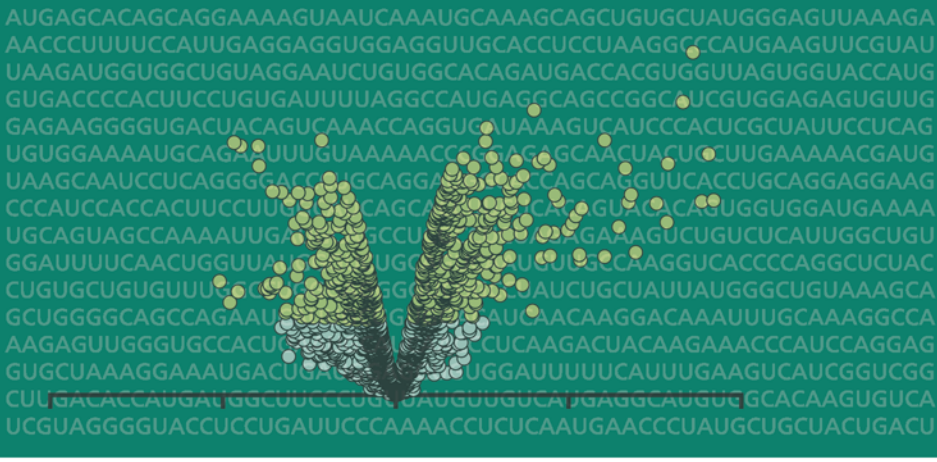
Alessandro Cellerino
Michele Sanguanini

Transcriptome Analysis

Introduction and Examples from the Neurosciences



EDIZIONI
DELLA
NORMALE

17

APPUNTI

LECTURE NOTES

Alessandro Cellerino
Scuola Normale Superiore
Piazza dei Cavalieri, 7
56126 Pisa

Michele Sanguanini
Gonville and Caius College
University of Cambridge
Trinity Street
CB2 1TA, Cambridge
Cambridgeshire, United Kingdom

Transcriptome Analysis
Introduction and Examples from the Neurosciences

Alessandro Cellerino
Michele Sanguanini

Transcriptome Analysis

Introduction and Examples from the Neurosciences

EDIZIONI
DELLA
NORMALE

isbn 978-88-7642-642-1 (eBook)
DOI 10.1007/978-88-7642-642-1

issn 2532-991X (print)
issn 2611-2248 (online)

Contents

Preface

In the broad panorama of academic publishing regarding next-generation sequencing, there is scarcity—if not a lack—of textbooks that tackle experimental issues, rationale of data analysis, and biological interpretation of genome-level data simultaneously.

The current textbook is not intended to be a cookbook of quantitative approaches in Biology, nor is it a rigorous collection of theorems. It aims to create a common language that can be useful to experimental biologists and data analysts: the former would be able to read papers critically based on transcriptomics and to judge independently whether conclusions are appropriate, to design their own transcriptomic experiments, and to create a dialogue with data-analysts to determine together the most appropriate approaches for the specific questions answered. The latter would have a glimpse of the experimental biologist's point of view, with numerous examples of how techniques that are familiar to an analyst have answered specific biological questions.

The target audience of this book are graduate students with a background in the Natural Sciences (for example, Cellular and Molecular Biology, Neurobiology, Evolutionary Biology, Applied Mathematics, Physics, or Chemistry) who are interested in acquiring the bases of next-generation RNA sequencing and transcriptome analysis and learning how these techniques can be used to derive new knowledge about the functional organisation of the nervous system. It is beyond the scope of this book to include a detailed review of RNA and neuron biology, or of the needed mathematical tools. So we will take for granted some basic knowledge in Molecular Biology, Neurobiology, Linear Algebra, and Calculus. Throughout the book, we made our best effort to complement the basics of quantitative analysis with relevant practical examples of how we used these tools to tackle specific questions in real laboratory life. This approach may seem too simplified for students trained in the quantitative sciences and still difficult to digest for 'wet-lab' biologists but represents—in our opinion—the best possible compromise.

The book is based mainly on notes for the Neurogenomics course that was held in 2015 at Scuola Normale Superiore in Pisa with some later integrations.

We would like to thank Dr. Francesco Neri and Prof. Michele Vendruscolo for meaningful comments on the manuscript, Dr. Marco Groth for providing useful RNA-seq data, and Mariateresa Mazzetto and Martino Ugolini for performing the data analysis that is shown in many figures. M.S. is also grateful to Prashanth Ciryam for his invaluable help in making this book a better read.

Pisa–Cambridge, February 2018

Alessandro Cellerino
Michele Sanguanini

Introduction: why studying transcriptomics?

Biological and physiological investigations classically involved the painstaking collection of measurements for single physical or chemical variables (pressure, volume, electrical potential, hormone concentrations, and so on). In more recent times, quantitative measurements of gene and protein expression became available and represented a small revolution in itself, since they allowed scientists to investigate the molecular landscape underlying biological or physiological processes for the first time. However, this approach, as widespread and successful as it has been, is centred on the selection of a handful of markers that is by its nature arbitrary.

The recent development of so-called 'omics' technologies has revolutionised biomedical research. These methods permit a quantitative assessment of the global molecular landscape associated with a biological phenomenon and provide an extremely powerful tool to identify the molecular upstream regulators as well as the downstream actuators of a given process. The whole potential of these techniques can be realised when these are coupled with the ever-growing toolbox of protocols available to experimentally modify gene expression in cells and even whole organisms, thereby offering the possibility to experimentally validate hypotheses derived from the global analysis. The great power of these experimental approaches lies in their *unbiased* nature. The experimenter assigns individual samples to different experimental groups and does not provide any further *a priori* hypothesis on the molecular mechanisms to be investigated. Therefore, the analysis may reveal novel and unexpected players. In recent years, the analysis of the *transcriptome* has become particularly widespread. Transcriptomics is the collective name for a host of techniques that allow a genome-wide estimate of transcript abundance (*i.e.* mRNAs) in a sample and is currently almost exclusively analysed through sequencing-based techniques (RNA-seq).

Genome-wide techniques are now indispensable tools for biomedical research; this is the main motivation for writing this book. However, analysis of genome-scale datasets has almost become a discipline

in and of itself, thus the processing, handling, and—most importantly—interpretation of these datasets is now well beyond the basic statistical knowledge of 'wet-lab' biologists. More importantly, there is a widespread misconception that this kind of analytics is a tedious, but straightforward, process. And it is not uncommon to see scientists providing biological samples to a facility specialised in genome-scale techniques with the expectation of receiving publication-quality results. The wrong assumption is that there is only one way of analysing these data, while, on the contrary, analysis of genome-scale data requires a long series of 'arbitrary' decisions that are dependent on the specific questions of interest. We chose to use the nervous system as the source of examples where analytical approaches were applied to gain biological insight.

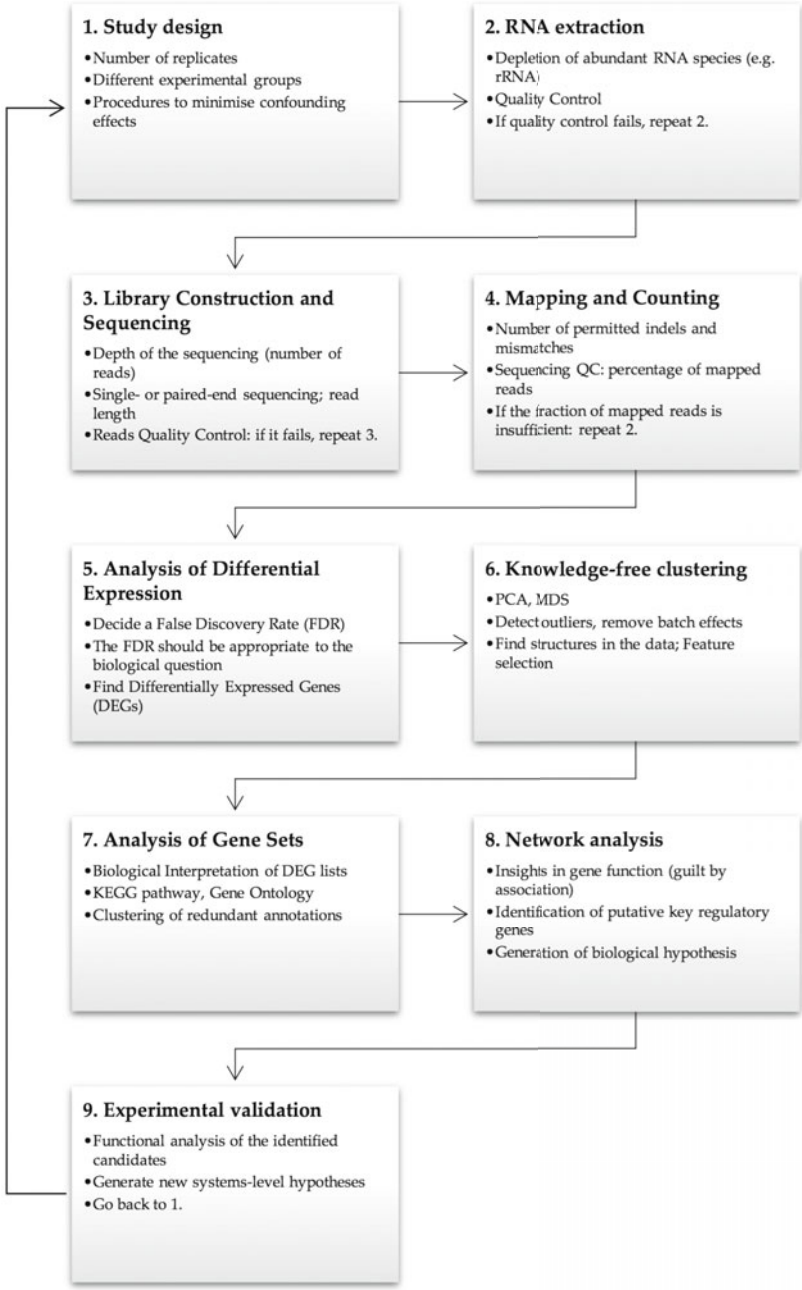
In this book, we will thus deal with three major levels and concepts.

A molecular level. The ribonucleic acid (RNA) is the carrier of the information flow that goes from the DNA to the phenotypes of a living system. This information can be carried under the form of *structural scaffolds* (an example is the ribosomal RNA), of *protein-coding sequences*, which are read by the ribosomes and translated in the amino acid alphabet that composes the thousands of proteins in a cell, or of *regulatory RNAs* that can modulate gene expression at various levels. The complete set of RNAs (transcripts) of a cell type is called the *transcriptome* and is the primary object of our analysis.

A quantitative level. Transcriptomics is a quantitative, data-oriented science: its raw input is a list of strings (the 'reads'), that must be assigned to an object (the 'genes' or 'transcripts' in the genome). Then the density of reads corresponding to each object is counted to calculate its *expression strength*; in other words, these data come in the form of a *vector* or a *matrix*. The numerosity of the datasets in neurogenomics is usually of thousands to tens of thousands of genes whose expression level is quantified in multiple samples and conditions. The task is to detect significant relationships between biological conditions and patterns of gene expression in the dataset. This high dimensionality can be handled only through specific methods of statistics and data science. Chapters 3 to 7 will be dedicated to presenting and understanding the basic tools which are commonly used to analyse transcriptomics data.

An organ and cellular level. The nervous system is the organ that integrates the stimuli from the external environment and the internal state of an organism to generate a consequential and contextualised response. *Neurons* are the fundamental cells and main computational units of the nervous system and encode the reaction to a stimulus

through a temporary change of their *excitation state*. This can be then transmitted through a *synapse* to other neurons to eventually reach the effector cells. Neurons are a unique class of cells due to their extreme variability in terms of functional, morphological and molecular phenotypes. The retina alone contains almost one hundred of different neuronal subclasses. The other cell types which compose the nervous system, such as *astrocytes* or *microglia*, however, are not bystanders of neural activity but have been associated with essential functions in the nervous system computation and plasticity. Due to their complexity and manifold functional outputs, the brain and the nervous system represent testing grounds where the full potential of genome-scale techniques can be employed. Chapters 8 and 9 will deal with the application of Neurogenomics to the study of the nervous system and neural function.



An example of pipeline for iterative transcriptome analysis

Chapter 1
A primer on data distributions and their visualisation

This chapter is intended for 'wet-lab' biologists to familiarise them with some fundamental concepts (*e.g.*, statistical distribution) and formats of data visualisation that will be used in many points of this book and that are necessary to critically read the scientific literature of interest.

1.1. Stochastic processes and Poisson distribution

When dealing with RNA-seq data, we will encounter many times the concept of the **stochastic process**, that we can define here as an ensemble of events that occur randomly in space and time (*random variables*). In 1898, the Russian statistician Ladislaus Bortkiewicz published the book *The Law of Small Numbers* where he analysed the now classical dataset of casualties in the Prussian cavalry corps from horse kicking measured over 20 years among 14 corps (Figure 1.1A).

Let's consider the frequency distribution of all the casualties in the Prussian cavalry corps (Figure 1.1B). It can be shown that, given a stochastic event whose probability of occurring is constant in the studied interval of times, the probability that it will occur n times in a fixed time interval τ is distributed according to a **Poissonian**

$$P(x_\tau = n) = \frac{\lambda_\tau^n}{n!} e^{-\lambda_\tau} \tag{1.1}$$

where λ_τ is the expected frequency of occurrence of the event in the time interval τ. It is thus possible to fit the casualties distribution with a Poissonian with $\lambda_{20yr} = 0.7$.

When we measure the concentration of a given mRNA in a sample by sequencing-based techniques, the process that we undertake is, as a matter of fact, analogue to sending a set of mRNA molecules one at a time through an array of 'mRNA detectors', each counting every time it is 'hit' by the specific mRNAs it is 'tuned' to. This is a stochastic process just like the killing of a Prussian horseman from the kick of a horse!

A

Year	GC	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C14	C15
1875	0	0	0	0	0	0	0	1	1	0	0	0	1	0
1876	2	0	0	0	1	0	0	0	0	0	0	0	1	1
1877	2	0	0	0	0	0	1	1	0	0	1	0	2	0
1878	1	2	2	1	1	0	0	0	0	0	1	0	1	0
1879	0	0	0	1	1	2	2	0	1	0	0	2	1	0
1880	0	3	2	1	1	1	0	0	0	2	1	4	3	0
1881	1	0	0	2	1	0	0	1	0	1	0	0	0	0
1882	1	2	0	0	0	0	1	0	1	1	2	1	4	1
1883	0	0	1	2	0	1	2	1	0	1	0	3	0	0
1884	3	0	1	0	0	0	0	1	0	0	2	0	1	1
1885	0	0	0	0	0	0	1	0	0	2	0	1	0	1
1886	2	1	0	0	1	1	1	0	0	1	0	1	3	0
1887	1	1	2	1	0	0	3	2	1	1	0	1	2	0
1888	0	1	1	0	0	1	1	0	0	0	0	1	1	0
1889	0	0	1	1	0	1	1	0	0	1	2	2	0	2
1890	1	2	0	2	0	1	1	2	0	2	1	1	2	2
1891	0	0	0	1	1	1	0	1	1	0	3	3	1	0
1892	1	3	2	0	1	1	3	0	1	1	0	1	1	0
1893	0	1	0	0	0	1	0	2	0	0	1	3	0	0
1894	1	0	0	0	0	0	0	0	1	0	1	1	0	0

B

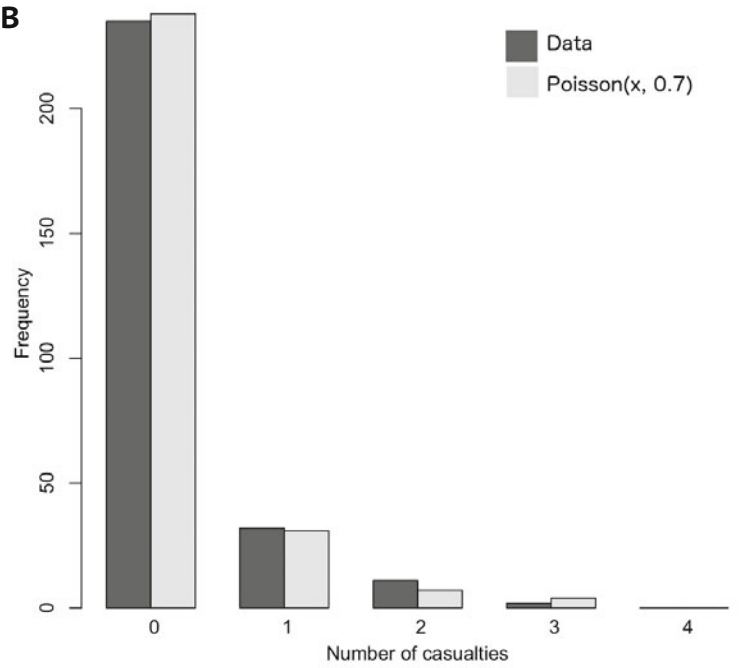


Figure 1.1. Data, distribution and Poisson fit ($\lambda = 0.7$) of the casualties in the Prussian cavalry corps from horse kicking measured over 20 years. Data from [3].

For each gene, the number of 'hits' is a measure of its concentration and from Equation (1.1) we can calculate the error of this measurement (see discussion of confidence intervals below).

1.1.1. Gaussian and t-Student distributions

Another fundamental distribution is the *normal* (or **Gaussian**) distribution, which is associated with a continuous probability density function. If x is a stochastic variable that is normally distributed, if one observes x once, the probability that its value falls within an interval $\mathcal{I}$ given the mean μ and the variance σ^2 of the distribution is

$$P(x \in \mathcal{I}|\mu, \sigma) = \int_{\mathcal{I}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x. \tag{1.2}$$

The distribution of many (approximately) continuous biological variables can be modelled using a normal density function: just think about the distribution of the heights in a group of people[1]. For neurogenomic applications, it is important to remember that the distribution of expression values obtained from *microarray* techniques is (after appropriate normalisation procedures) distributed like a Gaussian function. In a real experimental context, what we want to do is to make an estimation of the 'real' parameters of the population (μ and σ, see Figure 1.2A) from a finite, possibly small, amount of measurements. The appearance of the variables from a normal distribution strongly depends on the sample size (the *degrees of freedom*): if we sample a normally distributed population with unknown standard deviation, the statistical distribution we want to use to describe the subset of the population is the **t-Student distribution**. The t-Student distribution is commonly used to test the statistical significance of the deviation of an observation from the control distribution (see, for example, at page 96). As shown in Figure 1.2B, with the increase of the size of the sampled population, the t-Student distribution approximates the normal distribution.

1.1.2. Parameters of a distribution

Figure 1.2A tells us some other information about the distribution of a normal population: more than half of the sampled population (68% in case of a Gaussian with $\sigma = 1$) stays within one standard deviation from

[1] The reason why the Gaussian distribution is so widespread represented can be conduced to the **central limit theorem**, that states that the average of a series of random variables (*e.g.*, a series of Poissonians) is distributed, at its limit, according to a normal distribution.

the mean, 95% of the sampled population stays within two standard deviations, and almost the totality of the data (99.7%) are within three standard deviations. However, when we measure a variable (with biological replicates) of a given population we can have two possible scenarios:

- the measurements are **precise**—that is, each measurement gives an estimation of the population mean that is close to the others,
- the measurements are **accurate**—that is, the estimated population mean is close to the real mean.

To make a classic example: if we throw four darts at a dartboard aiming at the centre and we get all the darts in roughly the same position we made a *precise throw*; a set of throws that, once averaged, hits the centre of the board is *accurate*. In the same way, a measurement can be both precise and accurate, only precise/accurate, or neither of those. The **Confidence Interval** (CI) of a stochastic variable is the range of values wherein the

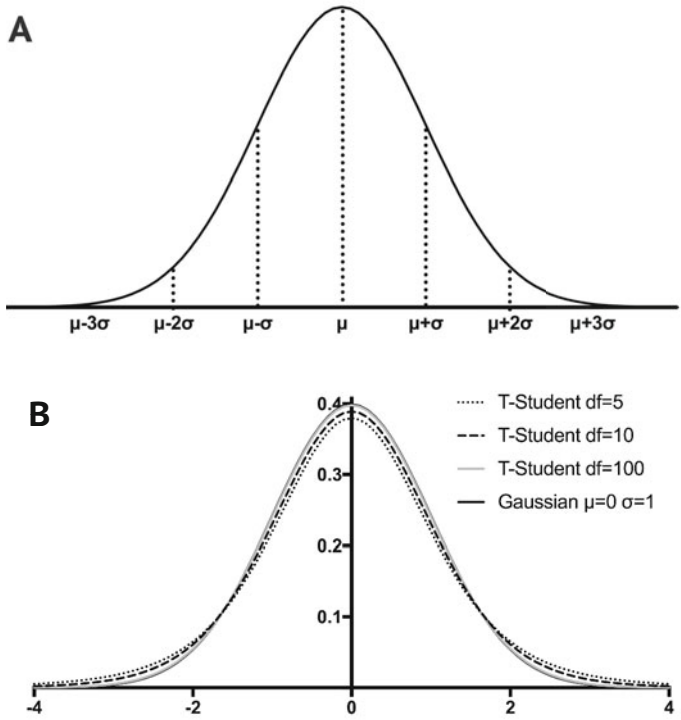


Figure 1.2. A) Gaussian distribution with unit standard deviation and its relationship with the data of the population. B) Comparison between a Gaussian distribution with zero mean and unit standard deviation ($y(x) = e^{x^2/2}$, solid black line), and associated left-tail Student t-distribution with three different degrees of freedom.

real value of the population would fall with a given probability. For example, if we have a population with estimated mean $\bar{x} = 2.3$ and a 95% confidence interval ($CI_{0.95}$) between 1.6 and 2.8, this means that there is a 95% confidence that we will find the real population mean μ_x in the $CI_{0.95}$ interval[2].

A common way to compute a CI from a population which is assumed to be distributed according to a normal distribution, but whose mean and the standard deviation are not known, involves the use of the t-distribution. Let's set our confidence level $1 - \alpha = 0.95$, and consider a set of n measurements (observed mean $\bar{x}$ and standard deviation s_x) from a population with real (unknown) mean μ and standard deviation σ. We want to find the interval of x for which we have a confidence of 95% that the real mean falls into this interval. This can be done through setting up a test of significance for the difference between the measured and the real mean, which is called **t-test**:

$$-t_{n-1} \leq \frac{\mu - \bar{x}}{s/\sqrt{n}} \leq t_{n-1} \tag{1.3}$$

where t_{n-1} is called the **critical value** from a one- or two-tailed t-distribution[3] with $n - 1$ degrees of freedom and $s/\sqrt{n}$ is called the **standard error of the mean**. Arranging Equation (1.3) in order to evidence the real mean we obtain

$$\bar{x} - t_{n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1}\frac{s}{\sqrt{n}} \tag{1.4}$$

and it is evident that, the higher the number of sampled measures, the shorter the CI (if the standard error is reasonably narrow), and the t value can be approximated with the cumulative normal distribution function $\Phi^{-1}(1 - \alpha)$.

In many cases in genomic analysis, more sophisticated methods are used to compute the CI of variables that cannot be assumed to be normally-distributed or whose distribution is not known. The concept of CI remains the same, however, and—regardless of the method used—it indicates the span of values within which the true value is expected to fall with a probability of 95%.

[2] However, in 5% of the cases, the chosen CI will not encompass the real value of the variable.

[3] For more information on the different cases of underlying distributions (one-tailed, two-tailed, paired, and so on) please refer to a textbook of biostatistics.

1.2. Representation of quantitative biological data

Now that we have introduced some basic notions of statistics, it is possible to thoughtfully describe a representation of data distributions that is widely used in genomics: the **box plot** (Figure 1.3A). The basic plot comprises three different components:

- the **whiskers** usually define the $5\% - 95\%$ percentile interval of the data (*i.e.* the CI), although in some cases they can represent the whole span of the distribution (*i.e.* minimum/maximum of the dataset);

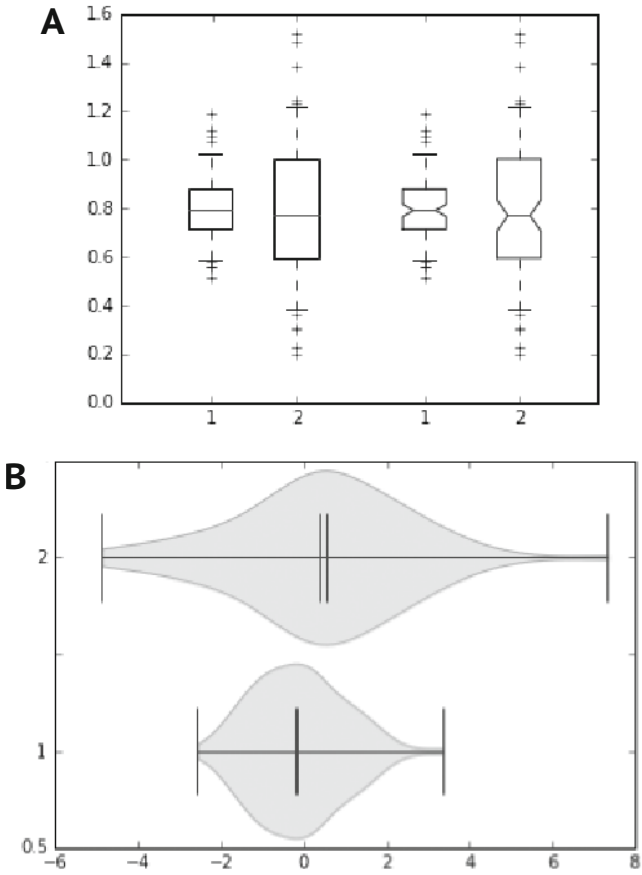


Figure 1.3. Commonly used graphic representation of data: A) classic (*left*) and notched (*right*) box plot with the whiskers representing the 5-95% interval and the outliers indicated with a '+', B) violin plot. In both the plots, the set 1 is composed of n=100 points randomly picked from a normal distribution with μ=0 and $\sigma = 1$, while the set 2 is picked from a normal distribution with μ=0 and $\sigma = 2$.

- the **box** defines the interval of the data between 25% (*first quartile*) and 75% (*third quartile*), also called the *interquartile range* (IQR), *i.e.* the span that the central 50% of the data falls within;
- the **bar** indicates the median (*second quartile*), *i.e.* the value that divides the distribution of the data in half, so that the probability for a point to fall on either side of it is 50%.

Additional features can be added to indicate data points outside of the $5\% - 95\%$ percentile interval (outliers), the mean of the set and so on. A variant of the box plot called **notched box plot** presents a narrowing of the box around the median ('notch') which represents the 95% CI of the median and thus it is a rough visual indication of probability that the difference between the medians of the two samples is not due to chance (statistical significance).

1.2.1. Violin plot

The **violin plot** is another very informative way to represent biological data (Figure 1.3B). It combines the information of a box plot (Figure 1.3A) with an estimate of the probability distribution function of the variable (see Figure 1.2). The whiskers of a violin plot can span the CI of from the minimum to the maximum value of the set, and usually the mean and/or the median of the set are shown. The density function of the dataset is shown perpendicularly to the whiskers. The major advantage of the violin plot—compared to the box plot—is the disclosure of the distribution density. This could reveal the presence of a *multimodal distribution* of the data—*i.e.* the frequency density of the set shows more than one peak, that would be hidden in a box plot representation.

1.2.2. Scatter plot

The **scatter plot** is a way to make *direct comparisons* between datasets which are composed of matrices $M \times 2$, for example where a given set of M features—*e.g.*, expression strengths, fold-changes and so on—is measured in two different conditions (young/old, normal/diseased, control/treated etc...). In a scatter plot, the value of the feature in each condition is a coordinate of a Cartesian plot. In Figure 1.4 a dataset with a clear monotonic relationship is shown with a scatter plot. Since gene expression data often range several orders of magnitude, it is often necessary to change the scale of the axes from linear to logarithmic. In the example visualised, this transformation makes it possible to detect a clear linear relationship between the logarithm of the values in X and the logarithm of the values in Y (a so-called power-law that is typical of many

biological systems, for example the relationship between body mass and metabolism).

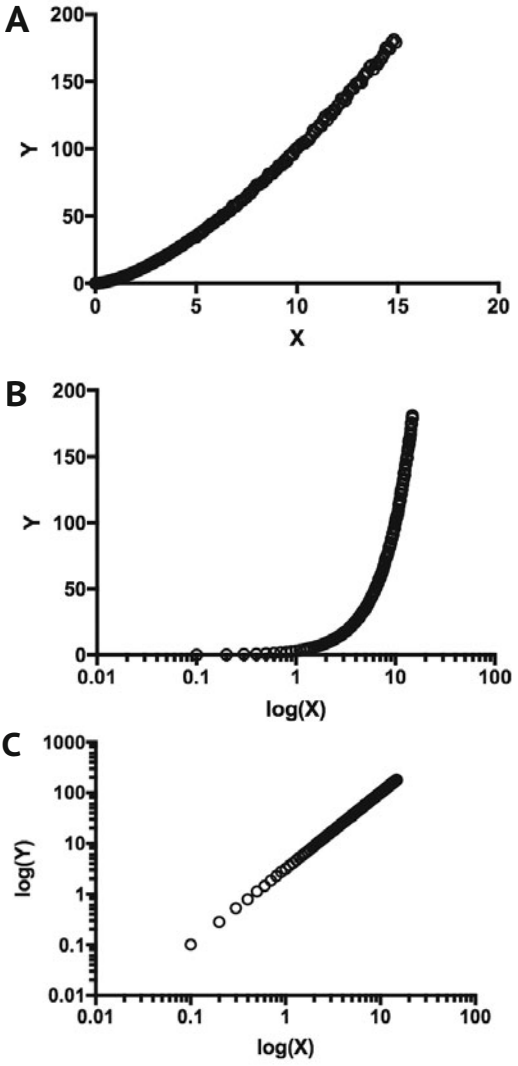


Figure 1.4. Scatter plot of a simulated dataset with different axis scaling: A) both X and Y linear, B) X logarithmic and Y linear, C) both X and Y logarithmic.

1.3. Lists of genes and Venn diagrams

We have so far considered a specific class of data, which can be described with frequency distributions and/or can be plotted in a Cartesian space.

The data we have previously seen could be, *e.g.*, a vector of abundance values for the transcripts associated with a given gene, or the average fold change associated with a given metabolic pathway in condition A against condition B. Let's now focus on a different type of data that could arise from a neurogenomic study: a **set of labels** (gene names, biological functions, and so on). If we consider the two lists of the genes that are differentially regulated[4] in the ageing human brain and in Alzheimer's disease (data derived from [9]), respectively. Given a single differentially expressed gene, there are three possible scenarios:

- the gene is up-/down-regulated in one of the two conditions only;
- the gene is up-/down-regulated in both conditions;
- the gene is up-regulated in one condition and down-regulated in the other.

An easy way to manipulate lists of differentially regulated genes is to consider each list as a **mathematical set**, where each gene is an element of the set. In this case the studied dataset can be written as follows:

$$\begin{aligned} A &= \{\text{gene} \mid \text{gene is up-regulated in Alzheimer's disease}\} \\ B &= \{\text{gene} \mid \text{gene is down-regulated in Alzheimer's disease}\} \\ C &= \{\text{gene} \mid \text{gene is up-regulated in ageing}\} \\ D &= \{\text{gene} \mid \text{gene is down-regulated in ageing}\} \end{aligned} \tag{1.5}$$

In the previous annotation, the list of genes that are up-regulated during both ageing and Alzheimer's disease can be represented as the *intersection* of the sets A and C ($A \cap C$)[5]. The easiest way to represent a group of sets and the operations on them (for example the common elements, and so on) graphically is the **Venn diagram**[6], where each element of the set is represented as a point in a closed region, and the common elements of two sets are included in the overlapping region between the graphic representation of the set (Figure 1.5). From the Venn representation of the sets of Equation (1.5) it is almost immediately noticeable that in the lists there are no genes that are up-regulated in Alzheimer's disease and down-regulated in ageing, or viceversa.

[4] See Chapter 4.

[5] Some trivial properties of the sets described above are that the intersection of sets of the genes up-regulated and down-regulated in the same condition (say, in ageing) has no elements, that is $A \cap B = C \cap D = \emptyset$, and thus that the genes which are differentially regulated only in Alzheimer's disease (but not during ageing) is $(A \cup B) \setminus (C \cup D)$.

[6] The curious reader can find an elegant representation of Venn diagrams on a stained glass window in the Dining Hall of Gonville and Caius College, Cambridge.

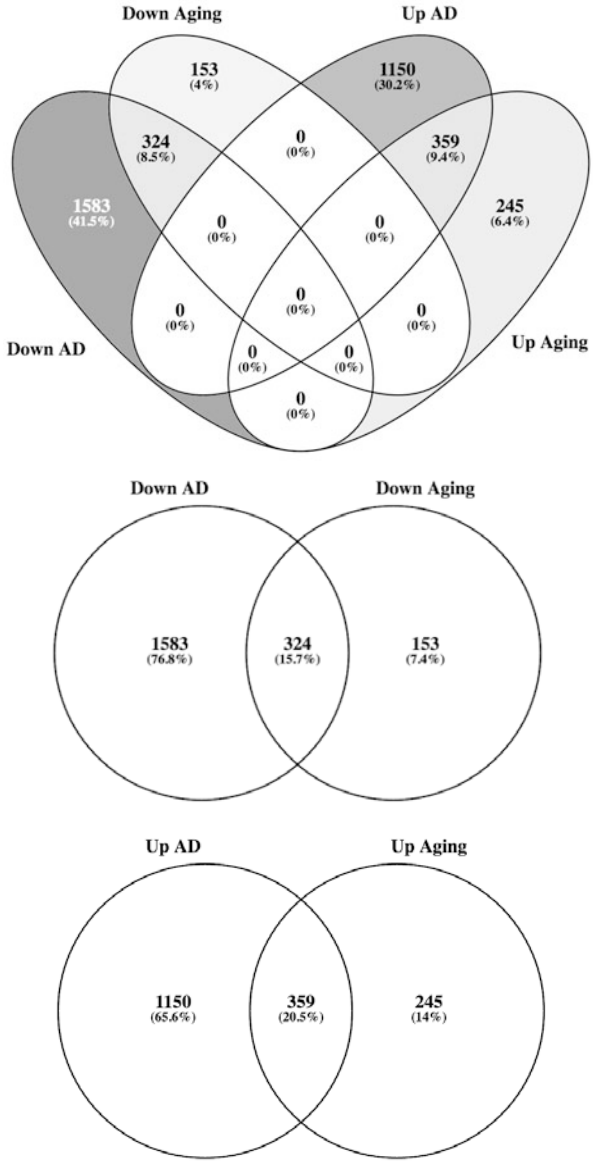


Figure 1.5. Venn diagram of up- and down-regulated genes in Alzheimer's disease (AD) and in ageing. Due to the absence of an intersection between the genes up-regulated in AD and down-regulated in ageing—or viceversa, the 4-way representation (*left*) of sets A, B, C and D (see Equation (1.5)) can be simplified in two 2-way representation of sets A-C and B-D (*right*). Venn diagram plotted using the Venny 2.1 tool (http://bioinfogp.cnb.csic.es/tools/venny/) from the data in [9].

Chapter 2
Next-generation RNA sequencing

2.1. Introduction

The Next-generation RNA sequencing (**RNA-seq**) is an high-throughput experimental technique (and design) that allows sequencing of cDNA at very high redundancy (depth), in the order of 10^7 individual sequences (reads) per sample. This technique can provide a quantification-by-sequencing of RNA abundance and analysis of RNA sequence variation from biological samples. The development of high-throughput technologies applied to biology has opened exciting avenues for computational and **data-driven biology**[1].

2.2. Advantages of RNA-seq

Compared to other high-throughput techniques, such as *cDNA microarrays*, an RNA-seq protocol offers the following **advantages**:

- the analysis of RNA transcripts can be both **qualitative** (allowing the assembly of transcriptomes from non-model species and the discovery of novel, previously not annotated transcripts), and **quantitative** (providing a measure of the relative quantities for different transcripts in a sample);
- the results show **single-base resolution**, which makes it possible to study single-base characteristics such as *RNA editing* or *Single Nucleotide Polymorphisms* (SNPs);

[1] Data-driven science is a paradigm shift where experimental activities are informed and motivated by unbiased analysis of large bodies of pre-existent global data (-omics). This approach is complementary to the hypothesis-driven approach that is typical of cell and molecular biology, where the experiments are tailored for testing a specific hypothesis derived from previous knowledge. The combination of these two approaches is expected to give a huge momentum to life sciences and neurosciences in particular.

- the *lack of pre-synthesised molecular probes* offers **flexibility in experimental design** and a potentially unbiased source of transcriptional information;
- it is not bound to existing species-specific 'hardware' (*e.g.*, microarrays) and it allows genome-wide quantification of transcript abundances in non-model species;
- there are virtually **no limits on sensitivity**[2].

Three steps are necessary in order to sequence the transcriptome of a biological sample: first the **RNA** has to be **purified**, second **an RNA (or cDNA) library needs to be synthesised**, and eventually the library is **sequenced**.

2.3. RNA purification

The first issue to be addressed while dealing with RNA purification is the **target RNA species of interest**. There are three major categories of RNA: protein-coding mRNAs ($\gg$ 100 bp), long non-protein coding RNA ($>$100 bp, such as rRNA, snRNA, lncRNA), and short non-coding RNAs ($\ll$ 100bp – tRNAs, miRNAs and related, snoRNAs, piRNAs, and so on). There are some technical differences in the purification procedures for the short RNAs as compared to long RNAs. Standard RNA purification by column-based commercial kits loses the majority of the small species (unless otherwise stated by the producer), so a guanidinium thiocyanate-phenol-chloroform extraction[3] coupled with specific silica membrane purification is usually adopted in order to extract small RNAs. These technical differences imply that *the small RNA-oriented RNA purification leads to a RNA sample which can be used to prepare both small RNA and total RNA libraries, while a standard extraction can be used when the small RNAs are not important for the study*. At the moment, a standard RNA-seq protocol requires a starting amount of 0.5-1μg of total RNA, but also lower amounts of RNA ($\sim 0.1\mu$g) can be processed in routine RNA-seq procedures. However, procedures of RNA amplification are available, that allow RNA-seq to be performed from very small amount of material and even single cells (see Chapter 9).

[2] In the sense that, by increasing coverage, it would be possible to detect even genes that are expressed as single RNA molecule per cell. However, it remains to be established where the the border lies between a biological meaningful signal and transcriptional noise.

[3] Commercially known as TRIzol® or QIAzol®.

2.3.1. RNA quality assessment

Once the RNA is extracted, it is mandatory to assess the purity and integrity of the sample. If chemical or biological contaminants are present, or the RNA is severely degraded, this could lead to poor quality or artefact-ridden RNA libraries and therefore to sequencing that is not representative of the true RNA populations of the sample.

The principal contaminant species are either related to the technical processing of the RNA sample (phenol, EDTA and so on) or to biological contaminants, such as proteins or genomic DNA. *A careful purification practice usually avoids these contaminants*: the assessment for contaminants is done through **UV spectrometry**. The peak of **absorbance** for **RNA** molecules is at **260 nm**, while the peak for the contaminants is usually distinct (see Table 2.1).

In order to evaluate the degree of purity of the RNA sample two *absorbance ratios* are used:

$\frac{\mathbf{A_{260}}}{\mathbf{A_{280}}}$ in order to quantify the protein contamination: a pure RNA solution has a ratio of ∼2.0, and a ratio of 1.8 is considered acceptable for many RNA-seq protocols. However, these values should be taken as a 'rule of thumb', due to the fact that the UV absorbance of RNA at 280 nm depends on many factors such as the pH, the elution buffer, and the base composition of the RNA analysed, since the different nucleotides have different values of absorbance.

$\frac{\mathbf{A_{260}}}{\mathbf{A_{230}}}$ quantifies the extraction solvent contaminants: a non-contaminated RNA solution has no peak at 230, and shows a 230:260:280 proportion which is about 1:2:1.

Contaminant	**λ peak**
Proteins	280 nm
Phenol	230 and 270 nm
Thiocyanate	230 nm
EDTA	230 nm

Table 2.1. Wavelength of the peak of absorbance for potential contaminants in RNA samples.

RNases are ubiquitous enzymes whose function is to degrade RNA: if the operator does not pay attention during the RNA sample handling, or the samples were stored inappropriately, some ambient RNases or RNases from the tissue itself that were not inactivated promptly could

contaminate the sample, thus degrading the molecules of interest[4]. The most commonly used method to qualitatively determine the level of **RNA degradation** is direct visualisation through *gel electrophoresis*. In intact RNA, the two bands of rRNA are sharp and the 28S is roughly twice as intense as the 18S. Degradation results in smears and a reduced 28S/18S ratio. In order to obtain a quantitative assessment of RNA integrity, the Agilent BioAnalyzer can be used. It performs an electrophoretic run in a microfluidic chip, thus profiling the RNA sample according to the molecular size and the fluorescence signal of an intercalating dye. The degradation is quantified in a score called the *RNA Integrity Number* (**RIN**) whose major proportion is determined by the shape of the peaks of the **18S** and **28S rRNAs**, which normally account for about 80% of the total RNA fraction. The RIN spans from a value of 1 (extremely degraded sample) to 10. For microarray applications, RIN is an extremely important parameter because it influences the intensity of the signal and binding to the array. For RNA sequencing, there is not a consensus on which is the minimum score required for further processing, especially since the RNAs need to be fragmented in order to create the sequencing library. A rule of thumb states that, for both microarray and RNA-seq, the RIN should be higher than 7; however in the case of RNA-seq a RIN of 4-5 could be acceptable upon protocol optimisation and without poly-A enrichment steps[5].

2.3.2. Abundant RNA species

A standard RNA sample is composed of

$\sim$ 83% rRNA
$\sim$ 15 % tRNA, and
$\sim$ 3% mRNA.

So, the fraction of mRNAs, which are usually the species of interest in RNA-seq experiments, is also the least represented in the total RNA pool. This fact implies that the RNA samples should not be sequenced *as they are*, but should be first treated to remove the most abundant species.

[4] To avoid the RNase contamination it is very important to keep the workbench clean (and possibly pretreated with RNase-inhibiting solutions) and to use nuclease-free **Diethylpyrocarbonate** (DEPC) **treated water**

[5] Of course these estimations should be taken *cum grano salis*, downstream quality controls (see next chapter) need to be performed in order to ensure that the sequencing was representative.

There are two opposing strategies to address this problem:

- to use **poly-dT primers** linked to magnetic beads in order to specifically enrich the mRNAs through binding their poly-A tails;
- to use specific **probes binding known abundant RNA** species to deplete them and therefore enrich the less abundant RNA species.

Both strategies have strengths and weaknesses. Isolating the mRNAs implies that all long non-coding RNAs devoid of a poly-A tail are excluded-not only the rRNAs and the tRNAs, but also rarer non-coding RNAs that may be important for regulation of gene expression. On the other hand, the exclusion of the abundant RNAs is possible only if their exact sequences are known. This is particularly problematic when working with non-model species. Furthermore, depletion of abundant species can never be complete and variable amounts will remain (for example, rRNA may be reduced from 80% to 30% but will still represent a sizeable proportion of reads[6]). Table 2.2 shows a comparison between the ups and the downs of the two strategies.

Characteristic	**poly-T prim.**	**RNA prob.**
All mRNAs are retained	yes	yes
Abundant rRNAs are removed	yes	yes
Other long noncoding RNAs are retained	no*	yes
DNA contamination are removed	yes	no
Applicable to every animal model	yes	not always

Table 2.2. Comparison between strategies for abundant RNA species depletion. *: there is a class of lncRNAs, called *lincRNAs* [56], that are poly-adenylated and are thus retained after poly-T primer purification.

2.3.3. Tissue-specific abundant RNA species

Ribosomal RNAs and tRNAs are not the only abundant RNA species which could determine an undersampling of the different transcripts in the transcriptome. In some tissues, there are 'structural' transcripts which could account for a predominant fraction of the sequenced dataset and thus must be depleted in order to sample the whole complexity of the transcriptome.

[6] In fairness, with the current rRNA depletion protocols it is possible to achieve a reduction of ribosomal RNA up to 99% in a consistent way.

2.3.3.1. Example 1: the globins mRNAs in blood RNA-seq. Blood is a liquid tissue composed of erythrocytes (92%), leucocytes (0.15%), and thrombocytes (7.75%), all contributing to the isolated RNA pool. Erythrocytes contain very high concentrations of haemoglobin, a heterotetrameric protein composed by two dimers of α- and β-globins, and it is reasonable to suppose that a great fraction of the mRNA present in the cells encodes for the α- and β chains of haemoglobin.

As an example, a sequencing of a polyA+ RNA library (a highly-enriched mRNA pool) from a mouse blood sample, resulted in $\sim$ 20% of uniquely mappable reads (reads mapping on the genome only once and with a small number of mismatches), a very minor fraction of non mappable reads, and about an 80% of *redundant mappable reads*—that is, reads that map to more than one position on the genome, as it would be the case for transcripts originating from the globin cluster. If we assign a gene to the aforementioned redundant reads, more than the 95% of them would originate from the α- and β-globin transcripts. This first sequencing proves that *it is necessary to deplete the globins transcripts from the RNA pool before sequencing*.

Reads	**% w/o depletion**	**% with depletion**
Uniquely mappable	$\sim$ 20%	$\sim$ 64%
Non mappable	$<$ 1%	$<$ 1%
Redundant mappable...	$\sim$ 80%	$\sim$ 36%
... of which mapped to globins	$>$ 95%	$\sim$ 15%

Data courtesy of Marco Groth, Leibniz Institute on Aging, Jena.

The depletion occurs with the same probe-bead system seen for the removal of abundant noncoding RNAs.

2.3.3.2. Example 2: actin and myosin mRNAs in muscle RNA-seq. In a second example (data courtesy of Marco Groth, Leibniz Institute on Aging, Jena), RNAs were extracted from the muscle of zebrafish. Actin and myosin are fundamental structural proteins for the cells in the muscle fibre, and for the cell in general, and their transcripts are expected to be highly represented in the muscle transcriptome. However, sequencing of the RNA library shows that these are abundant but not to the point that they dominate the set of reads. Redundant mappable reads (*i.e.* reads deriving from genes that are structurally redundant) and non-mappable reads are $<$20% of the total number of reads. Actin and myosin account only for $\sim$ 9.5% of the unique mappable reads (the 'working' dataset).

So, in this case, there is no need to deplete the actin and myosin transcripts, because their amount does not compromise the sampling of the transcript diversity in the library (issues of effects of abundant transcripts on library normalisation will be dealt with in Chapter 4).

2.4. Library preparation

Now we will deal with a fundamental step before performing the next-generation sequencing: the preparation of the RNA library. *From now on the described platform will be the Illumina (Solexa) sequencing*.

The rationale of the library preparation is to produce some *short cDNA fragments*, which are adapted and linked to the sequencer cartridge and then expanded through a peculiar type of PCR called *bridge amplification* to form clusters, each originating from a single DNA molecule (clusters can be thought as equivalent of DNA *clones*).

2.4.1. RNA fragmentation

The first step of the library preparation for Illumina sequencing is the **RNA fragmentation**: the length of the fragment is technically required to be smaller than 800 bp (otherwise the clustering by bridge amplification would be negatively affected—see below), however, the **optimal fragment size** depends on the **sequencing strategy** and the chosen **read length** (see Section 2.5.3). A usual read length is between 50 and 300 bp[7], and the sequencing could occur with two modalities (Figure 2.1):

Single-end sequencing where the sequencing occurs only at one end of the molecule: the minimum length of the RNA fragment should be the *read length*, so that the sequencing cycles are not wasted;

Paired-end sequencing where the sequencing occurs at both the ends of the target and the sequence information from both ends is connected: the minimum length of the fragment should be *twice the read length* in order not to have an overlapping sequencing[8].

Fragmentation can occur using many different methods, in particular **enzymatic** (using RNA endonucleases[9]), **chemical** (sample at 94 °C in

[7] It is important to notice that the longer the read, the lower the average quality of the last sequenced positions in the read (see page 28);

[8] The rationale for paired-end sequencing is to sequence the boundaries of a fragment so that it would be easier to reconstruct the source transcript. This is particularly useful in transcriptome assembly and analysis of alternative splicing. Having an overlapping sequencing would consume resources without getting additional information.

[9] The enzymatic fragmentation, however, suffers from a *composition bias* in hydrolysis, as some regions of the RNAs are more susceptible than others.

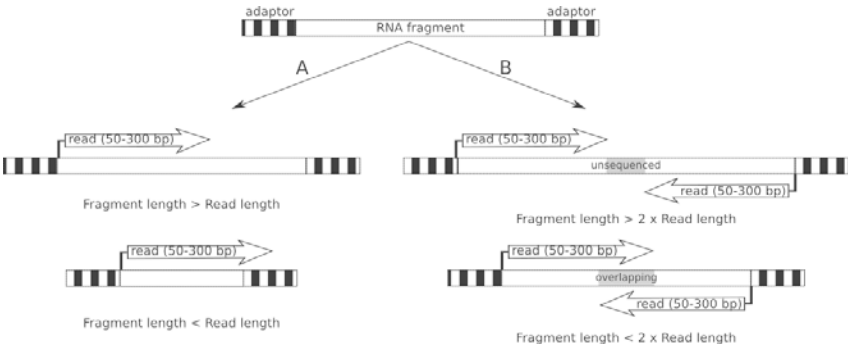


Figure 2.1. Illumina sequencing could occur with two set-ups: the single-end sequencing (A), where the sequencing occurs only from one side of the fragment, and the paired-end sequencing (B), where the sequencing occurs at both sides of the fragment and the two sequences are coupled in the downstream analysis. See text for further detail.

Tris-Magnesium salt buffer), or **mechanical** (sonication). The statistics of the fragment population—in particular the mean, median and size distribution—*evolve as a function of time*. For example, bringing the buffered RNA solution to 94 °C generates a population of fragments ranging from about 130 to 350 bp (with a median of 200 bp); after incubating the solution at 94 °C for 12 minutes, the population ranges from about 130 to 180 bp with a median of 140 bp. The median of the population scales down roughly linearly with time.

The quality of the fragmentation should be then directly assessed with gel electophoresis—such as using the previously described Agilent BioAnalyzer platform.

