the data while retaining relevant information of the original data. To that end, two methods can be applied: “feature selection” and “feature extraction”. The first one consists in selecting the best variables among the initial variables while the second one consists in generating, from the initial variables, new variables explaining a maximum of the variability of the original data. Feature extraction is more efficient than feature selection to describe the original data. Nevertheless, the created variables do not have a concrete meaning unlike the selected variables which are unchanged. In some cases, the new variables can still be interpreted because it is possible to know which variables make up the new variable and in what proportion. This is the case for the principal component analysis. However, this is not the case for the variables obtained with the UMAP algorithm.

In this project, two DR algorithms are used: Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). UMAP is a new algorithm that replaces the t-SNE algorithm that was the most commonly used technique for visualization of scRNA-seq data. UMAP is indeed better than t-SNE to represent global structures of data while being as efficient to highlight local structures. Moreover, UMAP is performed with a shorter runtime ¹²¹.

In our situation, the main advantage of DR is to visualize the data in 2D or 3D graphs using feature extraction as in Figure 3.4 and Figure 3.5. Indeed, the goal is to see how the different cell types are arranged and to visualize the developmental trajectories. Therefore, desired information can be observed even if the new variables do not represent genes anymore. However, feature selection is also used when it comes to reducing the number of variables considered where the meaning of the variables (genes) is important, namely to analyze the RNA velocity of cells and to perform

differential expression analysis. This filtering of genes helps to reduce noise and decrease the computational resources required.

Metacells

In order to smooth the data and reduce noise, a common technique is to use the k Nearest Neighbors algorithm (kNN). In short, the coordinates of each cell are replaced by the average obtained from the coordinates of this cell and those of its nearest neighbors. The drawback of this approach is that the independence between cells is lost since the same cell can be used to calculate the new coordinates of several cells. This approach is used in the two papers describing RNA velocity^{113,114}.

Our idea is to cluster the data into groups of fixed size k. The metacells are then the means of the different groups, which are disjoint from each other. In fact, metacells result from the fusion of the original cells composing these groups. Given that we want the groups to be constituted of cells with a similar transcriptomic profile, it is necessary to use a method that reduce the number of metacells created from dissimilar cells. In order to do that, the clustering step to generate the groups is performed as follows: the cell which is the furthest away on average from the others (among those still available) is selected, then the (k-1) closest cells among the remaining ones are put in its group. The coordinates of the metacell correspond to the average of the coordinates of the cells in the group. The RNA velocity of the metacell is then calculated from the average values of the number of spliced and unspliced mRNAs. Finally, the cell type assigned to the metacell is the majority cell type among the cells of the group. All cells in the new group are removed from the set of available cells and the process is repeated until no more cells are available. Thus, we continue to do kNN, but this time across groups of independent cells.

Using metacells instead of single cells is beneficial for different reasons. The main advantage is to decrease the noise of the dataset. Indeed, the goal of using means is to smooth the variations caused by the random binding of mRNAs to primers during the scRNA-seq. This way, underestimates offset overestimates while increasing the size of the metacell reduces the impact of these individual anomalies. Therefore, it is crucial that the cells composing the metacell share an equivalent transcriptomic profile. Otherwise, relevant information of each cell could be muted causing a misleading interpretation. In addition, working with metacells reduces the size of the dataset and therefore shortens calculation time. This is particularly true when it comes to compute the transitions matrix corresponding to a high transition. In fact, the complexity of this algorithm increases with the cube of the number of cells.

Second gene filtering

Given that RNA velocities are used to determine the long-term fate of cells by a series of little “jumps” in the high dimensional space, the prediction of fates is more efficient if the RNA velocities of cells sharing a similar transcriptomic profile define a "logical" flow. With this in mind, the velocity of each gene of each cell is compared with the velocity of the same gene of close cells. If the velocities of close cells are significantly more similar than velocities randomly chosen, the gene is preselected (Figure 3.6). This means that this gene is more involved in the differentiation dynamics than noise.

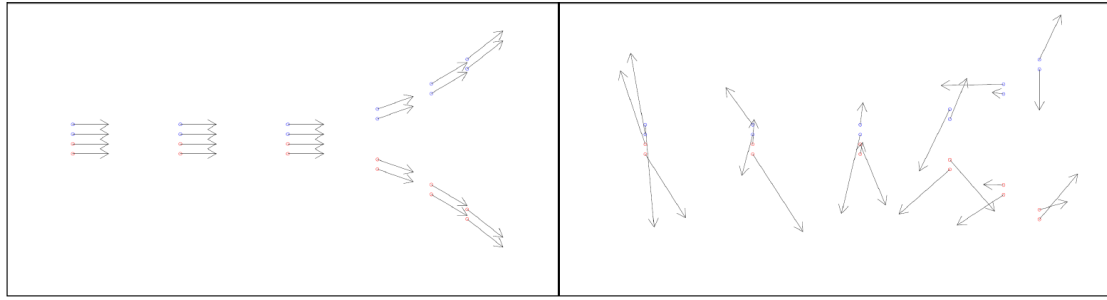


Figure 3.6: Two flows of RNA velocities for a same cellular dataset. Left: an ideal flow of velocities. Right: a noisy flow of velocities. Figure and explanations from “Using local variations of RNA velocity to filter genes”, figure 1 by Loïc Demeulenaere, 2021¹²².

In order to measure the dissimilarity of two vectors of the same dimension of which at least one is non-null, we use the following formula:

$$\Delta(\vec{x}, \vec{y}) := \frac{\|\vec{x} - \vec{y}\|}{(\|\vec{x}\| + \|\vec{y}\|)/2} = 2 \frac{\|\vec{x} - \vec{y}\|}{\|\vec{x}\| + \|\vec{y}\|},$$

Where the dissimilarity (Δ) between the vector \vec{x} and \vec{y} correspond to the length of their difference normalized by the mean of their norms ($\|\cdot\|$). In case \vec{x} and \vec{y} are equal to zero, we set $\Delta(\vec{x}, \vec{y}) = 0$. This formula limits the values that Δ can take between 0 and 2. Moreover, Δ is equal to 0 when \vec{x} and \vec{y} are equal and is equal to 2 when \vec{x} and \vec{y} are parallel, but with opposite directions.

This value of dissimilarity enables the comparison of velocities of cells. As mentioned above, it is desirable that neighboring cells have similar velocities. To do this, a local measure of velocity dissimilarity is defined for each cell as the mean of dissimilarity between that cell and its l nearest neighbors. Given that the calculation of Δ works in any finite dimension and that each dimension corresponds to a gene, it is possible to study genes individually by considering vectors whose dimension is equal to one. By doing this, each cell will be assigned a dissimilarity value for each gene. As a result, each gene will have its own distribution of local measures of dissimilarity, independent of other genes. In order to represent noise, a random measure of velocity dissimilarity is computed. For each cell, l random cells are selected instead of nearest cells. Then, the dissimilarity of their velocities is computed for each gene in the same way as neighboring cells.

Thanks to these two variables, it becomes possible to select genes. Indeed, for each cell we have a local dissimilarity and a random dissimilarity. The significance of the difference between local and random dissimilarities can then be assessed by paired-sample t-test.

The genes can then be sorted according to the p-values obtained. Once the genes are ranked, the local and random measures for cumulated genes can be computed. The idea is to perform t-tests again but this time by accumulating genes. First, the significance of the alignment is tested for the best gene alone, then the two best genes together, then for the three best genes and so on. Given that genes are ranked from the best to the worst in terms of the significance of the alignment, the significance of the alignment will globally increase until it reaches a maximum and then decrease progressively (Figure 3.7). The set of genes inducing the best significance of the alignment is selected and the other genes are removed from the dataset. Once the cells and genes are filtered, the database is stored in a loom file.

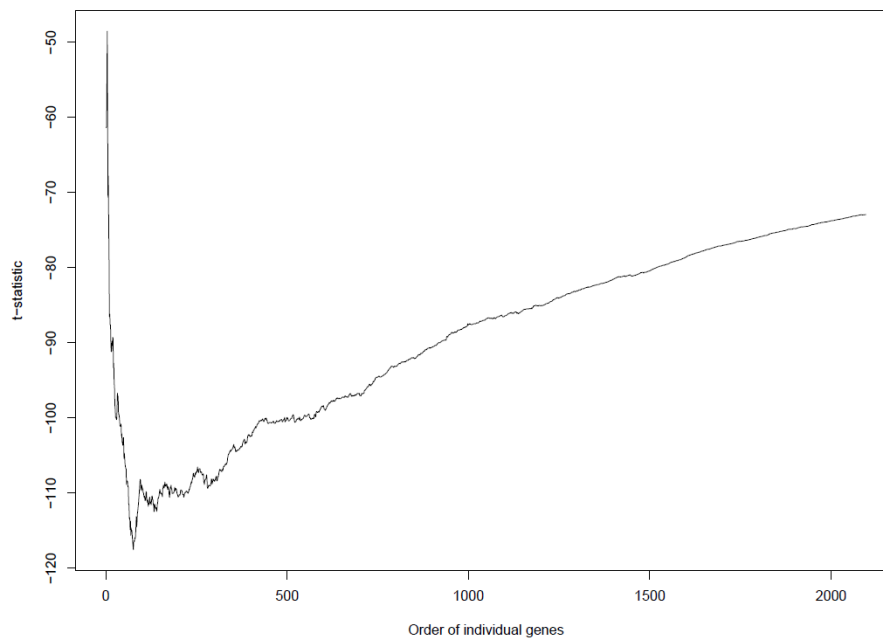


Figure 3.7: *t*-statistics for tests comparing local and random velocity dissimilarity for growing sets of genes. Figure and explanations from “Using local variations of RNA velocity to filter genes”, figure 2 by Loïc Demeulenaere, 2021¹²².

In our case, the second gene filtering was performed on metacells instead of cells. The reasoning is identical but the result varies according to the size of the metacells. When generating metacells of size 5, 125 genes are selected while with metacells of size 10, 76 genes are selected. In addition, the 76 genes of the metacells of size 10 are included in the 125 genes of the metacells of size 5.

Loom files

Loom file format is a specific file format used to store large biological datasets, generally obtained from omics studies. The data is organized in different matrixes while metadata is stored in rows and columns. What sets the loom format apart from other data storage methods is that it treats the data as a matrix and not as a table. This allows the columns (which represent cells) and rows (which represent genes) to be managed in the same way. Therefore, it is very easy to add metadata to both cells and genes unlike other types of data storage. Different kinds of data can be added such as the name of the cell, the UMAP coordinates of the cell, the cluster and thus the cell type of the cell, the stage of development, whether it comes from an organoid or not. As far as genes are concerned, their names and the chromosome on which they are located are examples of information that can be specified in the metadata of the loom file. Finally, the loom format was developed to efficiently access the desired columns and rows without having to load the entire matrix into memory¹²³.

Computation of RNA velocities

The predictions are based on the following concept: 1) the variation of quantity of unspliced mRNAs at the instant t is calculated by subtracting to the transcription rate at the instant t ($\alpha(t)$), the product of the splicing rate ($\beta(t)$) and the quantity of unspliced mRNAs ($u(t)$) at the instant t . This product represents the amount of mRNAs spliced per time unit. 2) The variation of quantity of spliced mRNAs at the instant t is calculated by subtracting to the product of the splicing rate ($\beta(t)$) and the quantity of unspliced mRNAs ($u(t)$) at the instant t , the product of the degradation

rate ($\gamma(t)$) and the quantity of spliced mRNAs ($s(t)$) at the instant t . The first product represents the amount of mRNAs spliced per time unit while the second represents the amount of mRNAs degraded per time unit¹²⁴.

$$(1) \quad \frac{du}{dt} = \alpha(t) - \beta(t) u(t) \quad (2) \quad \frac{ds}{dt} = \beta(t) u(t) - \gamma(t) s(t)$$

In order to solve these equations, some approximations are imposed. For a given gene, α corresponds to the rate of transcription and is assumed to be constant through time. The degradation rate is also constant through time and is represented by γ . Finally, the splicing rate is represented by β and is considered to be constant across all genes¹²⁴. Moreover, by measuring all rates in units of the splicing rate, we can set $\beta = 1$. The equations become:

$$(3) \quad \frac{du}{dt} = \alpha - u(t) \quad (4) \quad \frac{ds}{dt} = u(t) - \gamma s(t)$$

By definition, RNA velocity corresponds to the temporal derivative of the quantity of spliced mRNAs. It can be calculated by using the 4th equation. For each gene, the value of γ must thus be determined. This is possible from the population of cells in the steady state of each gene. Indeed, for this population, the amounts of spliced and unspliced RNAs stabilize in the long term, which implies that the derivatives become equal to zero. By inserting these values into the 4th equation, it becomes:

$$(5) \quad \begin{aligned} 0 &= u(t) - \gamma s(t) \\ \Leftrightarrow \gamma s(t) &= u(t) \\ \Leftrightarrow \gamma &= \frac{u(t)}{s(t)} \end{aligned}$$

Equation number 5 indicates that γ is the ratio of unspliced to spliced mRNA molecules in steady state populations. The cells at steady-state are those which intercept the line of equation $u(t) = \gamma s(t)$. In the article of La Manno et al.¹¹³, this line is approximated by linear regression of the cells belonging to the extreme quantiles. This is a linear regression of the amount of unspliced RNA against the amount of spliced RNA that is restricted to cells whose amount of spliced RNA is below their 5% quantile or above their 95% quantile¹²⁴. In our case, the line is obtained by linear regression of all the cells while imposing to pass by the origin of the axes. This allows the regression to be more robust against outliers. Once the degradation coefficient γ is defined for a gene, the RNA velocity of this gene can be computed for each cell of the dataset by using the 4th equation.

Computation of probability distributions using Markov chains

The following two equations are deduced from the properties of Markov chains. π_n represents the probability distribution after n transitions and P represents the transition matrix. The probability distribution is a row vector with a number of entries equal to the number of states. Each entry refers to a state and indicates the probability that the system is in that state. These properties are verified if n is a natural number.

$$(1) \quad \pi_{n+1} = \pi_n \cdot P \quad (2) \quad \pi_n = \pi_0 \cdot P^n$$

This means that 1) by multiplying the probability distribution after n transitions by the transition matrix, we obtain the probability distribution after $n+1$ steps and 2) by multiplying the initial probability distribution by the transition matrix raised to the n^{th} power, we obtain the probability distribution after n transitions. In addition to these two properties, a third property indicates that: for all states i and j and for any natural number $n \geq 1$, the coefficient in row i and column j of the matrix P^n is the probability of going from state i to state j in n transitions.

Since in our case the initial probability distribution is known and the transition matrix can be calculated using the different models described below, the probability distribution after n transitions can also be calculated. In fact, when the long-term future of a cell is studied, its initial state is known and therefore the third property can be used to determine the probability distribution after n transitions. Indeed, if we study the transcriptomic state after n transitions of the third cell in the database, its probability distribution corresponds to the third row of the transition matrix raised to the power n . This is a consequence of the second property. When the initial state is known, the probability distribution contains only zeros except one 1 at the position corresponding to the initial cell. The product of this row vector and the transition matrix raised to the n^{th} power results in a row vector corresponding to the row associated with the initial cell in the transition matrix raised to the n^{th} power.

Prediction models

The purpose of the models of prediction described in this section is to calculate the transition probabilities of each cell in the database. Among these models, two variables are systematically used: the RNA velocity vector of the studied cell and the vector connecting that cell to a given neighbor. Since these two vectors are defined in the same geometric space and have the same point of origin, namely the studied cell, it is possible to calculate the cosine of the angle between these two vectors. This value is obtained by dividing the scalar product of these two vectors by the product of their magnitude. This formula is obtained from the definition of the scalar product:

$$\vec{x} \cdot \vec{y} = \|\vec{x}\| \cdot \|\vec{y}\| \cdot \cos(\vec{x}, \vec{y})$$

Indeed, by dividing the two members of the equality by $\|\vec{x}\| \cdot \|\vec{y}\|$, we obtain:

$$\cos(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y} / \|\vec{x}\| \cdot \|\vec{y}\|$$

The utility of calculating the cosine is that it allows the measurement of the alignment of the two vectors. Indeed, the cosine can take values between -1 and 1. When the angle between the two vectors is equal to 0° , the value of the cosine is maximum and is therefore equal to 1. Then the cosine decreases as the amplitude of the angle increases until reaching -1 when the angle is the largest possible, namely 180° . Moreover, the value of the cosine is equal to 0 when the angle is equal to 90° .

