From Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

GLOBAL REGULATION OF GENE EXPRESSION IN STEM CELLS AND REGENERATION

Ilgar Abdullayev

Karolinska Institutet

Stockholm 2017

Cover art: “2017: A Newt Odyssey”. Chromosomes dancing around the monolith, inspired from the movie by Stanley Kubrick. Designed by Ilgar Abdullayev and illustrated by Dai Lu.

Published by Karolinska Institutet.
Printed by Eprint AB 2017

ISBN 978-91-7676-771-9

Global regulation of gene expression in stem cells and regeneration

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Ilgar Abdullayev

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i föreläsningssalen CMB, Berzelius väg 21

Fredagen den 8:e September 2017, kl 13:00

Principal Supervisor:
Rickard Sandberg
Karolinska Institutet
Department of Cell and Molecular Biology

Co-supervisor(s):
Pelin (Akan) Sahlén
Royal Institute of Technology
Science for Life Laboratory
Division of Gene Technology

Opponent:
Ali Mortazavi
University of California Irvine
School of Biological Sciences
Department of Developmental and Cell Biology

Examination Board:
Gerhart Wagner
Uppsala University
Department of Cell and Molecular Biology, Microbiology

Johan Holmberg
Karolinska Institutet
Department of Cell and Molecular Biology

Jussi Taipale
Karolinska Institutet
Department of Medical Biochemistry and Biophysics

To my family and my son Cansun

ABSTRACT

Rapid developments in genomics and transcriptomics fields have made it possible to ask new questions as well as solve various old problems in biology that were not achievable previously. Novel techniques such as RNA sequencing and Hi-C became available at the time I started my PhD. Therefore, in order to study regeneration in salamanders and genome-wide regulatory interactions in mouse embryonic stem cells, my first goals were to make use of these techniques. Regeneration in salamanders has not been fully understood despite being studied for a few centuries. One of the reasons was the scarcity of genomic data. We mainly solved this problem by providing a high-quality transcriptome of red spotted newt, using latest tools (Paper I). Combining Hi-C with promoter capture probes increased the resolution for finding regulatory interactions, mainly promoter-enhancer (distal element). One of the surprising discoveries was enhancer-enhancer interactions, which was actually due to imperfect promoter capture efficiency. Our method, HiCap (Paper II), had a highest resolution for locating enhancers, yet had a modest improvement over assigning enhancers to their closest gene. Further analysis of regulatory networks showed a strong connectivity of enhancers and promoters individually than promoter-enhancers together.

My last two projects involved studying gene regulation at a single cell level. The role of small RNAs in gene regulation in individual cells was not studied at that time. Aiming to shed a light on this, we developed a single-cell method for small RNAs, where I performed all the computational analysis (Paper III). This novel method, Small-seq, mainly revealed that microRNAs could be used to cluster different cell types. Since almost all of the available single-cell methods quantify polyadenylated RNAs (mainly mRNAs), Small-seq showed that one can get equally good clustering of cells using an order of magnitude less number of genes (about 200 microRNAs in human embryonic stem cells compared to a few thousand mRNAs). By making use of the newt transcriptome from Paper I, we aim to decipher the cellular composition of blastema – a small bud of cell mass formed on the amputation surface of regenerating newt limb. Adult newt limbs, upon amputation, undergo a precisely controlled "magic" of regenerating fully functional copy of its original limb. Newt cells are shown to dedifferentiate back to progenitor-like cellular state, populate and differentiate back to necessary cell types. The extend of this dedifferentiation and which cells contribute and how much is unknown. In paper IV, we have studied limb regeneration in newt and identified 8 cell types in blastema, where one cell type has significantly enriched for transposable elements, DNA fragments that are able to change their genomic positions, and has been shown to play a critical role in stem cell pluripotency, disease and development. Overall, this thesis covers studies of gene regulation in regeneration and several types of stem cells, both at an individual cell level as well as using millions of cells, by applying latest experimental and computational methods.

LIST OF SCIENTIFIC PAPERS

I. **Ilgar Abdullayev**, Matthew Kirkhama, Åsa K. Björklund, András Simon, Rickard Sandberg. (2013) A reference transcriptome and inferred proteome for the salamander Notophthalmus viridescens. Exp. Cell Res. 319:1187–1197doi:10.1016/j.yexcr.2013.02.013

II. Pelin Sahlén*, **Ilgar Abdullayev***, Daniel Ramsköld, Liudmila Matskova, Nemanja Rilakovic, Britta Lötstedt, Thomas J. Albert, Joakim Lundeberg and Rickard Sandberg. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. Genome Biol. 2015;16(1):156. 10.1186/s13059-015-0727-9

III. Omid R Faridani*, **Ilgar Abdullayev*,** Michael Hagemann-Jensen, John P Schell, Fredrik Lanner & Rickard Sandberg. Single-cell sequencing of the small-RNA transcriptome Nat. Biotechnol., 34 (2016), pp. 1264–1266

IV. Ahmed Elewa*, **Ilgar Abdullayev***, Åsa Bjorklund, Thomas Hauling, Heng Wang, Åsa Segerstolpe, Raquel Firnkes, Connie Xu, Nuria Oliva Vilarnau, Mats Nilsson, Rickard Sandberg & Andras Simon. [Manuscript]

* Equal contribution

SCIENTIFIC PAPERS NOT INCLUDED IN THE THESIS

Ahmed Elewa, Carlos Talavera-López, Heng Wang, Alberto Joven, May Penrad, Zeyu Yao, Neda Zamani, Yamen Abbas, Gonçalo Brito, **Ilgar Abdullayev**, Rickard Sandberg, Manfred Grabherr, Björn Andersson, András Simon. (2017). Reading and editing the Pleurodeles waltl genome reveals novel features of tetrapod regeneration. [Manuscript is under revision at *Nature*]

CONTENTS

LIST OF ABBREVIATIONS

3C	Chromosome conformation capture
3C-cap	3C with sequencing capture
4C	Chromosome conformation capture coupled with sequencing
bp	Base pair
cDNA	Complementary DNA
ChIA-PET	Chromatin interaction analysis by paired-end tag sequencing
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
DNA	Deoxyribonucleic acid
H3K27Ac	Histone 3 lysine 27 acetylation
H3K4me1	Monomethylated histone H3 lysine 4
H3K4me3	Trimethylated histone H3 lysine 4
HAT	Histone acetyltransferase
ISS	In situ sequencing
kb	Kilobase
lncRNA	Long non-coding RNA
mESC	Mouse embryonic stem cell
miRNA	microRNA
mRNA	Messenger RNA
ncRNA	Non-coding RNA
nt	Nucleotide
Pol	Polymerase
polyA	Polyadenylated
pri-miRNA	Primary miRNA transcript
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RPKM	Reads Per Kilobase per Million mapped reads
rRNA	Ribosomal RNA
scRNA-seq	Single-cell RNA sequencing
sdRNA	snoRNA-derived RNA

snoRNA	Small nucleolar RNA
TF	Transcription factor
tRNA	Transfer RNA
tsRNA	tRNA-derived small RNA
TSS	Transcription start site
UMI	Unique molecular identifier

1 INTRODUCTION

An organism consists of many different cell types that dramatically differ in both structure and function. Deoxyribonucleic acid (DNA) encodes all the RNA and protein molecules that are needed to construct an organism. However, the complete DNA sequence of any organism, aka genome – be it a few million nucleotides (nt) of simple bacterium or a few billion nucleotides of a human – does not enable us to reconstruct the entire organism no more than words in any dictionary enable us to speak an actual language. What matters in both cases is how to use those words in a dictionary or elements in DNA sequences. For example, a neuron and a fibroblast have so distinct functions that it is difficult to imagine they contain the same genome. These differences in structure and function are results of complex processes of cell differentiation where the genomic sequence is not changed, instead cells accumulate different sets of RNA and protein molecules.

Soon after completing the sequencing of the human genome, it became clear that only a minor fraction of the human genome encoded for proteins (Venter, 2001). Early experiments suggested that there are about 50,000 - 100,000 transcribed genes, but genome-wide studies showed that there are approximately 20,000 protein-coding genes in the human genome (Pertea & Salzberg, 2010) and the vast majority of those genes from earlier findings are alternative transcript variants of the same genes. This number was considerably lower than expected given the fact that less complex organisms such as fruit flies and round worms seemed to have a similar number of genes. This was contradictory to the assumption that the complexity of an organism was related to the number of protein-coding genes they encode. Furthermore, only 1-2% of the human genome consists of protein-coding genes (Claverie, 2005). It was proposed that the fraction of non-coding genes could contribute to the complexity of an organism and that many of these regions could function as regulatory elements or through transcription into non-coding RNAs (ncRNAs) (Taft, Pheasant, & Mattick, 2007). Enhancers, one of the key regulatory elements, acting by increasing the expression of a gene, could also be expressed. One of the enhancers we identified in Paper II, was validated by another group (Groff et al., 2016), and also worked as non-coding RNA. It is the Linc-p21 locus, encoding for a long non-coding RNA, which plays a significant role in p53 signaling, tumor suppression, and cell-cycle regulation – demonstrating the overlaps between functions as well as definitions of these regulatory players.

1.1 REGULATION OF GENE EXPRESSION

Regulation of gene expression occurs at different layers: including transcription, RNA processing, translation, transport, degradation and protein stability. However, since the entire process starts with transcription, the transcriptional regulation is one of the most crucial steps. The transcriptional machinery of eukaryotes involves two complimentary regulatory components: the *cis*-acting elements and the *trans*-acting elements (**Figure 1.1**). The *cis*-acting elements are DNA sequences in the genome (coding as well as non-coding part) located in the vicinity of a gene they are regulating. The epigenetic information could also be overlaid onto the *cis*-acting regulatory elements. This comprises chromatin modifications and remodeling which creates an accessible region in the DNA for factors (*trans*-acting) to initiate the transcription. On the other hand, some epigenetic processes prevent trans-acting factors from binding to DNA by making chromatin inaccessible. The *trans*-acting elements are transcription factors (TFs) or other DNA-binding proteins that recognize and bind to specific DNA sequences in the *cis*-acting elements to initiate, increase or suppress transcription. TFs may regulate multiple genes, work in a combinatorial or complex manner to bind to the cis-regulatory elements at multiple binding sites thereby generating a huge catalog of precise and unique control patterns.

1.2 TRANSCRIPTIONAL CONTROL OF GENE EXPRESSION

Gene expression begins with transcription in the nucleus of a cell. Transcriptional control determines where, when and how often a gene is transcribed. The part where a gene starts to be transcribed is called the transcription start site (TSS). This site is in the middle of a region called the core promoter (**Figure 1.1**). RNA polymerase (Pol), an enzyme that catalyzes RNA synthesis, forms a chemical bonds (binds) with promoter. There are three types of polymerases in metazoans which transcribe specific classes of RNAs. The first one, RNA Pol I, transcribes ribosomal RNAs (rRNAs) which make up one of the most important and complex molecular machines called ribosomes, that orchestrates the synthesis of proteins. Ribosomal RNAs are the most abundant class of RNAs in the cell, comprising of 80 % of the total RNA in a cell. rRNA genes are present in multiple copies in eukaryotic genomes (Stults, Killen, Pierce, & Pierce, 2007). The second type of polymerase, RNA Pol II, transcribes genes that produce messenger RNAs (mRNAs), long noncoding RNAs (ncRNAs), and some of the small regulatory ncRNAs. Lastly, RNA Pol III, transcribes transfer RNAs (tRNAs), which are RNA molecules performing the transfer of amino acids to the ribosome where protein polypeptides are synthesized.

