SPATIAL ANALYSIS OF COMPLEX BIOLOGICAL TISSUES FROM SINGLE CELL GENE EXPRESSION DATA

Clustering and visualizing functional tissues in *P. dumerillii*

JEAN-BAPTISTE OLIVIER GEORGES PETTIT

UNIVERSITY OF CAMBRIDGE

2014

ABSTRACT

This thesis revolves around such single cell gene expression datasets in the context of the marine annelid or ragworm *Platynereis dumerilii*, an important model organism, part of the lophotrochozoan taxon of the bilaterians. After describing single cell expression data acquired from Wholemount in Situ Hybridization assays for 169 genes as well as single cell RNA-seq data for 72 cells in the developing brain 48 hours post fertilization of *P. dumerilii*, I discuss the main advantages of both methods and propose a back-mapping method to generate a spatially referenced data set of whole transcriptomes at the single cell level.

As the spatial characteristics of the data are crucial to the work presented in this thesis, I also present a 3-dimensional visualization tool that facilitates greatly the upstream and downstream analysis of such datasets. The rest of the thesis focuses on answering the question of identifying cell types from single cell expression data, that is clustering cells together in a meaningful, functional way. Specifically, to take advantage of both the cells location within the tissue and the pattern of gene expression within each cell, I propose a statistical method based on Hidden Markov random fields to cluster the cells according to their gene expression patterns as well as their spatial localization. The method is validated by a simulation study and the quality of the results are compared to those of other clustering methods. Finally the method's output when applied to the *P. dumerilii* in-situ hybridization dataset are biologically validated and functional hypotheses about putative unstudied regions of the brain and their function are formulated.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

Part I

CELL TYPES AND SINGLE CELLS, A VERY SPATIAL RELATIONSHIP

DECLARATION

This thesis:

- is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specied in the text;
- is not substantially the same as any that I have submitted for a degree or diploma or other qualication at any other university; and
- does not exceed the prescribed limit of 60,000 words.

Cambridge, 2014

Jean-Baptiste Olivier
Georges Pettit

1

INTRODUCTION

The work presented in this thesis revolves around a few key biological concepts such as *tissue*, *cell type*, *spatial coherency*, *gene expression*. I define these notions in the Introduction and give an overview of the organism central to the work presented throughout this thesis *Platynereis dumerillii*.

Complex tissues, like the brain are coherent structures composed by large numbers of cells. These cells are not identical to one another. First, visually, cells composing the tissue may have different anatomical characteristics. Second, when observing the cells lives, different cells in different parts of the complex tissue will exhibit different behaviour. Consequently, biological *tissues* can be viewed as an interconnected mosaic of cells having different functions and working together to assume the global role of the tissue. When looking closely at this mosaic, it is sometimes possible to observe under a microscope [113] that the cells spatial organisation is not random, which allows to classify the cells into different categories by considering their appearance and behaviour. These categories are known as *cell types*. Consequently, the definition of a complex biological tissue may be an "ensemble of cells belonging to different cell types regrouped in the same spatial structure".

As an example, the pancreas shown in Figure 1 exhibits an islet architecture [21] with the vast majority of cells belonging to a first cell type organised in spatial structures called *acini* as well as islets of cells from an other cell type called *Langerhans islets* which are anatomically and functionally different.

Interestingly, these two cell types have very different function. Cells that compose the acini define the *exocrine* pancreas and release their product out of the body, namely digestive enzymes into the digestive track [89] whereas the ≈ 1 million cells [48] from the Langerhans islets define the *endocrine* pancreas releasing hormones such as insulin or glucagon into the body via the bloodstream. Importantly, as defined above, the complex tissue called pancreas has a coherent overall func-

tion which is the control of the digestion process. But when looking at the cell type level it is fascinating to see that a wide variety of functions are displayed.

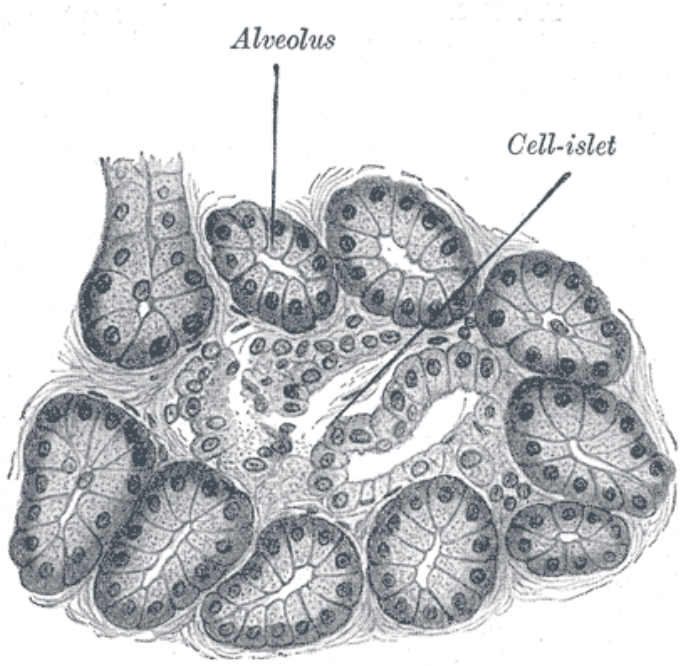


Figure 1: Illustrated section of a dog's pancreas. Acini and Langerhans Islets are indicated by arrows. Public domain work published in [110]

As shown in Figure 1, sub-tissues in complex biological tissues tend to be spatially coherent. In other words cells that belong to the same cell type are usually structured spatially and as a result it seems sensible to assume that cells spatially close to one another have a greater probability to belong to the same cell type. However, the spatial coherency of these sub-tissues is not necessarily always the same. Some cell types may consist of individual cells that are scattered inside another more spatially coherent tissue. An interesting example is the difference between the spatial coherency of cells forming the neuronal tissue in the brain and cells forming a well defined region in the brain like an exocrine gland. When asking the question: "is it likely that this particular cell is fully surrounded by cells belonging to the same cell type?", the extensions created by the axons of neurons (as shown in Figure 2) will decrease this probability. Indeed, axons will grow through other types of tissues to reach their destination [12, 27], making the overall spatial coherency of neural tissues smaller than

very well spatially defined tissues.

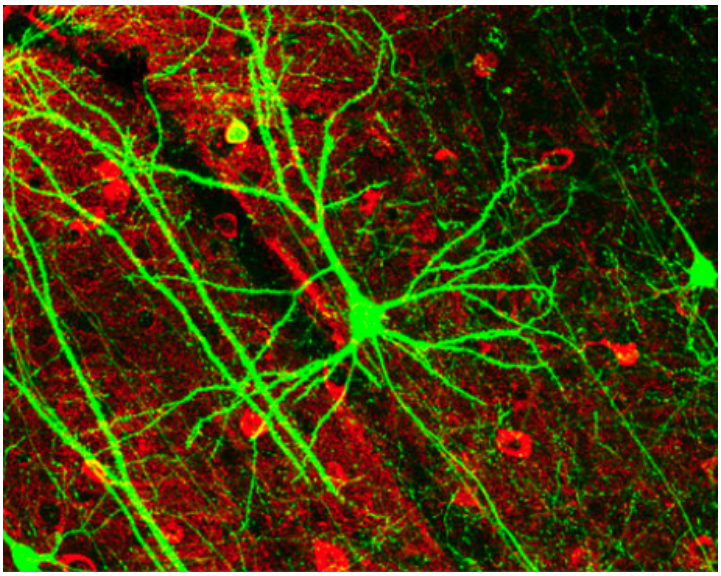


Figure 2: GFP staining of a pyramydal cell in mouse cortex showing dendrites and axons of a neuron linked and at the center the soma where the nucleus is located. This Figure has published under Creative Commons Generic license in [64].

Keeping this in mind will prove important to improve the results. This fact and its consequences on the work presented herein are detailed in chapters 4, 5 and 6.

So far, organs and cell types have been defined mainly by their anatomical traits. However, the functional heterogeneity of complex tissues goes further that simple anatomical traits. For example, in the case of the endocrine pancreas presented above, the Langerhans islets anatomically look like they are composed of a single coherent cell type. However, looking closely at the function of the cells composing them unveils a completely different story. Indeed, despite their apparent similarity, α-cells release the hormone glucagon into the bloodstream, a signal to increase the sugar level in the blood, when β-cells release the hormone insulin which has the exact opposite effect [28].

This very telling example shows that defining cell types based on their anatomical traits is clearly insufficient to characterize tissues. As a result, I will be interested in a trait that fundamentally define how cells are functioning, namely *gene expression*.

1.1 GENERALITIES ABOUT GENE EXPRESSION AND DEVELOPMENT

Throughout this thesis the term *cell* will be used to refer to eukaryotic cells and, more specifically, those of multicellular organisms. Every cell in a complex organism possesses the same genome, that is, the sum of all the genetic information contained in the cell (nucleus and other compartments). This fundamental homogeneity is in plain contradiction with the heterogeneity observed anatomically. If every cell has the exact same DNA, where does the great variability between cell types come from? In other words, what makes a neurone become a neurone and not a pancreatic cell? Answering this type of question defines the field of developmental biology.

A short answer to many developmental biology questions actually is: same genome but different pattern of gene expression. As a central cellular activity, numerous traits exhibited by cells throughout their life from their differentiation to their death, are defined by the way they express some specific parts of their genomes.

Of course, to understand what gene expression is, the notion of *gene* must first be defined. The precise definition of a gene is still controversial. The concept of a "*factor that conveys traits from parents to offspring*" was laid by Gregor Mendel in 1866 [73] when the accepted theory at the time was based on blending inheritance where the traits of the parents appeared mixed in the offspring following a continuous gradient. The most recently published definition of a gene followed the publication of the ENCODE project [37]. It states that a gene is "*a union of genomic sequences encoding a coherent set of potentially overlapping functional products.*"

Gene expression describes the way cells express their genes. Expression of a gene is the process of transcribing the DNA of that particular gene. It is interesting to note that there are several ways to look at gene expression. Indeed in a cell or tissue, at a given time point it is possible to examine whether a gene is expressed or not (binary expression) or how much a certain gene is expressed (quantitative expression). The product of gene expression is RNA molecules. Technically speaking, for transcription to occur in the nucleus a complex system of proteins will attach itself onto the DNA double helix. In the case of protein coding genes, this complex will the RNA polymerase II. This complex attached to the promoter region of the gene,

transcription can begin. The resulting RNA molecule will be a copy of the coding strand of the DNA created by creating the complementary sequence of the template strand. When the polymerase reaches an "end codon", transcription terminates and the RNA molecule is released.

A portion of the RNA molecules are translated into proteins that can have very different purposes. Some will serve directly in the cellular life as functional/structural agents (elements of the ATP synthase for example [17]), others will be excreted by the cell and will serve a purpose at the scale of the organism [58] others called transcription factors will have a regulatory effect on gene expression [75]. In other words the expression of gene G_a, coding for protein P_a might activate, accelerate, inactivate or decelerate the expression of gene G_b and potentially others. This outlines the complex interdependent regulatory system that is gene expression, see Figure 3. For precise examples of gene regulatory networks, see [42, 93, 41, 11].

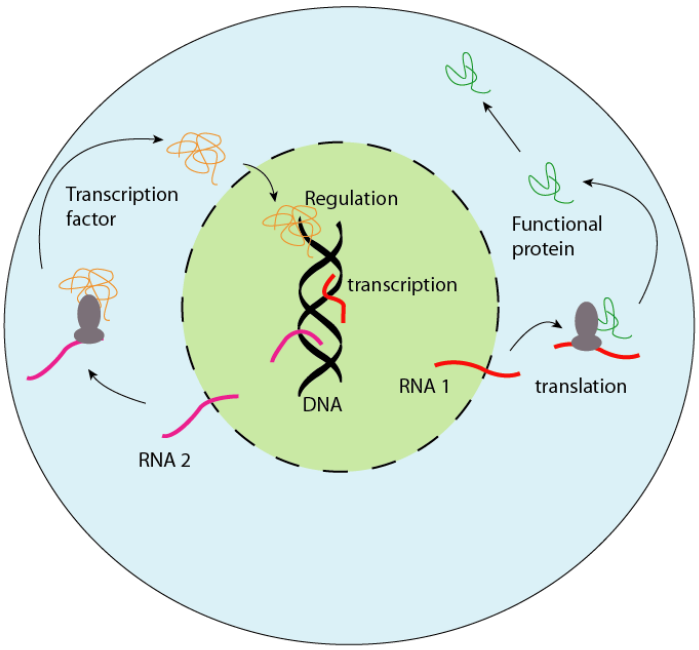


Figure 3: Gene expression and protein translation and gene regulatory networks. The schematics shows that genes in the DNA are transcribed to RNA molecules that are further translated outside the nucleus into proteins. Those proteins can serve various purposes inside the cell or come back to the nucleus to regulate gene expression.

During development, different mechanisms exist that allow cells to develop differently from one to the other. This is how the asymmetrical axis (dorso-ventral, and basal-apical) of the body are defined. The main mechanisms for two cells to take two different differentiation pathways, are signalling gradients and epigenetic control. They both act on the gene expression pattern of the cells. Epigenetic factors are directly coded onto the DNA structure but are not linked directly to the DNA sequence. Epigenetic modifications such as methylation or histone modifications influence gene expression by changing the accessibility of certain regions of the chromatin by modifying the chromatin state or the promoter strength [55]. Signalling gradients are environmental factors, they occurs when cells grow in a medium containing certain chemicals that will influence which gene they express [26], those signalling molecules are can either penetrate the cells and directly regulate transcription as transcription factors [35] or be recognized by some receptors at the surface of the cells that in turn will release transcription factors inside the cell [105].

Consequently, gene expression can be described as one of the key factors during tissue development. Therefore, the ability to study gene expression patterns has revolutionized the field of developmental biology. Technological innovation has been the main driving factor of this revolution. In the next section I will present two methods for assaying gene expression.

1.2 CAPTURING GENE EXPRESSION IN THE LABORATORY

1.2.1 *In-situ hybridization assays*

In-situ hybridization (ISH) is an experimental technique where the practitioner is able to determine in which cells of the tissue under study a particular RNA is present. As opposed to Southern blotting [94], ISH assays not only enable the experimentalist to know whether a gene is expressed or not, but also where in the tissue it is expressed. First proposed in 1969 by Pardue [81] and John [56] independently, in-situ hybridization (ISH) used radioactive tritium labelled probes on a photographic emulsion to reveal parts of the studied tissues where particular RNA or DNA sequences were present.

With the development of fluorescent labelling techniques [63, 85] allowing for faster, more sensitive and of course safer hybridization assays compared to radioactive probes [98], Fluorescent in-situ hybridization (FiSH) quickly became the standard technique to study gene expression in the spatial context of biological tissues. Importantly, using multiple fluorescent probes of different colours allowed the simultaneous localization of several RNA fragments within a tissue [78].

For small enough tissues under study, it is possible to hybridize the probes in the whole animal. This method is called Wholemount in-Situ hybridization (WiSH) and a 3-Dimensional representation of the expression map of a gene can be deduced using confocal microscopy to study the patterns of gene expression in the tissue slice by slice.

1.2.2 *RNA sequencing*

Whole Transcriptome Shotgun Sequencing (WTSS) also called RNA sequencing (RNA-seq) [76, 107] has developed alongside Next Generation Sequencing (NGS) techniques used to sequence genomic DNA. In RNA-sequencing, only the fraction of RNA molecules in the cell are targeted. Protein coding mRNA molecules can further be selected, they are separated from the rest by targeting the polyadenylated 3′ tail, a characteristics exhibited by to protein coding transcripts and a few other types of transcripts only (lncRNAs for example). Most current technique use magnetic beads to achieve this separation [77, 76].

Once isolated from a population of cells, transcripts undergo fragmentation to obtain an average length of 200-300 nucleotides. The next step is the reverse transcription, which creates a complementary DNA (cDNA) library using viral reverse transcriptase enzymes. After amplification using quantitative Polymerase Chain Reaction (qPCR), the cDNA library is ready to be sequenced by NGS technology.

NGS refers to numerous experimental techniques used to acquire the DNA sequences from a sample with a high throughput. I will not extensively describe all of these methods in this thesis but succinctly present some of them.

- Massively parallel signature sequencing (MPSS) [19] was the first NGS, developed in the 1990s. The method revolves around

beads that capture the DNA, adaptors biding and subsequent sequencing.

- 454 pyrosequencing [68] is a sequencing technique that revolves around luciferase binding, a light emitting protein to detect and sequence the nucleotides while the DNA molecules are amplified via Polymerase Chain Reaction (PCR)
- Illumina sequencing [14] is one of the mainly used NGS technology nowadays. Similarly to 454 pyrosequencing, dyes are used to detect the nucleotides of "DNA colonies" while they are amplified inside picoliter volume wells.

All the described sequencing methods generate large datasets of small reads, which need to be mapped back onto the reference genome of the considered species, providing this genome is available. In this case, the resulting dataset will reflect a snapshot of the whole transciptome in the studied cell population. In the case where the reference genome is not fully available, an alternative option is to map the reads back to a list of known gene sequences. The resulting dataset will represent a quantitative image of the considered genes in the cell population at one point in time.

Because of technical limitations in these sequencing protocols, until very recently the starting quantity of RNA had to be relatively important (this issue is discussed further in 2.2). This is why most of the published RNA sequencing studies use a population of cells as a starting point. This however, means that the gene expression landscape obtained as an output will represent an averaged expression over all the cells used as an input.

Importantly, when comparing RNA-seq to the previously described in-situ hybridization technique, if the methodological burden to analyse the expression of a lot of genes at the same time is greatly reduced, the spatial localisation of the cells is lost during the protocol.

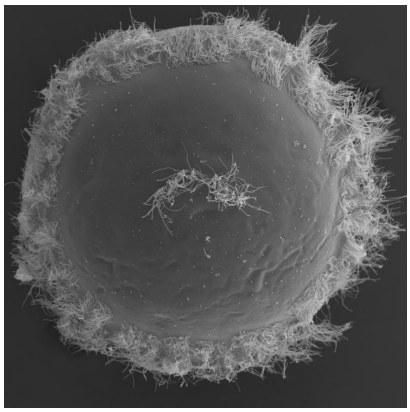
Developmental biology relies of numerous model organisms such as *Mus musculus*, *Caenorhabditis elegans*. The work presented in this thesis revolves around the marine annelid *Platynereis dumerillii*. In the next section, I will introduce generalities about this interesting model organism and in particular how its neural system is ideal to study brain development and evolution.

1.3 PLATYNEREIS DUMERILII, AN IDEAL ORGANISM FOR STUDYING BRAIN EVOLUTION

1.3.1 *General description*

Platynereis dumerillii is a marine annelid of the class Polychaeta, which has been established as one of the main marine animal models in the fields of evolutionary, and developmental biology as well as ecology, toxicology and neurobiology [49, 102, 45, 34, 38, 39].

P. dumerillii populates shallow (no more than 3m deep) ocean floors around the world. It is commonly found in the Mediterranean sea, the north Atlantic coast of Europe as well as in the shallow seas surrounding Sri Lanka, Java and the Philippines. Eggs, embryos and larvae are roughly 160µm long while the adults can measure up to 6cm in length.



(a) Larval form of *P. dumerillii*. Image: MPI for Developmental Biology.

(b) Adult *P. dumerillii*. Image: Arendt group, EMBL

Figure 4: *Platynereis dumerillii*'s larva and adult forms.

They are several reasons why *P. dumerillii* has been chosen as a model by numerous laboratories. Indeed, evolutionary wise, *P. dumerillii* shows several interesting characteristics. As a member of the bilaterians *P. dumerillii* has a defined bilateral symmetry. It belongs to the lophotrochozoan taxon of the bilaterians as opposed to most of the well established model animals which either belong to the ecdysozoans (*Caenorhabditis elegans*, *Drosophila melanogaster*) or the deuterostomes (mouse, human). *P. dumerillii* as one of the only lophotrochozoan model is essential to be able to use comparative approaches full range of bilaterians [39].

P. dumerillii also exhibits an exceptionally slow evolving nature and it has even been described as a "living fossil" for that reason [39]. For that reason, the numerous ancestral developmental characteristics of *P. dumerillii* are a snapshot of the common past of all bilaterians. An interesting example described in [32, 103] is the conserved molecular topography of the genes responsible for the development of the central nervous system between *P. dumerillii* and all vertebrates. This slow evolving nature makes *P. dumerillii* a better comparison with vertebrates than fast evolving species like *Drosophila* and nematodes where derived features can obscure evolutionary signal [39, 10].

Experimentally speaking, model organisms are chosen for several characteristics that make them easy to use in the laboratory, those characteristics include but are not limited to the size of the animal, the conditions required for the organism to develop, gestation and development time, ease to produce a new generation.

in that regard *P. dumerillii* is nearly an ideal animal. Even in their adult form, they are relatively small, they can easily be kept and bred in captivity producing offspring throughout the year [38]. Furthermore, the behavioural characteristics of *P. dumerillii*'s mating ritual have been well studied and can be reproduced on demand in the laboratory. The "nuptial dance" happens on the water surface. Males and females respectively release the sperm and eggs synchronously. This activity is synchronized by pheromones released into the water [115]. Over 2000 individuals can be produced within a single batch. Every new individual will undergo embryonic then larval development before reaching *P. dumerillii*'s adult form.

1.3.2 *Larval development*

Similarly to the other polychaetes, the larval development of *P. dumerillii* can be decomposed into three main anatomical stages, as detailed in [47]: the trochophore, the metotrochophore and the nectochaete. The trochophore is spherical and moves thanks to an equatorial belt of ciliated cells as well as an apical organ displaying a ciliary tuft [90, 79] as seen in Figure 4a and schematically in Figure 5. The metotrochophore stage is characterized by the development of a slightly elongated segmented trunk compared to that of

the trochophore [44]. The next developmental stage is referred to as the nectochaete larvae which resembles the adult (figure 4b) in many traits, especially with parapodial appendages used for swimming and crawling [44].

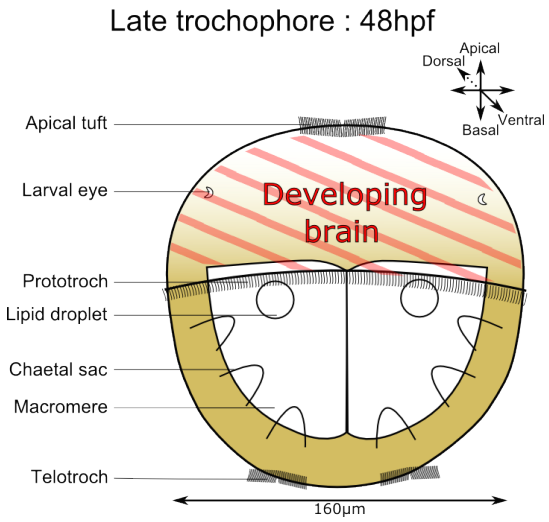


Figure 5: *Platynereis dumerillii*'s larvea development at 48hpf (late trochophore). Red stripes indicate the area that forms the developing brain of the larvae.

Aside from this purely anatomical description, an additional staging system exists and has become the norm for current studies. The development is measured in *hours post fertilization* (hpf) at 18°C.

A key factor making *P. dumerillii* such an interesting model to work with is the fact that after fertilization, the ≈ 2000 larva will start developing at the exact same time, in a synchronous fashion. Furthermore, the larval development of *P. dumerillii* follows a very stereotypical pattern with little variation from one individual to the other; this is true even between batches provided the temperature is kept constant [38, 34]. An illustration of this synchronous development is shown in Figure 6. This is a very important feature as it allows biologists to repeat experiments on several individuals at a very close developmental stage even if they are from different batches.