The notion of cosine appears in all models because it is the heart of the reasoning. In fact, it consists in increasing the probabilities that the studied cell transits to cells which are aligned with its velocity. Moreover, for a mathematical reason, only a limited number of neighbors are considered. Otherwise, it is possible that the distribution of long-term fates is uniform.

The other important notion that is absent from the first model is the consideration of the distance that separates the two cells. Since RNA velocity is an instantaneous value that can rapidly change

over time, its interpretation makes more sense when looking at the near future of the cell. In our database representing transcriptomic profiles, the notion of time can be replaced by that of distance. Indeed, it takes a certain time for a cell to change its transcriptome significantly. Following this reasoning, it is natural to think that the closer the cells are in terms of transcriptome, the higher the transition probability should be.

In addition to the choice of the model, the values of two parameters must be filled in by the user, namely the number of neighbors to be considered reachable in a transition and the minimum value that the cosine must have for the transition to be possible. Indeed, it is not absurd to think that the transitions leading to a movement in the opposite direction of the velocity should be less considered.

Since a transition matrix stores probabilities, the sum of the values of a row must be equal to 1. Therefore, it is necessary to divide the value obtained for each transition by the sum of all the values of this row. As a reminder, each row of a transition matrix is associated with a starting state, in our case the transcriptomic profile of a cell. Each column corresponds to the state reached after a transition. The intersection between a row "i" and a column "j" gives the probability that a cell in state "i" goes to state "j" in one transition.

First model: $e^{\cos(\vec{x}, \vec{y})}$

The first model is simply to calculate the exponential of the cosine of the angle between the two vectors described above. The notion of distance is not considered in this situation. Calculating the exponential of the cosine gives a positive value no matter what the value of the cosine is. This is useful for calculating probabilities. Moreover, the exponential has the effect of favoring even more the transitions corresponding to high cosines which corresponds to angles of small amplitudes and therefore indicates a good alignment between the two cells and the velocity vector.

Second model: $e^{\cos(\vec{x}, \vec{y})} \cdot e^{-\text{dist}^2}$

The second model aims to improve the first model by taking into account the distance between the two cells considered. Taking the exponential of the opposite of the distance allows to strongly penalize cells located at large distances from the studied cell. To increase this penalty, the square of the distance is used instead of the simple distance. $e^{-\text{dist}^2}$ has a value between 0 and 1. Indeed, if the distance is equal to zero, the expression becomes equal to the exponential of 0 which is equal to 1. Since the limit of the exponential function equals 0 as x approaches $-\infty$, when the distance increases, the expression tends towards a value equal to 0. In conclusion, in the second model, the value obtained in the first model is multiplied by a factor whose value, which is between 0 and 1, is defined by the distance between the two cells concerned.

Third model: $\cos(\vec{x}, \vec{y}) \cdot \left(1 - \frac{\text{dist}^2}{\text{max}^2}\right)$

The objective of the third model is to correct the biases present in the second model. Since $e^{\cos(\vec{x}, \vec{y})} \cdot e^{-\text{dist}^2}$ can be simplified to $e^{\cos(\vec{x}, \vec{y}) - \text{dist}^2}$, the two variables do not have an equal impact on the final value. Indeed, the value of the cosine of an angle is between -1 and 1 and that of the opposite of the square of the distance is between $-\infty$ and 0. If dist^2 takes values much higher than $\cos(\vec{x}, \vec{y})$, then the weight from the angle will be very small on the final results and the calculated probabilities will depend essentially on the distance. With this model, we will systematically impose a threshold greater than or equal to 0 on the minimal value of the cosine.

As a result, the exponential is no longer necessary to ensure that the values of the cosine of the angles are positive. The variable "max" corresponds to the highest distance between two points (cells) in the database. Therefore, the result of $\frac{\text{dist}^2}{\text{max}^2}$ is between 0 and 1. Thus, subtracting this value from 1 also gives a value between 0 and 1. The closer the distance is to the maximum distance, the closer the result of the fraction will be to 1. The factor corresponding to the weight will then be close to 0 which will strongly decrease the transition probability.

Long term fate

Compute the transition matrix to a high power

In order to know the long-term fate of cells, one solution is to calculate the transition matrix at a very high power. The matrix will then indicate for each cell the probability of becoming another given cell in the long term. By grouping all the cells corresponding to the same cell type in one column, a transition matrix that indicates for each initial cell, the probabilities of becoming each cell type in the long term can be created.

The problem with this technique is that it requires computing very high powers of very large matrices. Indeed, the computation time needed to perform the matrix product is proportional to the cube of the number of rows of this matrix. Considering our original database, the matrix contains around 30 000 rows which leads to huge computation times. Therefore, repeating this calculation a very high number of times to obtain high power is not realistic with the available computers. In this work, metacells of size 10 are used, which leads to a reduction of the database size by a factor of 10. The computation times are thus strongly reduced and it is not impossible to carry out these calculations. However, for complete databases, the size can be a real limiting factor.

Exponentiation by squaring

There is a method that allows to calculate powers faster than multiplying the base by itself until reaching the requested power. This method is called exponentiation by squaring. Let's imagine that we want to calculate our matrix to the power of 1024. By proceeding in the classical way, it would be necessary to carry out 1023 matrix products to obtain the desired result. With the exponentiation by squaring, the principle is to compute a succession of squares. In our example, we compute A^2 , then A^4 (which is the square of A^2), A^8 and so on until A^{1024} , with "A" being the initial matrix. This way, 9 matrix products are needed to obtain the desired matrix.

This method is particularly efficient for calculating a number of powers corresponding to powers of 2. This method can also be used even if the number to be reached is not a power of two but it requires more operations.

Simulations

Even using fast exponentiation, the computation of the transition matrix at very high power can be extremely long. Since the objective of this project is to develop and test models, it is not realistic to take the time to perform these very long calculations for each combination of parameters tested. An alternative to these complex calculations is the use of simulations.

Indeed, it is faster to simulate, for example, 100 transitions from the initial matrix than to calculate the matrix raised to the power of 100. The principle of a simulation is to make the initial cells evolve towards their potential future transcriptomic states while respecting the transition probabilities provided by the transition matrix. Once all the new states have been determined, the graphs are generated (see the simulation graphs section below), the list of current cells is replaced by the cells corresponding to the new states and the same process is repeated using the same transition matrix until the required number of steps is reached.

A compromise between these two techniques can also be found by computing a small power of the matrix and then performing simulations from that matrix. Indeed, since most of the transitions are not allowed, the initial transition matrix contains many probabilities equal to 0. This greatly alleviates the computation of the matrix product for the first few matrix products until the number of remaining zeros becomes too small.

Reproducibility of simulations

In order to interpret the results of the simulations, it is necessary to take into account the variability that exists between each simulation. The origin of this variability comes from the use of random numbers to choose among the different possible cell fates. To do this, a large number of simulations are performed. The simulations are characterized by the number of steps to be simulated and by the number of transitions performed in one step. The results of a simulation at a given step correspond to the number of cells per cell type at that given step. For each step of the simulations, the results of each simulation are grouped together in order to calculate the mean and the standard deviation. With these two values and the number of simulations, we can calculate the 95% confidence interval on the mean of the values by using the following formula where \bar{x} is the mean, t is the Student's t critical value, σ is the standard deviation and n is the number of simulations.

$$\bar{x} \pm t \frac{\sigma}{\sqrt{(n)}}$$

The reproducibility of the simulations should be measured for each combination of parameters tested. Since we have a sample of 100 simulations, the value of t is equal to 1.984 for a 95% two-sided confidence interval. This value is readily available online¹²⁵.

Simulation graphs

During the simulation, 2 graphs are produced per step. The first type of graphs indicates for each cell type, its initial number of cells and the distribution of their new cell type at the considered step. In this work, this type of graph will be called the proportion graph. The second type of graph represents for each cell type, the quantity of cells having this cell type at the studied step of the simulation. Once all the steps of the simulation are completed, all the graphs of the same type can be grouped together to create a gif that represents the evolution of the cells during the simulation. Finally, a graph comparing the initial quantities with the quantities obtained with different experimental conditions can be created for a fixed model.

The whole program has been written in the Python programming language. The main libraries used are: scVelo, LoomPy, Joblib, Scikit-Learn, SciPy, Pandas, Numpy and Matplotlib. If you want to use the code, feel free to contact me at this address: D.Aguilar@student.uliege.be

CHAPTER 4

Results

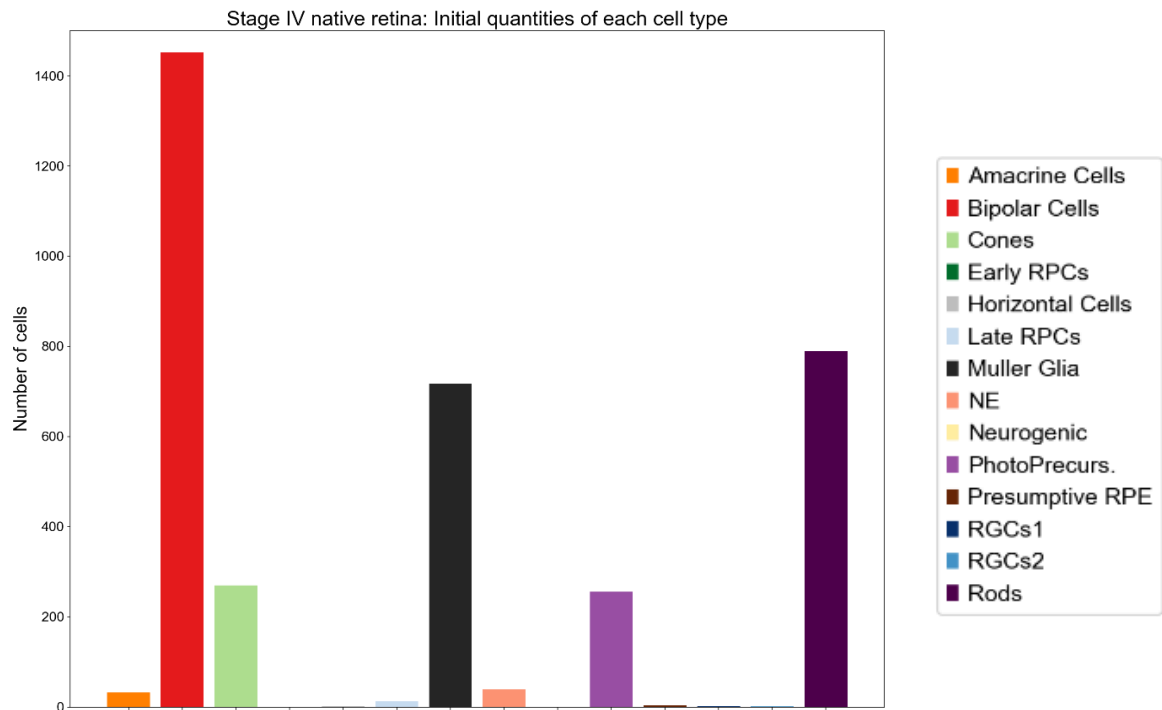


Figure 4.1: Graph showing the initial quantities of cells from native retina at the fourth stage of development by cell type. These are real cells and not metacells because in this case the original database is used.

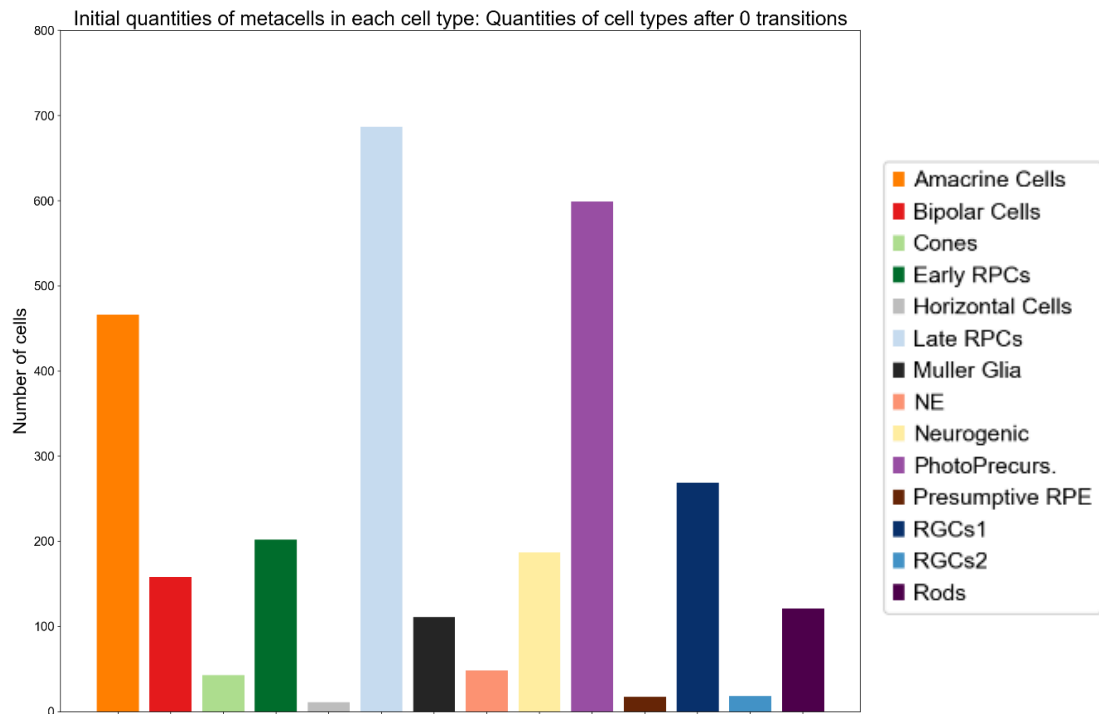


Figure 4.2: Graph showing the initial quantities of metacells per cell type in the database.

In this section, results of simulations of different models with different parameters will be shown. Before that, let's define an objective to reach during our simulations to be able to determine which model and which parameters are the best. The chosen objective corresponds to the proportions of each cell type at the last stage of development of our database. As mentioned in the introduction, retinal organoids tend among other things not to finish the differentiation and to lose some cell types. Therefore, only cells from native retinas will be considered to define the optimal proportions.

Among the 3576 cells from native retinas collected at the fourth stage, 32 are amacrine cells (0.9%), 1451 are bipolar cells (40.6%), 270 are cones (7.5%), none are ERPCs (0%), 2 are horizontal cells (0%), 13 are LRPCs (0.4%), 717 are Müller glia (20%), 39 are neuroepithelial cells (1.1%), none are neurogenic RPCs (0%), 256 are photoreceptor precursors (7.2%), 4 are presumptive RPE (0.1%), 2 are RGCs1 (0%), 1 is RGCs2 (0%) and finally 789 are rods (22.1%) (Figure 4.1).

By creating metacells of size 10, the new database contains 2937 metacells. Among these metacells, 466 are amacrine cells (15.9%), 158 are bipolar cells (5.4%), 43 are cones (1.5%), 202 are ERPCs (6.9%), 11 are horizontal cells (0.4%), 687 are LRPCs (23.4%), 111 are Müller glia (3.8%), 48 are neuroepithelial cells (1.6%), 187 are neurogenic RPCs (6.4%), 599 are photoreceptor precursors (20.4%), 17 are presumptive RPE (0.6%), 269 are RGCs1 (9.2%), 18 are RGCs2 (0.6%) and finally 121 are rods (4.1%) (Figure 4.2).