2.4.2. Reverse transcription

After the fragmentation step, the RNA templates should be converted to cDNA through the process commonly known as **reverse transcription**[10] (RT). At this point, several options for RT strategies are available:

- using **oligo-dT** to prime the poly-A tail of mRNAs;
- using **random primers**, an assortment of all the possible hexamers[11]. The probability of a random primer to bind to an RNA molecule is not

[10] The reverse transcription occurs using an *RNA-dependent DNA-polymerase*, also known as reverse transcriptase, primed with a DNA oligonucleotide.

[11] At each one of the six positions of the primer there is an equal probability to find one of the four bases, so there are 4^6 *i.e.* $\sim$ 4000 different primers.

higher in any specific region and therefore the random primer pool will provide relatively even coverage along the RNA population by priming randomly the RNA fragments;
- ligating **special adaptor oligos** at the end of the RNA using a T4 RNA-ligase, which can then be used to prime the RT.

These strategies have strengths and weaknesses which will be discussed below.

2.4.2.1. Oligo-dT priming cannot be used for RT of fragmented RNA. Oligo-dT priming starts the RT from the poly-A tail of the mRNA and is one of the oldest RT methods. The principal advantage of this method is that the probability of priming is not influenced in a major way by the RNA sequence since all mRNAs are primed, regardless of their upstream sequences, reducing possible biases in library construction. However, this approach has several drawbacks. The first, and most important for our purposes, is that in case of fragmented RNA, only the 3′ region attached to the poly-A tail would be primed, while *all the other fragments would be excluded*. A second similar limitation is that, even when the RNA sample is not fragmented, every long RNA which is not polyadenylated cannot be primed and therefore this method cannot be used when detection of ncRNAs is required. Moreover, this method can be applied only to eukaryotic mRNA[12]. Last but not least, the fact that the RT polymerase is not a 'fast' enzyme increases *the bias of the read distribution towards the* 3′ *end of the mRNA* as opposed to the 5′ end.

In case of using non-fragmented RNA to reverse transcribe the cDNA, the obtained library cannot be fragmented through chemical methods, but only with enzymatic (not advised) or mechanical procedures (such as nebulisation and sonication)[13].

2.4.2.2. Advantages and disadvantages of random primers. Random primers can be used to prime RT from fragmented RNA libraries due to the fact that the probability of their binding to RNA molecules is roughly equal along the 5′ - to 3′ extent for any RNA type (mRNA, lncRNA, and so on). For this reason, they also solve the 3′ positional bias of the oligo-dT priming.

However, there are two limitations to this method: the first one is that the random priming is not exactly even, but there are some *priming*

[12] In fact, bacterial mRNAs are not polyadenylated.

[13] However, physical fragmentation induces the loss of at least the 50% of the starting cDNA!

hotspots. Principally, that is because of the thermodynamic stability and the propensity of each hexonucleotide to interact with its complementary sequence. The second one is that the combination of fragmentation and RT (if not prepared *ad hoc*) usually leads to the *loss of strand information*—this means that, given a cDNA molecule, it is not possible to determine whether the cDNA was synthesised from the sense or the the antisense strand[14].

2.4.2.3. Use of end-ligated primers. Priming with pre-ligated oligos is used in Illumina RNA-seq preparation of miRNA libraries and in other RNA-seq platforms. This approach ligates an adapter with a known sequence to the 3′ of the RNA fragment using a T4-RNA ligase: the sequence can then be used to prime the RT. This approach combines the strengths of the oligo-dT and of random primer strategies: it can be applied to fragmented libraries, but it is devoid of hotspot bias and it retains strand-specific information.

2.4.2.4. Second strand synthesis and strand-specific libraries. At this point, the RNA library would likely consist of fragmented RNA-cDNA heterodimers. There are two possibilities of synthesising the second cDNA strand,

- to prepare a **strand-specific library**, where the information from the original strand of the transcript is retained,
- to synthesise the second cDNA strand without considering its orientation: this is commonly obtained through RNA nick synthesis, where an *RNase* degrades part of the second-strand RNA, the *E. coli* DNA polymerase I replaces it with a cDNA complementary strand, and a *DNA ligase* eventually joins the nicks.

Retaining the strand-specific information is an advantage during the transcriptome assembly from an unknown genome, or from a genome with poor gene annotation (for example when working with a *non-standard animal model*): this could reduce computing time—there is no need to reverse-complement the reads, because the orientation is known—and would identify the sense and anti-sense strand.

From an experimental point of view, the synthesis of a strand-specific library exploits the property of the *E. coli* DNA Polymerase I, which is able to incorporate in a DNA strand the deoxyuridine (dUTP) and the

[14] This information can be relevant because a significant proportion of protein-coding genes also generates some specific ncRNA transcripts from the opposite strand (antisense transcripts).

deoxythymine triphosphate (dTTP). The cDNA then undergoes a PCR using a polymerase which is blocked by dUTP, so that only the first strand (devoid of dUTP) can be amplified.

2.4.3. Addition of the adapters

The adapters are short, asymmetrical DNA fragments which are composed of various functional elements:

- two **amplification elements**, at the 5′- and at the 3′-terminus of the fragment, which are used for the clonal amplification of the construct;
- a **primary sequencing priming site**, juxtaposed to the insert, which is necessary to initiate the (single-end) sequencing reaction;
- an optional **inline barcode** (positioned at the 5′ of the insert), or
- a **inline index** (positioned at the 3′ end of the insert and sequenced in a separate reaction), which provides a unique label in order to discriminate reads obtained from different samples when these are pooled in the same lane of the sequencing flowcell (see text below);
- in case of paired-end sequencing, a **paired-end sequencing priming site**, which is necessary to initiate the coupled sequencing reaction.

There are many ways the adapters can be added to the fragments, such as PCR or direct ligation.

2.4.4. Quality control

Quality control of the library is an important step, which can be achieved with a quantitative real-time PCR (using adapter-specific primers) or through elecrophoresis assessment using the previously mentioned BioAnalyzer. The quality control step gives the possibility to 1) quantify the concentration of cDNA in the library and 2) to evaluate the presence of primer dimers (in case of blunt-end ligation, with BioAnalyzer). Table 2.3 shows the ups and downs for each quality control step.

	qPCR	BioAnalyzer
Precision of quantification	+++	+
Tuning curve needed	yes	no
Detection of adapter dimers	no	yes
Fragment size determined	no	yes

Table 2.3. Comparison between qPCR and the Agilent BioAnalyzer when used for the library quality control.

2.5. Library sequencing

The principle of Illumina sequencing is to incorporate fluorescent nucleotide analogues into the cDNA single-strand fragment at each sequencing cycle and to scan them in order to determine the sequence of the transcript (*sequencing by synthesis*). Here we will present the method briefly, since many resources (including the producer's) are available to obtain more details.

2.5.1. Bridge amplification and sequencing by synthesis

In order to sequence the library, the cDNA molecules are heat-denaturated at 95 °C and injected into the **flowcell** of the sequencer, a functionalised surface where the DNA functional groups ($3' \rightarrow 5'$ oriented) are able to bind the adapters[15]. A first cycle of cDNA elongation and denaturation immobilises a single-stranded copy of the library element on the flowcell, while the original library templates are washed away. Then, the free $5'$ end of the molecule bends and interacts with a free complementary functional group on the flowcell forming a bridge and a new elongation step occurs. After denaturation, this *bridge amplification* step leads to two $3'$-$5'$ oriented cDNA fragments. This process of **solid state amplification** is repeated until a series of **clusters** of the same molecule ($\sim 10^3$ molecules per cluster) is created[16]. The reverse fragments are cleaved and washed away, so that only the forward copies of the original template are preserved.

After cluster formation, a primary sequencing primer is added to the template so that it can seed further sequencing: at each cycle a reversible fluorescent probe (one for each nucleotide), complementary to the respective template position, is added to the growing read fragment sequence and imaged through laser excitation[17]. After reaching the desired read length, the sequenced fragment is denatured and washed away, and the index is primed and sequenced.

In case of *paired-end sequencing* a new step of bridge amplification is performed, and—just like in the previous sequencing step—the for-

[15] As a quality control step, some phage DNA (*PhiX*) is added in small amount: since the sequence of the phage DNA is known, this will ensure that no global sequencing errors occurred.

[16] The density of the initial library is a key player in the process: if the library is too concentrated, clusters will overlap making sequencing impossible; if the library is too diluted many positions on the flow cell will remain empty increasing the per base cost of the sequencing. For this reason, quantifying the library concentration is an important step.

[17] After one cycle of sequencing the fluorescence of the probe incorporated in that cycle is reversed, so as not to interfere with the imaging of the next template position.

ward template is cut and washed away. Then a secondary sequencing primer is added to the reverse template, the paired read is sequenced as previously seen, and, eventually, the second index is sequenced. The two indexes/sequences are then paired during the data analysis.

2.5.2. Single-end or paired-end sequencing?

The choice between single-end or paired-end sequencing (see page 17) should be made according to the experimental design and the biological question to address.

Paired-end sequencing is more cost- and time-intensive than single-end sequencing (virtually every fragment is sequenced twice), so it should be chosen only when the additional cost is justified by the advantage of retrieving the structural information provided by coupled sequencing. Two typical applications of paired-end sequencing are the assembly of a previously unknown transcriptome (particularly when the reference genome is not available) and the analysis of different splicing isoforms. In the latter case, single-end sequencing provides only limited information on splice junctions[18]. For routine gene expression analysis in organisms with known transcriptomes and annotated genome, the most common use of RNA-seq, it is usually sufficient to opt for a single-end sequencing.

2.5.3. Choosing the right read length

The **read length** is determined by the sequencing platform and currently is in the range 50-300. The optimal read length again depends on the biological question to address:

- a simple **gene expression analysis**, where mapping of the reads against an annotated reference genome is performed only for counting purposes doesn't require long reads (**50 bp** are sufficient to identify the *gene of origin*);
- an **analysis of splicing isoforms** would require longer reads ($\sim$ **100 bp**), because the longer the read the higher the probability that it includes a splicing junction, this analysis would also benefit from paired-end information;
- a **transcriptome assembly** is improved by the availability of RNA structural information, so the reads should be as long as possible ($>$**150-200 bp**), ideally with paired-end information.

[18] In this case the splicing information would be provided only by the reads which include the splicing junction (see Figure 4.1), but it would not include which are the exon combinations for a given RNA transcript.

2.6. Other applications of next-generation RNA sequencing

There are many further possible applications of next-generation sequencing of RNA to the neurosciences. The most common are:

Small RNA sequencing. The only difference with 'conventional' RNA-seq is that RNAs are size-sorted to enrich for specific classes of non-coding RNAs. A very common application is **miRNA-seq** (by selecting RNAs in the 18-30 bp length range) to profile microRNA expression. The statistical analysis of the data is very similar to RNA-seq, however, a complication is that miRNAs are subject to post-transcriptional events of RNA processing giving rise to a multitude of *isomiRs*.

Ribosome footprinting. In this technique, binding of RNA to the ribosome is stabilised by cycloheximide and the RNA is then enzymatically degraded. Only RNA molecules bound to ribosomes are protected and can be sequenced. This technique provides a quantification of actively-transcribed mRNAs (**translatome**[19]), has a codon-level resolution, and requires depletion of rRNAs—this step can be technically challenging. The use of the drug lactimidomycin, which induces the specific stalling of the ribosomes initiating the translation (but not the elongating ones), is very useful to identify alternative translational starts on the mRNAs [29].

RNA-immunoprecipitation sequencing. In this approach, antibodies that bind specific **RNA-binding proteins** are used in order to immunoprecipitate a specific population of RNAs before sequencing. The use of antibodies against components of the RNA-induced silencing complex (RISC) provides a picture of repressed RNAs that can be particularly useful in combination with translatome analysis and miRNA-seq.

Sequencing dependent on the RNA chemistry. The versatility of next generation RNA-seq platforms makes it possible to further investigate the molecular biology of RNA at a single-nucleotide level. Sample preparation protocols that focus on isolating the mRNA fragments containing the 5′ cap (CAGE-seq, DECAP-seq, ...) are able to distinguish gene Transcription Starting Sites (TSS); other protocols are tailored to recognise specific post-transcriptional modifications of RNA sequences, such as pseudouridine (Ψ-seq) or 2′-O-methylation (2OMe-seq), as the reverse transcriptase blocks its activity in the pres-

[19] A variation of the translatome analysis is the purification of polysomes by centrifugation methods.

ence of pseudouridine conjugated with the molecule CMC or of 2′-O-methylated nucleotides[20]. RNA-immunoprecipitation techniques using specific antibodies make possible to study other post-translational modifications such as N^1- and N^6-methyladenosine (respectively, m1A-seq and m6A-seq).

[20] Indeed, some protocols designed to study the secondary structure of mRNA molecules (structure-seq/DMS-seq, CIRS-seq) take advantage of this fact: the induction of chemical modifications, such as the selective formation of a CMC-Ψ complex (that occurs when Ψ is in a single-stranded position), or the 2-O-Methylation using the chemical DMS at the level of unpaired Cytosines and Guanosines.

Chapter 3
RNA-seq raw data processing

3.1. Introduction

The human genome contains more than 20000 protein-coding genes, but the complexity of the RNA population in any given human sample is at least one order of magnitude higher due to alternative splicing that generates different splicing isoforms. To this, one has to add an increasing number of non-coding RNAs and various forms of RNA editing. This high complexity poses important technical and computational questions such as,

- how 'deep' should the planned sequencing be (*i.e.* how many clusters should be sequenced from the cDNA libraries) to obtain a good representation of the transcript diversity?
- Is the processing of the dataset (*i.e.* the identification of the *gene of origin* for each sequence) feasible in terms of computation time?
- Can the complexity be reduced?

In this chapter the problems of **complexity** and of **mapping** the RNA-seq reads to a the reference genome will be addressed from a probabilistic and informational point of view. The issue of reducing the complexity will be dealt with in Chapters 5 and 6.

3.2. General quality assessment

Having good-quality data is important in every branch of life sciences. However, in high-throughput data analysis, quality becomes a fundamental issue which could undermine the validity of all downstream analysis if not addressed correctly. The raw data from a sequencer come in the **FASTQ format**[1] and are associated to a quality score called the **Phred**

[1] The FASTQ format could be considered an extension of the FASTA format—a widely used standard for nucleotide and protein sequence deposition—that includes a Phred quality score (Q) for each

Quality score (**Q**), whose classical definition, as applied to the Sanger sequencing, is

$$Q = -10 \log_{10}(P) \tag{3.1}$$

where P is the probability that the base-calling for a given nucleotide sequence is inaccurate[2] (in that sense, the Phred score is a measure of sequencing quality).

The threshold to determine whether a single base in a read is of good quality or not is obviously arbitrary, however a reasonable separation adopted by the FastQC quality control software[3] considers base calls to be:

good quality bases if $Q > 28$ (*i.e.* the confidence is of at least 99.8%);
fair quality bases if $20 < Q < 28$;
poor quality bases if $Q < 20$ (*i.e.* the confidence is less than 99%).

The threshold for acceptable base-calling also strongly depends on the application: when analyzing RNA-seq data to obtain expression values, lower Phred scores are acceptable (because only mapping is relevant and sequence variation is disregarded) while analysis of RNA editing or allele-specific transcription will require higher Phred scores. A common quality bias in Illumina high-throughput sequencing is that the quality

nucleotide of the sequence (see text). The major structure of a FASTQ file is as follows:

- 1st line: a **header** in the form

  @InstrumentID:RunID:FlowcellID:FlowcellLane:Tile:Xpos:Ypos pair:N:0:index

  includes the technical information (from left to right: the instrument, the run, the position in the sequencing cartridge) and other information (in case of paired-end sequencing, which part of the pair is sequenced—1 or 2, Y/N shows if the read has been filtered or not, 0...2n attests that 2n control bits are on, and an esanucleotidic control index—*e.g.*, ATGCAT);
- 2nd line: the **sequence** in the single letter nucleotide code (GTCA and N for undecidable positions);
- 3rd line: a **+**, optionally followed by the header;
- 4th line: the **quality scores** associated with the sequence.

[2] That is, a Phred score of Q=30, which is close to the boundary between good and average measures, means that there is a probability of 0.1% that the associated base is not accurate. There are slight differences between the Sanger-applied quality scores (currently adopted by Illumina, after v1.8) and the older Solexa/Illumina pipeline score which followed the definition

$$Q_{\text{solexa}} = -10 \log_{10}\left(\frac{P}{1-P}\right)$$

which differs from the reference Phred score in low-quality values (Q<18), but is asymptotically equivalent for higher values.

[3] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

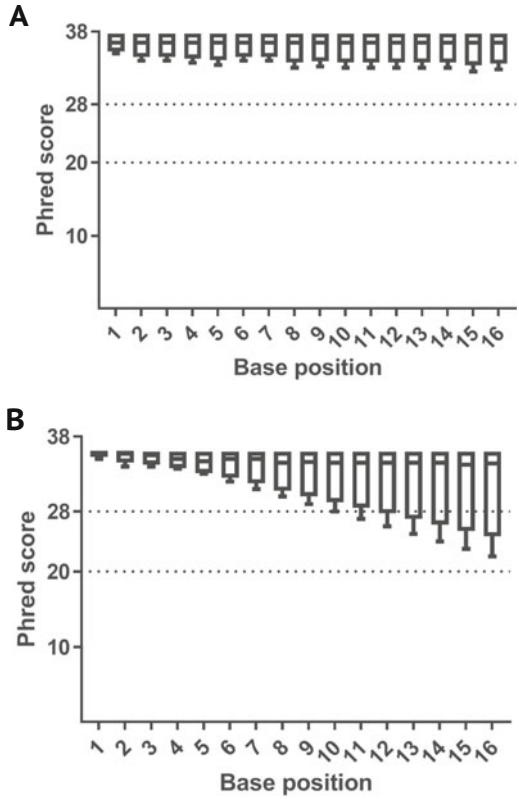


Figure 3.1. Simulated population distribution of Phred scores as a function of position A) in a good quality Illumina sequencing and B) in a poor quality sequencing. Notice the trend of lower Q scores for bases at higher positions in the read. For the reader who is unfamiliar with the boxplot representation, see Section 1.2 on page 6.

score of a base along the read is highly dependent on its position: the first bases have higher Q scores compared to the last bases (Figure 3.1).

The **single base Phred score** in the sequencing pool of reads can be used to assess **global quality of the sequencing run** by calculating the statistics of Phred scores across the set at different positions, or by plotting the distribution of the mean Phred score for each read (Figure 3.2).

Other important quality information is given by the **statistical properties of the sequences**: this is fundamental to assess the presence of **biological** or **technical sequence artefacts**. The underlying assumption is that a random sampling of the genome would imply a uniform

distribution of the different bases at a given position[4] and that no particular sequence pattern should be over-represented. If the frequency of the different bases as a function of position is not constant, or there are overrepresented sequences[5], technical (*e.g.*, Illumina primer dimers contamination, sequenced adapters and so on) or biological contaminations (*e.g.*, rRNA, mtDNA, polyA tails) are likely.

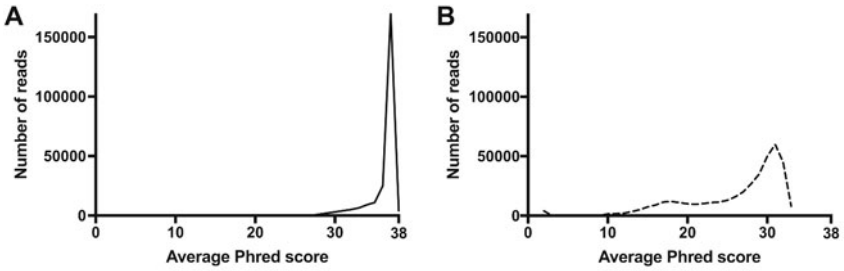


Figure 3.2. *Distribution of the mean Phred score for each read.* In the case of good quality sequencing (A), the distribution of the averaged per-read Q scores should be biased towards the values of $Q > 28$, with little or no population of lower quality scores (in particular with $Q < 20$), while poor-quality sequencing (B) has a wider average distribution.

3.2.1. The analysis of Kmer levels permits estimation of the presence of artifact sequences

A set of raw RNA-seq data could contain a certain amount of contaminant sequences, derived from the next-gen sequencing platform or from issues in sample handling. The presence of these contaminants can be detected as a change in the statistics of the base content in the reads. The easiest method to evaluate the presence of contaminant-related reads is to test for over-represented sequences (*e.g.*, duplicates), or to investigate the dataset for the presence of primer or adapter sequences. However this procedure depends on counting the occurrences of *exact sequences*. In the case of

[4] That is, at a given position of the read the probability of finding a G, C, T or A should be the same as their presence in the genome, so the dataset frequency of a given base in every position should be roughly constant.

[5] The quality control software FastQC (note 3) analyses which fraction of the dataset is composed of **duplicated reads** (usually a high amount of duplicated sequences, > 30%, is an indicator of some *enrichment bias*), the presence of **overrepresented sequences** in the reads (which could be associated with Illumina primers, or to other over-enriched molecular species), the presence of **adapter** sequences, and the analysis of **Kmer content** (see text for further details).

random errors, *e.g.*, due to

- sparse low-quality reads (more common in longer reads), or
- sequences which are inserted in the reads at different positions

the identity between the repeated sequences is lost, thus resulting in an underestimation of the presence of contaminants.

The analysis of the Kmer (read: k-mer) frequencies is a powerful and more unbiased approach for estimating the presence of contaminants. A Kmer is a string of K DNA bases: the Kmers derived from a read of length L are $L - K + 1$, so the number of Kmers in a dataset of R reads is $R \cdot (L - K + 1)$, and the complete set of DNA Kmers is composed of 4^K different strings.

If we assume a uniform distribution of different Kmers in the reads, the probability of finding n times a given Kmer in a certain part of the read after R reads is given by a binomial distribution, so one can carry out a *binomial test* to determine if there are overrepresented Kmers at a certain position, which is a flexible indicator for technical contamination.

3.3. Removal of artefacts

In case contaminants are detected after the first quality check of the sequenced dataset, the user should decide whether to remove the contaminants and continue the analysis, or to re-sequence the sample(s). Both options have drawbacks[6], but in case of bias correction there are at least four actions that can be undertaken.

Trimming The low-quality parts (*e.g.*, the endings) can be trimmed from the reads in order to get reads that can be mapped with more accuracy to the reference genome.

Clipping The adaptor sequences, that can be included in the length of the read when very short RNA fragments are sequenced, can be clipped[7]: it is crucial to remove this form of contamination when it is required to map short RNA fragments on the genome with high accuracy (*e.g.*, during the sequencing of miRNAs).

Filtering Corrupted sequences with low average quality value, or high amount of N (unassigned bases), Illumina derived sequences (*e.g.*,

[6] Rerunning the experiment is costly and may cause a time delay, but removing the contaminants from the dataset also causes a time delay and the resulting cleaned dataset might be suboptimal in terms of coverage.

[7] In some sources this process is called *trimming* of the adaptor sequences.

primers, or the internal control PhiX), artifactual low-complexity sequences, polyA stretches can be excluded from the dataset.

Biological contaminants removal Homology searching for conserved non-coding RNA sequences (such as rRNA, tRNA and so on) would help to filter out the most abundant sources of biological contamination.

After removing the artefacts, a good practice is to run the quality control pipeline again in order to test whether the quality issues have been solved.

3.4. Mapping the reads to the reference genome

Given a read, if we want to ask what gene it originates from, we need to find the best sequence match (*i.e.* to align with the biggest possible accuracy[8]) in the reference genome (or transcriptome). This process is called mapping. Many methods are available to identify the best match of a sequence in a database (the most used is probably BLAST[1]). However, these are designed to match sequences several hundred bps in length.

Mapping the reads of a typical RNA-seq experiment is a formidable task: a whole mapping would involve the matching of several tens of millions of short reads (30-100 bp) against a much larger genome (on the order of billions of base pairs in the case of vertebrate species). If, for example, 32 million reads (the size of the dataset corresponding to one sample in a typical RNA-seq experiment) need to be mapped and each mapping is completed in one second, the whole process would require roughly one year! So, the commonly used string alignment algorithms would be computationally too expensive and would not complete the task in a feasible time frame. In order to reduce the computational cost and time for the alignment, the **genome** should be reversibly **compressed** and associated with fast-to-consult reference indexes.

3.4.1. The Burrows-Wheeler transform

One of the most elegant approaches to solve the mapping problem applies a compression algorithm called the **Burrows-Wheeler Transform**[9]. Let's consider a string (S) of DNA, which is encoded by the four bases

[8] There are many possible principles to weigh the definition of a good alignment: for example the alignment should contain only few, short mismatches and prefer that lower-quality bases would mismatch compared to higher-quality bases.

[9] The Burrows-Wheeler Transform is an algorithm which has been developed in the early 1990s *specifically in order to compress a large amount of text*.

(ACTG) and the termination symbol ($, which could be in a genome the delimiter between two chromosomes):

$$A_0G_0C_0C_1A_1T_0G_1A_2\$_0$$

where the different subscripts are the **rank** of the base in the string[10]. If we take all the rotations of the string (*i.e.* all strings that are obtained by a recursive process where the last character is removed and placed at the first position of the string) and sort them in a consistent order, for example alphabetically (with crescent order $ACGT), we obtain the following array

$$\begin{array}{cccccccc|c}
\$_0 & A_0 & G_0 & C_0 & C_1 & A_1 & T_0 & G_1 & \mathbf{A_2} \\
A_2 & \$_0 & A_0 & G_0 & C_0 & C_1 & A_1 & T_0 & \mathbf{G_1} \\
A_0 & G_0 & C_0 & C_1 & A_1 & T_0 & G_1 & A_2 & \mathbf{\$_0} \\
A_1 & T_0 & G_1 & A_2 & \$_0 & A_0 & G_0 & C_0 & \mathbf{C_1} \\
C_1 & A_1 & T_0 & G_1 & A_2 & \$_0 & A_0 & G_0 & \mathbf{C_0} \\
C_0 & C_1 & A_1 & T_0 & G_1 & A_2 & \$_0 & A_0 & \mathbf{G_0} \\
G_1 & A_2 & \$_0 & A_0 & G_0 & C_0 & C_1 & A_1 & \mathbf{T_0} \\
G_0 & C_0 & C_1 & A_1 & T_0 & G_1 & A_2 & \$_0 & \mathbf{A_0} \\
T_0 & G_1 & A_2 & \$_0 & A_0 & G_0 & C_0 & C_1 & \mathbf{A_1}
\end{array}$$

which is called the *Burrows-Wheeler Matrix* (BWM) of S and where the last column (in bold font) is the Burrows-Wheeler Transform (BWT) of S. The BWT has the powerful property of retaining the information about the suffixes[11] of S (in italic font). It is easy to notice that the suffixes of S are sorted accordiing to the same sort order of the rotations of S. Moreover, the order of the ranks of the different bases is the same in the first column of the BWM and in the BWT[12]: this is the basis of the so called **Last-to-First** (LF) **mapping**. The LF function maps a given position of the BWT to the original string and thus can be used to 1) reconstruct the original string (Walk Left Algorithm), and 2) align a given query to S given the BWT. The equation which describes the function is

[10] The rank of a character in a string is the number of times the character has already occurred in the string. A character that is found in the string for the first time has rank 0, for the second time has rank 1, and so on.

[11] The suffix of a string is a substring which includes the end of the string itself, for example -ying is a suffix of both playing and copying

[12] If we consider the ranks of the sorted A bases in the first column of BWM (A ranked 2 at position 2, A ranked 0 at position 3 and A ranked 1 at position 4), and the ranks in the BWT (A ranked 2 at position 1, A ranked 0 at position 8 and A ranked 1 at position 9), it is visible that the order of occurrence of the ranks is the same.

the following:

$$LF(idx, char) = \text{rank}_{\text{BWT}}(idx, char) + \text{occupancy}(char) \qquad (3.2)$$

where *char* is the character which has to be mapped, rank_{BWT} is the rank of *char* in the BWT[13], and the occupancy is the number of characters lexically smaller of *char* which can be found in the BWT[14]. It can be recognised that the LF mapping of *char* is its position in the first column of the BWM.

Idx	F	BWT	rank_{BWT}	occupancy	$LF(idx, BWT)$
0	$\$_0$	A_2	0	1	1
1	A_2	G_1	0	6	6
2	A_0	$\$_0$	0	0	0
3	A_1	C_1	0	4	4
4	C_1	C_0	1	4	5
5	C_0	G_0	1	6	7
6	G_1	T_0	0	8	8
7	G_0	A_0	1	1	2
8	T_0	A_1	2	1	3

That is, given a character in the BWT, the LF function associated with that character is the value of the position of that same character in the first column of the BWM. This is fundamental, due to the property that if a character X has a position P in the first column of the BWM the character Y with position P in the BWT is going to be the character preceding X in the generating string—that is, the string has a form [...]YX[...].

[13] See note 10. Be careful because *the rank in the BWT is not necessarily the rank of the character in the string* (usually it isn't)!

[14] For example, the characters in the genome lexically smaller than G (if an alphabetical sorting is adopted) are $, A, and C. It's quite straightforward that

$$\text{occupancy}(char) = \sum_{char^* < char} (\text{rank}^{\max}_{\text{BWT}}(char^*) + 1).$$

Remember that here *both the rank and the raw position in the BWM start from 0*.

The *walk left* algorithm permits reconstruction of the generator string. Let's start from the end of string sign ($) in our example BWT, which has $\text{rank}_{\text{BWT}} = 3$, and let's calculate its LF value, which is 0. The character of the string with $\text{rank}_{\text{BWT}} = 0$ is A_2, which has a LF=1. Following the same procedure we thus find G_1 ($\text{rank}_{\text{BWT}} = 1$, LF=6), T_0 ($\text{rank}_{\text{BWT}} = 6$, LF=8), and so on. After three steps of the walk left algorithm[15] we have thus reconstructed the string $T_0G_1A_2\$$, which is exactly the four-characters suffix of the string S at page 32.

3.4.2. Application of the Burrows-Wheeler transform to genome mapping

After presenting the BWT, the next step is to understand how we can use it to reduce the computational time of genome mapping of RNA-seq reads. One can intuit that the BWT of the genome, whilst it maintains all the information of the originating genome, rearranges the indexes of the string in order to make the relation between characters of the string more accessible.

A fast way to search for a string $Q = q_0q_1q_2 \cdots q_{len(Q)-1}$ within a string S using the BWT is to apply a backward search approach, which involves a variant of the BWT data structure called the **FM** (*Full-text Minute-space*[16]) **index**, where the first column of the BWM and the BWT are stored, together with information of the offset of the suffixes[17]. Given a n-character string $Q = q_0q_1...q_{n-1}$ to query in a string S, the general

[15] The name derives from the fact that the proposed 'index translation' of a character from the rank in the BWM to the LF can be seen as going towards the left (*i.e.* the first column) of the BWM. This becomes explicit while visualising the first step of the algorithm.

F		BWT
$\$_0$	$\longrightarrow$	A_2
A_2	$\nwarrow^1$	G_1
A_0		$\$_0$
A_1		C_1
C_1		C_0
C_0		G_0
G_1		T_0
G_0		A_0
T_0		A_1

[16] But it was introduced in 2000 by Paolo Ferragina and Giovanni Manzini...

[17] The suffixes shown earlier in the BWM of the string S (page 33) can be associated with the **offset** of the suffix en respect of the first character of the string: the suffix which begins at the i^{th} character of the string will have offset i. So the suffix $ will have an offset equal to length(S)-1, while the suffix coincident with the whole string will have offset 0.

algorithm (in pseudo-code) is as follows:

```
top = 0
bottom = n
for char in Q
    top := LF(top, char)
    bottom := LF(bottom, char)
# [top, bottom) is the interval where
# to find the query after each step.
```

where in this case the function $LF(idx, char)$ is always applied *as if* the character in the BWT at position *idx* has value *char*.

As an example, let's try to find the query ATG in the string S (page 32). The first cycle of the algorithm would compute LF(0, G) = 6 and LF(n = 9, G) = 8, as if there was a G concatenated at the beginning and at the end of BWT[18]. The top and bottom pointers are then moved to position 6 and 7 (because the `bottom` value is excluded by the range), which are coincident to the interval of Gs in the column F.

	F	BWT		*Offset*
			$\leftarrow \text{top}_0$	
	$\$_0$	A_2		8
	A_2	G_1		7
	A_0	$\$_0$		0
	A_1	C_1		4
	C_1	C_0		3
	C_0	G_0		2
$\text{top}_1 \rightarrow$	G_1	T_0		6
$\rightarrow$	G_0	A_0		1
bottom_1	T_0	A_1		5
			$\leftarrow \text{bottom}_0$	

The second cycle of the algorithm takes into consideration the character T and evaluates the `top` = LF(6, T) = 8, and `bottom` = LF(7, T) = 9. In this example, there is no other T in the S string, and the bottom pointer goes outside the string, while the top pointer goes to position 8 of the F column (pointing to the only T).

[18] Caution is needed as LF(0,*char*) takes into account the original character at position 0 in the BWT, while computing the occupancy of *char*.

Query: ATG

	F	BWT		*Offset*
	$\$_0$	A_2		8
	A_2	G_1		7
	A_0	$\$_0$		0
$\text{top}_3 \Rightarrow$	A_1	C_1		4
bottom_3	C_1	C_0		3
	C_0	G_0		2
	G_1	T_0		6
	G_0	A_0		1
	T_0	A_1	$\leftarrow \text{top}_2$	5
			$\leftarrow \text{bottom}_2$	

The last step of the algorithm evaluates $\texttt{top} = \text{LF}(8, \text{A}) = 3$, and $\texttt{bottom} = \text{LF}(9, \text{A}) = 4$: the query process has ended, and the query is in the interval [3 4) of the BWM, that is at position 3 of the BWM:

Idx									
3	$\mathbf{A_1}$	$\mathbf{T_0}$	$\mathbf{G_1}$	A_2	$\$_0$	A_0	G_0	C_0	C_1 .

Given that the FM index includes the suffix index[19], it can also be established that the query string starts at position 4 of the original string

Idx	0	1	2	3	4	5	6	7	8
String	A_0	G_0	C_0	C_1	$\mathbf{A_1}$	$\mathbf{T_0}$	$\mathbf{G_1}$	A_2	$\$_0$

If the query was not present in the original string, GTG, the third step of the algorithm would estimate $\texttt{top} = \text{LF}(8, \text{G}) = 7$ and $\texttt{bottom} = \text{LF}(9, \text{G}) = 7$, that is the query is in the interval [7 7) which is a notation nonsense.

Query: GTG

	F	BWT		*Offset*
	$\$_0$	A_2		8
	A_2	G_1		7
	A_0	$\$_0$		0
	A_1	C_1		4
	C_1	C_0		3
	C_0	G_0		2
	G_1	T_0		6
$\text{top}_3 = \text{bottom}_3 \dashrightarrow$	G_0	A_0		1
	T_0	A_1	$\leftarrow \text{top}_2$	5
			$\leftarrow \text{bottom}_2$	

[19] In the shown example, it included the whole set of suffixes offsets.

The BWT query algorithm can be, extended to take into account other parameters as well, such as the Phred score of the aligned bases, or tolerance to mismatches/extensions, and so on.

3.4.3. Optimal storage of the suffixes index

In the example shown in Section 3.4.2, as soon as the backward search algorithm ends, we can associate the found string in the FM index with the sequence offset (*i.e.* the starting position of the query in the generator string), because all indexes of the suffixes associated with the BWM are stored within the FM data structure. However, storing all the suffixes indexes for the whole human genome would require about 12 Gb of memory, which is too much to be handled in a fast and economical way.

A solution to this problem is to store only an *evenly-spaced fraction of the suffix indexes*. If the found query is at a position which is not indexed, it would be sufficient to use the walk-left algorithm until an indexed position is reached.

`Query: ATG`

	F	BWT	*Offset*
	$\$_0$	A_2	8
	A_2	G_1	
	A_0	$\$_0$	
top $\Rightarrow$	A_1	C_1	
bottom	C_1	C_0	3
	C_0	G_0	
	G_1	T_0	
	G_0	A_0	
	T_0	A_1	5

In this case there is no saved suffix index for BWT position 3, so if we apply the walk left algorithm (page 3.4.1) we get position LF(3,C)=4. There is a saved suffix for BWT position 4, so we can apply the following

$$\text{offset}(\textit{query}) = \text{n}^{\text{o}}\ \text{steps}_{\text{wla}} + \text{offset}(\textit{char}_{\text{wla}}) \tag{3.3}$$

where $\text{n}^{\text{o}}\ \text{steps}_{\text{wla}}$ is the number of steps performed by the walk-left algorithm before finding a indexed BWT position (in the example $= 1$), and $\text{offset}(\textit{char}_{\text{wla}})$ is the value of the found offset (in the example, $= 3$). Applying the Equation (3.3) to the example thus gives $\text{offset}(\textit{query}) = 4$, which is the correct value of the index.

3.4.4. Structure of a SAM file

In RNA-seq applications, some flexibility in the mapping is allowed (a certain number of mismatches and indels is tolerated). After mapping the reads on the reference genome, the output is usually encoded in the SAM (**Sequence Alignment/Map**) **data format**. The key feature of SAM files is that they contain information on the position of the match in the genome and any mis-alignments between the individual read and the reference sequence. If we have the following alignment:

```
Alignment:  12345678901234 5678901234567890
Refer0001:  AGCATGTCAGATAG**GATAGCAGTGCTAGTA
Read001+:         TCAGATAGAGGATA*CAG
```