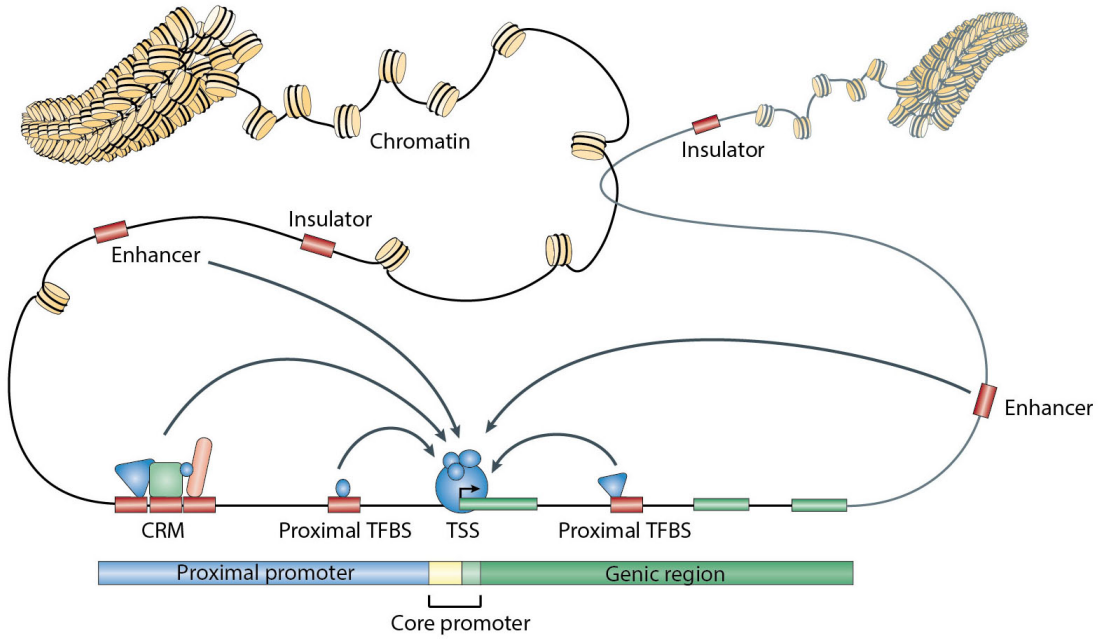

Figure 1.1: Schematic representation of gene regulation. DNA is wrapped around nucleosomes creating efficient and compact structure chromatin. Chromatin could be tightly organized (heterochromatin) or accessible to proteins in active form (euchromatin) cis-regulatory DNA sequences. These regulatory sequences are promoters (composed of proximal and core promoters), enhancers, insulators or silencers and binding of activating or repressive TFs can affect the rate of transcription initiation of the TSS either positively or negatively. Regulatory sequences such as enhancers could be located tens of hundreds ok kilobases (kb) away from their target promoters, as illustrated above. Figure modified from (Lenhard, Sandelin, & Carninci, 2012)

In order to bind and start transcribing, several other facilitating proteins are needed together with RNA polymerases. These proteins comprise general TFs which are able to bind the promoter region of all genes or many genes. The binding of the general TFs on their own results in low levels of transcriptional activity. This activity is increased or decreased by other sequence-specific TFs, estimated to be around 1400 in humans (Vaquerizas, Kummerfeld, Teichmann, & Luscombe, 2009), which bind to regions of the DNA including enhancers and silencers respectively. A gene can be regulated by several enhancer regions that may exist nearby or millions of nucleotides away from the gene, for example enhancer controlling the expression of sonic hedgehog (*SHH*) (Lettice et al., 2003). Most of the sequence specific TFs and the factors assembled at the promoter region interact with co-factors - proteins that do not directly bind to the DNA. Another multiprotein complex, mediator, interacts with both TFs and RNA Pol II, functioning as coactivator. Although the general TFs and the mediator complex are shared among all genes, TFs and cofactors can vary for the transcriptional machinery of each gene. Therefore, the change in the concentration of TFs and cofactors influences the timing and rate of transcription of genes, providing a mechanism of gene expression regulation.

Cis-regulatory DNA sequences are composed of two distinct elements: proximal elements (promoters) and the distal regulatory regions including enhancers, silencers, insulators and locus control regions (LCRs). These elements cooperatively act on their target genes and regulate their expression pattern (**Figure 1.1**).

1.3 PROMOTERS

The RNA polymerase II (Pol II) promoter regions are composed of two parts: the core promoter and the proximal promoter. Messenger RNAs, microRNAs and small nuclear RNAs are transcribed from Pol II promoters. The core promoter is the minimal part of the promoter enough to initiate the transcription by Pol II machinery and is located approximately 35 base pairs (bp) upstream or downstream of the TSS (**Figure 1.2**). The core promoter serves as the binding site of factors for assembly of the preinitiation complex (PIC) and it contains a few sequence elements. The consensus sequence of TATAAAA (TATA box) is located 26 to 31 bp upstream of the TSS, and its sequence may vary (Wong & Bateman, 1994). Though the TATA box was considered to be an essential part of the core promoter, it was discovered that only 24-32 % of the human core promoters contain the TATA box (Y. Suzuki et al., 2001; Yang, Bolotin, Jiang, Sladek, & Martinez, 2007). Another core promoter element is the TFIIB recognition element (BRE), located 3-6 bp upstream of the TATA box with the consensus sequence of G/C G/C G/A C G C C, is only recognized by TFIIB but not TFIID. The function of BRE is to repress basal transcription which is released upon the binding of activators. As the transcription start site is denoted as +1, the initiator element (INR) - simplest functional promoter that is able to direct transcription initiation without a functional TATA box, is placed from -2 to +4 having the consensus sequence of YYANWYY (Xi et al., 2007). TATA box and INR elements are most often found in promoters of protein-coding genes. The downstream promoter element (DPE), located downstream at +28 to +32 relative to the TSS, contains the consensus sequence of A/G GA/T C/T G/A/C, and functions in combination with the INR in TATA-less promoters (Hahn, 2004). The motif ten element (MTE) is an element located at +18 to +27 from the TSS, functioning independently from the TATA box and the DPE but cooperatively with the INR (C. Y. Lim et al., 2004). Another important element, the downstream core element (DCE), is located downstream of the TSS, which includes both MTE and DPE (D.-H. Lee et al., 2005). DCE is located at +10 to +45 relative to the TSS and its function is distinct from the DPE. All of these core elements (TATA box, INR, DPE, DCE and MTE) initiate the recruitment of transcription factor IID (TFIID) initiation complex to the promoter. It is believed that there is no universal core elements, and other core elements may still remain to be discovered (Gershenzon & Ioshikhes, 2005).

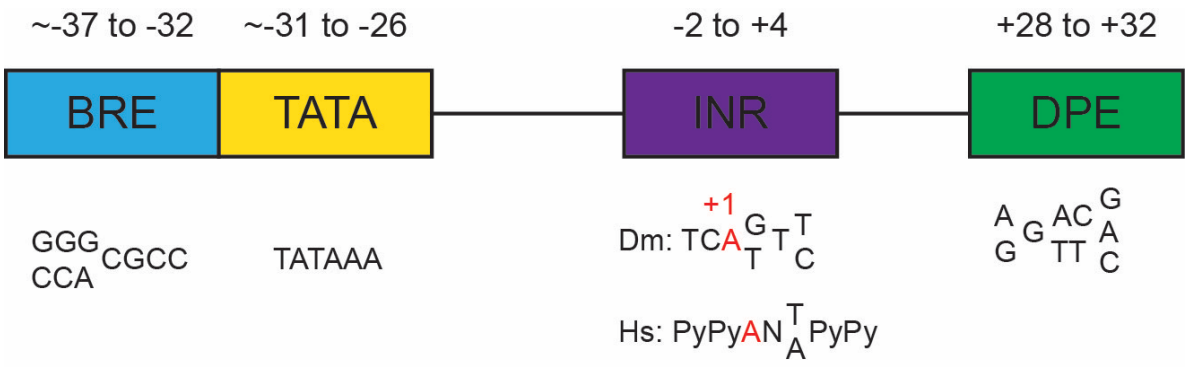

Figure 1.2: The core promoter for the RNA polymerase II. The relative positions of the core promoter elements: TATA box (TATA), initiator element (INR), downstream promoter element (DPE), and TFIIB recognition element (BRE) are shown. The consensus sequences of these elements are shown below each element. The transcription start site is indicated by "+1". Any specific core promoter may contain all, some, or none of these motifs. Inspired from (Butler & Kadonaga, 2002).

The proximal promoter is a DNA element located a few hundred bp to a few thousand bp at upstream of the core promoter and can be involved in altering the rate of transcription (Hurst et al., 2014). Interestingly, the ModENCODE consortium which was aiming to identify genome-wide functional elements in the genomes of *Caenorhabditis elegans* and *Drosophila melanogaster*, identified the proximal promoter element size as TSS ± 4000 bp (Huminiecki & Horbańczuk, 2017). A CpG island, 500-2000 kb GC rich sequence is considered as proximal element (Smale & Kadonaga, 2003). They are linked with approximately 60% of the human promoters. CpG islands contain multiple binding sites for the transcription factor Sp1 but the core elements have not been fully identified.

1.4 ENHANCERS

Enhancers are typically 50-1500 bp *cis*-acting DNA sequences that can increase the transcription of genes. They generally function regardless of orientation (whether they are upstream or downstream of their target promoters) and located at various distances from their targets (Shen et al., 2012; Visel, Rubin, & Pennacchio, 2009b). Enhancer was first identified in the tumor virus SV40 and was shown to increase transcriptional activities beta-globin gene (Banerji, Rusconi, & Schaffner, 1981). The simian virus SV40 enhancer contains 72 bp repeat sequences when deleted reduces the viral protein levels expressed in early stages of infection, eliminating the virus. After the discovery of the viral enhancer, the first enhancer in mice and humans were found to activate the immunoglobulin heavy chain gene in a tissue-specific (lymphocyte) fashion (Banerji, Olson, & Schaffner, 1983). A typical enhancer contains multiple transcription factor binding sites (TFBS) which are often conserved sequences with a certain degree of degeneracy which helps the binding of the TFs. Different TFBS are arranged

in a particular orientation to control the specificity of the enhancer. However, how an enhancer mediate activation of its target promoter is not fully understood yet.