The goal is to see how well our models recapitulate differentiation trajectories. To do so, 100 simulations were generated per experimental condition. The following results correspond to the average of the results obtained at the end of these 100 simulations. The 95% confidence interval is represented on each quantity graph. Additional graphs are available in the appendices.

The starting points of the simulations correspond to the 2937 metacells. Therefore, metacells belonging to mature cell types are present from the beginning of the simulation. If our model effectively recapitulates the differentiation pathways, then the cells belonging to mature cell types should not disappear, they should keep their cell type while progenitor cells should reach a mature cell type. Therefore, this information must be taken into account when comparing the proportions obtained from our simulations with the proportions of cells in the fourth stage presented above. Indeed, if the model works well, one should not expect the initial amount of amacrine cells (15.9%) to decrease to the proportion corresponding to the fourth stage (0.9%). As a reminder, the mature cell types of the retina are: amacrine cells, bipolar cells, horizontal cells, Müller glia, presumptive RPE, RGCs1, RGCs2 and rods.

For each model, the impact of the number of neighbors considered is evaluated. For all the following graphs, the threshold imposed on the cosine value is equal to 0.15. This means that neighbors with which the angle is greater than about 80° are not considered. Since there can be neighbors on both sides of the velocity vector, the angle in which the neighbors are accepted is twice as large and is therefore equal to 160° . This ensures that the movement is globally in the direction of the velocity. The neighbor filtering based on the cosine value is done after determining the k nearest neighbors. The number of potential transitions is therefore not the same for each cell. The number of considered neighbors is equal to 10 in the first condition, to 20 in the second and to 30 in the third. The choice of these values is arbitrary. 500 transitions were made in each simulation. The results obtained are stable as of about 200 transitions depending on the number of neighbors considered. By taking 500 transitions, we consider that the metacells have finished differentiating. Finally, for ease of explanation and biological interpretation, we will refer to the metacells as cells during the analysis of graphs.

Effect of the number of neighbors considered on the average cell number after 500 transitions

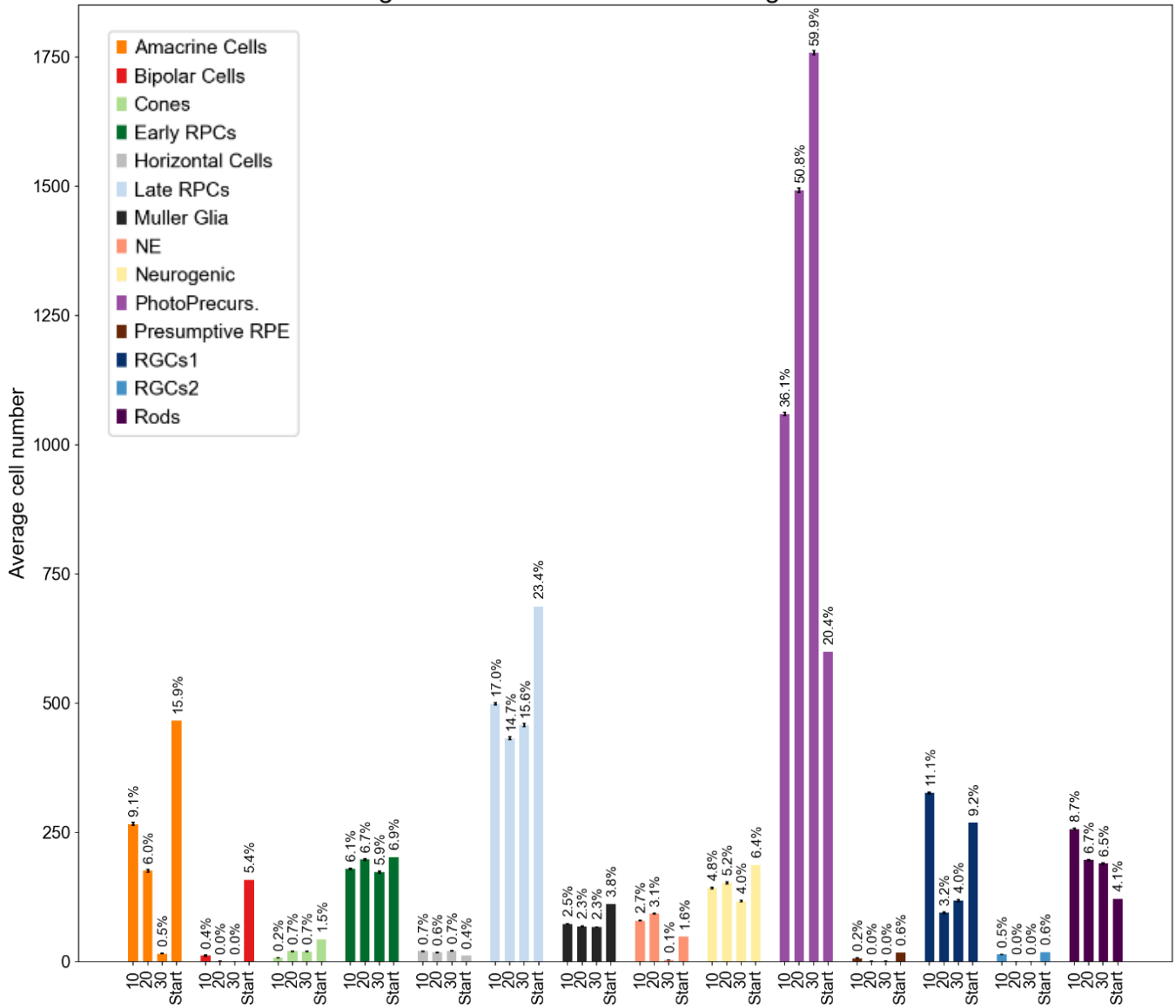


Figure 4.3: Graph comparing the average cell quantities per cell type at the end of 100 simulations (500 transitions) for the first model with 10, 20 and 30 neighbors. The initial quantities of each cell type are also indicated with the label "Start". The 95% confidence interval is indicated by the black bars.

First Model

First, let's observe under which experimental condition the quantity of progenitor cells decreases the most and the quantity of mature cells increases the most (Figure 4.3). In general, there is no obvious difference in the proportion of progenitor cells between the different conditions. Ideally, the cell types corresponding to progenitor cells should no longer contain cells. This is not the case for any of the three conditions. Only the neuroepithelial cells have completely differentiated in the third condition. In contrast to what is expected, the amount of certain progenitor cell types increased. This is the case for neuroepithelial cells in conditions 1 and 2 and for photoreceptor precursors in all three conditions. In addition, although the amount of cells in all the other progenitor cell types decreased, this decrease is small. Concerning the mature cell types, for all models a decrease in the number of amacrine cells, bipolar cells, cones, Müller glia, RGCs 2 and presumptive RPE is observed. For RGCs 1, the quantity slightly increased in condition 1 while it decreased in condition 2 and 3. The amount of rods increased in all conditions, especially in the first one. Similarly, the amount of horizontal cells also increased in all conditions.

Proportion graphs analysis

To better understand into which cell type the initial cells differentiate, proportion graphs are generated for each condition (Figure 4.4).

Mature cell types

Concerning the mature cell types, the more the number of neighbors increases, the less the cells tend to keep their original cell type. This is the case for bipolar cells, amacrine cells and RGCs2. The proportion of RGCs 1 to remain RGCs 1 decreases sharply when going from 10 to 20 neighbors but hardly changes between 20 and 30 neighbors. Similarly, the proportion of initial rods that are still rods at the end of the simulation decreases in the same way when considering 20 or 30 neighbors compared to 10. The majority of the cells belonging to the presumptive RPE change cell type in the first condition while in the second and third conditions, all of these cells change cell type. The proportion of initial horizontal cells that do not change cell type is close to 100% when considering 10 neighbors and decreases slightly when considering more neighbors. Only for cones does going from 10 to 20 or 30 neighbors increase the proportion of initial cones that keep their cell type. This proportion is lower for 30 neighbors compared to 20 neighbors. Finally, the proportion of Müller glia that do not change cell type is stable when varying the number of neighbors considered.

By looking at the cells that do not keep their original cell type, it can be observed that the photoreceptor precursor type is a common fate for a large part of the cells of almost all mature cell types. Among the three conditions, the only mature cell types not to produce a significant amount of precursors of photoreceptors are RGCs 1 and 2 in the first condition, presumptive RPE in conditions 1 and 2 and horizontal cells, especially in the first condition. In addition, it can also be observed that the proportion of cells that become photoreceptor precursors increases in every mature cell type when the number of neighbors considered increases. Moreover, under all three conditions a significant fraction of cones and bipolar cells tend to become rods while no rods become bipolar cells and almost none become cones. Besides, bipolar cells also generate some cones. In addition to photoreceptor precursors: amacrine cells generate mainly RGCs 1 and late RPCs in all three conditions.

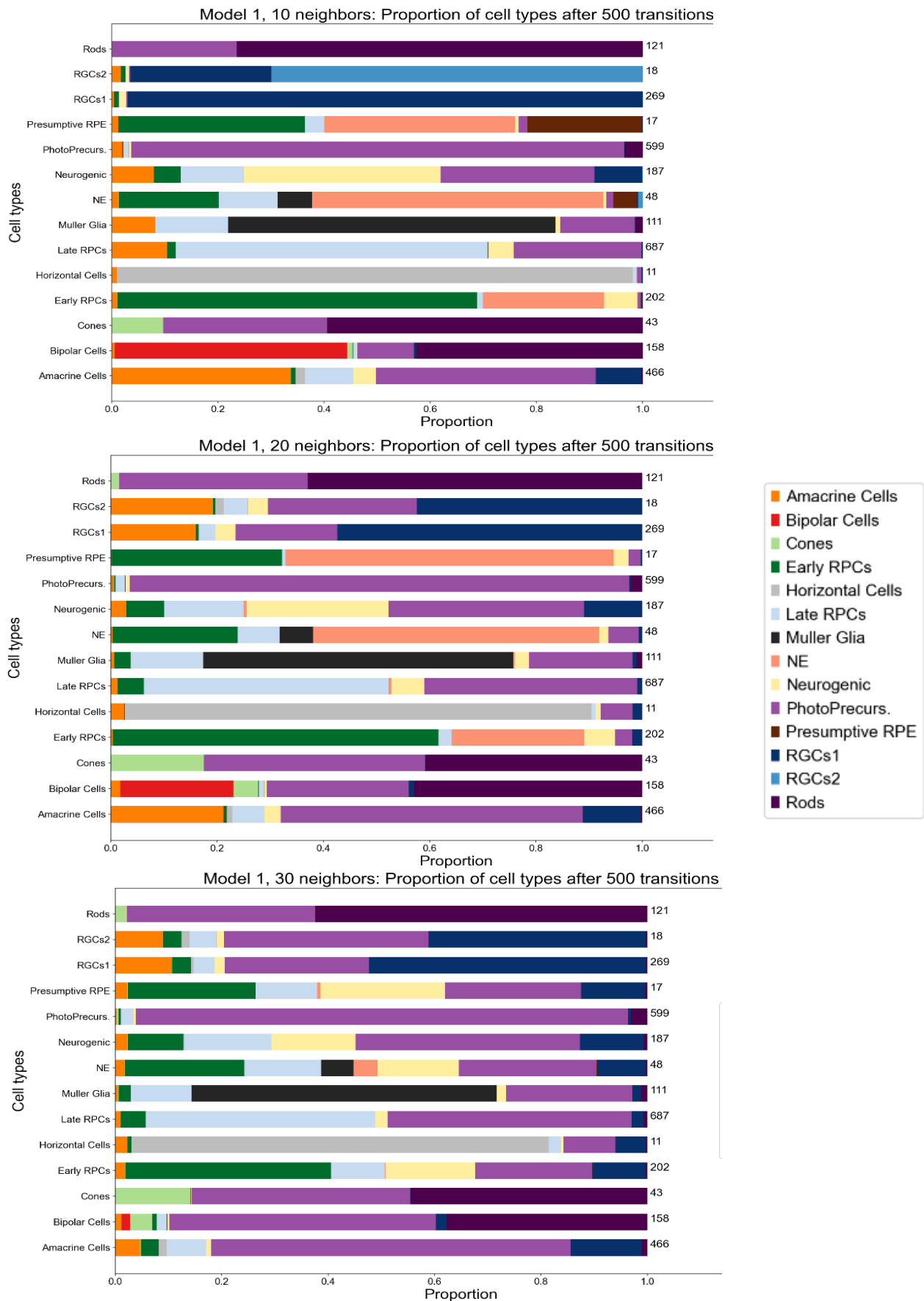


Figure 4.4: Set of three proportion graphs for the first model, each associated with the number of neighbors considered. Each graph indicates for each cell type, its initial number of cells and the distribution of their new cell type at the end of the simulations (500 transitions). The proportions shown are the average proportions obtained from 100 simulations.

Horizontal cells generate mainly RGCs 1 and amacrine cells in the second and third conditions. Müller glia generate mainly late RPCs in all three conditions, amacrine cells in the first condition and early RPCs in the second and third conditions. Presumptive RPE generate mainly early RPCs in all three conditions, neuroepithelial cells in the first and second conditions and late RPCs and neurogenic cells in the third condition. RGCs 1 generate mainly amacrine cells and late RPCs in the second and third conditions, neurogenic cells in the second condition and early RPCs in the third condition. RGCs 2 generate mainly RGCs 1 in all three conditions and amacrine cells and late RPCs in the second and third conditions.

Progenitor cell types

For progenitor cells, a considerable proportion of the initial cells does not differentiate. This is the case for all progenitors except neuroepithelial cells in the third condition. (1) This non-differentiation is particularly important for photoreceptor precursors in all three conditions. For the part that differentiates, let's see if differentiation has a biological meaning. (2) First, in the first and second condition, the early RPCs that differentiate become mostly neuroepithelial cells which is the opposite of what happens during retinal development. In the third condition, the fraction that differentiates is more important and the cell types obtained (late RPCs, neurogenic cells, photoreceptor precursors, RGCs 1 and amacrine cells) are coherent with the biology of the retina. (3) Concerning the late RPCs, only a very small portion "de-differentiates" into early RPCs. The other cell types (amacrine cells, neurogenic cells and photoreceptor precursors) formed are consistent with the biology of the retina. However, the amount of amacrine cells generated decreases in condition 2 and 3 compared to the first condition. (4) The neuroepithelial cells being the original precursor cells of the retina, they give birth to the other progenitor cells. This is the case in the three conditions and especially in the third one in which there are no neuroepithelial cells left at the end of the simulation. (5) Finally, the initial neurogenic RPCs differentiate into photoreceptor precursors and RGCs1 under all three conditions. The greater the number of neighbors considered, the greater the proportion of photoreceptor precursors generated and the lower the proportion of neurogenic cells that retain their cell type. A small proportion also differentiates into amacrine cells, especially in the first condition. Nevertheless, in all three conditions, part of the neurogenic cells "de-differentiate" into early and late RPCs.

Conclusion

In the 3 conditions, no cell type gives rise to bipolar cells, apart from the initial bipolar cells. This is also the case for the cones in the first condition. In the other two conditions, some cones are generated from bipolar cells and rods. The same observation can be made for horizontal cells in all three conditions, with only minimal production from amacrine cells. Similarly, RGCs2 are not produced in any of the three conditions. The formation of rods from photoreceptor precursors is minimal, with the majority coming from mature cell types, namely bipolar cells and cones. Amacrine cells are formed from both progenitor and mature cell types. In the first condition, the production of amacrine cells comes mainly from late RPCs and neurogenic cells, which is in agreement with the biology. However, in conditions 2 and 3, RGCs 1 and 2 have the highest proportions of amacrine cell formation. Contrary to what is expected, Müller glia are not formed from late RPCs but only from a small proportion of neuroepithelial cells in all three conditions. Presumptive RPE are generated by a small fraction of neuroepithelial cells only in the first condition.

