Gene regulation in embryonic development

Berg, P.R. van den

INTRODUCTION

Our bodies are unbelievably complex systems. An important contributor to this complexity is the enormous diversity of cell types. Each cell type has its own properties and functions, such as the transfer of information in the brain or the transport of oxygen in the bloodstream. Despite all this diversity, each cell in our body has the same origin: a single fertilized egg cell. Developmental biology is the science of life's journey from this single cell, to trillions of cells with hundreds of different cell types.

But what even defines a cell type? The exact answer to this is not trivial and can be approached at different levels. Functional specialization or location in the body can both be used to define a cell type. What a cell can do (which is largely determined by its protein composition) and where a cell is located (when and where it was formed during development) is all determined by the genome of that single fertilized egg cell.

To discover what parts of the genome, i.e. which genes, are active when and where in a developing embryo, is thus an important task of developmental biology. Various biochemical processes, collectively termed *gene regulatory mechanisms*, determine the amount of transcription, translation and protein degradation, resulting in precisely tuned protein abundances. These processes are at the center of defining a cell type, since they ultimately determine its protein composition. In this thesis we will study embryonic development with gene regulation as a common thread. Here we will introduce the different types of gene regulatory mechanisms discussed in this thesis, as well as some of the state-of-the-art experimental techniques we employ to study these mechanisms.

DNA methylation

The human genome comprises 46 DNA molecules, totaling two meters in length, which are packaged inside the cell's nucleus. How tightly different parts of the DNA are packed determines, at least in part, which genes are active. In turn, the gene activity profile is strongly correlated with the cell's type. Which parts of the DNA are accessible or inaccessible is not encoded in the DNA sequence itself but determined by chemical modifications of the DNA or proteins that are associated with it. The field of *epigenetics* studies how this meta-information is modified by the cell and how it determines when a gene is active. One layer of this meta-information is stored in the modifications of a specific DNA base. A methyl-group can be added and removed from cytosine and this triggers a whole cascade of other processes influencing DNA packing. Knowing where throughout the genome the DNA is methylated, its *methylome*, thus reveals vital information about how genes are regulated.

Transcriptional and translational regulation

The instructions on how to build proteins are stored in the genome, but these instructions are not read out directly. In between DNA and protein, messenger RNA (mRNA) acts as a carrier of these instructions. The process of copying DNA into mRNA is called transcription and reading mRNA's instructions for protein is called translation. Both of these are heavily regulated processes.

Transcription is the process of copying DNA sequences to the equivalent of mRNA. In contrast to DNA there can be thousands of copies of mRNA in the cell simultaneously. The amount of mRNA in the cell will largely determine the amount of the corresponding protein in the cell so controlling transcription is vitally important. There are many actors that bind DNA that are responsible for either promoting or repressing transcription. These molecules are called transcription factors and at any given transcription site there is typically a complex set of these present that fine-tune transcription levels. Conversely, each transcription factor can bind multiple locations in the genome. In a famous example, there is a set of transcription factors that, when introduced into the cell, can reprogram a mature cell type back into a progenitor cell. Another important role in transcriptional regulation is that of the enhancer. These regions of DNA are typically positioned far away from the transcription site but can nonetheless control the transcription level.

Translation is the process of converting genes encoded in mRNA into proteins. Akin to transcriptional regulation, translation can be regulated in multiple ways. Most regulatory mechanisms involve the binding of proteins or other RNAs to specific sequences in the mRNA. One mechanism studied in this thesis involves micro-RNAs (miRs). miRs are short pieces of RNA that bind to complementary sequences in their target mRNAs. They are known to either signal *slicer* proteins to degrade the mRNA or to simply block the translation machinery. When it comes to important cellular processes like differentiation, big changes in translational regulation is much less prevalent than in transcriptional regulation. However, in particular cases translational regulation plays a decisive role and should not be overlooked.

