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

Clustering and visualizing functionnal tissues in *P. dumerillii*

JEAN-BAPTISTE OLIVIER GEORGES PETTIT

UNIVERSITY OF CAMBRIDGE

2014 – version 0.9

[January 29, 2014 at 16:54 – classicthesis version 0.9]

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

Part I

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

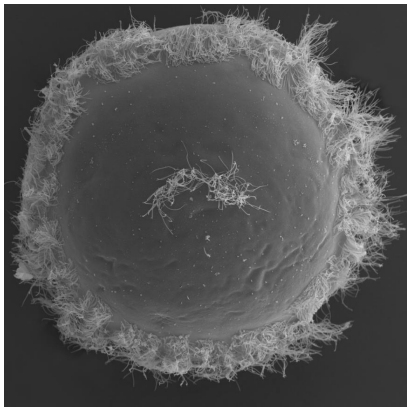
1 CAPTURING GENE EXPRESSION IN *PLATYNEREIS DUMERILLII*'S BRAIN

1.1 PLATYNEREIS DUMERILII, AN IDEAL ORGANISM OF BRAIN DEVELOPMENT STUDIES

1.1.1 *General description*

Platynereis dumerillii is a marine annelid of the class Polychaeta, it has been established as one of the main marine animal models in the fields of evolutionary, developmental and neurobiological biology as well as ecology and toxicology [29, 62, 26, 16, 19, 20]. As a member of the bilateria *P. dumerillii* has a defined bilateral symmetry.

P. dumerillii populates shallow (no more than 3m) hard 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 160 µm 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 1: *Platynereis dumerillii*'s larva and adult forms.

They are several reasons why *P. dumerillii* has been chosen as a model by numerous laboratories. In terms of evolution *P. dumerillii* shows several interesting characteristics. It belongs to the lophotrochozoan taxon of the bilaterian animals as opposed to most of the well established models animal which either belong to the ecdysozoans (*Caenorhabditis elegans*, *Drosophila melanogaster*) or the deuterostomes (mouse, human). Lophotrochozoans being extremely under represented, *P. dumerillii* as a model organism is essential to comparative approach

on bilaterian biology [20].

P. dumerillii also shows an exceptionally slow evolutionary lineage. It has even been described as a "living fossil" for that reason [20]. Therefore, the ancestral developmental characteristics exhibited by *P. dumerillii* translate into an image of the common past of all bilaterians. For example, an interesting example described in [14, 63] 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 evolutionary rate confers to *P. dumerillii* the advantage of being a link between fast evolving models like *drosophila* and vertebrates.

In terms of practicality, *P. dumerillii* can easily be kept and bred in captivity producing offspring throughout the year [19]. The behavioural characteristics of *P. dumerillii* mating ritual have been well studied. The "nuptial dance" happens on the water surface. Male and female release the sperm and eggs synchronously, respectively. This activity is synchronized by pheromones released into the water [68]. 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.

1.1.2 *Larval development*

Similarly to the other polychaetes, the larval development of *P. dumerillii* can be decomposed into three main anatomical stages: the trochophore, the metotrochophore and the nectochaete. The trochophore is spherical and moves via a equatorial belt of ciliated cells as well as an apical organ possessing a ciliary tuft [53, 46] as seen on figure 1a and schematically on figure 2. the metotrochophore stage is characterized by the development of a slightly elongated segmented trunk compared to that of the trochophore [25]. The next stage is the nectochaete larvae that resembles the adult (figure 1b) in most of the traits especially with parapodial appendages used for swimming and crawling [25]. This traditional subdivision has been applied to *P. dumerillii* [28].

Aside from this purely anatomical subdivision, 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 $\approx$ 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

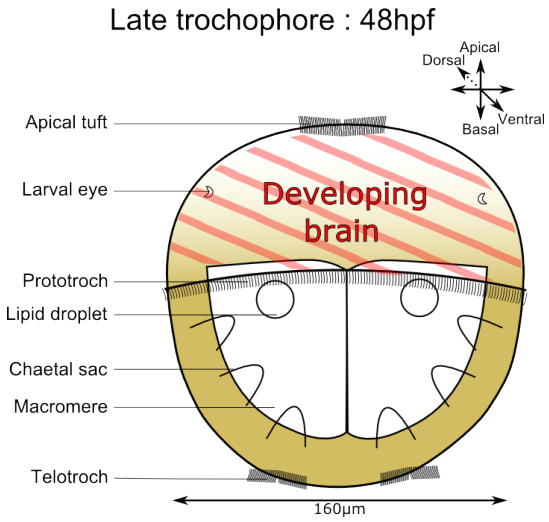


Figure 2: *Platynereis dumerillii*'s larvea development at 48hpf or late trochophore. Striped in red is indicated the area which forms the developing brain of the larvae.

pattern with very little variation from one individual to the other and even between batches provided the temperature is kept constant [19, 16]. An example showing the similarity between individuals during development can be seen on figure 3. 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.