Effect of the number of neighbors considered on the average cell number after 500 transitions

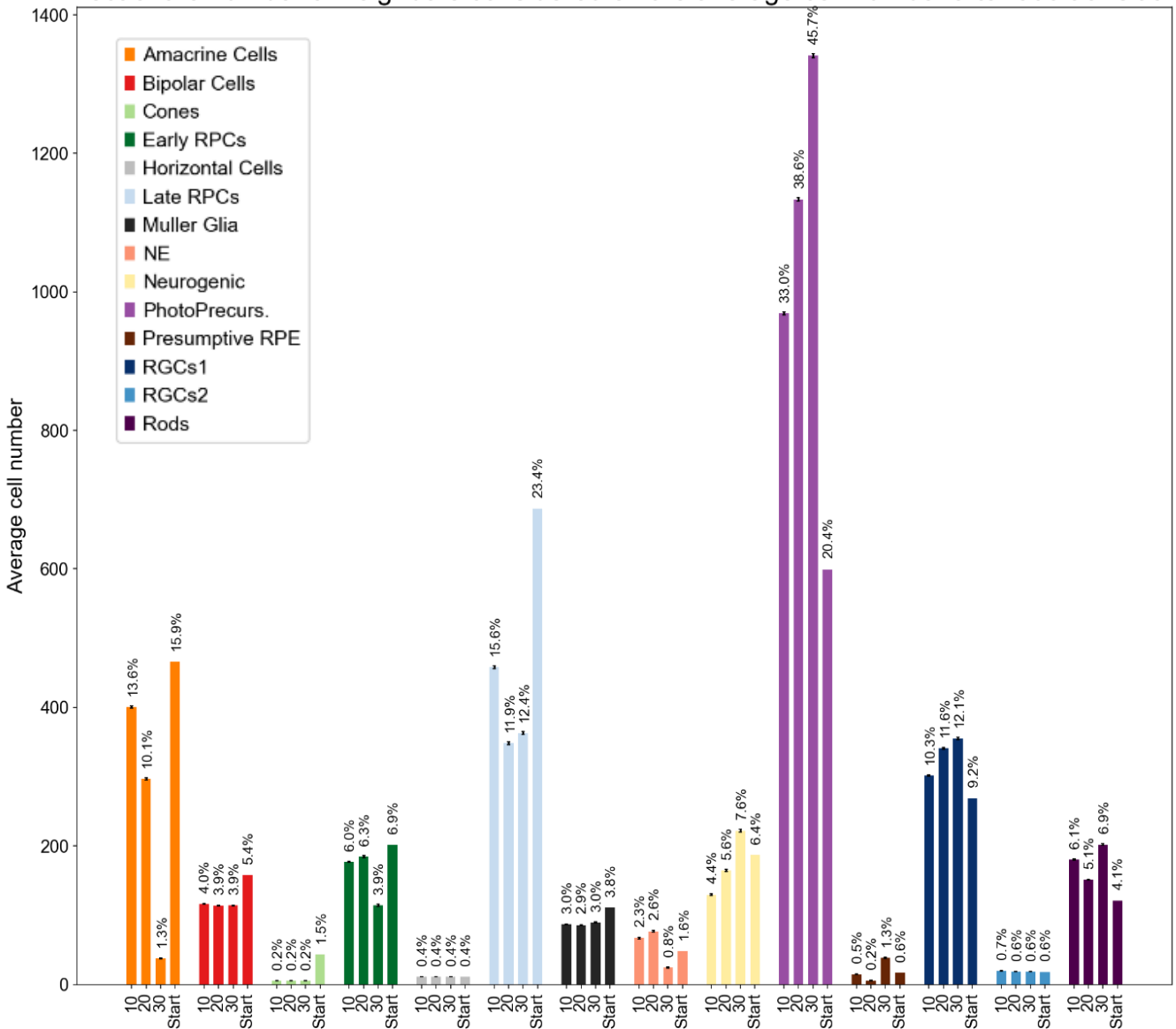


Figure 4.5: Graph comparing the average cell quantities per cell type at the end of 100 simulations (500 transitions) for the second model with 10, 20 and 30 neighbors. The initial quantities of each cell type are also indicated with the label "Start". The 95% confidence interval is indicated by the black bars.

Second Model

As for the first model, let's observe under which experimental condition the quantity of progenitor cells decreases the most and the quantity of mature cells increases the most (Figure 4.5). Although the amount of progenitor cells decreases for some cell types, there are still cells in each progenitor cell type at the end of the simulations for each condition. For early RPCs, a clear decrease in the third condition can be seen compared to the other two conditions where the final quantity is close to the initial quantity. For late RPCs, a decrease in cell number is observed in all three conditions, especially for conditions 2 and 3. In contrast, for neuroepithelial cells, neurogenic RPCs and photoreceptor precursors, the final amount of cells is greater than the initial amount for at least one condition. For the other conditions, the decrease is small. Concerning the mature cell types, we observe for all conditions, a decrease in the number of amacrine cells, bipolar cells, cones and Müller glia. Increasing the number of neighbors drastically decreases the final amount of amacrine cells while the amounts of bipolar cells, cones and Müller glia do not vary. The amount of rods and RGCs 1 increased in all conditions, especially in the third one. In the same way as the horizontal cells, the amount of RGCs 2 neither increases nor decreases.

Proportion graphs analysis

To better understand into which cell type the initial cells differentiate, proportion graphs are generated for each condition (Figure 4.6).

Mature cell types

For RGCs 1 and 2, Müller glia, horizontal cells, and bipolar cells, increasing the number of neighbors has almost no influence on the proportion of cells that retain their original cell type. These proportions are close to 100% for horizontal cells and RGCs 1 and 2, 75% for bipolar cells and 70% for Müller glia. For rods, the proportion of initial rods that are still rods at the end of the simulations varies with the number of neighbors. This proportion is highest in the third condition, followed by the first condition and finally by the second condition. Concerning the initial presumptive RPE, the proportion of these cells that do not change cell type is similar between the first and third condition. This proportion is higher than the proportion obtained in the second condition. The proportion of cones that are still cones at the end is the same in each condition. For amacrine cells, the proportion still being amacrine at the end of the simulations is slightly higher in the first condition than in the second. In the third condition, this proportion decreases strongly.

About the other generated cell types, the proportions generated are unchanged in all three model for RGCs 1, RGCs 2, Müller glia, horizontal cells and bipolar cells. RGCs 1, RGCs 2 and horizontal cells hardly produce any other cell types. Müller glia differentiate into late RPCs and photoreceptor precursors. Bipolar cells generate mostly rods and some photoreceptor precursors. Regarding rods, only photoreceptor precursors are generated. The lowest proportion corresponds to the third condition and the highest to the second condition. Regarding presumptive RPE, the remaining cells differentiate mainly into early RPCs and neuroepithelial cells in the first and second conditions and into neuroepithelial cells and neurogenic RPCs in the third condition. For cones, the other cell types generated are photoreceptor precursors and rods. Depending on the condition, the proportion corresponding to each type changes. The third condition is the one that produces the most rods

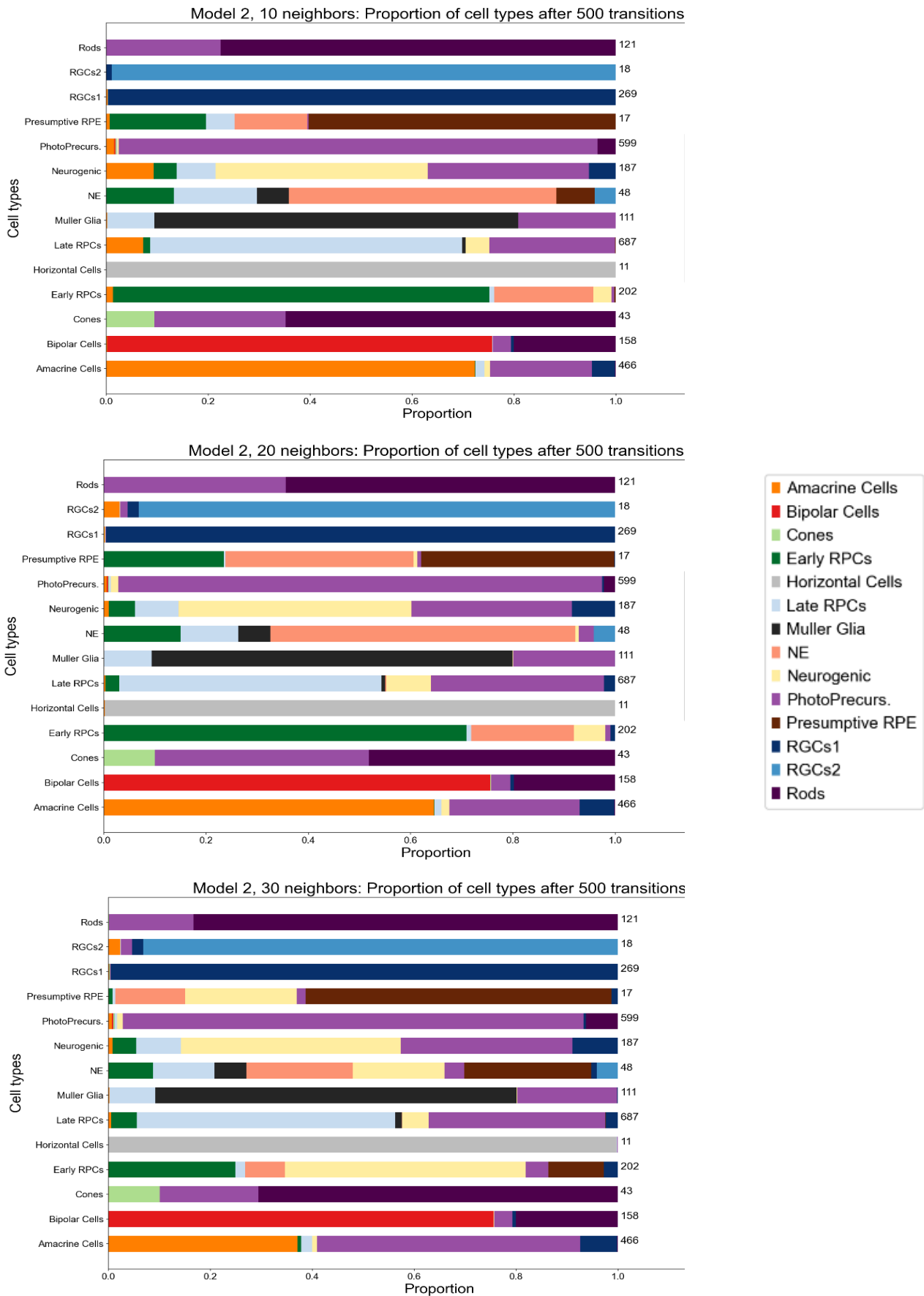


Figure 4.6: Set of three proportion graphs for the second model, each associated with the number of neighbors considered. Each graph indicates for each cell type, its initial number of cells and the distribution of their new cell type at the end of the simulations (500 transitions). The proportions shown are the average proportions obtained from 100 simulations.

and therefore, the least photoreceptor precursors while the second condition is the one that produces the most photoreceptor precursors and therefore the least rods. Finally, the other cell types generated from amacrine cells in all three conditions are mainly photoreceptor precursors and, to a lesser extent, RGCs 1.

Progenitor cell types

(1) Under all three conditions, almost all photoreceptor precursors fail to differentiate. (2) For neurogenic RPCs, more than half of the cells differentiate. The cell types generated are the same in all three conditions (photoreceptor precursors, RGCs 1, early RPCs and late RPCs) and are produced in the same proportions, with the exception of amacrine cells which are significantly produced only in the first condition. Since neurogenic cells are derived from both types of RPCs in the retina, generating early and late RPCs is contrary to what happens biologically. (3) Regarding neuroepithelial cells, more than half of the initial cells does not change cell type in the first and second conditions, while in the third condition, this proportion decreases to 20% of the initial cells. In all three conditions, early and late RPCs are generated as well as RGCs 2 and Müller glia in similar proportions across the 3 conditions. Presumptive RPE is generated in conditions 1 and 3 with an approximately three times higher proportion in condition 3 while some photoreceptor precursors are generated in conditions 2 and 3. Finally, neurogenic cells are produced only in condition 3. (4) The proportion of late RPCs not differentiating is greater than 50% in all three conditions, and is slightly higher in the first condition than in the other two. The cells that differentiate generate a high proportion of photoreceptor precursors, a low proportion of neurogenic cells and an even lower proportion of Müller glia in all three conditions. Some early RPCs are also generated from the initial cells. The proportion generated increases with the number of neighbors considered. Finally, some RGCs 1 are generated in conditions 2 and 3 while amacrine cells are generated in condition 1. (5) Concerning early RPCs, more than 70% of the initial cells do not differentiate in the first and second condition. In the third condition, this proportion is close to 25%. The other cell types produced are the same in the first two conditions, namely a majority of neuroepithelial cells and to a lesser extent, neurogenic RPCs. In the third condition, the majority of the cells become neurogenic RPCs, a small fraction becomes neuroepithelial cells, another fraction becomes photoreceptor precursors, another fraction becomes RGCs1 and finally a larger fraction becomes presumptive RPE.

Conclusion

There is still no production of bipolar cells, horizontal cells and RGCs 2 in the different conditions. Moreover, no condition generates cones. However, the majority of cells belonging to the mature cell types do not change cell type, with the exception of cones and amacrine cells. Also, those that change cell type mainly take the cell type of the closest progenitor and not a mature cell type, except for cones and bipolar cells that can become rods. Conversely, the condition for which progenitors change cell type the most is the third, especially in early RPCs and neuroepithelial cells. In general, progenitor cells tend to produce predominantly cell types consistent with those produced biologically in the retina, although some "de-differentiations" are observed, especially in early RPCs in the first two conditions. Although increasing the number of neighbors has a positive effect on the progenitor cells, it also has the disadvantage of strongly decreasing the amount of amacrine cells. Finally, late RPCs produce Müller glia but the proportion is very low.