The corresponding SAM format is as follows:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
Read001+ 99 Refer0001 7 30
          8M2I4M1D3M = 37 39 TCAGATAGAGGATACAG *
```

The sample SAM file above has two main components: the header and the alignment sections. The **header** usually has multiple lines with each line starting with an @ and encodes information from the sequencing platform, the technical characteristics of the alignment and so on[20].The **alignment** section shows the alignment of the read with the reference sequence and is followed by a line with *11 mandatory fields* (together with optional fields), in the order:

GNAME the name of the read, the same in the FASTQ file;
FLAG bitwise flag, a combination of bitwise flags where each bit value associates to the read a precise alignment property—*e.g.*, 99 means that the read is paired (0×1), mapped in a proper pair (0×2), whose mate read should be reverse complemented (0×20), and it is first in the pair (0×40);
RNAME is the reference sequence name as found by the alignment;
POS is the starting position of the alignment of the read in the reference sequence;

[20] For example: @HD–aligment header (sorting of aligments,...), @SQ–reference sequence dictionary (reference genome, species, length of reference read,...) @RG–read group (sequencing platform, read information,...), @PG–programme, @CO–comments.

MAPQ is the mapping quality in Phred-scale, that evaluates the probability that a (high-Phred value) base is mismatched;

CIGAR the CIGAR string states how well the read is aligned to the reference, for example 8M2I4M1D3M means that, starting from the POS value of 7, there are 8 matches (8M), 2 insertions (2I), 4 matches (4M), 1 deletion (1D) and 3 matches (3M);[21]

MRNM is the Mate Reference Name ('=' if the same of RNAME, '∗' if unknown), that is the name of the reference for the paired sequence;

MPOS location of start alignment of pair mate on the reverse strand;

ISIZE inferred insert size, that is the maximal coverage of the reference sequence RNAME given by all the mapped reads;

SEQ sequence of the read ('∗' if not stored);

QUAL are the quality scores of the read (Sanger Phred scores, as stored in the FASTQ file, '∗' if not stored).

3.5. Complexity and depth of the sequencing

A key issue is the estimation of the complexity (*i.e.* the number of different molecules) of the library one has sequenced. A dataset of reads from an RNA-seq experiment is derived from a library containing C distinct cDNA species. The problem is to estimate the molecular complexity of the original cDNA ensemble through this limited sampling (the mapped reads). As it will be shown later, this is important from a conceptual and a technical point of view, because an estimation of complexity allows the scientist to make an informed decision as to increase the depth of the sequencing (or not!), and is a quality checkpoint of the cDNA library[22]. We show here a solution which is mostly applied to next-generation DNA sequencing, but that can also be of use when building a transcriptome.

3.5.1. The Negative Binomial distribution is commonly used to estimate the complexity of a library

As a first example, we want to estimate the complexity of a library composed of C different molecules of DNA that differ in their concentration using a given amount of reads R. The easiest way to estimate C is to consider the detection of a read from a given cDNA to be a stochastic event. While the Poissonian would be the easiest model, a major assumption un-

[21] The insertions, or deletions, are always considered from the point of view of the aligned read (deletion to the reference, insertion from the reference).

[22] Usually, if the estimated complexity is particularly low, this should be taken as a strong indication that something went wrong in the experimental preparation of the library.

derlying this distribution is that each cDNA has a uniform representation (*i.e.* concentration) in the library, and this is clearly not the case in transcript populations[23]. A widely-used approach is to model the distribution of the reads using a statistically overdispersed distribution, such as the Negative Binomial model:

$$P(x_\gamma = x \text{ maps after } r \text{ non mapping}) = \binom{r+x-1}{x} p^x (1-p)^r \quad (3.4)$$

that describes the probability of having x reads mapped to the cDNA γ after a number r of failures, if p is the probability of mapping a read.

In order to explain why the Negative Binomial is so commonly used, let's consider the probability that a given read comes x times from the cDNA γ. This is likely to be a random process which follows a *Poissonian probability distribution*

$$P(x_\gamma = x|\lambda_\gamma) = \lambda_\gamma^x \frac{e^{-\lambda_\gamma}}{x!} \quad (3.5)$$

where λ is an estimation of the frequency of the event to randomly occur, which in the case of a library complexity estimation would be linked to the prevalence of the species γ in the library, the effective complexity C and the number of reads R. So the data that are mapped on the genome could be explained as a mixture of Poissonians $P(x|\lambda_\gamma)$, with λ_γ composed of a constant component and a variable λ, which is distributed as a gamma distribution[24], that is

$$P(\lambda|\alpha, \beta) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \quad (3.6)$$

where $\Gamma(\alpha) = (\alpha - 1)!$ for $\alpha \in \mathbb{N}$, α and β are distribution parameters to be fit from the data, and thus the *gamma mixture* will have the form:

$$P(x|\alpha, \beta) = \int_0^\infty \frac{\lambda^x e^{-\lambda}}{x!} \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \mathrm{d}\lambda = \mathrm{NB2}(x|\alpha, \beta) \quad (3.7)$$

which can be reconducted to a Negative Binomial distribution with the complexity of the library and the dispersion of the data as parameters (for further information see Appendix 3.1 at the end of the chapter).

[23] Along highly-expressed transcripts (tRNAs, rRNAs, main homoeostatic genes), there is a plethora of lowly expressed—and arguably more interesting—transcripts. This condition is called **overdispersion**, when the variance between a random sampling of the different transcripts is higher compared to the one (expected) from a Poisson distribution.

[24] The choice of this distribution is justified by the fact that the gamma distribution is the **conjugate prior** of an *a posteriori* Poisson distribution.

The approximate estimation of the complexity of the library ($\hat{C}$) is given by

$$\hat{C} = \frac{M}{1 - \text{NB2}(x = 0|R/\tilde{C}, \tilde{\omega})} \tag{3.8}$$

where $\hat{C}$ is the estimated complexity from the sequencing, $\tilde{C}$ is the value of complexity which is expected from the library, M is the number of unique molecules mapped after the sequencing, and NB2 is the probability value of no reads for a molecule assuming a negative binomial distribution with parameters $R/\hat{C}$ (with R the number of reads) and dispersion $\tilde{\omega}$ (see Equation (3.15)). An interpretation of the Equation (3.8) is that the number of molecules that can be mapped to a reference genome in a run of sequencing depends on the effective complexity of the library, multiplied by the probability of detecting a molecule[25].

3.5.2. Marginal value of additional sequencing

Now that we have a good model to estimate the complexity of the library from the data, *e.g.*, from a shallow number ($< 10^6$) of reads, we can use this parameter to answer an important question: *how deep should we sequence the library in order to get the most information in a cost-time/effective way?*

Given Equation (3.8), the number of molecules mapped after a round of sequencing is proportional to the fitted complexity of the library and the probability of not detecting any reads for a molecule after R reads, as modelled by the Negative Binomial distribution (Equation (3.14)). So the change in the number of different molecules detected after adding r more reads is given by:

$$M(r) = \hat{C} \cdot (1 - \text{NB2}(x = 0|(R + r)/\tilde{C}, \tilde{\omega})) \tag{3.9}$$

whose behaviour is shown in Figure 3.3. As it can be seen, the number of different transcripts detected initially rises quickly when the number of reads (coverage) is low. But after a certain number of reads the average increase in the observed distinct molecules becomes almost flat—thus meaning that a further increase in depth (and cost!) of the sequencing would not be worth the amount of additional information that can be obtained. Also, and equally importantly, the higher the dispersion, the higher the numbers of reads necessary to obtain a substantial flattening

[25] Which can be stated as the complementary probability of not having any read for a molecule. Please note that Equation (3.8) contains the estimation of library complexity in both of its members, so it does not have a trivial solution.

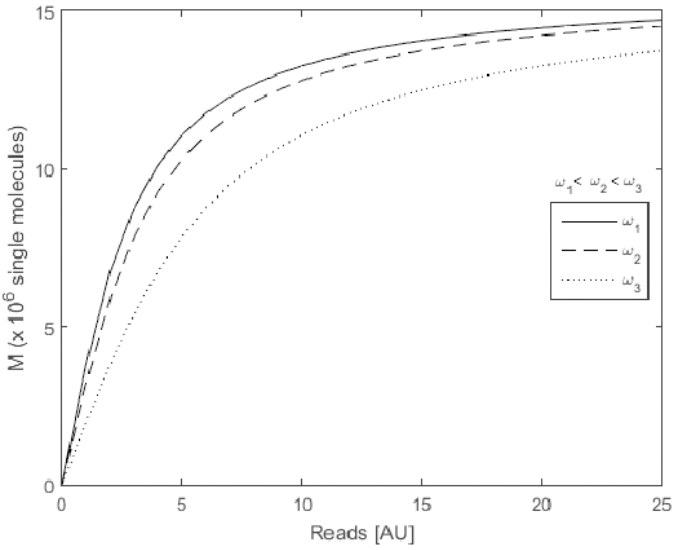


Figure 3.3. Qualitative behaviour of the number of expected single molecules to be observed (M) as a function of the total number of reads (R) in a synthetic library of $C = 15 \cdot 10^6$ unique molecules. The underlying model is a Negative Binomial distribution with different dispersion estimations ($\omega_1 < \omega_2 << \omega_3$).

of the curve. *In practical terms, a depth of 40 Million reads is generally considered adequate for a typical vertebrate RNA-seq experiment*.

Appendix 3.1 – Negative binomial derivation

Starting from Equation (3.7) it is possible to reconduct the gamma-Poisson mixture to the Negative Binomial distribution (Equation (3.4)), and, moreover, to reveal the relationship between the ancillary parameters of Equation (3.4) (r and p) and the hyperparameters of the gamma distribution which include the complexity (C) of the library and the dispersion of the original sample (ω).

Let's take out from the integral the functions which are not dependent on lambda and rearrange Equation (3.7) as follows:

$$P(x|\alpha, \beta) = \frac{\beta^\alpha}{x!\,\Gamma(\alpha)} \int_0^\infty \lambda^{x+\alpha-1} e^{-(\beta+1)\lambda}\, \mathrm{d}\lambda \qquad (3.10)$$

after resolving the integral of the form $\int x^n e^{-ax} \mathrm{d}x$ using $x + \alpha$ cycles of integration by parts and remembering that $\Gamma(x+1) = x!$ if $x \in \mathbb{N}$ we get

$$P(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(x+1)\,\Gamma(\alpha)} \left(\frac{1}{\lambda_*(\beta+1)}\right)^{x+\alpha} \Gamma(\alpha + x) \qquad (3.11)$$

which can be rearranged as

$$P(x, \lambda|\alpha, \beta) = \frac{\Gamma(\alpha + x)}{\Gamma(x + 1)\,\Gamma(\alpha)} \left(\frac{1}{\beta + \lambda_*}\right)^{x+\alpha} \beta^{\alpha} \lambda_*^{x}. \tag{3.12}$$

Given that

$$\frac{\Gamma(\alpha + x)}{\Gamma(x + 1)\,\Gamma(\alpha)} = \binom{\alpha + x - 1}{x} \tag{3.13}$$

we obtain the following:

$$\begin{aligned} P(x|\alpha, \beta) &= \binom{\alpha + x - 1}{x} \left(\frac{1}{\beta + 1}\right)^{x+\alpha} \beta^{\alpha} \\ &= \binom{\alpha + x - 1}{x} \left(\frac{1}{\beta + 1}\right)^{\alpha} \left(\frac{\beta}{\beta + 1}\right)^{\alpha} \\ &= \binom{\alpha + x - 1}{x} \left(\frac{1}{1 + \beta}\right)^{x} \left(1 - \frac{1}{1 + \beta}\right)^{\alpha}. \end{aligned} \tag{3.14}$$

So, if we apply the transformations $\frac{1}{1+\beta} \to p$ and $\alpha \to r$ to the previous equation we get the combinatorial form of the Negative Binomial distribution (Equation (3.4)).

One can also argue that the hyperparameter p (the probability of success) and r (the number of failures before success) are a function of the number of reads (normalised against the complexity of the library) and of the dispersion of the library. It can be shown that

$$r = \frac{1}{\tilde{\omega}} \tag{3.15}$$

$$p = \frac{R/\tilde{C}}{R/\tilde{C} + r} \tag{3.16}$$

where $\tilde{\omega}$ and $\tilde{C}$ are estimated iteratively from the data, for example using *Expectation-Maximisation algorithms*. For further information see [20, Chapter 8].

Chapter 4
Differentially expressed gene detection and analysis

4.1. Introduction

RNA-seq analysis of gene expression delivers, for any given sample, an estimate of the the abundance of transcripts derived from a gene (expressed as number of reads mapping to that gene). However, in order to derive useful biological information from these high-dimensionality data, it is necessary to compare these data with those obtained from other samples that differ in some biological variable of interest (*e.g.*, different tissues, conditions, time points, and so on). This requires a statistical model that—for each pairwise comparison—provides the probability (**p-value**) that the observed difference is due to chance (*i.e.* the probability that the measurements in the two groups are extracted from the same distribution). Only when this difference is significant (*i.e.* the p-value is sufficiently small), the gene can be defined as differentially expressed between the two conditions.

Due to the nature of the data, the methods of standard statistical cookbooks are not applicable to this problem. In this chapter, we will deal with how to detect differentially expressed genes (DEGs) from two RNAseq datasets.

4.2. Counting genes in a dataset

In Chapter 3, a simple strategy which employs the Burrows-Wheeler Transform to map many short reads to the reference genome was presented (page 33). Since the genome is annotated (*i.e.* the position of exons and introns are known, Figure 4.1) three different outcomes can follow from the mapping:

non-mappable reads reads that do not map to the reference genome (corrupted, artifactual, and so on);

mappable reads reads that map on a given site of the reference genome, *i.e*. reads from **unspliced** regions of the gene (*the vast majority of the reads* if short reads are sequenced);[1]

'splicing site' reads reads that map partially on two disjunct stretches of the reference genome (within the same gene), thus encompassing a **splicing site** of the gene[2].

The joint analysis of mapping and exon annotations makes it possible to **count** how many reads are associated with a given gene[3], which is reasonably a measure of the intensity of gene expression.

An interesting visualisation of mapping is the density of reads mapped along a genomic region. This can be obtained by 'piling-up' all reads mapping to the same region and counting (with base-resolution, if needed) the number of reads contained within a given sequence window. The cumulative overlap (or *coverage*) of the reads to the reference genome—when sequencing depth is sufficiently high—usually results in a roughly flat distribution along the exons of a given gene (with decreasing density in the vicinity of splice junctions) and much lower density on the introns. These **coverage vectors** are very compact ways to store the tallying of reads to the reference genome (they are integer vectors associated with each base pair in the chromosome), and are an estimation of the probability density for a given base pair to be represented in the sequenced dataset.

4.3. Detection of differentially expressed genes

Ideally, the process of gene counting can be compared to particle detection in physics. In the latter, a detector will count every time it is hit by a particle. In RNA-seq, we can image that all reads run sequentially through an array of detectors, each corresponding to one gene of the genome. Each detector will count how many times it was 'hit' by a read. So the processing of the sequenced dataset delivers a list of unique

[1] A special case of this scenario consists of the redundant mappable reads. That is, reads that map onto genes that are structurally repeated in the genome.

[2] This is an interesting piece of information when dealing with *different isoforms* of a given mRNA. A simple alignment algorithm would classify these reads as non-mappable: the algorithm requires special steps in order to detect the reads including a splicing site.

[3] The same dataset can be counted multiple times using different references. For example, one can use first using protein-coding RNAs, then non-coding RNAs, and finally transposable elements as reference. More importantly, the same dataset could (*and should*) be reanalysed if a new reference genome/transcriptome becomes available for the species of interest.

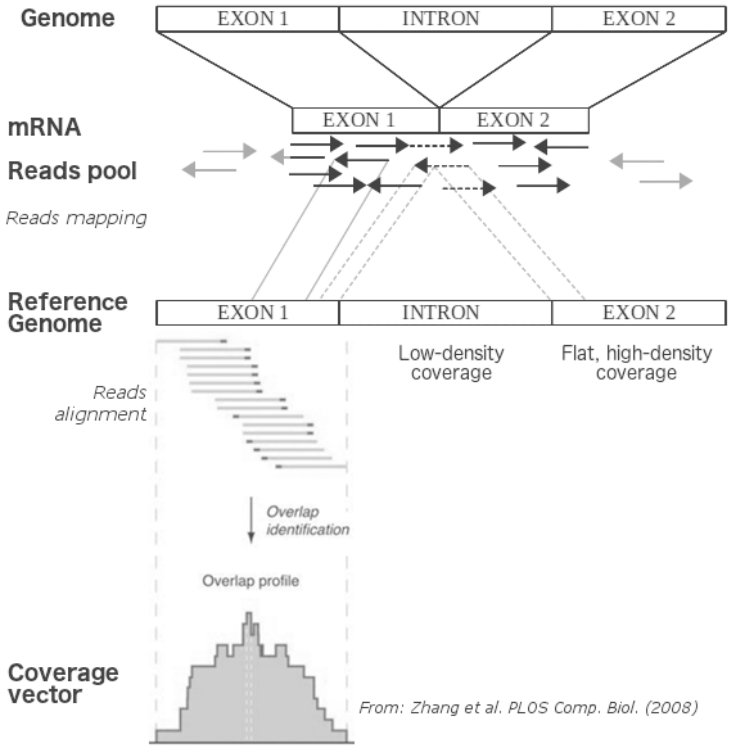


Figure 4.1. The mapping process can provide information on the splicing sites of the different transcripts. A mature mRNA is usually formed by splicing several exons together. The sequencing occurs randomly, starting from different base pairs of the corresponding cDNA, and could include (dashed black arrows) or not (solid black arrows) the splicing junction. In the former case, the read cannot be mapped completely to the same region of the reference genome, but to two disjunct, but neighbouring regions within the same gene; in the latter the read is mapped uniquely to a given region. A minority of reads (<10% in good-quality sequencing) cannot be mapped to the reference genome (solid grey arrows). The mapping density can be represented using a coverage vector (*vector reproduced from* [62] under a CC-BY License).

gene identifiers[4] each with an associated with an integer number (counts). Since transcripts have different lengths and libraries from different samples have different sizes (number of reads), some form of normalisation

[4] Every identifier is a string uniquely associated with a given genomic element (gene/exon/transcript) that allows an algorithm to retrieve the sequence and other information relative to the gene from a database. An example are the Entrez IDs, which refer to the NCBI database.

is needed to compare transcript abundance across samples. Three commonly used procedures are:

RPKM (*Reads Per Kilobase per Million mapped reads*) where mapped reads are first normalised to RPM (Reads Per Million, divided by the total number of reads in the library scaled by the factor 10^6), in order to compare datasets with different sequencing depths, and then for the length of the gene, assuming that a longer gene L would likely yield a larger amount of reads compared to a shorter gene S, when the expression levels of S and L are the same;

FPKM (*Fragment Per Kilobase per Million mapped reads*) have the same normalisation rationale of RPKM, however it is used in paired-end RNA-seq, so that two paired *reads* (or a single unmatched one) are considered a single *fragment* and are not counted twice;

TPM (*Transcript Per Million*), a relatively new normalisation strategy where the mapped reads are first normalised with the length of the gene and then with the total of the normalised reads scaled by the factor 10^6.

The inverted order of the normalisation steps is not the only difference between RPKM/FPKM and TPM. The TPM offers the great advantage that the sum of the TPM for a given sample is *always constant*, no matter the depth of the sequencing (this is not true for the RPKM values). A mathematical derivation of this claim is available in Appendix 4.1.

However, all three methods suffer from a possible limitation: if some highly-abundant transcripts change their level of expression—*e.g.*, due to the experimental condition—this can skew the entire distribution of reads leading to errors in the estimation of the library size. We will now present an early solution to this problem proposed by Anders and Huber [2]. In order to make a comparison between the data from different samples, perhaps with different sequencing depths, we have to model the properties of the set (see also Section 3.5.1 on page 40), in particular its *numerosity* and its gene expression *mean* and *variance*. In [2], the estimation of the numerosity of a dataset j is called **size factor** s_j. It is easy to recognise that the global expected value of gene expression for the dataset j is proportional to s_j, and thus this factor can be used to later normalise the single gene expression values. In fact, if we call $\mathbb{E}[K_{i,j}]$ the expected value of the gene i in the dataset of sample j and $\mathbb{E}[K_{i,j'}]$ the expected value of the gene i in the dataset of sample j' we have that

$$\frac{\mathbb{E}[K_{i,j}]}{\mathbb{E}[K_{i,j'}]} = \frac{s_j}{s'_j} \tag{4.1}$$

if the gene i is not differentially expressed in the samples j and j', or if j and j' are technical replicates. Anders and Huber make the reasonable assumption that the large majority of the genes are *not* differentially expressed. Therefore, in the case of two samples, the estimator of $\frac{s_j}{s_{j'}}$ is the median of $\frac{K_i j}{K_{ij'}}$ computed over all the expressed genes. In order to estimate the size factor of a sample j ($\hat{s}_j$) from its sequenced dataset $k_i j$, in the case of multiple samples, Anders and Huber propose to use a *pseudo-reference dataset* obtained from temperating the set with a geometric mean across all the samples and picking the median estimation (between all the genes) from this new dataset:

$$\hat{s}_j = \underset{i}{\text{median}} \left[k_{i,j} \sqrt[m]{\prod_{\nu=1}^{m} \frac{1}{k_{i,\nu}}} \right] \tag{4.2}$$

where m is the total number of samples. Empirical evidence shows that this approach can really provide better estimates of library sizes as compared to RPKM (Figure 4.2).

Let's now consider to have two different datasets, which have been prepared from samples with distinct conditions[5]. The first interesting biological question one could ask is *whether there are genes whose expression is modulated with high probability in response to the conditions*. This is a typical statistical question.

4.3.1. When is a difference significant?

To decide whether differences in gene expression are 'real', we need to **model the underlying statistics** of the dataset in order to determine when a given difference is unlikely to have occurred by chance. That is, when the difference of gene expression is **statistically significant**. The problem is simple in principle and for each gene we need to follow three steps:

1. Identify the probability density function which describes the distribution of the gene expression data.
2. Estimate the variance and mean of expression (and other relevant parameters) for each gene in the two conditions.
3. Calculate the p-value.

[5] *e.g.*, prenatal neurons/adult neurons, young nervous tissue/aged tissue, brain of healthy aged people/pathologically aged patients.

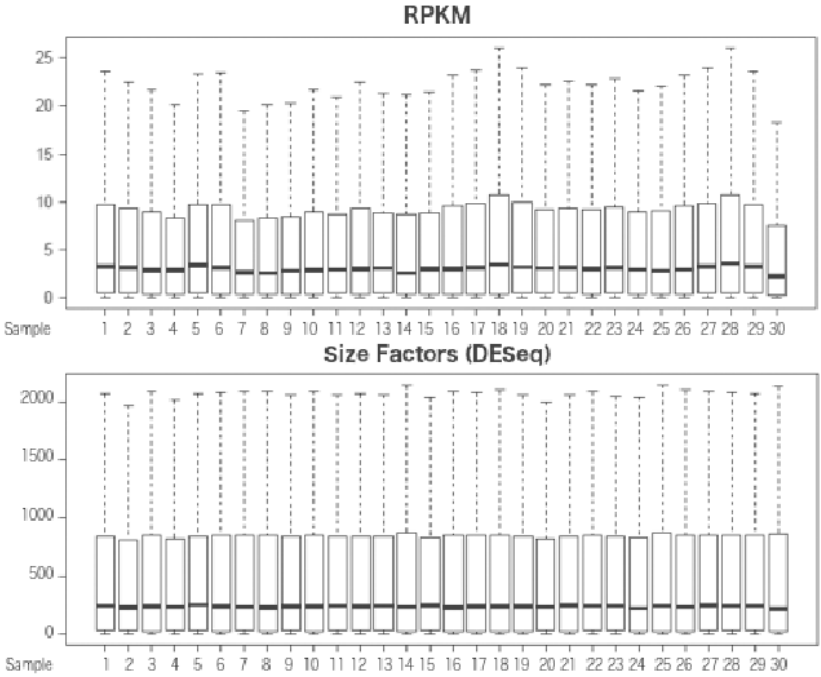


Figure 4.2. Comparison between two different normalisation methods: the RPKM and the 'size factors' normalisation step as described in [2] and discussed in the text. Each boxplot represents the distribution of expression values in one individual RNA-seq sample. The distribution of expression levels is represented as conventional boxplots (see Figure 1.3A at page 6), where the whiskers represent the 5-95% interval and there is no representation of the outliers. It is noticeable that the choice of a normalisation procedure can affect the distribution of expression values: while the RPKM method fails to produce homogeneous (*i.e.* correctly comparable) sets of transcript abundance, the 'size factors' method achieves a better result. Data from [6].

However, there are three major hurdles that complicate this approach:

- the distribution of the data (counts) clearly does not follow a Gaussian distribution (Figure 4.3), and the dynamic range of the data spans several orders of magnitude (seven in the example shown);
- normally, only a few replicates per condition are available, so a precise estimate of mean and variance is not possible;
- several thousand genes are compared in one experiment, so there is a fundamental necessity to *correct for multiple testing*.

4.3.2. Modelling the data distribution

As we mentioned in Chapter 3, sequencing can be seen as a series of Poissonian variables—as it represents a process of random sampling. Indeed, experimental tests have shown that *technical replicates* of the same

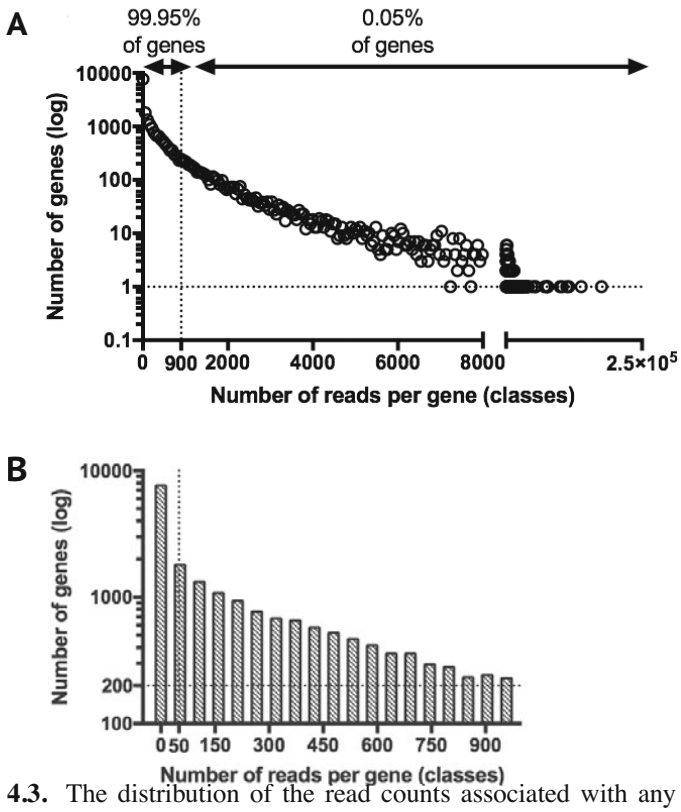


Figure 4.3. The distribution of the read counts associated with any gene in one example of RNA-seq dataset (in this case, the data derive from a *killifish* transcriptome) plotted in log scale to illustrate how the dynamic range spans several orders of magnitude. In this specific case, the depth of the sequencing was $\sim 20 \cdot 10^6$ reads. A) Global count distribution with associated cumulative frequency showing the extreme tail of the distribution: 99.95% of the genes are represented by less than 900 reads, however a few genes have counts > 9000. B) A close-up on the leftmost part of the distribution notes a strong prevalence of genes with less than 50 counts (vertical dotted line). The distribution can be fitted with a negative binomial.

experiment (*i.e.* replicate sequencing of the same library) are distributed according to a Poisson distribution [33]. A characteristic of Poisson distributions is that mean and variance are equal. Following Anders and Huber [2] again, we can partition the variance of (normalised) gene counts across samples in two components: the shot noise of the sequencing, that represent the poissonian technical variability, and the 'real' variance (*i.e.* related to biological processes)

$$\sigma^2_{i,j} = \underbrace{\mu_{i,j}}_{\textit{shot noise}} + \underbrace{s^2_j \nu_{i,\varrho(i)}}_{\textit{raw variance}} \tag{4.3}$$

where $\mu_{i,j}$ is the mean count of gene i in sample j, s_j is the size of sample j, and $v_{I,\varrho(j)}$ is a smooth function[6] of the expression level of gene i and further depends on the effect of the experimental condition of the sample j $\varrho(j)$.

As previously mentioned, the raw variance is the 'interesting' part of the variance and is due to biological variation and experimental condition. Our aim is to disentangle variation due to experimental conditions from technical and biological variation.

A Poisson process which occurs in a highly overdispersed sample (*i.e.* with high-variance, see note 23 at page 41) can be described as a **Negative Binomial**

$$R_{ij} = \mathrm{NB2}(\mu_{ij}, \sigma^2_{ij}) \tag{4.4}$$

where R_{ij} is the value of the read count in the sample j assigned to gene i, distributed as a Negative Binomial with parameters[7] μ_{ij} and σ^2_{ij}, which are respectively the 'real' mean and the 'real' variance of gene i in sample j. These population parameters are not known, but their value for each gene has to be estimated from the datasets. This is not easy since usually very few biological replicates are available.

Anders and Huber adopted a clever approach to solve this problem in their widely used DEseq package [2]. They again start from the assumption that the majority of genes are not differentially expressed. They further assume that the relationship between mean and variance is a 'smooth' function. So, to derive this key feature, they average the $\frac{\mu}{\sigma}$ values obtained from genes with similar expression levels and perform a local regression. (Figure 4.4). This function estimates the expected variance of expression for each gene. A very important aspect can be appreciated from Figure 4.4: at low expression values, the variance is dominated by the shot-noise (the Poisson and DEseq lines are very close) but at high expression values the raw variance is two orders of magnitude larger than the shot noise (note that the Y-axis is in log-scale).

[6] A smooth function is a function whose derivatives are continuous up to a certain order.

[7] A common parametrisation for NB2, derived from Equation (3.4) with $p = \frac{\sigma^2-\mu}{\sigma^2}$ and $r = \frac{\mu^2}{\sigma^2-\mu}$, is as follows:

$$\mathrm{NB2}(R = x|\mu, \sigma) = \binom{\frac{\mu^2}{\sigma^2-\mu} + x - 1}{x} \left(1 - \frac{\mu}{\sigma^2}\right)^x \left(\frac{\mu}{\sigma^2}\right)^{\frac{\mu^2}{\sigma^2-\mu}}.$$

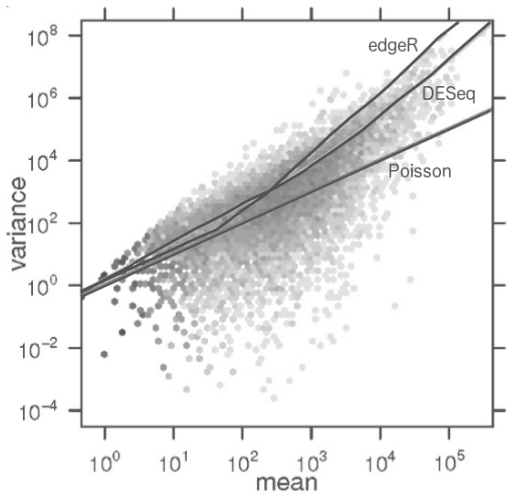


Figure 4.4. Fitting of the relationship between the common-scale variance from the scaled gene expression values (common-scale mean) of the genes in a *Drosophila* RNA-seq dataset. A Poissonian process would underestimate the relationship between mean and variance, because in a Poissonian trend μ is equal to σ^2—hence the straight line in the graph—while the DESeq approach (described by Anders and Huber [2] and partially discussed in the text) provides a better fit for the dataset. A second algorithm, edgeR [46], models the negative binomial by multiplying the variance with a constant instead of performing a local regression. This is shown to overestimate the μ/σ^2 dependence at high expression values and to underestimate it at low expression values. Figure adapted from [2] (http://creativecommons.org/licenses/by/2.0).

4.3.3. Testing the negative binomial model

Let's assume that we want to test whether a given gene i is differentially expressed between conditions A and B, where the number of replicates is m_A and m_B. This requires us to calculate

> the probability that the counts of i in condition A and B are equal to the measured values $K_{i,A}$ and $K_{i,B}$, under the hypothesis that the real expression values for i are the same in both conditions ($H_0 : \kappa_{i,A} = \kappa_{i,B}$).

$K_{i,A}$ and $K_{i,B}$ are respectively the expected values of m_A and m_B distinct Negative Binomial distributions whose parameters were estimated using the model described in Section 4.3.2. The paper from Anders and Huber [2] also describes how to calculate this probability under the aforementioned conditions. Other approaches use different models and methods to

obtain this probability[8]. Regardless of the statistical test used, the relevant point is that the probability we previously stated is the **probability of accepting the *null hypothesis*** (H_0), which is the logical negation of the working hypothesis. For each gene, one has to determine an 'idealised' statistical model of the data (*e.g.*, the *statistical test T*) that is associated with the probability distribution of the null hypothesis. The next step is to estimate the probability that the values obtained in the two conditions were extracted from this distribution: this is the **p-value** for the null hypothesis of the differential expression.

The gene is considered differentially-expressed if the p-value is smaller than an arbitrary significance threshold α. In this scenario, the null hypothesis is rejected under the assumption that the data are distributed as a Negative Binomial[9]. The value of α is usually 5% when the significance is evaluated for single hypothesis. However, in RNA-seq experiments, the number of genes that are tested simultaneously is in the order of 10^4, making it a clear case of **multiple testing**. In this case the probability of observing rare random events increases, so randomly occurred fluctuations could be erroneously detected as differential expression (type I error, or false positive). This problem is of paramount relevance not only in neurogenomics, but also in neuroimaging where a large number of pairwise comparisons between voxels are performed. It is thus necessary to correct the p-values for multiple comparisons, which would take into account the possibility of occurrence of type I errors.

The **family-wise error rate** (FWER) describes the probability of rejecting at least one true null hypothesis (*i.e.* calling DEG a gene that is not differentially-expressed). The **Bonferroni correction** ($\alpha' = \alpha/n$, where n is the number of multiple conditions) is proven to control the FWER, under the assumption all genes are expressed independently—a condition far from reality due to extensive gene co-expression patterns. Applying the Bonferroni correction to gene expression data, however, implies dividing α by $\sim 10^4$: few genes, if any, would remain DEG after this correction.

[8] This point of the statistical analysis, here dismissed with a few lines, is actually an absolutely central issue: different available programs make different assumptions to estimate mean and variance from the data, thus resulting in different performances. See, for example, [2]. Also, a completely different approach entails the transformation of the data into a nearly-normal distribution that allows us to use generalised linear methods developed for microarrays. A promising strategy is the use of Bayesian statistics—whose discussion would however greatly exceed the space of this book. The interested reader is referred to the DESeq2 follow-up article from Love, Huber and Anders [30].

[9] This assumption is not to be underestimated while evaluating the scientific value of a statistically significant finding. See also [58] for an extensive discussion on the meaning and the correct use of p-values in a scientific context.

The **false discovery rate** (FDR) is a less stringent condition compared to the FWER, as it is based on the concept that in high-dimensionality datasets one might tolerate a certain fraction of error. In fact, it is defined as the *proportion* of correct H_0 which have been rejected. So, an FDR with value 0.1 means that out of 1000 genes that are identified as DEGs, 100 are not real DEGs (obviously, by the chosen α). **FDR-controlled tests** for multiple conditions thus have **greater statistical power**, but also include **more false positives** (see also Appendix 4.2). A widely used FDR-controlled procedure is the **Benjamini-Hochberg-Yekutieli correction**. Let's consider n different null hypotheses H_0^1, H_0^2, ..., H_0^n with associated p-values P_1, P_2, ..., P_n and let's *rank* the p-values in crescent order $P_{<1>}$, $P_{<2>}$, ..., $P_{<n>}$ so that to each $i = 1...n$ would correspond a rank $r = 1...n$. Given a FDR (ϕ) and the ranked list of p-values, the Benjamini-Hochberg-Yekutieli procedure defines a $k \in \mathbb{N}$ so that

$$\max(k) \; : \; P_{<k>} \leq \frac{k}{n \cdot c(n)} \phi$$

where $c(n)$ is a function which takes into account the relationship between the tested conditions, and $c(n) = 1$ in case of independent conditions, or $c(i) = \sum_{i=1}^{n} i^{-1} \sim \ln(n)$ (for $n \gg 1$) if an arbitrary relationship is assumed between the conditions. The process thus rejects every null hypothesis associated with the ranked p-values $P_{<1>}$, ..., $P_{<k>}$.

A very important aspect of RNA-seq data, that derives directly from the relationship between shot noise and raw variance (Equation (4.3) as illustrated in Figure 4.4), is the relationship between expression level, fold change and p-value. This is normally expressed as an **MA plot** (Figure 4.5A) where expression is reported on the X-axis and fold change is reported on the Y-axis. DEGs are represented in a different shade of grey and the dashed line represents the 'border' of DEGs. It is immediately apparent that the lower the expression level, the higher the fold change needs to be in order to reach significance. An important corollary of this is that the number of DEGs that can be detected depends critically on the sequencing depth. So, in order to test the differential expression of genes with low expression levels high depths are necessary. Another very common representation of DEG analysis is the so-called volcano plot, where fold-change is represented on the X-axis and $-\log$ (p-value) on the Y-axis (Figure 4.5B).

4.4. Testing alternative splicing

Given the single-base resolution of the RNA-seq data, these also lend themselves to the analysis of differential splicing or of *differential exon*

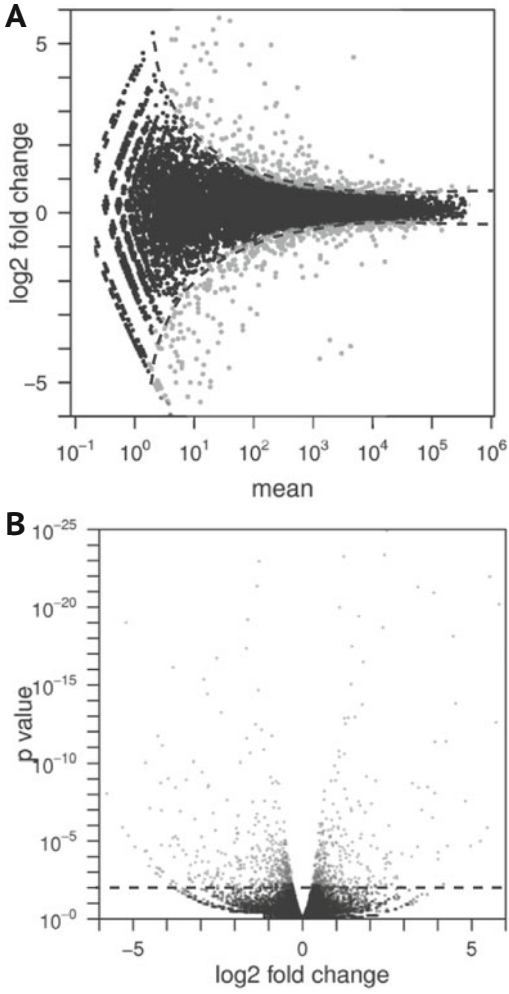


Figure 4.5. Statistical significance test of differential gene expression in two fly neural datasets. A) Plot of gene expression change as a function of mean read counts (*MA plot*). There is a strong correlation between statistical significance and read count—in other words, the statistical effect of the *shot noise* is higher in lower-represented transcripts. B) *Volcano plot* associated with the differential expression distribution A. Grey: statistically significant DEGs (Benjamini-Hochberg correction with 10% FDR, see Section 4.3.3), Black: non significant DEGs. Figure adapted from [2] (http://creativecommons.org/licenses/by/2.0).

usage. A key requisite for this analysis is a high-quality annotation of all exons and transcripts in the reference genome. Several different methods were proposed; for example when paired-end sequencing is available,

a sizeable proportion of paired reads will be placed in different exons facilitating this analysis (and also the annotation of previously unknown transcripts). An interesting approach to alternative splicing is taken by the package SplicingCompass [4]. In this case, the mean distribution of reads mapping to a gene is represented as a vector with n coordinates, where n is the number of annotated exons and the coordinates are the expression levels of each exon. The differential inclusion of an exon would determine a change in the *direction* of the vector that can be tested statistically.

Appendix 4.1 – Properties of TPM

The sum of the TPM for a given sample is *always constant*, no matter the depth of its sequencing, while this is not true for the sum of the RPKM values. Let's consider the RPKM for gene i in sample j to be

$$r_{i,j}^{\mathrm{RPKM}} = \left(\frac{R_{i,j}}{s \sum_i R_{i,j}} \frac{1}{L_i} \right)_j \tag{4.5}$$

where R_i is the number of reads mapped to i, s is the scaling factor, and L_i is the length of gene i, and its TPM as

$$r_{i,j}^{TPM} = \left(\frac{R_{i,j}}{L_i} \frac{1}{s \sum_i \frac{R_{i,j}}{L_i}} \right)_j . \tag{4.6}$$

Now let's consider the sum of the scores in the two cases:

$$\sum_i r_{i,j}^{\mathrm{RPKM}} = \left(\sum_i \frac{R_{i,j}}{s \sum_i R_{i,j}} \frac{1}{L_i} \right)_j = \left(\frac{1}{s \sum_i R_{i,j}} \sum_i \frac{R_{i,j}}{L_i} \right)_j \tag{4.7}$$

and

$$\begin{aligned} \sum_i r_{i,j}^{TPM} &= \left(\sum_i \frac{R_{i,j}}{L_i} \frac{1}{s \sum_i \frac{R_{i,j}}{L_i}} \right)_j \\ &= \left(\frac{1}{s \sum_i \frac{R_{i,j}}{L_i}} \sum_i \frac{R_{i,j}}{L_i} \right)_j = s^{-1}. \end{aligned} \tag{4.8}$$

So, the sum of the TPM values of a sample is always a constant number and in particular is the reciprocal of the scaling factor, thus it is usually a multiple of one million.

Appendix 4.2 – On errors and statistical tests

In Section 4.3.3 of this chapter we found the need to control for Type 1 errors in the statistical significance evaluation for multiple comparisons. The hypothesis testing involves rejecting or failing to reject the null hypothesis H_0.

Four scenarios can occur:

H_0 is $\rightarrow$ and is $\downarrow$	`True`	`False`
`Rejected`	**Type I Error** False positive	**Correct** True positive
`Failed to Rej.`	**Correct** True negative	**Type II Error** False negative

The **statistical power** β of a test (also called **sensitivity**) is defined as

$$\beta = 1 - \phi \tag{4.9}$$

where ϕ is the **false discovery rate** (see Section 4.3.3 at page 55). The power is the ability of the test to reject the null hypothesis, thus getting more sensitive in recognising the true positive data, but also more prone to false positives[10]. On the contrary, the **specificity** of a statistical test, which is the ability to recognise a true positive, depends on the significance threshold α (see page 55).

[10] A widely diffused minimum value for post-hoc power analysis of a statistical test is 80%.

Chapter 5
Unbiased clustering methods

5.1. DEGs analysis: the issue of complexity

After DEG analysis, the RNA-seq dataset becomes a matrix of vertically-organised vectors, where the rows correspond to the genes (with identifiers associated with gene annotation databases) and the columns to attributes such as fold-change (*e.g.*, up-regulated, downregulated, non-differentially regulated genes), p-value, FDR corrected p-value, or other relevant information and measures (Table 5.1). At this point, the numerosity of the data could easily overwhelm the human user and it is not unusual for a 'wet-lab' scientist to struggle and fail to extract relevant information from and make biological sense out of long lists of DEGs.

DEG ID (Ensembl)	Gene Name	Gene Description	Control	Treatment	$\log_2$ (FC)	p value
ENSDARG00000043856	amd1	adenosylmethionine decarboxylase 1	37.838	17.571	–1.107	0.037
ENSDARG00000087012	BX004816.3		0.816	0.201	–2.019	0.0215
ENSDARG00000052948	BX323876.2	similar to C–type natriuretic peptide 3	0.018	1.129	5.985	$5.10 \cdot 10^{-5}$
ENSDARG00000069620	BX323882.1		0.577	2.115	1.873	0.007
ENSDARG00000075757	BX510940.1	Gig2–like protein DreE	20.275	5.533	–1.874	0.042
ENSDARG00000058094	C19H1orf51	si:ch211–284a13.1	15.578	5.835	–1.417	0.004
ENSDARG00000045139	ca7	carbonic anhydrase VII	0.665	0.022	–4.91	$3.38 \cdot 10^{-8}$

Table 5.1. Example of output from a DEG-computing software.

It is a truth universally acknowledged that too much information is no information (at least for a scientist). In order to extract biologically relevant information from the detected DEGs, it is necessary to

- **reduce the complexity** of the data, *i.e.* to reduce the number of variables necessary to describe the dataset;
- **find structures** inside the data, using global methods in order to understand the relationships between samples, genes or group of genes: for example, to reveal a sub-structuring of the samples, to test whether

the experimental conditions influence global gene expression in a meaningful way[1], or to find groups of genes with similar expression patterns in order to find the function of non characterised genes ('**guilt by association**')[2].

Clustering and Principal Component Analysis (**PCA**) are two widely-used approaches to reduce complexity and detect structures within the data.

5.2. Clustering

The goal of gene expression clustering is to partition a group of genes into sub-sets, so that each sub-set contains genes that are 'similar' by some metric (*e.g.*, with similar expression pattern). This is ideally a straightforward approach to reduce the numerosity of the data and to capture some of the relationships between DEGs. The major challenges in clustering are

- *to define a measure of similarity between two genes (*i.e. *to define a distance*),
- *to apply the chosen similarity criterion in order to partition the data (i.e to define a clustering algorithm)*, and
- *to define what a cluster is*, bearing in mind that the issue of defining the optimal number of clusters is not trivial.

It is not hard to believe that this flexibility of definitions induced the development of many, equally valid, clustering strategies and algorithms, and there is indeed a vast literature on clustering methods[3]. There are four main categories of clustering algorithms that found application in transcriptome analysis: hierarchical, k-means, fuzzy c-means, and Self Organising Maps.

---

[1] For example, let's consider an experimental setup where there are RNA-seq datasets of samples from a disease model, a (healthy) control, and from the same disease model exposed to different treatments. A common question would be which treatment drives the transcriptome of the diseased model to be 'more similar' to the control one.

[2] A reasonable assumption is that some groups of co-regulated genes have similar functions. So, finding a non annotated gene with a similar expression pattern to other genes of a given class would suggest that the unknown gene is somehow related to that class. See also the example on regional gene expression and homogeneity at page 8.2.

[3] There is no precise and workable definition of cluster [11]. This fact implies that 1) different clustering algorithms are difficult to compare (unless they are based on a consistent set of assumptions), and 2) each clustering method is usually tailored to solve a specific problem. The user undergoing a clustering analysis should be careful to choose the best strategy according to the original data and the scientific question being addressed.

5.2.1. Hierarchical clustering

Hierarchical clustering organises the dataset as a **dendrogram**, where the root is the whole dataset, the leaves are the single genes, and each *internal node* (branching point) includes all the genes in the corresponding subtree. If we make an analogy to phylogenetic trees, nodes connecting two terminal branches define 'sister genes', and the subtree departing from more internal nodes define more distant relationships (sub-genus, genus, subfamily, family, and so on). In order to cluster the data the user should cut the dendrogram at a certain *internal node height* (which states how 'distant' are its child nodes) and consider the subtrees thereby obtained as the clusters of the dataset. Obviously, the choice where to cut the tree is totally arbitrary and so is the number of clusters obtained.

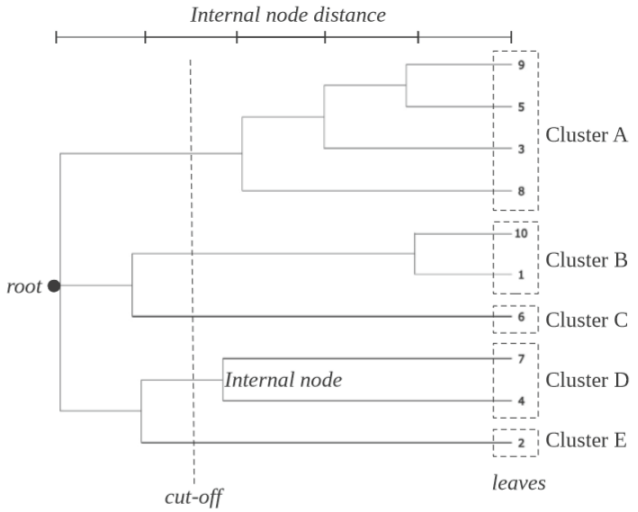


Figure 5.1. In hierarchical clustering the dataset (root) is organised in subtrees according to a correlation measure between the elements of the set and the clusters are obtained choosing a cutting threshold.