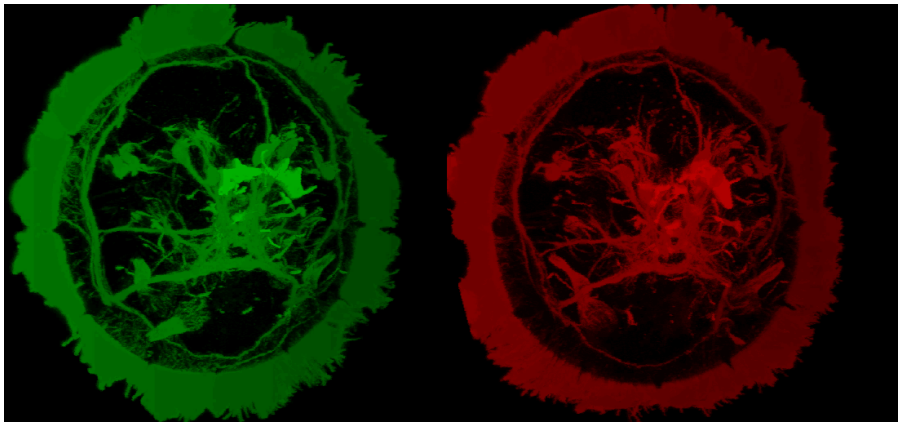


Figure 3: *Platynereis dumerillii*'s stereotypical and synchronous development. In green and red are two different *P. dumerillii* individuals' with the same gene expression being highlighted. They show extremely similar patterns of development.

Describing the entire development of *P. dumerillii* does not fall within the scope of this thesis. Indeed, we 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.

1.2 GENE EXPRESSION IN PLATYNEREIS' DEVELOPING BRAIN

1.2.1 *Platynereis' nervous development until 48hpf*

The main purpose of this thesis is not to fully understand the patterns of development in *P. dumerillii*'s larval brain. Therefore we will only give a brief summary of what the main component of the brain are at 48hpf, the time point we will be interested in in the next chapters. *P. dumerillii*'s larval brain development is detailed in [20].
From the early trochophore (24-26hpf) neural system development starts taking place. The apical ganglion forms at the apical tuft. It contains one sertonergic cell and a few neurons linked to the nerve of the ciliary band of the larva called the prototroch (see figure 2). This allows the first movements of the larve thanks to the cialited cells of the prototroch.
The mid-trochophore (26-40 hpf) sees the formation of the first cerebral commissure, it is a band of nerves interconnecting the ventral nerve cord and the brain. This trait is a typical feature of annelids' brains. 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 neutrites [20].
The data we will use in the rest of this thesis will not encapsulate the whole larvae, just the developing brain (see figure 2) thus excluding the ventral part of the nervous system. The best studied areas of the developing brain are the larval eyes, the developing adult eyes, the apical organ all located 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 [64].

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.

1.2.2 *Spatial organization of complex biological tissues like the brain*

This section is not intended to demonstrate a specificity of the *P. dumerillii*'s brain, it is meant to ask some of the fundamental questions that intrinsically motivate the work presented in the rest of this thesis. Complex tissues, the obvious example of which is the brain, could be viewed as an interconnected mosaic of cells having different functions, working together to achieve the global function of the organ.

If we look closely at this mosaic of cells, it is easy to observe that the spatial organisation is not random. Indeed, cells that serve the same function will often be close to each other, thus defining functional tissues. However, the spatial coherency of those tissues is not necessarily always the same. Some cell types may consist of cells that are scattered inside another more spatially coherent tissue. To illustrate that fact, 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 a 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 neurones will decrease this probability. Indeed, axons will grow through other types of tissues to reach their destination [10, 13], making the overall spatial coherency of neural tissues smaller than very well spatially defined tissues.

When trying to analyse the full structure of the brain with an automated method, keeping in mind that fact could prove important to improve the results. This fact and its consequences on the work presented in this method are further discussed in section cite section spatial clustering.

So far, we have only regarded organs and cell types as regarding their anatomical traits. But as mentioned in the introduction **??** the functional heterogeneity of complex tissues goes further that simple anatomical traits. We need to work on traits that fundamentally represent how cells are functioning.

1.2.3 *Generalities about gene expression and development*

Throughout this thesis the term cell will be 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 (what makes a neurone become a neurone and not a pancreatic cell) ?. Answering this sort of questions defines the field of developmental biology.

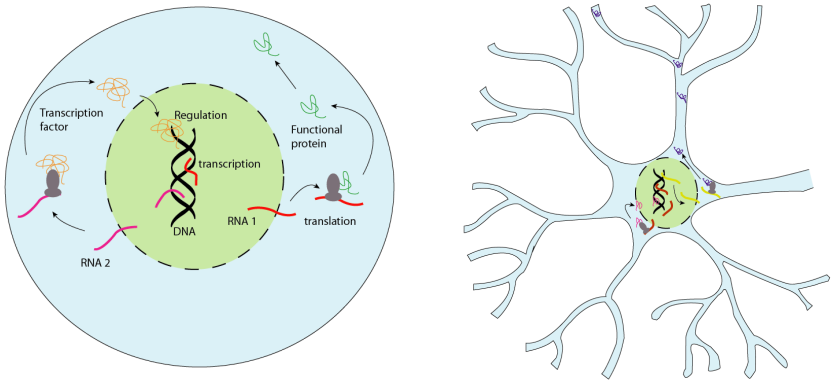
The short and rather complete answer to any developmental biology question actually is: same genome but different pattern of gene expression. As gene expression is the central, most important, most studied cellular activity, gene expression is indeed the common denominator of life as large parts of the mechanisms making up gene expression are actually shared by every living creature known to man.

Of course to understand what gene expression is, we must first define what genes are. 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 [41] 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 recent published definition of a gene followed the publication of the ENCODE project [18]. It states that a gene is "a union of genomic sequences encoding a coherent set of potentially overlapping functional products."