Effect of the number of neighbors considered on the average cell number after 500 transitions

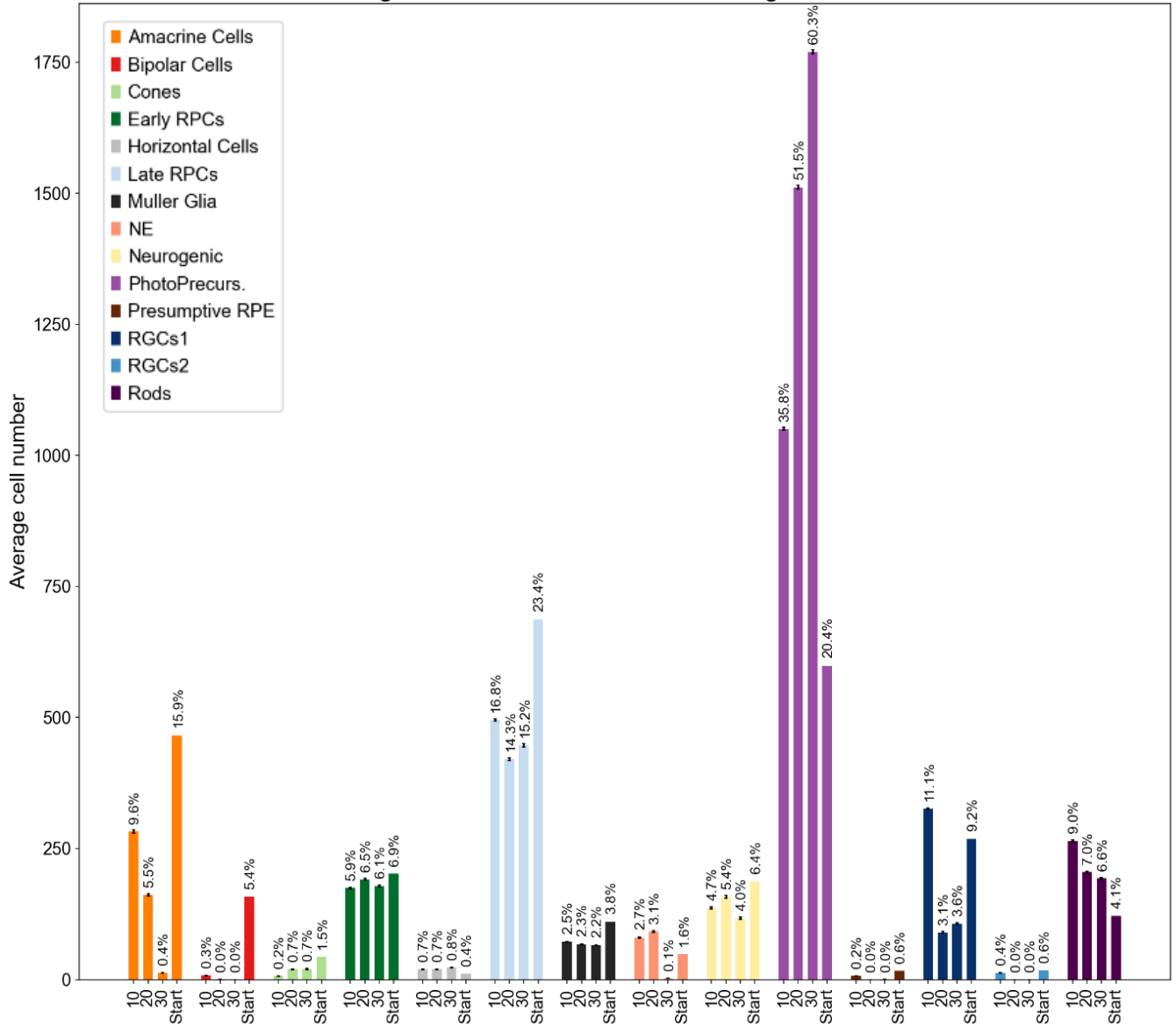


Figure 4.7: Graph comparing the average cell quantities per cell type at the end of 100 simulations (500 transitions) for the third model with 10, 20 and 30 neighbors. The initial quantities of each cell type are also indicated with the label "Start". The 95% confidence interval is indicated by the black bars.

Third Model

The graphs obtained with the third model (Figure 4.7, Figure 4.9) are the same as those obtained with the first model (Figure 4.3, Figure 4.4). This is the case for both the quantity and the proportion graphs. Therefore, the analyses of the results of the first model are also valid for the third model. The origin of this similarity is addressed in the discussion section.

2D UMAP analysis

In order to improve the understanding of the movements that take place during the simulations, a 2-dimensional UMAP representation of the database corresponding to the metacells is generated (Figure 4.8). Contrary to the first UMAP representation of the original database (Figure 3.5), only the genes that have passed the second filter are considered. With a few exceptions, the bipolar cells are all grouped in a region (blue frame) disconnected from the rest of the database. This region also contains an important part of the rods. There are also cones and a small fraction of total photoreceptor precursors. Another region separated from the main mass is that containing Müller glia (green circle). It is interesting to highlight the fact that in order to reach the amacrine cells, the progenitor cells must first pass through the photoreceptor precursor state. Similarly, a direct link between progenitor populations and horizontal cells is not present. Furthermore, passing through the neurogenic stage before reaching mature cell types does not seem to be necessary in late RPCs.

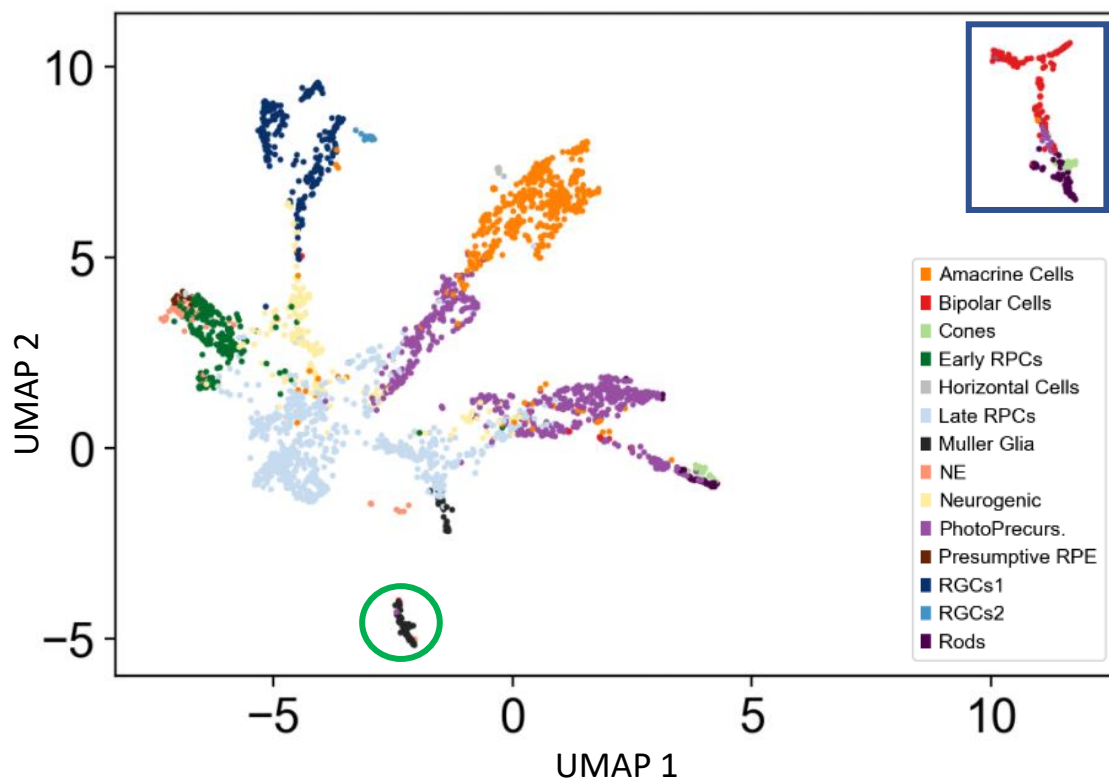


Figure 4.8: 2D UMAP manifold showing the arrangement of the 2937 metacells after the second gene filter is applied. Two groups separated from the main population are highlighted (blue frame and green circle).

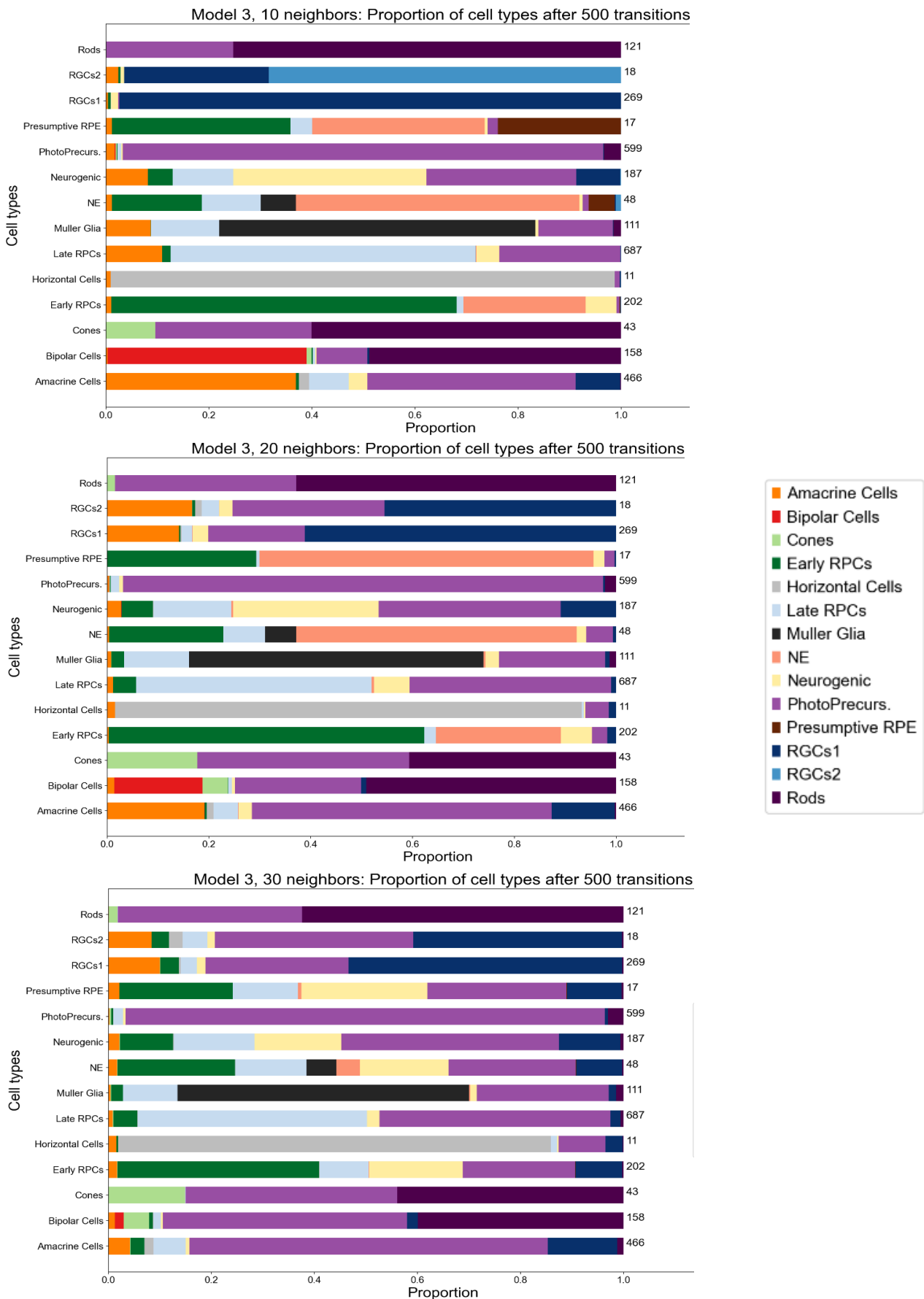


Figure 4.9: Set of three proportion graphs for the third model, each associated with the number of neighbors considered. Each graph indicates for each cell type, its initial number of cells and the distribution of their new cell type at the end of the simulations (500 transitions). The proportions shown are the average proportions obtained from 100 simulations.

CHAPTER 5

Discussion

Similarities between the first and the third model

The fact that models 1 and 3 produce the same results is unexpected. As is often the case in programming, the first reflex was to look for an error in the code responsible for the problem. Nevertheless, after several checks, no errors could be found. Assuming that the results obtained are not the consequence of an error, the formulas of the models have been compared to see under which conditions they give the same results.

$$\text{First model: } e^{\cos(\vec{x}, \vec{y})}$$

$$\text{Third model: } \cos(\vec{x}, \vec{y}) \cdot \left(1 - \frac{\text{dist}^2}{\text{max}^2}\right)$$

The first model takes into account only the cosine of the angle while the third model assigns a weight to the cosine of the angle and a weight to the distance. For the models to produce the same results, they must take the same variables into account during the calculation. For this to be true, the weight of the distance must be insignificant in the third model and therefore the term $\left(1 - \frac{\text{dist}^2}{\text{max}^2}\right)$ must be close to or equal to 1. It is important to specify that in the first model, the exponential of the cosine is used, while in the third model the value of the cosine itself is used. Therefore, the two terms are not equal. However, this does not prevent them from producing the similar results.

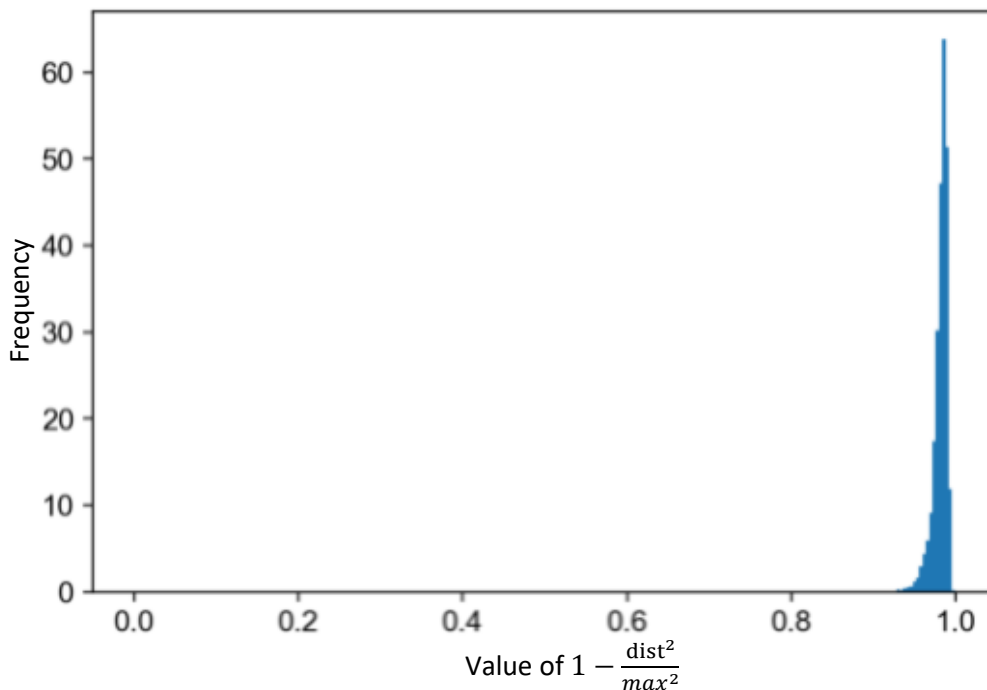


Figure 5.1: Graph illustrating the distribution of values taken by the term $\left(1 - \frac{\text{dist}^2}{\text{max}^2}\right)$ in the metacell database considering 10 neighbors

The distribution graph of $\left(1 - \frac{\text{dist}^2}{\text{max}^2}\right)$ (Figure 5.1) shows that the values of $\left(1 - \frac{\text{dist}^2}{\text{max}^2}\right)$ are close to 1 and therefore that $\frac{\text{dist}^2}{\text{max}^2}$ is close to 0. This means that for all considered neighbors of all cells, max^2 is much larger than dist^2 . However, the difference between max^2 and dist^2 depends on the database. In conclusion, this model needs to be redesigned in order to assign a better weight to the distance.

Model comparison

By comparing the results of this second model with those of the first model, some important points can be highlighted. First, although the decrease in amacrine cells is common to both models, a greater proportion of amacrine cells is retained in the second model. The same can be said for Müller glia. Similarly, the decrease of bipolar cells is much lower in the second model which allows to keep 80% of the initial cells whereas they almost all disappeared in the first model. The amount of RGCs 1 increases in the three conditions of the second model while in the first model, the amount decreased strongly for the second and third condition. Moreover, RGCs 2 do not disappear in the second model, unlike the first model. Finally, the decrease in the amount of precursor cells is more important in the second model, except for neurogenic RPCs where the amount is higher than the initial quantity in the third condition. It is important to specify that even if the final cell type of a cell is the same as its initial cell type does not necessarily mean that this cell does not move. To conclude, the fact that photoreceptor precursors fail to differentiate, regardless of the model and the number of neighbors, is very interesting.

Regarding the consistency of the results with the biology of the retina, the second model outperforms the first in its ability to generate Müller glia from late RPCs and presumptive RPE from neuroepithelial cells. The proportion of photoreceptor precursors that does not differentiate is the same for both models. However, the final amount of photoreceptor precursors is higher in the first model. This difference comes from the fact that in the first model, some mature cell types differentiate into photoreceptor precursors which is not desired. One of the reasons why there are more neurogenic RPCs in the second model is that a greater proportion of the initial neurogenic RPCs do not differentiate. However, in the first model, a larger fraction of the neurogenic cells de-differentiate to become early and late RPCs.

