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 24, 2014 at 23:02 – classicthesis version 0.9]

CONTENTS

[January 24, 2014 at 23:02 – classicthesis version 0.9]

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

Part I

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

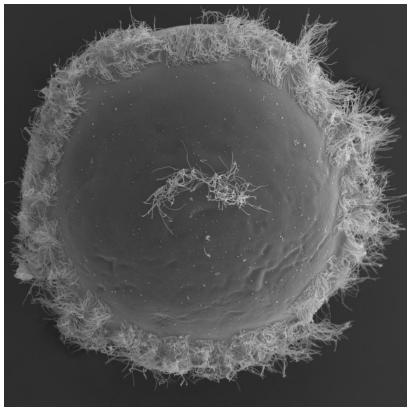
1 CAPTURING GENE EXPRESSION IN *PLATYNEREIS DUMERILLII*'S BRAIN

1.1 PLATYNEREIS DUMERILII, AN IDEAL ORGANISM OF BRAIN DEVELOPMENT STUDIES

1.1.1 *General description*

P. 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 [15, 39, 13, 5, 8, 9]. As a member of the bilateria *P. dumerillii* has a defined bilateral symmetry.

P. dumerilliipopulates 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.

P. dumerillii also shows an exceptionally slow evolutionary lineage. It has even been described as a "living fossil" for that reason [9]. This means that the ancestral developmental characteristics of *P. dumerillii* are at an image of the common past of all bilaterians. To illustrate this fact an interesting example described in [3, 40] 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 *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 [8]. The behavioural characteristics of *P. dumerillii* mating ritual have been well studied. The "nuptial dance" happens on the water surface, male and female releasing the sperm and eggs synchronously, respectively. This activity is synchronized by pheromones released into the water [44]. 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 [33, 27] 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 [12]. 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 [12]. This traditional subdivision has been applied to *P. dumerillii* [14].

Aside from this purely anatomical subdivision, an additional staging systems 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 pattern with very little variation from one individual to the other

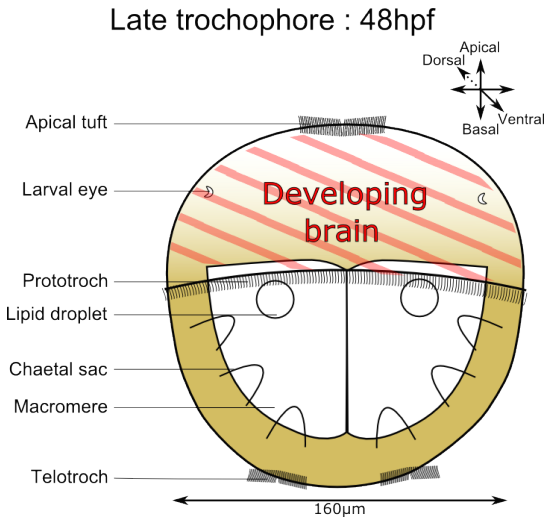


Figure 2: *Platynereis dumerillii*'s larvea development at 48hpf or late trochophore. Striped in red is indicated the area which forms the developing brain of the larvae.

and even between batches provided the temperature is kept constant [8, 5]. 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.

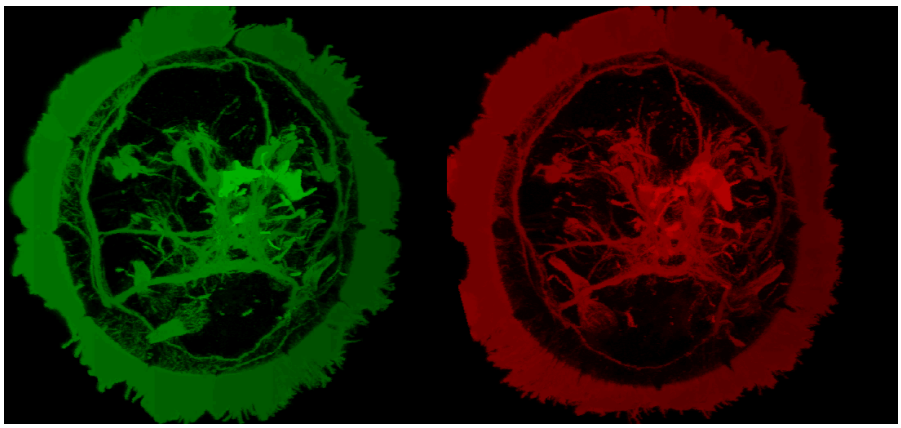


Figure 3: *Platynereis dumerillii*'s stereotypical and synchronous development. In green and red are two different *P. dumerillii* individuals' with the same gene expression being highlighted. They show extremely similar patterns of development.