Omics

To measure is to know, therefore science is forever pushing the limits of measurements. This thesis makes use of several high-throughput methods to measure the molecular profiles of cells. In order to discern what is being measured we use a set of suffixes. For instance we measure the gen*ome* with gen*omics* and the prote*ome* with prote*omics*. *-ome* relates to all the information the type of molecule holds while *-omics* describes the measurement of this information. In this thesis we will deal with the genome (the DNA sequence), the methylome (where the DNA is methylated), the transcriptome (the composition of mRNAs), the mirnome (the composition of miRs) and the proteome (the composition of proteins). All of these measurements have been made possible by recent scientific and commercial advances.

In particular DNA sequencing methods have recently driven transcriptomics and genomics. Mass spectrometry is one of the main tools to measure the proteome. Most of these -omics tools measure many cells to get accurate averages, but a recent trend is to measure single cells. Single cell -omics allows us to study cell populations without averaging out important variability between cells.

Thesis outline

In this thesis we explore the ways cells exert control over their states, particularly in the context of development. We mirror the central dogma of molecular biology in the order of the chapters: from DNA to RNA to protein.

In **Chapter 1** we look at the dynamics of DNA methylation in mouse embryonic stem cells (ESCs). It has been previously found that there are regions in the genome that are differentially methylated depending on the cell type. Bulk DNA sequencing based methods can only provide a static snapshot of average methylation levels. In particular, bulk methods obscure heterogeneity within the cell population, which might have important functional consequences. Here we used a locus-specific fluorescent reporter for DNA methylation to study two super enhancers sites. Our study reveals that their methylation level is highly dynamic and correlated with gene expression. Moreover, methylation changes occur independently on the two alleles (i.e. the mother's and father's copy of the super enhancers). In summary, Chapter 1 demonstrates a new and unique tool for studying enhancer DNA methylation in heterogeneous populations of cells.

In **Chapter 2** we look at how the human kidney develops. Most research on kidney development has been done in mice, even though there are many crucial differences between the species. We were one of the first to measure human fetal kidneys with a new state-of-the-art technique: single-cell RNA-sequencing. In this dataset we identified 22 distinct kidney cell types at five developmental ages. One of our key findings is that there are several subclasses of nephron progenitors, which give rise to the basic functional unit of the kidney. In summary, Chapter 2 creates a better picture of how the human kidney develops and how it compares to the mouse kidney.

In **Chapter 3** we look at the discordance between mRNA and protein abundance to discover mechanisms of translational regulation. mRNA changes are often used as surrogate for protein changes in the cell. However, protein expression over time does not directly correspond to mRNA expression due to finite rates of protein synthesis and degradation and a resulting delay. For example, mRNA expression may spike upwards in a matter of minutes, but protein expression may take a day to catch up, owing to a slow protein synthesis rate. To better understand these differences we measured both mRNA and protein in a mouse ESCs differentiation experiment. By modeling protein expression as a birth-death process we discern which gene is entirely determined by mRNA abundance and which is the target of post-transcriptional regulation. We further integrate a small RNA-sequencing dataset into

our model to identify miRs that may be responsible for some of this regulation. In summary, Chapter 3 determines the important differences between mRNA and protein dynamics and uses these differences to identify cases of post-transcriptional regulation.

In **Chapter 4** follows up on some of the results of the preceding chapter. The regulation that miRs exert on mRNAs adds a layer of complexity to gene regulation that is often ignored. This is in part due to the enormous number of possible miR-gene interactions and the ambiguity of the binding sites. In the previous chapter we used our birth-death model to predict several miR-gene interactions in mouse ESC differentiation. Here, we set out to validate some of these interactions by introducing mimics and inhibitors of these miRs into ESCs. We created reporter cell lines of miR activity to accurately select the timing and dosage of those reagents. We found that four out of six candidate miRs down-regulated their predicted target. In summary, Chapter 4 validates the birth-death model as a tool for finding miR-gene interactions.