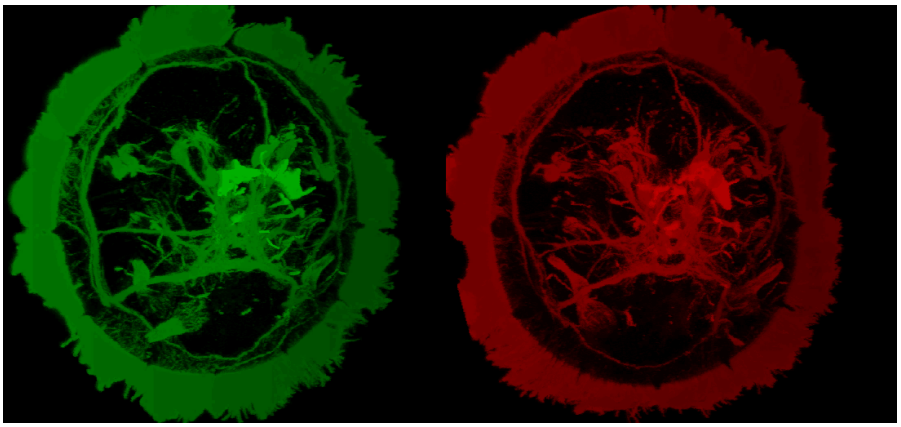


Figure 6: *Platynereis dumerillii*'s stereotypical and synchronous development. In green and red are two different *P. dumerillii* individuals' with the same gene expression being highlighted. They show extremely similar patterns of development for the nervous system.

1.3.3 *Platynereis' nervous development until 48hpf*

Describing the entire development of *P. dumerillii* does not fall within the scope of this thesis. Indeed, I will only be interested in the brain of *P. dumerillii*'s larvae at 48hpf. Therefore, it is important to have an anatomical idea of what the brain looks like at this time in development and what inherent characteristics will be the most interesting to investigate. *P. dumerillii*'s larval brain development is detailed in [39]. From the early trochophore (24-26hpf) neural system development starts taking place. The apical ganglion which contains one sertonergic cell and a few neurons linked to the nerve of the ciliary band of the larva called the prototroch forms at the apical tuft, (see Figure 5). This allows the first movements of the larvae thanks to the cialited cells of the prototroch.
The mid-trochophore (26-40 hpf) sees the formation of the first cerebral commissure: a band of nerves interconnecting the ventral nerve cord and the brain, which is a typical feature of annelid neurobiology. During this phase the apical ganglion becomes bigger with three more serotonergic cells.
The late trochophore (40-48hpf) sees the formation of the second commissure in the ventral nerve cord. It is at the end of this stage that the tissues of the brain become more complex with a notable increase in the number of neurites [39].
The data used in the rest of this thesis will not encapsulate the whole larvae, just the developing brain (see Figure 5) thus excluding the ventral part of the nervous system. The best studied areas of the de-

veloping brain are the larval eyes, the developing adult eyes and the apical organ on the dorsal side. On the ventral side are located the mushroom bodies, a pair of structures that are known to play a role in olfactory learning and memory in insects and annelids [104].

Consequently, even at this early stage in a relatively "simple" organism, the brain quickly becomes an extremely complex tissue. Cell types diverge and functional areas are formed. Before trying to understand more about *P. dumerillii*'s brain organization, it is interesting to ask the more general question of how complex tissues such as the brain are defined spatially.

1.3.4 *Building a image library of gene expression for Platynereis*

During his PhD, Raju Tomer and other members of the Arendt lab in EMBL, used wholemount in-situ hybridization to create an image library of gene expression in the brain of *P. dumerillii*. They were able to record gene expression in the full brain at 48hpf for 169 genes. In practice, each individual larvae was dissected to isolate the region containing the developing brain. Each brain was then stained with two different fluorescent probes corresponding to two messenger RNAs (mRNA). One of the genes is considered a reference, as it is always hybridized in all the assays (the main reference gene used was Emx) alongside another gene of interest, see Figure 7. Each brain was then visualized with laser confocal microscopy to reveal the gene expression patterns in the brain slice by slice, generating at the same time 3D coordinates for each slice.

As mentioned previously, the larval development of *P. dumerillii* is highly similar in every individual larvae. In the case of this study where WiSH was performed on many independent animals, the stereotypical development of *P. dumerillii* has proven essential. Indeed, having the same reference gene localized in all assays allowed Tomer to align all other gene expression patterns onto this scaffold. The result is an image library of 169 gene expression patters in the full brain of *P. dumerillii* with a exploitable spatial reference that allows for a very precise mapping.

However useful and practical WiSH may be, such assays are limited in terms of the number of genes one is able to study. Indeed,

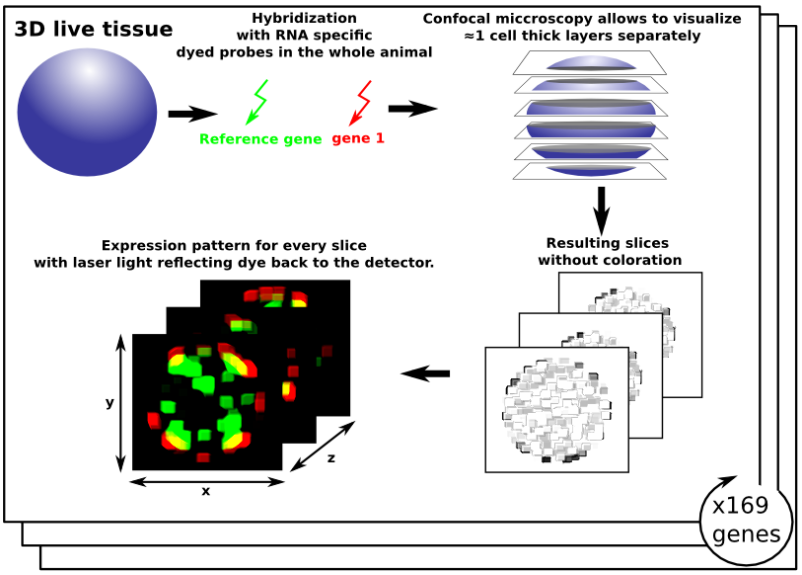


Figure 7: Wholemount in-situ hybridization ssays used to create a 169 genes catalogue of gene expression in the brain of *P. dumerillii*. From the live tissue cut into thin fixed layers, every slice is stained with a reference gene and a gene of interest that will reveal areas of expression under fluorescent microscopy. The process repeated 169 times for key genes in *P. dumerillii* neural development has been generated by [104]

each individual larvae only provides the expression of two genes, one being the reference. To overcome this problem, crucial developments in sequencing technologies have brought about a new way of studying the expression of the whole transciptomic landscape in a single assay,namely RNA sequencing.

1.4 SUMMARY

In the introduction, I have presented a central aspect of cell and developmental biology, namely, gene expression. I have given an overview of how cells express their genomes and what how the expression of specific signalling genes can influence the fate of cells during development. I have also described two methods that allow practitioners to capture gene expression from a biological tissue: in-situ hybridization and RNA-seq.

Subsequently, I described *Platynereis dumerillii* and the advantageous traits it exhibits for developmental biologists especially in the field of neural development. I have discussed the fact that anatomical traits

are not sufficient to fully comprehend the deep heterogeneous patterns of functionalities inside a complex organ such as the brain. In order push this understanding further I have discussed how gene expression levels can be used to characterise different tissues and how an image library of gene expression for 169 gene was generated by [104] in the full brain of *P. dumerillii* using WiSH.

So far, I have considered biology at the scale of the tissue, or the sub-tissue. However, the heterogeneity of complex biological tissues does not stop at this scale of study. In fact, with a top-down approach looking at big tissues and then separating them in smaller sub-tissues until "true" functional tissues are defined is an extremely complicated problem. A solution to this problem would be to reverse the approach from a top-down to a bottom-up mindset. This means reducing the scale of study to the smallest biological unit available, the single cell, defining the heterogeneity of gene expression at the single cell level and going back up to the functional tissue level from there. Instead of a fragmentation problem, this becomes a clustering problem, attaching single cells to a certain number of categories. In order to implement such an approach, single cell gene expression data is therefore needed.

This model animal, the image library of gene expression in its brain and the question of finding functional tissues and sub-tissues from single cell gene expression are the key concepts that motivate this thesis.

In Chapter 2, I describe how advances in both RNA-seq and in-situ hybridization have allowed the extraction of single cell gene expression data and how this data is analysed. I also describe how an "ideal" dataset of spatially referenced single cell expression data can potentially be create by mapping the results of single cell RNA-seq onto an in-situ hybridization scaffold.

In Chapter 3 I present a tool I developed to allow an easy visualization of 3D information and more specifically clustering results. This tool is central to the upstream and downstream analysis of the data and the results used in the thesis.

In Chapter 4 I present the theoretical background underlying the Hidden Markov random fields (HMRF) spatial clustering method I

developed and how this method diverges from previously described HMRF methods.

I then evaluate in Chapter 5 the performances of the HRMF method on simulated data. I describe how binarized gene expression data with a spatial component was simulated, then validate the parameters estimation by the method and finally compare the performance of the method to other non spatial clustering methods in terms of clustering quality and computing resources.

Finally, I present in Chapter 6 the results of the HMRF method when applied to the binarized single cell in-situ hybridization data in the brain of *Platynereis dumerillii*. First, I propose a scoring method in order to use the clustering results to functionally define the regions. I then validate the method further by describing known regions of the brain that are found by the method. Finally I present how functional hypothesis can be made for unstudied regions of the brain.

2

FROM TISSUE TO SINGLE CELL TRANSCRIPTOMICS, A PARADIGM SHIFT

2.1 SPATIALLY REFERENCED SINGLE CELL-LIKE IN-SITU HYBRIDIZATION DATA

2.1.1 *Dividing images into "cells"*

Because in-situ hybridization preserves spatial information in the tissue under study, measuring gene expression at single cell resolution from an image obtained through confocal microscopy is a matter of microscope performance and cell size. For big enough cells, single cell resolution has been documented as far back as 1989 [100, 86].

When considering the *P. dumerillii* brain dataset, with current microscope technology, achieving single cell level resolution on one particular image is feasible. However, the main limitation is analysing the quantity of data involved; indeed, each brain is separated into 20 slices, for 169 genes this yields 3380 images that require inspection. This technical bottleneck can be overcome with an automated way of analysing the fluorescence images. However this is not an easy task, as the computer program required needs to be able to *see* and divide the global picture into cells. Considering that all cells do not exhibit the same shape and size, constructing this *cell model* is a very complicated task.

It is for instance possible to highlight the limits of the cells and to automatically acquire those boundaries through computer vision methods. This process relies on targeting proteins in the membrane or in the extracellular matrix of the cells with specific fluorescent probes. Once the boundaries are acquired, defining every cell is a matter of finding enclosed spaces. To that end, numerous contour detection algorithms exist [66, 36, 9].

Unfortunately, a dataset with the cell limits highlighted does not yet exist for *P. dumerillii*'s brain, making a precise division of the im-

ages into cells very difficult. Instead, Tomer used a simple approach that divides the images into "cubes" [104].

2.1.2 *A simple cell model, the "cube" data*

Every slice of *P. dumerillii*'s brain being aligned onto the reference gene scaffold (see section 1.2) for all 169 genes, the "cube" model simply consists of dividing each image into squares approximately the size of an average cell. In the *P. dumerillii* dataset, the size chosen was 3 μm^2 [39]. Importantly, this is actually smaller than the average cell size in *P. dumerillii*'s brain. Each slice of the brain being approximately 3 μm thick, the resulting dataset, spatially referenced in 3D, will contain 3 μm^3 cubes, each of which is associated with the luminescence data for each of the 169 genes.

Of course this cell model is far from perfect: it assumes that every cell in the brain is roughly the same size and cubical, which is clearly not the case. Consequently, the "cube" model will introduce errors in the dataset. The first type of error occurs within areas where the genes under study are highly expressed. In that case, the light emission might contaminate the surrounding cubes that do not necessarily express the same gene (see Figure 8A). The second type of error is introduced by the choice of 3 μm^3 cubes. As they are smaller than the average cell, some cubes will fall on areas that may be artificially empty. Indeed, transcription in the cells mainly happens in the nucleus. mRNA molecules then travel to the cytoplasm to be translated but they are not evenly distributed across the cell; in particular for some large cells, parts of the cytoplasm may record no expression in a cell that actually contains a lot of transcripts (see Figure 8B).

Hence, the data will tend to exhibit spatial discontinuity and inconsistency. With this fact in mind, any automated way of interpreting this data, in the case of this thesis: clustering "cubes" into cell types, will have to take into account this spatial discontinuity and try as much as possible to smooth over those potential expression gaps.

However, even with this simple cell model, the data generated by [104] is highly valuable. Indeed, not only does this dataset give a snapshot of gene expression for 169 genes in the full brain of *P. dumerillii*,

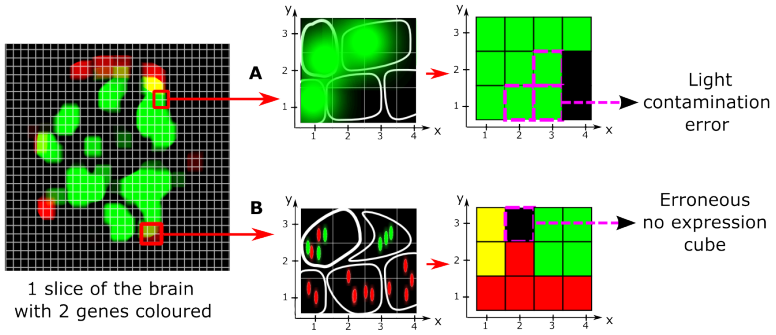


Figure 8: Errors introduced by the "cube" cell model. Path A shows how regions with highly expressed genes can introduce errors through light contamination. Path B shows how some cubes may appear artificially void of expression because of the uneven distribution of transcripts inside the cytoplasm especially for large cells.

it also attaches spatial information to each data point.

2.2 SINGLE CELL RNA SEQUENCING, BUILDING A MAP OF THE FULL TRANSCRIPTOME

2.2.1 *Single cell RNA sequencing*

The scale shift from tissue to single cell is harder to achieve in the case of RNA-seq. As described in the Introduction 1.2, an important factor for the success of RNA-seq assays is the input quantity of RNA to be sequenced. Taking mammalian cells as a reference, the quantity of RNA depends a lot on the cell type considered and can vary between 10 and 50 pg per cell, only 2% of which is mRNA [50, 52]. With such a small input quantity, distinguishing biological variation between different cells from the technical variation linked to mRNA capture rates and to cDNA amplification protocols is extremely challenging.

However, with the creation of new protocols [87, 99], and the rise of microfluidics to faciliate the extraction and sequencing of single cells [80], the last couple of years have seen a dramatic increase in the number of single cell RNA-seq based studies [53, 69, 112, 95, 33]. However, challenges still need to be overcome in order to analyse further complex tissues using such approaches.

2.2.2 *Mapping back gene expression to a spatial reference*

Single cell RNA-seq captures a snapshot of the entire transcriptome of a given cell at a given point in time. However, to analyse cells from a complex tissue, current protocols require that the tissue is reduced to a suspension of single independent cells. This prevents the user from keeping track of any spatial information about the cells. Hence, when analysing single cell RNA-seq data from a complex tissue, back-mapping every cell to its original location becomes a crucial problem.

In order to achieve this back-mapping, a reference is needed. This reference should consist of an independent assay where gene expression in the considered tissue is defined for enough genes at a spatially small enough resolution to find for each sequenced cell, if not its exact original location, at least a restricted region of the tissue from which the sequenced cell originated with a high probability.

Fortunately, in-situ hybridization assays provide exactly this type of data and I will present in the last section of this Chapter (2.5) a methodological proof-of-concept of this back-mapping in the brain of *P. dumerillii* with 72 sequenced single cells. However, before that , I will discuss the impact of the noise level in both in-situ hybridization and single cell RNA-seq assays on the quantitative trait of the resulting datasets.

2.3 ABOUT THE QUANTITATIVE TRAIT OF SINGLE CELL EXPRESSION DATA

2.3.1 *Light contamination in in-situ hybridization data*

The light intensity value obtained from in-situ hybridization assays can be considered as a quantitative measure of gene expression [34]. Indeed, the light emitted by every cell in the considered tissue is correlated with the number of RNA fragments of the gene of interest present in the cell as each fragment bound to a probe is an independent source of emission and the probes are hybridized in the cells in large excess. This means that if the targeted gene is highly expressed in a cell, there will be more sources of emission, thus making the overall light intensity captured on this area higher than in a cell ex-

pressing the gene at a low level.

As mentioned in a previous section 2.1, in-situ hybridization assays at the single cell level are prone to localized errors due to the cell model. One explanation for those errors, as shown in Figure 8B is the phenomenon of light contamination. When a large group of neighbouring cells express the same gene, the additivity of light intensity mentioned above means that even though the cells express the gene at the same rate, cells surrounded by a lot of other cells expressing the same gene will have an abnormally high light intensity readings due to light contamination from the adjacent areas. As a result, when considering an hypothetical circular portion of tissue where a gene is monotonously expressed, the recorded light intensity will show a gradient with the maximum localized on the circle's centre.

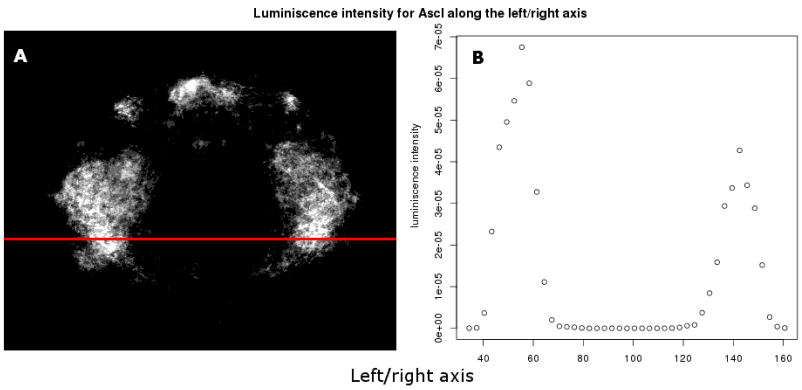


Figure 9: **Light contamination in in-situ hybridization luminescence data seen with the example of the gene Ascl.** Panel A shows the raw fluorescent microscopy capture of the gene's expression for one layer in the brain of Platynereis. Panel B shows the light intensity measured along the red line in panel A. Because of the small scale of study, cells surrounded by other cells expressing a particular gene will have higher intensity values because of nearby light contamination.

As shown in Figure 9, the issue of light contamination seems to occur when using the 3 μm^3 "cube" model. In this context, and because of the single cell scale of this study, considering the in-situ hybridization data as quantitative may introduce significant errors. In order to avoid this light contamination bias a solution is to transform the quantitative data into binary data where, for a given "cube", genes are simply expressed or not. The binarization method is described in the following section (2.4).

2.3.2 *Technical noise in single cell RNA-seq data*

Single cell RNA-seq is also prone to high levels of noise. This technical noise is caused by the minute amounts of starting RNA material. A study laid by Philip Brennecke, Simon Anders and Jong Kyoung Kim [18], proposes a statistical method to overcome this high noise level and distinguish between biological variation and technical variation in the gene expression levels.

To illustrate the dramatic increase in noise level, they used a series of dilution assays, reducing step by step (5000 pg, 500 pg, 50 pg, 10 pg) the input quantity of RNA fragments extracted from total *Arabidobsis thaliana* RNA with two technical replicates each time using the Tang protocol [99]. The authors of the study let me analyse this data, and after normalizing by the size factor using the Bioconductor package DESeq [8] the scatter plots shown in Figure 10 were generated.

It is clear from these dilution assays that the noise level is correlated with the input quantity. Even though highly expressed genes are consistently well quantified even with 10 pg input material, for most of the genes, with less than 50 pg input RNA it seems dangerous to assume the results of single cell RNA-seq as quantitative with the current technological capabilities.

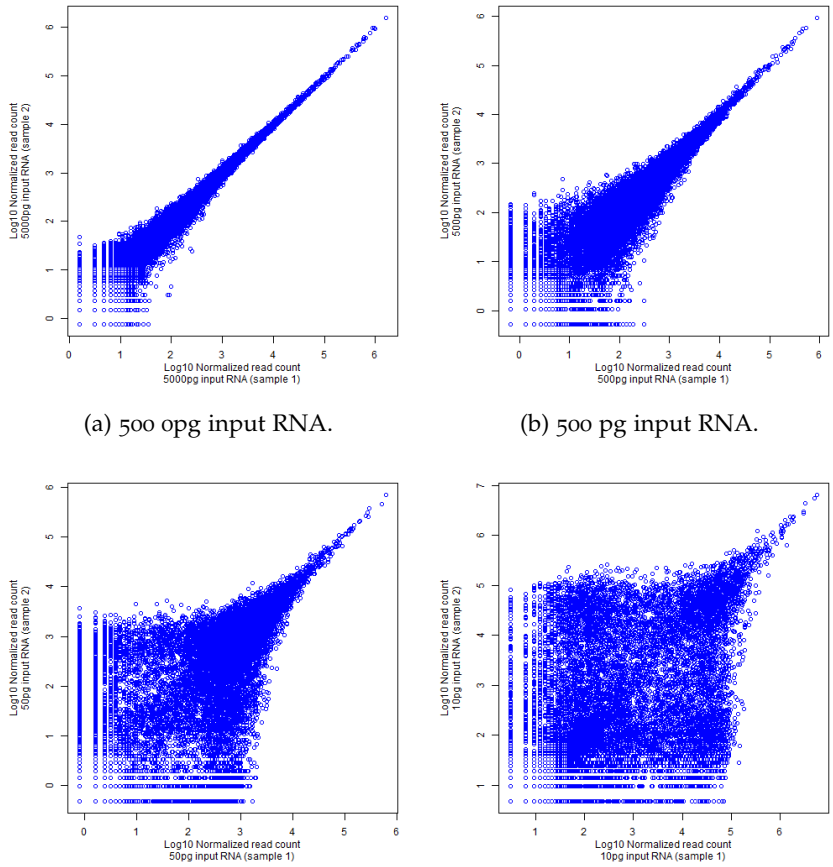
2.3.3 *Conclusions*

The paragraphs above have shown that neither in-situ hybridization nor RNA-seq data are fully quantitative when the scale is lowered to the single cell level. To avoid problems linked to the noise level in the rest of this study, a solution was to binarize single cell datasets. However, binarization is not a trivial problem as discussed in the following section.

2.4 BINARIZING GENE EXPRESSION DATASETS

2.4.1 *Binarizing in-situ hybrdization datasets*

As shown in Figure 9 and discussed in the previous section 2.3, the various problems linked to light contamination can be avoided by



(a) 500 opg input RNA.

(b) 500 pg input RNA.

(c) 50 pg input RNA.

(d) 10 pg input RNA.

Figure 10: Dilution series of total A. thaliana RNA

transforming the "quantitative" fluorescence information into binary data. In other words, if S is the set of all "cubes" in the brain, M the set of all the considered genes and $y_{i,m}$ the value retrieved from the in-situ hybridization data for "cube" $i \in S$ and gene $m \in M$, then $y_{i,m} = 1$ if gene m is expressed at site i, $y_{i,m} = 0$ otherwise. The binarization process itself is not trivial. Indeed, defining the light intensity threshold above which a gene is considered expressed is a complicated problem, especially for noisy data.

Looking at the density of intensities across all the "cubes" for each gene yielded two very different scenarios: some densities were separated into two clear peaks, making the threshold easy to find while others exhibited a single peak making it hard to choose a clear cut value as shown in Figure 15. After trying different thresholding methods based on those densities, I found, in collaboration with Kaia Achim and Maria Tosches from the Arendt group in EMBL Heidel-

berg that none of them resulted in binary expression that was satisfying for many genes when compared to the manual inspection of in-situ hybridization raw images. Considering that this binarized dataset will be the cornerstone of the work presented in this thesis, it was very important to achieve a high confidence thresholding. Given the small number of genes studied (169), and the collaboration with a team of biologists working specifically on *Platynereis dumerillii*'s brain, a manual thresholding approach was developed. Indeed, by going through the 169 genes one by one, it was possible to adjust the thresholds manually until the resulting binarized expression pattern corresponded perfectly to 1) the fluorescent stack images from in-situ hybridization data; 2) the biologically known expression patterns in the brain of *P. dumerillii* expected by the biologists.

