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.

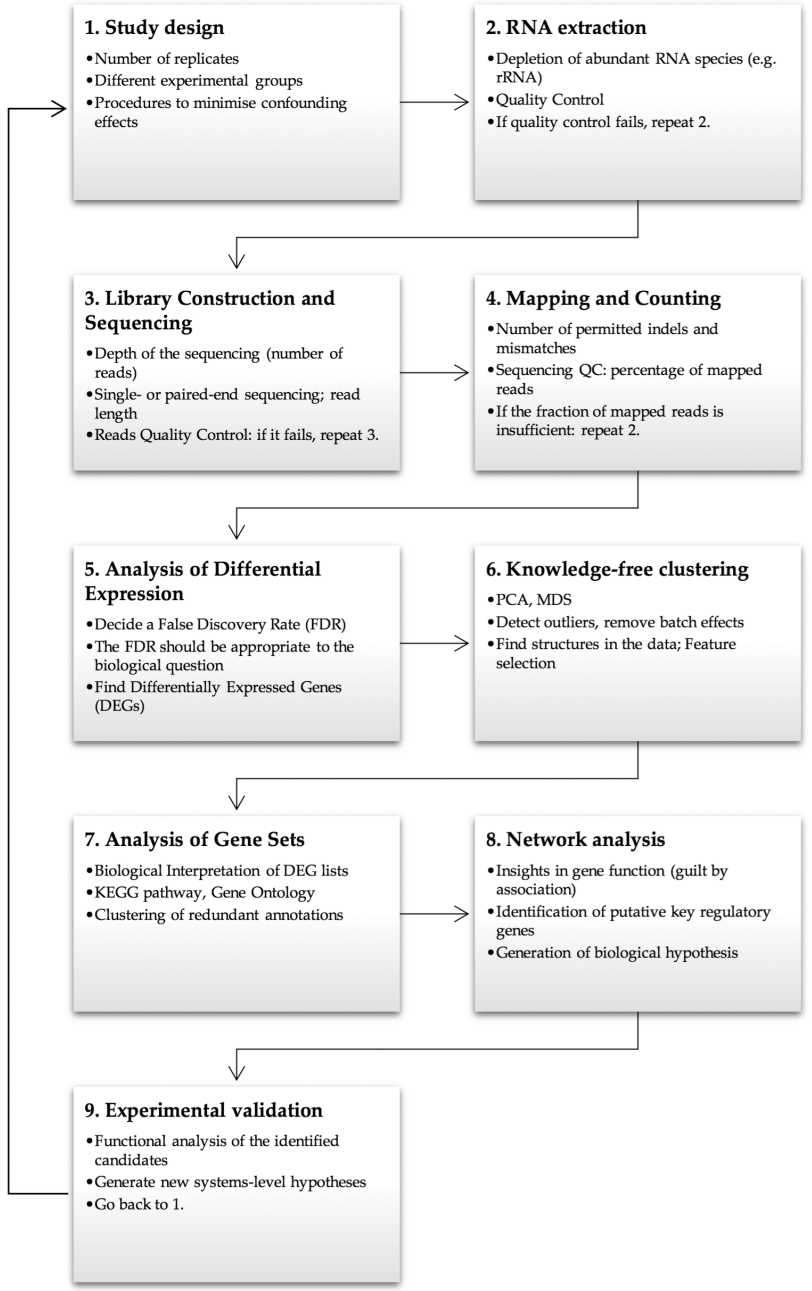
In this book, we will thus deal with three major levels and concepts.

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