The way in which enhancers stimulate transcription remains poorly understood and it is one of the main questions in the field (García-González, Escamilla-Del-Arenal, Arzate-Mejía, & Recillas-Targa, 2016). Since enhancers were first characterized based on their ability to increase the levels of target genes, the quantity of gene product is important. Historically, there have been several models suggested for understanding enhancer mode of action. Firstly, the proteins bound to promoters and enhancers may interact with each other by creating DNA loops (Rippe, Hippel, & Langowski, 1995; Saiz, Rubi, & Vilar, 2005) forming a multi-protein complex for transcription to take place. Secondly, the promoter and enhancer may not come close contact with each other, instead, the enhancer may direct the DNA element into specific regions in the nucleus where high concentrations of TFs are available to facilitate the transcription (Lamond & Earnshaw, 1998). Alternatively, enhancers may act via supercoiling of DNA, nucleosome remodeling and altering chromatin structure to create an accessible structure for recruitment of regulatory proteins to initiate transcriptions (L. A. Freeman & Garrard, 1992). More recently, two models have gained more attention to explain enhancer function: the binary model (Walters et al., 1995) and the progressive or rheostatic model (Ko, Nakauchi, & Takahashi, 1990). The binary model proposed that enhancers actually increase the probability of creating transcriptionally active loops rather than increasing the levels of gene expression (Bartman, Hsu, Hsiung, Raj, & Blobel, 2016; Fukaya, Lim, & Levine, 2016; Walters et al., 1995). However, the progressive model proposes that enhancers increase the number of RNA molecules transcribed from genes, but not the number of cells that initiate transcription (Chepelev, Wei, Wangsa, Tang, & Zhao, 2012). Currently, which of these models explain observed enhancer action is not fully resolved.

The identification of enhancers has been challenging for several reasons. First, enhancers are scattered across the genome that does not encode proteins (mainly 98 % non-coding). This means one should design genome-wide assays or computationally search for enhancers using billions of base pairs of sequences. Second, although it is known that they function in *cis*, their relative location to their target promoter (or promoters) is highly variable. They can be found a few kb up to a few million kb upstream or downstream of genes, as well as within introns. Moreover, they can bypass neighboring genes to regulate genes located more distantly along a chromosome, rather than acting on the closest promoter (Sahlén et al., 2015). And in some cases, single enhancers have been found to regulate multiple genes (Mohrs et al., 2001), which makes their functional annotation further complicated. Third, there are no known general sequence motifs or codes for enhancers, as opposed to the well-defined sequence code of protein-coding genes, making it extremely difficult if not impossible to computationally identify (with high confidence) enhancers from DNA sequence alone. Lastly, enhancers are known to be tissue-specific, so their activity could be restricted to a particular cell type, a time

point in life, or to specific physiological or environmental conditions. While this dynamic nature of enhancers permits their precise action (i.e. when, where and how much specific gene is expressed), it further complicates the discovery and functionally annotating of them.

Genome-wide studies of histone modifications have revealed new insights into transcriptional regulation (ENCODE Project Consortium et al., 2007; Roadmap Epigenomics Consortium et al., 2015). The first histone modification globally linked to distal regulatory regions was identified as monomethylated histone H3 lysine 4 (H3K4me1), whereas trimethylated histone H3 lysine 4 (H3K4me3) was predominantly enriched at gene promoter regions (Heintzman et al., 2007). Thus based on their histone H3K4 methylation status cannot be exclusively used to distinguish between enhancers and promoters, since histone H3K4me2 or H3K4me3 marks have also been detected in active enhancers (Barski et al., 2007; Core et al., 2014).

Enzymes with histone acetyltransferase (HATs) activity plays an important role in enhancer function. One of the well-studied of such enzymes is CBP/p300, a co-factor with HATs activity. It has been shown that p300 binding is an accurate predictor of in vivo enhancer activity in development (mouse) and 95 % of p300 in vivo binding is found at promoter distal regions (human) (Visel et al., 2009a; Yao et al., 1998). Furthermore, p300 binding sites overlap with DNase I hypersensitive sites (DHS) and expression of active genes during development (Visel et al., 2009a). It is proposed that p300 might recruit RNA Pol II to enhancers that are marked with H3K4me1 leading to transcribe those enhancer regions (eRNAs). Previously, eRNAs have been associated with enhancer function, but to what extend their involvement in enhancer function is still now well-understood (T.-K. Kim et al., 2010). Other HATs have also been shown to interact with enhancers (Krebs, Karmodiya, Lindahl-Allen, Struhl, & Tora, 2011). To help differential recruitment of cofactors different HATs are speculated to bind enhancer regions. Furthermore, HATs may modify TFs affecting their activity or protein interactions at their target enhancers.

Although genome-wide studies have shown correlation between enhancers and p300 or H3K4me1, this alone is not enough to accurately predict enhancer activity. Further studies showed the correlation between histone 3 lysine 27 acetylation (H3K27Ac) and active enhancers during ESC differentiation (Creyghton et al., 2010; Rada-Iglesias et al., 2011). The acetylation of enhancers may weaken nucleosome stability or make chromatin more accessible (Merika, Williams, Chen, Collins, & Thanos, 1998), which may help TFs to access their binding sites more efficiently.

1.5 NON-CODING RNAS

Non-coding RNAs (ncRNAs) are RNA molecules that are not translated into proteins. They have been divided into short ncRNAs (<200 nt) and lncRNAs (>200 nt) mostly due to limitations in column purification procedures. There are numerous ncRNAs available in the literature such as transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), microRNAs, small nucleolar RNAs (snoRNAs), small interfering RNAs (siRNAs), small nuclear RNAs (snRNAs), piwi-interacting RNAs (piRNAs), exRNAs and scaRNAs and the long ncRNAs such as Xist and HOTAIR (ENCODE Project Consortium et al., 2007). MicroRNAs are amongst the well-studied small ncRNAs. However, the lncRNAs, including the long intergenic ncRNAs (lincRNA), antisense RNAs (asRNAa) and intronic RNAs, are not as thoroughly investigated. It has been speculated lncRNA secondary structures might be conserved throughout evolution but not their sequences since many lncRNA sequences are poorly conserved (Johnsson, Lipovich, Grandér, & Morris, 2014). While miRNAs mainly function as post-transcriptional regulators of gene expression, lncRNAs can act both as positive and negative regulators, playing roles in epigenetic remodeling, chromatin structure and RNA stability (Vadaie & Morris, 2013).

1.6 MICRORNAS

A microRNA (miRNA) is approximately 22 nucleotides in length, small non-coding RNA molecule found in animals, plants and some viruses, and mainly functions in RNA silencing and post-transcriptional gene regulation. The first miRNA (lin-4) was discovered in *C. elegans* in 1993 by Ambros (R. C. Lee, Feinbaum, & Ambros, 1993). Although at the time it was not defined as a miRNA, lin-4 shared sequence complementarity and suppressed the mRNA of protein-coding gene lin14. For many years, this was considered as a unique case and no new miRNA was reported. Then in 2000 another miRNA, let-7, was reported, which played an important role in developmental timing in *C. elegans* and was shown to be highly conserved from nematode to human (Pasquinelli et al., 2000; Reinhart et al., 2000). Currently, there are 28,645 hairpin precursor miRNAs in Release 21 of the Mirbase database, expressing 35,828 mature miRNA products in 223 species (http://www.mirbase.org/). Out of these, 2588 mature miRNAs are identified in humans. About 30-60% of all human mRNAs are suggested to be under the regulatory control of miRNAs (Friedman, Farh, Burge, & Bartel, 2008).

Most miRNA genes are transcribed by RNA polymerase II and some of them by RNA polymerase III, producing primary miRNA transcripts (pri-miRNAs) that are long and might contain 5′ cap, polyA tail and 3′ modifications similar to pre-mRNAs (Cullen, 2004). In fact, many miRNA sequences are located within annotated genes for mRNAs (or other RNAs), which are often considered as host genes of these miRNAs. miRNA genes are not well defined experimentally and pri-miRNAs are not as extensively studied like mRNAs. About 40% of miRNA genes are estimated to lie within the introns or exons of other genes (Rodriguez,

Griffiths-Jones, Ashurst, & Bradley, 2004). Although it is possible that the miRNAs have their own promoter driving their expressions, it is often assumed that expression of host genes produces pri-miRNA transcripts that eventually processed into mature functional miRNAs.

In mammals, based on processing of primary transcripts, miRNAs are divided into two big classes, canonical and non-canonical. In the canonical pathway, the enzyme Drosha binds its regulatory subunit DGCR8, cleaving a pri-miRNA into hairpin structured precursor microRNA (pre-miRNA) which is approximately 60–70 nt long (Han et al., 2004; Y. Lee et al., 2003). As a result of cleavage by Drosha, the pre-miRNA often contains a 2-nt long 3′ overhang, and then it is exported from nucleus to the cytoplasm by Exportin5 (Exp5) (Yi, Qin, Macara, & Cullen, 2003). In the cytoplasm, dephosphorylation of GTP induces the release of pre-miRNA from Exp5 which then allows it to be cleaved by another RNase, Dicer, to produce a miRNA duplex intermediate of about 22 bp (Grishok et al., 2001; Ketting et al., 2001; Zhang, Kolb, Brondani, Billy, & Filipowicz, 2002). Finally, RNA induced silencing complex (RISC) containing the argonaute2 (Ago2) protein binds to the intermediate miRNA duplex and integrates the mature, single-stranded miRNA into the Ago:RNA complex (Hammond, Boettcher, Caudy, Kobayashi, & Hannon, 2001; Hutvágner & Zamore, 2002). The mature miRNA guides the RISC complex to 3'UTR of mRNAs the target mRNAs, where the recognition takes place primarily. The other strand, referred as the passenger strand, gets degraded due to its lower levels in the steady state and relative thermodynamic stability (Khvorova, Reynolds, & Jayasena, 2003). Sometimes both strands of the duplex become functional miRNA having two different target mRNAs. In non-canonical pathway, miRNA processing does not involve all of the factors from canonical pathway. For instance, some pre-miRNAs are produced by splicing, not by Drosha cleavage (Okamura, Chung, & Lai, 2008; Ruby, Jan, & Bartel, 2007) and pre-miR-451 is cleaved by Ago2, avoiding Dicer (Cheloufi, Santos, Chong, & Hannon, 2010). Some pri-miRNAs, for instance, endogenous shRNAs, siRNAs in mouse ES cells, are small hairpin RNAs that possibly serve as pre-miRNAs and Dicer can process them directly (Babiarz, Ruby, Wang, Bartel, & Blelloch, 2008). It is unclear how many non-canonical miRNAs are out there, but by using deep-sequencing experiments low abundance miRNAs are being identified and deposited to miRBase, though details on how these RNAs are processed has not been well-studied (Graves & Zeng, 2012).

1.7 TRNA-DERIVED SMALL RNAS

There are small RNAs that are derived from other non-coding RNAs. One of them is tsRNAs, 5'-phosphate, 3'-hydroxylated tRNA-derived small RNAs of about 30-34 nt in size (Haussecker et al., 2010). It has been previously shown that introduction of sperm tsRNA from high-fat diet mouse into normal zygotes changed the gene expression of metabolic pathways in early mouse embryos and created metabolic disorders (Q. Chen et al., 2016). Therefore, sperm tsRNAs could play an important role in epigenetic inheritance of diet-induced metabolic

disorders. There are two types of tsRNAs based on their biogenesis: Dicer-dependent and Dicer-independent. The Dicer-dependent tsRNAs can moderately down-regulate target genes in *trans* and been previously detected in mice but comprehensive structural and functional analyses had been lacking (Haussecker et al., 2010). In Dicer-independent biogenesis, a tRNA processing enzyme RNaseZ, an endonuclease, which processes the RNA so that it leaves a 3'-hydroxyl and 5'-phosphate at the cleavage site (Mayer, Schiffer, & Marchfelder, 2000).