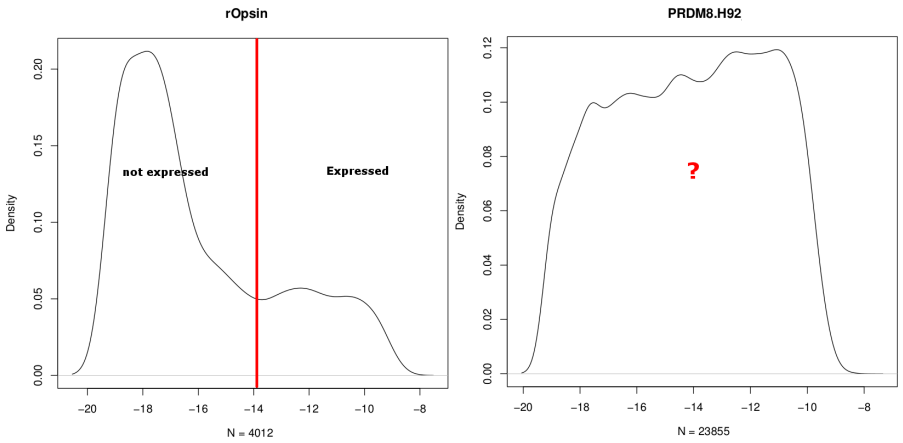
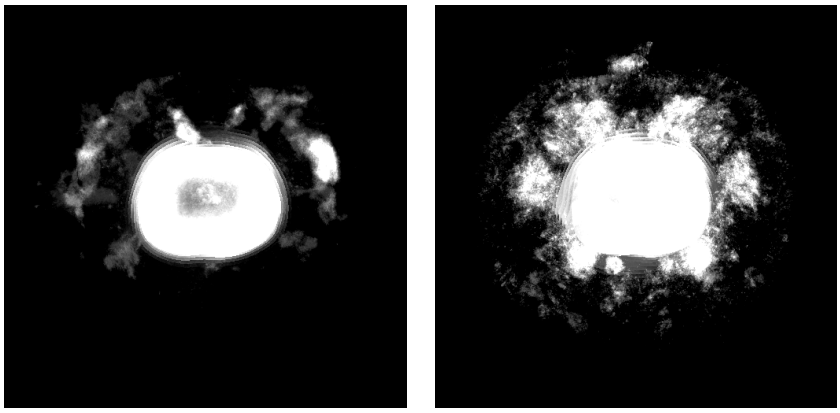


Figure 11: **Densities of log luminescence values for two genes (rOpsin, PRDM8) over the** 32,302 **cells.** For *rOpsin*, the density exhibits two clear peaks making the choice of a binarizing threshold easy. By contrast, for *PRDM8* there is no such clear threshold, making an automated binarization method hard to implement.

This method resulted in a high confidence binarized dataset for 86 genes. Several reasons explain why 83 genes out of the starting 169 were removed from the dataset. For some of the genes no good threshold could be found, this was due to high noise level in the in-situ hybridization images. Other images suffered from experimental errors that yielded blurred and unexploitable expression patterns. Finally some images were polluted by a well known experimental artefact linked to confocal microscopy imaging as shown in Figure 12.



(a) Z-stack of gene Smad23 showing an experimental artefact (round very bright spot in the middle).

(b) Z-stack of gene Syt showing an experimental artefact (round very bright spot in the middle).

Figure 12: Confocal microscopy experimental artefact for 2 genes of the original 169 studied genes.

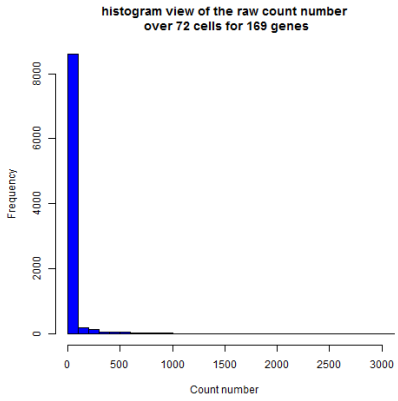
Although the aforementioned method resulted in a high quality binary dataset, it has been possible only because the number of genes considered was small. This will not be the case when dealing with RNA-seq data.

2.4.2 *Binarizing whole transcriptomes*

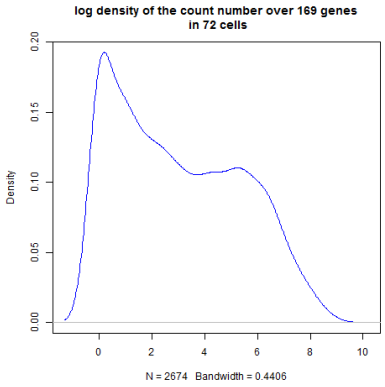
When dealing with whole transcriptomes, manually finding thresholds to binarize gene expression data is no longer a valid option due to the high number of genes considered. An automated method is thus required. I will discuss possible ways to binarize single cell RNA-seq data, presenting some results from a small number (72) of sequenced cells from the brain of *P. dumerillii*, where the expression levels of the 169 PrimR genes was recorded (see next section 2.5 for a detailed presentation of these data).

A naive approach would be to simply consider that as long as one RNA fragment mapped to a particular gene has been found in a cell, the gene is considered as expressed. Although such a method would be justifiable in the case of a perfect dataset, with no noise or errors, as discussed above 2.3 in the case of single cell RNA-seq the biases created by the mRNA capture rate are to high to rely simply on this method. Indeed, as a first approach on *P. dumerillii*'s dataset, we can see on Figure 13a that the value 0 represents a very dominant peak.

The problem in that case is that for read counts of 0 it is dangerous to consider the gene as non expressed when it could be lowly expressed.



(a) Histogram showing the frequencies of count values over 72 sequenced cells, with the fragments mapped to 169 genes.



(b) Density plot for the count values over 0 in the single cell sequencing dataset

Figure 13: Thresholding RNA sequencing data for *P. dumerillii*

Another option would be to find a global threshold over the complete dataset. The threshold $T > 0$ would represent the count number of reads for a particular gene and a particular cell needed to consider the gene as expressed. T could be inferred from the count density over all the genes and all the cells. The expected result would be a 2 peaks density with one peak corresponding to the non expressed count values, the second to expressed genes. The binary threshold would then be set between the first and second peak. Although more precise than the previous method, binarizing in such a manner may lead to numerous errors. Indeed, the underlying assumption behind this method is that all genes behave in a similar way. As Figure 13b shows, if a 2 peak behaviour is indeed present, the cut is not extremely clear and an important portion of count numbers actually fall in between the two peaks. This is due to the fact that all expressed genes are not expressed in the same way; some are expressed lowly some highly, which has a tendency to flatten the density curve making this thresholding method, if better, still not 100% reliable.

The more suitable approach to this thresholding problem would be to compute one threshold per gene based on the density curve for every gene across all cells. However, with 72 cells into consideration, considering the sparse nature of the count data, no significant results

can be extracted with this method on this particular dataset. However, I believe that one threshold per gene would prove a big improvement over the previously mentioned thresholding methods providing sufficient number of data points per gene.

2.5 PRELIMINARY RESULTS ON SINGLE CELL RNA-SEQ SPATIAL BACK-MAPPING

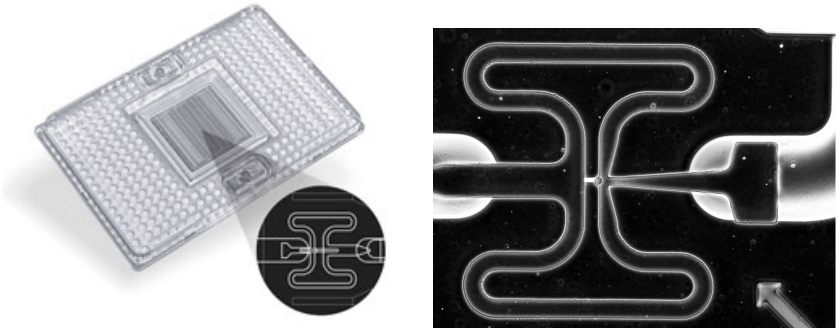
2.5.1 *Single cell RNA-seq in Platynereis' brain*

A collaborations with Kaia Achim in the Arendt lab in EMBL provided us with a unique RNA-seq dataset of 72 single cells from *P. dumerillii*'s 48hpf developing brain.

Experimentally, the work consisted in setting up *P. dumerillii* batches, picking up 50-100 individuals at 48hpf. These were washed in Ca-Mg free sea water and incubated in a mixture of pronase which breaks extracellular matrix and thioglycolate (helps to break the chorion). After this treatment, the trunks and epispheres (brains) were separated. 40-60 epispheres were then picked out, transferred to Phosphate buffered saline (PBS) and then incubated for 1 minute in PBS containing collagenase to break more extracellular matrix. After two PBS washes, the cells were dissociated by pipetting up and down then washed again in 1 ml of PBS and concentrated by centrifuging (1 min, 1000 rpm). Cells were re-suspended in 20 microliters of PBS, of which 5 microliters could be loaded on the capture chip.

Fluidigm's C1 Single-Cell Auto Prep System instrument with the Fluidigm Single-Cell Auto Prep IFC chip optimized for 10-17 micron cells were used as shown in Figure 14. The reverse transcription was performed using Clontech SMARTer Ultra Low Input RNA Kit and for on-chip PCR the Clontech ADVANTAGE-2 PCR kit. Sequencing libraries were prepared using Nextera DNA Sample Preparation kit from Illumina.

With two chips and a capture rate of 65%, 72 libraries were sequenced including 11 cells from first chip, 35 live single cells, 17 dead single cells, 3 wells containing 2 cells, one with 4 cells, and 3 unsure ones from the second chip resulting in 72 raw reads files as shown in Table 1.



(a) Fluidigm C1 chip with 96 wells. Image taken from Fluidigm website

(b) One cell from *P. dumerillii*'s brain captured in the chip. Image generated by Kaia Achim

Figure 14: Microfluidics single cell sequencing with C1 chips.

Chips used	2
Capture rate	65%
Libraries sequenced	72
Live single cells	38
Dead single cells	15
Debris + live single cell	4
Multiple cells	6
Debris/unsure	9

Table 1: Results over two C1 chips. The experiments were conducted by Kaia Achim.

Of course those results do not include the spatial localization of the cells as the protocol requires the separation of the coherent tissue into a cell suspension. As a crucial point in any dowstream analysis, being able to map back the single cells to their original location in the brain is required. To that end, I took advantage of the spatially localized in-situ hybridization described in the previous section.

2.5.2 *Mapping RNA-seq data back to PrimR in-situ hybridization assays*

Firstly, the RNA-seq raw reads were mapped to the 86 reference genes composing the in-situ hybridization data using Bowtie. The resulting dataset comprises the number of reads mapped back to each of the 86 genes in the 72 cells sequenced. In order to map back to the in-situ

hybridization data, the chosen approach consisted of extracting the genes that were the most specifically expressed for each sequenced cell, before comparing this specific fingerprint to the in-situ 3D data in order to isolate the regions of the brain where those specific genes are co-expressed.

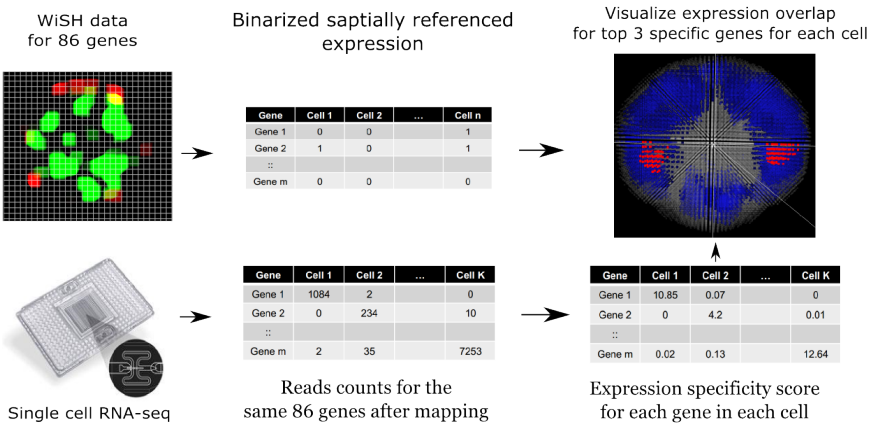


Figure 15: **Spatial back mapping of single cell RNA-seq.** The binarized in-situ hybridization data provides the spatial reference onto which the single cell results will be mapped. All the single cells sequenced are put together and a expression specificity score for each gene and each cell is computed. Tha mapping is realized using the top 3 or 4 most specific genes for each cell.

Given the set of 86 considered genes M, and the set of 72 cells C, with the read count matrix D of size $M \times C$, the expression specificity ratio $r_{m,c}$ can computed for each cell and each gene as :

$$r_{m,c} = \frac{D_{m,c}}{\frac{1}{\|C\|}\sum_{a \in C} D_{m,a}}$$

where $\|C\| = 72$ is the number of cells considered. Subsequently, for each cell, the genes with the highest specificity scores can be determined. On the one hand, this mapping method has the inconvenience of using the average expression level across all considered cells to compute the ratio r. This means that the ability to precisely infer the original location of each cell, in other words, the mapping quality, will depend on the overall sequencing quality. Furthermore, this method's performance relies on the assumption that the data are in fact a collection of cells from different cell types. However, given the experimental protocol described above, this seems to be an acceptable hypothesis. On the other hand, this mapping method has the advantage of not being sensible to technical noise in the RNA-seq protocol,

providing the technical noise between cells remains at a constant level. This justifies the use of the read counts in a quantitative way and not a binarized dataset.

The goals of this study were to validate the protocol used in order to obtain single cell RNA-seq results in *P. dumerillii*'s brain and to establish a methodological proof-of-concept on spatially mapping RNA-seq results onto in-situ hybridization data. I will present here a few examples of sequenced cells, their most specifically expressed genes and their resulting potential original location in the brain as well as the probable cell type they belong to.

In table 2 are shown the most specific genes for four of the sequenced cells. For each cell, this list of genes can be used to visualize the areas within the brain where they are co-expressed according to the in-situ hybridization data. A snapshot of this visualization is shown on Figure 16. In every case, simply looking at the three most representative genes seems to allow a clear localization of the sequenced cells. Of course this mapping is not at the single cell level, but having an idea of the tissue every cell originated from is already a nice proof-of-concept.

From the most specific genes to each cell and their potential localization, it is possible to hypothesize, using previous biological studies, the cell type of each sequenced cell. As shown in Table 2, for cell "X2C911L" the most specifically expressed gene "Emx" has been used as a reference gene to localize the mushroom bodies, a hypothesis which is compatible with the co-expression of "CALM.R29" and "Dach" [104]. Cell "X2C521L" expresses Wnt8 very specifically, a gene shown to be linked to lateral brain development. Cell "X2C61L" can be easily classified as a developing neuron. Indeed both VACht (Vesicular acetylcholine transporter) and ChaT (Choline acetyltransferase) are genes coding for enzymes interacting with the neurotransmitter acetylcholine. Finally cell "X2C241S" displays the specific expression of the gene "Mitf", one of *P. dumerillii* most studied gene and expressed solely in the developing adult eyes [60, 43].

Overall, the results of this back-mapping method look very promising. Although at the time of writing, this work is still in progress in collaboration with Kaia Achim and Detlev Arendt, I have been able so far to map-back 27 of the 38 live single cells to a defined area of overlapping expression in the brain using the top 3 or 4 most specific

x2c911l	x2c521l	x2c61l	x2c241s
Emx	Wnt8	VAChT	Mitf
CALM.R29	HEN1-Y61	ChaT	Otx
Dach	Gsx	LYamide	Tolloid-Y68
Mushroom body	Developing lateral brain	Differentiated neural tissue	Adult eye

Table 2: Top 3 most specific genes for 4 sequenced cells and the potential tissue they belong to. The resulting localization of those four cells infered from the in-situ hybridization data are shown in Figure 16.

genes.

I decided to check the probability of obtaining such a result. To this end, I randomly generated 1000 "top 3" genes out of the the 86 genes considered in the in-situ hybridization data and computed the corresponding number of "cubes" where there all 3 genes were expressed at the same time. I then compared this "by chance" number of overlapping cubes to the number of overlapping cubes with the top 3 genes generated from the 38 live single cells. The results are shown in Figure 17 and the results clearly comfort or approach.

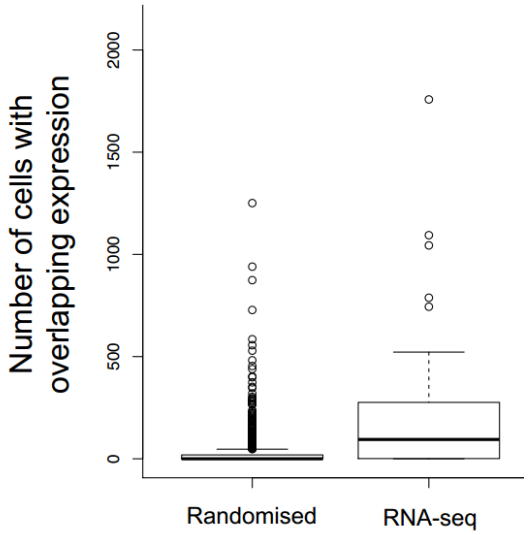


Figure 17: **Back-mapping by chance vs RNA-seq data.** This boxplot shows the number of "cubes" in which the top 3 genes overlap with randomly generated top 3 genes (left N=1000) and the top 3 genes obtained from the live single cells RNA-seq data(N=38).

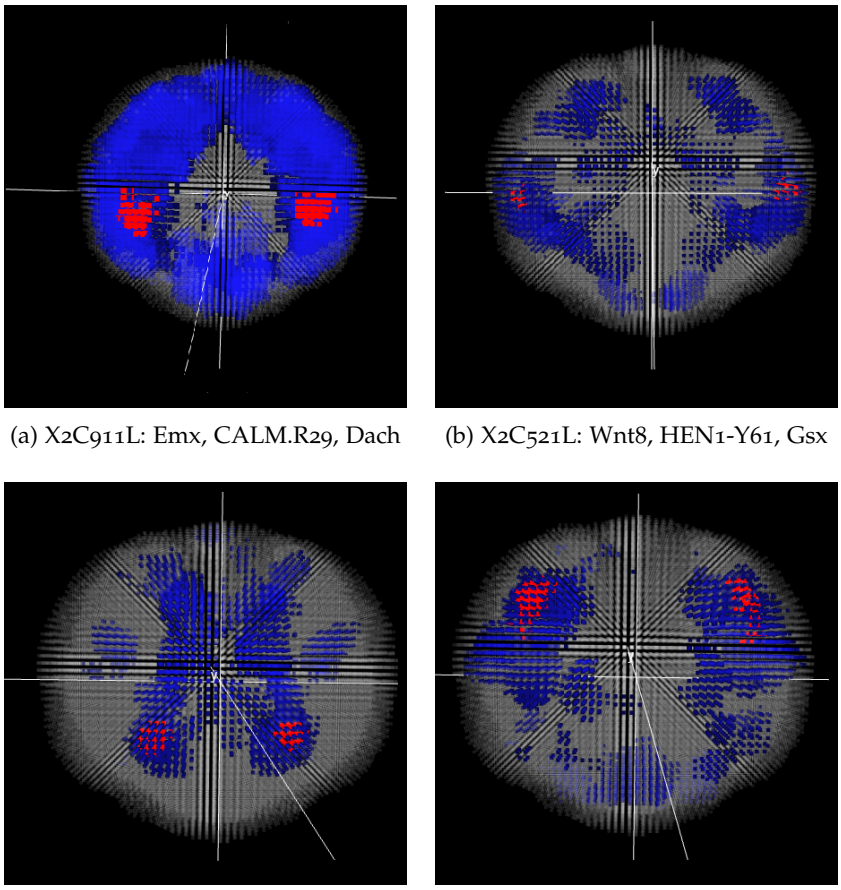
2.6 CONCLUSIONS

In this Chapter I described how the scale of gene expression studies has shifted from the tissue level to the single cell level. For two experimental protocols the generation of such data was presented and I explained why at the time of writing, it is still not safe to assume that single cell trascritomics datasets are quantitative. To avoid that problem, turning those datasets into binary gene expression is an attractive solution. However the binarization process is not trivial and I have presented ways to obtain a high confidence dataset.

One big advantage of in-situ hybridization assays is the fact that the spatial information stays attached to each gene expression fingerprint. Using this information, it was possible to allocate cells assayed via single cell RNA-seq to their spatial location in the brain using the most specifically expressed genes in every cell. To the best of my knowledge, an a posteriori localization of single cell RNA-seq data has never been presented before.

As mentioned previously in the Introduction (1), if the ability to study the heterogeneity of cell populations at the single cell level offers incredible possibilities for the future of developmental biology and potentially of cancer research, the development of new statistical methods adapted to this single cell scale, allowing conclusions to be drawn at the tissue level is crucial.

The work presented hereafter was done to answer simple but important questions: can known functional tissues of a complex organ like the brain be defined and localized from single gene expression data? Can unknown regions in such a complex tissue be detected and finally, is it possible to hypothesize the functional role of those unknown regions based on single cell expression data?



(a) X2C911L: Emx, CALM.R29, Dach

(b) X2C521L: Wnt8, HEN1-Y61, Gsx

(c) X2C61L: VAChT, ChaT, LYamide

(d) X2C241S: Mitf, Otx, Tolloid-Y68

Figure 16: Regions defined by the expression overlap of the top 3 scoring genes in [104] binarized in-situ hybridization data. The red colour shows the co-expression of the three considered genes, the blue areas are those where one or more of the three considrered genes are expressed but not all, in grey are the areas where none of the considered genes are expressed. The 4 figures are from a apical view with the dorsal side on top.

3

VISUALIZING TISSUES FROM 3D SINGLE CELL EXPRESSION DATA

3.1 ELEMENTS OF CLUSTERING FOR BIOLOGICAL TISSUES

3.1.1 *Motivations*

Following the conclusions of Chapter 2 in which I discussed the shift from the tissue to the single cell scale of study in developmental biological assays as well as the main challenges to analyse and interpret such data, the first question that seems natural to answer is the following: given single cell gene expression data only, is it possible to classify cells that are the most alike together and define the organization of complex biological tissues like the brain of *Platynereis dumerillii*?

This is fundamentally a clustering or a classification problem. Indeed, small units need to be put in a determined number of classes or clusters because they are alike in one way or another. As anatomical and functional information about some tissues in the brain of *P. dumerillii* is already available (see Chapter 1), the obvious validation for any clustering/classification method developed is to check that the single cell level information leads to the definition of these known tissue. Once this has been validated, it seems important to determine whether the single cell expression data *adds* to the known biology by redefining (subdividing for example) known tissues or finding new ones. Identifying putative cell type is already a major challenge but the question can be pushed further. As mentioned in the Introduction 1, because gene expression is a key process in a cell's life, studying the genes that characterize a cluster can provide insights into the functional role of the cells it contains.

3.1.2 *General considerations about clustering*

I mentioned before that methods developed to answer those questions could either be clustering or classification methods. In the field

of machine learning, these two notions are fundamentally divergent. Clustering describes a method that assigns points to an unknown number (*a priori*) of sets in an undirected way, while classification takes advantage of an already known classifier it assign in a directed manner, new elements to a determined number of clusters.

In my case, the number of tissues in the brain of *P. dumerillii* is unknown and there is no previous classification or a strong enough biological knowledge of each and every cell to opt for a directed classifier. Therefore the methods presented hereafter will be clustering methods with respect to the machine learning definition of the word.

As a general consideration about clustering, it is interesting to note that unless working on simple enough datasets, there is no perfect method. This is especially true when dealing with biological data, where the complexity and the noise level (see 2.3) tend to be extremely high.

To illustrate this notion, for the single cell expression data in *P. dumerillii*'s brain, the question of finding the "true" number of tissues is extremely complicated. Without any prior knowledge, the statistical methods to determine the number of clusters presented in this thesis will yield indications about what the optimal number of "tissues" is given the data and the model, which does not necessarily means that this number is biologically *true*.