Gene expression is the way cells express their genes. Expression of a gene is the process of transcribing the DNA of that particular gene. The product of gene expression is RNA molecules and there are several ways to look at gene expression. In a cell or tissue, at a given time point we can to choose to look whether a gene is expressed or not (binary expression) or how much a certain gene is expressed (quantitative expression).

Most 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 [11]), others will be excreted by the cell and will serve a purpose at the scale of the organism [34] others called transcription factors will have a regulatory effect on gene expression [42]. In other terms the expression gene G_a, coding for protein P_a might activate, accelerate, inactivate or decelerate expression of gene G_b and potentially others. This outlines the complex interdependent regulatory system that is gene expression, see figure 4. For precise examples gene regulation see [23, 56, 22, 9].

During development mechanisms exist that allow gene expression to become differential as the divisions occur. This is how the asymmetrical axis (dorso-ventral, and basal-apical) of the body are defined.



(a) Undifferentiated cell with a circular shape.

(b) Differentiated neuron, with typical axons development.

Figure 4: Cell types anatomical heterogeneity, 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.

The main mechanism involves chemical gradients. The first of these gradient has to come from the original cell which must contain some asymmetrically distributed chemical so that the first divisions lead to non identical cells. In the case of *Platynereis dumerillii*, the body axis are defined between 2hpf and 7hpf [20].

As described, gene expression is the key factor during tissue development. 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 we will present two methods to capture gene expression.

1.3 CAPTURING GENE EXPRESSION IN THE LABORATORY

1.3.1 *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 [57], ISH assays not only allow to know whether a gene is expressed or not, but also where in the tissue it is expressed. First proposed in 1969 by Pardue [48] and John [33] independently, in-situ hybridization (ISH) used radioactive tritium labelled probes on a photographic emulsion to reveal parts of the studied tissues where particular RNAs or DNA regions were present. With the development of fluorescent labelling techniques [37, 50] allowing for faster, more sensitive and of

course safer hybridization assays compared to radioactive probes [59], Fluorescent in-situ hybridization (FiSH) quickly became the standard technique to study gene expression in the spatial context of the biological tissue. Importantly, using multiple fluorescent probes of different colours allowed the simultaneous localization of several RNA fragments in the tissues [45].

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

During his PhD, Raju Tomer and other cite thesis members of the Detlev Arendt lab in EMBL, used Fluorescent in-situ hybridization to create an image library of gene expression in the brain of *P. dumerillii*. He was able to record gene expression in the full brain at 48hpf for 169 genes. In practice each individual larvae was dissected to isolate the brain, which was then cut into thin slices and fixed. Each individual slice was then stained with two different fluorescent probes corresponding to two messenger RNAs (mRNA). One of the gene is considered a reference, as it is always hybridized in all the assays (the main reference gene used was Emx) alongside an other gene of interest, see figure 5.

As mentioned previously, the larval development of *P. dumerillii* is highly similar in every individual larvae. In the case of this study requiring a lot of different assays conducted each time a on different animal, the stereotypical development of *P. dumerillii* has proven essential. Indeed, having the same reference localized in all the assays has allowed Tomer to align all the 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 fluorescent in-situ hybridization may be, such assays are limited in terms of the quantity of gene one is able to study. Indeed, each individual larvae only provides the expression of two genes, one being the reference. Crucial developments in sequencing technologies have brought a way to study the expression of the whole transciptome landscape in a single assay, RNA sequencing.

1.3.3 *RNA sequencing*

Whole Transcriptome Shotgun Sequencing (WTSS) also called RNA sequencing (RNA-seq) [43, 65] has developed alongside Next Generation Sequencing (NGS) techniques used to retrieve the genome (DNA). Instead, when preparing the starting material, only the RNAs are extracted. Protein coding mRNA can further be selected, they are separated from the rest by targeting the polyadenylated 3' tail, spe-

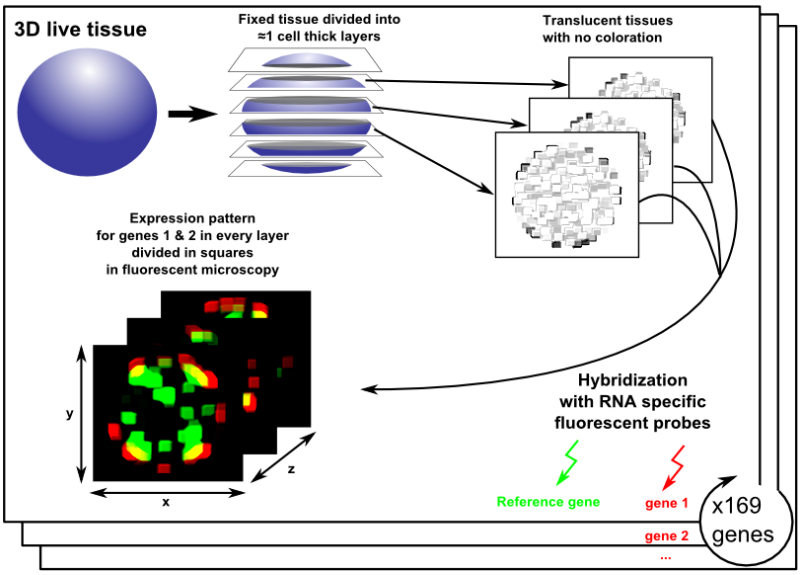


Figure 5: Fluorescent in-situ hybridization assays to create a 169 genes catalogue of gene expression in the brain of *P. dumerillii*. From the live tissue cut into thin fixed layers, every slice is stained with a reference gene and a gene of interest that will reveal the areas of expression under fluorescent microscopy. The process repeated 169 times for key genes in *P. dumerillii* development has been generated by [64]

cific to protein coding transcripts. Most current technique use magnetic beads to achieve this separation [44, 43].