1.8 SNORNA-DERIVED SMALL RNAS

Another class of ncRNA-derived small RNAs is snoRNA-derived RNAs (sdRNAs). There are two classes of sdRNAs. First, sdRNAs derived from H/ACA snoRNAs, are primarily 20–24 nt in length and originate from the 3′ end of snoRNAs. Second, sdRNAs derived from C/D snoRNAs, which are predominantly 17–19 nt or >27 nt in length (exhibiting a bimodal distribution) and mostly originating from the 5′ end of the snoRNAs (Taft et al., 2009). Due to high expression of some sdRNAs in human THP-1 cells, it is unlikely that these sdRNAs are result of RNA degradation (or RNA turnover), since their precursor snoRNAs are weakly expressed (Taft et al., 2009).

2 METHODS FOR STUDYING GENE REGULATION

This chapter is about several methods to study gene regulation discussed in this thesis.

2.1 QUANTIFYING RNA

Quantifying RNA enables us to understand many aspects of biological samples. Starting from northern blot (Alwine, Kemp, & Stark, 1977), one of the first and simplest methods for measuring RNA abundance using radioactively labelled RNA probes, followed by quantitative reverse transcription polymerase chain reaction (qRT-PCR) (W. M. Freeman, Walker, & Vrana, 1999) where RNA is converted to cDNA and measuring DNA amount using a dye, then later microarrays (Schena, Shalon, Davis, & Brown, 1995) using oligonucleotide probes for quantifying fluorescently labelled cDNAs (converted from RNA), nowadays we can measure RNA amounts from samples containing as little as pictograms of RNA using widely known technique called RNA sequencing (RNA-seq) (Lister et al., 2008; Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008; Nagalakshmi et al., 2008).

RNA-seq protocols starts with sample containing RNA. For that RNA needs to be extracted from biological samples. This could easily be done using standard column-based RNA extraction kits. Then either RNA is fragmented and then converted into cDNA (as in Illumina mRNA-seq protocols) or vice versa (as in Clontech Smarter protocols). One of the many modifications to the steps of standard RNA-seq protocol is incorporation of dUTP in the second-strand synthesis of cDNA, generating a strand-specific RNA-seq library (Parkhomchuk et al., 2009). Fragmenting cDNA followed by ligation of universal adapter sequences and DNA barcodes to the end of each cDNA fragment. cDNA gets amplified using PCR. At the end, the cDNA gets sequenced, producing millions of short reads, which is a partial readout of actual cDNA. Thus, RNA-seq does not sequence the entire RNA molecule or long cDNA converted version – it simply provides readout of small pieces (reads), but given a couple of millions of such reads it is possible to recapitulate the entire transcriptome.

Due to high costs of sequencing, often samples are pooled together – called multiplexing. DNA barcode is added to each sample to keep track of their sample identity. Sequencing machine reads that barcode as well and provides barcode information as well. After demultiplexing, samples are mapped to corresponding reference genome – i.e. reads from human samples are aligned (or mapped) to reference human genome, etc. Reference genomes should be downloaded and prepared (often indexed) according to sequence aligner's preferences. There are various publicly available sequence aligners for RNA-seq, such as TopHat (Trapnell, Pachter, & Salzberg, 2009), GSNAP (T. D. Wu & Nacu, 2010), STAR (Dobin et al., 2013), HISAT (D. Kim, Langmead, & Salzberg, 2015) and etc.

It is essential to check the quality of mapping. Fraction of uniquely mapped, multi-mapped and unmapped reads tell us a lot about the quality of the sample as well as the performance of the method. Looking at all samples together allows to set the right unique-mappability cut-off. Furthermore, we use FastQC program to check the average quality score for each base in the reads, calculate GC content, and find overrepresented sequences and etc. Sometimes, adapter and primer dimers take large fraction of samples (especially if starting RNA material is low) and FastQC can identify them as overrepresented. Also, there could be overrepresented reads from contamination by other species, such as bacteria. Using these kinds of feedbacks helps to design experiments better.

Mapping is usually followed by quantification of reads. For that we often use a metric called RPKM (reads per kilobase and million mapped reads), calculated by script (Ramsköld, Wang, Burge, & Sandberg, 2009) developed in our lab by Daniel Ramsköld. RPKM is calculated by using number of reads per gene, gene length (the part that could be uniquely mappable (Storvall, Ramsköld, & Sandberg, 2013)) and the total number of uniquely aligned reads (excluding reads that could not be assigned uniquely) per sample (sequencing depth). Eliminating the fact that samples will by chance have different total reads, and correcting for that in our analysis is called normalization. By that, we eliminate some of the technical differences. As a result, we get a table of expression values (RPKM) for genes and samples normalized by each gene length and sample depth. Now, we can compare samples, or sample groups, visualize their differences, cluster samples etc, depending on the need.

2.2 SINGLE CELL RNA SEQUENCING

A conventional RNA-sequencing protocols require high amount of input RNA, for example minimum of 0.1 ug of total RNA is need to perform Illumina TruSeq Stranded mRNA kit. For many applications, this protocol is fairly useful, but it has its own limitations. In order to obtain this mass of RNA, tens of thousands or even millions of cells must be utilized, resulting in the average profiling of the bulk samples, because often these cell populations are not homogenous. Thus, this approach eliminates and drowns the signal from rare cell populations in the initial sample. This could be a problem since those rare populations could carry critical information about the tissue being studied, for example, that could be a rare stem cell population that is composed of fewer cells that divide slowly, yet it is critical to replenish the tissue. This problem could be overcome by sorting cells before running the protocol, but cell sorting has its own limitations such as requiring fairly high number of cells to start with and a highly expressed cell surface marker unique to that population. Putting together, classical RNA-sequencing is still a powerful technique to study various aspects of biology considering its advantageous and limitations, pointing out the necessity for a new technique, such as single-cell RNA-sequencing, in which those constraints could be solved.

Several methods have been developed recently to overcome this problem and enable single cell sequencing from extremely low amounts (picograms) of RNA. An average single cell contains about 10 picogram of RNA, which is so low that with conventional RNA-seq methods this amount would be lost during pipetting steps. Therefore, single cell RNA-seq methods aim to minimize the RNA loss as much as possible, performing reactions in the same tube. This includes depositing the cell in a single tube, lysing, reverse transcribing and amplifying in the same tube. Furthermore, ribosomal RNA depletion and polyA enrichment steps are also omitted in order to prevent further loss. In order to evade sequencing ribosomal RNA, oligo dT primers are used to reverse transcribe the polyA containing RNAs, which are mostly mRNAs. As a result, with a few exception (Faridani et al., 2016; Sheng, Cao, Niu, Deng, & Zong, 2017), all single cell RNA-sequencing methods profile polyadenylated RNAs (Islam et al., 2011; Picelli et al., 2014) while missing all the non-polyadentylated RNAs.

Single cell RNA-sequencing methods apply different strategies to increase the amount of RNA enough for sequencing. One of the most widely used methods is PCR, which amplifies RNA exponentially, because it makes use of the newly synthesized DNA as a template too. Another method is called in vitro transcription, in which the RNA is transcribed from cDNA. This process is linear, which brings an advantage over PCR, since it is more robust in preserving the initial ratios between the gene products, because PCR can easily over-amplify even the small differences. However, in vitro transcription is slow and requires higher input material. One of the single cell RNA-seq methods – CEL-seq, uses one round of in vitro transcription in their protocol, barcodes the samples at the 3' end of the transcript, then pools samples and amplifies altogether(Hashimshony, Wagner, Sher, & Yanai, 2012). Due to the design, CEL-seq is biased towards the 3' end. On the other hand, single tagged reverse transcription (STRT) method adds barcode at the 5' end of the transcript, making it 5' biased method (Islam et al., 2012). STRT method also allows pooling multiple samples together because of initial barcoding step. Additionally, STRT method has incorporated a smart strategy of counting molecules using unique molecular identifiers – random 5 nucleotide long sequence added additional to sample barcode (Islam et al., 2012; Kivioja et al., 2011). This principle relies on the assumption that it is unlikely for two reads originating from two molecules of the same mRNA will contain identical UMIs, allowing us to use the number of UMIs as an absolute molecule count for each gene (Islam et al., 2014).

While both CEL-seq and STRT methods have biases towards both ends of the transcripts, another method, Smart-seq (Ramsköld et al., 2012) and Smart-seq2 (Picelli et al., 2014), has overcome this problem by sequencing the whole transcript. Smart-seq relies on template switching, which enables both first and second strand synthesis one after another in the same reaction tube, providing more even read coverage across transcripts than polyA-tailing methods

(Ramsköld et al., 2012). An improved version of this method, Smart-seq2, provides even better coverage than Smart-seq and also increased the sensitivity of detecting RNA molecules, which is 40 % in Smart-seq2 (Qiaolin Deng, Ramsköld, Reinius, & Sandberg, 2014; Picelli et al., 2014), whereas STRT captures only 12.8 % of RNA molecules (Macosko et al., 2015). However, since the entire transcript is being sequenced in Smart-seq approach, in order to be able to achieve desired read depth, there is a limitation in pooling multiple samples. Additionally, with Smart-seq2 one can study allelic gene expression and expression of different isoforms, neither of the other methods is able to provide this kind of information. Therefore, Smart-seq2 is well-suited for studying hundreds or even a few thousands of cells in depth, while 5' and 3' methods are extremely powerful to analysis of tens of hundreds of thousands of cells.

Another powerful method to study single cells is Drop-seq (Macosko et al., 2015), which encapsulates cells in tiny droplets for parallel analysis. This approach uses nanoliter-scale droplets – spherical compartments formed by combining aqueous and oil flows very precisely in a microfluidic device. These droplets enable performing reactions in nano- liter-sized reaction chambers. After dissociating a tissue, each individual cell gets encapsulated into a droplet together with a bead (microparticle) containing a barcoded primer. Cells get lysed inside the droplet, mRNAs bind to the primer sequences and reverse transcribed into cDNAs. These cDNAs are bound to the microparticles and carry a unique barcode, which allows to pool and amplify all the samples together. Each bead contains three parts, a common sequence called PCR handle to enable PCR amplification, a unique cell barcode and a unique molecular identifier to be able to digitally count mRNA molecules. At the end of each primer there is a stretch of 30 Thymine nucleotides called oligo dT which bind to the polyA tail of mRNAs and other polyadenylated transcripts. Samples get pooled, amplified and sequenced using NGS.

Recently, new technique (Small-seq) have been developed to capture small RNAs at a single cell level, which was not possible previously (Faridani et al., 2016). These small RNAs include micro RNAs (miRNAs), small RNAs derived from small nucleolar RNAs (sdRNAs) and small RNAs derived from transfer RNAs (tsRNAs). Incorporation of UMI sequences enabled counting number of molecules. Conventional miRNA protocols often include gel size selection which limits the automation. However, Small-seq overcomes this problem by skipping the size selection and blocking the most abundant ribosomal RNAs with blocking oligos. One of the biggest advantageous of this technique is being able to cluster different cell types using only a few hundred expressed miRNAs, as opposed to other single cell methods capturing few thousand expressed mRNAs. Small-seq contains reads from ribosomal RNAs and protein coding genes, which could come from degradation products of larger transcripts or could be novel small RNAs derived (properly processed by enzymes) from precursors and have some function. At this point, more experiments are required to validate these results further.