The usual input dataset for a clustering algorithm is a $n \times m$ matrix, where n is the number of samples (which usually correspond to different experimental/biological conditions), and m is the total number of genes to be clustered. The first step is to calculate the **pairwise distance** between all genes in the dataset, that provides a similarity measure between items in the set, in order to get a $m \times m$ **gene similarity matrix**. Here we describe some of the most applied distances. The **Euclidean distance** measures the geometric distance between two genes, that are represented as vectors in an Euclidean space of dimension n (where n corresponds to

the number of samples):

$$d_{ij}^{\epsilon} = \sqrt{\sum_{k=1}^{n} (q_{ik} - q_{jk})^2} \tag{5.1}$$

where q_{ik} is the expression level of gene i in the sample k. The **Pearson distance** (linear correlation) is a measure of how well the relationship between two gene expression profiles can be fitted by a straight line and the distance is 1 minus the correlation coefficient:

$$d_{ij}^{R} = 1 - \frac{\sum_{k=1}^{n} (q_{ik} - \mu_i)(q_{jk} - \mu_j)}{\sqrt{\sum_{k=1}^{n} (q_{ik} - \mu_i)^2} \sqrt{\sum_{k=1}^{n} (q_{jk} - \mu_j)^2}} \tag{5.2}$$

where μ_i is the mean of the expression value of gene i across all the samples. The **Spearman distance** (rank correlation) is a non-parametric distance, so it doesn't assume that the data are distributed normally:

$$d_{ij}^{S} = 1 - \frac{\sum_{k=1}^{n} (q_{(ik)} - \mu_{(i)})(q_{(jk)} - \mu_{(j)})}{\sqrt{\sum_{k=1}^{n} (q_{(ik)} - \mu_{(i)})^2} \sqrt{\sum_{k=1}^{n} (q_{(jk)} - \mu_{(j)})^2}} \tag{5.3}$$

whose formula is the same as that of the Pearson correlation, except the fact that the expression values q_{ik} are replaced by the correspondent *ranks*[4] $q_{(ik)}$. Figure 5.2 shows the behaviour of Spearman's and Pearson's correlation coefficients with different datasets. There are a variety of other possible distances that could be used on gene expression data, but these are applied less frequently.

After having applied the chosen distance to every pair of genes, we get an *upper-diagonal similarity matrix*[5] (S). The matrix is scanned in order to find the highest value (*i.e.* the most similar pair of genes): this pair of genes is linked by an internal node, which becomes a new 'virtual gene' whose distance to all other genes becomes the mean of the distance of the two original nodes $q_{(i,j)k} = \text{mean}(q_{ik}, q_{jk})$. Both genes i and j are replaced by the internal node (i, j) and his procedure is repeated $m - 1$ times until only the root remains[6]. A dendrogram is a tree-shaped graph

[4] Ranking a gene expression value means to associate the value with its position in a coherently ordered list (*e.g.*, decreasing expression level and so on).

[5] In fact, this matrix is symmetrical with respect to the diagonal, and the diagonal is composed of ones.

[6] This widely used procedure of hierarchical clustering was originally developed to reconstruct phylogenies and was first used by Eisen and colleagues in 1998 for gene expression data [13]. Some refer to it as *Eisen clustering*.

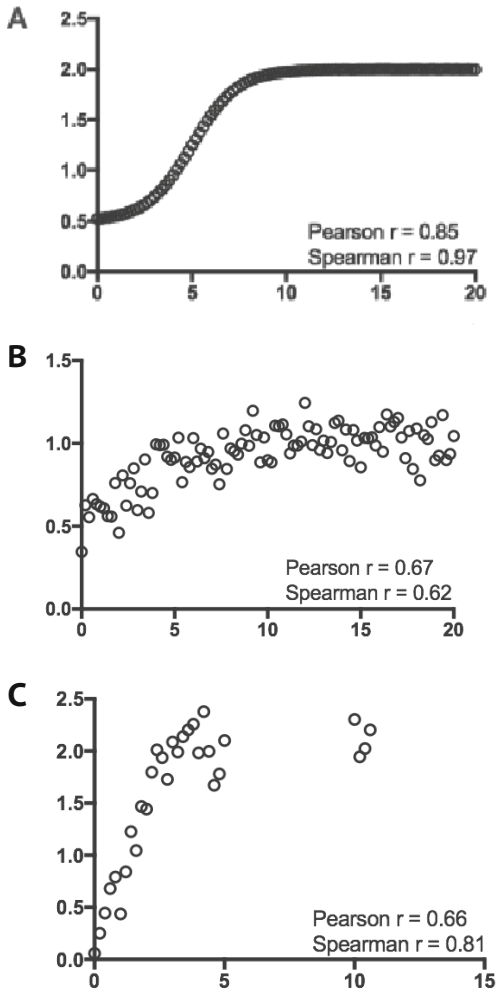


Figure 5.2. The main difference between Pearson's correlation and Spearman's correlation coefficients is that the former is very sensitive to a linear relationship between two variables (B) while the latter is less sensitive to linear relationships, but can detect any monotonic relationship (A and C).

that can be flipped at any node (imagine a physical tree with a rotating joint at every node). There are 2^{m-1} possible visualisations of a given tree and there is no privileged ordering of the leaves. So the user should apply a *simple ordering criterion*, *e.g.*, the averaged expression level of the gene, the time of maximal expression, and so on[7].

[7] Fitting an optimal linear ordering through minimisation of a cost function is possible, but computationally cumbersome and a waste of resources when done just for display purposes.

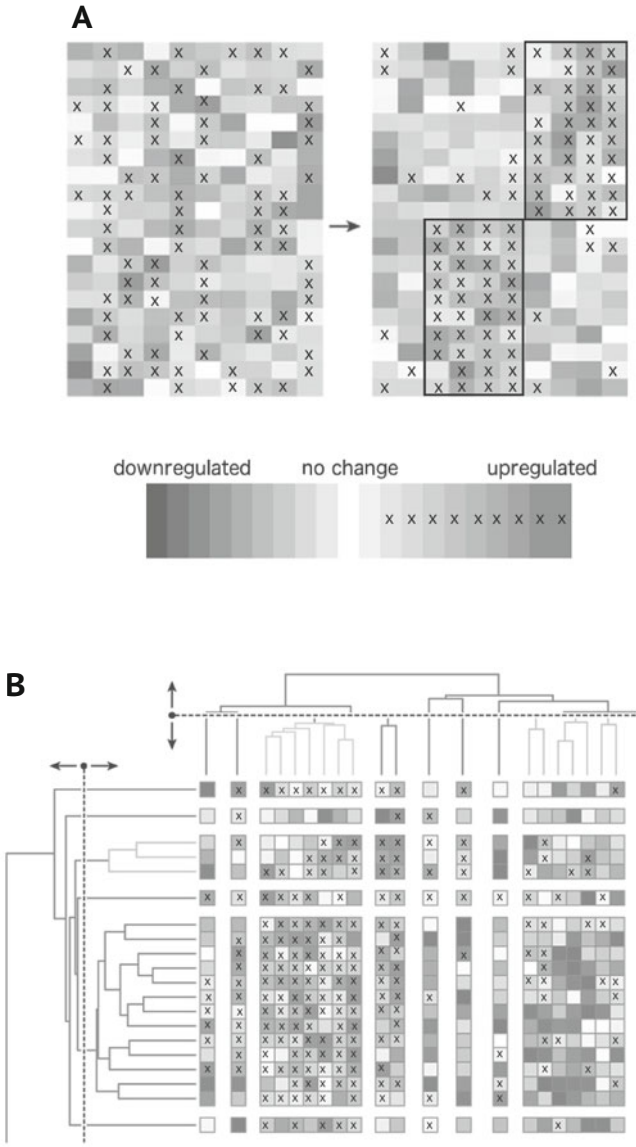


Figure 5.3. Heatmaps are a common way to visualise gene expression matrices—where the rows are usually the expression profiles of a given gene across different samples. In 1D maps (A) the samples are hierarchically clustered (black boxes) along one dimension—usually the samples—while in 2D maps (B) also the second dimension is ordered according to a hierarchical clustering. Adapted by permission from Macmillan Publishers Ltd: Nature Methods [16], © (2012).

The clustering is often coupled to a **heat map** of gene expression values, where different colours code for the intensity of expression. Since normally there are too many genes to be displayed as rows in a heat map, a very important issue is *feature selection*,, the process of selecting a subset of genes on which to perform the clustering. We will explore this issue in more detail in the PCA Section on page 70. (Figure 5.3A). Since the genes have different expression baselines and different fold changes, the rows in the heat maps are usually z-normalised. The z-normalised value of a gene i in the sample j is

$$z_{i,j} = \frac{k_{i,j} - \bar{k}_i}{\sigma_{k_i}}$$

where $k_{i,j}$ is the non-normalised gene expression value, $\bar{k}_i$ is the mean of the gene expression across the samples, and σ_{k_i} is its standard deviation.

Hierarchical clustering can also be used to cluster samples based on global gene expression and it is often used, for example, to subdivide tumour samples into different groups based on their expression patterns. Usually, the order of columns in the clustering is decided by the data analyst, for example corresponding to experimental series. It is also possible, however, to visualise a **two-ways clustering** where the samples are first clustered globally to decide the order of the columns (Figure 5.3B).

Hierarchical clustering is very important in quality control to detect **outliers** and **batch effects**. If one or a few samples show great distances from the other samples, they may be contaminated for various reasons and may be excluded from the analysis (outliers). If samples cluster according to the date of processing, then some systematic technical variability has influenced the results (batch effects). This effect can be later mitigated by PCA-based decomposition (see below).

5.2.2. K-means clustering

K-means clustering subdivides genes in a pre-determined number (k) of clusters. The clustering algorithm has three steps (Figure 5.4):

0. Initialise k cluster centroids at random positions (also called *seeding*);
1. Assign the genes to the closest centroid through minimisation of a distance measure

$$c_i = \min_{c_{(j)} \in \mathbb{K}} \left(\sum_f \| x_i^{(f)} - x_{c_{(j)}}^{(f)} \| \right) \tag{5.4}$$

where c_i is the centroid assigned to gene i, with features respectively $x_{c_{(j)}}^{(f)}$ and $x_i^{(f)}$;

2. Calculate the centroids of the clusters by averaging the features of the elements of the clusters

$$i \rightarrow c_{(j)}, \quad x_{c_{(j)}}^{(f)} = \text{mean}(x_i^{(f)}); \tag{5.5}$$

0′. Update the centroids.

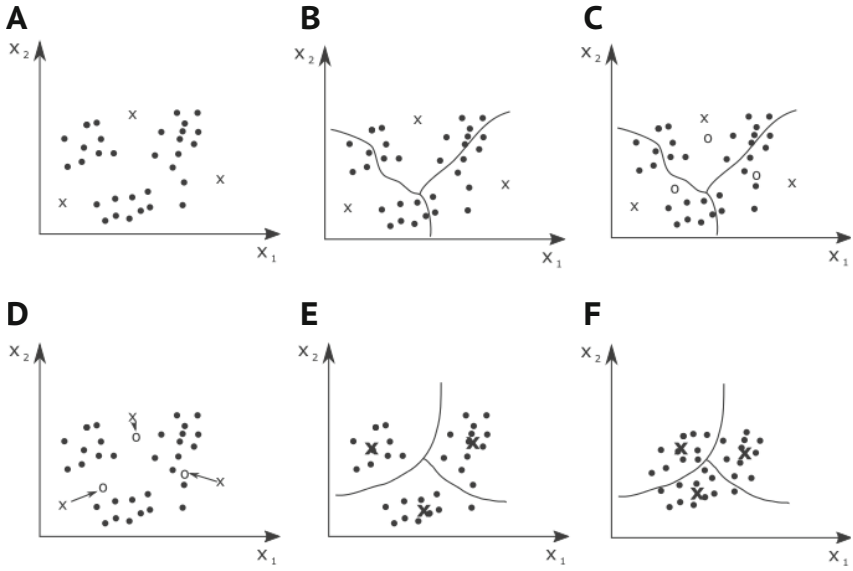


Figure 5.4. *Steps of the k-means clustering algorithm*. A) Random initialisation of *k=3* centroids. B) Assignment of the points of the dataset to the nearest centroid. C) Computation of the centroid of the found clusters. D) Update of all centroids. E) Clusters from a given initialisation of three centroids (**x**), however different initialisations (in particular with higher values of *k*) could result in different clusters. F) If we initialise three centroids, the algorithm will always find three clusters (centroid **x**), even if there are no noticeable subsets in the dataset.

This procedure is repeated until the assignment of the genes to the clusters is stable (*i.e.* no gene changes cluster). After k-means clustering, each gene is depicted as a point in a 2D graph where the distances between the points are equal to the calculated distance between the genes. Since different initialisations for the centroids might produce different final clusters, it is necessary to repeat clustering with different seeding initialisations, in order to check whether the clusters obtained are robust. It should be noted that k-means clustering will always divide genes into *k* clusters, even when the genes are homogeneously distributed (Figure 5.4F). It is possible to evaluate the 'goodness' of clustering objec-

tively using a clustering diagnostic such as the **Davies–Bouldin index** (DBI).

The DBI evaluates the relationship between the **intracluster distance** (that is, how much the genes in a cluster are packed around the centroid of the cluster)

$$S_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \|X_j - A_i\|_2 \tag{5.6}$$

where $\|\cdot\|_2$ is the Eucledian distance[8] between the feature vectors of the genes X_j and the centroid A_i, and the **intercluster distance** (how much close are the different clusters)

$$M_{i,j} = \|A_i - A_j\|_2 \tag{5.7}$$

which, again, is equal to

$$M_{i,j} = \sqrt{\sum_f \left(a_i^{(f)} - a_j^{(f)}\right)^2}. \tag{5.8}$$

A 'good' clustering would have its centroids as distant as possible, and the genes in a cluster packed around its centroid as much as possible. If we define the following index

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \tag{5.9}$$

R would be always positive and the closer to zero, the better the clustering of the elements of i and j. For every cluster i, let's pick the cluster j which is 'not a good distinct cluster' when compared to i (*i.e.* shows the highest value of $R_{i,j}$)

$$D_i = \max_{i \neq j} R_{i,j}, \tag{5.10}$$

the DBI is defined as the average of these worst-case scenario measures

$$\text{DBI} = \frac{1}{k} \sum_{i=1}^{k} D_i \tag{5.11}$$

and thus, the smaller the value of the DBI, the better defined and more distanced the clusters are.

---

[8] This can be generalised as a p-norm given by the formula:

$$\|X_j - A_i\|_p = \sqrt[p]{\sum_f \left(x_j^{(f)} - a_i^{(f)}\right)^p}.$$

It can be recognised easily that when $p = 2$, the p-norm is a Euclidean norm. When $p = 1$ the norm is called the *Manhattan distance*.

5.2.3. Fuzzy C-means

One of the limitations of the *k*-means algorithm is that a *gene can be assigned only to one cluster*. However, the transcription of a gene can be controlled by several independent mechanisms and these can be shared to various extents with other genes. For example, think about the binding of transcription factors to promoters and enhancers (different genes have different combinations of binding sites in their promoters). The **Fuzzy C-means** algorithm (FCM) overcomes this limitation allowing a gene to be member of more than one cluster.

The FCM is a three-step iterative procedure:

0. Initialise C random centroids;
1. Assign to every gene i a *membership weight* for the centroid c_j, which is a value $0 \leq w_i^{(c_j)} \leq 1$ which depends on the formula

$$w_i^{(c_j)} = \frac{1}{\left(\sum_{k=1}^{C} \frac{\|X_i - X_{c_j}\|}{\|X_i - X_{c_k}\|} \right)^{\frac{2}{m-1}}} \tag{5.12}$$

   where $\|X_i - X_{c_j}\|$ is the distance between all the features of the gene i and the centroid c_j, and m is a fuzziness parameter (if m=1 the FCM is reconducible to the k-means algorithm, while for $m \gg 1$ the membership for every gene gets small and the clusters become fuzzier);
2. The new centroids are evaluated according to a weighted mean using the membership of the genes

$$c_j := \frac{\sum_{i=1}^{n} (w_i^{(c_j)})^m \, X_i}{\sum_{i=1}^{n} (w_i^{(c_j)})^m} \tag{5.13}$$

   where n is the number of genes, and m is the fuzziness parameter;

0′. Update the centroids.

The algorithm will continue until convergence: like for the *k*-means clustering, the obtained clusters depend on the initial seed, so the procedure should be repeated to evaluate their robustness.

A particularly useful application of k-means and FCM clustering is in the analysis of temporal series or dose-dependent responses. They can, for example, distinguish genes that are regulated at different times after the onset of a stimulus. For example, Baumgart and colleagues [6] analysed the gene expression profiles in the killifish brain at different ages. Using FCM clustering they could identify genes that change direction of

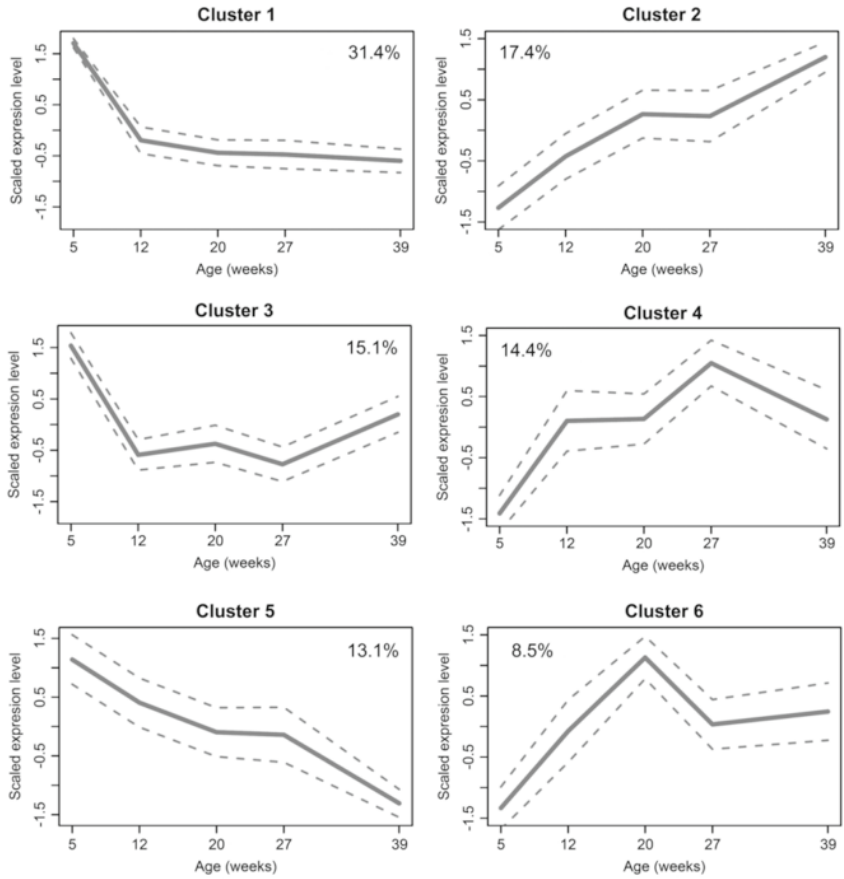


Figure 5.5. Fuzzy C-mean clustering of a RNA-seq dataset (killifish brain, 5 age steps from 5 weeks post hatching to 39 weeks). Notice that each cluster is prototypical of a different ontogenetic regulation. For example, both cluster 1 and cluster 5 contain genes that are down-regulated during ageing, but cluster 1 shows faster downregulation then does cluster 5. Interestingly, there are genes that show an inversion in the direction of their expression regulation (clusters 3 and 4).

regulation during ontogeny—*e.g.*, up-regulated during development and down-regulated during ageing or *vice versa* (Figure 5.5).

Given that both the *k*-means and the FCM algorithms divide the dataset in a given number of clusters, in a way the different centroids are 'forcing' a certain structure on the dataset. So if the number k (or C) of centroids is not reasonable given the 'real' structure of the dataset, the resulting clustering would be artefactual and fail to capture the biological phenomenon. Unfortunately, there is no consensus on which is the optimal method to calculate the optimal number of centroids. A common approach is to try

different centroid numbers and then to (hand-)pick the one that makes more biological sense to the analyst.

5.3. Principal component analysis

The Principal Component Analysis (PCA) is, as already mentioned, one of the most used approaches to 1) reduce the dimensionality of datasets and 2) reveal some hidden structure of the data. The aim of PCA is to take advantage of the strong patterns of correlation in gene expression levels *to find a small subset of dimensions which would describe as much of the variability in the dataset as possible*.

5.3.1. An intuitive view of PCA

The most straightforward way to visualise the aim of PCA is to consider a geometrical example, as pictured in Figure 5.6. So, if we imagine that

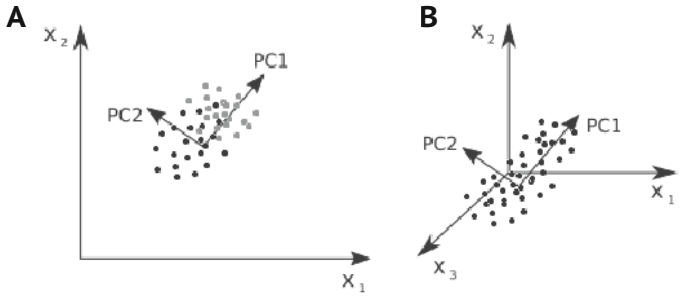


Figure 5.6. Simple geometric representation of PCA. A) New dimensions make possible to visualise the differences between two groups in the dataset (grey and black). B) PCA can be used to diminish the dimensionality of the dataset: in the example shown, all points in the dataset lie roughly on the same plane.

the data form an ellipsoid in an n-dimensional space[9], then PCA is a transformation that creates a new set of axis that corresponds to the axis of the ellipsoid. A quite naïve explanation is that it is 'easier' to build an ellipsoid from its axes rather than from a random set of *n* orthogonal axes. The major axis would be the first principal component (PC1) and it identifies the direction along which there is most dispersion of the data. Computing the PCA is, in first instance, a problem of dimensionality reduction. For example, the 3D dataset in Figure 5.6B was originally described with three features (x_1, x_2, and x_3), however the data are lying on a plane, which can be described by only two features—PC1 and

[9] This would roughly be the case if all variables are distributed according to a Gaussian function.

PC2, that could be obtained by a linear combination of x_1, x_2, and x_3. Therefore, PCA can be seen as a *rototranslation* of the axes that aligns them with the direction of maximum variability. In a typical RNA-seq experiment, the features (genes) are in the order of thousands whereas the samples are usually in the order of tens and normally correspond to different conditions. When applied to the analysis of RNA-seq data, PCA has two really useful properties:

1. due to very high levels of gene co-expression, PCA normally provides a tremendous **reduction in the dimensionality** of the data[10]. Usually the first ten components are sufficient to describe the dataset with good approximation and it is not uncommon for the first three components alone to retain 80% or more of the variance (see below for further discussion).
2. PCA can often **dissect the effects of different variables** on global gene expression patterns, as shown in Figure 5.7. In this experiment, neurons were generated by direct reprogramming of fibroblasts from donors of different ages. The aim was to test whether the fibroblast-derived neurons retained the age signature of the donor cells. From the PCA graph it is clear that the first two components (the x,y plane in the graph) capture the cell-specific components of gene expression and the third component (the z axis) describes an effect of age on global gene expression patterns that is very similar in both cell types. An inspection of the **loadings** (see Section 5.3.2) of PC3 would then easily reveal which genes contribute the most to this effect.

After having found the PCA components of the data, a widely used approach is to calculate the **explained variation** for each component. This is defined as the fraction of variance that a single principal component can account for (see the introductory explanation in Section 5.3.2). Let's consider Figure 5.8, which shows the retained variance distributions of two datasets. The first thing that can be noticed is that there is an enormous *reduction of dimensionality with contained loss of information*. At this point there are two possible scenarios: a high amount of the variance is explained by two or three components (Figure 5.8A), so that the other components could be discarded without losing a lot of information, or the variance is distributed among many of the principal components (*e.g.*, the first two components explain less than half the variability in the data, Fig-

[10] By definition, the PCA is a set of features which describe the highest variability in the dataset—assuming a linear correlation, so finding these features would automatically exclude redundant features and optimise the remaining ones.

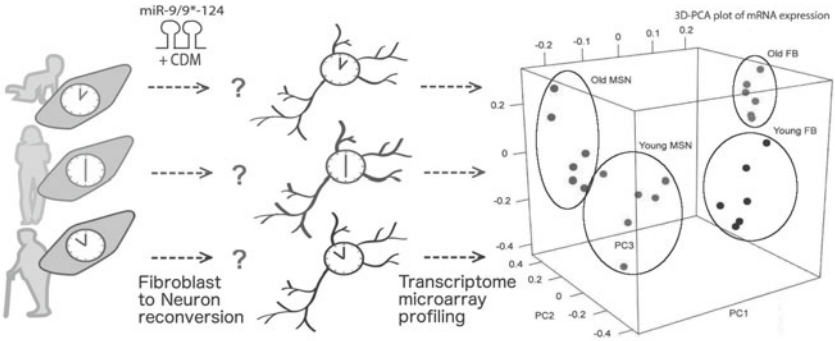


Figure 5.7. A PCA projection onto the first three principal components is able to distinguish the age and the cell type in the trascriptional profiles of human fibroblasts (young and aged) and miRNA-reconverted neurons (from either young or aged fibroblast). FB: host fibroblasts, MSN: directly-converted striatal medium spiny neurons. Adapted from [22]. https://creativecommons.org/licenses/by/4.0/

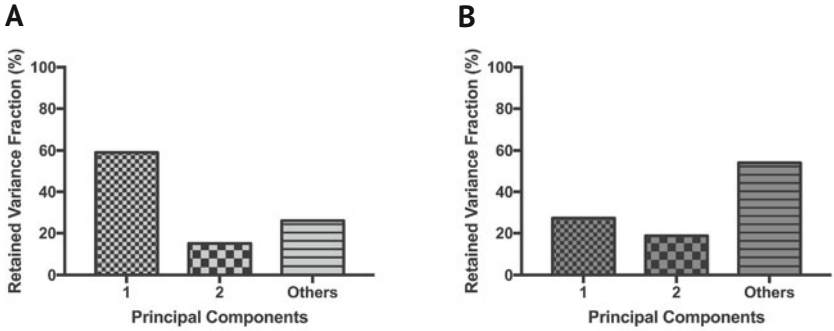


Figure 5.8. The retained variance distribution according to the different principal components (PCs) could vary massively with the dataset. A) The first two PCs of a killifish RNA-seq dataset (see Figure 5.12, analysis courtesy of Mariateresa Mazzetto) retain most of the variance. B) The distribution of the retained variance in the PCA on a dataset extracted from the Allen Human Brain Atlas (see Section 8.2 at page 122) shows that the PCs excluded from a 2D visualisation account for more than 50% of the whole variance.

ure 5.8B). In the first case, the data could be easily represented in a 2D/3D plot with only minimal loss of information; in the latter[11] a representation of the higher principal components would take into account roughly 50% of the total variance, so it would be better to represent the data with multiple combinations of principal components in order to look for different

[11] Indeed, this variance distribution is not uncommon when dealing with biological datasets. See for example the analysis of microarrays on breast cancer sample as shown in [45].

structures inside the data[12]. In particular, correlation only captures linear

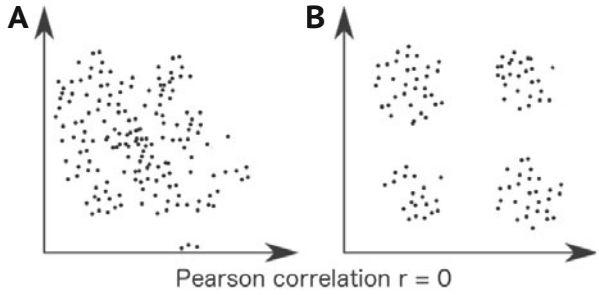


Figure 5.9. Limitations of linear correlation: both the point distributions in A and B have a null linear correlation, however while in A there is no evident relationship between the points, in B there is a marked nonlinear dependence.

relationships among the data and does not capture more complicated relationships (Figure 5.9). Indeed, PCA is tailored to normally-distributed data and may not perform well in the case of data with distributions that greatly deviate from normality. It should also be mentioned that PCA is sensitive to the size of the variables and so z-normalisation (see note 5.2.1 at page 65) should be performed before PCA.

An important property of PCA, in particular when dealing with high-throughput technologies, is the fact that, if an unfiltered systematic experimental artifact introduces a strong, highly-correlated source of variance, this will result in a dominant principal component. However, this PC can be subtracted from the data reducing this bias, that would normally mask the biological sources of variance. Indeed, in some instances, PCA-based methods were used to remove *batch effects*, systematic effects due to processing different samples at different times. Again, the raw data quality is a fundamental issue in RNA-seq data analysis.

Another key issue in PCA analysis is **feature selection**, that is performing a PCA analysis only on a subset of genes in order to obtain a clearer separation of different samples according to a chosen feature. Gene selection may be based on:

- the most expressed genes;
- the genes with the highest coefficient of variation;

[12] The linear correlation of the different principal components is zero, so each one could potentially capture independent sources of the variance in the dataset [45]. However this is valid in strict sense only if the variables are near-normally distributed, which is very often not the case. Nonetheless, in practical terms, PCA can provided useful information also for data that are clearly non-normal (see Figure 5.7 and Figure 5.10).

- the most differentially-expressed genes;
- the genes presenting characteristic defined *a priori* (*e.g.*, being included in a specific KEGG pathway, see Chapter 7, and so on).

This can be particularly useful when samples differ for more than one experimental condition, as exemplified by the example of Figure 5.10. In this case, samples from brain, liver, muscle and skin were analysed at five different ages in two different strains of the turquoise killifish. One can use either the genes differentially-expressed between the two strains, or those differentially-expressed due to age[13]. When performing PCA on these two sets, in both cases the first two components sharply separate samples originating from different tissues. In Figure 5.10A, PC3 captures a tissue-independent effect of genetic background but samples of different ages are not separated. On the other hand in Figure 5.10B a tissue-independent effect of age is clearly visible, but the samples are not separated according to the strain.

5.3.2. A computational approach to PCA

How can we find the set of dimensions which maximises the represented variance? Another geometric interpretation, as seen in Figure 5.11, could be helpful. The points in the dataset (a sample in our case) have n single-feature values, which are their 'projections' on the corresponding feature axis. So if we consider a new unit axis $\hat{\mathbf{u}}$ (which is a linear combination of the existing axis directions), projecting the points of the dataset on $\hat{\mathbf{u}}$ gives a new set of coordinates with associated statistics (*e.g.*, variance, mean, and so on). So the first principal component is the direction in which the variance of the projections of the points in the dataset on the new axis is maximal. If we remember that the distance of a vector $\mathbf{x}$ on a versor $\hat{\mathbf{u}}$ s given by

$$\mathbf{x_u} = \mathbf{x^T} \cdot \hat{\mathbf{u}} \tag{5.14}$$

where x^T is the transposed vector of $\mathbf{x}$, the problem to solve—when considering the whole dataset $\mathbf{X} = [\mathbf{x}^{(1)}\mathbf{x}^{(2)} \cdots \mathbf{x}^{(m)}]$ composed of m samples—can be stated as follows

$$\max_{\|\hat{\mathbf{u}}\|=1} \left(\frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}^{(i)^\mathbf{T}} \cdot \hat{\mathbf{u}})^2 \right) \tag{5.15}$$

[13] In this case, since several time points are available, instead of the approach described in Section 4.3.3—the test for differential expression in multiple conditions fitting a negative binomial—a *generalised linear model* is used to test a regression with age.

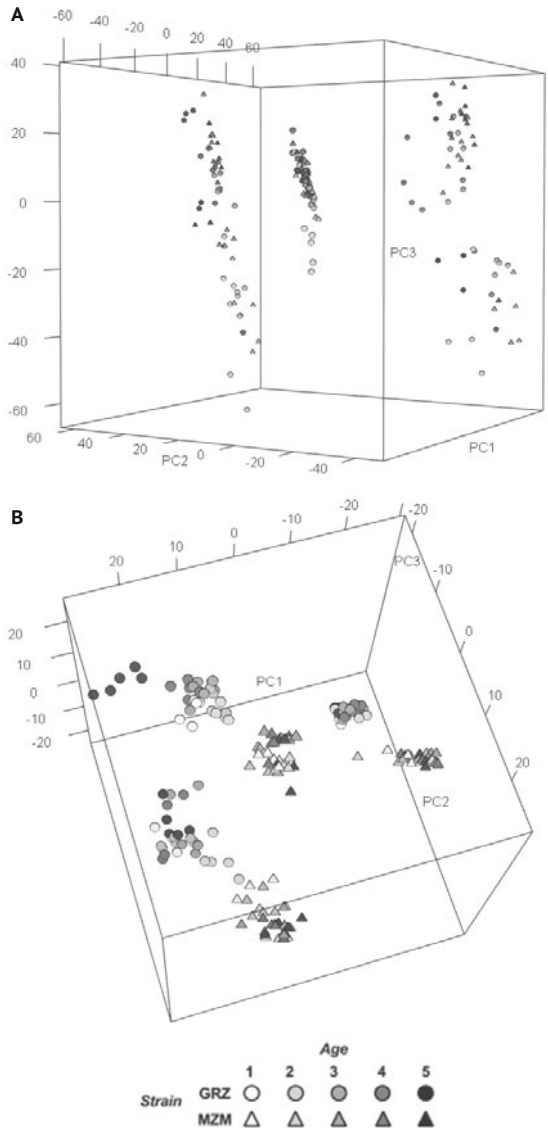


Figure 5.10. Example of feature selection in PCA. A) Feature selection for the genes involved in ageing makes it possible to distinguish the different ages along the third principal component. However, the samples from the different strains are mixed. B) On the other hand, feature selection for the strains is able to distinguish the samples from GRZ and MZM, but it loses the age separation. In both cases PC1 and PC2 resolve the tissue identity (liver, muscle, brain or skin). *Images and analysis courtesy of Mariateresa Mazzetto*.

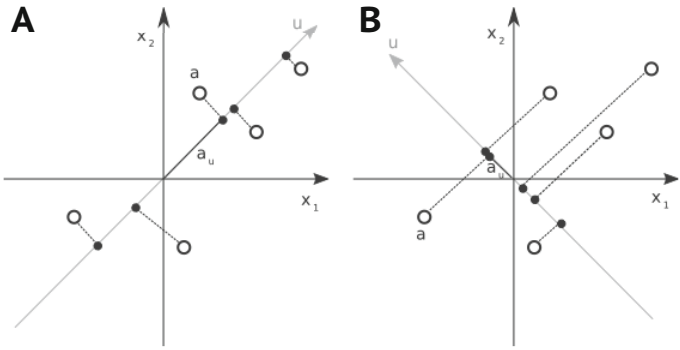


Figure 5.11. The projection of points of the dataset on a given axis u could result in a group of vectors with high-variance (A), or with a more uniform size (B). For example if we consider the point a and its projection on u (a_u), in the plot A) a_u is quite far from the origin and the global variance is quite high, while in the plot B) a_u is close to zero and has a quite homogeneous size compared to the other vectors.

with the condition $\|\hat{\mathbf{u}}\| = 1$, that is $\hat{\mathbf{u}}$ should be a unit vector, and where $\mathbf{x}^{(i)}$ is the vector of features associated with the condition i.

Let's now work on Equation (5.15):

$$\frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}^{(i)^{\mathbf{T}}}\cdot\hat{\mathbf{u}})^2 = \frac{1}{m}\sum_{i=1}^{m}(\hat{\mathbf{u}}^{\mathbf{T}}\cdot\mathbf{x}^{(i)})\,(\mathbf{x}^{(i)^{\mathbf{T}}}\cdot\hat{\mathbf{u}}) \tag{5.16}$$

which can be rearranged as

$$\max_{\|\hat{\mathbf{u}}\|=1}\left(\mathbf{u}^{\mathbf{T}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}\cdot\mathbf{x}^{(i)^{\mathbf{T}}}\right)\hat{\mathbf{u}}\right) = \max_{\|\hat{\mathbf{u}}\|=1}\left(\hat{\mathbf{u}}^{\mathbf{T}}\cdot\Sigma\cdot\hat{\mathbf{u}}\right) \tag{5.17}$$

where Σ is the *covariance matrix* of the data, under the assumption the dataset has zero mean[14].

It can be shown that the solution to Equation (5.17) is the **principal eigenvector** (*i.e.* the eigenvector with the highest eigenvalue[15]) of the

[14] In order to have zero-mean data, provided that the data points are distributed as a Gaussian, the following normalisation step is done:

$$x^{(i)} := \frac{x^{(i)} - \mu_i}{\sigma_i}$$

where μ_i is the mean and σ_i is the standard deviation for a given experimental condition. The resulting normalised dataset has zero mean and unit standard deviation.

[15] Given the matrix Σ, there is a set of vectors $\mathbf{u}_1, \mathbf{u}_2, ...$, and numbers $\lambda_1, \lambda_2, ...$, so that

$$\Sigma\cdot\mathbf{u} = \lambda\,\mathbf{u}$$

covariance matrix. *In general, the first k principal components of the dataset are the first k eigenvectors, as sorted by their associated eigenvalue*. The projection of the dataset $\mathbf{X}$ on the matrix of the eigenvectors (or of the selected principal components) $\mathbf{U}$ can be computed as follows:

$$\mathbf{X}_{\mathrm{PC}} = \mathbf{U}^{\mathbf{T}} \cdot \mathbf{X}. \tag{5.18}$$

From Equation (5.15) and note 15 of this chapter one can easily derive that the i^{th} eigenvalue of the covariance matrix Σ includes the variance of the dataset when projected on the i^{th} principal component.

Let's now try to compute the *total dispersion* (or variance) of the dataset in function of the known parameters of the covariance matrix starting from the following equation:

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 \tag{5.19}$$

where m is the number of samples in the dataset, $\mathbf{x}_i$ is the vector of gene expression in the sample i, $\bar{\mathbf{x}}$ is the vector of the mean of the gene expression values across the samples, and $\|\cdot\|$ is the norm operator. If we apply the vectorial form of Equation (5.18) to Equation (5.19) we have:

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} \| \sum_{j=1}^{p} \hat{\mathbf{u}}_j^{\mathbf{T}} (\mathbf{x}_i - \bar{\mathbf{x}}) \|^2 \tag{5.20}$$

where p is the number of principal components. So taking the sum out from the norm we get

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{p} \|\hat{\mathbf{u}}_j^{\mathbf{T}} (\mathbf{x}_i - \bar{\mathbf{x}})\|^2 = \sum_{j=1}^{p} \left(\frac{1}{m} \sum_{i=1}^{m} \|\hat{\mathbf{u}}_j^{\mathbf{T}} (\mathbf{x}_i - \bar{\mathbf{x}})\|^2 \right). \tag{5.21}$$

However, the expression in the parentheses is the variance of the projection of $\mathbf{X}$ on the principal component j, that is equal to the eigenvalue λ_j, so the expression is reduced to

$$\sigma^2 = \sum_{j=1}^{p} \lambda_j. \tag{5.22}$$

where λ is called the *eigenvalue* and $\mathbf{u}$ is called the *eigenvector* of Σ.

From the Equation (5.22) it is possible to estimate the **retained variability** of a given principal component $\hat{u}_h$:

$$\sigma_h^2 = \frac{\lambda_h}{\sum_{j=1}^{p} \lambda_j}, \tag{5.23}$$

σ_h^2 is also called *relative dispersion*.

Let's call Λ the diagonal matrix of the eigenvalues of $\mathbf{X}$. We define the matrix of the **component loadings** as the product

$$L = U \cdot (\Lambda)^{\frac{1}{2}} \tag{5.24}$$

and because Λ is a diagonal matrix we have that the i^{th} *component loading* is given by

$$\mathbf{l}^{(i)} = \hat{\mathbf{u}}^{(i)} \sqrt{\lambda_i} \tag{5.25}$$

where λ_i is the eigenvalue associated with the i^{th} principal component. The component loadings are a sort of *weighted principal component* which takes into account both the direction of maximal variability and the value of the retained variability. In every loading, the position $\mathbf{l}_j^{(i)}$ roughly corresponds to the 'explanatory contribution' of the component i to the gene j. So, if a biological feature is strongly correlated with a principal component, it is possible to recognise a set of genes which are likely to be involved in that process.

5.4. Multi dimensional scaling

The aim of Multi Dimensional Scaling (MDS) is to map a high-dimensional dataset to a lower-dimensional space (**MDS Maps**[16]) in order to maintain the relative 'general distances' between two points in each representation. MDS can be seen as an alternative data visualisation tool to PCA. However, if the PCA takes into account the *covariance matrix* of the dataset (see Equation (5.17)), the MDS can use any kind of similarity or dissimilarity matrix, of which the covariance matrix is essentially a particular case.

Let's build a distance matrix applying to the gene dataset (made of n genes) one of the measures of distance seen on page 61, and let's call $\delta_{i,j}$ the distance between the points i and j). The goal of MDS is to find a set of n vectors $\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_n} \in M$ so that

$$\min_{M} \left(\sum_{i<j} (\|\mathbf{p_i} - \mathbf{p_j}\| - \delta_{i,j})^2 \right) = \min(\varphi^2) \tag{5.26}$$

[16] These MDS maps are usually 2D or 3D for obvious reasons.

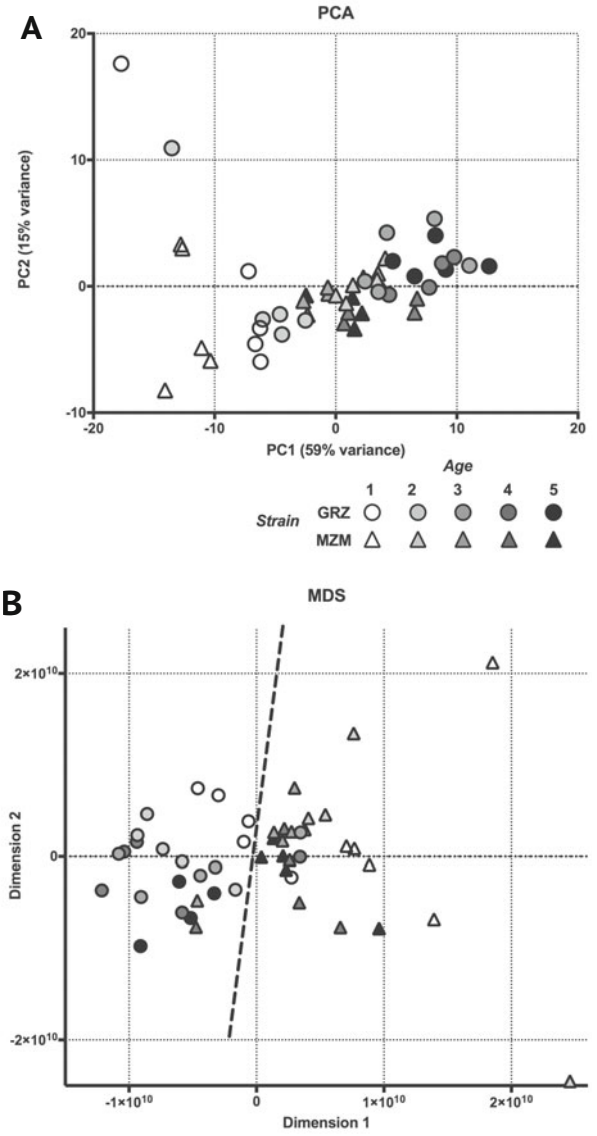


Figure 5.12. Principal Component Analysis (A) and Multi Dimensional Scaling (B) of an RNA-seq dataset including RNA samples from five different ages and two different killifish strains. Both techniques are able to discriminate the different ages within the reduced dimensions, however the MDS is also able to distinguish the samples from the two different strains (dashed line). *Images and analysis courtesy of Mariateresa Mazzetto*.

where $\|\cdot\|$ is the vector norm and φ is called **stress function**. The stress function can usually be minimised by numerical methods.

5.5. Nonlinear multidimensional mapping

T-Distributed Stochastic Neighbour Embedding (t-SNE). is a more advanced method of dimensionality reduction, which has proven to be particularly suitable for the analysis of **single-cell RNA-seq** to cluster large numbers of samples (*i.e.* cells). Since it easy to forecast that single-cell RNA-seq will become a ubiquitous technology in neuroscience, here we dwell in some details of this method.

The rationale of t-SNE is the same as that of MDS. That is, to create a low-dimensional, 2D or 3D, map in which each point is associated with a gene expression profile (for example of a brain region, of a cell line, etcetera). The distance between two points is related to the difference in their expression profiles. However, the way to compute the distance matrix between the gene profiles is not the minimisation of a cost function such as Equation (5.26), which includes a *similarity function* based on gene expression, but of a new cost function which is nonlinear and tempered by a **Gaussian kernel**. The aforementioned Gaussian kernel is a normal probability density function. Given a pair of gene distributions, the joint probability for the couple is higher if they are similar and lower if they are different:

$$p_{ij} = \frac{e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}}{\sum_{k \neq l} e^{-\frac{\|\mathbf{x}_k - \mathbf{x}_l\|^2}{2\sigma^2}}} \tag{5.27}$$

where $\mathbf{x}_i$ is the N-dimensional vector of the gene expression profile (for N genes) of the set $i \in \mathcal{D}$ and with a σ chosen in order to get $\sum p = 1$.

The aim of the Multidimensional Mapping is to get a *map*

$$\mathcal{M} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M\} \tag{5.28}$$

where M is the number of the gene profile sets, $\mathbf{y}_i \in \mathbb{R}^d$ is the *projection vector* of the dataset $\mathbf{x}_i$ on the d-dimensional map $\mathcal{M}$. Obviously, in order to create a map that can be visualised, the dataset should be projected on 2D or 3D vectors. Similarly to what we see in Equation (5.27), in order to measure the similarity between two vectors in the map[17] we can use the value of a probability distribution computed on the norm of the difference between a given couple (i,j) of sets. In t-SNE the distribution chosen in order to compute the similarity between the map vectors is a **Student**

[17] It must be emphasised here that each vector in $\mathcal{M}$ has a correspondent gene profile set in the starting dataset $\mathcal{D}$, so when $\mathcal{M}$ is a good projection of $\mathcal{D}$ given two sets i and j, if $\|\mathbf{x}_i - \mathbf{x}_j\| \sim 0$ then also $\|\mathbf{y}_i - \mathbf{y}_j\| \sim 0$.

t-distribution[18]

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|\mathbf{y}_k - \mathbf{y}_j l\|^2)^{-1}} \tag{5.29}$$

which is derived from a renormalised Student t-distribution density of probability function with a single degree of freedom (when a map $\mathcal{M}$ with $d = 2$ is computed). At this point, we need to define a **cost function** whose minimisation[19] will provide the value of the vectors $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M\}$, which would take into account

- the similarity measure for the sets in $\mathcal{D}$, which is distributed as a Gaussian function (Equation (5.27));
- the similarity measure for the vectors in $\mathcal{M}$, which is distributed as a t-distribution function (Equation (5.29)).

The cost function that will be chosen is the **Kullback-Leibler divergence**,. However, in order to understand the rationale underneath this choice, it would be better to introduce a few concepts of probability and information theory.

A **prior** probability distribution is the the function associated with an event before some evidence (*e.g.*, experimental data) is taken into account. It is apparent that Equation (5.29) describes a prior model of the system: a map is initialised and its probability density is assumed to be Student t-distributed without taking into account the experimental data. A **posterior** probability distribution is the conditional probability function that is assigned to an event after taking into account the relevant experimental data. Also in this case it is straightforward to associate the Gaussian kernel described in Equation (5.27) to a posterior distribution.

Now, the Kullback-Leibler divergence (also called *discrimination information*) is an indicator of the divergence between two probability distributions P and Q, and can be used as a test of difference between the prior (Q) and the posterior (P) probability distribution:

$$D_{KL}(P \| Q) = \sum_{i \neq j} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right). \tag{5.30}$$

[18] The choice of the t-Student distribution takes into account the fact that the number of degrees of freedom from the original dataset $\mathcal{D}$ (usually composed by many hundreds of sets, and possibly normally distributed) collapses to just 2 in case of a 2D map $\mathcal{M}$.

[19] The minimisation will be accomplished through numerical methods such as the *gradient descent* algorithm.

In a way, the value of the divergence describes how different the two distributions are: if $p_{ij} = q_{ij}$, then $D_{KL}(P\|Q) = 0$. So, a successful minimisation of Equation (5.30) through an iterative method would find a set of projections $\mathbf{y}_i$ which produce the probability distribution closest to the one determined by the dataset.

As it can be seen from Figure 5.13, tSNE is able to cluster related samples in a more robust way compared to PCA or MDS. This property makes this analysis tool particularly suited to analyse groups of samples whose

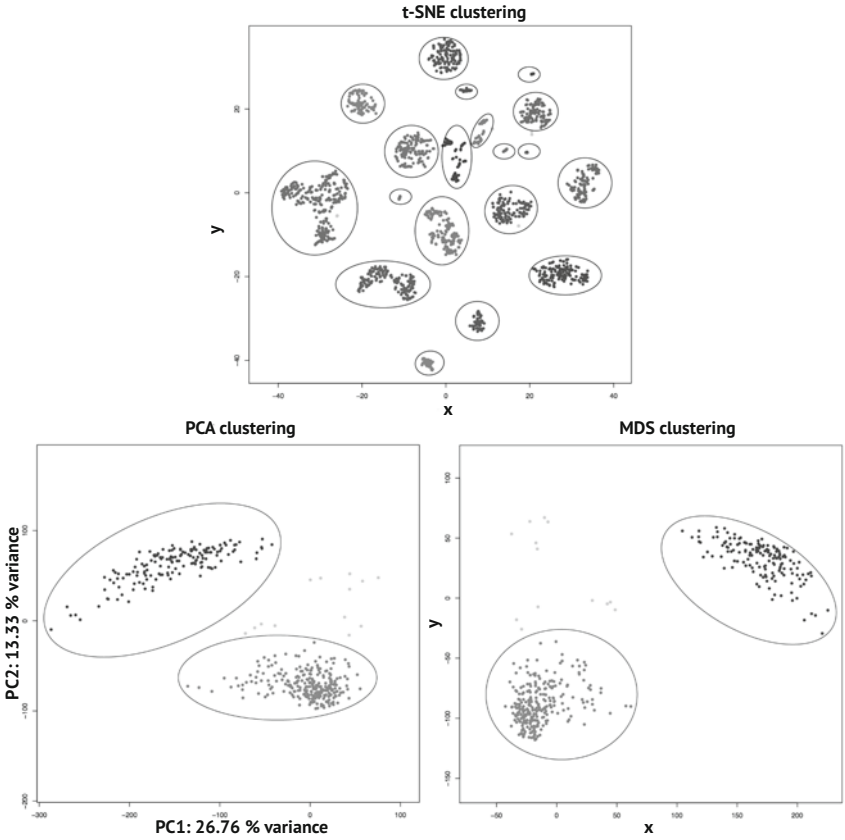


Figure 5.13. The tSNE analysis makes it possible to cluster the transcriptomes derived from a variety of human tissues with high precision [55]. Each cluster (found using the algorithm DBSCAN [15] and indicated with ellipses) corresponds to a different tissue, *e.g.*, brain, whole blood, muscle, and so on. As a comparison, neither PCA nor classic MDS are able to cluster the transcriptomes in a tissue-specific manner. *Figure courtesy of Martino Ugolini, dataset derived from* [55].

biological variation could be subtle, such as subpopulations of a specific cell type, or regions of a given tissue (such as the different brain cortical regions).

5.6. Self-organising maps

The Self-Organising Maps (SOM) are a particular **topology-oriented neural network**, which is formed by a 2D grid of interconnected artificial neurons whose aim is to map distinct features in the dataset [28]. In our case, each neuron is a centroid, and the features to find are the clusters which are recognisable in the data. SOM can be seen as another form of 2D clustering, but it is not applied often to gene expression data.

Chapter 6
Knowledge-based clustering methods

6.1. Introduction

In the previous chapter, we became familiar with unbiased methods for reducing the complexity of RNA-seq data. These methods cluster genes into groups based on co-expression patterns without taking any advantage of prior knowledge on gene function.

A complementary method for dimensionality reduction clusters genes into **gene sets** that are defined by *a priori* knowledge of gene function. This method is widely used to test whether some gene sets are **overrepresented** in a list of DEGs and can be seen as an obligatory step in the analysis of differential expression, but it is also possible to directly test whether a gene set (rather that individual genes) is up- or down-regulated according to a given condition (**gene set enrichment analysis**). The analysis of pathways and gene ontology is very often of great help in the attempt to make biological sense out of a long list of DEGs and, therefore, will be treated here in some detail.

6.2. Testing for gene set overrepresentation

The problem of testing the statistical significance (p-value) of the overrepresentation of a gene set can be illustrated with the following example. We have detected N down-regulated DEGs according to an experimental condition (*e.g.*, Alzheimer's disease brain sample vs. control) and n of these genes are part of a gene set of interest called G, for example the list of genes coding for post-synaptic density. The set of all the protein-coding genes detected in the experiment is called B (for background[1]). Gene set G contains s genes, while gene set B contains S genes. We need to calculate the probability of observing a ratio of n/N DEGs of a given

[1] The issue of the correct background gene set is often not given the due attention. In many cases, the entire set of protein-coding genes in the genome is used as a background.

set if the expected ratio is s/S if there is no enrichment[2]. This is the probability of observing n genes from G in a set of N genes randomly extracted from B and it is equivalent to the well-described 'urn problem': *we have an urn containing N marbles, K black and $N - K$ white. If we draw n marbles from the urn, what is the probability that k are black?* The problem is similar to the random sampling seen in Section 4.3.2 (underlying Poisson distribution). However—due to the fact that we are not replacing the extracted marbles—the situation is modelled by the **hypergeometric distribution**

$$p(k|n, K, N) = \frac{\binom{K}{n}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{6.1}$$

while the ratio between the *observed frequency* (n/k) and the *expected frequency* (N/K) is called **enrichment score**

$$E = \frac{\frac{n}{k}}{\frac{N}{K}}. \tag{6.2}$$

In the example, if p is smaller than the significance threshold α (see Section 4.3.3), then we can state that there is a statistically significant overrepresentation of post-synaptic genes in the genes down-regulated during Alzheimer's disease. It is easy to prove that the sensitivity of the test is reduced when either n or K are small, *i.e.* the smaller the gene set (*e.g.*, synaptic proteins vs. GABAergic receptors) and the smaller the number of DEGs, the larger the enrichment score E required to reach significance. So, when the number of DEGs is small, one may relax the significance (*i.e.* increase the False Discovery Rate—see Section 4.3.3) in order to improve the sensitivity of the test.

In our specific case, we may be tempted to say that postsynaptic genes *are downregulated* in the progression of Alzheimer's disease. This claim is not entirely correct: what we observed is an overrepresentation of this gene set in the DEGs; however, the expression value of the gene set averaged across all genes may not be different. The gene set overrepresentation test should be applied only to binary gene lists[3], while the differential expression of multiple gene sets can be directly tested by the Gene Set Enrichment Analysis (GSEA, see Section 6.5).

[2] In the case of the post-synaptic density example, s=187 (from the Gene Ontology entry 0014069 for *Homo sapiens*), while a tipical value of S is in the order of 10^4, so the expected ratio is of $\sim 10^-2$.

[3] That is, lists compiled using a yes/no criterion such as genes deleted or duplicated in the genome.

6.3. KEGG pathways

A natural way of clustering genes is grouping them based on their involvement in a given pathway (TP53 pathway, cell cycle, insulin pathway, etcetera). This approach is adopted by the *Kyoto Encyclopaedia of Genes and Genomes* (**KEGG**) [24, 25, 23], which makes it possible to visualise up- and down-regulated genes within a pathway through French-flag colouring (Figure 6.1).

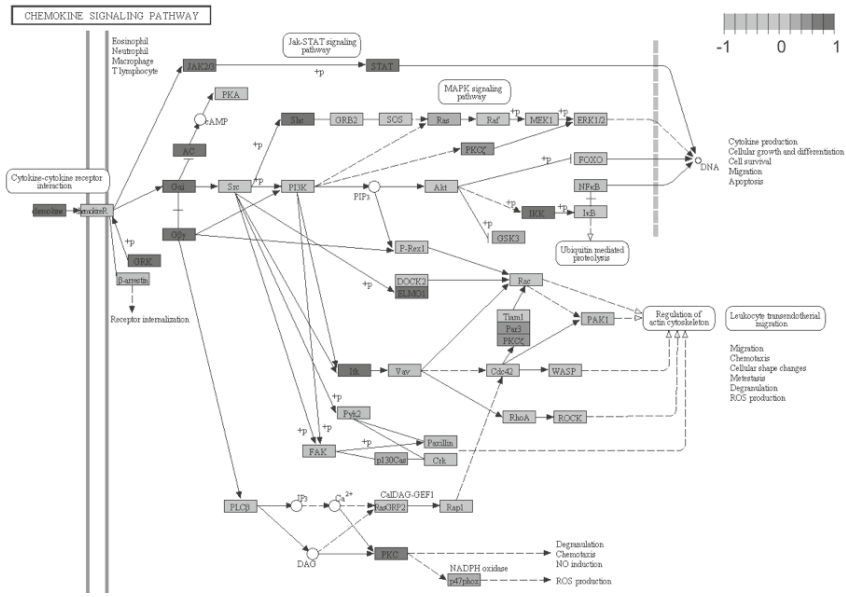


Figure 6.1. Example of KEGG pathway where each gene is coloured according to its value of differential expression. Reproduced with permission from the KEGG database.

It should be noted that a KEGG pathway can contain genes with opposite actions (*e.g.*, inducers or inhibitors of a target protein), as exemplified in the Notch signalling pathway (Figure 6.2). The Notch signalling pathway is highly conserved and is of fundamental importance to organismal development. The KEGG map contains Notch activators (the *γ-Secretase complex*, Delta, TACE) and inhibitors (Numb, Serrate). Therefore, an overrepresentation of the Notch pathway may not necessarily mean that the activity of Notch is enhanced. The representation of the pathway includes information on the cell localisation of the processes. Notch and Delta are plasma membrane proteins and mediate intercellular communication, the NICD (Notch Intracellular Domain, cleaved by

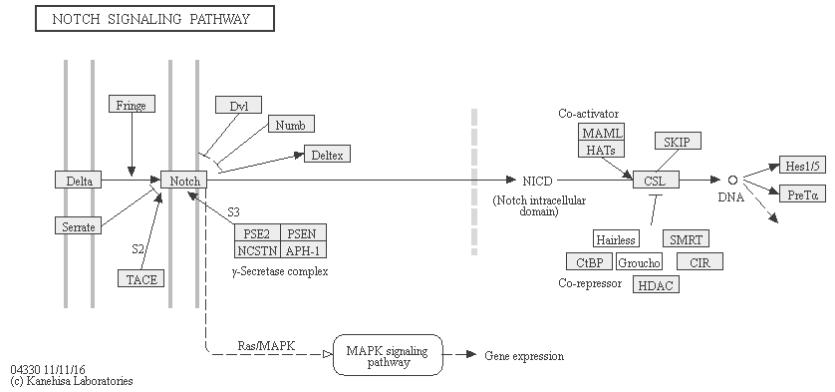


Figure 6.2. KEGG pathway of Notch: the cleaving of Notch protein often involves interaction with a membrane receptor of a different cell and exerts gene expression regulation in the nucleus. Reproduced with permission from the KEGG database.

the *γ-Secretase complex*) is a transcriptional regulator and acts in the nucleus.

In the KEGG database, each gene is *annotated* with one or more **unique pathway identifiers**. For example, the following list of annotations is associated with the gene coding for PSEN1 (Presenilin1, whose mutations are linked to familiar Alzheimer's disease):