Based on these observations, the second model seems to better recapitulate the different trajectories of retinal differentiation. Nevertheless, the results are not yet convincing, mainly because of the lack of photoreceptor and bipolar cell formation and the low production of Müller cells. Indeed, these cell types (rods, cones, Müller glia and bipolar cells) are the most present in the cell population corresponding to stage 4 (Figure 4.1). Moreover, the second model is biased. Indeed, the weight assigned to the distance can be much higher than the weight assigned to the cosine of the angle. Since this model is the one that gives the best results, it is not impossible that in our case, only considering the distances could give better results than using the RNA velocity. Nonetheless, the RNA velocity is still used to filter out the neighbors behind the cell thanks to the threshold set on the cosine value.

For this model, it is not easy to decide which condition gives the best results. Indeed, considering 30 neighbors, the quantity of cells is higher in the following mature cell types: rods, RGCs1 and presumptive RPE. Moreover, the progenitor quantity is lowest for early RGCs and neuroepithelial cells and is almost as low as condition 2 for late RPCs. However, condition 3 is also the one in which the majority of amacrine cells disappear. These cells become mainly photoreceptor precursors. Since the increase of photoreceptor precursors over the other conditions does not come from the differentiation of progenitor cells, the third condition cannot be considered better at generating photoreceptor precursors than the others.

Changing the number of neighbors therefore does not have an entirely positive or negative effect. A trade-off must be made based on the effects on each cell type.

Finally, the quality of the results does not depend solely on the models developed and the parameters used. Indeed, other factors that influence the results will be discussed in the sections below.

UMAP graph interpretation

The analysis of the UMAP graph is very interesting because it allows us to obtain additional information to that obtained from the proportion graphs. First, it helps to understand why the photoreceptor precursors do not differentiate, regardless of the model. Indeed, we can see that the group of cells containing bipolar cells, rods and cones is located very far from the photoreceptor precursors, thus making them inaccessible. This inaccessibility comes from the fact that the distance makes that the cells belonging to the distant cluster will never be part of the “n” nearest neighbors of the photoreceptor precursors. Second, the UMAP graph can be used to assess whether the selected genes correctly represent the biology of the retina. Since the biological differentiation pathways of the retina are known, it is possible to verify whether the different cell types are arranged consistently in our database.

However, the observations made should be interpreted with caution. Indeed, going from 76 dimensions to 2 necessarily leads to a loss of information. This is why it is not impossible that some connections between cell types exist in the database but are not visible on the UMAP.

Identification of spliced and unspliced mRNAs

One potential source of problems is the way the 10X methodology defines spliced and unspliced mRNAs. Indeed, using the Chromium Single Cell 3' Reagent Kit v2.0, the sequencing libraries generated produce reads of the following size: 26 base pairs for Read 1, 8 base pairs for the i7 index and 98 base pairs for Read 2 which corresponds to the insert¹²⁶. In mice, only 33.3% of exons are less than 100 base pairs long¹²⁷ (Figure 5.2). Therefore, for two thirds of the exons, Read 2 is not long enough to cover an exon/intron junction and thus splicing information cannot be obtained. It may be interesting to see how many reads are in this situation in our database. Indeed, since only reads categorized as spliced or unspliced are used in the velocity calculation, it is possible that there is a bias excluding genes whose exon at the 3' end is too long to determine splicing.

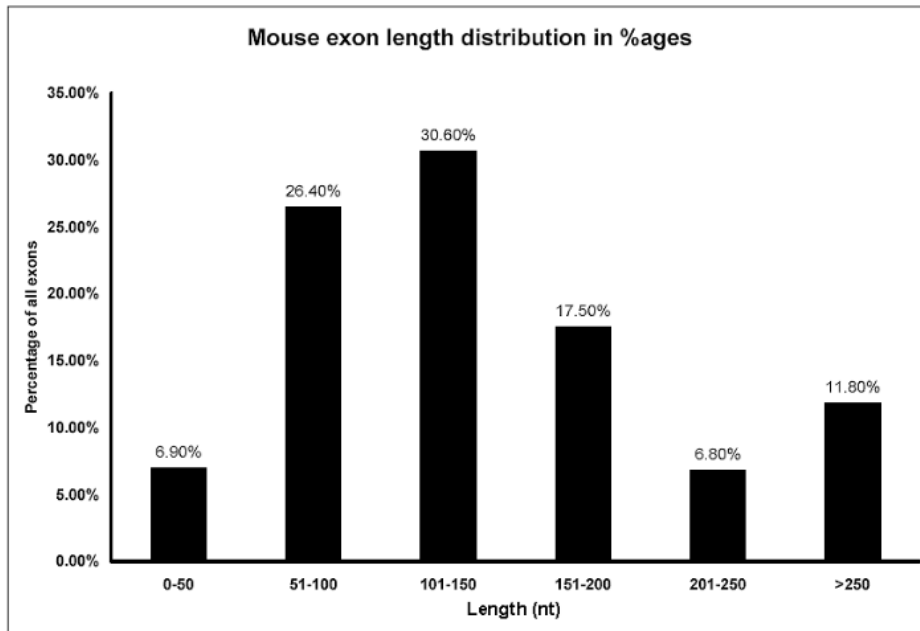


Figure 5.2: *Percentage distributions for exon length for mouse genome.* Figure and explanations from "An analysis on gene architecture in human and mouse genomes" figure 4c by Sakharkar et al., 2005¹²⁷.

Cellular Death

In our programs, cellular death is not modeled. Indeed, during a simulation, the final number of cells is equal to the initial number of cells. The effect of this programmed cell death is indirectly and partially taken into account when sequencing cells at a stage after apoptosis. Indeed, only cells that have survived this process are sequenced. However, all cells collected before the apoptosis stage are sequenced and therefore, those destined to die cannot reach their final stage since it is not present in the database.

This could be a reason why some of the progenitor cells do not differentiate into final cell types. Their velocity would make them go in a direction corresponding to the dead cells that are absent from our database and they would therefore be blocked.

Lack of cells

This section describes two different types of lack of cells. The first is simply a consequence of the metacell smoothing method. By creating metacells of size 10, one necessarily divides the number of cells of the dataset by 10. With fewer cells and thus fewer states to transit to, it is possible that the differentiation pathways can no longer be recreated via a succession of small transitions. Especially since decreasing the number of cells via the creation of metacells leads to an increase in the distance between the new metacells.

The other type of lack of cells corresponds to the absence in the initial database of cells allowing the transition between cells harvested at different stages of development. If some areas are inaccessible, it may be due to an excessive time gap between sampling steps. Indeed, since our model relies on a succession of small jumps to recapitulate the total movement, if too much time separates the different samples, then the transcriptomic profiles of cells are too different from one sample to another. This results in a large distance in the multidimensional space. This large distance acts on two levels. First, since the cells of different samples are far away, there is less chance that they are included in the "n" nearest neighbors of cells of other samples. Moreover, since models 2 and 3 penalize transitions according to distance, the probability of reaching these areas is low. Based on the results obtained, this could be the case for the fourth stage of development in our database.

Use of multiple databases

Since the goal of this project is to produce a tool that works not only on mouse retinal data but also on data from any other organ and any organism, it is important not to test all possible parameters on a case-by-case basis. Indeed, in addition to the considerable time that this would take, it would lead to what is called overfitting. The way our final model works would then be too much influenced by our initial data, it will have been tailor-made for this database and will certainly produce bad results when applied to another database. To avoid this problem, the 3 models can be tested on several different databases to see how they behave on each one and to identify the important parameters.

In addition to avoiding overfitting, using multiple databases has another advantage. If ever the current database on which we develop the tools in one way or another is not compatible with the theoretical model under development, the results obtained will never be satisfactory. By testing our model on several databases, it is possible to determine whether the problems encountered are systematic or only related to one database. For instance, if the sampling steps are too far apart in time, the cells will get stuck and the model will not work on that database.

Calculation of RNA velocity

Since RNA velocity is at the heart of our model, if it is not a faithful reflection of reality, the trajectories derived from these velocities will not correspond to real biological trajectories. In this section, the assumptions necessary to calculate RNA velocity are discussed.

A first assumption is that the transcription rate (α) of a gene is constant over time. The gene is therefore either transcribed or switched off. This assumption is not necessarily true. Depending on the needs of the cell, a gene may be transcribed at a certain intensity. In response to a specific signal, the intensity of transcription may increase. The intensity of transcription is in fact proportional to the number of RNA polymerases that generate mRNAs simultaneously¹²⁸.

Another assumption made is that the steady state is systematically reached. However, some genes can be transcribed over a very short time. Therefore, the steady state is not reached and consequently, the inferred degradation rate (γ) is biased. Since RNA velocity is calculated from the degradation rate, this has a direct impact on the quality of the velocity.

Finally, the assumption that the splicing rate (β) is the same for all genes is problematic. Indeed, splicing can begin before transcription is complete. This phenomenon is called co-transcriptional splicing. Although this process is common, it is not systematic¹²⁹. Genes for which splicing begins before the end of transcription are therefore spliced more rapidly than genes for which splicing begins only after transcription.

Low transcript capture per cell

The proportion of mRNA transcripts captured per cell during the single cell RNA sequencing depends on the Single Cell 3' reagent chemistry (v1, v2, or v3) used. The yield of v2, which was used to obtain our database, is about 15% against about 8% for v1 and 32% for v3¹³⁰. This means that only 15% of the total information is used to characterize the transcriptomic state of the cells. If this 15% is representative of the total mRNAs, the fact that the capture is incomplete is not a real problem.

However, if for any reason this 15% is not representative, then incomplete capture becomes a serious bias to consider. If genes involved in the differentiation process have low expression levels and are not efficiently captured, the information associated with these genes may be lost for a certain proportion of cells. Nevertheless, the majority of the transcriptome can still be recovered because multiple cells from the same population are sequenced, each having captured a different set of the total transcriptome¹³⁰. Quantifying the original quantities from these cells remains a complex task.

Selected genes

The choice of genes is essential for the proper functioning of the program developed, regardless of the model considered. As the distance between two cells depends on the space in which they are projected, it is important that the space chosen be a faithful reflection of biological reality. Since each dimension of this space corresponds to a gene, filtering genes can have a significant impact on the distance between cells.

Because the second gene filtering is based on the alignment of velocities, the existence of biases in the calculation of velocities may influence the genes chosen. It is possible that biologically relevant genes are rejected and conversely that genes that are not involved in development are retained. As explained in the "UMAP graph interpretation" section, visualization of the data in 2 or 3 dimensions can help determine if the set of genes studied is relevant.

Conclusion

The democratization of single cell RNA sequencing has led to the production of a large amount of single cell data by different laboratories around the world. Among these databases, some have been produced to study the developmental process of a tissue, an organ or even an entire organism. From these data, the objective of this work was to create a model recapitulating the differentiation trajectories present in the data by using the concepts of RNA velocity and Markov chains. However, this goal was not properly achieved. As explained above, there are several possible reasons for this failure. It is then necessary to try to correct these errors one by one until the model is functionally adept.

Outlook

First, the hypothesis that there are not enough cells in the database can be tested by generating smaller metacells or by using the original cells without making metacells. If this modification does not improve the results, then it is possible that the samples (stage I, II, III and IV) are too far apart in time, especially the fourth sample. To remedy this, the models could be tested on molecular atlases. It is already planned to test the models on the transcriptomic atlas of the *C. elegans* embryogenesis¹³¹.

If modifying the database still does not have a positive effect on the results, it may be worthwhile to test the operation of the models on tailor-made data to verify that the models work theoretically. If the models work theoretically but not on the different databases provided, then the problem may come from the techniques used to obtain these databases or from the calculation of RNA velocities.

A new method for calculating RNA velocities has been published in 2020 by Bergen et al.¹¹⁴. This one is based on other assumptions and would be worth to be tested to see if the new velocities induce better movements. If the results are still not convincing, then the methods used to obtain the databases may not be suitable for using our models. For example, by using the v3 kit, the proportion of the transcriptome captured per cell would increase from 15% to 32%, which is more than double. In addition, SMART-Seq sequencing enables the obtaining of the complete sequence of mRNAs which improves the distinction between spliced and unspliced. The impact of the assumptions made when calculating the RNA velocities could also be evaluated to see how much it influences the velocities obtained. Besides, the method for gene filtering developed in this work is not necessarily the best and can still be improved.

Finally, in order to better evaluate our models, a third type of graph could be generated. This one would represent the position of the cells on a UMAP graph at each step of a simulation. A gif showing the movement of the cells during the simulation could then be produced. This would allow to identify the regions where the cells get blocked. This would allow us to study the behavior of velocities in these regions.