Single-cell RNA-sequencing is a powerful technique allowing new discoveries that was initially not possible using bulk sequencing. By this we could study cellular heterogeneity at an unprecedented fashion, e.g. Human Cell Atlas (https://www.humancellatlas.org/) aims to comprehensive map and discovery of all cell types in human body. Discovering rare cell populations in general (Grün et al., 2015) and particularly for tumor formation and drug resistance (Patel et al., 2014) is also one of the key advantageous brought by single cell techniques.

2.3 SPATIALLY RESOLVED TRANSCRIPTOMICS

During the last century, optical microscopy and tissue staining has been widely used to study the tissue landscape, but these methods lack the elucidation of the genetic information. In order to obtain genetic information, tissues had to be dissociated and nucleic acid was extracted, which resulted in loss of spatial information. Most of the scRNA-seq methods also rely on dissociation of single cells from tissue resulting in loss of spatial information. There are two major approaches for spatial transcriptomics – imaging and sequencing based methods. Imaging based methods use fluorescently labelled DNA probes complementary to a target RNA sequence. In order to obtain sufficient signal, imaging-based methods, such as single molecule *in situ* hybridization (smFISH), hybridizes multiple probes to each of the target RNA sequences (Femino, Fay, Fogarty, & Singer, 1998). A new technology called multiplexed error-robust fluorescence in situ hybridization (MERFISH) can detect the position, identity and copy numbers of thousands of RNA molecules inside a single cell (K. H. Chen, Boettiger, Moffitt, Wang, & Zhuang, 2015).

There have been a few revolutionary methods in the transcriptomics field. One of them is spatially resolved in situ RNA and DNA molecule detection techniques, such as in situ sequencing (ISS) (Ke, Mignardi, Hauling, & Nilsson, 2016; Ke et al., 2013). ISS enables sequencing nucleic acids at a single cell level directly on the tissue slices. ISS is based on the use of padlock probes designed to bind specifically to a mRNA of interest and are circularized by a ligase upon binding. Nano blobs of DNA are generated by rolling circle amplification (RCA) of the circularized padlock probes. These blobs can be detected by hybridization of a fluorescencently labelled primers, which allow sequencing of the molecular barcode originally carried by the padlock probe. Another method is spatial transcriptomics (Ståhl et al., 2016), which allows studying expression of transcripts of tens of cells, preserving spatial localization in a given tissue section. First, freshly frozen tissue section is placed on a chip which contains an array of 100 μm unique sequence-barcoded oligo-dT capture probes containing sequencing adaptors. Then the image of the tissue is taken, recording the relative positions of cells to the array. Once the sample is permeabilized, the transcripts diffuse into the array. cDNA synthesis takes place on the chip, creating a library for sequencing. Since each read contains barcode carrying spatial information, they could be mapped back. However, currently spatial

transcriptomics cannot provide single-cell resolution. Although ISS is a promising tool, it is also less efficient due to bottlenecks in sample imaging, molecular processes, data handling, and interpretation.

2.4 STUDYING GENOME ARCHITECTURE

Simplistically, a genome is composed of long stretch of all DNA sequences in a cell. We usually work with linear DNA sequences, but in reality, chromatin (bundle of DNA and histone proteins) is compacted into precise three-dimensional (3D) structure that enables its function. Chromatin undergoes further condensation to create a structure called chromosomes. When working with genomes, we categorize the genomic data into chromosomes – linear DNA sequences. It is easier to identify and study the linear DNA sequence than its 3D structure. There are various methods to shed a light on 3D genome architecture. Starting with light microscopy, chromosomes were studied during metaphase of mitosis. Although microscopy techniques provide single cell measurements, they lack the resolution to identify interactions between specific regulatory elements, such as promoters, enhancers and etc. The development of the next generation sequencing (NGS) enabled us to study and begin to uncover 3D organization of a genome.

The first method to study the interaction of two genomic loci, chromosome conformation capture (3C), was developed by Job Dekker in 2002. 3C relies on strengthening the interaction between two genomic loci using formaldehyde cross-linking, followed by digestion of chromatin with restriction enzyme, performing proximity ligation where intra-molecular ligations are preferred over inter-molecular, and finally amplifying and detecting the ligated fragments using PCR with known primers. The restriction enzyme, Hind III, detects 6 bases, therefore called 6-cutter. 3C is considered as one-vs-one method, since it only allows to study two regions at a time, therefore, extremely low-throughput. More recently, Dekker developed a genome-wide method, called Hi-C, which allows to identify the interaction between all genomic loci, thus making it all-vs-all technique. Hi-C also starts with cross-linking of genomic material, which is like taking a snapshot of all the interactions at the time of formaldehyde treatment. Followed by digesting with restriction enzyme, but before proximity ligation, a key novelty of Hi-C was filling the digested DNA ends with biotinylated nucleotides. This allows the pull-down of biotinylated material, helping to get rid of all the background arising from not interacting regions. Biotinylated yet unligated DNA ends are also removed (by using the exonuclease activity of T4 DNA polymerase). Then ligation products are sheared into smaller fragments using sonication (sound waves breaking the DNA into pieces). Then adapters are added, amplified and library is sequenced like in standard protocols. Resulting libraries are sequenced by paired-end sequencing where a given DNA fragment is sequenced from both ends, carrying twice more information about that fragment. Interaction of two loci, when

captured by Hi-C, results in chimeric product, where both pieces could be computationally identified, which provides the basis for interaction of those two loci.

The restriction digestion step has been modified in order to increase the resolution of 3C-based methods. Initially, 3C as well as Hi-C, were based on 6-cutter restriction enzyme, which is the main determinant of the resolution. Using a 4-cutter restriction enzyme would increase the number of total restriction fragments about 16-fold, which in turn leads to 256-fold higher number of pairwise contact contacts. 4-cutter was first used in 4C method – circular chromosome conformation capture; also, known as chromosome conformation capture-on-chip (van de Werken et al., 2012; Z. Zhao et al., 2006). 4C investigates the interaction between single loci and the rest of the genome, making it 1-vs-all technique. One needs to sequence really deep if studying larger genomes with 4-cutters, since the total number of pairwise contacts would be extremely high. For instance, the finest resolution obtained in mammalian (human) genome just using standard Hi-C with 4-cutter, has generated 1 kb resolution by using 4.9 billion chromatin contacts (Rao et al., 2015). Using Hi-C with 4-cutter, on the other hand, is more suitable to study animals with smaller genomes, such as fly (Sexton et al., 2012), where the total number of possible pairwise contacts were significantly reduced compared to that of mammalian genomes. Furthermore, mechanical shearing (Fullwood et al., 2010) and enzymes such as DNase I (Ma et al., 2015) and micrococcal nuclease (MNase) (Hsieh et al., 2015) has been used to digest chromatin for 3C-based applications. Therefore, depending on which organism is being studied and the desired resolution, various versions of 3C techniques are available.

2.5 ASSEMBLING A NEW TRANSCRIPTOME

Genomic information for a particular organism is not always available due to various reasons, such as dealing with a non-model organisms, or high costs of sequencing extremely large genomes. In that case, studying a transcriptome, set of all transcribed genes in an organism, provides us valuable information. There are major challenges in transcriptome assembly (reconstruction) as opposed to genome assembly. While genomic sequencing depth is usually similar across the genome, the transcriptomic read coverage varies quite significantly, because the variation in coverage indicates the variation in gene expression. Transcriptomic data could also be strand-specific, unlike genome assembly data. Moreover, different isoforms from the same gene makes the reconstruction complicated, because they share the same exons, and may result in assembly of spurious or ambiguous transcripts that require further functional annotation. Reconstructing a transcriptome could be done either with the assistance of a genome, or using de novo approach without the help of reference genome. One of the most widely used de novo transcriptome assembler is Trinity (Grabherr et al., 2011).

There are three independent modules in Trinity: Inchworm, Chrysalis, and Butterfly, which are implemented sequentially. Trinity first creates individual de Bruijn graphs, which corresponds to complex transcriptional network for each gene or locus, and processes them independently. First, Inchworm generates unique sequences (contigs), which are often enough to define full transcripts for the dominant isoform, filtering out non-unique portion of isoforms. Chrysalis combines those contigs into complete de Bruijn graphs. Finally, Butterfly processes the de Bruijn graphs, locating the paths that read pairs take within the graph, eventually reporting full-length transcripts for alternatively spliced transcripts, and separating paralogous genes.

3 REGENERATION AND STEM CELLS

3.1 STEM CELLS

From evolutionary perspective, cells evolved as self-sufficient individuals, and these cells still dominate our planet. However, most of the cells in our bodies are specialized and they are part of a multicellular community. These cells have lost features required for surviving individually and instead obtained properties helping our bodies survive as a whole. According to conservative definition of cell types, there are more than 200 differently defined cell types in the human body that work in a collaborative manner (Alberts, Johnson, Lewis, Morgan, Raff, Roberts, & Walter, 2014a). Out of these cells, stem cells are the most interesting and important cell types. Stem cells are specialized in providing a fresh supply of differentiated cells, constantly replacing the tissues, repairing and regenerating whenever necessary. While many tissues renew, and repair themselves, some others do not. Therefore, once those cells are lost, they cannot be reversed, enabling the loss of function of that particular region permanently, causing blindness, dementia, deafness and etc. Although they share the same genome, stem cells, as well all the specialized (called differentiated) cells are enormously diverse in structure and function.

Embryonic stem cells (ES cells) are pluripotent stem cells derived from the inner cell mass (ICM) of a blastocyst, a mammalian embryo at an early stage. ES cells were first derived from pre-implantation mouse embryo in 1981 (Evans & Kaufman, 1981). Almost two decades later, a breakthrough embryonic research happened - human ES cells were derived from the blastocyst (Thomson et al., 1998). ES cells can differentiate into all types of the cells (cells from all three germ layers, i.e. ectoderm, mesoderm and endoderm) in the body. They can be grown and propagated *in vitro* culture media. While human ES cells are approximately 14 μm, mouse ES cells are smaller - approximately 8 μm (Zwaka & Thomson, 2003). There are mainly three TFs that are highly expressed in ES cells and play a crucial role in maintenance of ES cells. These are SOX2, OCT4 and NANOG. For instance, ES cells cannot be derived from the Sox2-defficient mouse embryos (Avilion et al., 2003). Furthermore, the deletion of Sox2 results in loss of pluripotency in ES cells and their ability to differentiate. Although overexpression of OCT4 can rescue the Sox2-defficiency phenotype, the overexpression of Sox2 downregulates the expression of its target genes such as *Nanog*, reduces pluripotency and induces differentiation (Kopp, Ormsbee, Desler, & Rizzino, 2008). These results indicate that precise control of expression levels of SOX2, OCT4 and NANOG is critical for the maintenance of stem cell renewal and pluripotency. Overall, ES cells have enormous potential in medicine since they could be used to repair damaged tissues, use as models to study genetic diseases and etc.

3.2 REGENERATION AND REPAIR

Many tissues in the body are not only self-renewing but also self-repairing, which is mainly due to stem cells and their control mechanisms that receive feedbacks regarding the regulation of their behavior and maintain of the homeostasis. However, natural repair mechanisms have limited capabilities. When neurons in our brain die (as in Alzheimer's disease) they are not replaced and when heart muscle dies due to lack of oxygen (as in a heart attack) it is not replaced with a new heart muscle. While some fish can regenerate rays of their fins (M. Suzuki et al., 2006), neonatal mice and human children can regenerate digit tips (Illingworth, 1974), the regeneration in majority of vertebrates is very limited and varies greatly.