```
04310 Wnt signaling pathway
04330 Notch signaling pathway
04722 Neurotrophin signaling pathway
05010 Alzheimer's disease.
```

This example illustrates one of the major difficulties in gene set analysis: *redundancy*. Since gene sets can be partially overlapping, the analysis of a gene list often results in the detection of a number of pathways that are to some extent related because they include the same DEGs. This issue becomes substantially more serious in the analysis of Gene Ontology (GO) terms (see below).

KEGG pathway analysis often offers a compact description of complex transcriptional regulation. The main disadvantage of using KEGG as a reference is that *the majority of genes are not mapped to any KEGG pathway* (*i.e.* do not have an annotation in KEGG), so only a minor proportion of DEGs can be analysed using this approach.

6.4. Gene ontology

Biological processes and structures are often organised in a hierarchical fashion:

> Neuron > process > synapse > postsynaptic density > neurotransmitter receptor > glutamatergic receptor > metabotropic glutamatergic receptor.

The aim of the Gene Ontology (GO) Project[4] is to develop a unified and *controlled vocabulary* of terms for describing gene product properties. The concept of 'ontology' has been borrowed from philosophy in order to define a set of terms and logical connections between terms associated with the entities of a certain domain of discourse. The ontology describes the formally acceptable structure and the relationships between its generating entities—in our case, the genes in a genome and/or the proteins they code for.

Notice that the previously given definition of ontology reminds us of the one given for a graph (see page 101). Indeed, the GO Project organises the annotation data according to a **taxonomy** of annotation classes, thus resulting in a **direct acyclic graph**[5], where the direction of the edges is usually the one from the leaves to the root of the graph. The GO graph imposes a *loose hierarchy* on the annotations where the child nodes (closer to the leaves) are more specialised than the parent ones (closer to the root), but where a single node can have more parents[6] and can have different logical relations with them. In Figure 6.3, we show an example of how the GO annotation is applied to the term 'centriole'. The GO annotation has three independent roots, called **domains**, which represent three different aspects of the biological properties associated with a gene product:

Cellular component terms are related to the localisation of the gene product in the topography of the cell, both at the anatomical level (*e.g.*, 'mitochondrion') or in macromolecular assemblies, such as 'ribosome' or 'proteasome';

[4] http://geneontology.org/

[5] This means that the relationships between the nodes of the graph have a direction (see also page 101) and that there is no directed cycle (*i.e.* there is no circuit along the allowed directions—see note 2 of Chapter 7). That is, the nodes are organised in a hierarchy where there are no connections from the 'upper' levels to the 'lower' ones.

[6] In a strict hierarchy there is only one 'upper' parent node. For example consider the Linnean taxonomy, where a *species* has only one *genre* (higher-level node), while a genre can have many species.

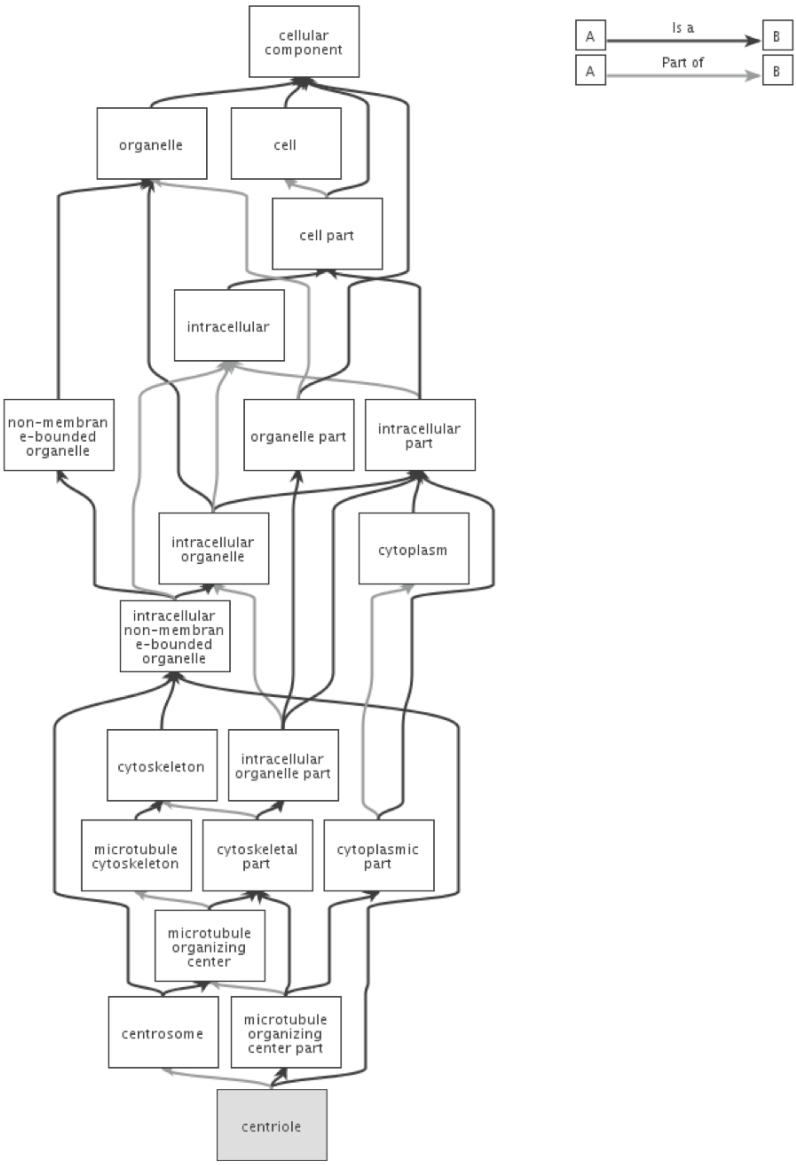


Figure 6.3. GO annotations of the term 'centriole'. The image was produced with QuickGO (http://www.ebi.ac.uk/QuickGO/).

Molecular function terms describe activities that occur at the molecular level, regardless of the performing agents (single proteins or complexes), and of the space, the time or the context where/when the activity takes place (*e.g.*, 'catalytic activity' or 'Toll receptor binding');

Biological process terms involve *processes* and *pathways*, that is an organised sequence of one or more assemblies of molecular functions (*e.g.*, 'signal transduction', or 'calcium-dependent cell-matrix adhesion').

Each GO term has a unique zero-padded seven digit identifier, called the accession term number. The organisation of the graph from the leaves to the root relies on a series of **logic relations** between child nodes and the parent ones. The majority of logical operations are grouped under the categories '`is a`', '`part of`', '`has part`', and '`regulates`'. Each term has a set of *direct relations* towards the parent nodes (or the child nodes), and of *inferred relations* towards non-adjacent nodes[7].

The `A is a B` relation implies that *A is a subtype of B* (for example: 'signal transduction' is a 'biological process', that is the signal transduction is a type of biological process).

The `A is part of B` relation implies that *A is necessarily part of B*: if A exists, then B also exists, but the converse is not necessarily true. For example 'nucleus' is part of 'cell', so if we have a nucleus, then we have also a cell, but the contrary is not valid—indeed, a cell could be devoid of nucleus (*e.g.*, the prokaryotes).

The `A has part B` is a relation *from parent node to child node* (while `is a` and `part of` are from child to parent) and is logically equivalent to `part of`: if A exists, then also B exists, but the contrary is not necessarily true. For example: the 'nucleus' has as a part 'chromosomes', because every nucleus contains at least one chromosome, but there are chromosomes that are not nuclear (*e.g.*, the bacterial chromosome).

The `A regulates B` relation is used in the graph of *biological process*, and has three possible components:

- `positively regulates`, for example when A induces B;
- `negatively regulates`, for instance when A inhibits B;
- `regulates`, when some members of A positively regulate B, while other members of A negatively regulate B.

Figure 6.4 shows maps of relations between terms—according to the GO definitions—and a few examples of the kind of relations that can be inferred.

Another important aspect of GO is that the level of evidence for a given annotation is reported by a three-letter code:

[7] An inferred relation applies the properties of the defined relations. For example: the relation `is a` have the transitive property, so given 'A is a B' and 'B is a C' it can be inferred that 'A is a C'.

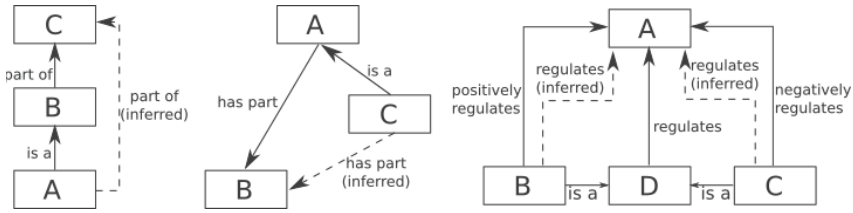


Figure 6.4. Examples of Gene Ontology relations and inferrable relations.

Experimental evidence codes
EXP Inferred from Experiment
IDA Inferred from Direct Assay
IPI Inferred from Physical Interaction
IMP Inferred from Mutant Phenotype
IGI Inferred from Genetic Interaction
IEP Inferred from Expression Pattern
Computational Analysis evidence codes
ISS Inferred from Sequence or Structural Similarity
ISO Inferred from Sequence Orthology
ISA Inferred from Sequence Alignment
ISM Inferred from Sequence Model
IGC Inferred from Genomic Context
IBA Inferred from Biological aspect of Ancestor
IBD Inferred from Biological aspect of Descendant
IKR Inferred from Key Residues
IRD Inferred from Rapid Divergence
RCA Inferred from Reviewed Computational Analysis
Author Statement evidence codes
TAS Traceable Author Statement
NAS Non-traceable Author Statement
Curatorial statement codes
IC Inferred by Curator
ND No biological Data available
Automatically-Assigned evidence code (the only class of codes not assigned by a GO curator)
IEA Inferred from Electronic Annotation

Since the GO Project has less stringent criteria for gene annotations than KEGG, more genes have an annotation in GO than in KEGG. Also, a gene has normally (many) more GO annotations as compared to KEGG ones. As an example, Table 6.1 shows the GO annotations of PSEN1 (whose KEGG annotation has been shown at page 88).

Molecular Function	**Evidence Code**		
PDZ domain binding	IPI	calcium channel activity	IMP
aspartic endopeptidase activity, intramembrane cleaving	IDA	endopeptidase activity	IDA
		protein binding	IPI
beta-catenin binding	IPI	cadherin binding	IBA
Biological process	**Evidence code**		
Cajal-Retzius cell differentiation	IEA	L-glutamate transport	IEA
Notch receptor processing	TAS	Notch signaling pathway	IEA
T cell activation involved in immune response	IEA	T cell receptor signaling pathway	IEA
activation of MAPKK activity	IEA	amyloid precursor protein catabolic process	TAS
amyloid precursor protein metabolic process	IDA	autophagosome assembly	IEA
beta-amyloid formation	IMP	blood vessel development	IEA
brain morphogenesis	IEA	Ca^{2+} ion transmembrane transport	IEA
canonical Wnt signaling pathway	IBA	cell fate specification	IEA
cellular response to DNA damage stimulus	IEA	cerebral cortex cell migration	IEA
choline transport	IEA	dorsal/ventral neural tube patterning	IEA
embryonic limb morphogenesis	IEA	endoplasmic reticulum Ca^{2+} ion homeostasis	IDA
endoplasmic reticulum Ca^{2+} ion homeostasis	IGI	epithelial cell proliferation	IEA
heart looping	IEA	*45 more annotations*	[· · ·]
Cellular component	**Evidence code**		
Golgi apparatus	IDA	Golgi membrane	IEA
aggresome	IDA	axon	IEA
azurophil granule membrane	TAS	cell cortex	IEA
cell junction	IDA	cell surface	IEA
centrosome	IDA	ciliary rootlet	IEA
dendritic shaft	IEA	endoplasmic reticulum	IDA
endoplasmic reticulum membrane	IEA	gamma-secretase complex	IDA
growth cone	IEA	integral component of membrane	IDA
integral component of membrane	TAS	integral component of plasma membrane	IDA
kinetochore	IDA	membrane	IDA
membrane raft	IDA	mitochondrial inner membrane	IEA
mitochondrion	IDA	neuromuscular junction	IEA
neuronal cell body	IEA	nuclear membrane	IDA
nuclear outer membrane	IDA	nucleus	IDA
plasma membrane	IDA	plasma membrane	TAS
presynapse	IEA	rough endoplasmic reticulum	IDA
smooth endoplasmic reticulum	IDA		

Table 6.1. GO annotation of the Presenilin-1 (PSEN1) gene.

Notice that the redundancy of GO annotations is considerably higher than for KEGG annotations. A second major difference between GO and KEGG annotations is that, due to parent-child relationships, GO has *multiple levels*. Therefore, in addition to an 'horizontal redundancy' (with almost-synonym terms at the same hierarchical level), there can be a 'vertical redundancy' where a term as well as some of its parent terms are overrepresented in a DEG list.

In practice, an analysis of GO terms associated with a gene list can easily result in a long list of terms that are more or less related. The major drawback of the redundancy of GO terms is that it makes it difficult to have a compact description of the biological processes under study, thus important aspects may be overlooked. There are at least two possible solutions to **reduce the redundancy** of a list of GO terms:

1. to *exclude all parent terms* of an enriched child term:
2. to *cluster related gene ontology terms* (several packages are available to perform this task, see Appendix 6.1 for one example).

6.5. Gene set enrichment analysis

The aim of *gene set enrichment analysis* (GSEA) is to provide a statistical framework to test whether a set of genes is collectively up- or down-regulated according to a given experimental condition. This is particularly relevant since small but consistent changes in expression for the majority of genes within a pathway may result in a clear biological effect that could not be detected by DEG analysis. Here we present two approaches to GSEA that use different statistical frameworks.

6.5.1. Non-parametric GSEA for multiple samples

The procedure described in this Section follows the seminal paper from Aravind Subramanian, Pablo Tamayo and colleagues [52].

Let's consider two sets: a list **B** of expressed genes and a set **S** of genes associated in a knowledge-based way to a feature of interest. Each gene is associated with an expression value in samples that differ for a biological condition. The aim of GSEA is to compute an **enrichment score** Ξ for the set **S** based on the expression values of its members.

A common procedure is to **rank-order** all expressed genes according to ϱ, a metric for the correlation of their expression with the biological variable of interest (fold-change, signed p-value, Spearman's correlation coefficient, and so on). Each gene in S will have a position in this ranked list. The aim of GSEA is to provide a statistical framework to test whether members of S occur more often than expected at the top or bottom of the rank-ordered list.

One can evaluate, for every position i in the list, the fraction of genes in **S** weighted according to ϱ ('hits', H) and the fraction of genes not in **S** ('misses', M). We define:

$$H(S,i) = \sum_{g_k \in S,\, j \le i} \frac{|\varrho|^p}{\sum_{g_k \in S} |\varrho|^p} \tag{6.3}$$

$$M(S,i) = \sum_{g_k \notin S,\, k \le i} \frac{1}{\#B - \#S} \tag{6.4}$$

$$\Xi(S,i) = H(S,i) - M(S,i) \tag{6.5}$$

where p is a coefficient[8] that determines the skewness of the weight distribution towards high absolute values of ϱ_j, $\#B$ and $\#S$ are the numerosi-

[8] If $p = 0$, the enrichment score becomes the standard *Kolmogorov–Smirnov statistic*, which is used in the Kolmogorov-Smirnov test in order to assess the goodness of fit to a dataset. In [52] $p = 1$ so the score is weighted using the feature-dependent correlation rank of the genes in list B.

ties of the sets (in Equation (6.4) it is assumed that $\#B \gg \#S$). Given the $\Xi(S, i)$ distribution, the enrichment score associated with a set S is:

$$\Xi(S) = \max_i (\Xi(S, i)). \tag{6.6}$$

The Ξ score, as calculated in Equation (6.5), has high values if the genes in the label set S are densely positioned at either end of the ranking of B, while it has low value in case of a random distribution (Figure 6.5).

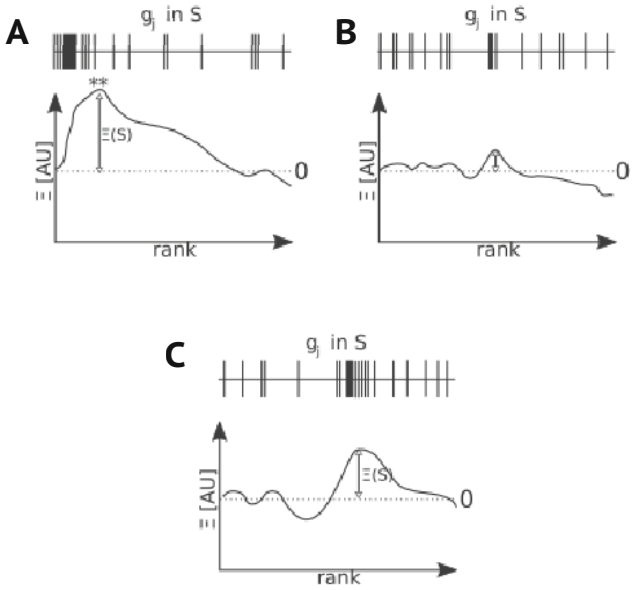


Figure 6.5. Qualitative behavior of the enrichment score distribution $\Xi(S, i)$ according to the different ranking in **B** of the genes in the label set **S**: A) if the genes are in the top ranking of **B**, the enrichment score will be high and associated with a low *p-value*, B) if the genes are distributed randomly, the Ξ score will be low, C) if the genes are distributed in a nonrandom way the enrichment score will be high, but the p-value would be high (null hypothesis failed to be rejected).

In order to evaluate the significance of an enrichment score, a **permutation test** is performed. This type of test is very widely used in genomics and provides a p-value without any assumption on the distribution of the variables. It uses a *random reassignment of phenotype labels*. This means that a pseudoreplica of the experiment is generated by creating one set that is the union of all samples and randomly extracting the pseudocases and pseudocontrols from this set. This procedure is equivalent to reshuffling the columns in the data matrix and recalculating the $\Xi^t(S)$. The

calculation of $\Xi^t(S)$ for many different permutations of the label assignment leads to an enrichment scores distribution $\Xi_0(S)$ for the null hypothesis 'the S set is *not* enriched in the sample', which can be used to evaluate the p-value of the enrichment for the label set S. If the statistical significance evaluation is made among more than two phenotypes, multiple hypothesis testing through evaluation of a *False Discovery Rate* (see Section 4.3.3 at page 55) is usually performed.

6.5.2. Generally applicable gene enrichment

In this case, we use an approach called meta-analysis, *i.e.* a summary of a number of different statistical tests[9]. Let's assume that we have a dataset composed of n controls and m cases. If we have a gene set **S**, a background set **B**, and two individual samples, one case and one control, we can compute the mean (μ_S) and variance (σ_S^2) of the fold changes for all genes in **S** between the two samples and compare these with mean (μ_B) and variance (σ_B^2) of the fold-changes for all genes in **B** between the two samples. It is now possible to use the conventional two-sample Student's t-test (or a non-parametric two-sample test) to calculate the p-value for $H_0 : \mu_s = \mu_B$. This procedure can be repeated for all possible pair-wise comparisons between a specific case (j) and n controls to calculate the mean of the log (p-value)

$$x_j = -\frac{1}{n} \sum_{i=1}^{n} \log (p_{ij}) \tag{6.7}$$

where p_{ij} is the p-value of the pair-wise test between the sample j and the i-th control sample. This is a summary measure of the deviation of case j from the controls. So, if we consider the sum of x_j values over all m cases,

$$X = \sum_{j=1}^{m} x_j \tag{6.8}$$

we obtain a new variable which is known to follow a Γ distribution (see Section 3.5.1) with parameters m and 1. This allows us to calculate

$$P(y > X) = \Gamma_{m,1}(y > X) \tag{6.9}$$

which is the probability that a random variable y distributed as a $\Gamma(m, 1)$ is larger than X. This probability is the p-value for the enrichment of gene

[9] This is often used in epidemiology when the results of different studies are combined.

set **S**. *False discovery rate* correction (see Section 4.3.3 on page 55) needs to be applied if more than one set is tested at the same time, as it would be the case for analysis of KEGG pathways or GO.

A major advantage of this approach (Generally Applicable Gene Enrichment, or GAGE) [31] is it can be used for paired samples (*e.g.*, leukocytes of a patient before and after a treatment). In this case, it refers to a paired t-test between the two samples.

Appendix 6.1 – GO redundancy reduction: clustering of annotations

In this section we will present a commonly used service for the clustering of Gene Ontology terms: the DAVID Bioinformatics Resources[10]. It is informative to show a specific example from [43]. In this case, a total of 4095 down-regulated DEGs were detected during ageing of the killifish brain. If we analyse the 10 most overrepresented terms among these DEGs (over 824 different terms), there is a high level of redundancy, as all terms are somehow related to either development or cell cycle:

GO Domain	Term	Number hits	%	p-value	Benjamini
Biological Process	cell development	273	9,3	2,9E-19	1,3E-15
Biological Process	neuron development	164	5,6	1,1E-17	2,4E-14
Biological Process	DNA replication	57	1,9	1,4E-17	2,1E-14
Biological Process	neuron projection development	136	4,6	2,0E-17	2,3E-14
Biological Process	nervous system development	306	10,5	3,8E-17	3,5E-14
Biological Process	cell morphogenesis involved in differentiation	123	4,2	3,8E-16	2,5E-13
Biological Process	neurogenesis	214	7,3	6,4E-16	4,4E-13
Biological Process	axon development	112	3,8	6,7E-16	3,8E-13
Biological Process	neuron projection morphogenesis	113	3,9	9,3E-16	4,5E-13
Biological Process	generation of neurons	199	6,8	1,5E-15	7,1E-13

On the other hand, if we perform a clustering of the terms, we obtain 373 clusters. In this case, we show only the most statistically significant term

[10] https://david.ncifcrf.gov/

for each cluster. Novel functions appear that were not included in the first list, such as hydrolase activity (in particular ATP-ases) or microtubule-related terms.

Cluster 1	**Enrich. score**	12.22	**n. terms**	24	
Biol. Proc.	cell development	*p-value:*	$2.9E-19$	*Benjamini:*	$1.3E-15$
Cluster 2	**Enrich. score**	11.91	**n. terms**	3	
Biol. Proc.	DNA replication	*p-value:*	$1.4E-17$	*Benjamini:*	$2.1E-14$
Cluster 3	**Enrich. score**	9.62	**n. terms**	12	
Biol. Proc.	mitotic cell cycle	*p-value:*	$1.6E-15$	*Benjamini:*	$6.9E-13$
Cluster 4	**Enrich. score**	7.86	**n. terms**	6	
Mol. Funct.	hydrolase activity	*p-value:*	$1.1E-10$	*Benjamini:*	$2.0E-7$
Cluster 5	**Enrich. score**	6.94	**n. terms**	3	
Mol. Funct.	Tubulin binding	*p-value:*	$2.7E-8$	*Benjamini:*	$1.0E-5$

Chapter 7
Network analysis

7.1. Introduction

Life is built on **functional interactions**—between molecules, macromolecular complexes, subcellular organells, cells and any other higher-level organisation. If we consider a set of genes and their expression changes across biological conditions, we could be interested to test whether these coordinated changes might suggest **functional interactions** among subsets of genes. The clustering methods we described in Chapter 5 are the standard methods to reveal structures within gene co-expression patterns. Since the early 2000s, **graph theory** has been increasingly applied to biological datasets in order to build **genome-scale networks** such as

- protein-protein interaction network, also called **interactome**,
- **knowledge-based networks**, such as KEGG pathways, which are built upon findings from the scientific literature, or
- **gene co-expression networks**.

In this chapter, we will focus on the applications of graph theory to the study of gene co-expression networks.

As a summary of the tools and the problems addressed in the first part of this book, we will present in detail the work from Baumgart and colleagues [6]. This paper clearly shows how information derived from RNA-seq can be linked to higher-level molecular, cellular and integrated functions. Moreover, it reveals how this experimental approach can identify key regulators of a biological process and how it can be used to study the perturbation that an external factor (*e.g.*, a drug) induces on the global transcription patterns of the organism.

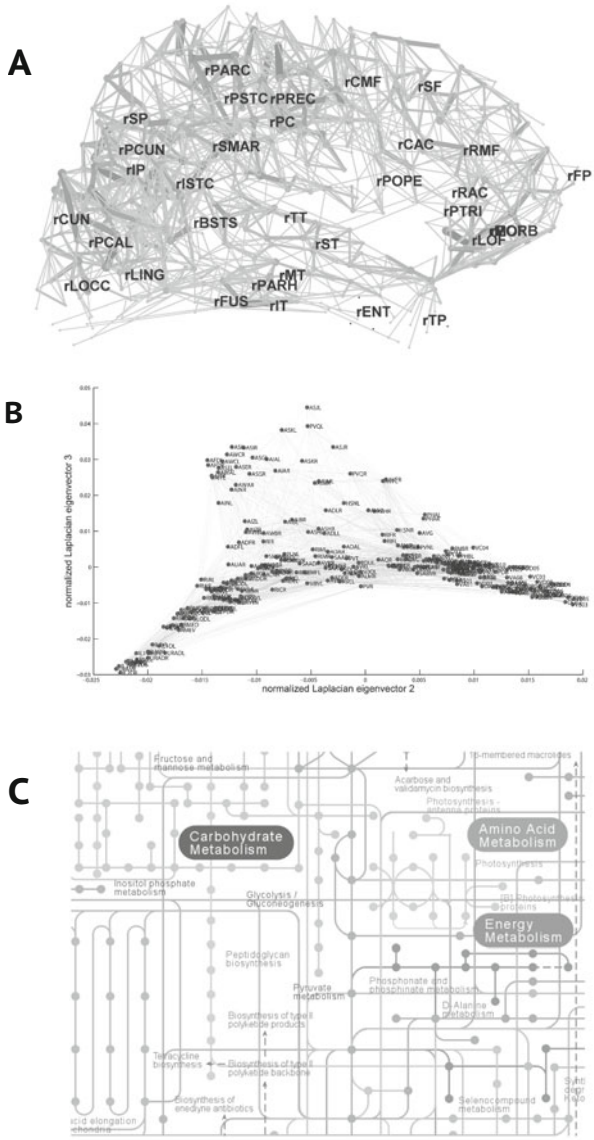


Figure 7.1. Examples of biological networks. A) The human **connectome** is a mesoscale map of connections between brain areas (for example, built from fMRI images), or B) The *C. elegans* connectome is a microscale network where each neuron of the worm is represented as an individual node). C) Portion of the human metabolic network from KEGG pathways. *Panel A adapted from* [18] *under a CC-BY Creative Commons Attribution License. Panel B adapted from* [57] *under a CC-BY Creative Commons Attribution License. Panel C adapted from the KEGG pathway hsa:01100* [24, 25, 23].

7.2. Biological networks

Many biological systems are organised in units (**nodes**) connected by links (**edges**). This is most obvious to neuroscientists that study connections between neurons or—at a higher level—between cortical areas (connectome, Figure 7.1A and B). A very well (almost comprehensively) described network is the metabolic network, where the nodes are metabolites and the edges are enzymatic reactions (Figure 7.1C). Intracellular transduction pathways also can be immediately represented as networks. A problem of high biological relevance is the identification of the 'key nodes' in these networks. In the case of gene networks, these nodes likely correspond to key regulators of the biological process of interest and are top candidates for downstream experimental validation. For these reasons, applications of graph theory to gene-expression data have gained significant momentum in recent years.

7.3. A primer on graph theory

Graph theory is a field of mathematics that deals with objects–or *graphs*–consisting in a series of points, called *nodes*, and connections between nodes, called *edges* (Figure 7.2). So, given a set of genes G and a list of

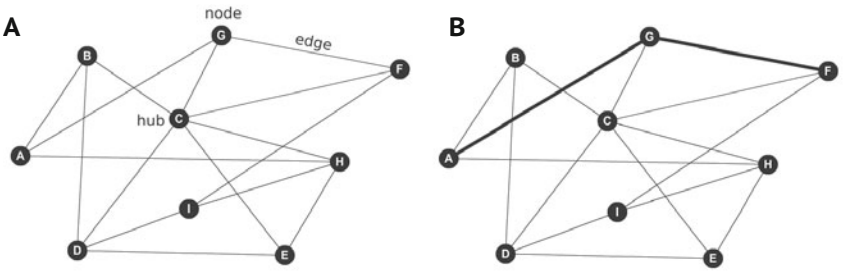


Figure 7.2. A) Basic characteristics of graphs: nodes, edges, hubs. B) Shortest path between node A and node F.

connections between them[1] defined as

$$C(G) = \{v, u \in G \,:\, v \neq u\} \tag{7.1}$$

the associated gene-wide network is the graph $W = (G, C)$. W is called an **undirected** graph if $\{u, v\} = \{v, u\}$, and a **directed** graph otherwise. Connections derived from gene co-expression data are usually

[1] The word is imprecise on purpose: the link between two genes could be a measure of the expression level or a direct interaction between the coded proteins—but this case is more complicated because it would likely need the extension of Equation (7.1) for the case $u = v$—or any other relationship.

undirected, meaning that *we cannot infer the direction/nature of the relationship between two genes directly*. On the other hand, neuronal connectivity patterns and signalling networks unequivocally define a directed network.

Other biologically relevant definitions of graph theory are:

hub: a hub is a node with ***high degree***. It occupies a top position in a list where nodes are ranked based on the number of their edges—a property also called *high connectivity*; in a gene co-regulatory network, this could be the case of key transcription factors or microRNAs;

weight: a weight function w_{uv} associates each edge $\{u, v\}$ with a real value, see for example the case of synaptic strength in a neuronal network. The resulting graph is thus called **weighted graph**;

path: a path from a node A to a node A' is a *walk* between adjacent nodes where each intermediate node is crossed only once. For example, this is the case for the biosynthetic pathway of a neurotransmitter[2];

shortest path: the shortest path from A to A' is the path which crosses the least number of nodes between A and A' (in unweighted graphs) or the path which shows the *least cumulative weight* (in weighted graphs).

In the following Sections we will discuss other important properties of graphs from a more quantitative point of view.

7.3.1. Algebraic graph theory offers some powerful tools to analyse graphs... and biological networks

A graph, such as the one shown in Figure 7.2, can be represented in a matrix form, called an **adjacency matrix**, where every element is defined as

$$a_{i,j} = \begin{cases} \alpha_{i,j} \text{ if } \quad \{i, j \in G, i \neq j\} \in W(G, C) \\ 0 \quad \text{otherwise} \end{cases} \tag{7.2}$$

where $\alpha_{i,j} = 1$ in the case of an unweighted network, or $\alpha_{i,j} = w(i, j)$ if the network is weighted.

[2] The first node can be crossed twice if it is coincident with the end of the path—in this case the path is called a *circuit*.

For example, the undirected graph of Figure 7.2 can be described by the following adjacency matrix:

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
A	0	1	0	0	0	0	1	1	0
B	1	0	1	1	0	0	0	0	0
C	0	1	0	1	1	1	1	1	0
D	0	1	1	0	1	0	0	0	1
E	0	0	1	1	0	0	0	1	0
F	0	0	1	0	0	0	1	0	1
G	1	0	1	0	0	1	0	0	0
H	1	0	1	0	1	0	0	0	1
I	0	0	0	1	0	1	0	1	0

which is a symmetrical $n \times n$ matrix with zero diagonal, where n is the number of the nodes in the graph.

The fact that a graph can be written as its adjacency matrix means that it is possible to explore the graphs' properties through algebraic manipulation. First of all, it is possible to define the **eigenvalues** (λ_e) and the **eigenvectors** ($\mathbf{v}_e$) of the adiacency matrix ($\mathbf{A}$):

$$\mathbf{A} \cdot \mathbf{v}_e = \lambda_e \mathbf{v}_e. \tag{7.3}$$

A powerful property of the eigenvectors of the adjacency matrix is that it is possible to define a **eigenvector centrality** measure, where the values of the all-positive valued eigenvector corresponding to the largest eigenvalue correspond to a measure of the importance of the node inside the graph[3]. For example, a numerical approximation of the eigenvectors of the shown adjacent matrix associated with the graph of Figure 7.2 gives

$$\lambda^*_{\max} = 3.8226 \tag{7.4}$$

[3] The definition of eigenvector centrality is quite self-referential: a node is important when it is linked to important nodes. It can be demonstrated that eigenvector centrality is a measure of how probable is that a random walk would pass through the node (*Gould index of accessibility*).

and so the associated eigenvector (normalised to the maximum element) is

$$\hat{v}^*_{\max} = \begin{pmatrix} 0.2464 \\ 0.2979 \\ 0.5053 \\ 0.3709 \\ 0.3274 \\ 0.2774 \\ 0.2731 \\ 0.3572 \\ 0.2669 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \end{matrix} . \tag{7.5}$$

So the node with highest rank is C, followed by D, H, E, B, F, G, I, and A has the lowest rank. Indeed, it can be seen that C is the hub of the graph of Figure 7.2, with six neighbours. Both D and H have four neighbours, however H is connected with low-rank nodes (A and I, which are not connected to the hub), so D has a higher eigenvector centrality. A similar reasoning would explain the other centrality scores.

7.3.2. Topological properties of a graph

In this Section, we will discuss about other important metrics aimed at describing 1) the general structure of the network, 2) the rank of a given node in the network, 3) the interconnectivity between the nodes of the network.

Neighbourhood and degree distribution. Two nodes i and j are adjacent nodes if there is an edge $\{i, j\} \in C(G)$. The **neighbourhood** of a node is the set of its adjacent nodes (also called *neighbours*). The numerosity of the neighbourhood is the **degree**[4] of the associated node. In a real-world analogy, the degree of a social network profile is defined by the number of its connections (friends, followers, and so on). As previously mentioned, the nodes with highest degree are called *hubs* of the network. It is also possible to define a **degree distribution** $P(k)$ as the ratio between the number of nodes with a given value of degree k and the total number of nodes:

$$P(k) = \frac{\#G_k}{\#G}. \tag{7.6}$$

The shape of the degree distribution provides key information as to the structure of the network. In a **random network**, each node has a given

[4] The term *connectivity* is used in the literature as a synonym of degree. However, in this book we will be more prone to use 'connectivity' in the context of weighted graphs, while 'degree' will be used for unweighted graphs.

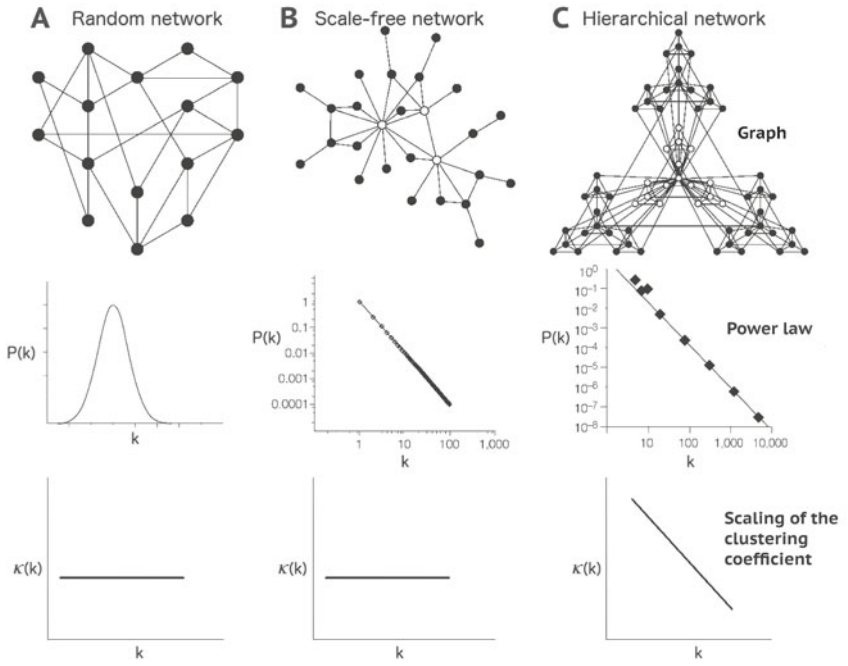


Figure 7.3. Graph, power law, and scaling of the clustering coefficient in A) a random network, B) a scale-free network, and C) a hierarchical scale-free network. The distribution of connectivity of a random network is bell-shaped and implies that individual nodes more frequently have a connectivity close to the average $\langle k \rangle$ of the network. The scaling of its clustering coefficient is flat, because there is no dependence between the degree of a node and its clustering coefficient. A scale-free network is characterised by a scale-free distribution of connectivity (see Equation (7.7)), so, if plotted on a log-log graph, it results in a linear function with slope $-\gamma$ that represents the scaling factor. A hierarchical architecture is observed in some biological systems and involves the presence of modular blocks that are added starting from a central hub. The connectivity distribution follows a power law and so does the distribution of the clustering coefficient—the more connected nodes show smaller clustering coefficients and vice versa. *Adapted by permission from Macmillan Publishers Ltd: Nature Rev. Genet*. [5], © (2004).

probability of being connected with another node according to a master *graph model*: the degree distribution for these networks is often **bell-shaped** (Figure 7.3A).

Remarkably, biological networks are usually *not* distributed as a random graph, but in many instances their degree distribution follows a **power law**

$$P(k) = k^{-\gamma} \tag{7.7}$$

where γ is a power factor determined by the network. The most evident result from the power law distribution of the degrees is that *in a biological network there are few highly-connected nodes and a majority of low-degree nodes*[5]. Networks whose degree distribution follows a power law are called **small-world networks** (or **scale-free** networks). It has been demonstrated that this network structure is an optimal compromise between robustness against random failures, information capability, and cost of the network. A **rich-club small-world network** is a network wherein the highest degree nodes are also strongly interconnected.

Connectivity in a weighted network. A weighted network presents a nonbinary connectivity matrix, so the concept of degree cannot be applied to measure the strength of connection of a node in a weighted network[6]. We can thus define a concept analogous to degree: the **connectivity** of a gene i, which is defined as

$$k_i = \sum_{i \neq j} \alpha_{ij} \tag{7.8}$$

where $\alpha_{i,j}$ is the value associated with the edge $\{i, j\}$ in the weighted adjacency matrix. If we define the ***maximum connectivity*** of the network as the highest value of connectivity of the nodes in the network, we can also define a **scaled connectivity** measure of the gene i

$$K_i = \frac{k_i}{\max_i (k_i)} \tag{7.9}$$

as the ratio between the connectivity of the node and the maximum connectivity of the network.

Measures of centrality of the nodes. In the previous paragraph, we were introduced to the concept of *eigenvalue centrality*. Here we will define other measures which can be used to 'rank' the nodes in a graph. What does it mean for a gene to be 'central'? In the case of eigenvalue centrality, a central gene is a gene that is linked to other central genes (see note 3). If we define a central gene as a gene which is 'close' to the other genes (*i.e.* the average shortest path is as small as possible), we get a

[5] To be precise: real biological networks follow a **truncated power law distribution**, which has value zero for $k > \max_i (k_i)$.

[6] The weight of a connection can be considered as the likelihood of passing from one node to another during a random walk. Now let's think about a node with a large neighbourhood of 'weak' connections and one with a smaller number very strong connections (high probability of being in the walks from the neighbouring nodes): which is the real hub?

closeness centrality measure

$$C_i = \frac{1}{\sum_{j \neq i} s_{ij}} \tag{7.10}$$

where s_{ij} is the length of the shortest path (see Figure 7.2) between nodes i and j. The closeness centrality coefficient (in connected graphs) has values $0 < C_i \leq (\#G)^{-1}$, that is between a node which has few connections with peripheral nodes to a node which has connection with every node in the network. In a real-world analogy, the path is the number of shares a post needs to reach website **B** from website **A**, so a website is central if most of the shares transit through it.

Another possible definition of centrality of a node i is the **betweenness centrality**, which measures the fraction of shortest paths in the network passing through a node i:

$$B_i = \sum_{j \neq i \neq k} \frac{\#s_{jk}(i)}{\#s_{jk}} \tag{7.11}$$

where $\#s_{jk}$ is the total number of shortest paths between the nodes j and k, and $\#s_{jk}(i)$ is the number of shortest paths which pass through the node i. This measure of centrality is particularly important in connectomics because areas with high betweenness centrality connect cortical modules processing multimodal information.

Clustering. Another interesting property of the nodes of a network is the clustering coefficient, which evaluates the **interconnectivity between the neighbours** of a node i

$$\kappa_i = \frac{2n_i}{k_i\,(k_i - 1)} \tag{7.12}$$

where k_i is the degree of gene i, and n_i is the number of links connecting the k_i neighbours of i with other neighbours of i. The clustering coefficient has value 0 in the case of a star-shaped network (none of the neighbours are interconnected) and maximum value when every node of the neighbourhood has a connection with every other node. The distribution of clustering coefficients can provide important information regarding the organisation of the network. If a network has a *modular structure*, *i.e.* it is composed of modular blocks that are added to a central hub (Figure 7.3C), then the relationship between clustering coefficient and connectivity is that of a power law. It is of note that the hierarchical organisation of the metabolic network was demonstrated using this method [42].

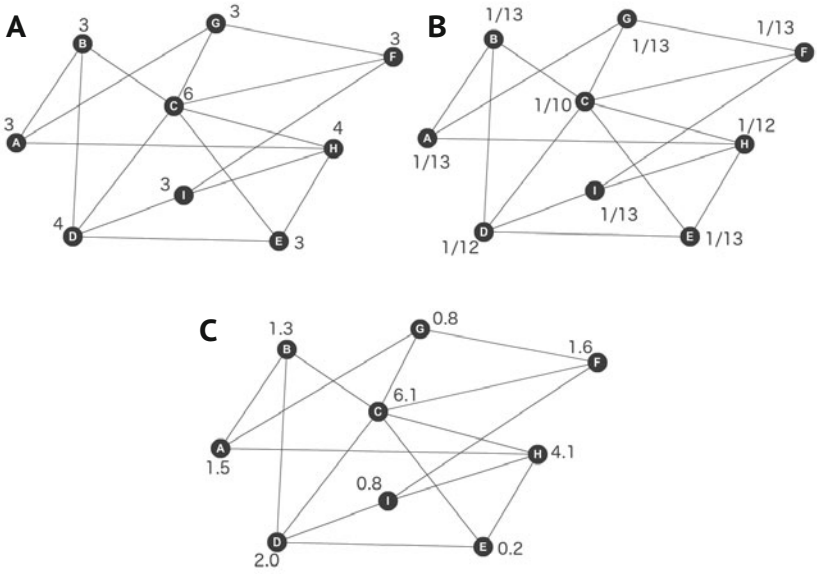


Figure 7.4. Different measures of centrality in the network shown in Figure 7.2. A) Degree centrality, B) closeness centrality, and C) betweenness centrality. Each node of the graph has been associated with the corresponding measure. It can be seen how each centrality measure depicts some part of the properties of the network.

7.4. Weighted gene coexpression network analysis

RNA-seq experiments provide a series of expression levels for n transcripts in m different conditions; that is, a $D_{n\times m}$ dataset matrix.

As already seen in Chapter 5, we can use Pearson's correlation coefficient as a metric for the co-expression between two genes. We then generate a gene co-expression matrix $\mathbf{R}_{n\times n}$ according to the coefficients

$$\rho_{ij} = \frac{\sum_{k=1}^{n}(q_{ik}-\mu_i)(q_{jk}-\mu_j)}{\sqrt{\sum_{k=1}^{n}(q_{ik}-\mu_i)^2}\sqrt{\sum_{k=1}^{n}(q_{jk}-\mu_j)^2}} \tag{7.13}$$

where q_{ik} is the expression of the gene i in sample k and q_{jk} is the expression of the gene j in sample k; μ_i and μ_j are the means across samples of the genes i and j respectively. After the construction of $\mathbf{R}$, the next step is to determine the adjacency matrix of the network $A_{n\times n}$. There are two approaches:

- a **hard thresholding**, that constructs an unweighted matrix

$$a_{i,j} = \begin{cases} 1 \text{ if } \quad \rho_{ij} \geq \vartheta \\ 0 \text{ otherwise} \end{cases} \tag{7.14}$$

where ϑ is a threshold value which determines the connections between genes to be discarded;

- a **soft thresholding**, where the value of the adjacency matrix is a function of the correlation coefficient

$$a_{ij} = a(\rho_{ij}) \tag{7.15}$$

and thus determines the construction of a weighted network.

Choosing to build an unweighted graph has some important drawbacks: because every edge whose correlation value falls below ϑ is considered *noise*, but a fraction of the excluded edges might carry biological information. Soft thresholding keeps all the information from the RNA-seq dataset (noise included), so a good practice is to emphasise the high-correlation edges and to 'punish' the weaker correlations—though without annihilating them. This can be achieved by applying an exponential function to the correlation coefficients

$$a_{ij} = \rho_{ij}^{\beta} \tag{7.16}$$

where the value of $\beta \geq 1$ can be found using the **scale-free topology criterion** (found in [61]) or set to $\beta = 6$, again following [61]. The scale-free topology criterion states that the index β should be chosen in a way that the obtained network has a scale-free degree distribution[7]. After having computed the adjacency matrix with a soft threshold we have obtained a *weighted gene co-expression network*.

7.4.1. Module analysis

A powerful way to start the downstream analysis of the weighted network is to divide the network into **modules**. That is, into *groups of genes which share high levels of co-expression*. This adjacency matrix is, apart from the power coefficient β, perfectly analogous to the correlation matrix seen in Chapter 5 and could be used as such for hierarchical clustering. However, in WGCNA, the clustering is based on the **topological properties** embedded in the adjacency matrix.

Let's now introduce a new measure, called **topological overlap** (ω) between two nodes i and j, which measures how much the neighbourhood

[7] The argument for the scale-freedom of the obtained network could sound a bit circular, however it is reasonable, as the majority of complex networks seen in biology behave like a scale-free network. In practice, a β is chosen, the network is computed, then the log/log plot of $P(n_k)$ as a function of the degree k is plotted. For the network to be accepted, the plotted function should be interpolated by a linear function with correlation coefficient $R^2 \geq 0.8$ and slope roughly equal to -1.

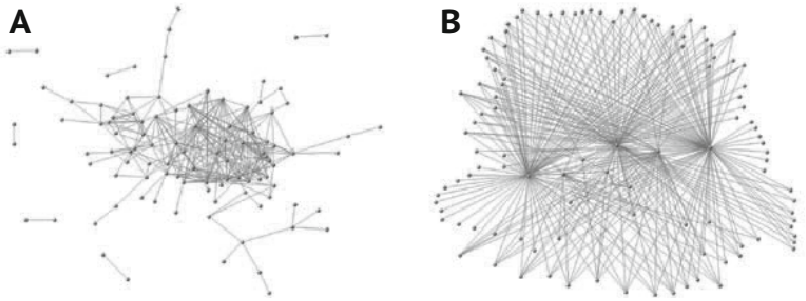


Figure 7.5. The shape of a gene co-expression network varies dramatically according to the topology matrix that is applied: A) gene correlation distance, B) topological overlap dissimilarity as described in [61]. It is apparent that the network built using the gene correlation appears more disconnected and shows poorer evidence for the nodes' centrality. In the one that considers the topological overlap of each node, all the nodes are connected and a scale-free organisation of the network is evident. *Figure adapted with permission from* [36], © (2006) *National Academy of Sciences, U.S.A.*

of the two nodes overlap. In the case of a unweighted network it has the form

$$\omega_{ij} = \frac{a_{ij} + \sum_k a_{ik} a_{jk}}{\min(k_i, k_j) + 1 - a_{ij}} \tag{7.17}$$

where k_i is the connectivity of the node i. In a real-world analogy, given two persons A and B, ω measures how many social network connections they have in common divided by the total number of connection of the one that has the least amount. The topological overlap takes values $0 \leq \omega_{ij} \leq 1$ when $0 \leq a_{ij} \leq 1$ (see [61] for further information), so it is possible to define a **topological overlap dissimilarity** distance

$$\delta_{ij}^{\omega} = 1 - \omega_{ij} \tag{7.18}$$

which can be used as a distance matrix for a **hierarchical clustering** of the nodes (genes) of the network. The rationale for using topological overlap as a measure for clustering—and not, for instance, a simple correlation measure—is that nodes with similar function usually present connections with similar neighbours. Figure 7.5 shows the differences in the gene co-expression network when using a correlation matrix (A) and a topological overlap matrix (B). The clusters obtained in a hierarchical clustering using the topological overlap dissimilarity distance are the modules of the network.

7.4.1.1. Correlation of a module with a phenotypic trait. The second step of module analysis is to find whether the genes of a module are related to

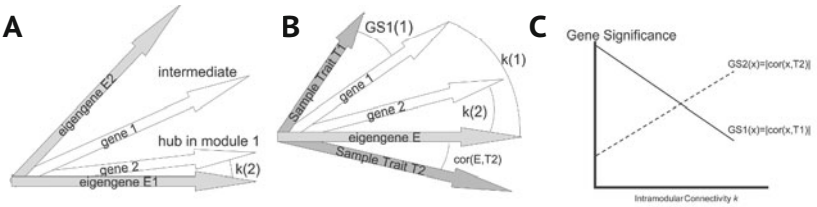


Figure 7.6. Hub genes in a module, eigengenes, and their trait-significance. A) Each gene can be ideally represented as a vector. The eigengene can be seen as a 'representative vector' of the module and the genes that are closest to its expression pattern can be seen as *module hubs*. Genes that are more peripheral in the module show an intermediate expression pattern. It can be shown [21] that the intramodular connectivity of a gene in a module is a function of its distance from the module eigengene (due to the way a topological overlap distance matrix is built). B) It's possible to assign a module a score of significance in relation to a given biological/clinical trait. This can be evaluated using the correlation of the trait with the module eigengene, or to a gene member. In the panel, for example, GS(1) is the gene significance of the gene 1 with the trait of interest T1. C) If a module is strongly associated with a trait T2, but not with a different trait T1, then the value of the gene significance for a given gene in the module is a function of its intramodular connectivity: the higher the connectivity, the higher the significance value with the module-associated trait and the lower with the non-module trait. *Figure adapted from* [21] *under a CC-BY Creative Commons Attribution License*.

a given biological phenomenon (*e.g.*, ageing, neurodegeneration, neural function, and so on). A straightforward approach is to define a **trait-based significance measure**. For each sample, we can assign information on one or more traits (*e.g.*, age, disease status, treatment). So each gene is associated with a vector of its expression related to the variation of the phenotypic trait. One can then study the correlation coefficient between the expression level of each gene i included in a module κ and the biological trait of interest[8] τ. The significance ψ of a gene i in relation to a trait is defined as an exponential function of the aforementioned correlation coefficient:

$$\psi_i^{(\kappa)} = |\rho_{i,\tau}|^b. \tag{7.19}$$

Thus, the **module significance** is easily calculated as the average of the significance measures of the genes in the module

$$\Psi_\kappa = \frac{\sum_i \psi_i^{(\kappa)}}{\#\kappa}. \tag{7.20}$$

[8] As τ could be either a quantitative or a qualitative trait, the user has to find the best way to encode the variation of the trait in the correlation measure.

Once having found the modules which are highly correlated with the studied trait, the analysis could continue with, for example:

- the **Gene Ontology** analysis of the members of the module,
- the dissection of the module in **sub-modules**;
- the detailed analysis of the topological properties of the module;
- ...

7.4.1.2. Dimensionality reduction in module analysis. After having having sectioned the network into the modules, it could be interesting to *further condense* the information present in the module, thus finding a collective description of the module expression pattern. If we consider the *expression matrix of the module* (a $\#\kappa \times m$ matrix), we can apply **singular value decomposition**[9], a procedure that is similar to principal component analysis:

$$\mathbf{D}_\kappa = U \Lambda V^T \tag{7.21}$$

where U is a $\#\kappa \times m$ matrix with orthnormal columns, Λ is an $m \times m$ diagonal matrix, whose values are incidentally the square root of the eigenvalues of the intracondition correlation matrix, V is an $m \times m$ orthogonal matrix of the corresponding eigenvectors. The eigenvector $v_e \in V$ correspondent to the highest eigenvalue in Λ is called **eigengene** of the module. The values of the eigengene offer a collective representation of the expression pattern of the genes in the module. For example, the eigengene values have been associated with the trait participation in trait-significant modules (Figure 7.6).