CHAPTER 6

Bibliography

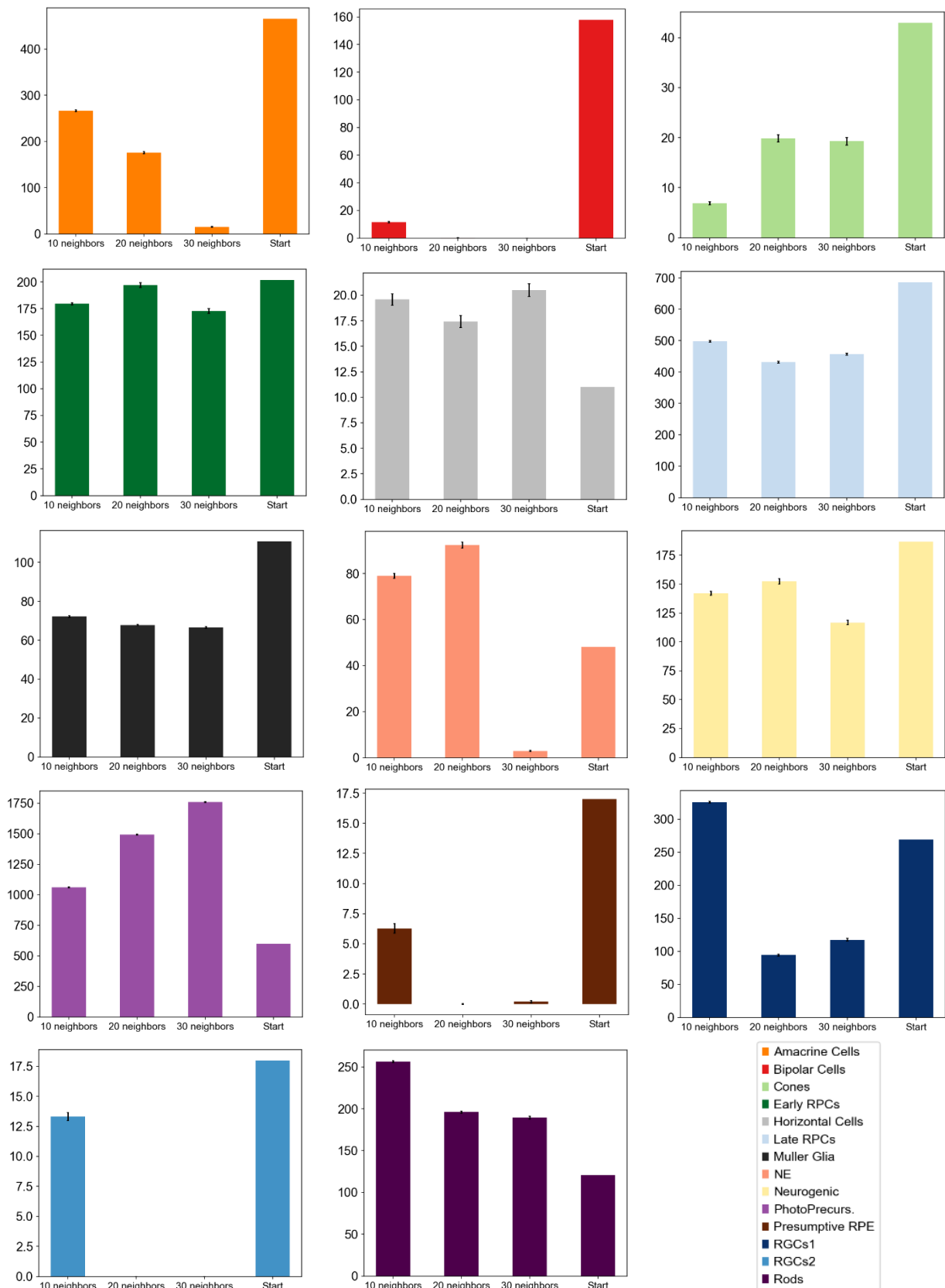
1. Motulsky, A. G. Genetics of complex diseases. *Journal of Zhejiang University. Science. B.* **7**, 167–168 (2006).
2. Aymé, S. & Rodwell, C. The European Union Committee of Experts on Rare Diseases: Three productive years at the service of the rare disease community. *Orphanet Journal of Rare Diseases* vol. 9 30 (2014).
3. Human Genome Project: Sequencing the Human Genome | Learn Science at Scitable. <https://www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828/>.
4. OMIM Entry - # 268100 - ENHANCED S-CONE SYNDROME; ESCS. <https://www.omim.org/entry/268100>.
5. Smith, G. D. *et al.* Genetic epidemiology and public health: Hope, hype, and future prospects. *Lancet* vol. 366 1484–1498 (2005).
6. Scheuner, M. T., Yoon, P. W. & Khoury, M. J. Contribution of Mendelian disorders to common chronic disease: Opportunities for recognition, intervention, and prevention. *American Journal of Medical Genetics* **125C**, 50–65 (2004).
7. Katsanis, N. The continuum of causality in human genetic disorders. *Genome Biology* **17**, 1–5 (2016).
8. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics* vol. 132 1077–1130 (2013).
9. Gurdon, J. B. The Developmental Capacity of Nuclei taken from Intestinal Epithelium Cells of Feeding Tadpoles. *Development* **10**, 622–640 (1962).
10. Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J. & Campbell, K. H. S. Viable offspring derived from fetal and adult mammalian cells. *Nature* **385**, 810–813 (1997).
11. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).
12. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell* **131**, 861–872 (2007).
13. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).
14. Yamanaka, S. Induced pluripotent stem cells: Past, present, and future. *Cell Stem Cell* vol. 10 678–684 (2012).
15. Halevy, T. & Urbach, A. Comparing ESC and iPSC—Based Models for Human Genetic Disorders. *Journal of Clinical Medicine* **3**, 1146–1162 (2014).
16. Kim, H. & Kim, J. S. A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics* vol. 15 321–334 (2014).
17. Sandoe, J. & Eggan, K. Opportunities and challenges of pluripotent stem cell neurodegenerative disease models. *Nature Publishing Group* (2013) doi:10.1038/nn.3425.
18. de Wert, G. & Mummery, C. Human embryonic stem cells: Research, ethics and policy. *Human Reproduction* vol. 18 672–682 (2003).
19. Israel, M. A. *et al.* Probing sporadic and familial Alzheimer’s disease using induced pluripotent stem cells. *Nature* **482**, 216–220 (2012).
20. Yagi, T. *et al.* Modeling familial Alzheimer’s disease with induced pluripotent stem cells. *Human Molecular Genetics* **20**, 4530–4539 (2011).
21. Yahata, N. *et al.* Anti-A β drug screening platform using human iPSC cell-derived neurons for the treatment of Alzheimer’s disease. *PLoS ONE* **6**, (2011).
22. Brennand, K. J. *et al.* Modelling schizophrenia using human induced pluripotent stem cells. *Nature* **473**, 221–225 (2011).
23. Devine, M. J. *et al.* Parkinson’s disease induced pluripotent stem cells with triplication of the α -synuclein locus. *Nature Communications* **2**, (2011).
24. Hofer, M. & Lutolf, M. P. Engineering organoids. *Nature Reviews Materials* 1–19 (2021) doi:10.1038/s41578-021-00279-y.
25. Zachos, N. C. *et al.* Human enteroids/colonoids and intestinal organoids functionally recapitulate normal intestinal physiology and pathophysiology. *Journal of Biological Chemistry* vol. 291 3759–3766 (2016).
26. Fatehullah, A., Tan, S. H. & Barker, N. Organoids as an in vitro model of human development and disease. *Nature Cell Biology* vol. 18 246–254 (2016).
27. Clevers, H. Modeling Development and Disease with Organoids. *Cell* vol. 165 1586–1597 (2016).
28. Kleinman, H. K. & Martin, G. R. Matrigel: Basement membrane matrix with biological activity. *Seminars in Cancer Biology* vol. 15 378–386 (2005).
29. Aisenbrey, E. A. & Murphy, W. L. Synthetic alternatives to Matrigel. *Nature Reviews Materials* vol. 5 539–551 (2020).
30. Hoon, M., Okawa, H., della Santina, L. & Wong, R. O. L. Functional architecture of the retina: Development and disease. *Progress in Retinal and Eye Research* vol. 42 44–84 (2014).
31. Holmes, D. Reconstructing the retina. *Nature* **561**, (2018).
32. O’Hara-Wright, M. & Gonzalez-Cordero, A. Retinal organoids: a window into human retinal development. *Development (Cambridge, England)* vol. 147 (2020).
33. Remington, L. A. *Visual System. Clinical Anatomy and Physiology of the Visual System* (2012). doi:10.1016/b978-1-4377-1926-0.10001-3.
34. Cepko, C. Intrinsically different retinal progenitor cells produce specific types of progeny. *Nature Reviews Neuroscience* **15**, 615–627 (2014).
35. LifeMap Sciences | LifeMap Sciences. <https://www.lifemapsc.com/>.

36. Clark, B. S. *et al.* Comprehensive analysis of retinal development at single cell resolution identifies NFI factors as essential for mitotic exit and specification of late-born cells. *bioRxiv* 1–20 (2018) doi:10.1101/378950.
37. Kim, D. S., Matsuda, T. & Cepko, C. L. A core paired-type and POU homeodomain-containing transcription factor program drives retinal bipolar cell gene expression. *Journal of Neuroscience* **28**, 7748–7764 (2008).
38. Brzezinski, J. A., Uoon Park, K. & Reh, T. A. Blimp1 (Prdm1) prevents re-specification of photoreceptors into retinal bipolar cells by restricting competence. *Developmental Biology* **384**, 194–204 (2013).
39. Samuel, A., Housset, M., Fant, B. & Lamonerie, T. Otx2 ChIP-seq Reveals Unique and Redundant Functions in the Mature Mouse Retina. *PLoS ONE* **9**, e89110 (2014).
40. Brzezinski, J. A. & Reh, T. A. Photoreceptor cell fate specification in vertebrates. *Development (Cambridge)* vol. 142 3263–3273 (2015).
41. Dorval, K. M., Bobechko, B. P., Ahmad, K. F. & Bremner, R. Transcriptional activity of the paired-like homeodomain proteins CHX10 and VSX1. *Journal of Biological Chemistry* **280**, 10100–10108 (2005).
42. Dorval, K. M. *et al.* CHX10 targets a subset of photoreceptor genes. *Journal of Biological Chemistry* **281**, 744–751 (2006).
43. Livne-Bar, I. *et al.* Chx10 is required to block photoreceptor differentiation but is dispensable for progenitor proliferation in the postnatal retina. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 4988–4993 (2006).
44. Brzezinski IV, J. A., Lamba, D. A. & Reh, T. A. Blimp1 controls photoreceptor versus bipolar cell fate choice during retinal development. *Development* **137**, 619–629 (2010).
45. Katoh, K. *et al.* Blimp1 suppresses Chx10 expression in differentiating retinal photoreceptor precursors to ensure proper photoreceptor development. *Journal of Neuroscience* **30**, 6515–6526 (2010).
46. Swaroop, A., Kim, D. & Forrest, D. Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. *Nature Reviews Neuroscience* vol. 11 563–576 (2010).
47. Cheng, H., Khan, N. W., Roger, J. E. & Swaroop, A. Excess cones in the retinal degeneration rd7 mouse, caused by the loss of function of orphan nuclear receptor Nr2e3, originate from early-born photoreceptor precursors. *Human Molecular Genetics* **20**, 4102–4115 (2011).
48. Cheng, H. *et al.* Photoreceptor-specific nuclear receptor NR2E3 functions as a transcriptional activator in rod photoreceptors. *Human Molecular Genetics* **13**, 1563–1575 (2004).
49. Cheng, H. *et al.* In vivo function of the orphan nuclear receptor NR2E3 in establishing photoreceptor identity during mammalian retinal development. *Human Molecular Genetics* **15**, 2588–2602 (2006).
50. Chen, J., Rattner, A. & Nathans, J. The rod photoreceptor-specific nuclear receptor Nr2e3 represses transcription of multiple cone-specific genes. *Journal of Neuroscience* **25**, 118–129 (2005).
51. Peng, G. H., Ahmad, O., Ahmad, F., Liu, J. & Chen, S. The photoreceptor-specific nuclear receptor Nr2e3 interacts with Crx and exerts opposing effects on the transcription of rod versus cone genes. *Human Molecular Genetics* **14**, 747–764 (2005).
52. Hao, H. *et al.* Transcriptional Regulation of Rod Photoreceptor Homeostasis Revealed by In Vivo NRL Targetome Analysis. *PLoS Genetics* **8**, e1002649 (2012).
53. Tarchini, B., Jolicoeur, C. & Cayouette, M. In vivo evidence for unbiased ikaros retinal lineages using an ikaros-cre mouse line driving clonal recombination. *Developmental Dynamics* **241**, 1973–1985 (2012).
54. Kolb, H., Nelson, R., Ahnelt, P. & Cuenca, N. Cellular organization of the vertebrate retina. *Progress in Brain Research* **131**, 3–26 (2001).
55. Curcio, C. A. *et al.* Distribution and morphology of human cone photoreceptors stained with anti-blue opsin. *Journal of Comparative Neurology* **312**, 610–624 (1991).
56. Ingram, N. T., Sampath, A. P. & Fain, G. L. Why are rods more sensitive than cones? *Journal of Physiology* vol. 594 5415–5426 (2016).
57. Demb, J. B. & Singer, J. H. Functional Circuitry of the Retina. *Annual Review of Vision Science* vol. 1 263–289 (2015).
58. Amacrine cell. <https://www.imaio.com/en/e-Anatomy/Anatomical-Parts/Amacrine-cell>.
59. Liu, J. *et al.* Tbr1 instructs laminar patterning of retinal ganglion cell dendrites. *Nature Neuroscience* **21**, 659–670 (2018).
60. Bringmann, A. *et al.* Müller cells in the healthy and diseased retina. *Progress in Retinal and Eye Research* vol. 25 397–424 (2006).
61. The Development of the Eye. http://education.med.nyu.edu/courses/macrostructure/lectures/lec_images/eye.html.
62. Sadler, T. No Title. in *Langman's medical embryology (11th. ed.)*. 295–299 (Lippincott William & Wilkins, 2006).
63. Eiraku, M. *et al.* Self-organizing optic-cup morphogenesis in three-dimensional culture. *Nature* **472**, 51–58 (2011).
64. Simó, R. *et al.* The Retinal Pigment Epithelium: Something More than a Constituent of the Blood-Retinal Barrier-Implications for the Pathogenesis of Diabetic Retinopathy. *Journal of Biomedicine and Biotechnology* **2010**, (2010).
65. Beatty, S., Boulton, M., Henson, D., Koh, H. H. & Murray, I. J. Macular pigment and age related macular degeneration. *British Journal of Ophthalmology* vol. 83 867–877 (1999).
66. Beatty, S., Koh, H. H., Phil, M., Henson, D. & Boulton, M. The role of oxidative stress in the pathogenesis of age-related macular degeneration. *Survey of Ophthalmology* **45**, 115–134 (2000).
67. Beatty, S. *et al.* Macular pigment and risk for age-related macular degeneration in subjects from a northern European population. *Investigative Ophthalmology and Visual Science* **42**, 439–446 (2001).