Some animals do far better than humans in regenerating their entire organs, such as whole limbs, after amputation. There are some invertebrate species that can even regenerate the entire tissues of their body from a single somatic cell. A freshwater flatworm, *Schmidtea mediterranea*, or *planarian*, is a centimeter-long organism capable of extraordinary regeneration capability: a small tissue section taken from almost any part of the body will restructure itself and will give rise to a completely new animal. When this animal is starved, it goes through a process called degrowth, where the number of cells are reduced without losing the proper body proportion (Alberts, Johnson, Lewis, Morgan, Raff, Roberts, & Walter, 2014b). These flatworms can reduce their body size down to twentieth of its original size and will grow back when the necessary nutrients are available. This phenomenon is explained by cell cannibalism, where differentiated cells die and the recycled nutrients are absorbed by neoblasts, undifferentiated stem cells constituting about 20% of the cells in the body. As a result, neoblasts can grow, divide and differentiate into necessary cells replenishing the body. This is an incredible ability, and perhaps somehow linked to regeneration, without affecting survival or fertility of the animal.

Furthermore, some vertebrates such urodele amphibians (salamanders) show remarkable regenerative abilities. They can regenerate many organs such as limb, heart, brain, lenses and etc. One of the widely studied salamanders are newts, belonging to salamander subfamily Pleurodelinae. Newts go through full metamorphosis, unlike axolotls that reach adulthood without going through metamorphosis. Regeneration mechanism can vary between larva and adult newts. It was previously shown that newts can switch the cellular mechanism for limb regeneration from a progenitor-based mechanism (larval mode) to a dedifferentiation-based one (adult mode) (H. V. Tanaka et al., 2016). They demonstrated that while adult newts use muscle cells in the stump during limb regeneration, larval newts recruit satellite cells for the same purpose.

3.3 SALAMANDER LIMB REGENERATION

Limb regeneration was first studied by Spallanzani almost 250 years ago in his 'Reproduction of the Legs in the Aquatic Salamander' within his *An Essay on Animal Reproductions*. Surprisingly, many of the most significant features of limb regeneration defined by Spallanzani still remain unresolved today. During limb regeneration process in adult salamanders, upon limb amputation, the differentiated cells seem to return to an embryonic-like state by first forming a blastema - a small outgrowth that looks like embryonic limb bud. The blastema then grows and its cells differentiate to form a correctly patterned replacement for the limb that has been lost, in what looks like a recapitulation of embryonic limb development (**Figure 3.1**).

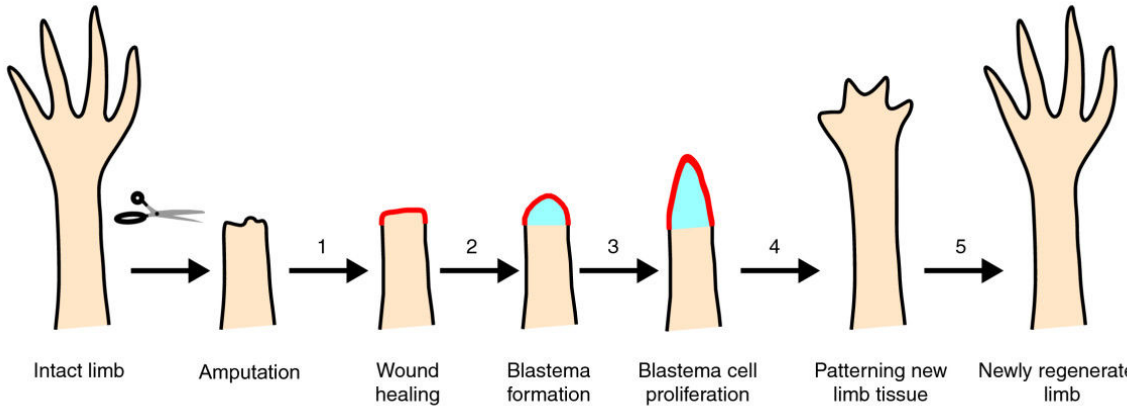

Figure 3.1: Key events of during salamander limb regeneration (Whited & Tabin, 2009). After amputation, the wound gets covered (step 1) by epidermal cells (called wound epidermis) migrating from the stump and forms the apical epidermal cap (AEC, red). Then around post-amputation day 14-19, the cells below AEC gives rise to blastema (blue) beneath the AEC (step 2) (H. Wang & Simon, 2016; C.-H. Wu, Huang, Chen, Chiou, & Lee, 2015). Blastema continues to proliferate (step 3) and starts to differentiate into diverse cell-types within the newly formed limb (step 4). A newly formed limb continues to grow until it gets the shape of fully functional original limb (step 5). The entire process takes about a month in adult newts.

Skeletal muscle cells seem to be one of the largest contribution to the blastema. Upon reentering the cell cycle, these multinucleate cells first dedifferentiate, and breakdown into mononucleated cells. These cells then proliferate and ultimately redifferentiate giving rise to one or more final cell types. Whether they redifferentiate only into muscle, or many other types of cells in limb is not fully understood. Current lineage tracing experiments performed using genetic markers, indicate that (contrary to previous belief) these cells are restricted. This means

muscle-derived cells give rise only to muscle, epidermal cells give rise to epidermal cells and connective-tissue cells only create connective tissues. Unlike flatworm, the cells in the adult vertebrate are less adaptable: they can work in coordination to replace or regenerate to varying degree, but each cell type is further away from being totipotent. Therefore, why salamanders and newts can regenerate many body parts, including an entire limb, still remains a profound mystery in biology.

The length it takes for a salamander to regenerate varies by species, body size and age. As they get older, salamander's ability to regenerate declines, but old salamanders are still able to continue to regenerate missing or damaged tissues. Typically, smaller larval salamanders regenerate faster than terrestrial salamanders (Young, Bailey, & Dalley, 1983). While a juvenile axolotl can regenerate a limb in approximately 40-50 days, terrestrial ones take much longer. Different terrestrial ambystomatid species show a great range of variation in their regeneration rate: *Ambystoma tigrinum* regenerates a limb in 155-180 days; *A. texanum* in 215-250; *A. maculatum* in 255-300; and *A. annulatum* does so in 324-375 days (Young et al., 1983). So, aging, body size and different species contribute to variation in regeneration rate, in addition to probably individual variation within the same species.

4 AIMS

The overall aim of this thesis was to study gene regulation, therefore we needed to develop molecular and computational tools. The goal at the beginning of my doctoral study was to develop both experimental and computational methods, implement, generate and analyze the data. Furthermore, we were also aiming to implement these methods and shed a light on stem cells biology and regeneration in salamanders.

4.1 SPECIFIC AIMS

The specific aims of the individual papers in this thesis are:

Paper I: Our goal was to reconstruct *de novo* transcriptome for red-spotted newt *Notophthalmus viridescens*, therefore I optimized dUTP method for strand-specific RNA-seq library generation and used the state-of-the-art computational approaches such as Trinity software.

Paper II: In order to generate genome-wide map of regulatory interactions with a high enhancer resolution, we optimized and combine Hi-C with sequence capture protocols, implemented on mouse embryonic stem cells.

Paper III: We aimed to study the role of small RNAs in individual cells, therefore, we developed single-cell small RNA-sequencing method, implemented in human embryonic stem cells and generated computational pipeline for the analysis of the data.

Paper IV: With the purpose of deciphering the heterogeneity in the newt blastema, our goal was to generate single-cell RNA-sequencing library for regenerating limb from red-spotted newt *Notophthalmus viridescens*, and analyze the single-cell data.

5 RESULTS AND DISCUSSION

5.1 PAPER I

Although salamanders have been studied for couple of centuries, we still have limited information about how they regenerate their body parts. The salamander regeneration research was partially hindered due to unavailability of comprehensive genomic information at the time I started my PhD. We aimed to create a good quality transcriptome for the community. I first experimented on extracting total RNA from the salamander red spotted newt *Notophthalmus viridescens*, and performed standard Illumina library preparation. A new technique, referred as dUTP method (Parkhomchuk et al., 2009), was published for preparing directional RNA-seq libraries for sequencing, and later comparative analysis of strand-specific RNA sequencing methods showed that dUTP method is the best amongst other directional protocols (Levin et al., 2010). This meant one would know which strand the read is originating from. The loss of RNA transcript polarity is one of the weakness of RNA-seq technique, because strand information is normally lost during the library preparation due to conversion of RNA (which is directional and single-strand) into double-strand DNA where both DNA strands are amplified during PCR. At the time Illumina TruSeq RNA Sample Prep Kit was just released. I modified TruSeq kit according to dUTP protocol and created strand-specific libraries for newt tissues. We believe having strand-specific library considerable helped the transcriptome reconstruction.

The transcriptome reconstruction was performed without the help of reference genome. We used Trinity (Grabherr et al., 2011) software which has been widely used amongst scientists for *de novo* transcriptome reconstruction. We pooled all reads from 6 samples and performed reconstruction. A few times the run failed due to memory problems, because Trinity is very memory-intensive program, especially the Inchworm and Chrysalis steps. A simple recommendation is to have 1 Gb of RAM per 1 million read pairs. A resulting transcriptome was a fasta file of hundreds of thousands of contigs. With the help of my main supervisor Rickard Sandberg, we developed a pipeline for validation of these contigs. Aligning contigs to publicly available DNA and protein sequences enabled us to assess the quality of transcriptome as well functionally assign newt transcripts. There were very few newt cDNAs and ESTs in Genbank at the time for direct comparison between newt contigs and newt transcripts. Therefore, comparing to other newt species such as axolotl, Japanese fire belly newt and Iberian ribbed newt as well as closest amphibian frog helped us further assess the transcriptome and inferred proteome of red spotted newt. It was interesting to observe that newt and frogs had similar numbers of proteins with different PfamA domains, yet there were about thousands of newt proteins that did not have any PfamA domain (**Paper I, Figure 2c**). This could be due to

false assembly by Trinity or from a more optimistic point of view, novel putative newt-specific proteins.

Indeed, some people in salamander field believes in the importance of newt-specific genes. An important one of those genes is Prod1 - retinoid-inducible gene encoding for a glycolipid-anchored protein, a member of the Three Finger Protein (TFP), which is expressed in a graded manner on the proximodistal axis (Geng et al., 2015; Kumar, Gates, Czarkwiani, & Brockes, 2015). Prod1 has been specifically linked to salamander regeneration, and identified in nine different salamander species, yet they have not identified any TFP member corresponding to Prod1 in non-salamander species (Kumar et al., 2015). When the expression of Prod1 was disrupted by injecting synthetic mRNAs to the fertilized newt eggs, the digit formation was not detected and Bmp2-positive cells were eliminated. Given that Bmp2 is a critical cytokine for the digit formation in amniotes, lacking Bmp2-cells and blocking digit formation indicated the necessity of Prod1 expression (Kumar et al., 2015). The same study also notes that loss of Prod1 has no effect on embryonic, larval or limb development before the stage of condensation of the radius/ulna and digits I and II. In general, more studies are needed to investigate the further role of newt-specific genes including Prod1.