Having the 'condensed' expression values of the eigengene of the module, it is possible, instead of calculating the module significance according to Equation (7.20), to correlate the eigengene with the trait of interest. Moreover, one can determine which are the nodes in the module whose expression pattern is closer to the eigengene (*i.e.* the correlation $\rho_{v_e,i}$ has the highest value, Figure 7.6A). These nodes are called **module membership hubs** and are often important in the function of the module or as biological markers (Figure 7.6B–C).

[9] Singular value decomposition is a generalisation of *eigendecomposition*, which is the decomposition of a square matrix as

$$M = V L V^T$$

where L is a diagonal matrix with the eigenvalues of M and V is the matrix composed by the corresponding eigenvectors.

7.5. In conclusion: an explained example of the power of (neuro)genomics analysis

In order to illustrate how the techniques we presented in the previous chapters can be applied to a scientific question, we will comment on the analysis of the dataset reported in the work of Baumgart and colleagues [6].

In this study, a longitudinal RNA-seq design was applied to the turquoise killifish. The aim of the study was

1. to test whether groups of fish with different longevity showed differences in gene expression at an early age and
2. to correlate these gene expression profiles with the 'age at death' in order to identify predictors of lifespan at early age.

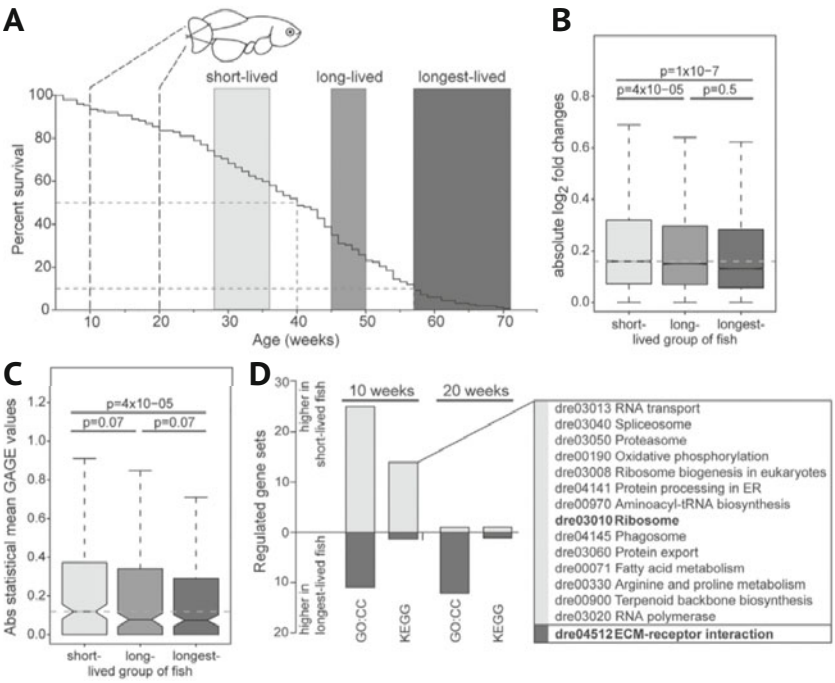


Figure 7.7. A) Average lifespan of the studied killifish population: *short lived* fish have a lifespan below the median, while the *longest lived* fish are in the top 10% of lifespan. B) The fold-change gene expression variation in the samples from short lived fishes is higher compared to the variation in the longest-lived fish. C) An even higher difference in the fold-change variation is evident when applied to a gene enrichment analysis on each KEGG pathway. D) Number of differentially expressed gene sets between short- and longest lived- fish at 10 (the list of the KEGG pathways is enclosed in the inbox) and 20 weeks. *Image adapted from* [6], *http://creativecommons.org/licenses/by/4.0/*.

The study analysed 130 individual fish. For each animal two fin biopsies were taken: one at age 10 weeks (25% of the median lifespan) and and one at 20 weeks (50% of the median lifespan). Fins regenerate in fishes and this procedure does not compromise survival. 45 individuals were selected based on the age of death (AoD) and divided into three equal groups: short-lived, long-lived and the longest-lived (Figure 7.7A).

As a first step, the authors calculated the median absolute fold-changes in gene expression between 10 and 20 weeks for the short-, long-, and longest-lived groups of fish. The largest differences were observed in the short-lived fish and the smallest were observed within the longest-lived fish ($p = 10^{-7}$, Wilcoxon signed-rank test, Figure 7.7B). Note from the panel that these differences are small in size, but given the large number of genes (>20000) that were compared, the statistical significance is very high. This also demonstrates that the global rate of age-dependent gene modulation is slower in the longest-lived group. A biological interpretation of these data is that the biological time 'ticks slower' in the longest-lived population and this is responsible for the longer lifespan.

In a second approach, a Generally Applicable Gene Enrichment (GAGE) analysis was applied (see Section 6.5.2) and, for each KEGG pathway, the absolute statistical mean of the fold-changes between 10 weeks and 20 weeks was computed for the three groups separately. This was largest in the short-lived and smallest in longest-lived groups of fish ($p = 4 \cdot 10^{-5}$, Wilcoxon signed-rank test, Figure 7.7C). It should be noted that grouping genes into KEGG pathways resulted in larger effect size. However, since the number of KEGG pathways for fishes is only 136, the p-values are larger than at the gene-level comparison. An intuitive approach to this dataset would be to identify genes differentially expressed among the three groups. The analysis of differential expression using DESeq detected a small number of DEGs (<100) due to the small size of the fold changes at the individual gene level. However, analysis by GAGE detected 15 differentially-expressed pathways (FDR < 0.05) that are more than 10% of all KEGG pathways in fishes. This illustrates how small changes at the individual gene level can generate a robust signal if they are consistent within a pathway.

A survey of the differentially-expressed KEGG pathways provides a synthetic biological interpretation of the data that points out three major cell processes:

global control of gene expression 'RNA polymerase', 'Spliceosome', 'RNA transport';

mRNA translation 'Ribosome biogenesis', 'Ribosome', 'Aminoacyl-tRNA biosynthesis';

processing and trafficking of proteins 'Protein processing in the Endoplasmic Reticulum', 'Protein export and recycling of proteins: Proteasome and Phagosome'.

All these pathways are expressed at higher levels in the short-lived individuals (Figure 7.7D). A biological interpretation is that higher activity of biosynthetic pathways early in life is correlated (and maybe predisposes) with short lifespan. Furthermore, these pathways are differentially-expressed at 10 weeks, but not at 20 weeks, *i.e.* the differences in gene expression between the short- and longest-lived fish are largest at the earliest time point. Intuitively, differences in gene expression between individuals that differ in their ageing rate should become larger as age progresses, but these data are rather reminiscent of the 'hourglass' model of comparative embryology[10]. This view is consistent with transcriptome studies of the human brain[11]. The biological interpretation of these data is that the conditions favouring longevity are expressed early in life and they leave a functional trace even after they vanish.

What are the dynamics of age-dependent regulation of those genes whose expression is correlated with lifespan? This question can be answered using k-means clustering (Section 5.2.2 at page 65) on a second cross-sectional dataset, where the gene expression was measured in samples from three tissues (brain, liver, and skin) obtained from animals that were euthanised at either of five time points ranging from young to very old age. In the longitudinal study a number of genes (n=688) showed a negative correlation with lifespan. The expression profiles of these genes in the cross-sectional study were clustered into three clusters. The first cluster showed up-regulation in all three tissues (261 genes, Figure 7.8A) and these genes can be considered markers of the biological age of skin; functionally, they may promote skin ageing. The second cluster (36%, 251, Figure 7.8B1) showed a U-shaped profile with a sharp decay between 5 and 12 weeks and up-regulation at later time points in all three tissues. This behaviour may indicate that these genes show different profiles in short- and longer-lived individuals and the population pattern may indicate higher mortality of those animals with early down-regulation. This behaviour is most evident in genes coding for RNA transport proteins (KEGG dre03013, Figure 7.8C).

[10] The model, originally proposed to compare the morphology of embryos of different species, postulates that *phenotypes are more divergent at early developmental stages, converge at mid-development and diverge again later*.

[11] See also the paper from Colantuoni and colleagues [10] described in Chapter 8 at page 131.

What is the prevalent function of the genes showing U-shaped behaviour? This question is best answered by testing the overrepresentation of GO terms. This cluster of genes is significantly enriched in transcripts coding for proteins of the mitochondrial inner membrane (N=20, GO:0005743, Fold-enrichment=7.2, FDR=$1.3 \cdot 10^{-9}$, Fisher's exact test, Benjamini-Hochberg correction). This is confirmed by a complementary approach where the differential expression of all genes of the GO term 'mitochondrion' is tested by GAGE (FDR= 0.0078 for short- versus longest-lived fish; Figure 7.8B2). Please note the complementary nature of these two tests. In the first case, a gene set is obtained by a method that provides a binary output (a gene is either contained within a cluster or it is not). In the second case, a statistical test is performed using the fold-changes of all genes belonging to a given gene set as input.

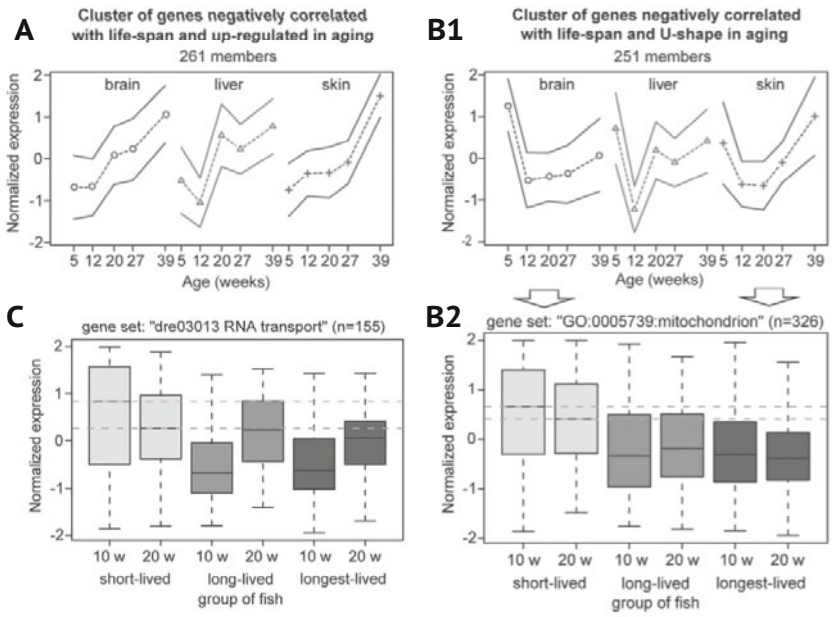


Figure 7.8. K-means clustering of the gene profiles along ageing shows the presence of clusters that are negatively correlated with lifespan and either A) up-regulated in ageing, or B1) U-shaped (*i.e.* up-regulated in development and ageing but down-regulated during maturity) in three different tissues. B2) The normalised expression of a GO term overrepresented in the U-shaped cluster shows an early correlation with the lifespan of the killifish. C) The early down-regulation of the KEGG pathway related to RNA transport is related to the lifespan of populations of killifish with different lifespan. *Image adapted from* [6], *http://creativecommons.org/licenses/by/4.0/*.

WGCNA (see Section 7.4) is a third approach to identify genes and pathways related to a biological condition[12]. The first step is the selection of the features (*i.e.* the genes) that are going to build up the gene co-expression network. The analysis was thus restricted to the 936 genes whose expression is correlated with age of death (Spearmann correlation, $p < 0.05$). WGCNA identified five modules and for each of these modules the correlation of the eigengene with age of death was computed (Figure 7.9A). Module 2 (149 genes) shows the largest absolute eigengene correlation with age of death (r =−0.45, $p = 10^{-5}$, Benjamini-Hochberg correction, Figure 7.9B). Its most overrepresented terms are for complex I of the respiratory chain (N=6, GO:0045271, $p = 6 \cdot 10^{-5}$, Fisher's exact test, Benjamini-Hochberg correction) and the members of this complex occupy central positions in the network (Figure 7.9C).

The WGCNA analysis provided a clear **working hypothesis**: the reduction of complex I activity should increase fish lifespan. This is also a *testable hypothesis*, since complex I of the respiratory chain can be potently inhibited by small molecules, such as rotenone (ROT, see [49]). The treatment of killifish with a dose of rotenone corresponding to 0.1% of the median lethal concentration (LC50) induced a life span extension of ∼15% that was statistically significant (log-rank Test, $p = 0.0181$).

Is this life-extension correlated with changes in age-related phenotypes? The transcriptome can be considered as a *multidimensional phenotypic trait* that describes the global response of an organism (or of cells) to a biological conditions or a treatment. Five animals treated with 15 pM ROT were taken after four weeks of treatment and RNA-seq was performed on brain, liver and skin samples. These were compared with animals of the same age treated with vehicle and with animals treated with vehicle for four weeks starting at age 5 weeks (young controls). The authors analysed the response to ROT of all genes differentially expressed with age (FDR-corrected $p < 0.05$, EdgeR and DEseq) detected in comparisons of old versus young controls (brain: 1436 DEGs, liver: 839, skin: 1830). In Figure 7.10, these data are shown as 2D plot with $\log_2$ of the expression ratio old controls/young controls on the X axis and $\log_2$ of the expression ratio old ROT-treated/old controls on the w axis. In all three tissues, the vast majority (∼90%) of genes up-regulated during ageing were down-regulated by ROT and vice versa resulting in highly-significant negative regression. This demonstrates that treatment with

[12] In this case, the biological condition is the lifespan.

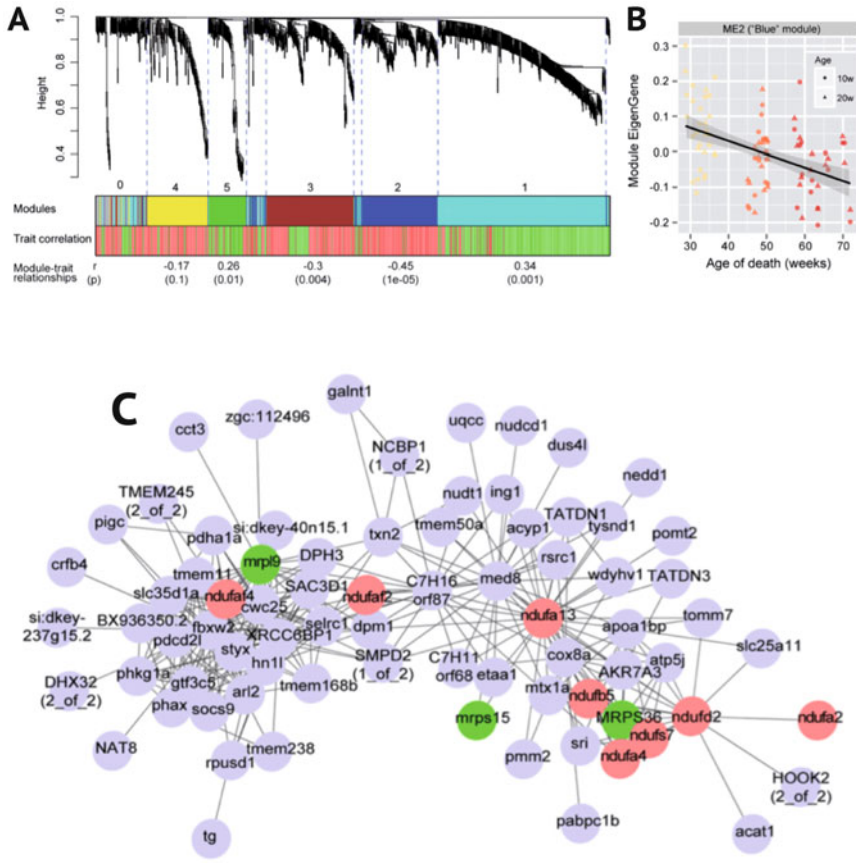


Figure 7.9. WGCNA analysis on the *N. furzeri* dataset. A) Dendrogram with module recognition and associated module-trait correlation. Green modules are positively associated with the age of death (*i.e.* their genes are highly expressed in fish that died later), while red modules are negatively correlated. B) As an example of such correlation, the expression levels of the eigengene for the blue module (genes negatively correlated with life expectancy) is shown as a function of the age of death. In the plot the components measured at adult age (10 weeks, •) or in an aged fish (20 weeks, ▲) are distinct. As can be seen, the expression of the eigengene at either 10 or 20 weeks is on average higher in fish that die earlier (30 weeks) and lower in fish that die later (70 weeks). C) Network representation of the most connected genes in the blue module using a topological overlap measure. The genes of Complex I (involved in cellular respiration) are shown in red and the mitochondrial ribosome genes are shown in red. *Image adapted from* [6], *http://creativecommons.org/licenses/by/4.0/*.

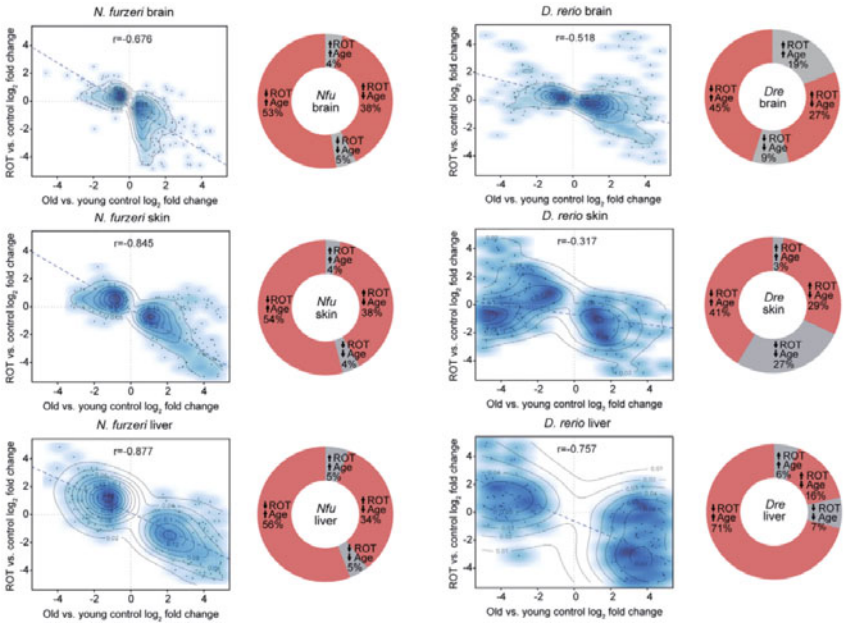


Figure 7.10. Effect of rotenone on the gene expression profiles in brain, skin, and liver of *N. furzeri* and of zebrafish (*D. rerio*). The plots show each DEG as a point whose coordinates are given by its $\log_2$ fold-change during ageing or after treatment with rotenone. The strong negative correlation (*dashed line*) between the DEG fold-changes is an indication that rotenone has an 'opposite' global effect on the transcriptomic profile of the different tissues compared to ageing. *Image adapted from* [6], *http://creativecommons.org/licenses/by/4.0/* .

ROT shifts the global gene expression towards a pattern more typical of younger ages (rejuvenation).

In summary, this study illustrates how the techniques presented in the previous chapters can be used to identify key regulators of specific biological phenomena and to generate hypotheses that can be tested experimentally.

Chapter 8
Mesoscale transcriptome analysis

8.1. Introduction

In the current chapter, we will discuss some publications that applied the previously described data analysis methods to genome-scale expression datasets in order to investigate aspects of brain organisation at the mesoscale level (*i.e.* at the level of areas and their connections). The high-throughput technique for quantification of gene expression used in most of these works is the **cDNA microarray** that, until very recently, represented the technique of choice to obtain genome-scale gene expression data. The output of a microarray experiment is an $n \times r$ matrix, where n is the number of (oligonucleotide) probes printed in the microarray chip (note that multiple probes may be associated with a single mRNA) and r is the number of samples (brain regions) analyzed, and the elements of the matrix are normalised hybridisation signal intensities for the different probes (again, a single transcript may be represented by multiple probes). The structure of the dataset is very similar to the output of RNA-seq, so the down-stream analysis uses the same methods. A notable difference is that normalised signal intensities for microarrays are near-normally distributed.

We wish to remark that, given the complexity of the brain at meso-, micro-, and subcellular-scale, the neurosciences represent a playground where the full potential of genome-wide technologies can be challenged. We provide examples spanning from neurobiology to cognitive science, in order to give the reader a hint of the full breadth of possible applications of transcriptomics at all different scales. The number of publications using these techniques is booming; it is not the aim of this book to provide a comprehensive review of the literature. On the contrary, we want to provide a few discussed examples to help the reader develop the capacity to approach these types of publications critically.

8.2. A comprehensive dataset of the human brain transcriptome

The Allen Human Brain Atlas is the most comprehensive database of spatially-resolved transcript expression in the human brain currently available[1]. At the moment, it includes microarray data from over 150 brain regions from both hemispheres of two adult healthy individuals [19] (the RNA samples from a part of these regions were analyzed with RNA-seq too), and from the left hemisphere of four further adult healthy individuals. Post-mortem NMR scans of these brains (while still in the skull) were obtained before the explant; therefore, by interpolation, the precise 3D origin of all regional samples can be identified. This produces 3D reconstructions of gene expression patterns. The database allows a description of molecular differences between regions of the human brain at an unprecedented level of spatial resolution and represents an interface between genomics and brain imaging.

In their first presentation of the Allen Human Brain Atlas dataset, Hawrylycz and colleagues [19] obtained the DEGs (see Chapter 4, and Sections 5.2 and 5.3) between different brain regions in the two full emispheres. They also applied the *Weighted Gene Co-expression Network Analysis* (or WGCNA, see Section 7.4 at page 108) to the entire dataset in order to find gene co-expression modules across different brain regions. This initial analysis on the data produced some interesting insights on the global and local structure of the human brain transcriptome and hints at the wealth of information that can be harnessed from this dataset.

Conservation of transcriptional profiles The first question that can be addressed from the Allen Human Brain Atlas dataset is: to what extent does the regional gene expression differ 1) between individuals, or 2) between the hemispheres of one individual? Hawrylycz and colleagues found a very high correlation in transcript expression when the same regions were compared between two individuals (Pearson coefficient $r = 0.98$), and a significantly high correlation ($r = 0.46$) in the DEGs computed between two given regions in the two individuals. These data indicate a remarkable stability of the human brain transcriptome. A very interesting finding was obtained when inter-hemispheric differences in gene expression levels were analyzed: no statistically-significant difference between paired left and right brain regions was confirmed in both individuals. Apparently, the well-established functional distinction between left and right hemispheres is not reflected in differences in tran-

[1] http://brain-map.org/

script expression that can be detected with conventional transcriptome analysis tools. This datum suggests that the molecular correlates of brain lateralisation are subtle and a higher power (*i.e.* more individuals, more depth) is needed to detect them. A later study [40] confirmed that interhemispheric differences of gene expression in the human brain are subtle and mostly below statistical significance.

Regional gene expression and cortical homogeneity A number of gene co-expression modules reproducible in the two individuals were identified by WGCNA (Figure 8.1A and B). Some modules were more highly expressed (*i.e.* the eigengene had higher values) in specific brain regions (*e.g.*, a *striatum* module, a *choroid plexus* module, and so on, see Figure 8.1C) and some modules showed clear overrepresentation for markers of specific cell-types (microglia module, cortical neurons module, or oligodendroglia subcortical-enriched module, see Figure 8.1B). Interestingly, the analysis of DEGs showed that almost half of these genes were poorly annotated at the time, but also that roughly 10% of those unknown genes were part of cell-type specific modules. This information can be used to impute to these genes the function associated with the module that contains them[2]. Using these data, it is also possible to compute a brain-wide dissimilarity matrix using as criterion the number of DEGs detected between each pair of brain regions (Figure 8.2A). This matrix reveals that subcortical regions show large degrees of dissimilarity on small spatial scales, as expected by the presence of many spatially-segregated nuclei with different functions and cytoarchitectures. On the other hand, the cerebellar- and the neo-cortical samples present much smaller regional variations in their gene expression profiles, probably due to their stereotypical cytological organisations. Some important exceptions to the neocortical uniformity are the primary sensory cortices and the temporal pole. In particular, the primary visual cortex stands out as an outlier (Figure 8.2B, see also Section 9.3.1 at page 154). The reasons for this are not clear, but could be related to the specialisation required to receive a large thalamic input and the expansion of layer IV that is typical of the primary sensory cortices.

Local patterning reproduces the main hippocampal divisions The authors performed an unsupervised hierarchical two-way clustering (see Chapter 5 on page 65) using microarray data derived from the hippocampal regions (CA1-CA4, dentate gyrus and subiculum). A two-way clus-

[2] For example, an unknown gene which correlates with the oligodendrocyte module may be associated with the specific functions of this cell type, such as membrane wrapping or biosyntheis of myelin (guilt by association).

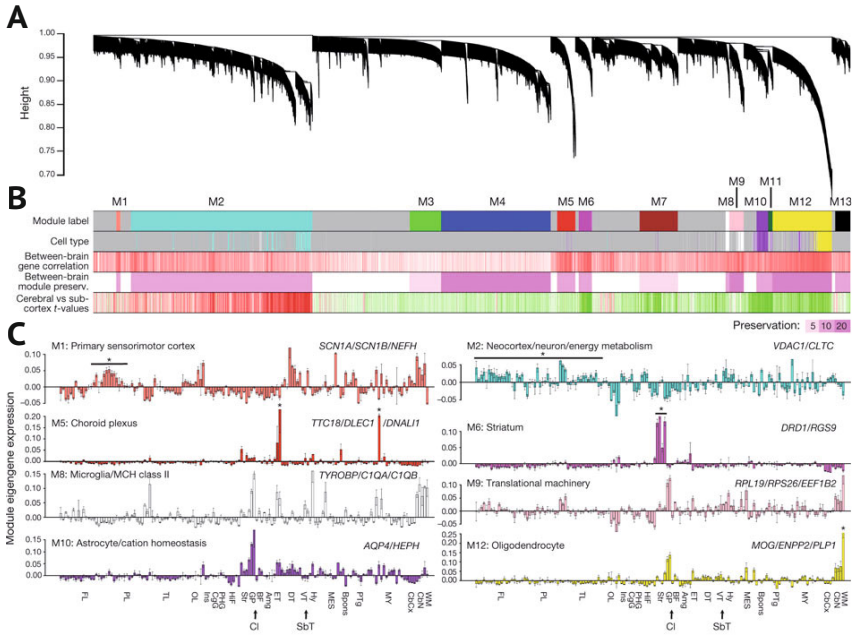


Figure 8.1. WGCNA analysis on the Allen Human Brain Atlas dataset. A) Dendrogram of the gene set hierarchical clustering across brain regions, with B) associated division in modules, enrichment of cell-type specific markers (yellow: oligodrendrocytes, purple: astrocytes, white: microglia, turquoise: neurons), value of correlation of the expression of a gene between the brains of two different individuals, overall preservation of the modules—darker modules are more preserved, thus less variable interindividually—and the value of association of the genes in the modules to cortical (red) or subcortical (green) regions. C) Expression of the module eigengene in the different analysed brain regions, with the areas of interest marked with an asterisk. The plot is essentially a graphic representation of the vector of the eigengene—that has n components, where n is the number of brain regions and each k_i value can be seen as the contribution of the region i to the eigengene—of the module. On the top-right of each histogram there are the genes whose expression profiles correlate more strongly with the eigengene. *Adapted by permission from Macmillan Publishers Ltd: Nature* [19], © (2012).

tering generates clusters on both the different genes (interregional expression profiles) and the regions (intraregional expression profiles) and then orders these in a matrix of transcriptional profiles in the different samples. This analysis reveals that the different subregions of the hippocampus can be separated based on their expression profiles[3]. This strongly

[3] That is, the sample clustering produces a CA1 cluster that contains all samples whose origin is in the CA1, a CA2 cluster, a DG cluster and so on. Interestingly, the CA3 and the CA4 regions are not separated by the clustering, which confirms previous findings that the CA3 and CA4 regions do not have a functional distinction.

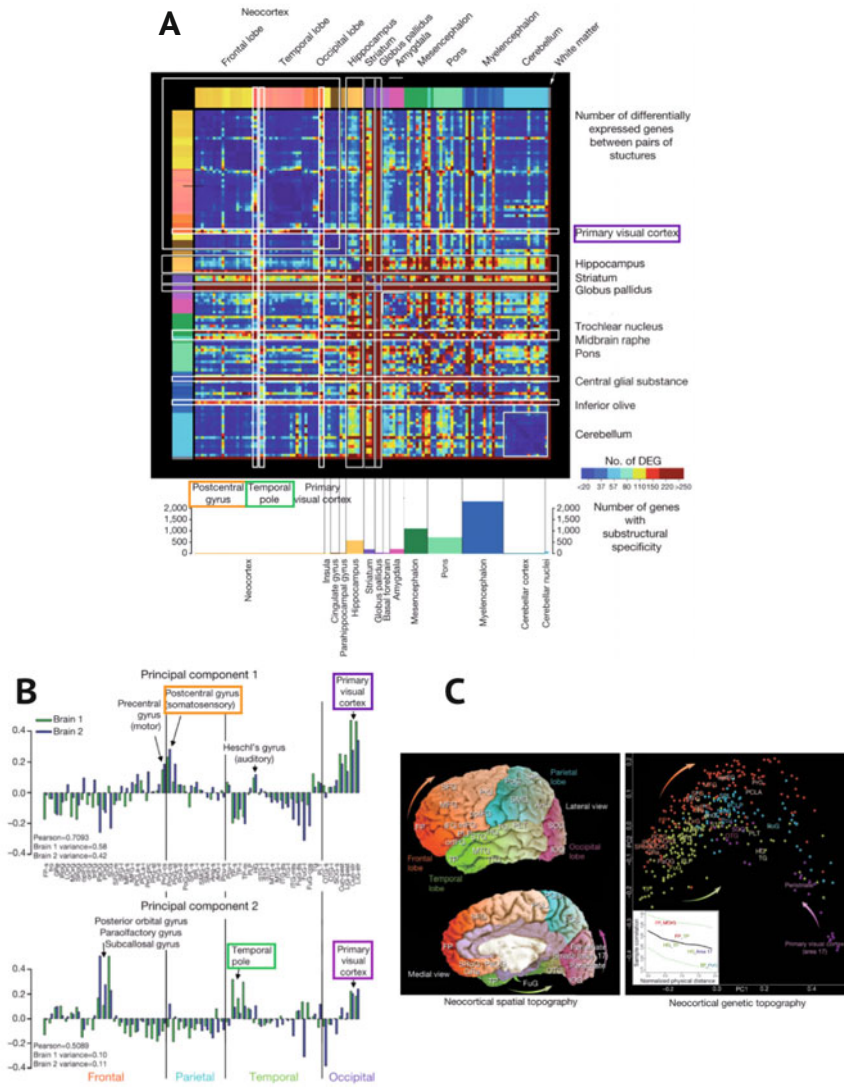


Figure 8.2. A) Matrix representing the numbers of differentially expressed genes in each pair of analysed brain regions. Notice that the neocortex and cerebellum are transcriptionally very homogeneous—with a few exceptions: the primary visual cortex (purple), the temporal pole (green), and the postcentral gyrus (orange). The variability of gene expression between subregions of a given anatomical structure can be quantified in the associated histogram (bottom panel). B) Histogram that represents the Pearson correlation coefficient between the gene expression profile of a cortical area and the first two principal components of the PCA computed on the subset of the cortical samples. It can be noticed how brain regions with higher variability in panel A also show a higher correlation with (*i.e.* contribution to) the first two principal components. C) The MDS projection of the gene expression profiles from different cortical regions (*right*) reflects the spatial topology of the neocortical regions (*left*). *Adapted by permission from Macmillan Publishers Ltd: Nature* [19], © (2012).

suggests that each subregion of the hippocampus has a specific transcriptional fingerprint. An independent validation of this concept can be easily obtained by classical staining techniques such as immunohistochemistry. Although this may appear as a trivial control to a student trained in Neuroanatomy, in Systems Biology it is fundamental to experimentally validate conclusions derived from computational analysis of high-throughput datasets.

The neocortical transcriptome reflects the cortical topology As already mentioned in the previous section, the gene expression patterns in the neocortex are quite homogeneous on a spatial scale—excluding a few exceptions. It is however still possible to select a set of genes with the largest variation of expression across the different cortical samples (*feature selection based on maximum variance*). Howrylycz *et al*. chose 1000 highly variable genes in the 56 sampled cortical regions and performed a PCA (see Section 5.3 at page 70) on the resulting matrix. It is understandable that the first two principal components—which take into account more than half of the variance in each brain—point out the regions with highest variance (see ***Regional gene expression and cortical homogeneity***). To obtain a compact graphical representation of the dissimilarity between samples, they applied multidimensional scaling (MDS, see Section 5.4 at page 78). Amazingly, the relative positions of the different samples in the MDS plane reflect the spatial distances between the associated brain regions and their relative organisation with a reasonable accuracy. In particular, there is an almost linear correlation between the 'genetic distance' expressed in the MDS plane projection and the MRI-related 'physical distance' between the regions. The goodness of fit between these coordinates is slightly less than 30% (Figure 8.2C)!

8.3. Evolutionary biology

Conservation and evolution of gene networks in humans and apes

Another early example of the application of WGCNA to neurogenomics was provided by Oldham and colleagues [36] in the field of Evolutionary Biology. Humans and apes show a high similarity at both genomic (∼98% DNA base pair conservation between humans and chimps) and gene expression level, but the differences from a neurobiological point of view are striking. It has been classically recognised that potential driving forces of the peculiar evolutionary changes in the human brain are 1) gene loci with *higher selective pressure* and 2) differential species-specific spatio-temporal regulation of mRNA transcription. As a result, these two points might induce different properties in the transcriptome network.

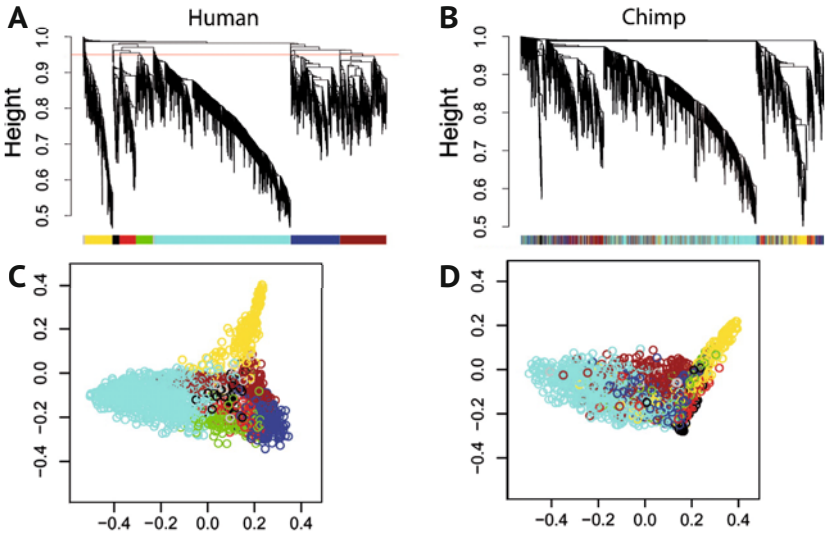


Figure 8.3. WGCNA analysis on microarray datasets derived from human and chimpanzee brain regions. A) Dendrogram and module division derived from the human brain samples. B) Dendrogram from the chimp dataset, with each gene coloured according to the module of the associated human homologue genes. C) Multi-Dimensional Scaling of the genes in the human dataset, coloured according to their module membership and D) of the genes in the chimp dataset, coloured according to the membership of their human homologue. The plot illustrates the conserved and the less conserved modules between humans and chimpanzees. *Figure adapted with permission from [36], © (2006) National Academy of Sciences, U.S.A.*

Oldham and colleagues analysed microarray data from human and chimpanzee samples originating from matched brain regions (Broca's area, anterior cingulate cortex, primary visual cortex, prefrontal cortex, caudate nucleus, and cerebellar vermis). A WGCNA approach on the human dataset identified six gene co-expression modules (Figure 8.3). Each of the eigengenes of these modules was 'highly represented' in a different brain region (pooling human and chimpanzee samples in the analysis) with the exception of a smaller module, whose most central genes are associated with myelination and glia[4]. Some of those modules (*e.g.*, turquoise/cerebellum, yellow/caudate nucleus) are strongly conserved between humans and chimps, while others—particularly the *purely cortical* ones—are weakly conserved (Figure 8.3). The conservation of a mod-

[4] The association between transcriptome and regional patterning was strongly confirmed later on in the Allen Human Brain Atlas paper [19] (see page 123).

ule is assessed using the correlation between the intra-module connectivity score k_{in} (see Equation (7.8)). The score is the sum of the weights between the gene and its neighbours that are included in the same module. For example, when a human gene *i* is strongly connected with genes in its module, its $k_{in}(i)$ will be high; in case its chimpanzee homologue gene *i'* shows fewer/less strong connections with the genes in its corresponding module, its $k_{in}(i')$ will be low. This would be a hint of a variation of the functionality of *i* in the network during the evolution of the human brain[5]. The power of the analysis of the intramodular connectivity variation is that, once we have compiled the lists of human genes that are less conserved in each module, it is possible to study their biological function using the knowledge-based dimensionality reduction methods discussed in Chapter 6 (*e.g.*, Gene Ontology, or KEGG pathways). For example, the genes in the 'blue' module, which is cortical, are the least conserved in their connectivity between human and chimpanzee brains and show some overrepresented GO categories such as

- protein transporter activity,
- microtubule cytoskeleton,
- ion transporter activity, particularly the *electron transport chain* (11 genes).

Interestingly, the electron transport chain has been linked in the scientific literature with an accelerated evolution in anthropoid primates [38].

As we have seen in Chapter 7, a module hub gene is a gene with a central position within a module in the network[6]. It is thus important to study the conservation and the variation of the connectivity patterns of the hubs in the different modules in order to recognise human-specific network connections. Oldham and colleagues defined a *human specificity score* HS according to the variation of the topological overlap $\omega_{i,j}$ (see Equation (7.17) at page 110) of a given gene couple (i,j):

$$\mathrm{HS}_{i,j} = \frac{\omega_{i,j}^{\mathrm{human}} / \langle \omega^{\mathrm{human}} \rangle}{\omega_{i,j}^{\mathrm{human}} / \langle \omega^{\mathrm{human}} \rangle + \omega_{i,j}^{\mathrm{chimp}} / \langle \omega^{\mathrm{chimp}} \rangle} \tag{8.1}$$

where $\langle \omega \rangle$ is the mean topological overlap in a given gene co-expression network. The HS score for each couple of genes is then thresholded at

[5] This difference is more striking if we consider that in some cases there is a strong difference in the k_{in} ranking between the chimpanzee and human modules of a given gene, while the transcriptional level of the two homologue genes are constant in the two species. For example, Oldham and colleagues discuss the case of the cortical gene NRG1.

[6] See also the Figure 7.5B at page 110, which shows the organisation of the network from [36] when a topological overlap metric is applied.

0.8. This means that a connection between two genes *i* and *j* is deemed 'present in human and absent in chimpanzees' if the value of the normalised topological overlap for the couple in chimpanzee is less than 25% of the normalised topological overlap in the human network. The matrix derived from the intra-modular HS can be then used to build a human-specific network. This can be used to evaluate which modules are richer in human-specific interactions and whether these interactions converge on single hubs of the system or are evenly spread in the module. Interestingly, the percentages of human-specific connections in each module correlate with the 'evolutionary specificity' of each brain region, with the cortex module showing the highest degree of human-specific connections (17.4%), while the more ancient caudate nucleus and cerebellum show 7.8% and the 4.5% of human-specific connections, respectively. Moreover, a large fraction of the human-specific connections are converging onto a specific set of *hub genes*. It is likely that these genes—in some cases poorly characterised or with unknown function—underwent some important changes in recent human/chimpanzee evolution. These changes could be, for example, an *increase in the expression level* of the transcript, or a change in its interaction partners. Figure 8.4 shows the analysis made by Oldham and colleagues in order to investigate why some genes show a higher *differential intra-modular connectivity*

$$\Delta k(i) = \log\left(k_{in}(i)\right) - \log\left(k_{in}(i')\right) \tag{8.2}$$

where *i* is a human gene and *i'* is its chimpanzee homologue. Plotting the differential connectivity against the differential expression (Figure 8.4A) shows a very significant positive correlation (Spearman coefficient 0.32, $P < 2.2 \cdot 10^{-6}$). A second feature that can be investigated *in silico* is whether the differences in the intramodular connectivity could be due to genetic modifications such as *in-dels*, inversions, and other structural variations. In order to test the effect of these modifications on the co-expression network, the authors aligned the exonic sequences of a fraction of the genes that show either a higher or similar connectivity in humans as compared to chimpanzees and then evaluated the presence of gaps in the sequences (Figure 8.4B). The average number of gaps in the sequences of genes with high $\Delta k(i)$ is three times higher than the number of gaps in the genes with $\Delta k(i) \sim 0$. This indicates that genomic rearrangements have some effects on the structure and the evolution of gene co-expression networks.

A third element of change that could contribute to the differential connectivity of a gene is the divergence in the associated protein sequence—which could potentially change its interaction partners and thus the as-

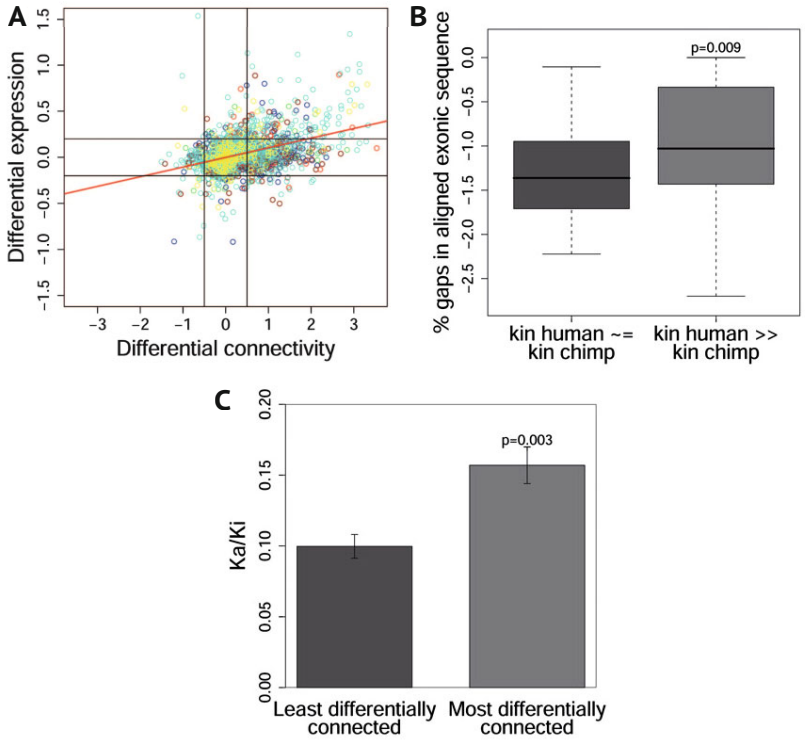


Figure 8.4. Properties of the genes that show human-specific connectivity. A) Genes that are more differentially connected tend also to be differentially expressed, each gene is associated with its module colour. B) The analysis of the gaps in the alignment of the exon sequences of homologue genes suggests that the genes with higher differential connectivity are likely to be found in regions that have experienced more genomic rearrangements. C) Genes with more human-specific connections show an accelerated protein sequence divergence compared to the genes with conserved network connectivity. *Figure adapted with permission from* [36], © (2006) *National Academy of Sciences, U.S.A.*

sociated gene co-expression. A metric for the **protein sequence divergence** is the ratio between the rate of non-synonymous nucleotide substitutions[7] (K_a) between i and i' and the control rate of nucleotide substitution in interspersed repeats in a region of 250 kBp centered around each gene (K_i). If K_a/K_i is very low, it means that the protein sequence is undergoing a strong purifying selection, so that non-synonym mutations are negatively selected. On the other hand, if K_a/K_i is high, there could

[7] A non-synonymous mutation is a nucleotide mutation in the coding sequence that determines a change in the encoded polypeptide sequence. A mutation in the protein coding sequence which does not change the aminoacidic sequence is called *synonymous mutation*.

be either a relaxation in the selection or a positive selection for different coding sequences. Figure 8.4C shows that the average protein sequence divergence of the genes in the most differentially connected quartile is significantly higher than the genes in the least differentially connected quartile. This indicates that genes that experienced an accelerated evolution in the human lineage also reshaped their connectivity in gene co-expression network in a way that is possibly related to the new functions they acquired.

8.4. Systems biology

Evolution of gene expression in the human prefrontal cortex during development and ageing

In 2011, using microarrays, Colantuoni and colleagues analysed gene expression in samples derived *post-mortem* from the human prefrontal cortex of 269 healthy individuals with age spanning from a few post-natal days to over 70 years. A number of 38 fetal (weeks 14 to 20) samples were included in the analysis. Moreover, the researchers obtained a genotype of $>$ 600,000 Single Nucleotide Polymorphisms for each subject.

The first question they asked is how the rate of expression change[8] evolves during development and ageing. The quantification of the rate change was done by applying linear models to the gene expression data. As can be seen in Figure 8.5A, the rate of expression change varies strongly during fetal development (6 weeks) and slows markedly during the post-natal months (infant, 0-6 months) and continues to slow down during childhood and adolescence[9], and stays almost constant in adulthood (20-40s). In aged brains, however, a reversal of the trend in expression change is visible, with an acceleration that persists throughout ageing. In addition, the grey histogram in Figure 8.5A shows how many genes invert their post-natal expression profiles—*e.g.*, genes that were down-regulated in infancy start being up-regulated with age or *vice versa*. These phenomena are consistent with the previously mentioned 'hourglass model' of comparative embryology (see note 10 at page 115).

Colantuoni and colleagues were able to show a global effect of development and ageing in the transcriptional profiles of the human prefrontal cortex using two of the previously presented unbiased dimensionality re-

[8] The rate of expression change computes how fast the expression of a given gene changes in a given time window—in the case of the presented paper, the absolute expression change is represented in a $\log_2$ scale, that is the number of doublings/halvings per year.

[9] It's worth noticing that the absolute rate of expression charge changes of one order of magnitude between infancy and childhood.

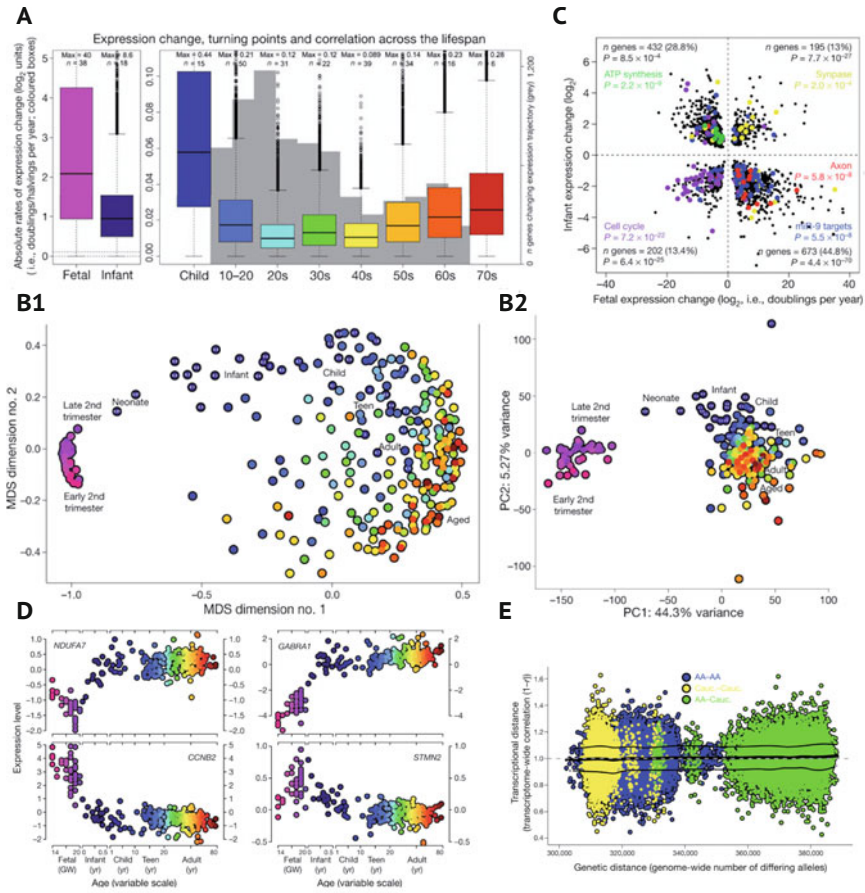


Figure 8.5. Evolution of gene expression in the human prefrontal cortex. A) Expression change and correlation across the human lifespan. See main text for further detail. B) Dimensionality reduction techniques such as Multi-Dimensional Scaling (B1) and Principal Component Analysis (B2) are able to discriminate samples from different developmental stages and age. C) Comparison between the expression trajectories in fetal development and the ones in infant development (see main text). D) Expression trajectory of four representative genes during development and ageing. E) The comparison between genetic distance (number of different SNPs from two individuals) and transcriptional distance (transcriptome-wide inverse linear correlation) shows that, overall, genetic variability does not significantly affect the gene expression architecture of the individual. *Adapted by permission from Macmillan Publishers Ltd: Nature* [10], © (2011).

duction techniques: Multidimensional Scaling (MDS, see Section 5.4 at page 78) and Principal Component Analysis (PCA, see Section 5.3 at page 70). In panels B1 and B2 of Figure 8.5 it is possible to see how both MDS and PCA are able to roughly distinguish the different age stages of the dataset. Moreover, the difference in the global transcriptional profile

between the fetal and post-natal samples is visible (along MDS dimension 1 and PC1), and the fact that the second component of the plot is able to discriminate the fetal samples from the early second trimester and the ones from the late second trimester. It is important to remark that samples of increasing age position themselves along the second component of the projections 'going' upward initially, and later show a downward trend. This is the reflection of inversions of the direction in gene expression that are described below. The strong difference in global expression profiles in the fetal and the infant samples makes it reasonable to compare the expression change during pre- and post-natal neurodevelopment (Figure 8.5C). The 'trajectory' of the brain samples at different ages in the space of the principal components would suggest that an **expression trajectory** can be associated with each gene or functional group of genes (Figure 8.5D). In Figure 8.5C there are four quadrants, corresponding to genes that either keep the expression change of the fetal development (although with a decreased absolute value in infancy), or to genes that invert the trajectory in infancy. Interestingly, a pathway analysis on the genes that present a statistically significant time-dependent slope at both fetal and infant stages shows that genes in a same pathway are more likely to be represented in the same quadrant. For example, the genes in the cell cycle pathway are mostly present in the quadrant of negative expression change in both ages (the neurons are post-mitotic cells), or the ATP-synthesis-related genes decrease during fetal development but then increase again with post-natal development[10]. Interestingly, it is apparent that the targets of some specific miRNA (*e.g.*, the miR-9 target mRNAs) show a similar quadrant clustering.

Furthermore, Colantuoni and colleagues investigated the influence of the genetic background on RNA expression levels in the human prefrontal cortex transcriptome. They used the SNP data collected from a DNA microarray in order to confirm the strong role that single SNPs have in affecting the RNA expression of individuals and to evaluate a score of **genetic distance** (*i.e.* the genome-wide number of differing alleles). They then compared the genetic distance to the **transcriptional distance** between each pair of individuals[11] and found, surprisingly, that *there is no association of genetic distance with their prefrontal cortex transcrip-*

[10] This can be explained by the fact that in fetal development the consumption of energy is mostly due to the cell replication, while in the post-natal brain it's the neural maturation and activity that drives up the energy consumption.

[11] In the paper the transcriptional distance is a transcriptome-wide linear correlation as in Equation (5.2) at page 62.

tomes (Figure 8.5E). The lack of association is confirmed even after restricting the analysis of genetic distance only to the SNPs that are associated with statistically significant modifications of genetic expression. A possible conclusion of this finding is that, despite genetic differences between individuals, the human genome produces a **robust molecular architecture** in the brain—or at least in the prefrontal cortex—throughout ageing, so that population-wide variations in gene sequences do not influence the transcriptional architecture at a global scale.

8.5. Molecular neurobiology

RNA-seq makes it possible to investigate a non-canonical form of RNA, the circular RNAs

Circular RNAs (**circRNAs**) are a relatively recently discovered splicing variant of RNAs, where the 3′ terminus of the last exon is backspliced with the 5′ terminus of the first exon (‘*head-to-tail splicing*’) resulting in a circularised RNA. It has been proposed that this species of RNA function as miRNAs and RNA-binding protein **molecular sponges**, thus indirectly modifying the cellular processing of the associated linear mRNAs. Recent evidence suggests that these RNAs might undergo cap-independent translation [48]. You and colleagues [60] applied an RNA-seq strategy on whole-tissue extracts and on hippocampal cell cultures in order to investigate the biology of circRNAs. This strategy was then complemented with classical experimental validation, such as Fluorescence *in situ* hybridisation (FISH).

The authors sequenced the total RNA samples (depleted of the rRNAs, as seen in Section 2.3.2 at page 14) from five different mouse tissues: brain, heart, liver, lung, and testis. In order to compute the fraction of the RNAs that are circularised, one should look for reads that partially cover the first and the last exon of a given gene—that is, the ones resulting from a head-to-tail splicing event—and compare them to the reads that are aligned to the gene and the canonical exon-exon junctions. From this sequencing it is evident that *the brain is highly enriched in circRNAs*, compared to the other tissues, at three different levels.