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 [9].
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 that link to the nerve of the ciliary band of the larva called the prototroch (see figure 2 for better understanding). 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 brain starts to become more complexity with a notable increase in the number of neutrites.
The data we will use in the rest of this thesis will not encapsulated the whole larvae, just the brain (see figure 2) thus excluding the ventral part of the nervous system. The best studied areas of the brain are the larval eyes, the developing adult eyes, 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 [41]. A schematic representation of those areas is shown on figure FIGURE brain areas.

Even at a very early stage in a relatively simple organism, the brain quickly becomes a 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 about 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, the spatial organisation of this mosaic is not random. Cells that serve the same function will often be close from each other, thus defining functional tissues. However, the spatial coherency of those tissues is not necessarily always the same. Some cell types could be formed of cells 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 the mushroom bodies. When asking the question, is it likely that this cell is fully surrounded by 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, making the overall spatial coherency of "neuronal" tissue 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*

When speaking about developmental biology it it should be noted that the term “cell" will be referring 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 indeed gene expression is the central, most important, most studied cellular activity. Gene expression even is general 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 [23] 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 [7]. It states that a gene is "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 directly serve in the cellular life as functional/structural agents (elements of the ATP synthase for example) others will have a regulatory effect on gene expression. In other terms the expression gene a, coding for protein A might activate, accelerate, inactivate or decelerate expression of gene b and potentially others. This outlines the complex interdependent regulatory system that is gene expression. For precise examples gene regulation see [11, 34, 10, 2].

Add figure for gene expression Camille

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.

The main mechanism involves chemical gradients. The first of these gradient has to come from the original cell which musty 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 [9].

As described, gene expression is the key factor during tissue development. The ability to study gene expression patterns has revolutionized the fields 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 expressed. As opposed to Southern blotting, 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 [29] and John [19] independently, in-situ hybridization (ISH) used radioactive tritium labelled probes on a photographic emulsion to reveal on which chromosome particular genomic components were located. With the development of fluorescent labelling techniques [20, 30] allowing for faster, more sensitive and of course safer hybridization assays [36] compared to radioactive probes, 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 [26].

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

During his PhD, Raju Tomer and colleagues cite thesis member 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 (RNAm). One of the gene is considered a reference, as it is always hybridized in all the assays (the main reference gene used was Emx) along an other gene of interest,

see figure 4.

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 pattern in the full brain of *P. dumerillii* with a exploitable spatial reference that allows for a very precise mapping.

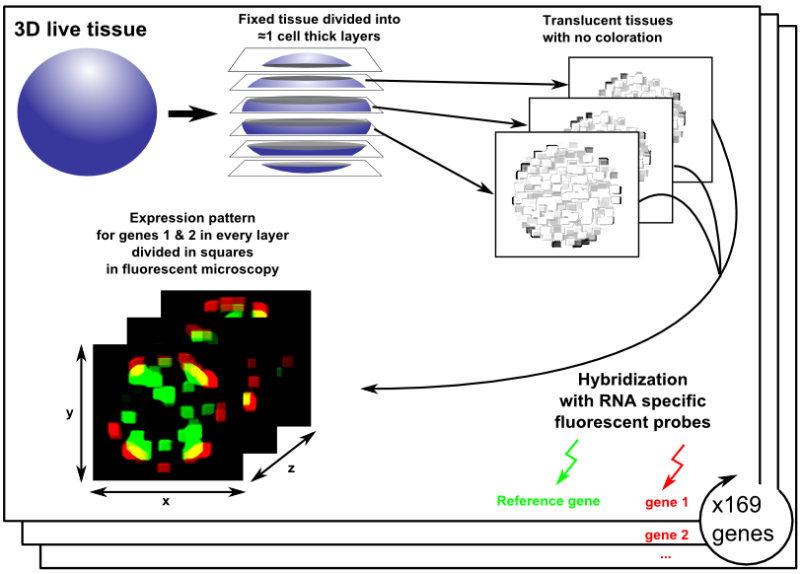


Figure 4: Fluorescent in-situ hybridization assays to create a 169 genes catalogue of gene expression in the brain of *P. dumerillii*. From the live tissue cut into thin fixed layers, every slice is stained with a reference gene and a gene of interest that will reveal the areas of expression under fluorescent microscopy. The process repeated 169 times for key genes in *P. dumerillii* development has been generated by [41]

However useful and practical fluorescent in-situ hybridization may be, such assays are limited in terms 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) [24, 42] has developed alongside Next Generation Sequencing (NGS) technique used to retrieve the genome (DNA). Instead, when preparing the starting material, only the RNAs are extracted. If interested in protein coding messenger RNA, they are separated from the rest by targeting the polyadenylated 3' tail, specific to protein coding transcripts. Most current technique use magnetic beads to achieve this separation [25, 24].