For a complex dataset such as brain tissue, a crucial matter in the upstream and the downstream analysis is to be able to visualize the data. This is especially true when the considered tissue is in 3 dimensions. For example, an important part of the upstream analysis to clustering is to visualize the gene expression patterns in the brain in order to find the right binarization thresholds. In downstream analysis, after developing a clustering method, seeing the resulting clusters and their localizations in the brain is also very important. Therefore, I developed a tool for 3D dataset visualization taking advantage of the latest developments in browser based technologies to create the software "bioWeb3D". This work has been published in [84].

3.2 VISUALIZING CLUSTERING RESULTS IN 3D WITH BIOWEB3D

3.2.1 *Background*

Visualisation is a key challenge in the analysis of large biological datasets, especially when analysing organized structures with distinct sub-clusters [91]. This is particularly important when analysing 3-Dimensional (3D) datasets. When a biological process or feature has been described spatially by a set of 3D referenced points, either via laboratory work (confocal microscopy for example) or generated within a simulation, with some data attached to each point in space, the first step in interpreting the data is to visualise it. Once the data are visualised and their quality assessed, downstream analysis can proceed. For example, a typical second step is to cluster the observations into different classes based upon the information associated with each point; those results will also need visualisation.

While various 3D visualisation tools have been developed, they have typically been made available via a locally installed piece of software such as BioLayout Express^{3D}[40], Arena3D [82], 3D Genome Tuner [106], Amira 3D [96], V3D [83], the Allen Brain Atlas [65] or Cytoscape [92]. These tools are very complete and usually complex to operate for non-expert users. Moreover, they require installation on every machine they are used on, which makes sharing inconvenient. To address this issue, some 3D visualisation tools have been built online and are accessible through the browser directly, such as AstexViewer [46], which is utilised by the Protein Databank Europe via a Java Applet. More recently, visualisation tools developed using HTML5/WebGL capabilities have been described, although they have focused on very specific applications, such as analysing radiology data [62].

Importantly, before bioWeb3D [84], no tool has allowed biologists to view their own 3D data directly online in an easy, fast, interactive and secure way. Using WebGL and the JavaScript 3D library Three.js, bioWeb3D aims to be a simple, generic, tool for tackling this problem.

3.2.2 *Implementation*

bioWeb3D allows the user to represent any 3D dataset on their browser by defining only two files. The two files can either be formatted as JSON or XML files, two widely used structured formats on the web [109] [7], or directly as Comma Separated Values files (CSV).

The first file used by the application, referred to as the *dataset file*, contains the spatial coordinates of every point in the dataset. The second type of file used, the *information layer* file, describes one or several information layers that are associated with every point defined in the first file. For example, if each point defines the location of a cell within a tissue, the second file could describe whether a particular gene is expressed in each cell. That way the tissue expression profile can be represented in the spatial context of the tissue.

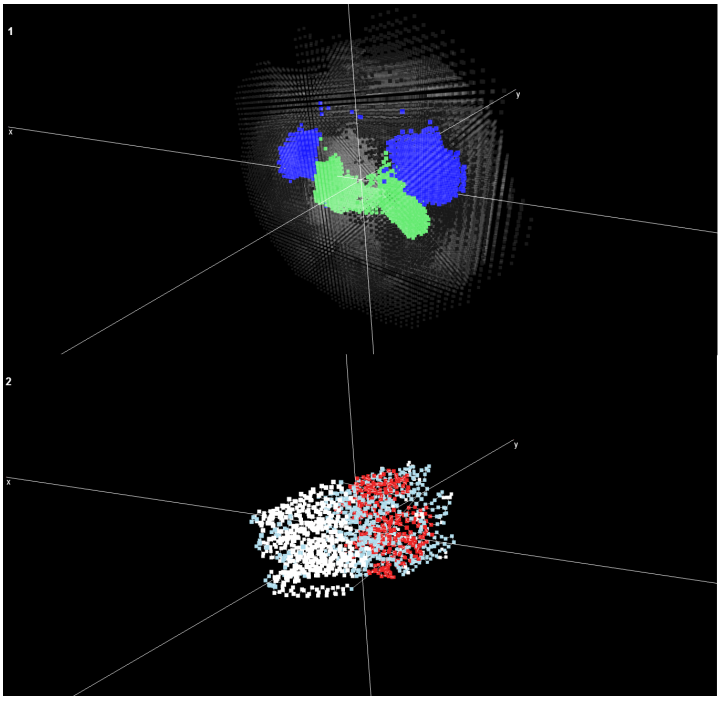


Figure 18: bioWeb3D allows several datasets to be visualized at the same time in up to 4 different “worlds"

Datasets can be viewed and compared in up to four “worlds" (each world refers to a separate visualisation sub-window) at the same time (see Figure 18). Although browser based, the application, fully written in Javascript, does not need to send any data to the host server.

Instead, the modern internet browser's local file system reading capabilities are used through the HTML 5 FileReader functionality. This allows the application to handle, in a very short period of time, large datasets while ensuring that the privacy of the data is maintained. Although the focus is on making bioWeb3D simple and easy to use, some options are available to customise how datasets are represented. The application can be used to visualise sequential information, such as 3D protein structures, in which case a solid line can be drawn between the points (Figure 18 (right)). In other situations, such as when a population of cells is considered, the points are viewed as individual particles. The information layers are visualised by colouring the 3D points according to the class that each point belongs to.

bioWeb3D is fully written in HTML/Javascript. It relies heavily upon a relatively recent 3D javascript library called Three.js [5]. This library is used as the main interface between WebGL (cross-platform, royalty-free web standard for a low-level 3D graphics API) [6] and javascript. More specifically, bioWeb3D allows the generation and manipulation of simple Three.js objects. Indeed the primary challenge associated with the creation of bioWeb3D has been to design interactions between the 3D visualisation and the user interface in the most efficient way.

The 3D data are rendered using simple 2D quadrilaterals positioned in the 3D space according to their coordinates. This simple technique has been selected to keep bioWeb3D as light-weight as possible whilst ensuring good quality visualisation performance and fluidity.

JSON is the recommended format to input files into bioWeb3D because of its rigorous structure and its fast object generation, which is directly built into all of the primary internet browsers' interpreter. Compared to other data-interchange languages, such as XML, JSON is also easily human readable thanks to a light-weight syntax.

However, some applications might output data only in an XML format and not JSON, as the latter is generally more web oriented. For this reason bioWeb3D can also accept XML as an input format.

Furthermore, much data generated in the biological sciences is stored within CSV files. Converting CSV documents to the JSON or

XML format is not always trivial. In order to facilitate this process, the application is also able to directly render simple CSV files that follow a certain format as an input. The file formats to input data into bioWeb3D are described with examples in Appendix A.

3.2.3 *Results*

The goal of bioWeb3D is to allow scientists unfamiliar with visualisation software to explore 3D data very quickly without having to install any software. To illustrate its utility I used bioWeb3D to visualize some preliminary results within the single cell gene expression data of *P. dumerillii*'s brain. In the context of bioWeb3D, the locations of the "cubes" are used to generate the "Dataset" file and information about the sets of cells that define clusters with similar gene expression profiles are used to generate the "Information Layer" file. In Figure 19 the results are illustrated — each point represents a pseudo-cell and its colour indicates the class (or cluster) to which it belongs to, here only two clusters are highlighted.

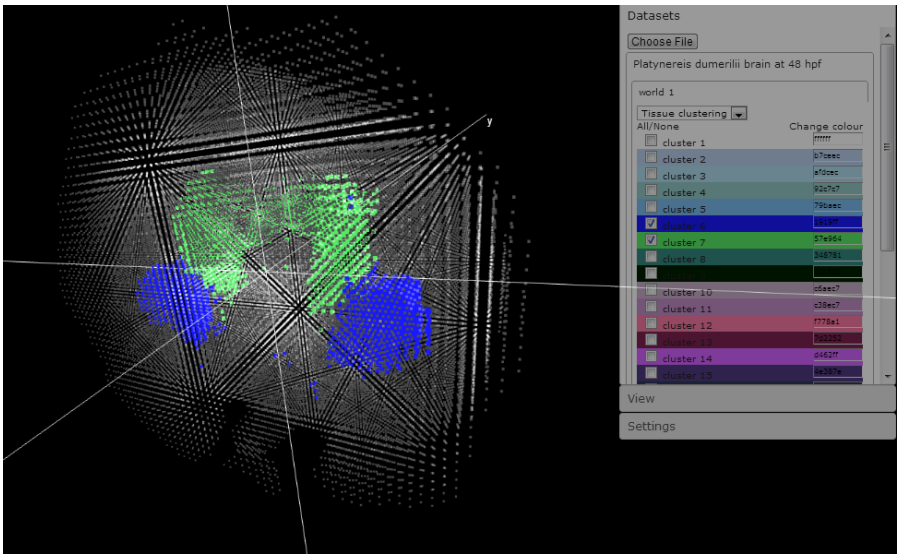


Figure 19: The 3D location of cells within the brain of the marine annelid *Platynereis dumerillii* is shown. Two classes are displayed (in green and blue) along with the shadow of the remaining cells. The User interface is visible on the right of the screen and can be hidden. See 1 for a presentation of *P. dumerillii* and Chapter 2 for detailed presentation of the data.

bioWeb3D can be used to visualise datasets derived from a wide variety of biological assays. Examples are shown on the Github wiki [3], where a 3D representation of a Principal Component Analysis

(PCA) carried out with R and the 3D structure of a protein extracted from the PDBe database are displayed.

More generally, the user can interact with the visualisation via an interface on the right of the screen, which contains three panels as shown on Figure 20. In the “dataset" panel, the user can choose the *datasets* and *information layer* files that should be represented in each world. This panel also allows the user to show/hide specific classes of the selected information layers. Each dataset file entered will create a new sub-panel where the user can input *information layer* files for that world. Selecting an *information layer* in the drop-down list will display the data in the current world and generate a list of classes that the user can modify regarding their visibility and colour. The “View" panel enables the user to choose which of the worlds are shown on the screen, ranging from 1 to 4. Finally, the “Settings" panel provides the user with a number of options that affect all worlds and all datasets, such as modifying the axes scales, modifiying the transparency and size of raw data points and information layer coloured points. The user can also choose to enable centering of the data around 0 or leave the coordinates as inputted.

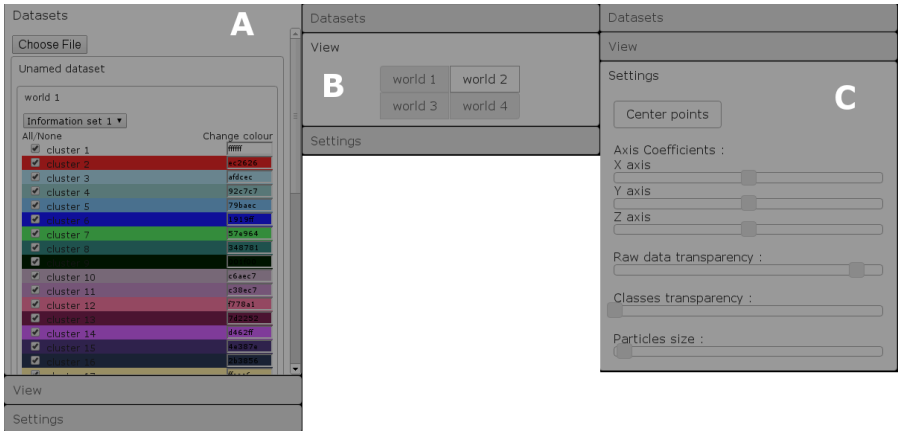


Figure 20: The three control panels to control visualization in bioWeb3D. A: the datasets panel, where new datasets and new information layer files can be inputted. From the dataset panel, information layers can be selected and unselected, and datasets attached or detached from the “worlds". B: the view panel where the user can choose the worlds to display. C: the settings panel where the user can customize visualization.

3.2.4 *Discussion*

Many 3D visualisation software tools, most of which require local installation, exist and provide similar functionalities with standard 3D format input such as Wavefront .OBJ. Some are extremely generic and powerful like Blender or Amira 3D. However, these tools are not typically oriented towards a scientific audience. Moreover, those that are more focused on science are often targeted towards a very specific application, especially in the medical sciences [106]. In this context, I believe that bioWeb3D can be useful as it is completely generic and browser based. It should also be noted that recent browser improvements regarding GPU acceleration through the WebGL paradigm allow bioWeb3D to visualise several hundred thousand points. Additionally, local software is usually platform specific, which is not the case for browser based applications.

As mentioned previously, browser based 3D visualisation tools currently exist mainly in the form of Java Applets. This technology has attracted much criticism in 2012 regarding security flaws, leading the "United States Computer Emergency Readiness Team" to advise that all Java Applets should be disabled due to current and future Java vulnerabilities [4]. The development of WebGL technology is viewed by many as a candidate for replacing Applets.

The main current limitation of a WebGL based application is the machine and browser compatibility. Only computers with fairly recent graphic cards will be able to run a 3D environment. It should also be noted that Microsoft has notified the developer community that Internet Explorer is not scheduled to support WebGL in the near future. However, importantly, Chrome, Firefox, Safari and Opera all now support WebGL applications. Moreover, WebGL is also supported on mobile platforms such as iOS or Android [2].

As a fully open source software, the source code for bioWeb3D is available on Github [3], a web platform that allows interested parties to collaborate on the development of the project. In the wiki page "Contribute to bioWeb3D", directions to alter or add capabilities to bioWeb3D are provided for users who wish to get involved.

3.2.5 *Conclusions*

bioWeb3D is designed to be a simple and quick way to view 3D data with a specific focus on biological applications. Being browser-based, the software can be easily used from any computer without the need to install a piece of software. Importantly, bioWeb3D has been designed to offer a very straightforward and easy-to-use working environment. Despite current limitations in terms of compatibility or rendering performance for large numbers of points, I believe that bioWeb3D will enable non-experts in 3D data representation to quickly visualise their data and the information attached to it in many biological contexts, thus facilitating downstream analyses.

3.2.6 *Availability and requirements*

The full source code is available on the Github page of the project [3]. A live version of the software is online [1]. You will require a graphical card and a browser with WebGL capabilities to run bioWeb3D.

3.3 NON SPATIAL CLUSTERING METHODS

Being able to visualize clustering results will be key in analysing any method's output from a biological perspective. Of course the next step is to actually develop a clustering method that would be able to cluster binarized single cell gene expression data. I will briefly describe in the next paragraphs existing methods that are able to cluster single cells together based solely on gene expression patterns without considering the spatial structure of the data.

3.3.1 *Hierarchical clustering*

The first method I took an interest in was hierarchical clustering (hClust) [57]. Indeed, in the field of molecular biology and biology in general, this clustering method is extremely popular mainly because it is relatively straight forward to use, and because the obtained dendrogram helps the downstream analysis of the data.

hClust relies on the computation of a distance matrix. TO calculate this I considered the matrix D of the in-situ hybridization data, with 86 columns corresponding to the 86 genes considered and 32, 203

rows corresponding to every cell in the dataset. The computation of the distance matrix was performed using the *dist* function in R with the *euclidean* metric (or *Manhattan* which is equivalent for a binary dataset) in order to compute the $32,203 \times 32,203$ matrix of distances between rows.

It is interesting to note that this step, in addition to being computationally expensive, creates a very large object in memory making it a limiting factor for very large datasets. Based on this distance matrix, the hierarchical clustering can take place using the "hclust" function in R. With the "complete" option turned on, the resulting object will be a dendrogram representing the hierarchical classification of all $32,203$ cells as shown in Figure 21.

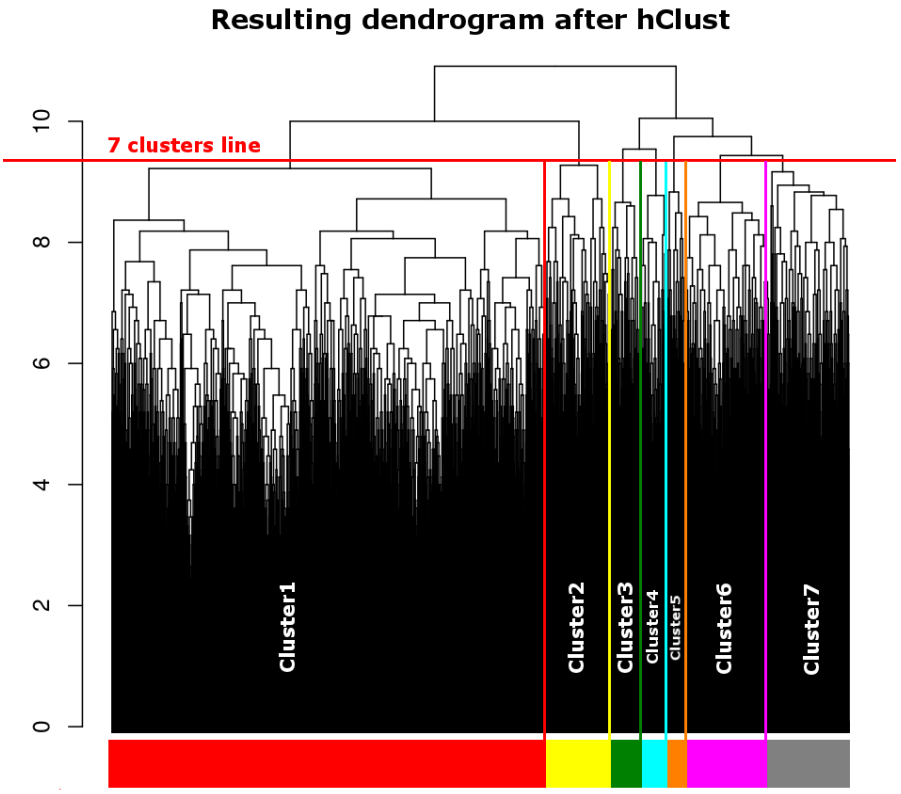


Figure 21: Dendrogram outputted by the hClust clustering method. This is the full dendrogram, to get the resulting clustering partition, the tree must be cut. The red line shows the cut needed to get 7 clusters out of the dendrogram. Importantly, a dendrogram does not provide any rationale about the best number of clusters.

Importantly, once this dendrogram is outputted, the hClust method gives no indication on where to "cut" the dendrogram to output the actual clustering results. This highlights a very important issue in all clustering methods, namely identifying the optimal number of clusters needed for a certain dataset when there is no prior information. For the rest of this thesis this crucial parameter will be referred to as K.

3.3.2 *Other clustering methods adapted to gene expression data*

There are several other clustering methods that could be applied in order to create a partition of gene expression data. They include but are not limited to K-means clustering [67] and independent mixture models [31]. Those methods are well suited to cluster gene expression data. However, they all fail to take into account the spatial information linked to the gene expression dataset.

3.4 DISCUSSION

3.4.1 *Spatial clustering techniques*

Using a clustering method that would in addition to the gene expression pattern of each cell, take into account its spatial localization, that is the context of each cell regarding the other cells around it, could theoretically improve the clustering performances for several reasons.

As mentioned in the introduction 1, many of the underlying processes implicated in tissue development the best example of which is the process of asymmetrical cell divisions, lead to highly structured spatial organization. In this context and without any other prior knowledge, not taking into account the spatial localization of each cell into the clustering scheme would seem to be an unexploited potentially important information.

Furthermore, as discussed in Chapter 2, single cell gene expression datasets whether they are generated from in-situ hybridization (see details on Figure 8) or single cell RNA-seq (Figure 10), are prone to errors and incoherency. A clustering method that would be able to compensate "erroneous" data points by taking into account the spa-

tial context of each cell, could potentially dramatically decrease the effect of noise level upon the clustering results. Additionally, from a downstream analysis perspective, as far as general hypothesis about clusters are concerned, "smoother"/less scattered clusters are easier to interpret from a direct visualization of the results.

3.4.2 *Hidden Markov random fields for clustering*

In order to utilise both the spatial and the gene expression information, it was decided to extend a graph theoretical approach developed for image segmentation to reconstruct noisy or blurred images [29], a method that finds its roots in the field of statistical mechanics as the Ising model [51] and its generalization, the Potts model [111]. The core concept of this method is to use an Expectation-Maximization (EM) procedure to estimate the parameters of a Markov Random Field based model using mean-field approximations to estimate intractable values as described in [24].

This approach exhibits several important advantages as will be described in detail in the next sections. Indeed, in addition to providing a way to take into account the spatial information in the clustering results, it also offers some nice features in terms of downstream analysis through the analysis of the optimal parameters upon convergence.

In the next three sections, I will first describe the theoretical framework underlying this model followed by an assessment of the method's performances compared to other non spatial clustering method on simulated data. Finally I will show and analyse the clustering results obtained through the developed method on the single cell in-situ gene expression data in *P. dumerillii*'s brain described in Chapter 2.

4

HIDDEN MARKOV RANDOM FIELDS FOR BIOLOGICAL DATA CLUSTERING

This Chapter gives a theoretical overview of a Hidden Markov Random Field based approach that is designed to cluster single cell in-situ hybridization gene expression data "cubes" as described in Chapter 2, into K clusters ($K \in 2, 3, ...$). Subsequently, we will describe our approach for estimating K.

4.1 MARKOV RANDOM FIELDS

4.1.1 *Neighbourhood systems*

Let S be a finite set of sites, each of which represents one "cube" of data. Given the 3D coordinates of each site, the first challenge that needs to be overcome in order to use the spatial characteristics of the data in the clustering scheme is to express the data and their spatial relationship in mathematically formal manner. To this end, starting from the spatial coordinates in 3D of each "cube", instead of a list of isolated measurements, it is possible to build a connecting graph representing the same data and the spatial dependence between the "cubes". In the context of this study, each node of the graph will represent a "cube" in the single cell expression data. Nodes that are linked together by an edge will be spatially dependent upon each other.

With prior biological data, one can manually create the spatial dependency graph by linking nodes together that are known to be functionally similar. In the case of this study however, no such prior knowledge being available, it is necessary to define the spatial dependences in a different way.

The central hypothesis while developing this method is to assume that "cubes" that are close to one another are more likely to belong to the same cell type (i.e cluster). Consequently, these spatial dependencies will be incorporated into a *neighbourhood graph* where "cubes" close to each other will be joined.

In the case of this study, because of the cell model used the graph will be a regular grid. In this context, there are several ways to translate the spatial relationship into a neighbourhood graphs depending on the number of neighbours considered for each site. As shown in Figure 22, the choice between a first or a second order neighbourhood system is purely technical. However, having more neighbours for each site will increase the complexity and ultimately the computational burden. Indeed, in G, a *clique* c is a subset of nodes that are all interconnected, i.e it is possible to go from any nodes in c to any other node in c by simply following one single edge.

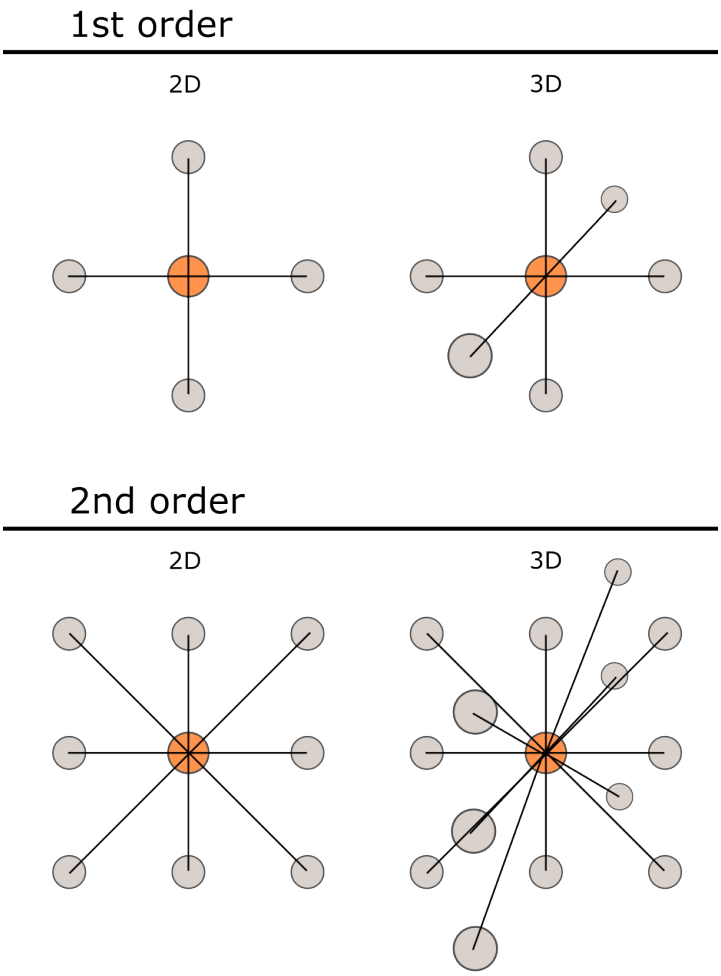


Figure 22: **First order and second order neighbourhood systems.** In the first order neighbourhood system, each site in the graph is linked to a maximum of 6 other sites in 3D while in the second order neighbourhood system each site can be linked to a maximum of 14 other sites. The Markov property on the graph implies that the state of any node (the orange one for example) can be fully determined by knowing the state of its neighbours (the grey ones).