Relative abundance The relative abundance of reads related to circRNAs in the brain is about 0.08%, while the testis, which is the tissue with second highest value, has an abundance of around 0.03%.

Hosting genes The authors found that about 20% of the genes detected in the brain extract produced circRNAs. In contrast, other tissues had either slightly more (again, the testis) or fewer than $< 10\%$ genes that produced them.

Tissue-specific circRNAs An analysis of the number of tissue specific circRNAs again shows that the brain hosts the highest number of tissue-specific genes from which circRNAs originate (> 200), while the testis is second (around 150 genes) and the other analysed tissues have $\ll 50$ genes.

A GO analysis of the genes that produce circRNAs more frequently shows that there is a strong enrichment of transcripts related to synaptic function–such as 'presynaptic active zone' (~ 5 fold enrichment), 'post-synaptic density' (~ 4 fold enrichment), or 'synapse' (~ 3 fold enrichment). It is well known that some neuronal transcripts are transported into neuronal processes (see also Section 9.2.1 at page 145). In order to prove experimentally that the circRNAs in the brain are enriched in the neurite compartment, You and coworkers extracted the synaptic fractions from mouse and rat brains either through the isolation of synaptosomes or with the microdissection of the neuropile in the hippocampus[12]. The RNA-seq data from these experiments show that most circRNAs can be reliably detected in both synapse-enriched preparations and that the genes hosting the circRNAs are conserved in mouse and rat. A further experimental validation of the neuritic localisation of circRNAs was provided through the hybridisation of a fluorescent probe designed against specific circRNA targets in hippocampal cell cultures. The experiment confirmed the presence of circRNAs at the soma-dendritic level.

Due to the observation that the regions around head-to-tail splicing junctions are highly evolutionarily conserved in the genes hosting circRNAs, You *et al*. analysed whether the circRNAs profile of neurons would change 1) during neural development[13], and 2) after homoeostatic plasticity protocols[14]. As expected, there was a massive regulation of different synaptic circRNA species, which suggests that this species of RNA has a functional role in neuronal activity.

More recently, it was shown that loss of a circular RNA causes brain dysfunction providing direct evidence of the functional importance of cir-

[12] The hippocampus has a highly-segregated structure, with a core zone which contains all the somata (*stratum piramidalis*) and the surrounding region which contains the majority of the neurites (*stratum oriens* and *stratum radiatum* in particular).

[13] In particular between post-natal day 0 (P0) and P10, when the first wiring of synaptic connections occurs.

[14] Homoeostatic plasticity events occur when there is global scaling of the synaptic strength of the synapses of a given neuron. The exposure of a cell culture to the GABA-A receptor agonist *bicucullin*—*i.e*. a molecule that mimics the action of the neurotransmitter GABA, which is inhibitory to adult neurons—triggers a general downscaling of the amplitude of the excitatory post-synaptic potentials.

cular RNAs in the nervous system [39]. In this case, the mechanism of action was a post-transcriptional regulation of two microRNAs. However, there is also some evidence that circRNAs can be translated in axons [48] and at least one protein derived from circRNA was detected in fly synapses [37]. The mechanisms by which circRNAs control neuronal function are manifold. It is easy to predict that in the future circRNAs will become a popular object of investigation in the neurosciences.

8.6. Brain networks

Correlation in gene co-expression is associated with synchronous activity in cortex networks

As a last example of the application of transcriptomics to study brain function, we will describe the work by Richiardi and colleagues [44]. The anatomical and functional connections between brain areas can be described using the same formalism that is used to describe gene-coexpression networks and network theory finds important applications in the study of the brain *connectome*. An obvious question arises: are the cortical modules identified by analysis of the connectome and those identified by analysis of gene co-expression patterns *related* in the human brain? The spatial resolution of the Allen Human Brain Atlas has allowed us to answer this highly relevant question for the first time. Richiardi *et al*. associated the networks of brain activity derived from resting-state functional Magnetic Resonance Imaging (fMRI[15]) of 15 healthy subjects with the corresponding gene expression data from the Allen Human Brain Atlas ([19], see also Section 8.2 in this chapter).

Using Independent Component Analysis[16], the researchers built a series of highly reproducible resting-state brain networks. Out of these,

[15] Functional MRI is a neuroimaging technique that reveals the areas of brain activation during a task or in the resting state relying on the association between enhanced neural activity in a brain tissue and increased local blood flow. In this note, we will describe the biophysical signal underlying BOLD (*blood-oxygen-level dependent*) fMRI—there are, however, different ways to get a signal of brain activity in MRI, for example using the diffusion of water molecules. The BOLD signal is present due to the chemical properties of haemoglobin. This molecule can exist in two different states, either coordinated with a oxygen molecule (oxyhaemoglobin, oxHb) or non-complexed (deoxyhaemoglobin, dHb). While oxHb is a diamagnetic molecule (so it doesn't interact with magnetic fields), **dHb** is a **paramagnetic** molecule so it reorients itself in presence of a magnetic field, thus forming a perturbation in it. The emission of a pulse of radio waves excites the hydrogens of the molecule, that then reemits a radio wave due to the the relaxation of the protons to the ground state. The relative abundance of oxHb:dHb in a specific neural tissue makes possible to build up an image of the pattern of activation in the brain.

[16] The ICA is a method similar to PCA that tries to separate the different components of a multivariate signal, through the assumptions that 1) the sources of the signal are non-Gaussian and 2) each source is statistically independent from the others.

they chose four large and non-overlapping networks: the dorsal default mode, salience, sensorimotor, and visuo-spatial networks. Each node of the networks derived from fMRI was then linked to the Allen Human Brain Atlas regions using normalised Montreal Neurological Institute coordinates. The non-cortical regions (together with the deep grey tissue such as the hippocampus) were excluded from the analysis, due to the fact that some gross differences in the transcriptomes of these tissues (compared to the cortical samples) are likely to arise from the differences in the tissue ontogeny. The chosen functional networks include 241 samples from the Allen Human Brain Atlas dataset.

Richiardi and coworkers then built a network where nodes are brain regions and edges are based on the correlation of gene expression across samples. In this network, the strength of the connection between two brain regions is related to the similarity in their expression patterns, computed through the Pearson correlation coefficient between each pair of brain regions[17] (i, j)

$$\rho_{ij} = \frac{\sum_{genes}(q_i - \mu_i)(q_j - \mu_j)}{\sqrt{\sum_{genes}(q_i - \mu_i)^2}\sqrt{\sum_{genes}(q_j - \mu_j)^2}} \tag{8.3}$$

where q_i is the expression value of a given gene in brain region i and μ_i is the average gene expression in the region. If a correlation value ρ_{ij} is negative, it is set to zero (that means there is no edge between the two regions in the correlated expression network).

In order to determine whether the gene expression correlation is higher in functionally connected regions—as found in the resting state fMRI networks, which are non overlapping—Richiardi and coworkers defined a measure of the 'strength fraction', which evaluates how much the nodes in a specific fMRI network $N \subseteq \mathcal{N}$ are connected in the gene correlation expression compared with the regions that aren't in any functional network ($\bar{\mathcal{N}}$, ~ 1500 nodes)

$$S_N = \frac{\sum_{i \in N} k_i}{\sum_{i \in \bar{\mathcal{N}}} k_i} \tag{8.4}$$

where $k_i = \sum_j \rho_{ij}$ is the weighted connectivity of the brain region i (see also Equation (7.8) at page 106). High values of S_N mean that the correlation of the gene expression between the regions in a specific functional network is higher compared to the correlation in the non-network

[17] The co-expression network for each single patient was computed separately.

regions[18]. In order to test the significance of the strength fraction for a given network, the authors used a permutation test (see Section 6.5), where the nodes of the set $\mathcal{N}$ are randomly reassigned, so that a density distribution of the values of the strength fractions S'_N can be estimated and the p-value computed. The strength fraction analysis showed that the functional networks have a significantly higher relative correlation in the gene expression compared to the one due to chance ($P < 10^{-4}$).

From gene expression correlation networks it is possible to determine the top-ranking genes that might affect the connection between functional brain networks and gene expression correlation[19]. The consensus list included 136 genes, mostly involved in the regulation of cationic channels[20] or other membrane receptors such as the γ-aminobutyric acid receptor GABRA5. The consensus list was used to validate the hypothesis that the expression values of a specific set of genes correlated with the functional connectivity of different brain regions using two approaches:

- using a dataset that pairs genetic variability measuring single nucleotide polymorphisms (SNPs) with the resting state fMRI recordings of 256 adolescents (equally distributed between males and females), the authors confirmed that sequence variations in members of the consensus gene list are associated with variations in the 'strength fraction' of the functional network computed for each individual;
- comparing the Allen Mouse Brain Atlas [34] (the mouse equivalent to the Allen Human Brain Atlas) to the Allen Mouse Brain Connectiv-

[18] In the Allen Human Brain Atlas annotation, some samples are associated with the same region, and these regions are likely to be strongly connected according to their gene expression correlation. In order not to boost the value of S_N, in case N would include nodes with a high number of biological replicates, Richiardi and coworkers removed the edges connecting samples with the same regional ontology.

[19] The detailed description of the methodology used to get the final gene list, the *list intersection discovery test*, is beyond the aim of this book. It is sufficient to mention here that the authors chose to use a False Discovery Rate of 5% (see also Section 4.3.3 at page 55); that is, it is expected that up to 5% of the genes in the final list are false positives.

[20] The statistically significant GO annotations in the consensus list are:

Molecular Function 'Voltage-gated cation channel activity'
Cellular Compartment 'Ion channel complex', 'Potassium channel complex', 'Voltage-gated potassium channel complex', 'Extracellular region', 'Plasma membrane', and 'Cation channel complex'
Biological Process no significant term.

Using a cell type-specific mouse transcriptome database [8] and the homologues of the consensus list, the authors were able to find that $\sim$ 30% of the genes were overexpressed in the neurons, $\sim$ 14% in the astrocytes and a similar fraction in the oligodendrocytes (the remaining genes were not significantly overexpressed in the mentioned dataset). It is thus arguable that the connection between the brain activity correlation and the regional gene expression is mostly due to the neuronal function.

ity Atlas [35], which provides a mesoscale model of axonal connectivity[21], Richiardi and colleagues were able to produce a tissue gene co-expression network using the orthologues of the consensus list as the 'feature selection' gene set. They then compared it with the graph of mouse brain connectivity. The correlation between the graph built using the consensus list and the mesoscale connectome is significantly higher than the correlation with a gene expression network built with an equally sized set of randomly selected genes (P=0.022 or P=0.011 according to the connectome model used). It is quite interesting to notice that the correlation between the resting state patterns of brain activity and the gene expression profiles seems to be evolutionarily conserved, as the human gene consensus list is significantly associated with brain connectivity in rodents.

This study represents an outstanding achievement of neurogenomics and shows the full potential of combining genomic and functional studies.

[21] Mesoscale cortical connectivity maps are available from high-throughput tract-tracing experiments in mice and represent a highly valuable dataset since they provide a direct measure of physical connections between brain areas that is not possible in humans. The mouse isocortex was then parcellated in 38 mesoscale regions, which means the resolution of this part of the study is somehow lower compared to the 88 single regions of the human study.

Chapter 9
Microscale transcriptome analysis

9.1. Introduction

In the previous chapter, we have analysed some important applications of RNA-seq in the neurosciences. A common feature of the described papers is that the spatial resolution of the analysed transcriptome corresponds to a brain subregion (*e.g.*, cerebellum, cortex, hippocampus, and so on). However, this type of approach can neither discriminate the signals originating from different cell types (neurons astrocytes, other glia cells) or subtypes[1] nor can account for cell-to-cell variability in gene expression.

A growing number of techniques were devised to increase the spatial resolution of the RNA-seq studies *down to the single-cell transcriptome*. In this chapter, we will describe two different approaches to studying the differential expression of genes at the microscale:

- the use of *fluorescence-assisted cell sorting* (FACS) or *RNA immunoprecipitation* strategies, that makes it possible to isolate and sequence RNAs from specifically tagged cell types and obtain **cell population-specific transcriptomes**;
- **Single-cell RNA-seq** that combines either microfluidics or FACS with cDNA amplification in order to quantify genome-scale transcript abundance from single cells (or single nuclei).

We can easily predict that the future of RNA-seq studies of the nervous system will be increasingly characterised by datasets at the single cell level. However—before undergoing the analysis of these experimental approaches—it is necessary to emphasise two important issues: **coverage** and **signal-to-noise ratio**. A single-cell RNA-seq dataset comprises

[1] It is worth remembering that the central nervous system is home to extreme variability in cell morphologies and behaviours. Consider, for example, the different sub-classes of inhibitory interneurons.

a large number of samples (*up to 100,000 with the most recent technologies*), each corresponding to a single cell, and it would be unsustainable to produce the usual 40-50 million reads for a single sequencing experiment. This implies that the number of reads per sample is reduced to a handful of million reads. Moreover, the challenge of extracting, capturing, and amplifying the mRNA molecules from a single cell could make it impossible to detect many low- to even moderately-expressed genes, or could lead to an artefactual enrichment of some species[2] [51].

9.2. Microscale transcriptome strategies involving ribosome immunopurification

In recent years, the interest in profiling the molecular fingerprint of specific classes of cells in the brain has driven the creation of several strategies for enriching a sample with RNAs derived from specific cell populations. A general blueprint of these microscale (but not single-cell) methodologies—often called *translating ribosome affinity purification*, or TRAP—is as follows:

1. find a promoter that can drive the expression of a transgene in the cell population of interest;
2. tag the ribosome of these cells with a *molecular tag*[3], such as a fluorescent protein [12] or haemagglutinin (HA) [48], nanobodies [14], and so on, usually fused to **ribosomal protein L10a**. An alternative choice is to use an *endogenous tag*—such as the phosphorylation of the ribosomal subunit S6—to specifically select the ribosomes of neurons that are activated after a specific stimulus [27];
3. purify the tagged ribosomes through immunoprecipitation[4], extract the bound mRNA and then perform microarray analysis or RNA-sequencing. This technology allows to quantify the transcriptome of neurochemically-defined cells (cholinergic cells by using the ChAT promoter, GABAergic cells by using the GAD promoter, and so on).

[2] These problems can be addressed through careful quantification of the technical variability in library construction and the sequencing, for example using exogenous RNA *spike-in* or molecular barcodes. For more information see the review from Stegle and colleagues [51].

[3] The strategy developed by the GENSAT project (http://www.gensat.org/), for example, applies a Bacterial Artificial Chromosome (BAC) strategy with promoters that drive the expression of the transgene in different cell types [12]. An alternative approach is to use a transgenic mouse with a floxed ribosome transgene (*i.e.* its sequence flanked with two Cre-LoxP recombination sites) and to express the CRE recombinase with a cell-specific promoter [48].

[4] For example, in the case of ribosomes including a EGFP-tagged L10a, the immunoprecipitation is performed using beads coated with antibodies against GFP that thus bind the GFP-tagged ribosomes and pull down the ribosome-bound mRNAs upon centrifugation [12].

TRAP and similar methods are emerging techniques alternative to fluorescence-based cell sorting and are particularly useful for two reasons. The first is that they allow us to enrich minute amounts of RNAs, such as those originating from rare cell populations or even from specific cellular compartments such as the axon terminals [48]—though *cDNA amplification* may be necessary, according to the amount of total mRNA extracted. The second is that TRAP specifically targets the ribosome-bound transcriptome (also called **translatome**) and therefore identifies actively-translated RNAs. This is particularly relevant since proteins, and not mRNAs, are the biologically-active molecules. Moreover mRNAs are subject to translational control. Given that proteomics techniques have less depth and are not as standardised and easily available as RNA-seq, analysis of the translatome is a step towards an analysis of the biologically-active molecules. In the following Sections, we will describe two applications of this method.

In 2014, Ekstrand and colleagues published a method that combines the retrograde tracing of the pre-synaptic afferent neurons to a given brain region with ribosome immunoprecipitation, thereby allowing the characterisation of the transcriptome of a specific population of projecting neurons [14]. The specific molecular tagging of the afferents can be achieved using a *retrograde tracing virus*[5] which encodes the construct for the fluorescent protein EGFP (Figure 9.1A). This allows scientists to express a kind of molecular bait specifically in the population of pre-synaptic neurons which are afferent to a specific brain region, in the case of this study the Nucleus Accumbens.

In order to capture the ribosomes of the afferent neurons, a strategy was devised to cross-link the ribosomes with the retrogradely expressed GFP and then to purify the GFP-ribosome complex (Figure 9.1B). Ekstrand and colleagues thus fused a camelid nanobody (a small protein acting as an antibody) that specifically binds GFP to the ribosomal large subunit protein L10a. This fusion construct can be expressed in all brain neurons—using, for example, the human synapsin promoter—or can be expressed in chemically-defined neuronal types, such as the dopaminergic neurons, by using the dopamine transporter promoter. After the

[5] Some viruses with neuronal tropism, such as the rabies virus, infect the neural terminals at the entry point—naturally a bite or a scratch, experimentally the point of injection—then uses the *retrograde transport mechanisms* of the neuron to reach the soma. Here they hijack the transcriptional and translational machinery of the host in order to replicate, so the spread of the infection proceeds from the post-synaptic neuron to the pre-synaptic one. Moreover, it is possible to modify the retrograde virus so that it is able to infect effectively only once: this makes it possible to trace with precision the afferent neurons to a given region.

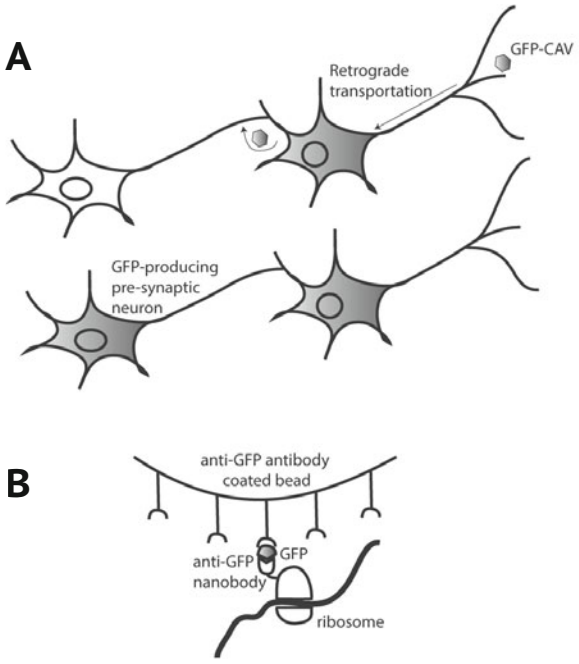


Figure 9.1. A) The mechanism of infection of some viruses, such as the rabies virus or the *canine adenovirus type 2* (CAV, used in [14]), can be used to trace the pre-synaptic neuronal afferents (left neuron) of the neurons in a brain region (right neuron) through the expression of a fluorescent marker. For more information see text. B) Immunoprecipitation strategy employed in [14]: the candidate pre-synaptic neurons specifically express the ribosomal subunit L10a fused to an anti-GFP camelid nanobody, the retro-infected neural afferents start producing GFP, which is then avidly bound to the ribosomes through the nanobody. The GFP-ribosome complex can then be immunoprecipitated using a standard bead coated with anti-GFP antibodies.

injection of the viral retromarker CAV-GFP, they then focused on two populations of neurons: one that is composed of dopaminergic afferents from the Ventral Tegmental Area (VTA) to the Nucleus Accumbens and another population of melanin-concentrating hormone (MCH) neurons from the Lateral Hypothalamus (LH). The GFP-ribosome complexes were immunoprecipitated from brain sections of the VTA/LH, the RNA extracted, sequenced, and the dataset analysed in order to provide a comprehensive molecular characterisation of the afferents from these regions. Some of the novel molecular markers identified, such as the expression of the protein p11 in a subset of hypocretin-producing cells in the LH, were then validated using confocal imaging of brain slices from the region of interest.

9.2.1. RNA-seq can be used to study local translation in developing and mature visual circuits

The development of the visual system is a textbook example of how a neuronal circuit is constructed in a multistep process. The axons of the retinal neurons that will relay the processed visual information to the higher brain centres (ganglion cells, RGC) must reach the first information relay station in the superior colliculus (SC) and, starting from a coarse spatial organisation, must produce a fine topographic projection. This process can be roughly described with a four-step model (Figure 9.2A).

Axon growth The axons elongate and orientate their growth towards the correct region of the colliculus guided by molecular cues.

Circuit formation The axons of the optic nerve that reached their termination zone in the SC, start *branching* and form synapses (*synaptogenesis*) with the local target neurons.

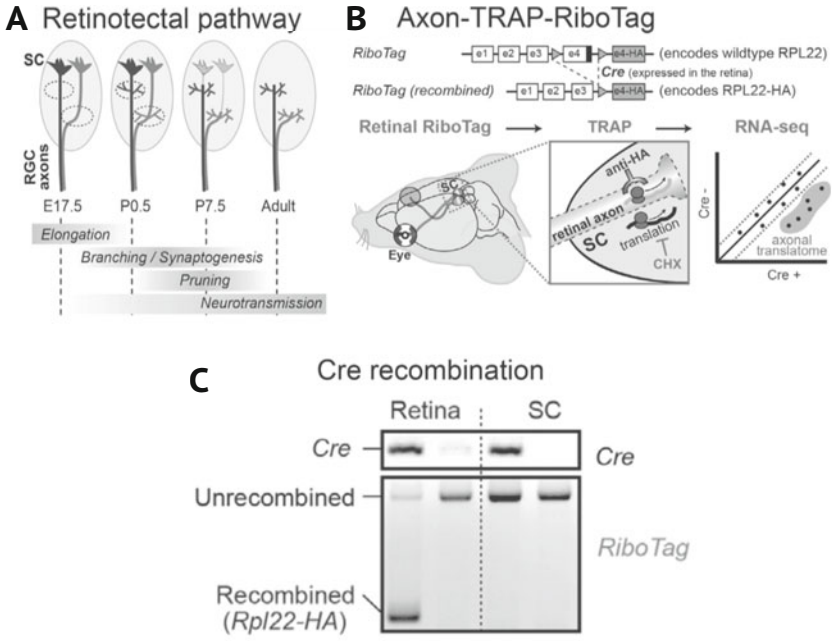


Figure 9.2. Schematic representations of A) the processes that occur during the development of the retinotectal/retino-collicular circuit (see text for further information) and B) the Axon-TRAP strategy in order to express haemagglutinin-(HA-)tagged ribosomes in the retina ganglion cells. C) PCR on genomic DNA that demonstrates the specificity of the recombination of the HA-tagged ribosome in the retina. *Figure adapted from* [48] *under a CC-BY license (http://creativecommons.org/licenses/by/4.0/).*

Circuit maturation The early process of synapse formation is imprecise and redundant. This means that 1) the early wiring is not topologically accurate—that is, there are axon branches which are initially found outside their presumptive adult termination zone—and 2) more synapses are formed than the amount that is necessary for visual processing. During its maturation, the circuit undergoes extensive activity-dependent *synaptic plasticity* and *pruning* (*i.e.* the deletion of non functional branches or synapses).

Maintenance Neurons are post-mitotic cells and, in the mammalian visual system, there is no adult neurogenesis in the retina. So, once the retino-collicular projections have been formed, these connections must be maintained over the entire lifespan and their *synaptic transmission* is strictly regulated. This implies a continuous turnover of synaptic, structural and metabolic proteins localised to the axon terminal.

It is known that in neurons, which are highly compartmentalised cells, the phenomenon of local translation is fundamental for many cellular functions, from neuritic development to synaptic plasticity. Ribosomes and mRNAs are localised in the axon growth cone and at the mature synaptic terminus. Moreover, a number of mRNAs that are locally translated in the axon have been identified in different structures and conditions. However, a global characterisation of axonal mRNAs that are locally translated is missing. The analysis of the axonal translatome at different times during the development of retinofugal connections is of particular interest, because changes in gene expression can be related to well-characterised anatomical and physiological transformations that have a functional relevance.

Shigeoka and colleagues [48] developed a TRAP strategy to target the question of how the local translation process in the RGC axons evolves during development and in adulthood (axonTRAP). Similarly to the general scheme that we presented in the previous section, the axonTRAP method requires the expression of an haemagglutinin-tagged ribosome in the cells of interest. Taking advantage of the large distance between the RGC body and the superior colliculus, RGC terminals can be isolated physically, and the ribosomes immunoprecipitated using anti-HA antibodies cannot be contaminated by somatic ribosomes of the RGCs. As can be seen from Figure 9.2B, the method which has been used to specifically tag RGC cells is Cre-LoxP recombination. In the knock-in mouse model that was used, the gene coding for the ribosomal large subunit protein L22 carries the last 'wild type' exon of the coding sequence (exon 4) floxed and collated with a sequence coding for exon 4 fused with an

HA tag. In the absence of the Cre recombinase, the endogenous exon 4 is transcribed and the translation of L22 protein stops at the native stop sequence; in presence of Cre, whose transcription can be controlled using specific promoters[6], the 'wild type' exon 4 is deleted from the genome, and the fusion HA-L22 ribosomal proteins are produced only in the region of Cre expression (see Figure 9.2C for the demonstration of specificity via PCR on genomic DNA[7]).

One of the major sources of noise in the axonTRAP protocol is the nonspecific binding of mRNA from the SC cells to the component of the immunoprecipitation system. In fact, axon ribosomes are extracted from a much larger pool of non-tagged ribosomes and RNA species. In order to determine the level of background noise, Shigeoka and coworkers performed a control immunoprecipitation and sequencing on the SC of Cre-negative mice, arguing that this step would filter out all the spurious components from the final signal.

Local translation is a tightly regulated process, and there are multiple molecular mechanisms that can stall or prevent the translation of a ribosome-bound RNA. So a second source of potential noise is the signal that comes from non-actively translating ribosomes. Notice that this signal is considered noise because the object of study is the set of RNAs actively translated (**translatome**) and not the set of species that are available at a given time in the axon (transcriptome). A way to quantify the fraction of the total axonTRAP signal that is due to ribosome-bound non-translating RNA is to perform a **ribosome run-off assay** followed by immunoprecipitation[8]. As shown in Figure 9.3A, after ribosome run-off assay the majority of the mRNA signal disappears or diminishes while a minor fraction of ribosomes ($\sim$ 15%) are insensitive to elongation in vitro and thus likely to be stalled during translation.

In order to identify how the axonal translatome changes in time during the genesis of the retino-tectal pathway, the authors extracted RNA from tagged axons at four time points: embryonic day 17.5 (E17.5, *elongation*), post-natal day 0.5 (P0.5, *axon branching*), P7.5 (*pruning*), and adult (*maintenance*). The strict spatial division in the retino-tectal path-

[6] For example, Shigeoka and colleagues used a promoter that induced the transient expression of the Cre recombinase in the neural progenitors

[7] In [48] the authors have demonstrated the specificity of the expression using TEM with immunogold labelled anti-HA antibodies and anti-HA immunohistochemistry.

[8] A ribosome run-off assay tests translational elongation *in vitro*. Only translationally active ribosomes can perform translation in the test tube until they detach from the mRNA. So, the ribosome immunoprecipitate should contain only the mRNAs that are in stalled ribosomes.

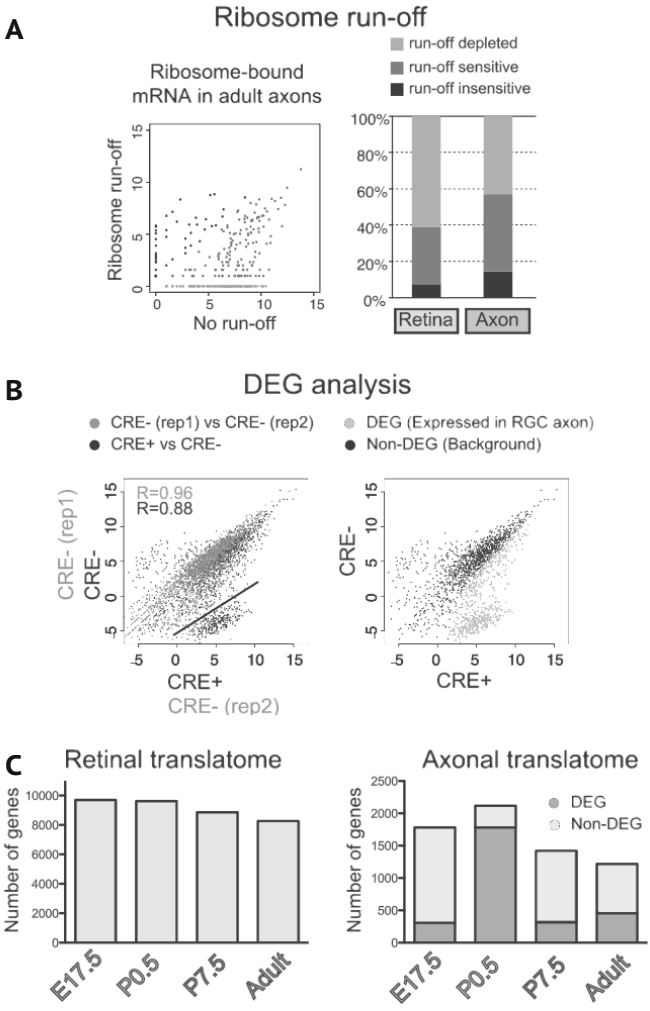


Figure 9.3. A) Ribosome run-off assay allows us to quantify the fraction of mRNAs that are likely to be present in the axon but whose translation is stalled. B) A scatter plot of the gene expression in two biological replicates of the Cre-negative control immunoprecipitation shows high consistency in the detected mRNAs, while a comparison between Cre-positive and Cre-negative samples shows a population of axon-enriched mRNAs (left panel). The right panel shows the DEGs (grey) between the RGC axons and the spurious background (black). C) While the soma shows a consistent translatome during the development, a high fraction of the (smaller) axonal translatome consists of DEGs. *Figure adapted from* [48] *under a CC-BY license (http://creativecommons.org/licenses/by/4.0/).*

way makes it possible to specifically isolate the axons from the cell bodies of RGCs—the soma can be used too as an internal control for the axonal specificity of the identified translatome. The mRNAs were extracted, sequenced, and the DEGs were computed between the ribosome-captured RNAs and the negative control of the $Cre^{-/-}$ mice (Figure 9.3B). As can be seen from Figure 9.3C, there is a peak in the complexity of the translatome (*i.e.* the number of axon-enriched transcripts) during the early steps of wiring of the neural circuit (P0.5), then it decreases during maturation and in adult life. In the case of the somatic translatome, on the other hand, there is no detectable change in complexity over the studied time periods.

From this axonTRAP experiment, it was possible to confirm the hypothesis that local translation occurs in axons even in adult age, due to the presence of translationally-active ribosomes and of DEGs with respect to the soma. Moreover, the analysis of the axonal translatome across all stages showed a relatively small set of mRNAs translated locally at every time point ($\sim$ 27%). When contrasted with the observation that the complexity of the somatic translatome in RGCs is not changing between E17.5 and adulthood, these data support a model of **active regulation** of the local translatome during axon development and maintenance.

A Gene Ontology (GO) analysis of the more represented terms of the axonal translatome at different times confirmed that axonTRAP is highly specific. In fact, there is an extremely low presence of terms related to specifically somatic functions such as 'chromosome', 'spliceosome', 'nuclear lumen', and an enrichment in the terms related to the synaptic function, such as 'synapse', 'actin cytoskeleton', and (obviously) 'axon'. When the GO analysis is applied to the axonal translatomes at different time points, perhaps unsurprisingly, the most represented and statistically significant terms are related to the main process undergoing at that developmental stage, like 'neuron projection morphogenesis' at E17.5, 'neuron projection development' at P0.5, 'neuron remodelling' at P7.5, and 'regulation of synaptic transmission, GABAergic' during adulthood. Interestingly, the most enriched KEGG pathways in the dataset are 'Parkinson's disease' (E17.5, P0.5, P7.5), 'Huntington's disease' (E17.5, P0.5), 'Oxidative phosphorylation' (P0.5, adult), and 'ribosome' (P0.5). As shown in Figure 9.4A, proteins that are of fundamental importance in neurodegenerative disorders, such as huntingtin, prion protein, amyloid β precursor protein, and tau are locally translated in the axon.

Once the set of DEGs at different time points is known, it is possible to evaluate whether the mRNAs which are known to be regulated from spe-

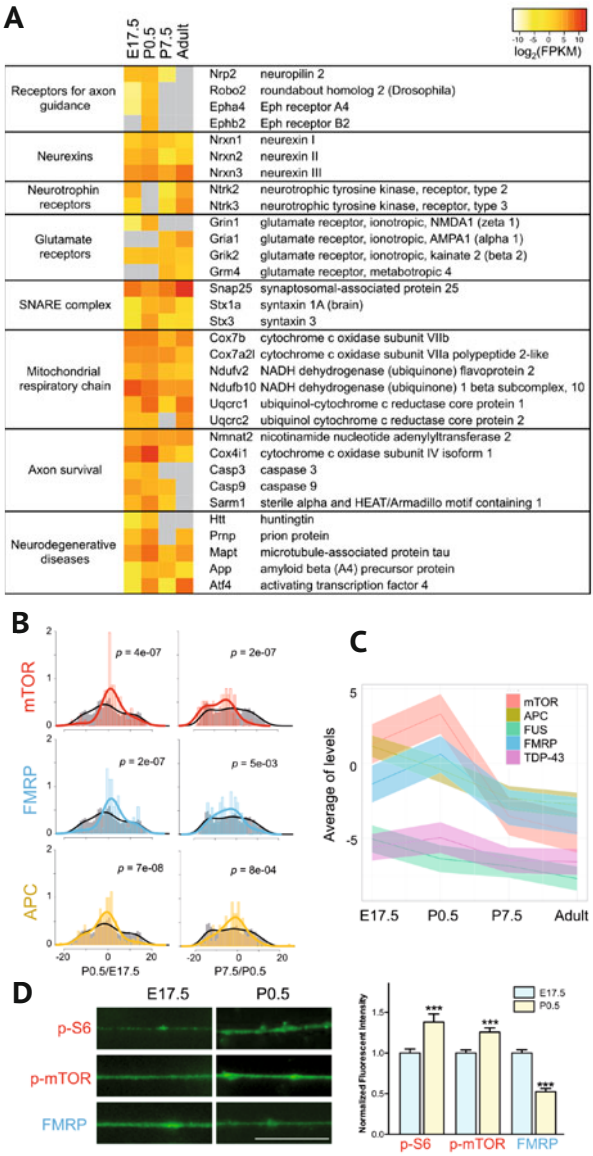


Figure 9.4. A) Fold change of the translation profiles of key proteins in various synaptic pathways during axon development. B) Change of distribution of the fold-change of the known mRNA targets of three translational regulators (mTOR, FMRP, APC) in the transition to early branching (P0.5/E17.5) and to the circuit maturation stage (P7.5/P0.5). C) Average of levels of the target proteins of different translation factors in the different axon developmental stages. D) Fluorescence profile of tectal axons at two different developmental stages for phosphorylation-activated ribosomes (p-S6), mTOR (p-mTOR), and for the presence of RNA binding protein FMRP. See main text for further information. *Figure adapted from [48] under a CC-BY license (http://creativecommons.org/licenses/by/4.0/).*

cific translational factors (such as mTORC1, FMRP, APC, TDP-43, and FUS)[9] show different dynamics from E17.5 to adulthood. Figure 9.4B–D show how the mRNAs that are regulated by different RNA binding proteins are differentially translated at the analysed time points from different perspectives. Figure 9.4B for example, plots the fold-change in the FPKM values of the axonal translatome in two consecutive stages A and B

$$x_i = \log_2 \left(\frac{\mathrm{FPKM}_A(i)}{\mathrm{FPKM}_B(i)} \right) \tag{9.1}$$

the cumulative values of x_i in the translatome form a density plot whose peak represent the trend in translational regulation from developmental stage A (*e.g.*, P0.5) and stage B (*e.g.*, E17.5). It can thus be seen that the targets of both FMRP and mTORC1 show a peak of translational regulation at P0.5, while for APC the peak of translation is during the elongation stage and steadily decreases thereafter. From Figure 9.4C, where the average FPKM value for the set of genes regulated by a given protein is plotted at any given time point, one should also appreciate that FUS and TDP-43 do not have a strong developmental effect on the local translatome. Indeed, a similar analysis can be performed with any selected set of mRNAs: for example, the authors found that the axonal translation of the targets of the miRNA miR-1 decreases with time. As an experimental validation of their finding, they analysed the fluorescence profiles in axons at two different developmental stages—E17.5 and P0.5—of phosphorylated mTOR (p-mTOR), phosphorylated S6 (p-S6, see also page 142), and FMRP immunohistochemistry (Figure 9.4D). The normalised quantification shows that the fluorescence of p-S6 and p-mTOR increases, thus reflecting an increase in the general local translation and in mTORC-mediated translation. The fluorescence of FMRP decreases with time, as FMRP is a translation inhibitor; this is consistent with the observed increase in the translation of its targets at P0.5.

All the results obtained up to this point could also have been gathered using a different strategy of transcriptional profiling, such as cDNA microarrays. A decisive advantage of RNA-seq, however, is that the unbiased nature of the sequencing can reveal the existence of previously unknown transcripts—such as novel splicing variants (see Figure 9.5A)

[9] The *mammalian target of rapamicin complex 1* (mTORC1) is known to be a regulator of activity-dependent local translation, *fragile X mental retardation protein* (FMRP) is a mRNA binding protein associated with the Fragile X syndrome, and *adenomatous polyposis coli* has been recently associated with the regulation of the microtubule assembly in the growing axon. The mRNA binding proteins TDP-43 and FUS are both associated to the axon-related neurodegenerative disorder Amyotrophic Lateral Sclerosis (ALS).

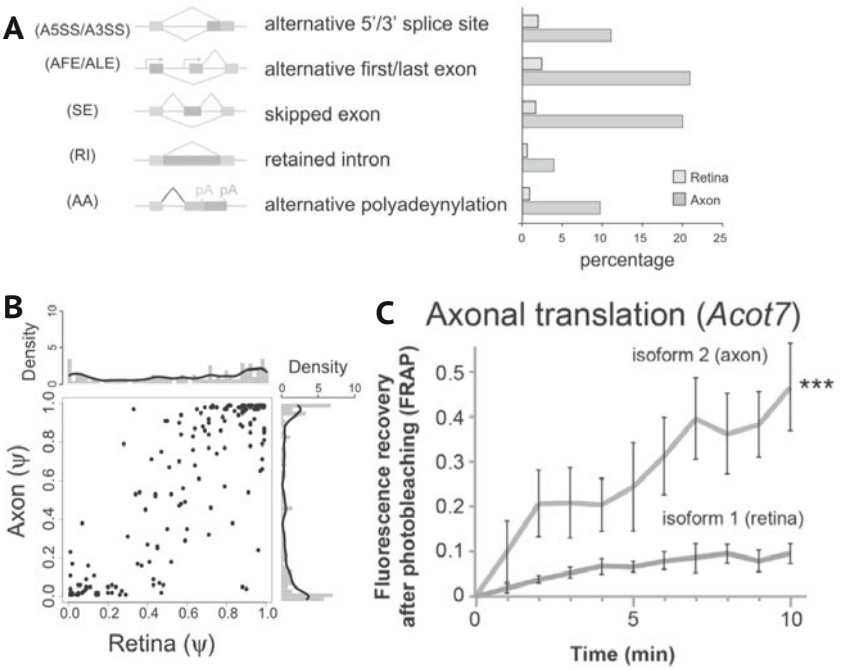


Figure 9.5. A) Schematic representation of the different splicing variants and their relative abundance as found in the local translatome and in the somatic (retina). B) Density plots of the 'percentage spliced in' (Ψ) coefficient for the transcripts in the translatome that show two alternative splicing variants. C) The use of the 5′-UTR of either the retinal or axonal splicing variants of *Acot7* (a gene computed in Ψ) coupled with a fluorescent reporter induces a fast FRAP signal in the axon terminus, implying that the local translocation of the axon mRNAs also depends on differential splicing of the RNA species. *Figure adapted from* [48] *under a CC-BY license (http://creativecommons.org/licenses/by/4.0/).*

in the local translatome. Shigeoka and coworkers decided to address the problem as to whether the distribution of the various mRNA splicing variants differs in the axonal as compared to the soma. The answer is that not only there is a high representation of different splicing variants[10] in the axonal translatome (Figure 9.5A), but the non-coding regions of these variants are likely to play a functional role in the translocation of the different constructs (Figure 9.5C, see below). In order to obtain a quantitative assessment of differential splicing in the axons, the authors

[10] Moreover and unexpectedly, the authors detected some potential back-splicing events in three genes, which would indicate the presence of **ribosome-bound circular RNAs** (see also Section 8.5 at page 134).

focused on a set of splicing isoform couples, of which at least one member is present both in the axonal and somatic translatomes. It is then easy to compute the relative proportions of the two isoforms in either translatome, expressed as a 'percentage spliced in' (Ψ) coefficient. In the retina (*i.e.* in the somatic compartment) there is an even distribution of Ψ coefficients from the different events, while in the axons the density distribution of the fractions is polarised towards 0% and 100% (Figure 9.5B)—meaning that only one of the alternative transcripts is present in the axon. Eventually, from the different splicing isoforms which are specific to the axon translatome, the authors were able to compute some sequence **consensus motives** that are putatively linked to the specific axonal translocation. The functional relevance of these motives for axonal translation could be proven by fusing the 3′-UTR or 5′-UTR from the axonal or the somatic splicing variants of some genes to a fluorescent reporter (myr-d2EGFP) with very slow diffusion[11]. When the fluorescence in the axon is photobleached, axon-specific phenomena induce a faster fluorescence recovery (FRAP signal) as compared to the construct with somatically-enriched UTRs. Since the fluorescent construct cannot diffuse from the soma, the recovery of fluorescence in the axon demonstrates specific mRNA translocation and local translation.

9.3. Use of FACS to obtain the transcriptome at the cell-population scale

Another strategy that could be used to derive transcriptomic data from specific cell populations is to apply a procedure of **whole-cell fluorescence labelling**—either by genetic methods, tract-tracing techniques, or immunostaining with fluorophore-conjugated cell-specific antibodies—followed by **Fluorescence-activated cell sorting** (FACS). Cell sorting separates one (or multiple) cell species of interest according to their fluorescence and light scattering profile. This approach, combined with microarray analysis, was applied by Sugino and colleagues [53] to produce a dataset of brain cell-specific transcriptomes[12]. The insights into the biology of neurons derived from this dataset will be described in the next section.

[11] The myr-d2EGFP fluorescent reporter is post-translationally modified with a 14 carbon unsaturated fatty acid (myristic acid), that induces a propensity to interact with the cell membrane. Moreover, this protein carries a destabilising tag that reduces its half-life to 2h.

[12] Each distinct cell type (distinguished by the area of isolation and the transgenic line used) was moreover associated with an electrophysiological activity profile.

9.3.1. Network analysis of cell-population transcriptomes and novel cytological properties of neurons

A systems biology approach to high-throughput datasets makes it possible not only to provide biological interpretations for global expression patterns, but also to produce biological hypotheses that can be tested experimentally at a molecular and/or cellular level. Winden and colleagues [59] applied the WGCNA approach to a transcriptomic dataset of 12 specific mouse neuronal subtypes. This approach identified 13 gene co-expression modules (see also Table 9.1). They also showed that:

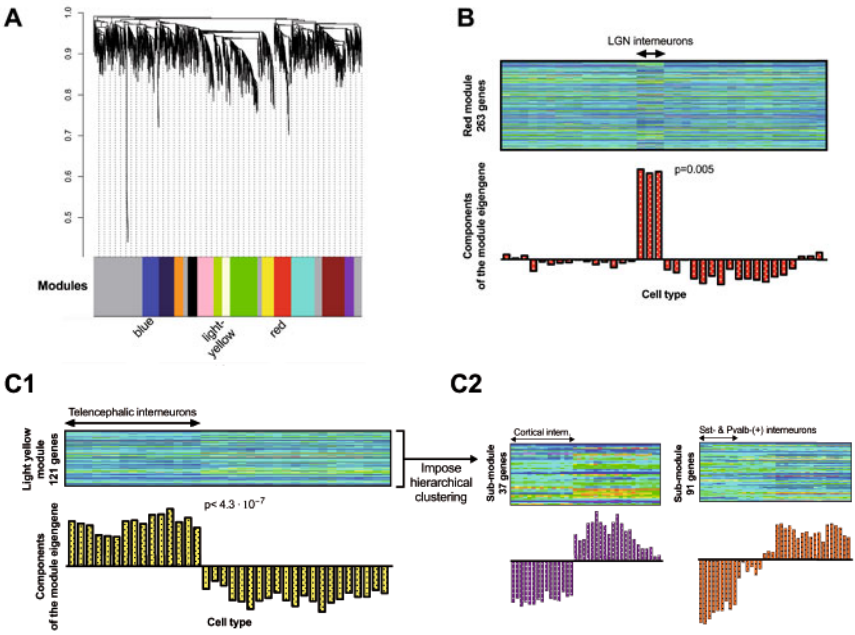


Figure 9.6. A) Dendrogram and WGCNA module division of the microarray dataset from various mouse brain cellular subtypes. B) As an example, the expression profiles of the different cell population samples of the 263 gene members of the 'red' module is plotted as a heatmap. It is easy to notice how the module genes are strongly coexpressed in specific cell types (the Lateral Geniculate Nucleus interneurons in this case). This can also be confirmed from the histogram showing the components of the eigengene of the module in the different cell populations. C) Modules with generic cell specificity—such as the 'light yellow' one, whose eigengene is expressed positively in the samples derived from telencephalic interneurons (C1)—can be further divided in coexpression sub-modules that show an eigenmodule representation of specific interneuron types (C2). *Figure generated from raw data in* [59].

Module	Top hubs	GO	Cell types	$\rho_{<\text{firing}>}$
1 black	Diras1, Plk2, Mast3	Cellular protein metabolism	*Amygdala and hippocampal pyramidal neurons*	−0.66
2 blue	Hadhb, Gas6, Ppp1cc	Mitochondria	*Sst- and Pvalb-positive interneurons; Somatosensory layer V pyramidal neurons*	0.60
4 green	Crym, Dnclc1, Klhl2	Synaptic transmission	*Glutamatergic neurons*	−0.34
6 light yellow	Arx, Dlx1, Nxph1	Signalling/ Signal transduction	*Telencephalic interneurons*	p>0.05
11 red	Tacr3, Lhx1, Sdsl	Lipid biosynthesis	*LGN interneurons*	p>0.05
12 turquoise	Uqcrfs1, Atp5b, Idh3a	Mitochondria	*Pvalb-positive interneurons; Somatosensory layer V pyramidal neurons*	p>0.05

Table 9.1. Characterisation of selected WGCNA modules from different classes of mouse neurons with eigengene (hub), gene ontology (GO), Cell type enrichment, and correlation of the module with the average firing rate ($\rho_{<\text{firing}>}$). Table generated from raw data in [59].

- some modules are strongly correlated either with molecular traits (*e.g.*, glutamatergic vs GABAergic neurons), or with regional/developmental origin (*e.g.*, telencephalic vs diencephalic) of the cells, thus pointing out important *global differences in the transcriptional profiles of distinct neuronal classes*;
- *the eigengene/eigenhubs of each module associate the modules to specific ontology*: for example, the module with higher expression in lateral geniculate nucleus (LGN) interneurons also shows enrichment for genes within the 'Lipid biosynthesis' GO term and shows the LIM Homeobox 1 (*Lhx1*) and Tachykinin Receptor 3 (*Tacr3*) genes as hubs;
- a less specific module, associated with the various 'telencephalic interneuron' classes, could be further divided into sub-modules that show the regional and developmental origin of the different interneuronal classes;