There were some additional factors that affected the transcriptome assembly. First of all, longer library insert size resulted in fewer and longer contigs. If the insert was shorter than 200 bp, since we sequenced reads 100 bp from both ends, there would be a lot of overlapping reads wasted. Fortunately, the insert size we obtained was about 500 ± 25 bp. Having sharp insert size distribution was thought to help the reconstruction and recommended by Trinity. We performed de novo assembly using fewer samples for various reasons. First, to avoid memory problems in Trinity, then to be able to obtain tissue-specific transcriptome (particularly brain-specific) and finally to compare Trinity assemblies as a function of input reads. We concluded that assembly with all samples generated transcriptome comprehensive enough to cover all the Gencode cDNAs (**Paper I, Figure 1**), whereas tissue-specific transcriptomes covered impressively high number of cDNAs, lacking only a few cDNAs. In fact, a recent study evaluated the completeness of publicly available salamander transcriptomes, where newt transcriptome from **Paper I** had a BUSCO (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) score of 87% completeness, while the transcriptome from the competing study (Looso et al., 2013) by Braun lab got only 30% (Table 1 from (Bryant et al., 2017)). Overall, these results indicated that our newt transcriptome project was well-designed, analyzed using state-of-the-art techniques and contained comprehensive genomic information for protein-coding and non-coding genes as well as proteins (inferred) of red spotted newt, where a few key biological questions (such as importance of miRNAs, cell cycle inhibitors, tissue-specific genes and UTRs) were discussed (see **Paper I** for details).

5.2 PAPER II

Enhancers are distal elements important for regulation of gene expression. They have been effectively identified during the last decade using various techniques such as ChIP-seq, DNase hypersensitivity assays, STARR-seq and etc. Although these techniques help us locate the enhancers, by design they are not able to tell which genes are transcriptionally regulated by which enhancers. Since we know that enhancers regulate genes that are closest to them (considering linear genome), as well as genes located further away in the linear genome. They might also regulate more than one genes, for example, closest gene and gene(s) further away. Therefore, there was a need for a technique to sort out this problem.

Developing a technique able to identify genome-wide regulatory interactions was my main project. When I started my PhD in 2011, we discussed current state-of-the-art methods together with my co-supervisor Pelin (Akan) Sahlén. Initially, she has tried to implement 3C technique and was planning to couple 3C with sequencing capture (3C-cap). In fact, from the beginning, throughout all our experiments we always performed 3C as a control. Theoretically, 3C should also work in order to obtain regulatory interactions. To our surprise, majority of the interactions we obtained from 3C spanned relatively short linear distance in the genome, meaning the distal region identified was very close (often less than a kb) to the promoter. We did not trust those interactions, since they could arise from unligated fragments and be considered as background. Since similar methods were also ignoring such interactions, for instance, if ChIA-PET (Fullwood et al., 2009) identified interactions (between two PETs) were shorter than 3 kb, they were considered self-ligation PET, and were removed. Even with a lot of self-ligation products, 3C-cap still contained valuable regulatory information, and perhaps we underestimated that. Years later, the first method of such kind (called Capture-3C) that was published (Hughes et al., 2014) was indeed a much less genome-wide version of our 3C-cap. Not surprisingly, we calculated that more than 90% of interactions identified by Capture-3C was short-range (**Paper II, Supplementary figure 1**). One possible explanation for Capture-3C interactions being short is that, capture probes are much more likely to bind to self-ligation products than a chimeric ligation product carrying "useful" information between two distant DNA fragments. Perhaps self-ligation products will form near perfect complementarity with capture probes. Pelin referred this as "self-ligation products are like magnets to the capture probes". Chimeric ligation products, instead, would bind loosely to the capture probes and might fall off spontaneously, being replaced by self-ligation products that have a higher affinity towards capture probes.

Despite a lot of failed experiments, we continued to optimize and develop HiCap where we combined Hi-C with sequence capture. Unfortunately, we later find out that, our failed experiments were mainly due to miss-communication between authors. One of the authors who was performing the first step of Hi-C was supposed to provide us formaldehyde cross-linked

nuclei, but instead was providing formaldehyde cross-linked cell lysate. We were implementing Hi-C assuming the starting material was nuclei, but it was not just nuclei but the whole content of the cell - all sorts of proteins, RNA and organelles were present and interfering Hi-C reactions. After optimizing many steps in Hi-C, such as using different ligation enzymes, ligation conditions, dTNPs, T4 DNA polymerase for removal of biotinylated but unligated ends, phenol purification step and etc, finally we decided there might be something wrong with starting chromatin material. I solved this problem by first culturing new cell line, U2OS - a cancer cell line, and Hi-C worked well on them. Then I isolated nuclei from the starting material – cross-linked cell lysate, then implemented HiCap and it worked well. Although it was a good learning experience for me to be able to optimize steps in experiments, we lost about a year during this period.

Introduction of a few novel steps enabled HiCap to identify genome-wide regulatory interactions with highest resolution. The resolution means what is the average fragment length that contains a regulatory region. Using 4-cutter restriction enzyme enabled HiCap to have about 8x higher resolution for promoters and about 5x more resolution for identifying distal regions or distal elements (potential enhancers) compared to competing method CHi-C (Schoenfelder et al., 2015) which employed 6-cutter restriction enzyme.

In HiCap we observed many potentially interesting results but we did not have enough evidence to support the claims that comes along with those findings. For instance, we did not observe any significantly different connectivity between super enhancers and their target promoters compared to that of other enhancers (data not shown). Super enhancers are region of the genome containing multiple enhancers that is mutually bound by an array of TFs to drive transcription of key genes involved in determining identity of a cell (Hnisz et al., 2013; Whyte et al., 2013). Furthermore, when we look at the distribution of interaction distances to TSS and the direction from TSS, there was imbalance or unequal distribution between upstream and downstream interactions. No matter how we binned this observation stayed valid, and we are not sure how to explain that. Moreover, we observed promoter-promoter and enhancer-enhancer interactions being stronger than promoter-enhancer interactions. This could be open to speculation about the structure and robustness of the regulatory network. Also, in case of multiple enhancers regulating the same gene, often some of those enhancers were also connected to each other. This made us to speculate the existence of functional chromatin units, we called them “chromatin flowers” (resembling cloverleaf), where multiple loops of varying sizes are coming together to connect a single gene. We did not perform additional experiments. Moreover, we assessed the sequence conservation of HiCap identified enhancers. Although the vertebrate phastcons conservation scores were not significantly higher in enhancers compared to the scores of same enhancer-sized random regions in the mouse genome, when we computationally looked for distribution of highest conserved smaller regions within the enhancer vs random regions, enhancers clearly contained, on average, significantly more

conserved regions. TFs binding motifs are usually smaller 6-8 nt regions, thus one can imagine that the distal elements that HiCap identified, which are actually Dpn II fragments, do not necessarily embody functional enhancers in vivo, but they rather contain regions where TFs bind, which is conserved amongst vertebrates.

It was important to show that enhancers are not necessarily only connected to the closest genes. Although 65% of enhancers were connected to the closest gene, there were thousands of long-range interactions showing modest enrichment for genes that become upregulated upon TF perturbation (over-expression) similar to that of the closest genes. Thus, HiCap should be taken in to consideration together with techniques like ChIP-seq in order to have both closest and long-range interactions. None of the other papers performed such vigorous computational experiments assessing the quality and importance of the non-closest vs closest interactions. Although HiCap showed modest predictive power, it increased the resolution to identify regulatory regions beyond any other methods available at the time.

5.3 PAPER III

Single-cell RNA-sequencing is a powerful technique to study cellular heterogeneity, characterize cell types in unprecedented detail and identify rare cell phenotypes, however current methods have been able to profile only mRNAs (Hashimshony et al., 2012; Islam et al., 2011; Patel et al., 2014; Picelli et al., 2014; Ramsköld et al., 2012; Sandberg, 2014). This has been a limitation in the protocols, rather than a choice. Non-coding RNA studies have indicated the importance of ncRNAs and their regulatory function, thus, including them in profiling single cells would help distinguishing cellular phenotypes easily. Especially, small non-coding RNAs such as miRNAs play an important role in cells, yet they have been lacking in current single-cell methods. There was a huge need to develop single-cell RNA-seq method covering small-RNAs. In **Paper III**, with the vision and expertise of Omid R. Faridani, we developed a method we forgot to name in the published paper, nevertheless we call it Small-seq (Faridani et al., 2016).

Small-seq incorporated many novelties. First of all, by skipping the gel purification step combined with rRNA masking, we are able to profile all small RNAs that are properly processed having 5’ phosphate and 3’ hydroxyl group. Furthermore, by incorporating UMIs we are able to count the number of RNA molecules in a given cell. Removing the biases introduced by PCR enables cleaner downstream analysis such as cell clustering and differential gene expression (**Paper III, Figure 1k and 1h**).

We were interested in understanding human embryonic stem cell (hESC) regulation via small RNAs. For that we generated very comprehensive annotation database by combining transcripts from Gencode, FANTOM, Mirbase, Repbase, Gtrnadb and etc. During some analysis, we considered all the RNAs (**Paper III, Supplementary Figure 4a**), and counted number of molecules for all. We noticed that most of the molecules are coming from miRNAs, sdRNAs, tsRNAs and small RNAs derived from protein-coding genes. Thus, we focused on mainly these small RNAs. Surprisingly, miRNAs showed great potential in separating clusters of different cell types, comparable to mRNAs, and performed better than other small RNAs (**Paper III, Figure 1k, Supplementary Figure 6**). Analysis of heterogeneity in hESCs revealed that miR-375 and miR-371-3 cluster showed variation in expression across individual primed hESCs, but not naïve ESCs. The variability of miR-371-3 cluster has been observed previously in human pluripotent stem cell lines (H. Kim et al., 2011) but not within hESC population. We also performed the similar variability analysis on sdRNAs and tsRNAs (**Paper III, Supplementary Figure 8**). Majority of the small RNAs did not vary considerably further from the expected (data not shown) amongst primed and naïve hESC populations.

One of the important aspects of a new method is its sensitivity and accuracy. In this context sensitivity indicates quantitative measure of how well Small-seq captures the total expressed genes in a given cell. We performed serial dilution experiments using HEK293T cells. We detected about 450 miRNAs (expressing more than 1 molecule) from 1,000 ng total RNA down to 1 ng. After that, we observed technical losses, and at 0.01 ng we observed about 40% of mature miRNAs. We concluded that Small-seq has about 40% sensitivity – meaning approximately 40% of miRNA molecules (as well as other small RNA molecules) expressed in an individual cell is captured (**Paper III, Figure 2a)**. Furthermore, variation in miRNA expression increased for the lowly expressed genes, yet the biological variation of miRNA expression for individual HEK293T cells was above technical noise, even for the lowly expressed genes (**Paper III, Figure 2d)**. Compared to bulk data, Small-seq generated very similar fraction of differentially expressed genes (**Paper III, Supplementary Figure 5**).

Overall we developed a sensitive and novel single-cell RNA-seq method. I developed a new computational pipeline designed for the purpose of analyzing Small-seq data. Not having RNA size selection step allows the automation of the protocol. However, Small-seq does not provide expression of large RNAs such as mRNAs and long non-coding RNAs, therefore, it could be used in conjunction with other single-cell method in order to fully understand biology of single cells.

5.4 PAPER IV

We believe one could have a better understanding of salamander limb regeneration by understanding the cellular composition of regenerating tissues. Although this project started a few years back with a slightly different setup, we soon got stalled by the difficulties during blastema dissociation and picking up cells. Later, the project got picked up and revived by Ahmed Elewa. We took advantage of single-cell RNA-sequencing method Smart-seq2 and created libraries from 19 dpa (day post-amputation) newt blastema in 2 x 384 cell plates.