Let C be the set of cliques of G. In a first order neighbourhood system, C is therefore the set of all sites alone and all the pairs of sites that are neighbours of one another. These are first and second order cliques, containing one or two sites. In a second order neighbourhood graph however, the set of all cliques in G also contains 3rd and 4th order cliques. Because the method implies iterating over the set of all cliques of the graph as I will detail in the next paragraphs, I decided to use a first order neighbourhood system to decrease the computational burden.

4.1.2 *Field distribution*

Let a Random Field Z be defined as a set of random variables $Z = \{Z_i, \forall i \in S\}$ where $Z_i \in 1, ..., K$. For every site $i \in S$, let $N(i)$ represent the set of its neighbours (see 4.1.1) and $z_{S-\{i\}}$ a realization of the field restricted to $S - \{i\} = \{j \in S, j \neq i\}$. Z is a *Markov Random Field* if and only if it follows the Markov property at every site :

$$\forall i \in S, P_G(z_i \mid z_{S-\{i\}}) = P_G(z_i \mid z_j, j \in N(i)) \tag{1}$$

Equation (1) states that the realization of the field, z_i at any site $i \in S$ can be fully determined using only the state of its neighbours $N(i)$. In other words the probability that a "cube" is in a given state depends only upon the state of its neighbours.

The Hammersley-Clifford theorem state that if Z is a Markov Random Field, the joint distribution of the field P_G follows a Gibbs distribution such that :

$$\begin{aligned} P_G(\boldsymbol{z}; \boldsymbol{\beta}) &= W(\boldsymbol{\beta})^{-1} \exp\left(-H(\boldsymbol{z}; \boldsymbol{\beta})\right) \\ &= \frac{e^{-H(\boldsymbol{z};\boldsymbol{\beta})}}{\sum_{z'} e^{-H(\boldsymbol{z};\boldsymbol{\beta})}} \end{aligned} \tag{2}$$

with $H(\boldsymbol{z})$ the Energy function summed over the cliques C of the graph G. Since we are working with a first order neighbourhood system, C is the set of all pairs of sites (i, j) that are neighbours. We chose to consider H as a function of vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ containing K parameters as detailed in the next paragraph, and $v_{i,j}$ a potential function set to 1 in our method because we are working on a regular grid graph, if the distances between sites were heterogeneous we could

have used this function to weight the spatial dependence between sites.

$$H(z) = -\sum_{i \in S} \beta_{z_i} \sum_{j \in N(i)} v_{i,j} \times \mathbf{1}_{[z_i = z_j]} \tag{3}$$

The denominator in (2) where z' represents all the possible realizations of the field is a normalizing constant referred to as $W(\boldsymbol{\beta})$.

4.1.3 *Single and multiple beta models in a biological context*

This model is closely related to a K-colour Potts model [111]. However, the unusual nature of the data used in this thesis led to the idea of extending the model commonly used in the field of image segmentation. In particular, the K-colour Potts model defines a single spatial coherency parameter β that is shared by all clusters [97, 116]. Importantly, the method presented here was extended by assigning one β per cluster so that:

$$\boldsymbol{\beta} = (\beta_1, ..., \beta_K)$$

Interestingly, equation (3) is an decreasing function of every component of $\boldsymbol{\beta}$. Indeed, for a particular cluster $h \in K$, a high value of β_h will accentuate the increase of the likelihood of the model through equation 2 when cluster h is spatially coherent. In other words, when a site has all its neighbours clustered in cluster h, classifying the site in cluster h as well -making cluster h spatially more coherent- will have an impact on the likelihood of the model proportional to the value of β_h. This Energy function thus favours spatially regular partitions and a higher value of β_h, with $1 \leqslant h \leqslant K$ will amplify the smoothing effect, or coherence over cluster h.

The reason why the model was extended to a multiple $\boldsymbol{\beta}$ parameter model, is inherent to the data used in this thesis. The first motivation is purely cytological. Indeed, in a biological context, it is expected that some tissues will be more spatially coherent than others. As mentioned in Chapter1 and visualized in Figure 3, tissues composed of different cell types may interact differently with their neighbours. For example, differentiated neural cells with long axons are likely to be in contact with numerous other cell types they go through.

The second motivation for the extended model finds its root in the cell model described in 2.1. Indeed, as described in Figure 8, some "cubes" may have inconsistent gene expression patterns. This sort of errors in the data will introduce spatial incoherency in the gene expression patterns. I also mentioned in Chapter 2, that the rate of errors linked to the experimental protocol may be dependent upon the cell type considered. Indeed, the errors described in Figure 8 are respectively more likely to arise in cell types with small and big cells. Therefore, I believe that allowing one spatial parameter for each cluster enables a better smoothing of these experimental errors by accounting for cell type specificity.

4.1.4 *Field parameters*

The field distribution contains K unknown parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)$ that have to be estimated by the model. It is important to note that $W(\boldsymbol{\beta})$ is summed over all possible realizations of the field Z, which is an exponentially complex sum as the cardinality of S rises. Therefore the computation of the normalizing factor becomes intractable very quickly. To address this problem, we are going to need to make some approximations in order to compute this quantity (see Mean Field Approximations).

4.2 THE EMISSION MODEL

We have described the field distribution of a Markov Random Field representing our graph, we now need to describe the relationship between Z and the data.

4.2.1 *Conditional independence in the observed data*

As Z is unknown a priori and represents the partition, let Y be a set of random variables representing the observations (the in-situ hybridization data). The model requires a conditional independence assumption with regard to the observations Y given the partition Z so that, with f_{z_i} the density function relative to cluster $z_i, i \in S$ (the realization of the field at node i):

$$\begin{aligned} p(\mathbf{y} \mid z; \Theta) &= \prod_{i \in S} p(y_i \mid z_i; \Theta) \\ &= \prod_{i \in S} f_{z_i}(y_i \mid z_i; \Theta) \end{aligned} \tag{4}$$

Equation 4 defines one unknown parameter per cluster: $\Theta = (\theta_1, ..., \theta_K)$. It is interesting to note that this part of the model is equivalent to an independent mixture model [72]. Indeed, Markov random fields can be viewed as independent mixture models where Z is a set of independent, identically distributed random variables, which happens when $\beta = 0$.

Given a particular cluster $h \in 1, ..., K$ and M the set of considered genes, the expression of each gene $m \in M$ in cluster h is modelled by a Bernoulli distribution with parameter $\theta_{h,m}$. This leads to one unknown Bernoulli parameter per gene per cluster so that :

$$\begin{aligned} \Theta &= (\theta_1, ..., \theta_K) \\ &= \begin{pmatrix} \theta_{1,1} & \dots & \theta_{1,K} \\ \vdots & \ddots & \vdots \\ \theta_{M,1} & \dots & \theta_{M,K} \end{pmatrix} \end{aligned}$$

4.2.2 *Full likelihood of the Hidden Markov random field model*

The conditional density function $f_i, i \in S$ can be expressed as :

$$\begin{aligned} f_i(y_i \mid z_i; \Theta) &= f_i(y_i \mid z_i; \theta_{z_i}) \\ &= \prod_{m \in M} \theta_{z_i,m}^{y_{i,m}} \times (1 - \theta_{z_i,m}^{1-y_{i,m}}) \end{aligned} \tag{5}$$

Looking at both fields Z and $Y \mid Z$ together, the complete likelihood of the model is expressed as :

$$\begin{aligned} P_G(\mathbf{y}, z \mid \Theta, \beta) &= f(\mathbf{y} \mid z, \Theta) P_G(z \mid \beta) \\ &= W(\beta)^{-1} \exp\{-H(z \mid \beta) + \sum_{i \in S} \log(f_i(y_i \mid z_i, \theta_{zi}))\} \end{aligned} \tag{6}$$

Because equation (6) is a Gibbs distribution, using the Hammersley-Clifford theorem we can conclude that the conditional field Z, Y is another a Markov Random Field with the Energy function

$$H(z.y \mid \beta, \Theta) = H(z \mid \beta) - \sum_{i \in S} \log(f_i(y_i \mid z_i, \Theta))$$

In our case, the goal is to recover the unknown realization of $Z : z$. To this end we need to maximize the values of all the parameters of the model $\psi = (\Theta, \beta)$, and to chose the optimal number of clusters, K.

4.3 PARAMETER ESTIMATION USING THE EM ALGORITHM

As mentioned before, the aim is to assign each cell i to one of the K possible clusters. To do so, it is interesting to consider the Maximum Posterior Marginal (MPM) that maximizes $P(Z_i = h|y, \psi)$, where the ψ are unknown and need to be estimated. To this end, the Expectation Maximisation [31] (EM) principle can be applied. The EM algorithm consists in the Expectation step (E) where the expectation of the model's likelihood with the current parameters is computed and the Maximization step (M) where the parameters values that maximize the model's expectation computed in the E step are found. The two steps are repeated until a convergence factor is reached.

4.3.1 *Initialization*

The first step of the algorithm is to initialize the model's parameter. To this end, it is possible to directly assign values for ψ^0, or to generate an initial clustering z^0 from which the initial ψ^0 will be derived.

I decided to use a combination of both approach, with arbitrary values assigned to β^0, typically 0, and the use of an initial clustering to compute the values of Θ^0. Indeed, for $h \in 1, ..., K, m \in M$, because of the emission model, each $\theta_{h,m}$ is the probability of gene m to be expressed in cluster h. Consequently, given a clustering z^0 with function $Expr_{h,m}$ the number of cells expressing gene m in cluster h and function Num_h the total number of cells in cluster h:

$$\theta^0_{m,h} = \frac{Expr_{h,m}}{Num_h}$$

4.3.2 *E step*

In the E step the parameters are fixed and the expectation of the model's likelihood $Q(\psi \mid \psi^l)$ at iteration $l > 0$ can be derived from equation 6 as:

$$Q(\psi \mid \psi^l) = \sum_z p(z \mid y; \psi^l) \log p(y, z; \psi)$$

Which can be further decomposed in :

$$Q(\psi \mid \psi^l) = \underbrace{\sum_z p(z \mid y; \psi^l) \log p(y \mid z; \Theta)}_{R_y(\Theta \mid \psi^l)} + \underbrace{\sum_z p(z \mid y; \psi^l) \log p(z \mid \beta)}_{R_z(\beta \mid \psi^l)} \tag{7}$$

Equation (7) allows to consider separately R_y and R_z.
R_y can be re-written using equation(4) as:

$$\begin{aligned} R_y(\Theta \mid \psi^l) &= \sum_z p(z \mid y; \psi^l) \sum_{i \in S} \log f_{z_i}(y_i; \Theta) \\ &= \sum_{i \in S} \sum_{h=1}^{K} [\log f_h(y_i; \Theta)] \; p(Z_i = h \mid y; \psi^l) \end{aligned}$$

Therefore, we know that in the M step I will need to compute the following probability:

$$t_{i\,h}^{l+1} = p(Z_i = h \mid y; \psi^l)$$

Computing this conditional probability is problematic because of the dependence between neighbouring "cubes", and computing an exact value is computationally expensive. Indeed, each point being dependent upon its neighbours, and the neighbours being themselves dependent upon their neighbours, unsurprisingly computing these conditional probabilities becomes exponentially complex as the number of connected nodes in the graph grow. Additionally for R_z, as mentioned previously, it is also necessary to compute the value of the normalizing constant $W(\beta)$.

To compute those quantities, approximations are needed. Methods to do so include Besag's pseudo-likelihood [15] to compute $W(\beta)$,

and simulating the posterior distribution of Z given y with the parameters at iteration l, with a Gibbs sampler to estimate t_{ih}^{l+1} [25].

However, another method exists, the mean field approximation originally proposed in the field of statistical mechanics. Since then, it has been used in a variety of fields including computer vision [114] and more recently to approximate the distribution of both $W(\beta)$ (with a single β) and t_{ih}^{l+1} [117]. I present here the extension of this method to a model with one β parameter per cluster.

4.4 MEAN FIELD APPROXIMATIONS

The idea behind this approximation is to compute intractable quantities at any point $i \in S$ by setting all the other sites in the field to their mean values. Keeping in mind the Markov property expressed in equation (1), when considering a single site $i \in S$, setting all the other sites in the graph to a defined value is equivalent, in the case of an MRF, to setting only the values of $N(i)$.

When computing t_{ih}^{l+1}, the mean fields approximation yields the following fixed point equation for $i \in S$ and $1 \leqslant h \leqslant K$ [29]:

$$t_{ih}^{l+1} \approx \frac{f_h(y_i; \theta_h^l) \exp\{\beta_h^l \sum_{j \in N(i)} t_{jh}^{l+1}\}}{\sum_{u=1}^{K} f_u(y_i; \theta_u^l) \exp\{\beta_u^l \sum_{j \in N(i)} t_{ju}^{l+1}\}} \tag{8}$$

For the normalizing constant $W(\beta)$, by applying the mean-field approximation, using equation (3), $W(\beta)$ can be written as:

$$W(\beta) = \sum_{z'} \exp(-H(z')) \approx \sum_{i \in S} \sum_{z_i} \exp(-H(z_i)) = \sum_{i \in S} \sum_{z_i} \exp(\beta_{z_i} \sum_{j \in N(i)} \mathbb{1}[z_i = z_j])$$

With this new set of equations, it becomes possible to estimate all quantities needed in the E step in order to compute the model's expectation.

4.5 M STEP

After the E step, maximizing ψ is relatively straightforward. Equation 7 yields:

$$\Theta^{l+1} = \arg\max_{\Theta} R_y(\Theta \mid \psi^l)$$
$$\beta^{l+1} = \arg\max_{\beta} R_z(\beta \mid \psi^l)$$

For Θ, once the the $t^l_{ih} = p(Z_i = h \mid y; \psi^l)$ have been computed during the E-step, those probabilities may be used to assign each cell to its most probable cluster at step l. This step classifies this method as hard clustering, this is a technical choice but of course the probabilities t^l_{ih} could instead be used to compute a fuzzy clustering criterion for instance, the Hathaway's fuzzy clustering criterion as described in [29].

Once the new partition is created, the values of Θ that maximize the model's expectation can be computed iteratively for cluster $h \in 1, ..., K$ and gene $m \in M$. Specifically if $Expr_{h,m}$ denotes the number of cells expressing gene m in cluster h and Num_h denotes the total number of cells in cluster h, we can write:

$$\theta^{l+1}_{m,h} = \arg\max_{\Theta} R_y(\Theta \mid \psi^l) = \frac{Expr_{h,m}}{Num_h}$$

In order to maximize β^{l+1}, an iterative approach such as the gradient ascent algorithm, the positive version of the gradient descent algorithm [23] was used for each $\beta^{l+1}_h, h \in [1, K]$ over the function $R_z(\beta \mid \psi^l)$.

The described EM algorithm leads, after convergence, to a partition over K clusters that finds a local maximum of the model's expectation. Importantly, the maximum reached is only a local one as indeed the EM algorithm does not guarantee to reach the global maximum of a function. It is interesting to note that this fact makes the initialization of the algorithm a crucial step. I will discuss this further in Chapter 5.

4.6 ESTIMATING K

Without any prior knowledge, choosing the right number of clusters K is challenging. I decided to use an *a posteriori* method relying on the final log Likelihood of the model derived from equation (6):

$$\log L(\psi) = \log P_G(y, z \mid \Theta, \beta)$$

Because $\log L(\psi)$ monotonically increases with the number of parameters of the model, we employed a penalized likelihood approach to infer the number of clusters. Specifically, if P is the total number of parameters in the model and N the cardinality of S, I calculated the Bayesian Inference Criterion (BIC) as:

$$-2 \log L(\psi) + P \log N$$

By computing the final likelihood for a large range of possible K values, the minimal resulting BIC will be chosen as the optimal number of classes, $\hat{K}$. When applied to the biological data however, this approach is not ideal as I will describe later in this thesis (see Chapter 6) but yields good results when applied to simulated data (see Chapter 5).

4.7 SUMMARY

The goal was to allocate the $S = 32,203$ "cubes" described above in Chapter 2 into K clusters, where K is unknown, using the binarised matrix of $M = 86$ gene expression measurements, Y. To incorporate spatial information into the clustering scheme, I assumed that Z, the (latent) vector of length S that describes the allocation of cells to clusters, satisfies a first-order Markov Random Field (MRF), where the probability that a cell is allocated to a given state depends only upon the states of its immediate neighbours (Figure 22). Additionally, within cluster h ($h \in 1, ..., K$), I assumed that the expression of gene m follows a Bernoulli distribution with parameter $\theta_{m,h}$. The $M \times K$ matrix Θ denotes the full set of Bernoulli parameters. In a typical MRF, the degree of spatial cohesion is determined by a single parameter β, which is assumed to be constant for all clusters [97, 116]. However, in the context of tissue organisation, it is reasonable to expect that the degree of spatial cohesion will differ between clusters; consequently, a separate value of β is estimated for each of the K clusters.

To estimate the parameters of the model an Expectation-Maximisation (EM) based approach has been used in conjunction with mean-field approximations to infer intractable values [24]. Finally, to choose the optimal number of clusters, K, I used the Bayesian Information Criterion (BIC).

The next step is to validate the method's behaviour and to assess the quality of the results compared to the other non-spatial clustering schemes described in Chapter 3.

5

METHOD VALIDATION AND PERFORMANCE ANALYSIS ON SIMULATED DATA

5.1 SIMULATING DATA WITH A SPATIAL COMPONENT

Simulating data with a spatial component is a non-trivial problem. Existing methods rely on MCMC approaches as described in [25]. However, in the *P. dumerillii* in-situ data, with a relatively large number of nodes in the graph ($\sim 34,000$), this is computationally expensive [13]. To overcome this problem, I exploited the fact that the *Platynereis* dataset already possesses a spatial component. As outlined in Figure 23, the simulation starts by clustering the gene expression data using different values of K with the HMRF method described in chapter 4 and by storing the resulting parameter estimates. Subsequently, I use the values of the estimated parameter $\boldsymbol{\Theta}$ to simulate binarised gene expression data from K clusters where, for cluster h, the expression of gene m is simulated from a Bernoulli distribution with parameter $\theta_{m,h}$ as described in 5.1.1. This non-spatial simulated data is then reintroduced in the spatial context of the biological data (5.1.2) leading to a simulated dataset with all parameters being fully determined. In the next paragraphs, I will describe each step of this simulation scheme.

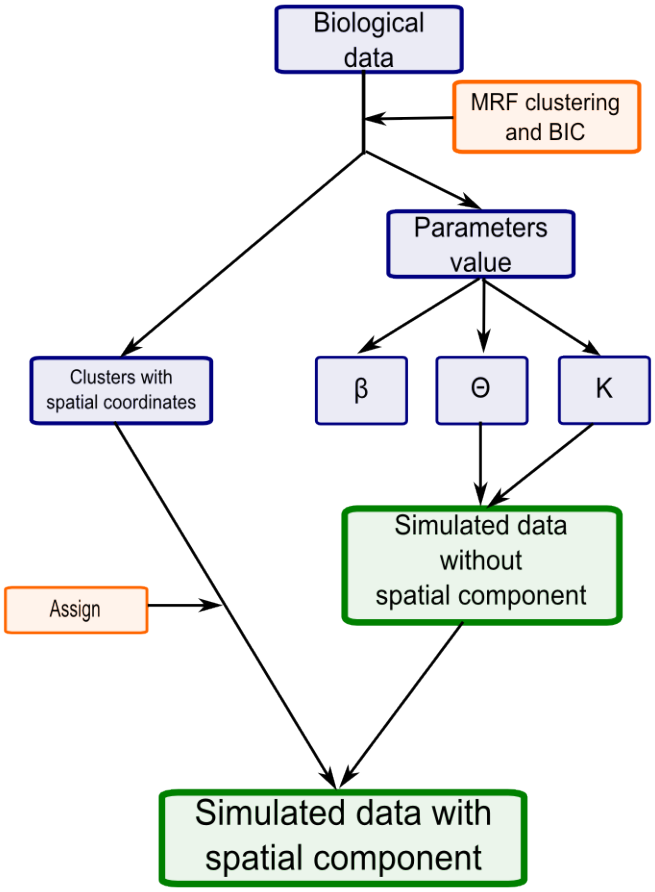


Figure 23: **Simulation scheme used to generate gene expression data with a spatial component and known parameters.** The values of Θ are used to generate a dataset of clusters with the same gene expression profile as the reference. Each simulated cell is then assigned to its corresponding spatial location so that the simulated data keeps the spatial structure of the biological data.

5.1.1 *Simulating non-spatial gene expression data*

The first step of the simulation scheme is to simulate binary gene expression data for $S = 32,203$ sites and $M = 86$ genes belonging to K clusters. Each cluster will be assigned N_h sites with $h \in [1, K]$. Given the emission model described in Chapter 4, for each gene and each cluster, a $K \times M$ matrix $\boldsymbol{\Theta}$ is needed where each $\theta_{h,m}$ represents a Bernoulli parameter corresponding to the probability that each site in cluster h expresses gene m.

In order to generate a biologically coherent Θ matrix, I applied the HMRF method to the true biological data for K clusters and used the resulting Θ matrix to simulate new data. The clustering of the biological data also generates the number of cells per cluster $N_h, \forall h \in K$.

Once the parameters values are available it is relatively straight forward to simulate a vector of gene expression for the N_h sites in each cluster.

5.1.2 *Introducing a known spatial context*

Each simulated data point is then assigned to the same spatial location as the corresponding "cube" in the biological dataset, meaning that both the simulated and the biological datasets have the same neighbouring graph. By using simulated gene expression data equivalent in the same spatial context as the true data, the hypothesis is that the set of parameters β will stay relatively stable when the simulated data is clustered. Consequently, the values of β obtained after clustering the true data may be used as reference values.

5.1.3 *Expected results*

Given this simulation scheme, the expected result after clustering the simulated data, is a strong conservation between the "true" values $\hat{\psi}$ obtained from clustering the biological data and the estimated values $\widetilde{\psi}$ obtained after clustering the simulated data.

5.2 COMPARING CLUSTERING RESULTS USING THE JACCARD SIMILARITY COEFFICIENT

5.2.1 *Theoretical problem in comparing clustering results*

To compare clustering results, several metrics exist to estimate the similarity between two lists of clusters. One of the most widely used ones is the Jaccard coefficient [54]. For two clustering results, for instance the output of two approaches when clustering the same data A and B, the Jaccard coefficient quantifies the similarity between A and B. In other words, the higher the Jaccard coefficient between A

and B, the more similar the two clustering sets are. $J(A, B)$ is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Although theoretically very simple, in practice computing this metric is not trivial. Indeed, depending on the clustering method used and on the initialization, even if the clustering results are 100% identical, they may be misaligned.

This means that for method A *cluster 1* could for example be *cluster 5* in method B. In order to compute the Jaccard coefficient and compare clustering results, it is necessary to be able to align different sets. This problem may be solved by computing the Jaccard coefficient for all possible combinations of clusters between A and B. However, computationally when K increases, computing the Jaccard coefficient for this exponentially increasing number of combinations quickly becomes expensive. Consequently, I developed a similarity specificity matrix approach as described in the next paragraph to align the clustering results before computing the Jaccard coefficient.