- surprisingly, the term 'mitochondrion' is enriched in two distinct modules—but only one shows correlation of the eigengene with the mean firing rate of the different neuronal classes.

The last finding is particularly interesting. It is reasonable to couple a high activity rate and a high energy requirement (*i.e.* mitochondrial activity), however the existence of two distinct modules of which only one is associated with the activity of the neuron lends credence to the hypothesis that neurons contain two different populations of mitochondria; these

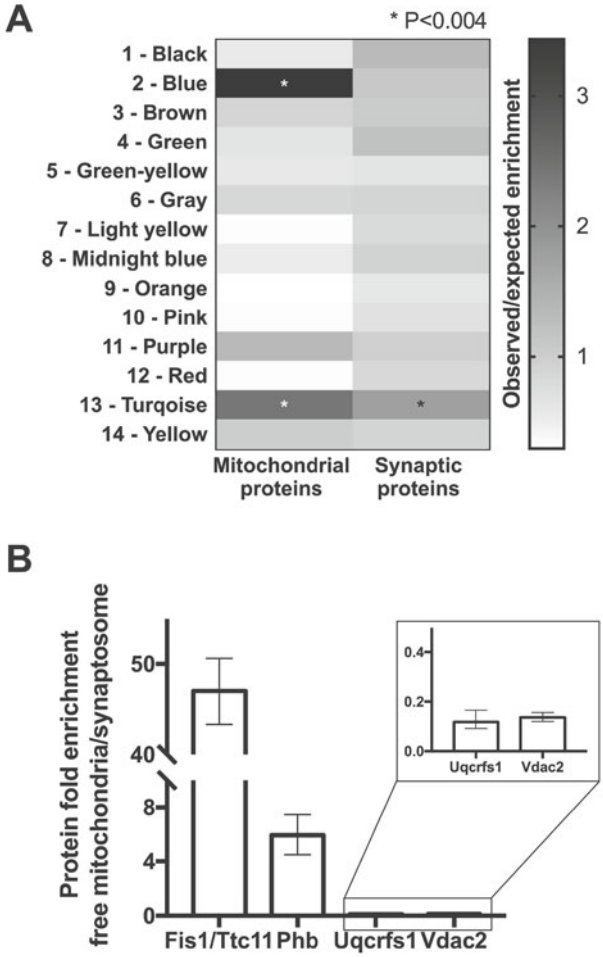


Figure 9.7. A) By analyzing the membership of proteins from the synaptic and mitochondrial proteomes in the modules, it becomes evident that two modules are highly enriched in mitochondrial proteins (#2 and #12—see also Table 9.1), but only one is also significantly enriched for synaptic proteins ($p<0.004$, Bonferroni correction). B) Quantification of western blots of proteins extracted after subcellular fractionation provides information on the relative fold change between the 'free mitochondrial' fraction—*i.e.* somatic, Cyc(+) and Syp(–)—and synaptosomal fraction, Cyc(+) and Syp(+), for hub proteins of the two modules. Fis1/Ttc11, and Phb are hub proteins of the 'blue' module, while Uqcrfs1 and Vdac2 are hub proteins of the 'turquoise' module. *Figure generated from raw data in* [59].

might differ in their physiology and subcellular localisation, possibly one confined to the soma and the other enriched in dendrites/synapses[13].

To validate the hypothesis that the distinct mitochondria-enriched modules are associated with different compartments in the neurons, Winden and colleagues performed an analysis on the membership enrichment of synaptic proteins and of mitochondrial proteins in the different modules (Figure 9.7A). They demonstrated that only one of the two mitochondria-associated modules is also enriched in synaptic proteins. Further experimental validation—using western blot (Figure 9.7B) and immunocytochemistry—shows that the hub proteins in the 'somatic' mitochondria module are depleted in the dendrites, while the opposite is true for the proteins in the 'synaptic' module. In conclusion, neuronal mitochondria present a peculiar transcriptomic fingerprint according to their localisation, and it is likely that they are specialised on a molecular level for their local activity. This is an example of how a quantitative approach can deliver novel biological hypotheses amenable to experimental validation.

9.4. Single-cell RNA-seq strategies

The strategies that were presented in the previous sections produce transcriptome expression profiles with resolution at the cell population level. However, these approaches show two major limitations:

1. TRAP techniques are not general and targeting a specific cell population requires a dedicated promoter or marker; so, targeting some populations may first require the identification of a suitable molecular tag;
2. isolation from a population of cells is still a *bulk technique* and the potential biological variability between single cells in what is assumed to be a homogeneous population is lost through averaging. This also implies that, if substructures exist within the population, these are not detectable.

These two caveats can be overcome using **single-cell RNA sequencing** platforms, where profiling of cDNA derived (and amplified) from single cells is performed.

The single-cell approach creates some new technical and analytical challenges both in the 'wet lab' and in the subsequent computational anal-

[13] It was already known that soma and dendrites presented local populations of mitochondria, however no global transcriptomic difference had been reported among mitochondria from different subcellular compartments.

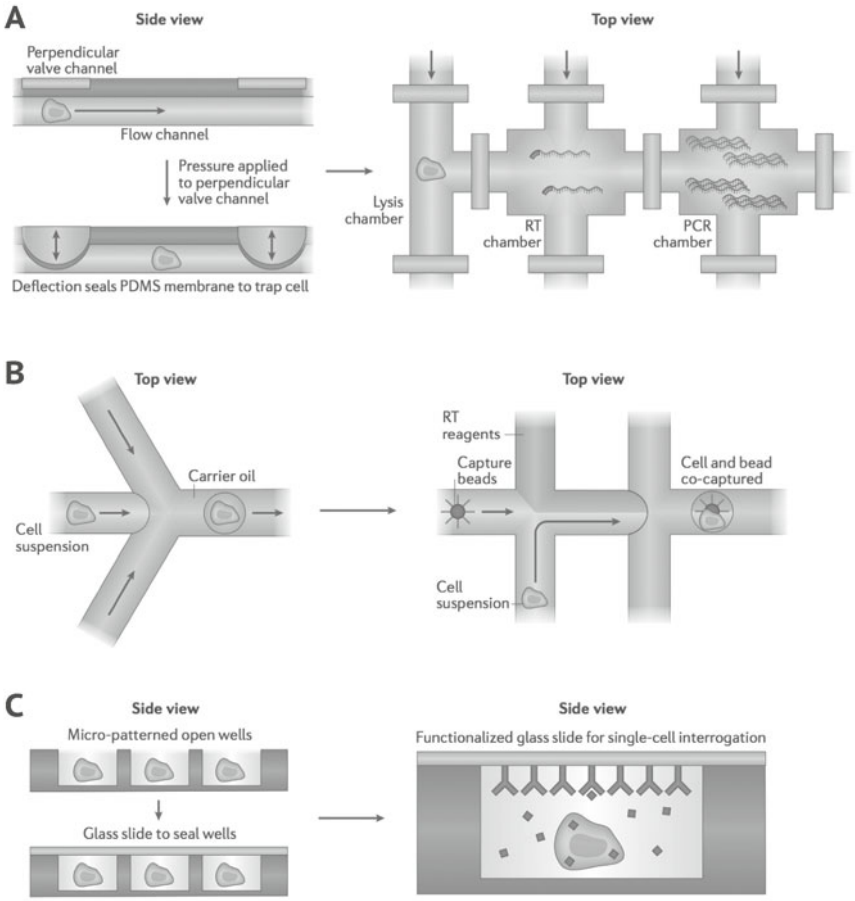


Figure 9.8. Microfluidic devices that are commonly applied to single-cell RNA-seq. A) The combined use of pumps and microfluidic channels forms isolated chambers where the different steps of cell capture and RNA extraction can be performed. These systems allow for tight control of the parameters, however they are more complex to design/produce and are difficult to scale up. B) Devices that produce nanodroplets of aqueous solution encapsulated in an inert lipid solution can be used to capture single cells and process the extraction of their RNA with a really high throughput (up to thousands of cells). The main downside of this technique is the low level of control the user has over the droplets (whose manipulation is essentially stochastic). C) Chips with printed micro-wells can be loaded with low-density solutions of cells using the force of gravity, then closed and manipulated with the reagents of interest. This strategy has a lower throughput than the droplets and gives less control than the valve-assisted, it is thus optimal for minimalistic protocols. *Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics* [41], © (2017).

ysis as compared to conventional 'bulk transcriptomics'. Firstly, the cells to be studied must be dissociated (this is tricky in the case of adult brain tissues) and isolated so that no mixing of multiple cells in a single unit occurs. This can be achieved through the use of different *microfluidic chips*, such as microfluidic channels coupled with valves (Figure 9.8A), lipid-water nanodroplets (Figure 9.8B), or micro/nanowell chips (Figure 9.8C). Secondly, the cell mRNA must be extracted with minimum degradation, captured, barcoded (*i.e.* a cell-unique sequence must be attached to the fragment to be sequenced in order to deconvolute the sequencing data and assign sequences to individual cells), reverse-transcribed, and amplified in order to obtain a library that can be sequenced. Spike-ins and unique molecular identifiers are normally used to test for sensitivity and specificity. Finally, since the number of reads obtained is relatively low, normalisation of the data is a serious issue.

9.4.1. A day in single-cell RNA sequencing

Providing a detailed description of single-cell RNA-seq protocols is beyond the aim of this chapter and is also not meaningful in such a rapidly evolving field (for further reference see the review from Prakadan and colleagues [41]). In this Section, however, we will describe one specific example of single-cell sequencing, as it is shown in Figure 9.9. The presented protocol is called InDrop [26] and can be used to derive a 3′-end counting of the transcriptome (see also note 14) potentially from thousands of single cells.

Let's consider a microdroplet microfluidic device, where the combination of an aqueous solution channel—whose solution is derived from the mixture of three converging channels, one for lysis/reverse-transcription

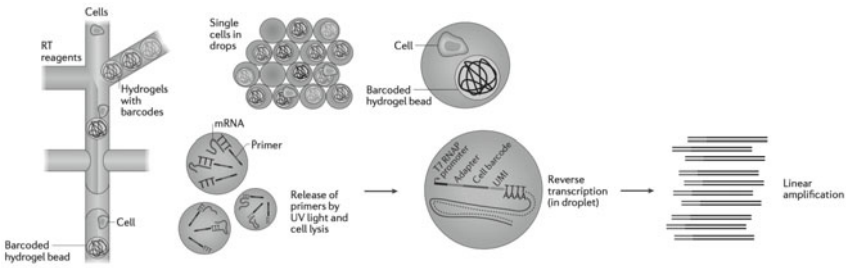


Figure 9.9. Example of a single-cell RNA-seq protocol that applies microdroplets (InDrop, [26]) to prepare barcoded samples for next generation RNA-seq. *Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics* [41], © (2017).

reagents, one for the isolated cells, and one for hydrogels of barcodes[14] — and a perpendicular channel of an inert lipid solution that makes the rapid generation of a series of droplets possible. A certain fraction of these droplets will not be 'productive' (*i.e.* will not contain both the barcode hydrogel and a single-cell), however, due to the design of the microfluidic chip, about the 90% of the productive droplets will contain a single barcode hydrogel and a single cell and are thus suited for single-cell specific RNA extraction and barcoding. After the lysis of the captured cell, its mRNA is freely available in the droplet: the use of a UV-light stimulus triggers the dissociation of the barcoded hydrogels, thus making the 3′-end primers available for association with the mRNAs and reverse-transcription. The droplets contain now cDNA that present a single cell-specific barcode and can be used to prepare a next-generation sequencing library.

9.4.2. A closer view on the taxonomy of mouse primary visual cortex as revealed by single-cell RNA-seq

The nervous system is composed of a plethora of different cell types, from neurons of different neurochemical classes (glutamatergic, GABAergic, secreting different neuropeptides, and so on), morphology and functions, to glial cells. An exhaustive catalogue of the cell diversity of the brain is far from being available and it is not known what the smallest partition of different nervous system cells with a biological meaning is. Performing extensive single-cell RNA-sequencing on different neural tissues is an effective experimental strategy to answer these questions.

Tasic, Menon and colleagues [54] focused on the reconstruction of a well-studied and easily accessible cortical region: the (adult male) mouse primary visual (V1) cortex, whose architecture is not columnar, unlike the human V1, but mixed/microcolumnar. In order to select both abundant and rare cells, the authors exploited the availability of a variety of transgenic mouse lines where a fluorescent reporter tdTomato is expressed by a defined cell type[15]. The V1 cortical samples were then microdissected,

[14] The barcoded hydrogels are microdroplets made of a photo-sensitive polymer that is covalently linked to reverse-transcription primers that carry a specific barcode sequence. These primers also present a poly-T tail that specifically targets the poly-A tail of mature mRNAs. This also means that the sequenced reads will map on the exon corresponding to the 3′-UTR of the gene (3′-end counting), so each captured gene is counted only in a digital way, without distinctions between RNA isoforms.

[15] In practical terms, this is obtained by crossing a line expressing the recombinase Cre in specific subsets of cortical cells with a transgenic line that carry a 'floxed' construct of the fluorescent reporter, which can thus be expressed only in the Cre-expressing cells. In order to get more specific cell populations, the authors also used a fluorescent reporter with nested flanking of *loxP* (Cre re-

treated with a protease mix (pronase) that permits the production of a single-cell suspension through serial trituration of the tissue, and single cells were then isolated using FACS.

After FACS, the mRNA was extracted, reference RNAs were added as a control *spike-in*, the mixture was reverse transcribed using special primers that flank the 5′- and 3′-ends of the cDNA with a stretch of identical sequences, and amplified using multiple cycles of PCR. This process results in whole transcriptome amplification. From the cDNA, single-cell tagged libraries were generated and sequenced with a minimum depth of $\sim 3.8 \times 10^6$ reads.

The single cell RNA-seq data were subsequently analysed according to a combined iterative PCA (iPCA) and iterative WGCNA (iWGCNA) pipeline. The steps of iPCA are:

1. identify the genes with variance larger[16] than the technical noise (using the variance of reads derived from the *spike-in* RNA as reference, see [7]). This procedure results in a data matrix that associates to each cell a vector of values corresponding to the counts of the genes selected based on the chosen CV threshold.
2. The matrix is log-transformed and z-normalised (see page 65), then the PCs are computed (see Section 5.3 at page 70). The associated real-part ordered list of eigenvalues of the matrix (the *spectrum*) is then likely to show a shoulder, that is when the absolute value of the real part of the eigenvalues starts 'dropping'. A reasonable choice in order to perform the subsequent clustering is to pick the principal components up to the spectrum shoulder.
3. Using the selected components, generate a distance matrix, where each point is determined as a weighted Euclidean distance (see page 62) between two cells in the PCA-projected space; the weight of each principal component is the real part of its corresponding eigenvalue.
4. Perform a hierarchical clustering of the cells (see Section 5.2.1 at page 61) and then split the cells into two groups according to the top branch of the dendrogram.

combinase) and *FRT* (Flp recombinase), or of *loxP* and *rox* (Dre recombinase) sequences. In this way the combination of a Cre line with a Dre/Flp line would result in the isolation of more specific cell lines—*i.e.* lines that are positive for both promoters, the promoter of Cre and the one of Dre/Flp, instead of a single one.

[16] The threshold for the variance was set to each one of four percentages above the coefficient of variation (CV) associated with the noise: 0%, 25%, 50% or 100% above the noise. The effective threshold was chosen after downstream validation of the P-values for the segregated clusters. That is, the threshold for the step was retroactively chosen as the one which provided more statistically significant clustering after step 5 of iPCA. In general, according to the authors [54], when two thresholds would determine statistically significant clusters, they resulted in identical clustering.

5. Assess the significance of the binary split after computing the p-value. The null hypothesis is that the dataset of the two groups of cells is derived from the same multivariate Gaussian curve (and not from two distinct Gaussian curves).
6. Repeat the procedure using one of the split groups found after the previous step as starting point until one of four *termination criteria* is met: there are no cells with variance greater than the noise, there is no shoulder in the principal component spectrum, none of the thresholds return a statistically significant split (considering $p < 0.01$), or if the starting group of cells at the point 1 is smaller than an arbitrarily small number (in [54] it is set as 4).

The final result is a dendogram similar to those conventionally obtained from hierarchical clustering.

The iWGCNA is applied independently to the starting dataset following a procedure conceptually similar to the one of iPCA. A detailed description of the procedure is beyond the aim of this book.

As previously discussed in Chapter 5, a major drawback of 'hard' clustering methods is that data points (in this case, single cells) can be assigned to only one cluster, creating a problem for cells with intermediate phenotypes. In order to avoid a binary cluster assignment, Tasic, Menon and colleagues applied a machine learning classification method based on a cross-validation algorithm (*random-forest classifier*[17]). In a nutshell, for each pair of clusters, their cell members are pooled and divided into five groups (each group containing 20% of the cells). The classifier is then trained with four out of five groups (80% of the cells assigned) and assigns labels to the remaining group of cells using the learned rules. This procedure is then repeated for each possible partition of the data, so that each cell is reassigned to a cluster. Considering that the partitioning can affect the training of the algorithm and thus the cell assignment, the partition and label assignment procedure was performed a total of ten times. After this, a vector of 10 entries, corresponding to the cluster assigned in different runs of the classifier, is associated with each cell:

core cells of a cluster are those that have been assigned to that cluster in 10/10 runs

[17] In machine learning, a **classifier** is an algorithm that assigns predefined labels to data points after being 'trained'. During training, the algorithm is exposed to a group of points which already have been labelled, for example by a human expert, and is left to change its internal parameters so that it is able to re-assign the same starting label to the data points. In order to evaluate the quality of the training of the algorithm, a *cross-evaluation set* is used. The algorithm is asked to classify a group of points whose label has been previously assigned and the degree of identity of the classification is then evaluated.

intermediate cells have been assigned to both clusters: the strength of membership of an intermediate cell to a cluster can be defined as the frequency of assignment (*e.g.* 7/10).

This procedure to identify 'core' and 'intermediate' cells can be extended to all possible pairs of clusters. The procedure previously described divides the 1679 cells in the dataset into 49 clusters. Remarkably, the vast majority of cells (1424) are core cells, and only 255 cells have been assigned to more than one core cluster.

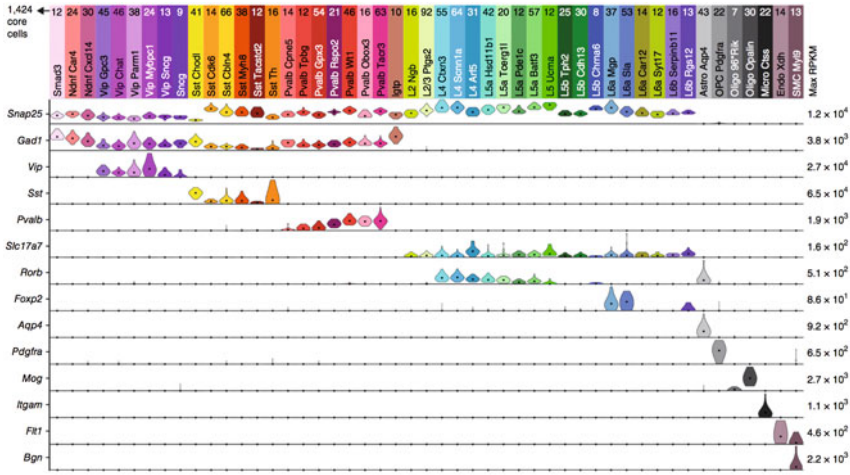


Figure 9.10. Cluster division (*top*) and associated violin plots for known molecular markers. The distribution of the violin plot for each gene is adjusted according to the maximum RPKM (*right*). *Snap25* is a pan-neuronal gene; Gad1 is a pan-GABAergic marker; *Vip*, *Sst*, and *Pvalb* are markers associated with GABAergic cell populations that express the neuropeptides vasointestinal peptide, somatostatin, and parvalbumin; *Slc17a7* is a pan-glutamatergic marker; *Rorb* and *Foxp2* are associated with distinct cortical layers; *Aqp4* is an astrocyte marker, *Pgdfra* is associated with the oligodendrocyte precursor cells, *Mog* is an oligodendrocyte marker, *Itgam* is specific to microglia, *Flt1* is a marker of endothelial cells, and *Bgn* is specific to smooth muscle cells. *Adapted by permission from Macmillan Publishers Ltd: Nature Neuroscience* [54], © (2016).

Once the robustness of the clusters is reaffirmed, each cluster can be screened for the expression of specific **molecular markers** to define the identity of (core) cells of the clusters (Figure 9.10). The protein *Snap25*, which is a key mediator in the exocytosis of synaptic vesicles, can be used to recognise *neuronal clusters*; the cells in these neuronal clusters can then express either *Gad1* (Glutamate Decarboxylase 1), an enzyme which is involved in the production of the neurotransmitter GABA, or *Slc17a7* (also called Vesicular glutamate transporter 1), a fundamental

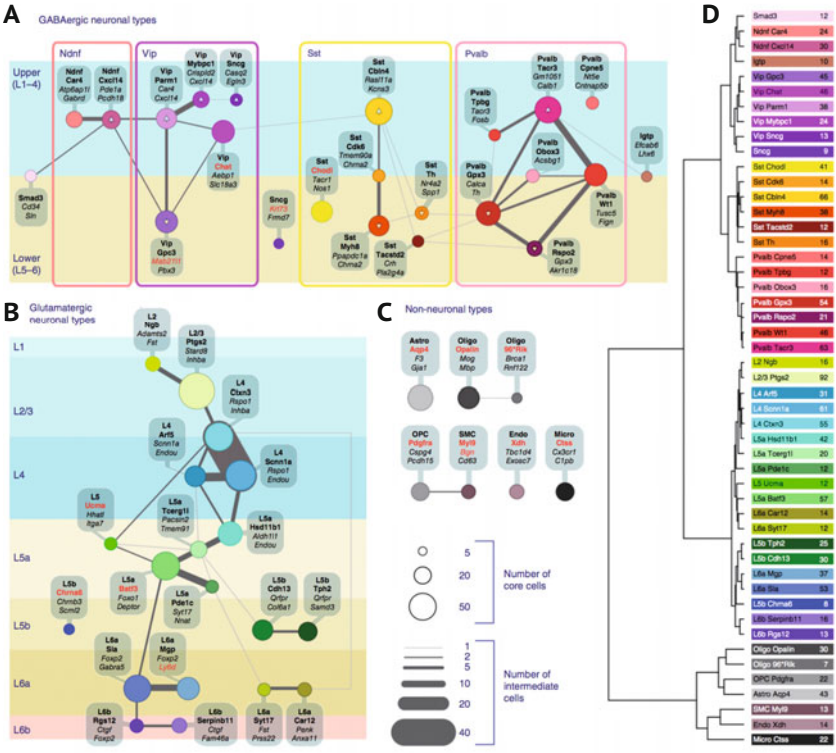


Figure 9.11. Summary of the cell types of the adult mouse visual cortex and their relationships. The size of the circles indicates the number of cells in the core cluster, while the width of the connecting lines indicates the number of intermediate cells. A) Different clusters of GABAergic neurons expressing the same neuropeptides can be predominantly found in distinct layers, such as the upper ones (L1-4) or the lower ones (L5 and L6). B) The relationship between the clusters associated with pyramidal excitatory neurons is strongly dependent on which layers of the cortex their core cells are localised in. C) The glia cell clusters are mostly isolated. D) Hierarchical clustering of the different core cell types. *Adapted by permission from Macmillan Publishers Ltd: Nature Neuroscience* [54], © (2016).

factor in the production of glutamate-containing synaptic vesicles[18]. The main result of this paper is that neuronal diversity in V1 exceeds previous predictions (Figure 9.11).

A total of 19 clusters of glutamatergic neurons and a total of 23 clusters of GABAergic neurons were identified. In particular, populations that

[18] In other words, cells that express the Snap25^{+}/Gad1^{+}/Slc17a7^{-} markers are GABAergic neurons (*i.e.* inhibitory in an adult brain), cells that are Snap25^{+}/Gad1^{-}/Slc17a7^{+} are glutamatergic (*i.e.* excitatory) neurons, and cells that do not express the synaptic vesicle marker Snap25 (Snap25^{-}/Gad1^{-}/Slc17a7^{-}) are non-neuronal cells.

were thought to be homogeneous, such as those inhibitory neurons expressing vasointestinal peptide (Vip) or somatostatin (Sst), were demonstrated to contain six different clusters each. Pyramidal neurons of layer V were subdivided into eight different clusters. In the case of inhibitory neurons, it became clear that cells expressing the same molecular markers, but localised in upper or lower layers, have different molecular fingerprints, as demonstrated by multiple immunohistochemistry. For example, two clusters of VIPergic cells are characterised by the mutually-exclusive expression of synuclein gamma (*Sncg*) or Cysteine Rich Secretory Protein LCCL Domain Containing 2 (*Crispld2*), and these cells are only present in the upper layers (Figure 9.11A). Overall, histological and molecular analysis demonstrated that there is a tendency for the enrichment of cells of different core clusters in one or two adjacent cortical layers, with more widespread intermediate cells bridging different layers/clusters (Figure 9.11A-B). This suggests that the molecular fingerprints of both excitatory and inhibitory interneurons are somehow specific for the localisation and the physiological function in the cortex.

The data obtained by single cell RNA-seq enable us to explore global properties of gene expression at the population scale as well. For example, Tasic, Menon and colleagues show how neurons express more genes than glia cells when sequenced at the same depth. Also, neurons express more genes at low- to medium-levels while glia tend to express a reduced amount of genes, but at higher expression levels. This dataset also allows us to probe differential splicing: at least 320 genes are differentially spliced in cell types at various levels of the presented taxonomy.

An interesting example is the case of the glutamate AMPA receptors *Gria1* and *Gria2*, that show a cell type-specific alternative splicing of two consecutive exons (*flip* and *flop*, as previously reported [50]). The two isoforms encode receptor units with different electrophysiological properties (Figure 9.12).

To conclude the paper, the authors demonstrated that 1) the sequencing data can be integrated with retrograde tracing to define the axonal projection patterns associated with different cell clusters, and 2) it is possible to integrate electrophysiological data into the description of specific cell clusters. In order to isolate the V1 pyramidal neurons that project to specific brain regions, Tasic, Menon and coworkers used retrograde labelling from the thalamus or from the contralateral cortex[19]. The tagged

[19] The labelling was done using a CAV expressing the recombinase Cre (see also Figure 9.1) in a transgenic mouse *Ai14* carrying a the red fluorescent protein tdTomato including a floxed STOP cassette. In presence of the recombinase Cre, the STOP cassette is excised from the genome and the

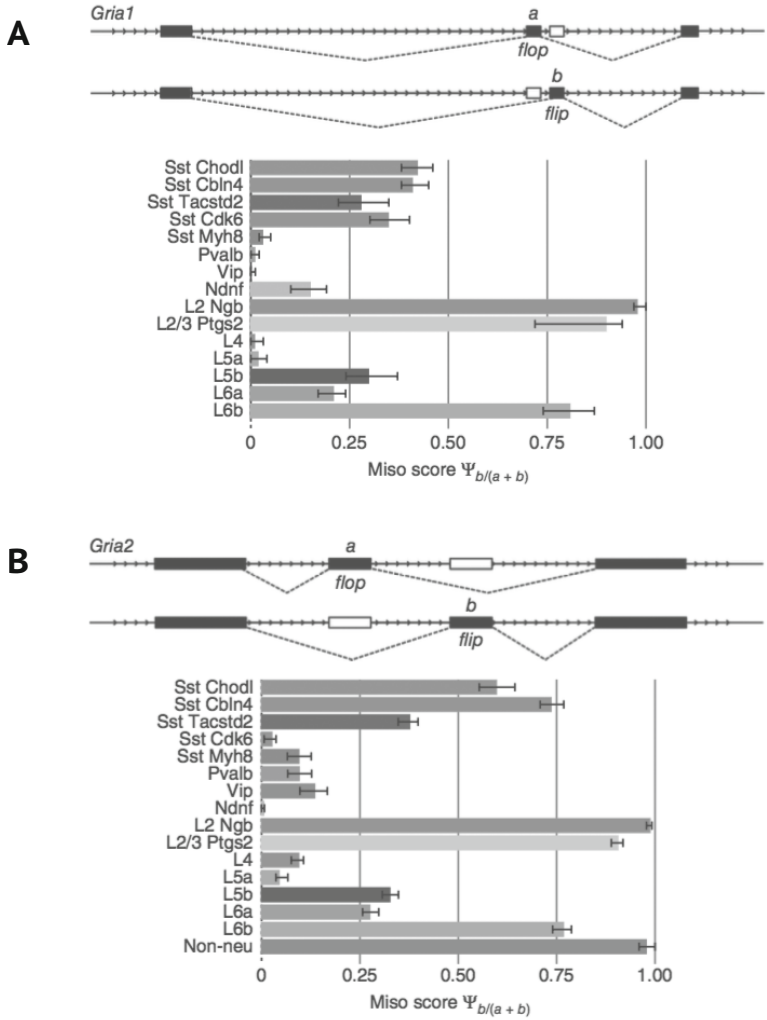


Figure 9.12. 'Flip-flop' architecture of the gene for the AMPA glutamate receptors *Gria1* (A) and *Gria2* (B), with associated Miso score of differential splicing $\Psi = b/(b+a)$ in different cell clusters ($\Psi = 0$ if only the flip variant is expressed, and the opposite is true when $\Psi = 1$). The differential slicing of two consecutive exons ('flip' and 'flop') produces two variants of the receptor, each presenting distinct biophysical properties. *Adapted by permission from Macmillan Publishers Ltd: Nature Neuroscience* [54], © (2016).

cells were then isolated by FACS and sequenced with the same protocol described before. Using their molecular profiles, these cells can be as-

target cell becomes fluorescent. In the presented experimental setup, the cells that become Cre^+ are the ones that project to the area injected with the canine adenovirus.

signed to one or more of the previously described core clusters using the same random forest classifier. The two classes of retrogradely-labelled cells are assigned to small non-overlapping subsets of the cell clusters previously localised in layer 5 and layer 6. This experiment allows the authors to attach functional information to these clusters, which would have been otherwise described only in molecular terms.

To prove the second point, the authors concentrated on a class of GABAergic neurons that are characterised by the expression of the neuropeptide neuron-derived neurotrophic factor (*Ndnf*). These cells are grouped into two separate core clusters (*Ndnf-Car4* and *Ndnf-Cxcl14*) that are localised in the uppermost layer of the cortex (L1) and express the marker Reelin (*Reln*). From this information the authors were able to assign the identity of *neurogliaform cells*. The production of transgenic mice obtained by crossing the Ai14 line with a line expressing Cre under the *Ndnf* promoter makes it possible to express a fluorescent tag in the cells of the two *Ndnf* clusters. Electrophysiological analysis through patch clamping of the tdTomato-tagged cells and 3D morphological reconstruction using fluorescence microscopy of cells intracellularly filled with a dye confirmed that the members of the two *Ndnf* clusters are electrophysiologically and morphologically distinct.

The results of this landmark paper demonstrate that the use of single-cell technologies can greatly improve our understanding of the nervous system by providing an unbiased account of cellular diversity and combining morphological, electrophysiological, and molecular information.

9.5. Other applications of single-cell RNA-seq to the nervous system

Two recent papers applied scRNA-seq to the retina. The retina is the region of the brain that has been best characterised in terms of cellular diversity and connectivity[20]. Single-cell technologies have recapitulated the diversity of retinal neurons in an unbiased manner and have, moreover, identified novel molecular aspects of retinal diversity [32]. In a follow-up study, the authors concentrated on retinal bipolar cells, a population of retinal cells that was thought to have been exhaustively classified. The molecular classification identified all previously observed types, as well as two novel types, one of which has a non-canonical morphology and position [47].

[20] In a sense, the retina is an outpost of the CNS 1) that is easily accessible to the experimenter and 2) can provide a 'simplified' image of the properties of neuronal computation and rules of connectivity—as it is essentially a thin and highly-ordered layer of nervous cells.

Another method worth mentioning combines classical 'pulse-chase' experiments, which use nucleotide analogues to label newborn neurons, with FACS sorting of nuclei rather than cells. Using this method—named Div-seq, for sequencing of dividing cells—it has been possible to study at single-cell resolution the transcriptome of an exceedingly rare population: the newborn neurons of the adult hippocampus [17].

Conclusion: new tools, new challenges

In Chapters 8 and 9 we showed that next-generation RNA-seq provides the neuroscientist with a wide range of new tools that allow for comprehensive molecular characterisation, from the whole organ to a single cell or even subcellular compartment. In this embarrassment of riches, the careful choice of a relevant scientific question and an informative experimental design is of extreme importance. As single-cell technologies will certainly become more standardised and prevalent in the neurosciences, we must consider the trade-offs: do we better tackle our biological questions by trading depth for single-cell resolution, or by employing whole-genome analyses that provide extreme depth at the expense of spatial resolution?

In general, these types of technologies deliver more information than you may desire. See an RNA-seq experiment as a hike through a forest: in some cases you need to focus on single trees ('is there a trail maker on the trunk?'), but in others you will have to consider distinct patches of the forest as single units (with specific branch density, and so on) to keep yourself from getting lost.

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *Basic local alignment search tool*, Journal of Molecular Biology (3) **215** (1990), 403–410.

[2] S. Anders and W. Huber, *Differential expression analysis for sequence count data*, Genome biology (10) **11** (2010), R106.

[3] D. F. Andrews and A. M. Herzberg, "Data: a Collection of Problems from Many Fields for the Student and Research Worker", Springer Science & Business Media, 2012.

[4] M. Aschoff, A. Hotz-Wagenblatt, K.-H. Glatting, M. Fischer, R. Eils, and R. König, *Splicingcompass: differential splicing detection using rna-seq data*, Bioinformatics (9) 29 (2013), 1141–1148.

[5] A.-L. Barabasi and Z. N. Oltvai, *Network biology: understanding the cell's functional organization*, Nature Reviews Genetics (2) **5** (2004), 101.

[6] M. Baumgart, S. Priebe, M. Groth, N. Hartmann, U. Menzel, L. Pandolfini, P. Koch, M. Felder, M. Ristow, C. Englert, A. Cellerino, *et al.*, *Longitudinal rna-seq analysis of vertebrate aging identifies mitochondrial complex i as a small-molecule-sensitive modifier of lifespan*, Cell Systems (2) **2** (2016), 122–132.

[7] P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, *et al.*, *Accounting for technical noise in single-cell rna-seq experiments*, Nature Methods (11) **10** (2013), 1093.

[8] J. D. Cahoy, B. Emery, A. Kaushal, L. C. Foo, J. L. Zamanian, Karen S. Christopherson, Y. Xing, J. L. Lubischer, P. A. Krieg, S. A. Krupenko, A. Sergey *et al.*, *A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function*, Journal of Neuroscience **28** (2008), 264–278.

[9] P. Ciryam, R. Kundra, R. Freer, R. I. Morimoto, C. M. Dobson, and M. Vendruscolo, *A transcriptional signature of alzheimer's disease is associated with a metastable subproteome at risk for aggregation*, Proceedings of the National Academy of Sciences, (17) **113** (2016), 4753–4758.

[10] C. Colantuoni, B. K. Lipska, T. Ye, T. M. Hyde, R. Tao, J. T. Leek, E. A. Colantuoni, A. G. Elkahloun, M. M. Herman, D. R. Weinberger, *et al*., *Temporal dynamics and genetic control of transcription in the human prefrontal cortex*, Nature (7370) **478** (2011), 519.

[11] P. D'haeseleer, *How does gene expression clustering work?*, Nature Biotechnology **23** (2005), 1499–1501.

[12] J. P. Doyle, J. D. Dougherty, M. Heiman, E. F. Schmidt, T. R. Stevens, G. Ma, S. Bupp, P. Shrestha, R. D. Shah, M. L. Doughty, *et al*., *Application of a translational profiling approach for the comparative analysis of cns cell types*, Cell (4) **135** (2008), 749–762.

[13] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns*, Proceedings of the National Academy of Sciences, (25) **95** (1998), 14863–14868.

[14] M. I. Ekstrand, A. R. Nectow, Z. A. Knight, K. N. Latcha, L. E. Pomeranz, and J. M. Friedman, *Molecular profiling of neurons based on connectivity*, Cell (5) **157** (2014), 1230–1242.

[15] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al*., *A density-based algorithm for discovering clusters in large spatial databases with noise*, In: "KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining", Vol. 96, 1996, 226–231.

[16] N. Gehlenborg and B. Wong, *Points of view: heat maps*, Nature Methods (3) **9** (2012), 213.

[17] N. Habib, Y. Li, M. Heidenreich, L. Swiech, I. Avraham-Davidi, J. J. Trombetta, C. Hession, F. Zhang, and A. Regev, *Div-seq: Single-nucleus rna-seq reveals dynamics of rare adult newborn neurons*, Science (6302) **353** (2016), 925–928.

[18] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns, *Mapping the structural core of human cerebral cortex*, PLoS Biology (7) **6** (2008), e159.

[19] M. J. Hawrylycz, E. S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. Van De Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley, *et al*., *An anatomically comprehensive atlas of the adult human brain transcriptome*, Nature (7416) **48** (2012), 391.

[20] J. M. Hilbe, "Negative Binomial Regression", Cambridge University Press, 2011.

[21] S. Horvath and J. Dong, *Geometric interpretation of gene coexpression network analysis*, PLoS Computational Biology (8) **4** (2008), e1000117.

[22] C. J. Huh, B. Zhang, M. B. Victor, S. Dahiya, L. F. Batista, S. Horvath, and A. S. Yoo, *Maintenance of age in human neurons generated by microrna-based neuronal conversion of fibroblasts* Elife 5 (2016), e18648.

[23] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, *Kegg: new perspectives on genomes, pathways, diseases and drugs*, Nucleic Acids Research (D1) **45** (2016), D353–D361.

[24] M. Kanehisa and S. Goto, *Kegg: kyoto encyclopedia of genes and genomes*, Nucleic Acids Research (1) **28** (2000), 27–30.

[25] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, *Kegg as a reference resource for gene and protein annotation*, Nucleic Acids Research (D1) **44** (2015), D457–D462.

[26] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, *Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells*, Cell (5) **161** (2015), 1187–1201.

[27] Z. A. Knight, K. Tan, K. Birsoy, S. Schmidt, J. L. Garrison, R. W. Wysocki, A. Emiliano, M. I. Ekstrand, and J. M. Friedman, *Molecular profiling of activated neurons by phosphorylated ribosome capture*, Cell (5) **151** (2012), 1126–1137.

[28] T. Kohonen, *The self-organizing map*, Neurocomputing (1-3) **21** (1998), 1–6.

[29] S. Lee, B. Liu, S. Lee, S.-X. Huang, B. Shen, and S.-B. Qian, *Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution*, In: "Proceedings of the National Academy of Sciences" (37) **109** (2012), E2424–E2432.

[30] M. I. Love, W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for rna-seq data with deseq2*, Genome Biology (12) **15** (2014), 550.

[31] W. Luo, M. S. Friedman, K. Shedden, K. D. Hankenson, and P. J. Woolf, *Gage: generally applicable gene set enrichment for pathway analysis*, BMC Bioinformatics (1) **10** (2009), 161.

[32] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, *et al.*, *Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets*, Cell (5) **161** (2015), 1202–1214.

[33] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, *Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays*, Genome Research (9) **18** (2008), 1509–1517.

[34] L. Ng, A. Bernard, C. Lau, C. C. Overly, H.-W. Dong, C. Kuan, S. Pathak, S. M. Sunkin, C. Dang, J. W. Bohland, *et al*., *An anatomic gene expression atlas of the adult mouse brain*, Nature Neuroscience (3) **12** (2009), 356.

[35] S. W. Oh, J. A. Harris, L. Ng, B. Winslow, N. Cain, S. Mihalas, Q. Wang, C. Lau, L. Kuan, A. M. Henry, *et al*., *A mesoscale connectome of the mouse brain*, Nature (7495) **508** (2014), 207.

[36] M. C. Oldham, S. Horvath, and D. H. Geschwind, *Conservation and evolution of gene coexpression networks in human and chimpanzee brains*, In: "Proceedings of the National Academy of Sciences", (47) **103** (2006), 17973–17978.

[37] N. R. Pamudurti, O. Bartok, M. Jens, R. Ashwal-Fluss, C. Stottmeister, L. Ruhe, M. Hanan, E. Wyler, D. Perez-Hernandez, E. Ramberger, *et al*., *Translation of circrnas*, Molecular Cell (1) **66** (2017), 9–21.

[38] D. Pierron, D. E. Wildman, M. Hüttemann, T. Letellier, and L. I. Grossman, *Evolution of the couple cytochrome c and cytochrome c oxidase in primates*, In: "Mitochondrial Oxidative Phosphorylation", Springer, 2012, 185–213.

[39] M. Piwecka, P. Glažar, L. R. Hernandez-Miranda, S. Memczak, S. A. Wolf, A. Rybak-Wolf, A. Filipchyk, F. Klironomos, C. A. C. Jara, P. Fenske, *et al*., *Loss of a mammalian circular rna locus causes mirna deregulation and affects brain function*, Science (6357) **357** (2017), eaam8526.

[40] M. Pletikos, A. M. Sousa, G. Sedmak, K. A. Meyer, Y. Zhu, F. Cheng, M. Li, Y. I. Kawasawa, and N. Šestan, *Temporal specification and bilaterality of human neocortical topographic gene expression*, Neuron (2) **81** (2014), 321–332.

[41] S. M. Prakadan, A. K. Shalek, and D. A. Weitz, *Scaling by shrinking: empowering single-cell'omics' with microfluidic devices*, Nature Reviews Genetics (6) **18** (2017), 345.

[42] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, *Hierarchical organization of modularity in metabolic networks*, Science (5586) **297** (2002), 1551–1555.

[43] K. Reichwald, and Petzold, Andreas and Koch, Philipp and Downie, Bryan R and Hartmann, Nils and Pietsch, Stefan and Baumgart, Mario and Chalopin, Domitille and Felder, Marius and Bens, Martin

and others *Insights into sex chromosome evolution and aging from the genome of a short-lived fish*, Cell **163** (2015), 1527–1538.

[44] J. Richiardi, A. Altmann, A.-C. Milazzo, C. Chang, M. M. Chakravarty, T. Banaschewski, G. J. Barker, A. L. Bokde, U. Bromberg, C. Büchel, *et al.*, *Correlated gene expression supports synchronous activity in brain networks*, Science (6240) **348** (2015), 1241–1244.

[45] M. Ringnér, *What is principal component analysis?* Nature Biotechnology (3) **26** (2008), 303.

[46] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, *edger: a bioconductor package for differential expression analysis of digital gene expression data*, Bioinformatics (1) **26** (2010), 139–140.

[47] K. Shekhar, S. W. Lapan, I. E. Whitney, N. M. Tran, E. Z. Macosko, M. Kowalczyk, X. Adiconis, J. Z. Levin, J. Nemesh, M. Goldman, *et al.*, *Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics*, Cell (5) **166** (2016), 1308–1323.

[48] T. Shigeoka, H. Jung, J. Jung, B. Turner-Bridger, J. Ohk, J. Q. Lin, P. S. Amieux, and C. E. Holt, *Dynamic axonal translation in developing and mature visual circuits*, Cell (1) **166** (2016), 181–192.

[49] T. P. Singer and R. R. Ramsay, *The reaction sites of rotenone and ubiquinone with mitochondrial nadh dehydrogenase*, Biochimica et Biophysica Acta (BBA)-Bioenergetics (2) **1187** (1994), 198–202.

[50] B. Sommer, K. Keinanen, T. A. Verdoorn, W. Wisden, N. Burnashev, A. Herb, M. Kohler, T. Takagi, B. Sakmann, and P. H. Seeburg, *Flip and flop: a cell-specific functional switch in glutamate-operated channels of the cns*, Science (4976) **249** (1990), 1580–1585.

[51] O. Stegle, S. A. Teichmann, and J. C. Marioni, *Computational and analytical challenges in single-cell transcriptomics*, Nature Reviews Genetics (3) **16** (2015), 133.

[52] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*, In: "Proceedings of the National Academy of Sciences" (43) **102** (2005), 15545–15550.

[53] K. Sugino, C. M. Hempel, M. N. Miller, A. M. Hattox, P. Shapiro, C. Wu, Z. J. Huang, and S. B. Nelson, *Molecular taxonomy of major neuronal classes in the adult mouse forebrain*, Nature Neuroscience (1) **9** (2006), 99.

[54] B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, *et al.*, *Adult mouse*

cortical cell taxonomy revealed by single cell transcriptomics, Nature Neuroscience (2) **19** (2016), 335.

[55] E. Taskesen and M. J. Reinders, *2d representation of transcriptomes by t-sne exposes relatedness between human tissues*, PloS one (2) **11** (2016), e0149853.

[56] I. Ulitsky and D. P. Bartel, *Lincrnas: genomics, evolution, and mechanisms*, Cell (1) **154** (2013), 26–46.

[57] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii, *Structural properties of the Caenorhabditis elegans neuronal network*, PLoS Computational Biology (2) **7** (2011), e1001066.

[58] R. L. Wasserstein and N. A. Lazar, *The asa's statement on p-values: context, process, and purpose*, 2016.

[59] K. D. Winden, M. C. Oldham, K. Mirnics, P. J. Ebert, C. H. Swan, P. Levitt, J. L. Rubenstein, S. Horvath, and D. H. Geschwind, *The organization of the transcriptional network in specific neuronal classes*, Molecular Systems Biology (1) **5** (2009), 291.

[60] X. You, I. Vlatkovic, A. Babic, T. Will, I. Epstein, G. Tushev, G. Akbalik, M. Wang, C. Glock, C. Quedenau, *et al.*, *Neural circular rnas are derived from synaptic genes and regulated by development and plasticity*, Nature Neuroscience (4) **18** (2015), 603.

[61] B. Zhang and S. Horvath, *A general framework for weighted gene co-expression network analysis*, Statistical Applications in Genetics and Molecular Biology (1) **4** (2005).

[62] Z. D. Zhang, J. Rozowsky, M. Snyder, J. Chang, and M. Gerstein, *Modeling chip sequencing in silico with applications*, PLoS Computational Biology (8) **4** (2008), e1000158.

Index

LECTURE NOTES

This series publishes polished notes dealing with topics of current research and originating from lectures and seminars held at the Scuola Normale Superiore in Pisa.

Published volumes

1. M. TOSI, P. VIGNOLO, *Statistical Mechanics and the Physics of Fluids*, 2005 (second edition). ISBN 978-88-7642-144-0
2. M. GIAQUINTA, L. MARTINAZZI, *An Introduction to the Regularity Theory for Elliptic Systems, Harmonic Maps and Minimal Graphs*, 2005. ISBN 978-88-7642-168-8
3. G. DELLA SALA, A. SARACCO, A. SIMIONIUC, G. TOMASSINI, *Lectures on Complex Analysis and Analytic Geometry*, 2006. ISBN 978-88-7642-199-8
4. M. POLINI, M. TOSI, *Many-Body Physics in Condensed Matter Systems*, 2006. ISBN 978-88-7642-192-0
   P. AZZURRI, *Problemi di Meccanica*, 2007. ISBN 978-88-7642-223-2
5. R. BARBIERI, *Lectures on the ElectroWeak Interactions*, 2007. ISBN 978-88-7642-311-6
6. G. DA PRATO, *Introduction to Stochastic Analysis and Malliavin Calculus*, 2007. ISBN 978-88-7642-313-0
   P. AZZURRI, *Problemi di meccanica*, 2008 (second edition). ISBN 978-88-7642-317-8
   A. C. G. MENNUCCI, S. K. MITTER, *Probabilità e informazione*, 2008 (second edition). ISBN 978-88-7642-324-6
7. G. DA PRATO, *Introduction to Stochastic Analysis and Malliavin Calculus*, 2008 (second edition). ISBN 978-88-7642-337-6
8. U. ZANNIER, *Lecture Notes on Diophantine Analysis*, 2009. ISBN 978-88-7642-341-3
9. A. LUNARDI, *Interpolation Theory*, 2009 (second edition). ISBN 978-88-7642-342-0

10. L. AMBROSIO, G. DA PRATO, A. MENNUCCI, *Introduction to Measure Theory and Integration*, 2012.
    ISBN 978-88-7642-385-7, e-ISBN: 978-88-7642-386-4
11. M. GIAQUINTA, L. MARTINAZZI, *An Introduction to the Regularity Theory for Elliptic Systems, Harmonic Maps and Minimal Graphs*, 2012 (second edition). ISBN 978-88-7642-442-7, e-ISBN: 978-88-7642-443-4
    G. PRADISI, *Lezioni di metodi matematici della fisica*, 2012.
    ISBN: 978-88-7642-441-0
12. G. BELLETTINI, *Lecture Notes on Mean Curvature Flow, Barriers and Singular Perturbations*, 2013.
    ISBN 978-88-7642-428-1, e-ISBN: 978-88-7642-429-8
13. G. DA PRATO, *Introduction to Stochastic Analysis and Malliavin Calculus*, 2014. ISBN 978-88-7642-497-7, e-ISBN: 978-88-7642-499-1
14. R. SCOGNAMILLO, U. ZANNIER, *Introductory Notes on Valuation Rings and Function Fields in One Variable*, 2014. ISBN 978-88-7642-500-4, e-ISBN: 978-88-7642-501-1
15. S. DIPIERRO, M. MEDINA, E. VALDINOCI, *Fractional Elliptic Problems with Critical Growth in the Whole of* $\mathbb{R}^n$, 2017.
    ISBN 978-88-7642-600-1, e-ISBN: 978-88-7642-601-8
    G. PRADISI, *Lezioni di metodi matematici della fisica*, 2012 (reprint 2018) ISBN: 978-88-7642-441-0
16. A. LUNARDI, *Interpolation Theory*, 2018 (third edition).
    ISBN 978-88-7642-639-1, e-ISBN: 978-88-7642-638-4
17. A. CELLERINO, M. SANGUANINI, *Transcriptome Analysis. Introduction and Examples from the Neurosciences*, 2018.
    ISBN 978-88-7642-641-4, e-ISBN: 978-88-7642-642-1

Volumes published earlier

G. DA PRATO, *Introduction to Differential Stochastic Equations*, 1995 (second edition 1998). ISBN 978-88-7642-259-1

L. AMBROSIO, *Corso introduttivo alla Teoria Geometrica della Misura ed alle Superfici Minime*, 1996 (reprint 2000).

E. VESENTINI, *Introduction to Continuous Semigroups*, 1996 (second edition 2002). ISBN 978-88-7642-258-4

C. PETRONIO, *A Theorem of Eliashberg and Thurston on Foliations and Contact Structures*, 1997. ISBN 978-88-7642-286-7

Quantum cohomology at the Mittag-Leffler Institute, a cura di Paolo Aluffi, 1998. ISBN 978-88-7642-257-7

G. BINI, C. DE CONCINI, M. POLITO, C. PROCESI, *On the Work of Givental Relative to Mirror Symmetry*, 1998. ISBN 978-88-7642-240-9

H. PHAM, *Imperfections de Marchés et Méthodes d'Evaluation et Couverture d'Options*, 1998. ISBN 978-88-7642-291-1

H. CLEMENS, *Introduction to Hodge Theory*, 1998. ISBN 978-88-7642-268-3

Seminari di Geometria Algebrica 1998-1999, 1999.

A. LUNARDI, *Interpolation Theory*, 1999. ISBN 978-88-7642-296-6

R. SCOGNAMILLO, *Rappresentazioni dei gruppi finiti e loro caratteri*, 1999.

S. RODRIGUEZ, *Symmetry in Physics*, 1999. ISBN 978-88-7642-254-6

F. STROCCHI, *Symmetry Breaking in Classical Systems*, 1999 (2000). ISBN 978-88-7642-262-1

L. AMBROSIO, P. TILLI, *Selected Topics on* "Analysis in Metric Spaces", 2000. ISBN 978-88-7642-265-2

A. C. G. MENNUCCI, S. K. MITTER, *Probabilità ed Informazione*, 2000.

S. V. BULANOV, *Lectures on Nonlinear Physics*, 2000 (2001). ISBN 978-88-7642-267-6

Lectures on Analysis in Metric Spaces, a cura di Luigi Ambrosio e Francesco Serra Cassano, 2000 (2001). ISBN 978-88-7642-255-3

L. CIOTTI, *Lectures Notes on Stellar Dynamics*, 2000 (2001). ISBN 978-88-7642-266-9

S. RODRIGUEZ, *The Scattering of Light by Matter*, 2001. ISBN 978-88-7642-298-0

G. DA PRATO, *An Introduction to Infinite Dimensional Analysis*, 2001. ISBN 978-88-7642-309-3

S. SUCCI, *An Introduction to Computational Physics*: – Part I: *Grid Methods*, 2002. ISBN 978-88-7642-263-8

D. BUCUR, G. BUTTAZZO, *Variational Methods in Some Shape Optimization Problems*, 2002. ISBN 978-88-7642-297-3

A. MINGUZZI, M. TOSI, *Introduction to the Theory of Many-Body Systems*, 2002.

S. SUCCI, *An Introduction to Computational Physics*: – Part II: *Particle Methods*, 2003. ISBN 978-88-7642-264-5

A. MINGUZZI, S. SUCCI, F. TOSCHI, M. TOSI, P. VIGNOLO, *Numerical Methods for Atomic Quantum Gases*, 2004. ISBN 978-88-7642-130-0