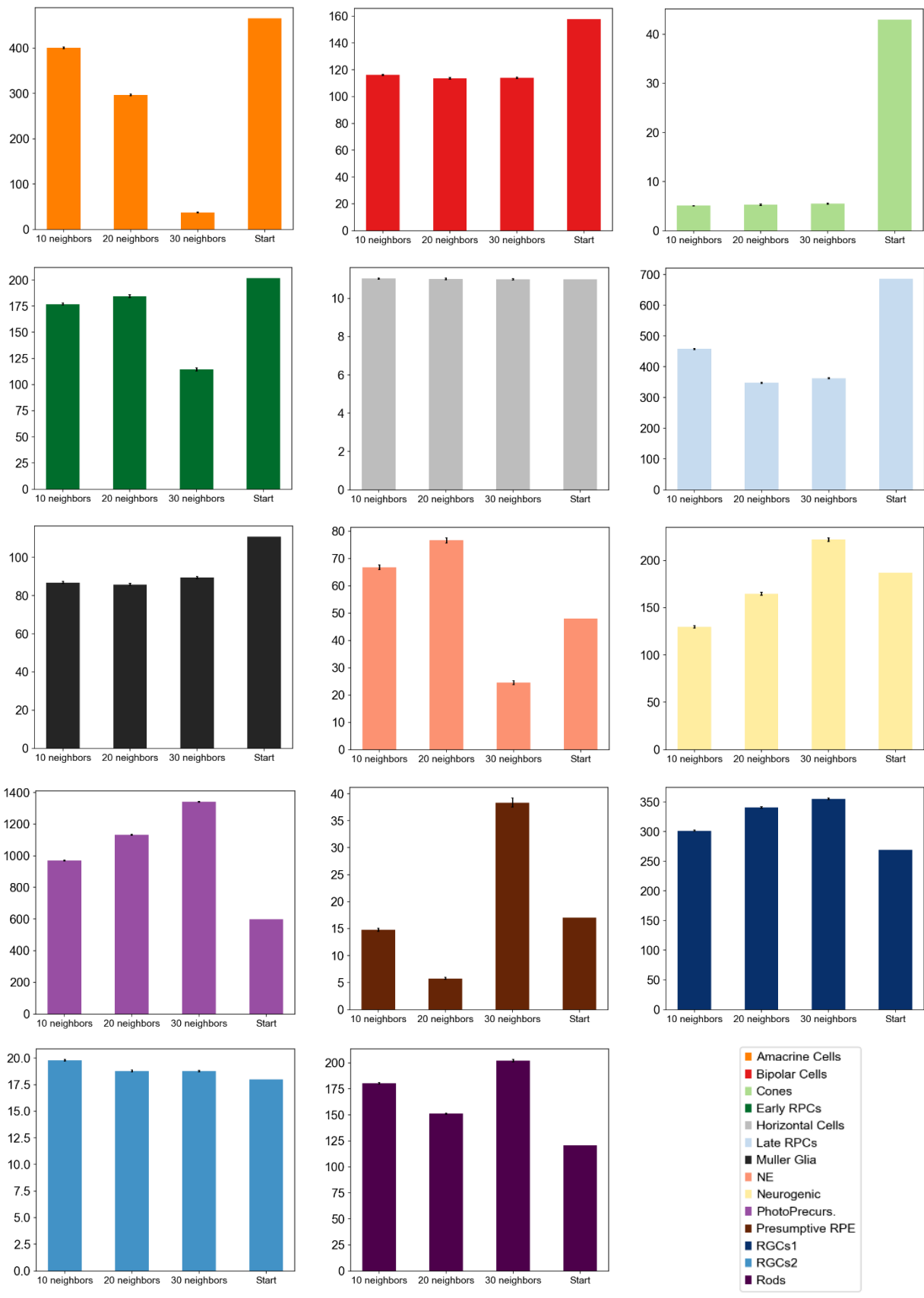
68. Hargrave, P. A. Rhodopsin structure, function, and topography: The Friedenwald lecture. *Investigative Ophthalmology and Visual Science* **42**, 3–9 (2001).
69. Bok, D. The retinal pigment epithelium: A versatile partner in vision. in *Journal of Cell Science* vol. 106 189–195 (Company of Biologists Ltd, 1993).
70. Steinberg, R. H. Interactions between the retinal pigment epithelium and the neural retina. *Documenta Ophthalmologica* **60**, 327–346 (1985).
71. Nguyen-Legros, J. & Hicks, D. Renewal of photoreceptor outer segments and their phagocytosis by the retinal pigment epithelium. *International Review of Cytology* vol. 196 245–313 (2000).
72. Vecino, E., Hernández, M. & García, M. Cell death in the developing vertebrate retina. *International Journal of Developmental Biology* **48**, 965–974 (2004).
73. Braunger, B. M., Demmer, C. & Tamm, E. R. Programmed cell death during retinal development of the mouse eye. *Advances in Experimental Medicine and Biology* **801**, 9–13 (2014).
74. Applebury, M. L. *et al.* The murine cone photoreceptor: A single cone type expresses both S and M opsins with retinal spatial patterning. *Neuron* **27**, 513–523 (2000).
75. Haverkamp, S. *et al.* The primordial, blue-cone color system of the mouse retina. *Journal of Neuroscience* **25**, 5438–5445 (2005).
76. Nikonov, S. S., Kholodenko, R., Lem, J. & Pugh, E. N. Physiological features of the S- and M-cone photoreceptors of wild-type mice from single-cell recordings. *Journal of General Physiology* **127**, 359–374 (2006).
77. Röhlich, P., van Veen, T. & Szél, Á. Two different visual pigments in one retinal cone cell. *Neuron* **13**, 1159–1166 (1994).
78. Eldred, K. C. *et al.* Thyroid hormone signaling specifies cone subtypes in human retinal organoids. *Science* **362**, (2018).
79. Kallman, A. *et al.* Investigating cone photoreceptor development using patient-derived NRL null retinal organoids. *Communications biology* **3**, 82 (2020).
80. Reese, B. E. Development of the retina and optic pathway. *Vision Research* vol. 51 613–632 (2011).
81. Dacey, D. M. Primate retina: Cell types, circuits and color opponency. *Progress in Retinal and Eye Research* **18**, 737–763 (1999).
82. Wassle, H. & Boycott, B. B. Functional architecture of the mammalian retina. *Physiological Reviews* vol. 71 447–480 (1991).
83. González-soriano, J. Morphological types of horizontal cell in rodent retinae: A comparison of rat, mouse, gerbil, and guinea pig. *Visual Neuroscience* **11**, 501–517 (1994).
84. Provis, M. *Development and degeneration of the macula CLINICAL AND EXPERIMENTAL. Clinical and Experimental Optometry* vol. 88 (2005).
85. Curcio, C. A. & Allen, K. A. Topography of ganglion cells in human retina. *Journal of Comparative Neurology* **300**, 5–25 (1990).
86. Sridhar, A. *et al.* Single-Cell Transcriptomic Comparison of Human Fetal Retina, hPSC-Derived Retinal Organoids, and Long-Term Retinal Cultures. *Cell Reports* **30**, 1644-1659.e4 (2020).
87. Zhong, X. *et al.* Generation of three-dimensional retinal tissue with functional photoreceptors from human iPSCs. *Nature Communications* **5**, 1–14 (2014).
88. Brooks, M. J. *et al.* Improved Retinal Organoid Differentiation by Modulating Signaling Pathways Revealed by Comparative Transcriptome Analyses with Development In Vivo. *Stem Cell Reports* **13**, 891–905 (2019).
89. Kaya, K. D. *et al.* Transcriptome-based molecular staging of human stem cell-derived retinal organoids uncovers accelerated photoreceptor differentiation by 9-cis retinal. *bioRxiv* 733071 (2019) doi:10.1101/733071.
90. Kuwahara, A. *et al.* Generation of a ciliary margin-like stem cell niche from self-organizing human retinal tissue. *Nature Communications* **6**, 1–15 (2015).
91. Gonzalez-Cordero, A. *et al.* Recapitulation of Human Retinal Development from Human Pluripotent Stem Cells Generates Transplantable Populations of Cone Photoreceptors. *Stem Cell Reports* **9**, 820–837 (2017).
92. Lowe, A., Harris, R., Bhansali, P., Cvekl, A. & Liu, W. Intercellular Adhesion-Dependent Cell Survival and ROCK-Regulated Actomyosin-Driven Forces Mediate Self-Formation of a Retinal Organoid. *Stem Cell Reports* **6**, 743–756 (2016).
93. Völkner, M. *et al.* Retinal Organoids from Pluripotent Stem Cells Efficiently Recapitulate Retinogenesis. *Stem Cell Reports* **6**, 525–538 (2016).
94. da Silva, S. & Cepko, C. L. Fgf8 Expression and Degradation of Retinoic Acid Are Required for Patterning a High-Acuity Area in the Retina. *Developmental Cell* **42**, 68-81.e6 (2017).
95. Georges, A. *et al.* Combined analysis of single cell RNA-Seq and ATAC-Seq data reveals putative regulatory toggles operating in native and iPSC-derived retina. 1–32 (2020) doi:10.1101/2020.03.02.972497.
96. *The Power of Single Cell Partitioning.* (2020).
97. Capowski, E. E. *et al.* Reproducibility and staging of 3D human retinal organoids across multiple pluripotent stem cell lines. *Development (Cambridge)* **146**, (2019).
98. Chichagova, V. *et al.* Human iPSC differentiation to retinal organoids in response to IGF1 and BMP4 activation is line- and method-dependent. *Stem Cells* **38**, 195–201 (2020).
99. Mellough, C. B. *et al.* An integrated transcriptional analysis of the developing human retina. *Development (Cambridge)* **146**, (2019).

100. Wang, L. *et al.* Retinal Cell Type DNA Methylation and Histone Modifications Predict Reprogramming Efficiency and Retinogenesis in 3D Organoid Cultures. *Cell Reports* **22**, 2601–2614 (2018).
101. Littink, K. W. *et al.* Autosomal recessive NRL mutations in patients with enhanced s-cone syndrome. *Genes* **9**, (2018).
102. Haider, N. B. *et al.* Mutation of a nuclear receptor gene, NR2E3, causes enhanced S cone syndrome, a disorder of retinal cell fate. *Nature Genetics* **24**, 127–131 (2000).
103. Oh, E. C. T. *et al.* Transformation of cone precursors to functional rod photoreceptors by bZIP transcription factor NRL. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 1679–1684 (2007).
104. Bohrer, L. R. *et al.* Correction of NR2E3 Associated enhanced S-cone Syndrome Patient-specific iPSCs using CRISPR-Cas9. *Genes* **10**, (2019).
105. Milam, A. H. *et al.* The nuclear receptor NR2e3 plays a role in human retinal photoreceptor differentiation and degeneration. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 473–478 (2002).
106. Schorderet, D. F. & Escher, P. NR2E3 mutations in enhanced S-cone sensitivity syndrome (ESCS), Goldmann-Favre syndrome (GFS), clumped pigmentary retinal degeneration (CPRD), and retinitis pigmentosa (RP). *Human Mutation* vol. 30 1475–1485 (2009).
107. Audo, I. *et al.* Phenotypic variation in enhanced S-cone syndrome. *Investigative Ophthalmology and Visual Science* **49**, 2082–2093 (2008).
108. Escher, P. *et al.* Mutations in NR2E3 can cause dominant or recessive retinal degenerations in the same family. *Human Mutation* **30**, 342–351 (2009).
109. Stone, E. M. *et al.* Clinically Focused Molecular Investigation of 1000 Consecutive Families with Inherited Retinal Disease. *Ophthalmology* **124**, 1314–1331 (2017).
110. Single Cell Gene Expression - 10x Genomics. <https://www.10xgenomics.com/products/single-cell-gene-expression>.
111. Head, S. R. *et al.* Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques* **56**, 61–77 (2014).
112. Georges, M. *Essential human genomics*. (2019).
113. la Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
114. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology* (2020) doi:10.1038/s41587-020-0591-3.
115. Paulo Czarnewski & Jules Gilet. 10. Trajectory inference analysis of scRNA-seq data - YouTube. https://www.youtube.com/watch?v=XmHDexCtjyw&list=PLjiXAZO27eIC_xnk7gVNM85I2IQI5BEJN&index=11&ab_channel=ChipsterTutorials (2019).
116. *PROTOCOL STEP 4-Library Construction Chromium Single Cell 3' Reagent Kits v2 User Guide*. www.10xgenomics.com/trademarks. (2019).
117. Sequencing Read Length | How to calculate NGS read length. <https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html>.
118. The MGH NextGen Sequencing Core | Core Services. <http://nextgen.mgh.harvard.edu/illuminaChemistry.html>.
119. Panariello, F. *Single-Cell RNA Sequencing and Data Analysis. Theory Refresher and Software Overview: Cell Ranger*. https://elixir-iib-training.github.io/2019-05-07-pozzuoli-singlecell/pres/Panariello_Theory_refresher_and_cellranger_overview.pdf (2019).
120. How are barcodes classified as cell-associated? – 10X Genomics. <https://kb.10xgenomics.com/hc/en-us/articles/115003480523-How-are-barcodes-classified-as-cell-associated->.
121. Dr. Juliansyah Noor. Evaluation of UMAP as an alternative to t-SNE for single-cell data. *Journal of Chemical Information and Modeling* **53**, 1689–1699 (2019).
122. Demeulenaere, L. *Using local variations of RNA velocity to filter genes*. (2021).
123. Loom file format specs — loompy 3.0.6 documentation. <http://linnarssonlab.org/loompy/format/index.html>.
124. Manno, G. la *et al.* Supplementary Note 1: RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
125. Gerstman, B. B. StatPrimer (Version 7.0) t Table. <https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf> (2016).
126. Sequencing Requirements for Single Cell 3' -Specifications -Sequencing -Single Cell Gene Expression -Official 10x Genomics Support. <https://support.10xgenomics.com/single-cell-gene-expression/sequencing/doc/specifications-sequencing-requirements-for-single-cell-3>.
127. Sakharkar, M. K., Perumal, B. S., Sakharkar, K. R. & Kanguane, P. An analysis on gene architecture in human and mouse genomes. *In Silico Biology* **5**, 347–365 (2005).
128. Pérez-Ortín, J. E., Medina, D. A., Chávez, S. & Moreno, J. What do you mean by transcription rate?: The conceptual difference between nascent transcription rate and mRNA synthesis rate is essential for the proper understanding of transcriptomic analyses Insights & Perspectives J. E. Pérez-Ortín *et al.* *BioEssays* **35**, 1056–1062 (2013).
129. Alpert, T., Herzog, L. & Neugebauer, K. M. Perfect timing: splicing and transcription rates in living cells. *Wiley Interdisciplinary Reviews: RNA* vol. 8 (2017).
130. What fraction of mRNA transcripts are captured per cell? – 10X Genomics. <https://kb.10xgenomics.com/hc/en-us/articles/360001539051-What-fraction-of-mRNA-transcripts-are-captured-per-cell->.
131. Packer, J. S. *et al.* A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single cell resolution. **365**, (2020).

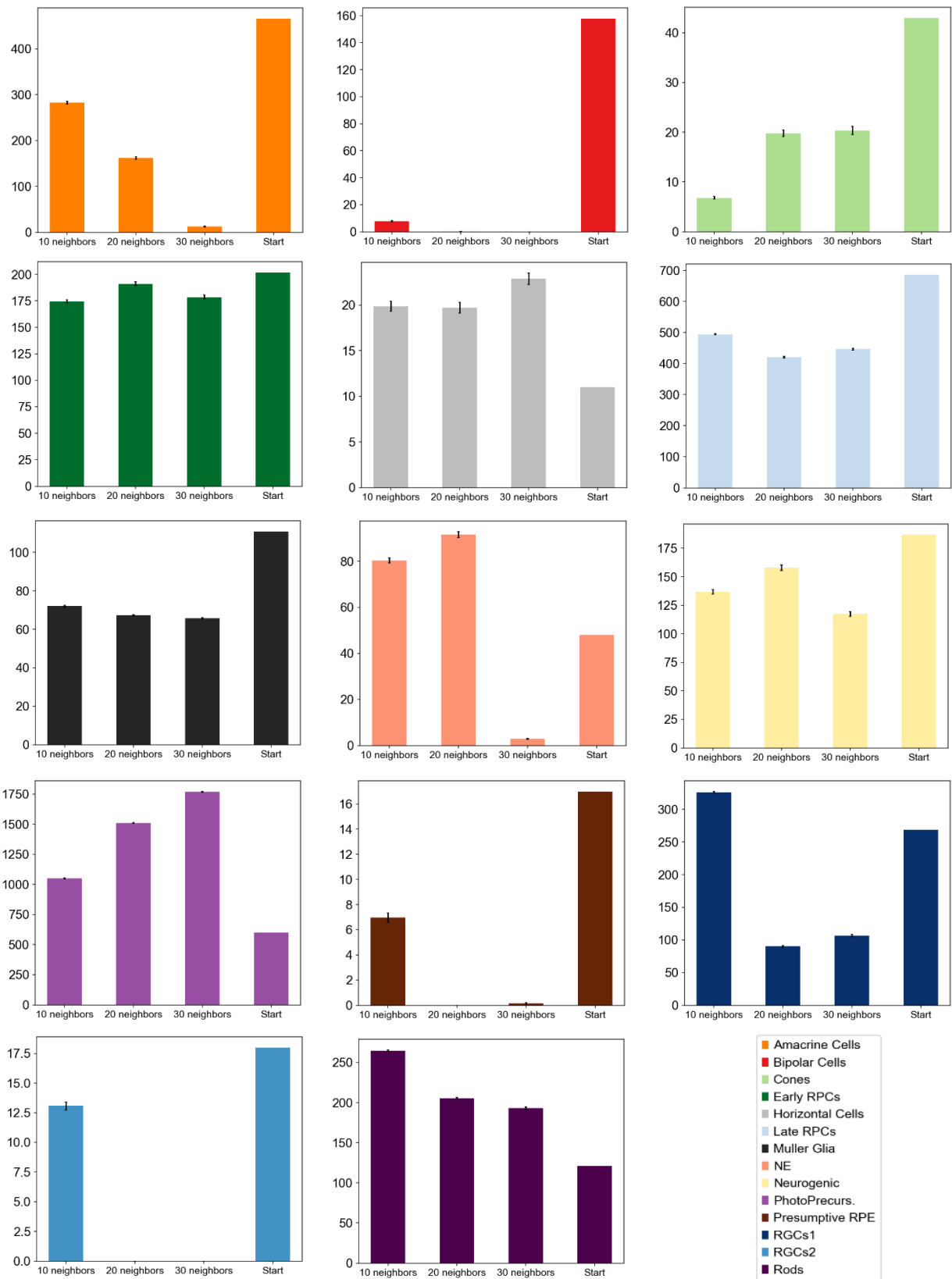
Appendices



Supplementary Figure 1: These graphs show the average quantities obtained with the first model after 500 transitions per cell type from 100 simulations for the three conditions as well as the initial quantity. The 95% confidence interval is indicated by the black bars.



Supplementary Figure 2: These graphs show the average quantities obtained with the second model after 500 transitions per cell type from 100 simulations for the three conditions as well as the initial quantity. The 95% confidence interval is indicated by the black bars.



Supplementary Figure 3: These graphs show the average quantities obtained with the third model after 500 transitions per cell type from 100 simulations for the three conditions as well as the initial quantity. The 95% confidence interval is indicated by the black bars.