5.2.2 *Alignment via similarity-specificity matrix*

The "count" matrix D and the "similarity/specificity" matrix H for comparing two clustering outputs, z and z', each with K clusters, so that $z = \bigcup_{h \in [1,K]} c_h$ and $z' = \bigcup_{h \in [1,K]} c'_h$ are defined as:

$$D = \begin{pmatrix} |c_1 = c'_1| & \dots & |c_1 = c'_K| \\ \vdots & \ddots & \vdots \\ |c_K = c'_1| & \dots & |c_K = c'_K| \end{pmatrix}$$

and

$$H_{ij} = \frac{D_{ij}}{\sum_a D_{aj} \sum_b D_{ib}}$$

With z the set of reference clusters (in the case of the simulation study, z is the set of "true" clusters obtained after clustering the biological data), for each row of the matrix H, the column with the highest value is selected as the corresponding cluster.

In the case of two sets of clusters being extremely similar the alignment is successful and no information is lost. However some errors

may arise if the two cluster sets are substantially different. For example if one cluster h_{z_1} in z is split into two clusters $h_{z'_4}, h_{z'_5}$ in z', the alignment process will assign them both to h_{z_1}, meaning that, because K is the same for z and z' one cluster in z will have no corresponding cluster in z'. In such cases, some information will be lost during the alignment process.

This type of error is very hard to avoid without controlling the initialization of the clustering, which would bias the results. Therefore, the Jaccard coefficient will not necessarily be linearly correlated with the similarity between the reference clusters and the clusters under study, instead it will have a tendency to worsen faster than the dissimilarity due to the alignment step. It remains however a good indicator of the divergence between clustering sets. An example of cluster alignment is shown through the values of $\boldsymbol{\Theta}$ by comparing Figure 24a and Figure 24b.

Now that I have established a method to compare cluster sets, I need to validate the correct estimation of the model's parameters, as described in the next section.

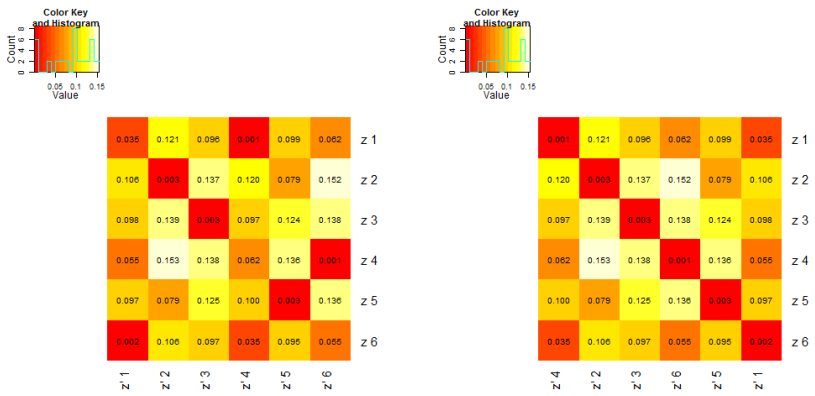
5.3 VALIDATION OF PARAMETERS ESTIMATION AND MODEL SELECTION

5.3.1 *Estimation of* $\boldsymbol{\Theta}$

To validate the consistency in estimating the values of $\boldsymbol{\Theta}$, I compare the "true" values used to simulate the data with the values obtained after clustering the simulated data.

A simple example with $K = 6$ is presented in Figure 24: each cell of the heatmaps $HM_{h,h'}$ with $h, h' \in K$ represents the mean of the vector of pairwise differences $\theta_{h,m} - \theta_{h',m}$. Figure 24a shows these values before alignment and Figure 24b after. As expected, after alignment the small values are aligned in the diagonal showing that each cluster h' exhibits highly similar values of $\boldsymbol{\Theta}$ compared to its corresponding cluster h in the reference.

It is also interesting to note that comparing the similarity between the inferred and true clusters with the Jaccard coefficient implicitly



(a) "Proximity" between values of Θ before alignment

(b) "Proximity" between values of Θ after alignment

Figure 24: Validating the estimation of Θ for $K = 6$. On the x axis are shown the 6 clusters obtained after clustering the simulated data. On the y axis are shown the 6 "true" reference clusters. Each cell of the heatmap corresponds to the mean of the absolute pairwise (with respect to the 86 genes considered) difference between "true" and simulated Θ values. A small number means that the difference between the reference Θ values and the ones obtained after clustering the simulated data is very small.

assesses the accuracy of the estimation of Θ: if the inferred and true clusters are identical, the estimates of Θ must be equal to the true values. In practice, a Jaccard coefficient of 1 implies perfect agreement. Figure 28 shows the value of the Jaccard coefficient for several clustering methods including the HMRF (red points). The very high value of the Jaccard coefficient suggests that the values of Θ are consistently estimated for $K \in [4, 80]$ (see 5.4 for the full description of Figure 28)

5.3.2 *Estimation of* β

To determine how accurately the values of β are estimated, I compare the true and inferred mean values of β for different values of K, as shown in Figure 25 (red and green dots). I chose to compare the mean values instead of comparing β in a pairwise manner because of the alignment errors described in the previous paragraph. Indeed, for example if one true cluster is split in two after clustering the simulated data, there isn't a simple rule about how the value of β should be distributed between the two resulting cluster. In this context using the mean value allows a consistent comparison of the overall value of

β, even though the sensibility for extreme values is lost.

Analysing Figure 25, I observe that the values of β increase with K, which is to be expected since more clusters implies the existence of more transition areas (sites where neighbours do not belong to the same cluster). Because the inner spatial structure of the data is conserved regardless of K, this males an increase of each component of β necessary to maintain the optimal spatial coherency of the full model.

Interestingly, Figure 25 also shows a slight but consistent underestimation of β. This can be explained by noting that the simulation scheme used may reduce the spatial coherency within clusters. Specifically, as illustrated in Figure 26, clusters may not display homogeneous expression of a given gene: instead, depending upon the value of θ, a gene will be expressed only in a fraction of cells. In reality, the cells in which such genes are expressed may have a coherent spatial structure within the biological cluster that is lost in the simulation, thus explaining the consistently smaller value for β that are estimated.

To explore this further, I performed a second simulation using the parameter values estimated from the first simulation as a reference. In this context no further loss of spatial coherency was expected, which was indeed confirmed as shown by the blue curve in Figure 25.

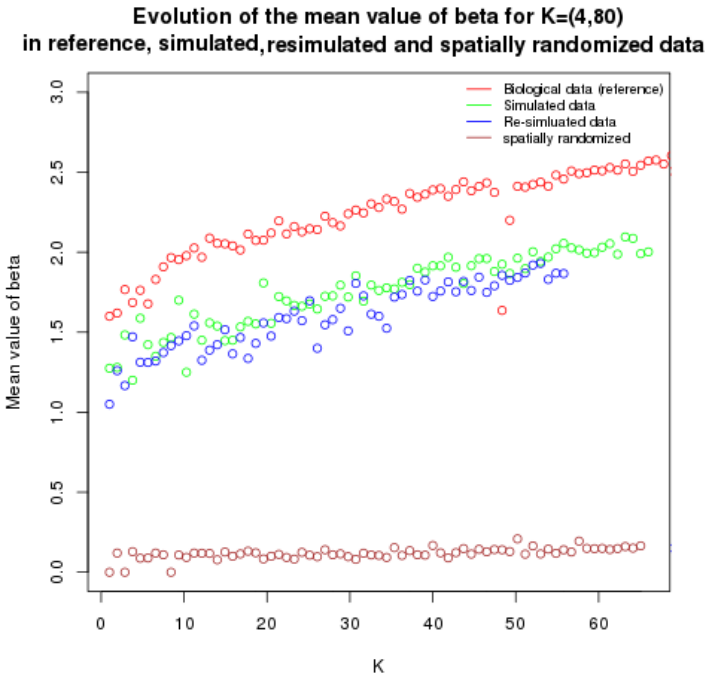


Figure 25: **Validating the estimation of beta.** This Figure shows the evolution for $K \in [4, 80]$ of the mean value of β across all the clusters. The red dots represent the biological data clustering (i.e the reference in our simulations scheme). The green dots represent the results obtained after clustering simulated data, which shows an underestimation of β. To confirm that this underestimation come from the simulation scheme and not the clustering method, the simulated data was used as the reference to generate a "second generation" of simulated data, suppressing the simulation scheme bias (see Figure 26). The results of this re-simulation are shown by the blue dots, which exhibit no underestimation of β. Finally the brown dots represent the mean value of β on the same simulated data but spatially randomized, as expected the β are now estimated to 0.

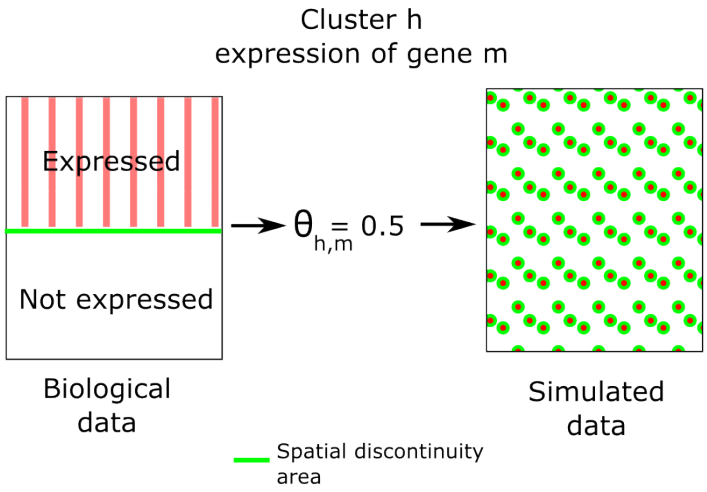


Figure 26: **Decrease in spatial coherency due to the simulation scheme.** For an example cluster h, gene m may only be expressed in half of the cells. This will yield $\theta_{h,m} = 0.5$. However, in the biological data, the cells expressing gene m may be spatially coherent (i.e., located close to one another), leading to a reduced area of expression discontinuity (the green line). By contrast, in the simulated data the expression of such a gene will lose its spatial coherency, leading to an increased area of expression discontinuity. The number of cells having a neighbour with some differences in the gene expression pattern is directly linked to the value of β_h through the energy function described in Chapter 4. This explains the underestimation of β observed in Figure 25.

To validate further our estimation of β, I randomized the coordinates of the simulated "cubes" to lose any spatial component before re-clustering the data. As expected, we observed that the estimates of β were very close to 0 for all clusters (Figure 25, brown dots) , as well as there being very similar Jaccard coefficient values (relative to the true values) for the independent mixture and the MRF model as shown in Figure 28B. Both of these observations provide confidence in our assertion that the model is able to consistently estimate the values of β and that the spatial component of the model plays an important role in the fit.

5.3.3 *Choosing K*

Finally, assessing the ability of the model to choose the correct number of clusters, K is crucial. To this end, the "true" number of clusters underlying the simulated data $\hat{K}$ was compared to the inferred value,

$\widetilde{K}$ obtained after applying the BIC method (see Chapter 4). The results for two representative choices of K are shown in Figure 27 and demonstrate that our clustering approach, in conjunction with the BIC, is able to accurately determine the optimal number of clusters.

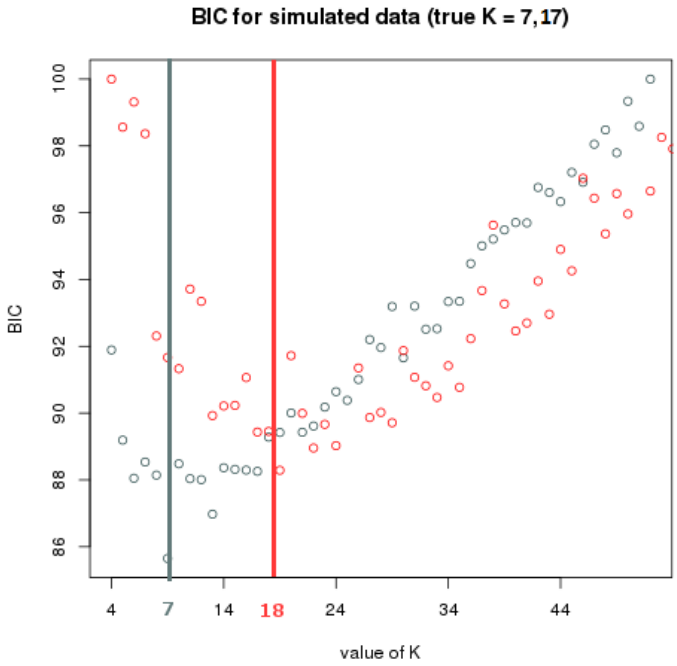


Figure 27: **Estimating the BIC from the simulated data.** The BIC is plotted on the y-axis for different values of K on the x-axis. The red and the grey points correspond to the BIC estimated when the underlying data have 17 and 7 clusters, respectively. The minimum BIC value is 18 and 7, respectively, suggesting that the MRF approach in conjunction with the BIC accurately estimates the optimal number of clusters.

5.4 METHOD PERFORMANCE AND INITIALIZATION

As pointed out previously, initialization is a key step of the HMRF clustering. Working on the simulated data allows a comparison of clustering results generated with a variety of initialization schemes.

5.4.1 *The EM principle and local maximum*

As explained in Chapter 4, the HMRF clustering I developed relies on a Maximum Posterior Marginal (MPM) approach, and the EM al-

gorithm is used to estimate the unknown parameter values. The likelihood function that needs to be maximised may possess a varied set of stationary points. Thus, convergence to the global maximum with the EM algorithm, depends strongly on the parameters initialisation. To overcome this problem, different initialisation strategies have been proposed and investigated (see for instance [16, 59, 72]).

Indeed, if the procedure is initialized with a set of clusters that are close to a local maximum in the likelihood function, the EM algorithm will converge to this local maximum and will never reach the global maximum of the model.

5.4.2 *Random initialization vs Hclust initialization*

To shed some light on the initialization scheme issue I compared two theoretically opposed initialization schemes :

- A random approach: 10.000 random initialization were generated for $K \in [4, 70]$ and for each, the initial likelihood of the model was computed. The initialization with the highest initial likelihood was selected to start the EM algorithm.
- A directed approach: the data were clustered using the non-spatial hClust method described in Chapter 3 and the resulting set of clusters were used to initialize the EM algorithm.

The results are shown in Figure 28 (black and green dots on panel A). Looking at the effect of the initialization scheme on the quality of the resulting clusters via the Jaccard coefficient for $K \in [4, 70]$ it is clear that, unsurprisingly, considering that the EM algorithm does not guarantee to reach the global maximum, the random initialization scheme performs better than the directed initialization scheme. Indeed, for the HMRF randomly initialized, the average Jaccard coefficient is around 0.8 when it averages only 0.6 when initialized with hClust.

Following this observation, for the rest of this thesis and especially in Chapter 6, the HMRF method will be randomly initialized.

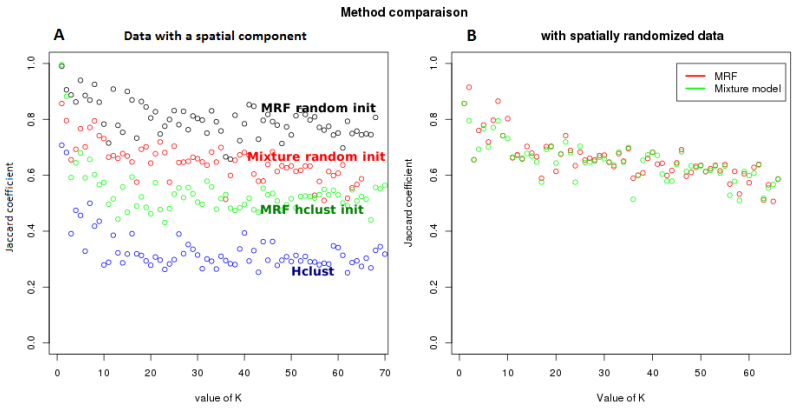


Figure 28: **Jaccard coefficient between true and resulting clusters on the simulated data with different methods and initializations.** Panel A compares the performance of the MRF method with a random initialization with an independent mixture model also with a random initialization, the MRF method initialized with the hClust classification and hClust alone on data simulated with a spatial component. Panel B shows the Jaccard coefficient for the MRF method and independent mixture model both with a random initialization; in this case both methods are applied to simulated data that lacks a spatial component.

5.5 METHOD PERFORMANCE COMPARED TO HCLUST AND INDEPENDENT MIXTURE MODELS

Because the simulated data provides a clear set of true clusters and parameter values, it allows the HMRF clustering to be compared to other clustering methods in terms of clustering quality via the Jaccard coefficient. Additionally, computing time being a key factor for large datasets, I provide in this section some key figures about the execution time of the different methods.

5.5.1 *Quality of the clustering results*

The resulting Jaccard coefficients obtained by comparing the true clusters and the clusters generated by different approaches are shown in Figure 28 for $\tilde{K} \in [4, 70]$. The HMRF method, when used with a random initialization scheme, has an average Jaccard coefficient of 0.8, and clearly demonstrates better performance than the other methods. The second best performing method is the independent mixture model with a random initialization, which has an average Jaccard coefficient of 0.7. Since the independent mixture approach is equivalent

to the MRF with all the β parameters set to zero (i.e., without a spatial component) this suggests that accounting for the spatial aspect yields improved results. hClust also performs relatively poorly with an average Jaccard coefficient around 0.4.

Although as mentioned previously, the Jaccard coefficient may not be linearly correlated with the quality of the clustering results because of the alignment step which can create biases, this simulation study shows that the HMRF consistently outperforms the other methods tested.

5.5.2 *Computing time*

I have assessed in the previous paragraph that the developed HMRF clustering method yields better results in terms of clustering than the other tested methods. However, the fact that the method takes into consideration the spatial dependencies between data points means that it is will be computationally more expensive than non spatial methods, especially as the number of sites increases.

The number of clusters K also has an important influence on the computing time. Given a fixed number of sites $S = 32.203$, I ran the HMRF, and the mixture model methods on simulated datasets for $K \in [4, 60]$ and have obtained, the computing times shown in Figure 29 (on the same machine).

Because of the necessity to estimate every component of $\boldsymbol{\beta}$ at every step through a gradient ascent algorithm (see Chapter 4), and the increased complexity of computing the likelihood of the model when K increases, it is unsurprising to see that the HMRF approach necessary computing time has an exponential relationship with K. On the other hand, the independent mixture model approach does not need to perform these calculations, and exhibits a linear evolution. However, as seen in the previous paragraph, the spatial component of the model seems to improve the clustering quality significantly. Consequently, the HMRF approach might prove useful when K is relatively low. Indeed, until $K = 30$ the required computing time required for the HMRF is not dramatically higher compared to the mixture model approach. In practice, this is likely to be the case as complex biologi-

cal tissues are not made of hundreds of sub-tissues.

The computing time required for hClust is quite high as shown is Figure 29 (blue line), but is constant for any number of clusters $K \in [2, S]$. Indeed, once the dendrogram is computed (see Chapter 3), cutting the clustering tree to any number of clusters is trivial. It is also interesting to note that the clustering results for all values of K are computed in one run with hClust when the other methods need a full run for each value of K. Although I have shown in Figure 28 that the clustering performances of hClust are clearly below those of the HMRF and the independent mixture models methods, the fact that the results for all values of K are obtained at once, regardless of the relatively high computing time, might be advantageous in some cases.

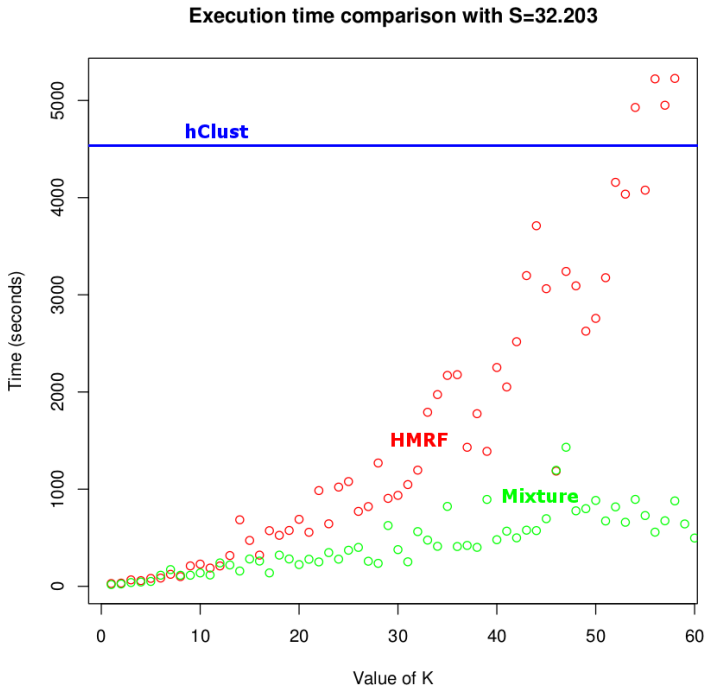


Figure 29: **Computing time required by different clustering methods for** $K \in [4, 60]$ On the x axis is shown the value of K used to cluster the 32.203 data points. The red dots represent the computing time required by the HMRF method, the green dots by an independent mixture model approach and the blue line for hClust.

5.6 SUMMARY

In this Chapter, I have presented the simulation study conducted to evaluate the performances of the HMRF clustering method described in Chapter 4. First, I have described the method used to simulate data with known spatial characteristics. Subsequently, I validated that all the parameters of the model were estimated correctly and consistently by the EM algorithm. I then discussed the performance of the method regarding the initialization scheme, and concluded that a selected random initialization yielded the best results. Finally I used the simulated data to compare the performances of the HMRF method compared to other clustering methods, the result of which was that even though the method is quite costly in terms of computing time when the number of clusters increases, it clearly outperforms both the independent mixture model and the hClust clustering methods in terms of accuracy.

Having validated the method, I will present in the next Chapter the results obtained when the HMRF clustering was applied to the single cell in-situ hybridization data described in Chapter 2.

6

HMRF CLUSTERING IN THE BRAIN OF *PLATYNEREIS DUMERILLII*

6.1 CHOOSING K WITH THE BIC ON BIOLOGICAL DATA

After validating the HMRF method using simulated data as detailed in Chapter 5, I now present the clustering results when the method is applied to the single cell expression dataset in *P. dumerillii*'s brain. Before interpreting the biological meaning of the inferred clusters, the first step is to choose K. To this end, as presented in Chapter 4 I applied the BIC method.

However, as shown in Figure 30 (grey dots), the BIC does not reach a clear minimum but instead reaches a plateau after a given number of clusters. This is most likely due to the highly, but not perfectly symmetrical nature of the brain: with a small K, the same "tissue" on both the left and the right hand side of the brain will belong to the same cluster. However, because the two sides of the brain are not perfectly symmetrical, as K increases the left and right part of the same "tissue" will be clustered separately. As a result, the likelihood continues to increase sufficiently to explain the flattened BIC curve.

Moreover, this hypothesis seems to be confirmed by the fact that when computing the BIC on the right and left side of the brain separately, the curve has in both cases a clear minimum as shown in Figure 30 (red and green dots). Given this, I opted to choose K as the point where the BIC curve reaches a plateau. *Consequently for the rest of the Chapter, I considered the clusters identified for* $K = 33$. Importantly, the BIC starts to rise after $K = 66$, which seems to confirm, assuming the symmetry hypothesis described above, that $K = 33$ is a sensible choice.