Once isolated from a population of cells, transcripts undergo fragmentation to obtain an average length of 200-300. The next step is then reverse transcription, which will create a complementary DNA (cDNA) library using a 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 need 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 being to map the reads back to a list of known gene sequences, for instance the 169 genes studied by [41] (PrimR genes). The obtain 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 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 functionality 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 [41] 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 tissue and then separating them in smaller sub-tissues until "true" functional tissues are defined is an extremely complicated approach. A solution to this problem would be to actually reverse the approach from a top-down to a bottom-up mindset. This mean 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

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 [38] with some work specifically directed towards achieving this single cell resolution [31].

When considering [41] 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.

Possibilities exit to highlight the limits of the cells and to automatically acquire those boundaries through computer vision. They rely 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 [21, 6, 1].

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 [41].

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 sim-

ply consists in dividing each image into square 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 those 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 are 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 5A. 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 5B.

Hence the data will tend to show some discontinuity and inconsistency spatially. With that fact in mind, any method hat 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.

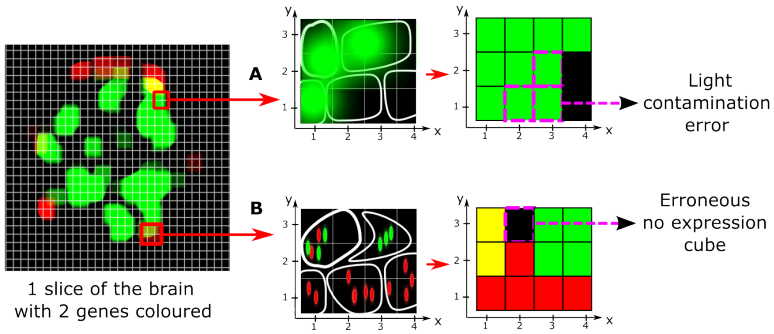


Figure 5: Errors introduced by the "cube" cell model. Path A shows how regions with highly expressed genes can introduce errors through light contamination. Path B shows how .

However, even with this simple cell model the data generated by [41] 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

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 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 [16, 17]. 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 [32, 37], and the rise of microfluidics protocol to faciliate the extraction and sequencing of single cells [28], the last couple of years have seen a dramatic increase in the number of single cell RNA-seq based studies [18, 22, 43, 35, 4]. Challenges remain 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.

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 small enough resolution spatially to find for each sequenced cell, if not the exact original location of the cell, at least a spatially restricted region of the brain 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

Light contamination in in-situ hybridization data

The fluorescence value obtained from in-situ hybridization assays can be considered as quantitative [5]. 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 sources 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 on 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 5B 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 reading 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 6, 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.

Technical noise in single cell RNA-seq data

- FIG 6 : show "typical" correlation plot from single cell RNA-seq with the noise increasing when reducing starting material - Both methods are currently unreliable quantitatively => need to binarize

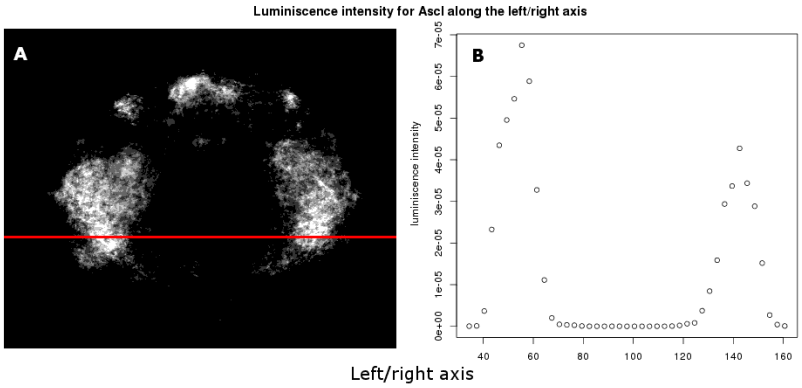


Figure 6: **Light contamination on in in-situ hybridization luminescence data seen on the example of gene Ascl.** Panel A shows the raw fluorescent microscopy capture of the gene's expression for one layer in the brain of *P. dumerillii*. Panel B shows the light intensity measured along the red line 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.

2.4 BINARIZING GENE EXPRESSION DATASETS

Binarizing in-situ hybrdization datasets

As shown in Figure 6 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 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 find a clear cut value as shown in figure 7. 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 dumerilii*'s brain, we decided to opt for a manual approach to thresholding. Indeed, by going through the 169 genes one by one, and adjusting 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.