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 will create 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.

This will generate a large dataset of small reads, that needs to be mapped back onto the reference genome of the considered species, providing this genome is available. In that case, the resulting dataset will reflect a snapshot of the whole transciptome in the studied cell population. However, in the case of *P. dumerillii*, this reference genome is not fully available yet, an alternative option is to map the reads back to a list of known gene sequences, for instance the 169 genes studied by [64] (PrimR genes). 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 this sequencing protocol, 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 cell 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 the 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.

1.4 CONCLUSIONS

In this chapter we have presented *Platynereis dumerillii* and the advantageous traits it exhibits for developmental biologists especially in the field of neural development. We 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 go push this understanding further we need to take an interest in what defines the life of tissues and sub-tissues, gene expression. We have also described two methods that allow practitioners to capture gene expression from a biological tissue, and how an image library of gene expression for 169 gene was generated by [64] in the full brain of *P. dumerillii*.

So far, our scale of study has been the tissue, or the sub-tissue. However, as mentioned in the introduction **??**, 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 we can work with, the single cell, define the heterogeneity of gene expression at the single cell level and work our way up to the functional tissue level. Instead of a fragmentation problem, we would have a clustering problem, attaching single cells to a certain number of categories. In order to implement such an approach, what we need is single cell gene expression data.

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 keeps the studied tissue spatially untouched, achieving single cell gene expression resolution from one image obtained through fluorescent microscopy is a matter of microscope performance and cell size. For big enough cells, single cell resolution has been documented as far as 1989 [61] with some work specifically directed towards achieving this single cell resolution [51].

When considering [64] dataset, with current microscope technology, achieving single cell level resolution in *P. dumerillii*'s brain on one particular image is feasible. However, our main limitation is the quantity of data involved, indeed, each brain is separated into 20 slices, for 169 genes. 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 [39, 17, 8].

Unfortunately, a dataset with the cells limits highlighted does not yet exist for *P. dumerillii*'s brain, making a precise division of the images into cells very difficult. Instead, Tomer used a basic approach to divided the images, the "cube" model [64].

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 1.3) for all 169 genes, the "cube" model simply consists in dividing each image into squares approximately the size of an average cell. In our dataset, the size chosen was 3 μm^2. 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, referenced on a 3-dimensional axis, will contain 3 μm^3 cubes, each of which is attached to the luminescence data for 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 florescence my contaminate the cubes around that do not necessarily express the same gene see figure 6A. 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 then travel in 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 6B.

Hence the data will tend to exhibit spatial discontinuity and inconsistency. With that fact in mind, any method that we develop using this data, 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 [64] 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*, it also attaches spatial information to each data point.

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

2.2.1 *Sequencing single cell RNA contents*

The scale shift from tissue to single cell is harder to achieve in the case of RNA-seq. As described in the previous section 1.3, an important

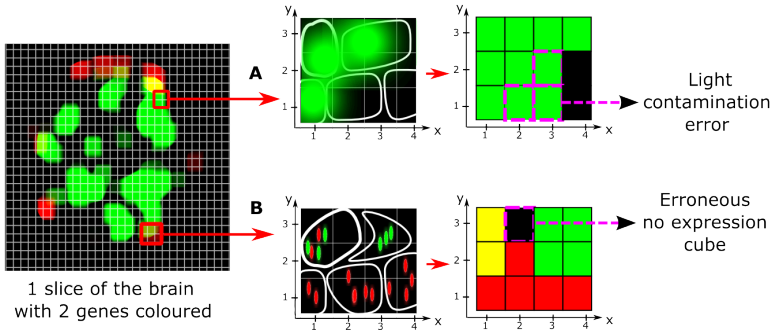


Figure 6: Errors introduced by the “cube" cell model. Path A shows how regions with highly expressed genes can introduce errors through light contamination. Path B shows how some ubes may appear artificially void of expression because of the not even distribution of transcripts inside the cytoplasm especially for large cells.

factor for the success of RNA-seq assays is the starting 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 30 pg per cell, only 2% of which is mRNA [30, 31]. With such a small input quantity, distinguishing biological variation between different cells from the technical variation linked to cDNA amplification protocols has long been impossible to achieve.

However, with the creation of new protocols [52, 60], and the rise of microfluidics to faciliate the extraction and sequencing of single cells [47], the last couple of years have seen a dramatic increase in the number of single cell RNA-seq based studies [32, 40, 67, 58, 15]. However, challenges are yet to be overcome to be able to analyse further complex tissues from whole transcriptomes obtained from single cell RNA-seq, one of which is the loss of spatial reference induced by the current protocols.

2.2.2 *Mapping back gene expression to a spatial reference*

Single cell RNA-seq achieves to capture 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 from keeping track of any spatial information about the cells. Hence, when analysing single cell RNA-seq data from a complex tissue, we need to be able to map back every cell to its original location.

In order to achieve this back-mapping, a reference is needed. This reference should consist in 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 the ex-

act original location of the cell, 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 we 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.