The main output of the method is a list of $S = 32.203$ cluster assignments, that is, the cluster each "cube" of the in-situ hybridization data belongs to. With this output and the spatial coordinates of each cube, it is easy to use the tool bioWeb3D presented in Chapter 3 to visualize the clusters in the brain. However, downstream analysis solely

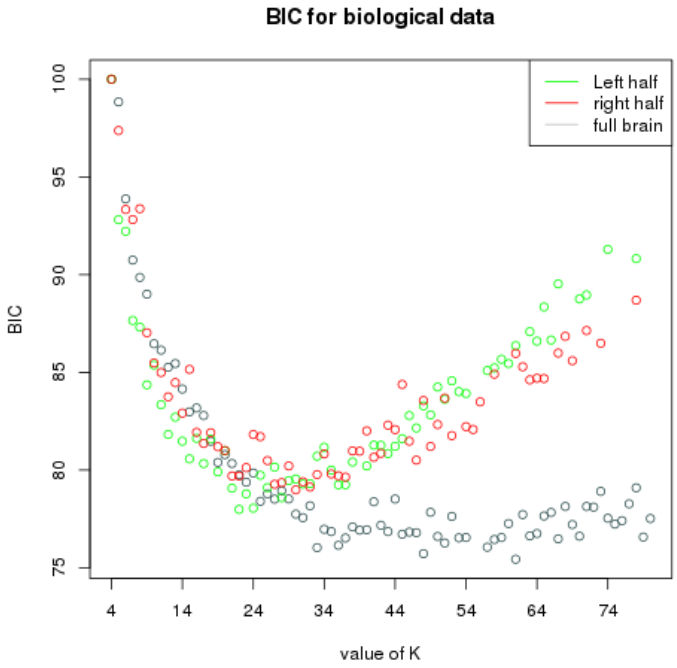


Figure 30: **BIC results on biological data.** Results are shown for $K \in [4, 80]$ (x axis) with the full brain, and for the left and right half separately. The y axis shows the BIC value as a % of the highest BIC value for each dataset.

based on the spatial localization of the clusters is insufficient, and it is possible to take advantage of the model's output parameter values to analyse and interpret the biological meaning of the resulting clusters.

6.2 PARAMETERS INTERPRETATION

Parameter Θ can be used to shed some light on the biological meaning of the inferred clusters. As described in the previous chapters, for $h \in K$ and $m \in M$, $\hat{\theta_{h,m}}$ is the probability that gene m is expressed in a given cell contained in cluster h. Therefore, the values of Θ provide a link between the mathematical model and downstream biological interpretation.

However, in practice, not all of the 86 genes will provide insight into the biological function of a given cluster. For instance, in the case of a ubiquitously expressed gene, g, the value of $\theta_{h,g}$ will be high

for all clusters. To overcome this problem, I developed a score, S, for each gene, m and each cluster h, where:

$$s_{hm} = \frac{\theta_{h,m}}{\sum_a \theta_{a,m}}.$$

For each gene, m, and cluster, h, s_{hm} is large if gene m is specific to cluster h. Consequently, the top scoring 3 or 4 genes for each cluster will represent a specific stereotypical expression pattern that will help infer or confirm the identity of the functional tissue represented by each cluster.

6.3 FINDING KNOWN BIOLOGICAL STRUCTURES TO VALIDATE THE METHOD

To validate the downstream analysis approach presented, I first considered some well characterised regions within the *Platynereis* brain.

6.3.1 P. dumerillii's *eyes*

Arguably the best-studied regions of the brain in *Platynereis* are the eyes: the brain has 4 eyes, two larval and two adult, and their locations and expression fingerprints are well known. As shown in Figure 31, our approach generates two spatially coherent clusters that correspond to each of these regions. Importantly, the best scoring genes that characterise these clusters are biologically meaningful: *rOpsin* and *rOpsin3*, both members of the well-described *opsin* family of photosensitive molecules [101, 88], best distinguish the adult eye and larval eyes respectively, consistent with the in-situ data images shown in Figure 32.

6.3.2 *Mushroom bodies*

As well as the eyes, a second region of the *Platynereis* brain, the mushroom bodies (which corresponds to the pallium, layers of neurons that cover the upper surface of the cerebrum in vertebrates [104]), are also clearly identified by our approach (Figure 34). They have been described anatomically and molecularly in *P. dumerillii* in [104], in this paper the author also defines the mushroom bodies as a ventral regions, subset of the expression pattern of BF1 defined specifically by the same genes as the top scoring genes that define my cluster,

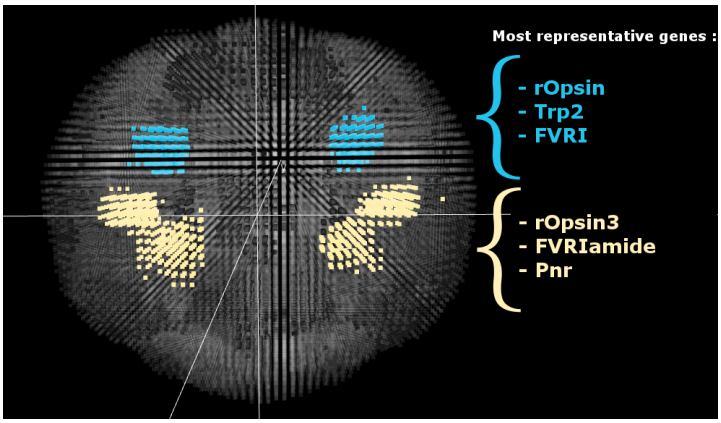


Figure 31: **Eyes in the brain of Platynereis as clustered by the HMRF method.** Adult (blue cluster) and larval (yellow cluster) eyes in separate clusters with their top 3 most representative genes.

Emx, Wnt8. As shown schematically in Figure **??**rom [104], the localization of the mushroom bodies (MB) is coherent with the inferred cluster in Figure 34.

6.3.3 *Motor regions*

Experimentally, when the brain is dissociated from the rest of the larvae, the most basal part contains the developing motor regions. These motor regions are clustered together as shown in Figure 35. Indeed, the green cluster defines a region on the basal side of the larvae that can be associated both by its localization and by its most representative genes (*MyoD* [108, 74] and *LDB3* [61, 71]) with the starting point of the developing muscles of the adult animal. *MyoD* has been shown to play a key role in the differentiation of muscles during development in vertebrates and invertebrates [108, 74] and *LDB3* codes for the protein LDB3, which interacts with the myozenin gene family that has been implicated in muscle development in vertebrates [71].

Again, it is possible to cross-validate the function of this region against previous studies. In this case, [39] used muscle specific phalloidin staining at 48hpf to visualize the developing motor region in the larvae's brain, the result of which is reproduced from [39] in Figure 36.

The eyes, the mushroom bodies and the developing motor regions validation provided good confidence that the HMRF method yielded sensible results and that the gene scoring developed was able to suc-

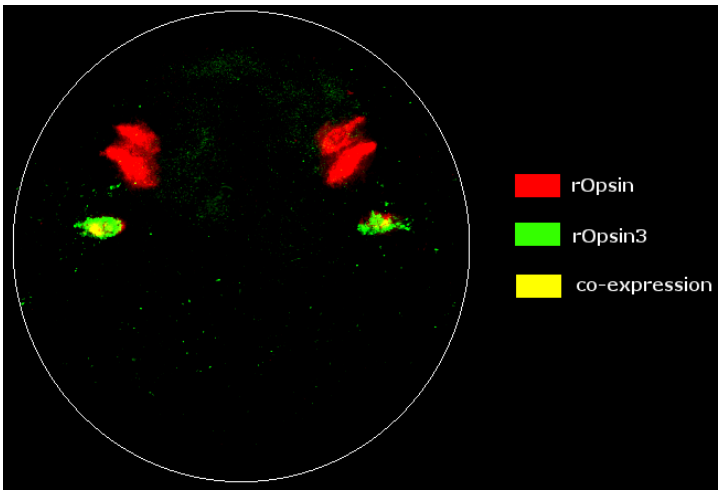


Figure 32: **In-situ hybridization image for rOpsin and rOpsin3 in the full brain at 48hpf (Apical view).** Z-projection of the expression of rOpsin (red) in both the adult eyes and the larval eyes, rOpsin3 (green) specifically in the larval eyes and co-expression areas in some areas of the larval eyes in the full brain of *Platynereis* at 48hpf. The white circle is a schematic outline of the brain. This image been obtained directly from the data obtained in [104].

cessfully define a specific gene expression fingerprint for each cluster. However, of the $K = 33$ clusters, some of the defined regions are not easily recognizable when compared against the known biology of *P. dumerillii*. These regions are very interesting as they may represent previously unstudied sub-populations of cells in the brain.

6.4 GENERATING FUNCTIONAL HYPOTHESES ABOUT UNKNOWN BIOLOGICAL TISSUES

As well as identifying clusters corresponding to known cell types, I also identified clusters that might correspond to less well studied sub-types with specific biological functions.

Given the location of the eyes (Figure 31) and the developing muscles (Figure 35), the location of the pink cluster in Figure 37 is interesting. This cluster surrounds the larval eyes, the adult eyes and reaches the developing muscles described above. Looking at the most representative genes for this pink cluster, it is interesting to note the presence of *Phox2*, a homeodomain protein that has been shown to be necessary for the generation of visceral motor-neurons (neurons of the central nervous system that project their axons to directly or indi-

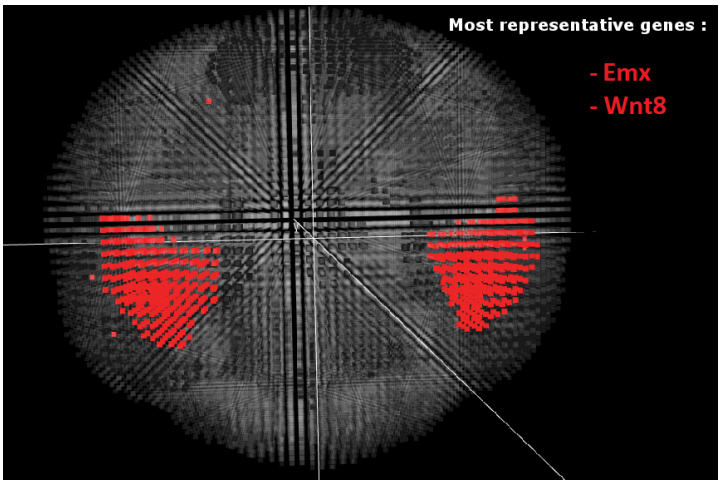


Figure 33: **Mushroom bodies in the brain of Platynereis as clustered by the HMRF method.** Mushroom bodies and their most representative genes.

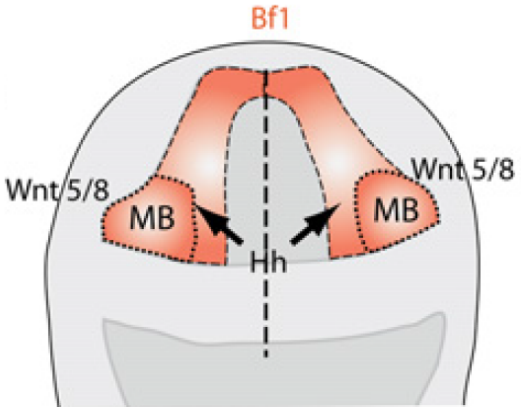


Figure 34: **Schematic representation of the mushroom bodies in the brain of Platynereis by [104].** MB: mushroom bodies

rectly control muscles) as described generally in [22] and in *Drosophila* [20]. The second most representative gene, *COE*, has also been shown to play a role in *Platynereis* and *Drosophila* neural tissue development [30]. In this context, although we lack biological validation, we can hypothesise that the cells within this particular cluster could be developing neurons that link the eyes to the muscles of *Platynereis*.

Although this hypothesis remains purely speculative and would need validation in the laboratory, this example is an interesting proof-of-concept that this clustering method can prove useful for hypothesis generation. Indeed, the analysis of the parameter values and the spatial localization attached to the clusters has allowed me to place

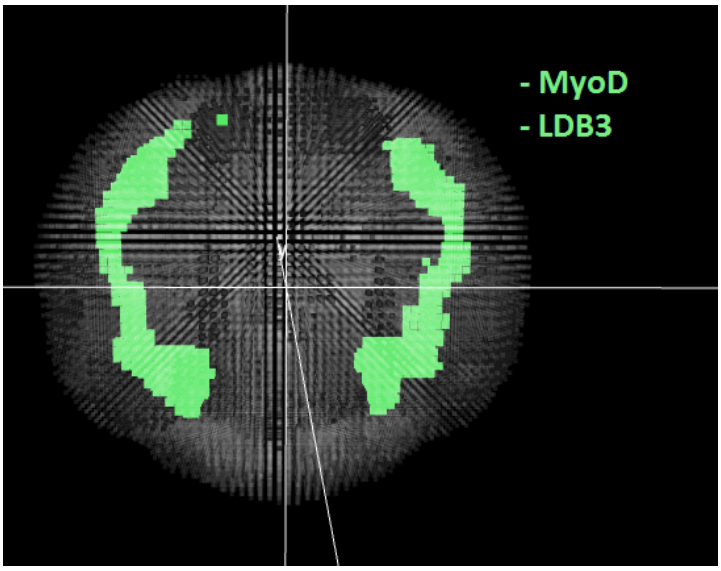


Figure 35: **Developing motor region in the brain of Platynereis as clustered by the HMRF method.** Basal motor regions and their most representative genes.

with a reasonable level of confidence a functional hypothesis about a tissue that was not clearly defined either spatially or functionally. It is also interesting to note that hClust does not separate this putative region or the developing muscles when clustering the same data with the same number of clusters.

6.5 SUMMARY

I applied the HMRF clustering method described in Chapter 4 to the binarized in-situ hybridization data presented in Chapter 2. I described how the BIC method was adapted to chose the optimal number of clusters $K = 33$.

In order to analyse the resulting clusters, I developed a scoring method based on a "specificity" matrix which extracts the most specific genes for each cluster. Thanks to the 3D visualization tool described in Chapter 3, and the most specifically expressed genes for each cluster, I was able to validate the method biologically by localizing 3 well studied regions of the brain. Additionally I checked that the top 3 genes were consistent with the known biology of *P. dumerilii*.

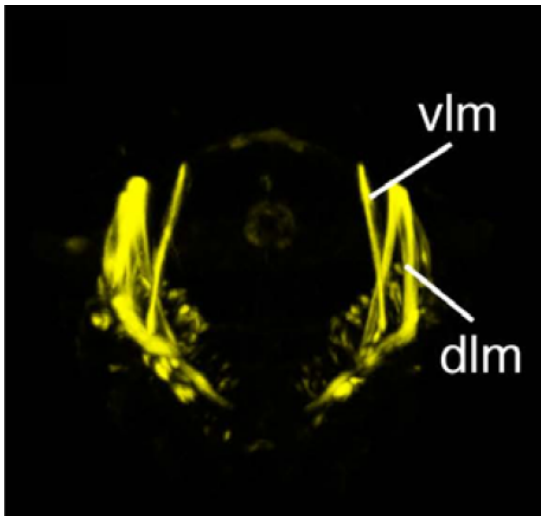


Figure 36: **Developing motor region in the brain of Platynereis visualized in-situ by phalloidin staining.** This Figure is reproduced from [39]. vlm: ventral longitudinal muscle; dlm: dorsal longitudinal muscle.

Furthermore, I demonstrated how my approach allows the generation of functional hypothesis about regions that are not well known. In particular, I discussed a previously unstudied tissue that may consist of developing neurons directly linking the eyes to the developing muscles.

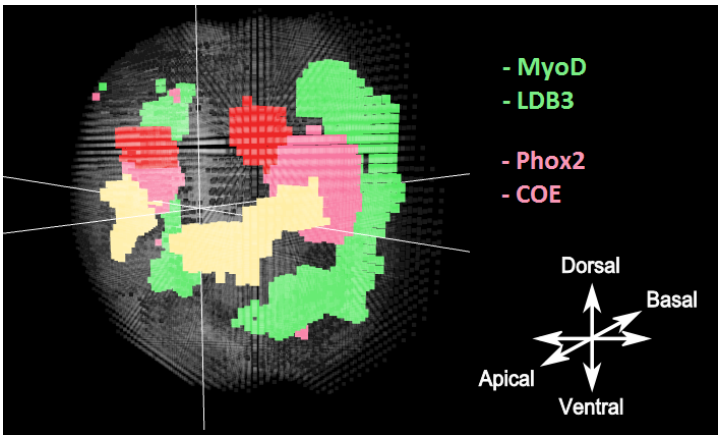


Figure 37: **A putative tissue of developing neurons between the eyes and the larvae's developing muscles.** The yellow and red clusters are the eyes as seen in Figure 31. The green cluster represents the developing muscles on the basal side of the larvae, as the location and the most specific genes strongly suggest. The pink cluster is a putative tissue that makes an interesting link between the eyes and the muscles. The most representative gene of this tissue is Phox2, a homeodomain protein required for the generation of visceral motor-neurons in *Drosophila* [20]

7

CONCLUSIONS AND FUTURE WORK

7.1 CONCLUSIONS

7.1.1 *Summary*

Throughout this thesis I tried to highlight my main original contributions while presenting work from collaborators and other groups. I especially tried to convey how my work is part of a global effort to investigate spatially referenced gene expression patterns at the single cell level in complex tissues.

At the time of writing, the field of transciptomics undergoes major evolutions every year with sequencing methods ever more accurate and requiring less and less starting materials. However, so far no gene expression assay has been able to achieve single cell resolution of the whole transcriptome while at the same time keeping the spatial organization of the tissue intact. On the one hand, in-situ hybridization succeeds for the spatial component but is not suitable to generate full transcriptomes while, on the other hand, single cell RNA-seq can provide with full single cell transcriptomes but no spatial reference.

The original contribution presented in Chapter 2 alongside methods to process single cell gene expression data, has been to propose a way to combine those the two gene expression capture methods in order to create spatially referenced full transcriptomes by mapping single cell RNA-seq results onto a spatial gene expression library generated via in-situ hybridization assays. I have demonstrated the efficiency of the method on a preliminary dataset of single cell RNA-seq data in the brain of *P. dumerillii*. This is very recent, and an ongoing effort with my collaborators. However, so far I have been able to map around 75% of sequenced single cells to a restricted area in the brain defined by the in-situ data. At the time of writing, a manuscript to submit this work as a research paper is being prepared.

I have presented in Chapter 3 some background about clustering single cells based on their gene expression patterns and laid the ques-

tion of visualization to analyse spatially referenced data in a 3D tissue. To answer this question I have presented a software tool that I have developed: "bioweb3D". This tool enables scientists who are not visualization experts to easily visualize 3 dimensional data in their web browser without the need to install any software on their computer or to send any data to a remote server. This tool however fairly basic has been of great use for my own work and as a results multiple Figures (16, 37, 35) in this thesis are screen shots of "bioweb3D". I was also pleased following publication of this tool in *Bioinformatics* [84] to see a few users (around 20 each month) regularly visiting and using "bioweb3D".

The questions about the limitations of single cell expression data as well as the visualization of 3D tissues answered in Chapter 2 and Chapter 3, I have introduced in Chapter 4, what has been my main contribution to the field of transcriptomics during the time of my PhD. The HMRF clustering method I have adapted and extended and applied to the single cell gene expression data, was chosen to answer the question of defining cell types from a bottom up approach at the scale of a very complex tissue where the number of cell types is unknown. The method's main advantage is to account for the spatial characteristics of the sites in order to favour spatially smooth clustering results. Not only does this spatial smoothness help the visual downstream analysis, I also hypothesize that considering cells that are spatially close from one another more likely to belong to the same cell type as biologically relevant and potentially important to reduce the impact of experimental noise on the clustering results.

The HMRF method and specifically the impact of the spatial component on the clustering was evaluated and there were good indication that the approach improves the overall quality of clustering as demonstrated by the simulation study detailed in Chapter 5. Indeed, when compared to other non spatial clustering method suitable for single cell gene expression datasets, the HMRF method performs consistently better, which in my opinion justifies the long work put into applying and extending this image analysis method to cluster a completely different type of data. The method was also assessed in term of computational burden, and although legends of the perfect clustering method may circulate, they are unfortunately unfounded. In the case of the HMRF, the price for the improved clustering quality compared to non spatial methods is paid with an exponentially increasing

computing time with regard to the number of clusters, which limits this approach to datasets where the true number of clusters is less than 40 when clustering tens of thousands of sites for around a hundred genes.

The last chapter (6) details the results when I applied the method to the spatially referenced single cell gene expression data from the brain of *Platynereis dumerillii*. The outcome was good and in particular, I was able to localize well studied structures of the brain in order to validate the method biologically. Furthermore, I described how some clusters localized in poorly documented regions of the brain could be analysed by taking advantage of the model's final parameters. Indeed, I detailed how a specificity score for each gene and each cluster was computed in order to extract the most specific genes to each cluster. This allowed me to characterize a previously unstudied area of the brain and to formulate an hypothesize about the function of its cells.

Even with these satisfactory results there is still work to be done for the clustering method and the single cell RNA-seq back mapping described in Chapter 2.

7.2 FUTURE WORK

7.2.1 *Single cell RNA-seq back-mapping*

The single cell back-mapping method detailed in Chapter 2, is the most recent work presented in this thesis. It s an ongoing project with new data available every month. Consequently, the results shown are preliminary and I expect that the method will be developed much further over the next few months. Particularly, a combinatorial approach to the selection of the most specific genes used to do the spatial mapping is currently being investigated.

7.2.2 *HMRF future developments*

The main areas where the HMRF method needs to be improved is in terms of adapting the emission model described in Chapter 4 to whole transcriptomes and quantitative gene expression data. Indeed,

as explained in Chapter 2, my work has been focused in this thesis on clustering binarized data for 89 key developmental genes.

With an hypothetical dataset in the brain of *P. dumerillii* containing the quantitative expression levels for whole transcriptome (more than 10.000 genes) of each cells, how would such a dataset impact the theoretical framework described in Chapter 4?

The quantitative aspect of the data would require a modification of the emission model. Indeed if Bernoulli distributions for each gene in each cluster is suited binarized data, for quantitative data it would be sensible to use Poisson or negative binomial distribution. Indeed, these are the mainly used approaches to model gene expression in the literature [70, 8].

Furthermore, the fact that instead of a few selected genes, cells have to be clustered considering their whole transcriptome is represents a theoretical issue because genes are not independently expressed. Indeed, as explained in the Introduction (1), gene expression is a highly regulated mechanism exhibiting high inter-dependence between genes. More particularly some expressed genes code for regulatory factors that will influence the level of expression of other genes. This inter-dependency is in plain contradiction with the conditional independence hypothesis for gene expression used in the HMRF model (as described in Chapter 4).

This issue may be resolved by developing methods to analyse whole transcriptomes upstream of the clustering to automatically select an ensemble of genes that are at the same time independent and extremely representative of the overall expression patterns in the studied tissue.

Another possible way to deal with this problem would be to try to extend Markov random fields without the conditional independence hypothesis. However, I will not be the judge of the theoretical feasibility of this idea.

Of course maybe such an approach won't be needed as it is possible that the errors introduced by some genes inter-dependence will represent a negligible bias compared to the signal brought by thousands of genes. Simulation studies may be able to decide this issue.

Part II

APPENDIX

A

INPUT FILE FORMATS FOR BIOWEB3D

A.1 DATASET FILE SPECIFICATION

When the user adds a new *Dataset* file, a new Dataset section is created in the “Data" panel of the application. Each dataset file contains one dataset.

A.1.1 *JSON format*

The *dataset* file should have a root object called “dataset" which contains:

- The “name" property of the dataset (*e.g.*, “my dataset");
- The “chain" parameter, which should be set to *true* if the points are connected (the default value is *false*) - the data will be considered sequentially, with each point connected by a solid line to the previous and next point according to its order in the dataset file;
- The “points" property, which is a two dimensional array representing a list of (x,y,z) vectors that define the co-ordinates of the points.

Listing 1 is an example of a minimal 3 points dataset file.

A.1.2 *XML format*

The *dataset* XML format used is very similar to the previously defined JSON format. The file must have a root object called “<dataset>" which contains:

- The “<name>" property of the dataset (*e.g.*, “my dataset");
- The “<chain>" parameter, which should be set to *true* if the points are linked (the default value is *false*) - the data will be considered sequentially, with each point connected by a solid

Listing 1: Json dataset file

```
{ "dataset" : {
      "name" : "my superb dataset",
      "chain" : true,
       "points" :
        [
          [
            0.5,
            100,
            -50.5
          ],
          [
            200,
            10,
            0.0
          ],
          [
            3,
            250.15,
            15
          ]
        ]
     }
}
```

Listing 2: XML dataset file

```
<?xml version="1.0" ?>
<dataset>
        <name>my superb dataset</name>
        <chain>true</chain>
        <points>
                <point>
                        <x>0.5</x>
                        <y>100</y>
                        <z>-50.5</z>
                </point>

                <point>
                        <x>200</x>
                        <y>10</y>
                        <z>0.0</z>
                </point>

                <point>
                        <x>3</x>
                        <y>250.15</y>
                        <z>15</z>
                </point>
        </points>
</dataset>
```

line to the previous and next point according to its order in the dataset file;

- The “<points>" property, which contains all the single “<point>" elements that define the dataset. Each “<point>" has three properties to define its spatial location, namely “<x>", “<y>" and “<z>".