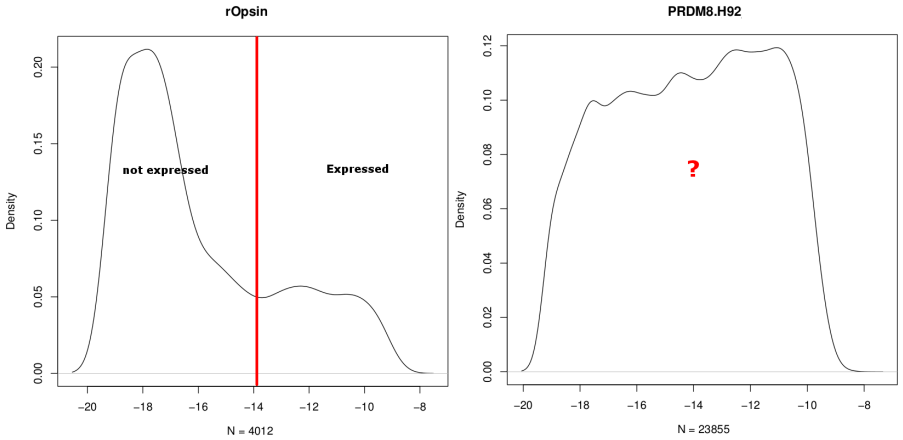


Figure 7: Light intensity densities found for genes rOpsin and PRDM8.H92 accross *P. dumerillii*'s whole brain on a logaritmic scale. N for each graph is the number of "cubes" in the dataset where the fluorescence value is higher than 0. On the one hand, the log density shows two clear peaks for rOpsin, making the choice of a expression threshold easy. PRDM8.H92 on the other hand does not display such a clear cut threshold.

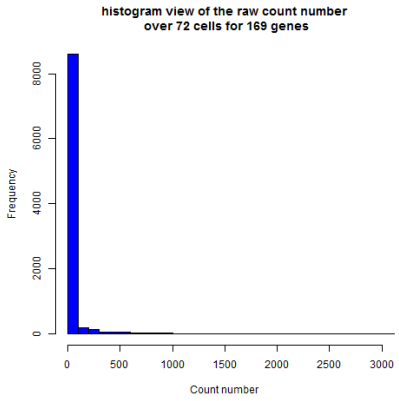
This method resulted in a high confidence binarized dataset for 86 genes. Several reasons explain why 83 genes out of the starting 169 were removed from the dataset. For some of the genes no good threshold could be found, this was due to high noise level in the in-situ hybridization images. Other images 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.

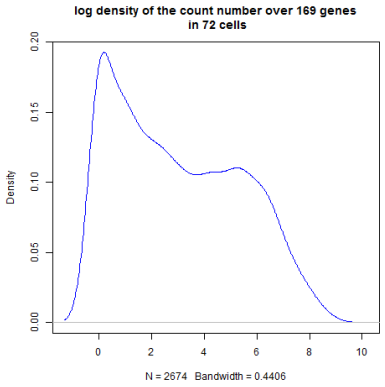
Binarizing whole transcriptomes

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 8a 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.



(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 8: 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 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 manner may lead to numerous errors. Indeed, the underlying assumption behind this method is that all genes behave in a similar way. As figure 8b shows, if a 2 peak behaviour is indeed present, the cut is not extremely clear and a good number of count numbers actually fall in between the two peaks. This is due to the fact that all expressed genes are not expressed 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 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

Single cell RNA-seq in Platynereis' brain

Collaborations with the Kaia Achim in the Detlev lab in EMBL as well as Luis R. Saraiva in the Marioni lab have provided us with a unique RNA-seq dataset of 72 single cells from *P. dumerillii*'s 48hpf developing brain. The method used was the Tang protocol [37] within a microfluidics pipeline .

get info on sequencing

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.

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 a the count number for each of the

Listing 1: A floating example

```
for i:=maxint to 0 do
begin
{ do nothing }
end;
```

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.

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.

Given the set of 169 considered genes M, and the set of 72 cells C.
- Present John's method - FIG 6: find a nice way to show a few good examples of mapping

3

HIDDEN MARKOV RANDOM FIELD BASED CLUSTERING FOR SINGLE CELL GENE EXPRESSION DATA

3.1 MARKOV RANDOM FIELD PRIOR DISTRIBUTION

Part II

APPENDIX

APPENDIX

BIBLIOGRAPHY

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

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

[43] 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.

[44] 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