2.3 ABOUT THE QUANTITATIVE TRAIT OF SINGLE CELL EXPRESSION DATA

2.3.1 *Light contamination in in-situ hybridization data*

The fluorescence value obtained from in-situ hybridization assays can be considered as quantitative [16]. Indeed, the light intensity emitted by every cell in the considered tissue is correlated with the number of RNA fragments 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 expressing 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 punctual errors due to the cell model. One of the culprit for those errors, as shown on figure 6B is the phenomenon of light contamination. When a large group of neighbouring cells express the same gene, because of the additivity of light intensity mentioned above, 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.

As shown on figure 7, we can confirm that light contamination issue on the in-situ hybridization data in *P. dumerillii*'s brain. In that context, and because of the single cell scale of our study, considering the in-situ hybridization data as quantitative may have introduced significant errors. In order to avoid this light contamination bias we decided to transform the quantitative data set into a binary data set where for a given "cube", genes are simply expressed or not. The binarization method is described in the following section 2.4.

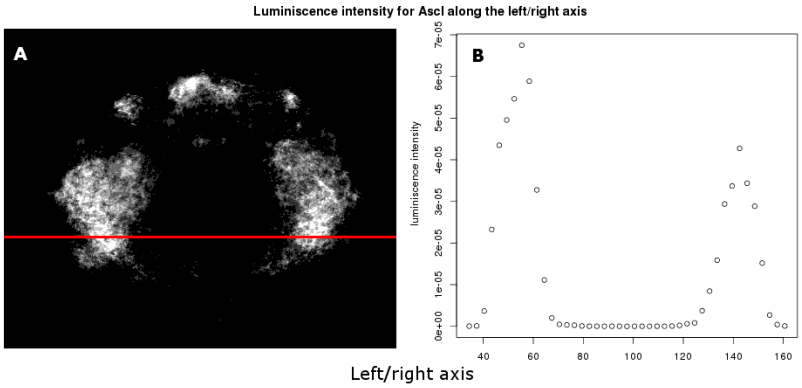


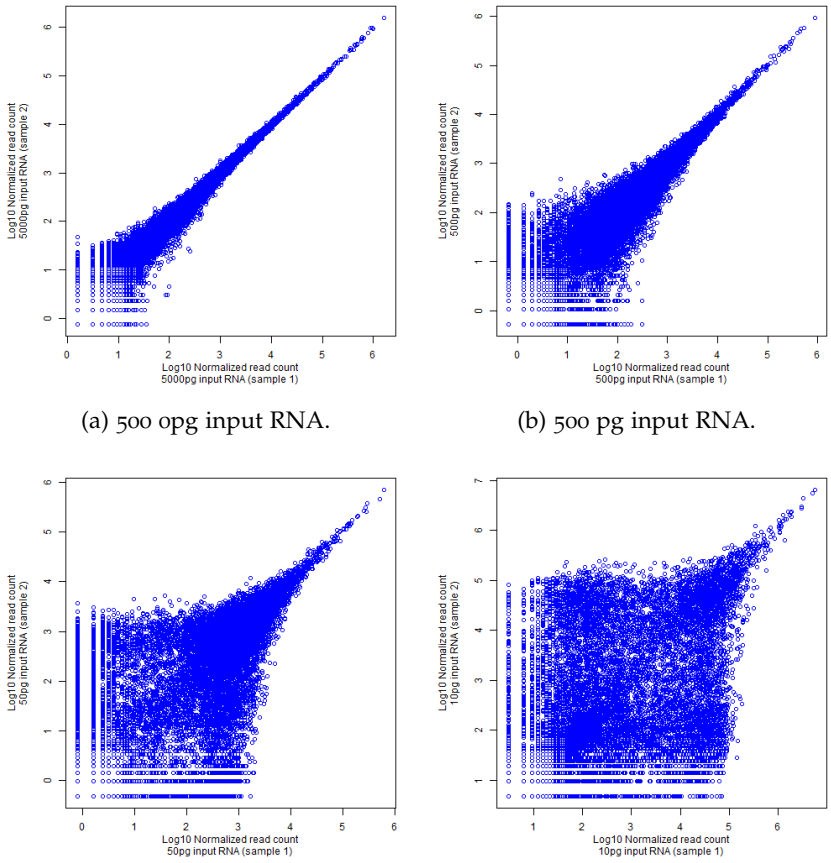
Figure 7: **Light contamination on in in-situ hybridization luminescence data seen on the example of gene Ascl.** Panel A shows the raw fluorescent microscopy capture of the gene's expression for one layer in the brain of *P. dumerillii*. Panel B shows the light intensity measured along the red line in panel A. Because of the small scale of study, cells surrounded by other cells expressing the same gene will have a higher intensity values because of nearby light contamination. Even though there might be an actual gradient in the expression level bewteen the cell in the middle and the others there is no option to separate it from the light contamination component.

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 lead by Philip Brennecke1, Simon Anders and Jong Kyoung Kim [12], 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 series of dilution assays, reducing step by step (5000 pg, 500 pg, 50 pg, 10 pg) the input quantity of RNA fragments extracted from *Arabidobsis thaliana* with two technical replicates each time using the Tang protocol [60]. The authors of the study let us analyse this data, and we were able after normalizing by the size factor using the Bioconductor package DESeq [7] to generate the scatter plots shown on figure 8.

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.



(a) 500 opg input RNA.

(b) 500 pg input RNA.

(c) 50 pg input RNA.

(d) 10 pg input RNA.