Listing 2 contains the same minimal dataset as Listing 1 but formatted in XML.

A.1.3 *CSV format*

Each line represents a point and the three coordinates on each line must be separated by “comma" characters.
As an example, listing 3 carries the same information as the JSON file

Listing 3: CSV dataset file

```
0.5,100,-50.5
200,10,0.0
3,250.15,15
```

in Listing 1. We note that although the spatial information remains the same it is not possible to set a name or to connect the points within a CSV file input.

A.2 INFORMATION LAYER FILE SPECIFICATION

The *Information layer* file contains information about the points described in the Dataset file. The information in this file has to be given in the same order as the points defined in the Dataset file.

A.2.1 *JSON format*

The *information layer* files must have a root element named “information". Since one information file can define multiple information sets, the structure below “information" is a list. Each element of the list is structured as follows:

- The “name" property (optional);
- The “numClass" property, which indicates the number of different classes the data will be assigned to;
- The “labels" property, which defines a list of names for the “numClass" classes previously defined (optional);
- The “values" property, which defines the class of each point in the dataset. As points do not have single IDs, this property must be in the same order and have the same length as the points defined in the *dataset* file.

For example coming back to the 3 points defined in Listing 1, two information layers could correspond to:

- one clustering algorithm that puts the first two points together in class one and the third point alone in a second class
- a second clustering algorithm that puts each point in a separate class

Listing 4: JSON information layer file

```
{ "information" :
  [
    {
      "name": "clustering algo 1",
      "numClass": "2",
      "labels" : [
        "Category 1",
        "Category 2"
      ],
      "values": [
        1,
        1,
        2
      ]
    },
    {
      "name": "clustering algo 2",
      "numClass": "3",
      "values": [
        1,
        2,
        3
      ]
    }
  ]
}
```

In this case the Information layer file would look like Listing 4.

A.2.2 *XML format*

The *information layer* XML format used is very similar to the previously defined JSON format. The *information layer* files must have a root element named “<information>". Since one information file can define multiple information sets, the structure below “<information>" is a list of “<set>" elements. Each “<set>" element is structured as follows:

- The “<name>" property (optional);
- The “<numClass>" property, which indicates the number of different classes the data will be assigned to;

Listing 5: XML information layer file

```
<?xml version="1.0" ?>
<information>
        <set>
                <name>clustering algo 1</name>
                <numClass>2</numClass>
                <labels>
                        <label>Category 1</label>
                        <label>Category 2</label>
                </labels>
                <values>
                        <value>1</value>
                        <value>1</value>
                        <value>2</value>
                </values>
        </set>
        <set>
                <name>clustering algo 2</name>
                <numClass>3</numClass>
                <values>
                        <value>1</value>
                        <value>2</value>
                        <value>3</value>
                </values>
        </set>
</information>
```

- The “<labels>" property, which contains as many individual “<label>" properties as the number of different classes. Each “<label>" defines the names for one class (optional);
- The “<values>" property, which contains all the single “<value>" properties, each one defining the class of each point in the dataset. As points do not have single IDs, the “<value>" properties must be in the same order and have the same length as the points defined in the *dataset* file.

Listing 5, carries the exact same information as Listing 4.

CSV FORMAT Each column represents the class to which a point belongs. The separation character between columns must be a “comma". Listing 6, carries the same information as Listing 4. Note that it is not

Listing 6: CSV information layer file

```
1,1
1,2
2,3
```

possible to use the “labels" or “name" properties available in Listing 4 within a CSV information layer file.

BIBLIOGRAPHY

[1] bioweb3d online. URL http://www.ebi.ac.uk/~jbpettit/bioWeb3D.

[2] Compatibility table for webgl. URL http://caniuse.com/webgl.

[3] bioweb3d on github. URL http://github.com/jbogp/bioWeb3D.

[4] United states computer emergency readiness team. URL http://www.kb.cert.org/vuls/id/636312.

[5] Three.js - javascript 3d library. URL http://mrdoob.github.com/three.js/.

[6] Webgl 1.0 specification. URL https://www.khronos.org/registry/webgl/specs/1.0/.

[7] Xml applications and initiatives. URL http://xml.coverpages.org/xmlApplications.html.

[8] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010.

[9] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33 (5):898–916, 2011.

[10] Detlev Arendt. Platynereis dumerilii: a "living fossil" elucidating the evolution of genomes and of the CNS. *Theory in Biosciences*, 124:185–197.

[11] James E Balmer and Rune Blomhoff. Gene expression regulation by retinoic acid. *Journal of lipid research*, 43(11):1773–1808, 2002.

[12] WILLIAM P Bartlett and GARY A Banker. An electron microscopic study of the development of axons and dendrites by hippocampal neurons in culture. i. cells which develop without intercellular contacts. *The Journal of neuroscience*, 4(8):1944–1953, 1984.

[13] Alexandre Belloni and Victor Chernozhukov. On the computational complexity of mcmc-based estimators in large samples. *The Annals of Statistics*, pages 2011–2055, 2009.

[14] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[15] Julian Besag. Statistical analysis of non-lattice data. *The statistician*, pages 179–195, 1975.

[16] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003.

[17] Paul D Boyer. The atp synthase-a splendid molecular machine. *Annual review of biochemistry*, 66(1):717–749, 1997.

[18] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*, 2013.

[19] Sydney Brenner, Maria Johnson, John Bridgham, George Golda, David H Lloyd, Davida Johnson, Shujun Luo, Sarah McCurdy, Michael Foy, Mark Ewan, et al. Gene expression analysis by massively parallel signature sequencing (mpss) on microbead arrays. *Nature biotechnology*, 18(6):630–634, 2000.

[20] J Briscoe, L Sussel, P Serup, D Hartigan-O'Connor, TM Jessell, JLR Rubenstein, and J Ericson. Homeobox gene nkx2. 2 and specification of neuronal identity by graded sonic hedgehog signalling. *Nature*, 398(6728):622–627, 1999.

[21] Marcela Brissova, Michael J Fowler, Wendell E Nicholson, Anita Chu, Boaz Hirshberg, David M Harlan, and Alvin C Powers. Assessment of human pancreatic islet architecture and composition by laser scanning confocal microscopy. *Journal of Histochemistry & Cytochemistry*, 53(9):1087–1097, 2005.

[22] JF Brunet and A Pattyn. Phox2 genes-from patterning to connectivity. *Current opinion in genetics & development*, 12(4):435, 2002.

[23] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96. ACM, 2005.

[24] Gilles Celeux, Florence Forbes, and Nathalie Peyrard. Em procedures using mean field-like approximations for markov model-based image segmentation, 2001.

[25] B CHALMOND. An iterative gibbsian technique for reconstruction of m-ary images. *Pattern recognition*, 22(6):747–761, 1989.

[26] Howard Y Chang, Jen-Tsan Chi, Sandrine Dudoit, Chanda Bondre, Matt van de Rijn, David Botstein, and Patrick O Brown. Diversity, topographic differentiation, and positional memory in human fibroblasts. *Proceedings of the National Academy of Sciences*, 99(20):12877–12882, 2002.

[27] RJ Colello and RW Guillery. The early development of retinal ganglion cells with uncrossed axons in the mouse: retinal position and axonal course. *Development*, 108(3):515–523, 1990.

[28] Linda S Costanzo, You Save, and Tell A Friend. Brs physiology (board review series), 2006.

[29] M. Dang and G. Govaert. Spatial fuzzy clustering using EM and markov random fields. *International Journal of System Research and Information Science*, 8(4):183–202, 1998.

[30] Adrien Demilly, Elena Simionato, David Ohayon, Pierre Kerner, Alain Garcès, and Michel Vervoort. Coe genes are expressed in differentiating neurons in the central nervous system of protostomes. *PloS one*, 6(6):e21213, 2011.

[31] Arthur P Dempster, Nan M Laird, Donald B Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.

[32] Alexandru S Denes, Gáspár Jékely, Patrick RH Steinmetz, Florian Raible, Heidi Snyman, Benjamin Prud'homme, David EK Ferrier, Guillaume Balavoine, and Detlev Arendt. Molecular

architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. *Cell*, 129(2):277–288, 2007.

[33] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167): 193–196, 2014.

[34] Adriaan WC Dorresteijn. Quantitative analysis of cellular differentiation during early embryogenesis of platynereis dumerilii. *Roux's archives of developmental biology*, 199(1):14–30, 1990.

[35] AJ Durston, JPM Timmermans, WJ Hage, HFJ Hendriks, NJ De Vries, M Heideveld, and PD Nieuwkoop. Retinoic acid causes an anteroposterior transformation in the developing central nervous system. 1989.

[36] Jianping Fan, David KY Yau, Ahmed K Elmagarmid, and Walid G Aref. Automatic image segmentation by integrating color-edge extraction and seeded region growing. *Image Processing, IEEE Transactions on*, 10(10):1454–1466, 2001.

[37] EA Feingold, PJ Good, MS Guyer, S Kamholz, L Liefer, K Wetterstrand, FS Collins, TR Gingeras, D Kampa, EA Sekinger, et al. The encode (encyclopedia of dna elements) project. *Science*, 306 (5696):636–640, 2004.

[38] Albrecht Fischer and Adriaan Dorresteijn. The polychaete platynereis dumerilii (annelida): a laboratory animal with spiralian cleavage, lifelong segment proliferation and a mixed benthic/pelagic life cycle. *Bioessays*, 26(3):314–325, 2004.

[39] Antje Fischer, Thorsten Henrich, and Detlev Arendt. The normal development of platynereis dumerilii (nereididae, annelida). *Frontiers in zoology*, 7(1):31, 2010.

[40] Tom C Freeman, Leon Goldovsky, Markus Brosch, Stijn Van Dongen, Pierre Mazière, Russell J Grocock, Shiri Freilich, Janet Thornton, and Anton J Enright. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS computational biology*, 3(10):e206, 2007.

[41] Clay Fuqua, Matthew R Parsek, and E Peter Greenberg. Regulation of gene expression by cell-to-cell communication: acyl-

homoserine lactone quorum sensing. *Annual review of genetics*, 35(1):439–468, 2001.

[42] Manfred Gossen and Hermann Bujard. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proceedings of the National Academy of Sciences*, 89(12): 5547–5551, 1992.

[43] Keren Guy. Development and molecular characterization of adult and larval eyes in platynereis dumerilii (polychaeta, annelida, lophotrochozoa). 2008.

[44] Valentin Häcker. *Die pelagischen Polychaeten-und Achaetenlarven der Plankton-expedition...* Lipsius & Tischer, 1898.

[45] Jörg D Hardege. Nereidid polychaetes as model organisms for marine chemical ecology. *Hydrobiologia*, 402:145–161, 1999.

[46] M. J. Hartshorn. AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aided Mol. Des.*, 16(12):871–881, Dec 2002.

[47] Carl Hauenschild, G Czihak, A Fischer, and R Siewing. *Platynereis dumerilii: mikroskopische Anatomie, Fortpflanzung, Entwicklung*. Fischer, 1969.

[48] Bo Hellman. Pulsatility of insulin release-a clinically important phenomenon. *Upsala journal of medical sciences*, 114(4):193–205, 2009.

[49] Thomas H Hutchinson, Awadhesh N Jha, and David R Dixon. The polychaete platynereis dumerilii (audouin and milne-edwards): a new species for assessing the hazardous potential of chemicals in the marine environment. *Ecotoxicology and environmental safety*, 31(3):271–281, 1995.

[50] Norman N Iscove, Mary Barbara, Marie Gu, Meredith Gibson, Carolyn Modi, and Neil Winegarden. Representation is faithfully preserved in global cdna amplified exponentially from sub-picogram quantities of mrna. *Nature biotechnology*, 20(9): 940–943, 2002.

[51] Ernst Ising. Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, February 1925. ISSN 0044-3328. doi: 10.1007/bf02980577. URL http://dx.doi.org/10.1007/bf02980577.

[52] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research*, 21(7):1160–1167, 2011.

[53] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 2013.

[54] Paul Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[55] Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33:245–254, 2003.

[56] HA John, ML Birnstiel, and KW Jones. Rna-dna hybrids at the cytological level. *Nature*, 223(5206):582, 1969.

[57] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[58] ET Kaiser and FJ Kezdy. Amphiphilic secondary structure: design of peptide hormones. *Science*, 223(4633):249–255, 1984.

[59] Dimitris Karlis and Evdokia Xekalaki. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3):577–590, 2003.

[60] Zbynek Kozmik, Jana Ruzickova, Kristyna Jonasova, Yoshifumi Matsumoto, Pavel Vopalensky, Iryna Kozmikova, Hynek Strnad, Shoji Kawamura, Joram Piatigorsky, Vaclav Paces, et al. Assembly of the cnidarian camera-type eye from vertebrate-like components. *Proceedings of the National Academy of Sciences*, 105 (26):8989–8993, 2008.

[61] Jennifer Krcmery, Troy Camarata, Andre Kulisz, and Hans-Georg Simon. Nucleocytoplasmic functions of the PDZ-LIM protein family: new insights into organ development. *Bioessays*, 32(2):100–108, 2010.

[62] Dinesh B. Kulkarni, Mahesh M. Doijade, Chetan S. Devrukhkar, Ganesh R. Zilpe, and Rajesh R. Surana. Article: Netraris - a

web based dicom viewer. *International Journal of Computer Applications*, 48(24):40–44, June 2012. Published by Foundation of Computer Science, New York, USA.

[63] JE Landegent, De Wal, N Jansen In, RA Baan, JHJ Hoeijmakers, and M Van Der Ploeg. 2-acetylaminofluorene-modified probes for the indirect hybridocytochemical detection of specific nucleic acid sequences. *Experimental cell research*, 153(1): 61–72, 1984.

[64] Wei-Chung Allen Lee. *Cellular and molecular analysis of neuronal structure plasticity in the mammalian cortex*. PhD thesis, Massachusetts Institute of Technology, 2006.

[65] E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, L. Chen, T. M. Chen, M. C. Chin, J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. Desaki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H. W. Dong, J. G. Dougherty, B. J. Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields, S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halverson, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T. Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee, K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels, J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J. Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchalski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno, K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A. Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K. R. Stumpf, S. M. Sunkin, M. Sutram, A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson, L. R. Varnam, A. Visel, R. M. Whitlock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, M. Wood, M. B. Yaylaoglu, R. C. Young, B. L. Youngstrom, X. F. Yuan, B. Zhang, T. A. Zwingman, and A. R. Jones. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, Jan 2007.

[66] Hui Li, BS Manjunath, and Sanjit K Mitra. A contour-based approach to multisensor image registration. *Image Processing, IEEE Transactions on*, 4(3):320–334, 1995.

[67] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[68] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437 (7057):376–380, 2005.

[69] Georgi K Marinov, Brian A Williams, Kenneth McCue, Gary P Schroth, Jason Gertz, Richard M Myers, and Barbara J Wold. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and rna splicing. *Genome research*, pages gr–161034, 2013.

[70] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[71] Nicola Marziliano, Savina Mannarino, Luisa Nespoli, Marta Diegoli, Michele Pasotti, Clara Malattia, Maurizia Grasso, Andrea Pilotto, Emanuele Porcu, Arturo Raisaro, et al. Barth syndrome associated with compound hemizygosity and heterozygosity of the taz and ldb3 genes. *American Journal of Medical Genetics Part A*, 143(9):907–915, 2007.

[72] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[73] G Mendel. Versuche ber pflanzen-hybriden. verh. *Naturforsch. Ver. Brnn*, 4:347, 1866.

[74] Alan M Michelson, Susan M Abmayr, Michael Bate, A Martinez Arias, and Tom Maniatis. Expression of a myod family member prefigures muscle pattern in drosophila embryos. *Genes & development*, 4(12a):2086–2097, 1990.

[75] Pamela J Mitchell and Robert Tjian. Transcriptional regulation in mammalian cells by sequence-specific dna binding proteins. *Science*, 245(4916):371–378, 1989.

[76] Ryan D Morin, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor J Pugh, Helen McDonald, Richard Varhol, Steven JM Jones, and Marco A Marra. Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. *Biotechniques*, 45(1): 81, 2008.

[77] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[78] PM Nederlof, D Robinson, R Abuknesha, J Wiegant, AHN Hopman, HJ Tanke, and AK Raap. Three-color fluorescence in situ hybridization for the simultaneous detection of multiple nucleic acid sequences. *Cytometry*, 10(1):20–27, 1989.

[79] Claus Nielsen. Trochophora larvae: Cell-lineages, ciliary bands, and body regions. 1. annelida and mollusca. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 302 (1):35–68, 2004.

[80] Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98, 2010.

[81] Mary Lou Pardue and Joseph G Gall. Molecular hybridization of radioactive dna to the dna of cytological preparations. *Proceedings of the National Academy of Sciences*, 64(2):600–604, 1969.

[82] G. A. Pavlopoulos, S. I. O'Donoghue, V. P. Satagopam, T. G. Soldatos, E. Pafilis, and R. Schneider. Arena3D: visualization of biological networks in 3D. *BMC Syst Biol*, 2:104, 2008.

[83] Hanchuan Peng, Zongcai Ruan, Fuhui Long, Julie H Simpson, and Eugene W Myers. V3d enables real-time 3d visualization and quantitative analysis of large-scale biological image data sets. *Nature Biotechnology*, 28(4):348–353, 2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2857929&tool=pmcentrez&rendertype=abstract.

[84] Jean-Baptiste Pettit and John C Marioni. bioweb3d: an online webgl 3d data visualisation tool. *BMC bioinformatics*, 14(1):185, 2013.

[85] D Pinkel, J Landegent, C Collins, J Fuscoe, R Segraves, J Lucas, and J Gray. Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences*, 85(23):9138–9142, 1988.

[86] Lars K Poulsen, Gwyn Ballard, and David A Stahl. Use of rrna fluorescence in situ hybridization for measuring the activity of single cells in young and established biofilms. *Applied and Environmental Microbiology*, 59(5):1354–1360, 1993.

[87] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, et al. Full-length mrna-seq from single-cell levels of rna and individual circulating tumor cells. *Nature biotechnology*, 30(8):777–782, 2012.

[88] Nadine Randel, Luis A Bezares-Calderón, Martin Gühmann, Réza Shahidi, and Gáspár Jékely. Expression dynamics and protein localization of rhabdomeric opsins in platynereis larvae. *Integrative and comparative biology*, 53(1):7–16, 2013.

[89] Rayner Rodriguez-Diaz, Midhat H Abdulreda, Alexander L Formoso, Itai Gans, Camillo Ricordi, Per-Olof Berggren, and Alejandro Caicedo. Autonomic axons in the human endocrine pancreas show unique innervation patterns. *Cell metabolism*, 14 (1):45, 2011.

[90] Greg W Rouse. Trochophore concepts: ciliary bands and the evolution of larvae in spiralian metazoa. *Biological Journal of the Linnean Society*, 66(4):411–464, 1999.

[91] O. Rubel, G. H. Weber, M. Y. Huang, E. W. Bethel, M. D. Biggin, C. C. Fowlkes, C. L. Luengo Hendriks, S. V. Keranen, M. B. Eisen, D. W. Knowles, J. Malik, H. Hagen, and B. Hamann. Integrating data clustering and visualization for the analysis of 3D gene expression data. *IEEE/ACM Trans Comput Biol Bioinform*, 7 (1):64–79, 2010.

[92] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–504, 2003.

[93] Kazuo Shinozaki, Kazuko Yamaguchi-Shinozaki, and Motoaki Seki. Regulatory network of gene expression in the drought and cold stress responses. *Current opinion in plant biology*, 6(5): 410–417, 2003.

[94] Edwin Mellor Southern. Detection of specific sequences among dna fragments separated by gel electrophoresis. *Journal of molecular biology*, 98(3):503–517, 1975.

[95] Anders Ståhlberg, Vendula Rusnakova, and Mikael Kubista. The added value of single-cell gene expression profiling. *Briefings in functional genomics*, 12(2):81–89, 2013.

[96] Detlev Stalling, Malte Westerhoff, Hans-Christian Hege, et al. Amira: A highly interactive system for visual data analysis. *The visualization handbook*, 38:749–67, 2005.

[97] Badri Narayan Subudhi, Francesca Bovolo, Ashish Ghosh, and Lorenzo Bruzzone. Spatio-contextual fuzzy clustering with markov random field model for change detection in remotely sensed images. *Optics & Laser Technology*, 57:284–292, 2014.

[98] RR SWIGER and JD TUCKER. Fluorescence in situ hybridization: A brief review. *Environmental and molecular mutagenesis*, 27 (4):245–254, 1996.

[99] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[100] Diethard Tautz and Christine Pfeifle. A non-radioactive in situ hybridization method for the localization of specific rnas in drosophila embryos reveals translational control of the segmentation gene hunchback. *Chromosoma*, 98(2):81–85, 1989.

[101] Akihisa Terakita. The opsins. *Genome biology*, 6(3):213, 2005.

[102] Kristin Tessmar-Raible and Detlev Arendt. Emerging systems: between vertebrates and arthropods, the lophotrochozoa. *Current opinion in genetics & development*, 13(4):331–340, 2003.

[103] Kristin Tessmar-Raible, Florian Raible, Foteini Christodoulou, Keren Guy, Martina Rembold, Harald Hausen, and Detlev

Arendt. Conserved sensory-neurosecretory cell types in annelid and fish forebrain: insights into hypothalamus evolution. *Cell*, 129(7):1389–1400, 2007.

[104] Raju Tomer, Alexandru S Denes, Kristin Tessmar-Raible, and Detlev Arendt. Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell*, 142(5):800–809, 2010.

[105] Hsien-yu Wang, David C Watkins, and Craig C Malbon. Antisense oligodeoxynucleotides to gs protein α-subunit sequence accelerate differentiation of fibroblasts to adipocytes. 1992.

[106] Qi Wang, Qun Liang, and Xiuqing Zhang. 3d genome tuner: Compare multiple circular genomes in a 3d context. *Genomics Proteomics Bioinformatics*, 7(3):143–146, 2009.

[107] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[108] Harold Weintraub, Robert Davis, Stephen Tapscott, Matthew Thayer, Michael Krause, Robert Benezra, T Keith Blackwell, David Turner, Ralph Rupp, Stanley Hollenberg, et al. The myod gene family: nodal point during specification of the muscle cell lineage. *Science*, 251(4995):761–766, 1991.

[109] Erik Wilde. Putting things to rest. *Transport*, 15(November): 567–583, 2007.

[110] Peter L Williams et al. Gray's anatomy. 1980.

[111] Fa-Yueh Wu. The potts model. *Reviews of modern physics*, 54(1): 235, 1982.

[112] Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, Ying Lian, Xiaoying Zheng, Jie Yan, et al. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 2013.

[113] Barbara Young, Phillip Woodford, and Geraldine O'Dowd. *Wheater's functional histology: a text and colour atlas*. Elsevier Health Sciences, 2013.

[114] Alan L Yuille. Generalized deformable models, statistical physics, and matching problems. *Neural Computation*, 2(1):1–24, 1990.

[115] Erich Zeeck, Tilman Harder, and Manfred Beckmann. Uric acid: the sperm-release pheromone of the marine polychaete platynereis dumerilii. *Journal of Chemical Ecology*, 24(1):13–22, 1998.

[116] Hua Zhang, Wenzhong Shi, Yunjia Wang, Ming Hao, and Zelang Miao. Spatial-attraction-based markov random field approach for classification of high spatial resolution multispectral imagery. *Geoscience and Remote Sensing Letters, IEEE*, 11(2):489–493, 2014.

[117] Jun Zhang. The mean field theory in em procedures for markov random fields. *Signal Processing, IEEE Transactions on*, 40(10): 2570–2583, 1992.