Figure 8: Dilution series of total A. thaliana RNA

2.3.3 *Conclusions*

The paragraphs above have shown that neither in-situ hybridization nor RNA-seq data can be safely assumed as quantitative when the scale is lowered at the single cell level. To avoid nonsensical conclusions linked to the noise level in the rest of the study, a solution was to binarize any single cell dataset that we were dealing with. 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 7 and discussed in the previous section 2.3, we decided to avoid the various problems linked to light contamination by 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, we found 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 9. After trying different thresholding methods based on those densities, we found that the resulting binary expression was not satisfying for a large number of genes. 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 biologist working specifically on *Platynereis dumerillii*'s brain, we decided to opt for a manual approach to thresholding. Indeed, by going through the 169 genes one by one, we adjusted the threshold manually until the resulting binarized expression patter 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* validated by the biologists.

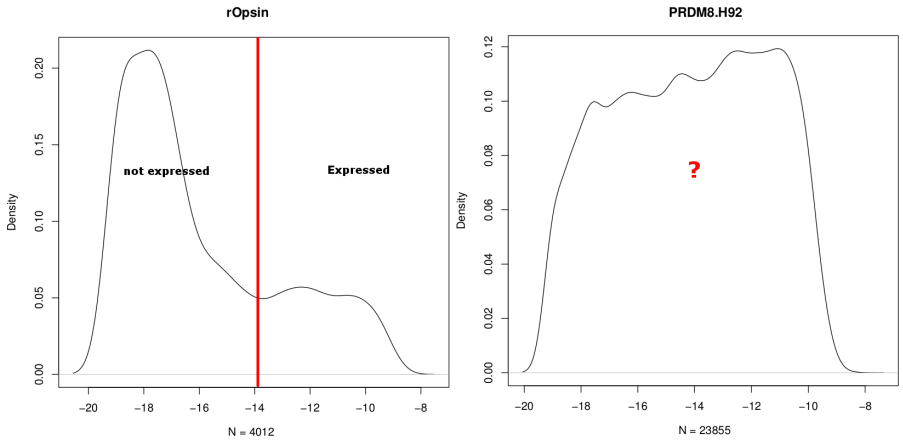


Figure 9: Light intensity densities found for genes rOpsin and PRDM8.H92 accross *P. dumerillii*'s whole brain on a logaritmic scale. N for each graph is the number of "cubes" in the dataset where the fluorescence value is higher than 0. On the one hand, the log density shows two clear peaks for rOpsin, making the choice of a expression threshold easy. PRDM8.H92 on the other hand does not display such a clear cut threshold.

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 resulting in blurred and unexploitable expression patterns. Finally some images where polluted by a well known experimental artefact linked to fluorescent microscopy imaging.

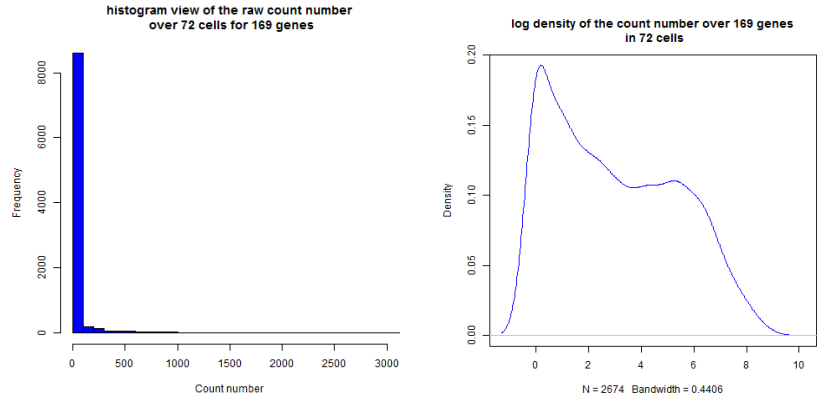
Although the aforementioned method resulted in a high quality binary dataset, it has been possible only because the number of genes considered was small. This will not be the case when dealing with RNA-seq data.

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. As we did not have access to a large single cell RNA-seq dataset to test the methods presented here, we will discuss possible ways to binarize single cell RNA-seq data, presenting some results from a small number (72) of sequenced cells in the brain of *P. dumerillii*, containing the count number for 169 genes (see next section 2.5 of a detailed presentation of this 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 noise level generated by the currently available sequencing technologies would prove too high to rely simply on this method. However, as a first approach on our dataset, on figure 10a see as predicted a very dominant peak for the value 0. The problem remains that for very small count numbers it seems dangerous to set the gene as expressed.

Another option would be to find a global threshold over the complete dataset. The threshold $T > 0$ would represent the count number of RNA fragments 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 curve, one peak would correspond 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) 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 10: Thresholding RNA sequencing data for *P. dumerillii*

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 10b 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 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 accross all cells. However, with 72 cells into consideration, considering the sparse nature of the count data, we cannot show any results with this method on our dataset. We believe however, 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 MAPPING SINGLE CELL RNA-SEQ DATA IN FROM PLATYNEREIS' BRAIN

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

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

Experimentally, the done 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 (breaks extracellular matrix) and thioglycolate (helps to break the chorion). After this treatment, the trunks and epispheres 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 washing, the cells were dissociated by pipetting up and down then washed again 1 ml of 1xPBS and concentrated by centrifuging (1 min, 1000 rpm). Cells were re-suspended in 20 microliters of 1xPBS, 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. 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 tubes from 2 cells, one with 4 cells, and 3 unsure ones from the second chip resulting in 72 raw reads files.

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, we need to be able to map back the single cells to their original location in the brain. To that end, we took advantage of the spatially localized in-situ hybridization described in the previous section.

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

We started by mapping the RNA-seq reads to the 169 reference genes composing the in-situ hybridization data using Bowtie.cite nuno pipline. The resulting dataset is the count number for each of the 169 genes in the 71 cells sequenced. In order to map back to the in-situ hybridization data, our approach consisted in extracting the genes that were the most specifically expressed for each sequenced cell, then compare 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.

Given the set of 169 considered genes M, and the set of 72 cells C, with the read counts matrix D of size $M \times C$ we can compute for each cell and each gene, the expression specificity ratio $r_{m,c}$ defined as :

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

By looking for each cell, at the genes with the highest ratio we can define the what genes are the most specifically expressed in this cell compared to the other cells sequenced. On the one hand, this mapping method has the inconvenient of using the average expression level across all considered cells to compute the ratio r. This means that the mapping quality of each cell will depend on the overall sequencing quality. Furthermore, this method performance relies on the assumption that we do have in fact a collection of cells from different cell types, 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 explains 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. We 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 1 are shown the most specific genes for four cells sequenced. With that list of genes that are supposed to be the most representative ones for the four cells considered, we visualized in 3D the areas where those genes are co-expressed in the brain of *P. dumerillii* according to the in-situ hybridization data. A snapshot of this visualization is shown on figure 11. 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, we can hypothesize from previous biological studies the cell type of each sequenced cell. As shown in table 1, for cell "X2C911L" the most specifically expressed gene "Emx" has been used as a reference gene to localize the mushroom bodies, an hypothesis which is compatible with the co-expression of "CALM.R29" and "Dach" [64]. Cell "X2C521L" expresses Wnt8 very specifically, a gene shown to

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 1: 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 on figure 11.

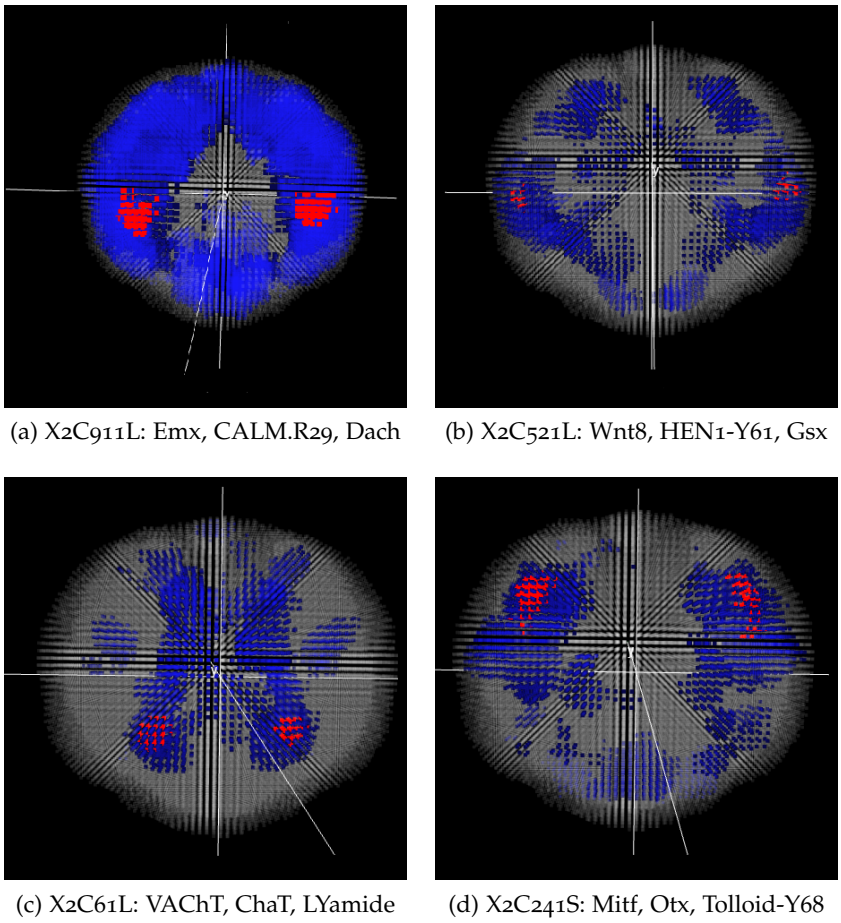
be linked to lateral brain development. Cell “X2C61L" can be easily classified as a developing neuron. Indeed both VACht (Vesicular acetylcholine transporter) and ChaT (Choline acetyltransferase) are genes coding for enzymes interacting with the neurotransmitter acetylcholine. Finally cell “X2C241S" displays the specific expression of the gene “Mitf", one of *P. dumerillii* most studied gene and expressed solely in the developing adult eyes [35, 24].

2.6 CONCLUSIONS

In this chapter we presented how the scale of gene expression studies has shifted from the tissue level to the single cell level. For two experimental protocols we have described how such data can be obtained and why at the time of writing, it is still not safe to assume the single cell datasets as quantitative. To avoid that problem, turning those datasets into binary gene expression is an attractive solution. However the binarization process is not trivial and we 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, we were able to replace single cell RNA-seq data in a spatial context, using the most specifically expressed genes in every cell. To our best knowledge an a posteriori localization of single cell RNA-seq data had never been presented before.

As mentioned previously 1, if the ability to study the heterogeneity of cell populations at the single cell level is offers incredible possibilities for the future of developmental biology and 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.



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

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

(c) X2C61L: VAChT, ChaT, LYamide

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

Figure 11: Regions defined by the expression overlap of the top 3 scoring genes in [64] 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.

The work presented thereafter was done to answer simple but important questions: can we define and localize known functional tissues of a complex organ like the brain from single gene expression data? Can we detect unknown regions in such a complex tissue and finally can we hypothesize the functional role of those unknown regions based on the single cell expression data?

3

CLUSTERING TISSUES FROM SINGLE CELL EXPRESSION DATA AND VISUALIZING THEM IN A 3D SPACE

3.1 ELEMENTS OF CLUSTERING FOR BIOLOGICAL TISSUES

3.1.1 *Why cluster?*

- Single cell data has a lot of potential but: - Big data general problem - need to be able to come back from the single cell level to the tissue for : a consistency check and improve the knowleddge at the tissue level from single cell level

3.1.2 *General considerations about clustering*

- No perfect method - directed vs undirected - choosing the number of cluster, a key issue

3.2 VISUALIZING CLUSTERING RESULTS IN 3D WITH BIOWEB3D

3.2.1 *Background*

Visualisation is a key feature in the analysis of large biological datasets, especially when analysing organized structures with distinct sub-clusters [54]. 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 the quality assessed, downstream analysis can proceed. For example, on obvious second step is to cluster the observations into different classes based upon the information assiciated 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}[21], Arena3D [49], 3D Genome Tuner [**?**], the Allen Brain Atlas [38] or Cytoscape [55]. Other 3D visualisation tools have been built online and are accessible through the browser directly, such as AstexViewer [27], which is utilised by the Protein Databank Europe via a Java Applet. More recently, visuali-

sation tools developed using HTML5/WebGL capabilities have been described, although they have focused on very specific applications, such as analysing radiology data [36].
Importantly, as yet no tool has allowed biologists to view their own 3D data directly online in an easy, fast and 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 files, a widely used structured format on the web [66] or directly as Comma Separated Values files (CSV).

The first file used by the application, referred to as the “dataset file", contains the 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 the points 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.

Datasets can be viewed and compared in up to four “worlds" (each world refers to a separate visualisation sub-window) at the same time. 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 links can be drawn between the points. In other situations, such as when a population of cells is considered, the points can be left unlinked as individual particles. The information layers are visualised by colouring the 3D points according to the class that each point belongs to.

TECHNOLOGICAL OVERVIEW 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]. 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.

DEFINING THE INPUT FILES FORMATS Using the JSON format to input files into bioWeb3D is recommended because of its rigorous structure, which allows fast Javascript object generation within the browser interpreter. Compared to other data-interchange languages, such as XML, JSON is also easily human readable thanks to a light-weight syntax. It is also supported by all of the primary internet browsers.
However, much data generated in the biological sciences is stored within CSV files. Converting CSV documents to the JSON format used in this application is not always trivial. In order to facilitate this process, the application is also able to interpret simple CSV files following a certain format as an input.

3.2.2.1 *Dataset files specification*

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

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 linked (the default value is *false*) - the data will be considered sequentially, with each point linked 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 **??** is an example of a minimal 3 points dataset file:

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

file in Listing **??**. Do note that although the spatial information remains the same it is not possible to set a name or to link the points within a CSV file input.

3.2.2.2 *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.

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 **??** and Listing **??**, 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

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

CSV FORMAT Each column will represent in which class each point belongs. The separation character between columns must be a “comma". Listing **??** carries the same information as Listing **??**. Note that it is not possible to use the “labels" or “name" properties available in Listing **??** within a CSV information layer file.

3.2.3 *Results and Discussion*

BASIC USAGE In figure 1 is visualised as an example, the brain the marine annelid *Platynereis dumerilii* each point represents a cell in the brain, the colour of which indicates the class it belongs to. The user

can interact with the visualisation via an interface on the right of the screen, which contains three panels. 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 simultaneous worlds. 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.

BIOWEB3D AND LOCAL SOFTWARE 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. 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 medical sciences [**?**]. In this context, we 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.

BIOWEB3D AND JAVA APPLETS 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.

CURRENT LIMITATIONS 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. It could

also be important to mention for eventual future developments that webGL is supported on mobile platforms such as iOS or Android. [2]

3.2.4 *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 the current limitations in terms of compatibility or rendering performances for large numbers of points, we 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 context, thus facilitating downstream analyses.

3.2.5 *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

3.3.1 *Hierarchical clustering*

- General description of the method - Computing distance matrix on binary expression data - Hclust method to use (full or partial) - Discuss the choosing the K with hClust (dendrogram is not informative)

3.3.2 *Independent mixture models*

- General statistical framework - Present the model - Gene independence hypothesis (this will be discussed further in the next chapter) - Likelihood of the model - EM algorithm to maximize the parameters (theta) - Choosing K with the BIC

3.4 DISCUSSION

3.4.1 *Spatial clustering techniques (hierarchical, model based)*

- Limits of non spatial methods on noisy data (cite the "cell model of chapter 1" paragraph) - Not using the spatial data seems silly when we do have it (playing chess blind comparison ?) - Some clustering methods are able to take into account the spatial localization of the

data points as well as the expression data - Quickly present spatial methods (spatial hclust), Mixture with a spatial component

3.4.2 *Method chosen*

- MRF because theta parameters are informative, easy to compute a likelihood and to choose K

Part II

APPENDIX

APPENDIX

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] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. doi: 10.1186/gb-2010-11-10-r106. URL http://genomebiology.com/2010/11/10/R106/.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

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

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

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

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

[58] 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.

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

[60] 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.

[61] 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.

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

[63] 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.

[64] 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.

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

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

[67] 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.

[68] 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.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

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