First, in order to work with a more comprehensive transcriptome, by using a software package Corset, we combined transcriptomes of two publicly available datasets (Abdullayev, Kirkham, Björklund, Simon, & Sandberg, 2013; Looso et al., 2013) and newly generated in-house transcriptome from regenerating newt limb. This procedure resulted in more 431,864 contigs with N50 value of 1,297 nt. We mapped reads to this new transcriptome using STAR (Dobin et al., 2013), quantified contigs using RSEM (Li & Dewey, 2011) and annotated. We tried a few different ways of annotation methods and although results (data not shown) indicated that combination of gene ontology (GO) terms from Trinotate BLAST, Trinotate Pfam assignments, MSigDB (using human ortholog mapping) and curated gene sets from MSigDB performed the best, GO terms from Trinotate BLAST would have been good enough. Then we used the PAGODA package (Fan et al., 2016) to identify statistically significant excess of coordinated variability in dataset, where GO terms are considered "overdispersed" when their explained variance (by the first PC) is significantly higher than expected (with multiple correction). We evaluated the results considering a few parameters, and finally ended up with 8 clusters. We could have performed a more systematic way of finding the best number clusters to decide. Depending on how you set the hierarchical clustering one could get different number of clusters. Then we run t-distributed stochastic neighbor embedding (tSNE) clustering method and the clusters from tSNE overlapped well with clusters from PAGODA (**Paper IV, Figure 1a**). Additionally, we identified differentially expressed genes using SCDE (Kharchenko, Silberstein, & Scadden, 2014) package. We were hoping to get some clear insights into identities of the 8 clusters using both enriched GO terms (PAGODA output) and differentially expressed genes (SCDE output), however, this has been challenging due to incomplete annotation and perhaps biology of blastema.

Cells in blastema have supposedly lost their original identities and have dedifferentiated back to stem-cell like progenitor cells. This seem to be reflected in our results: significantly enriched GO overlapped a great deal between clusters, did not show many GO terms specific to cell types, except a few (**Paper IV, Figure 1b**). Cluster 1 and 8 doesn't seem to have clear function. Cluster 2 showed a very clear enrichment for GO terms reflecting transposable element (TE) activity. This remains as the most interesting cluster we have identified in this project. Cluster 3 has mainly DNA repair related GO terms, cluster 4 has immune response and cluster 5 has

splicing related GO terms. Cluster 6 could be a connective tissue or collagen. One of the top differentially expressed genes in cluster 2 was MARCS transcript. Interestingly, MARCS-like protein has been implicated in stimulation of cell cycle in axolotl limb regeneration (Sugiura, Wang, Barsacchi, Simon, & Tanaka, 2016).

Having some candidate transcripts, we wanted to study further and validate our results by performing *in situ sequencing* (ISS) experiments. Our collaborators at Mats Nilsson's lab designed primers and performed ISS experiments, first for housekeeping genes then on regenerating newt samples. The results are very preliminary, but promising. First of all, all the primers designed to detect markers genes were successfully hybridized their targets and we could detect in situ maps for all (**Paper IV, Figure 3a**). Furthermore, we identified TE overexpressing cell markers (such as MARCS, DMBT1 and etc) at several locations in the tissue samples *in situ* (**Paper IV, Figure 3b**). Further experiments indicated that these TE-overexpressing markers are expressed throughout the limb regeneration process, however, there was an uncertainty in the distribution of expression pattern. Overall, in this project we have generated some candidate marker genes and their corresponding clusters, but more experiments are needed to identify the cell types, validate and visualize their expression pattern along the regenerating limb, since there could be difference in the cellular composition along blastema proximal-distal axis (for instance, Pax7+ satellite cells were observed along skeletal muscle fiber cells in a more proximal part to the amputation plane) (H. V. Tanaka et al., 2016).
Summary and Future Perspectives

6 SUMMARY AND FUTURE PERSPECTIVES

My projects are centered on understanding how genes are expressed and controlled encapsulating regeneration in newts, regulatory interactions in mouse ES cells and small RNAs in human ES cells. Given more time and resources, it would be interesting to combine these areas of research, for instance, studying small RNAs and regulatory regions in newts would have enormous implications for the field – which seems a bit divided into understanding the "magic powers" of newts: some believes key to regeneration relies on newt-specific genes, whereas some believe regulation and gene-wiring is the answer. Since I'm also involved in the sequencing of newt genome, it is fair to assume that after genome is publicly available, the field will move towards genome-wide epigenetic studies, small RNA studies (especially miRNA and tsRNA).

Since young children has been shown to regenerate their fingertips if the stump skin is not stitched together (Whited & Tabin, 2009), this gives us a hope for finding ways of reviving lost regeneration abilities in humans. Especially, if we could understand the early stages of wound healing in both regenerating and non-regenerating circumstances, we might have a better chance of finding how to heal a wound in a way that leads to formation of a blastema rather than a scar tissue.

Single-cell RNA-sequencing has become widely-used and affordable. With the establishment of Human Cell Atlas consortium, the field is headed towards identification and detailed molecular characterization of all cell types in a human body. This will open up many opportunities for studying rare cell populations involved in cancer metastasis, tumor resistance, better characterization of tissue function and etc.

On the other hand, despite the significance of enhancers in gene regulation, the field has not progressed as fast as other fields, such as scRNA-seq. Higher resolution TF binding profiles and expression data for many cell types provided by ENCODE consortium (ENCODE Project Consortium, 2012) and others opened up emergence of machine learning tools in modern biological research areas. This will likely to progress and expand since we are generating more data, new methods, new types of data. So, there will be a need to make use of combining different data types to predictive on work on models.

7 ACKNOWLEDGEMENTS

Thanks my dear supervisor Professor **Rickard Sandberg**. You have been an amazing supervisor over the years, guiding through projects, helping with the programming and computational issues. Thanks for sending me to world-class courses and conferences. I always get motivated by talking to you, get a better perspective where you often remind me how to think critical and how to prioritize things. Your friendly approach makes it easy and natural to share personal stories with you. Special thanks for all the support during challenging times. You have been a great example for how to be successful scientist, while keeping personal life and work in balance. I am still amazed by your vision for science, leadership skills, academic network, and consistent drive and determination. I wouldn't ask for a better boss.

Thank you my co-supervisor **Pelin Sahlén**. You were the first person who truly guided me through the world of experimental laboratory work. I learned many valuable lessons from all those painful HiCap optimizations. Thanks for cheering me up every time an experiment works, creating an emotional connection with the projects. My mentor, **Örjan Wrange**, thank you for your valuable advice and guidance. Despite few meetings, your extensive experience taught me a lot about how academia works. Thank you.

Thanks to my collaborator **András Simon** for introducing me to wonders of salamander regeneration and being such an inspirational researcher. Thanks to the people in Simon lab: **Heng Wang, Matthew Kirkham** and **Ahmed Elewa** for such a great collaborative work and providing newt samples. **Ahmed**, besides work, I feel very fortunate to have met you, shared stories and drinks with you. Thanks for all the motivational inputs during the thesis writing.

One of the most important thanks goes to **Ingemar Ernberg**, my masters co-supervisor who recommended me to Rickard. Thanks for the support over the years, advising us to look at the big picture and not afraid of asking important questions, such as "what is life". Thanks to my masters supervisor **Erik Aurell**, I learned the essentials of systems biology and bioinformatics. Working on the microarray project with you during summer opened up many opportunities for me. Also thanks to **Aymeric Fouquier d'Herouel** and **Ziming Du** for the help during that time and **Qin Li** for her support and friendship since then.

Big thanks and gratitude to all the current and previous members of Sandberg lab, for their support and friendship over the years. Thank you **Omid Faridani**! I felt very comfortable after you started working in our lab, and we have become very close since then. Thank you for being such a wonderful friend first, sharing happy moments and supporting through hard times. I will very much miss our coffee/tea breaks. It is also pleasure to work with you and to discuss stimulating topics such as fancy methods, entrepreneurship and solving cancer. **Mtakai Ngara**, thanks for your friendship. You are such a gentle and caring person who is laid back but also have many deep layers. Thank you for pushing me to go to gym, well, pushing each other perhaps. **Helena Storvall**, thanks for your friendship and bringing your pleasant attitude to our lab for many years. You were exceptional in being a good phd student while not missing any extracurricular social activities. Well done! I'm sure you are having the same success in your

work place. I look forward to continue our friendship and resume our nerdy discussions about life, astronomy and physics. Thank you **Daniel Edsgärd**, for being such a good friend and a remarkable bioinformatician. I appreciate all the efforts for teaching us some statistics, programming and of course dancing. I truly enjoy our fun socializing activities as well as sharing personal stories with you. I'm extremely grateful to **Daniel Ramsköld** for hands-on help me to learn depths of Python programming. You are such a wonderful friend, and smart person whom I can always count on getting help almost immediately. Thank a lot **Åsa Segerstolpe**, for being very kind and supportive friend, as well as an expert experimental scientist. I wish you a very happy family life and all the best in your new job. **Björn Reinius**, you are such an interesting character who not only brings joy and delight to our lab but also quality research. Thank you! You are very close to have your own group, and I believe you will be a very prosperous researcher. **Qiaolin Deng**, you have been very helpful over the years about research, how to raise kids etc. Thank you for all. You have a big responsibility now with having your own group and raising two kids. I sincerely wish you the best in both. Thank you **Per Johnsson**, for being a very caring, calm and very friendly colleague. Congratulations on having second child now. I appreciate all the tips about writing thesis. Tesekkurler **Ersen Kavak**, for giving valuable tips early on during my phd. You paved the way of becoming successful bioentrepreneur and I'm sure we will hear more success stories in the future. Thanks **Gösta Winberg** for guiding the lab with your deep expert knowledge, sometimes fun and twisted sense of humor and of course reminding us the importance of free food. Oh beloved **Sven Sargasser**, you brought joy to our lab, helped me early on in my experiments and gave valuable parenting tips. Thanks and I promise we will go to fishing one day. Thank you **Åsa Björklund** for all the expert bioinformatics help and being a good example of how to be successful programmer in our field. **Michael Hagemann-Jensen**, besides having remarkable social skills such as drinking and dancing, you are also dedicated to your work, and thanks for awesome collaboration on Small-seq and many more methods to come. Thank you dear **Athanasia Palasantza**, aka **Sissy**, for your friendship and support. Although it was brief, you had a very successful academic track record and I wish you all the best in your new career. I'm sure you will bring the best. **Husain Talukdar**, thank you for your friendship since our masters education and tips on writing thesis. **Gert-Jan Hendriks,** you are a great addition to our lab and I appreciate our pleasant discussions over lunch. **Sophie Petropoulos**, you are hard-working researcher and also fun person to talk to. **Marlene Yilmaz**, thanks for showing dance moves as well as sharing stories during good and difficult times. You were very dedicated to your work, so I believe you will continue doing that wherever you go. Thank you **Anton** – the first person who understood the cover of this thesis, **Leo** – enthusiastic and friendly new member of our lab, **Ping** – pleasant and hard-working colleague, and beloved members of Deng lab **Geng, Shangli, Yu** for keeping the lab cheerful and lively.

Also thanks to everyone at **CMB**, those we have shared any time together, be it like a conference, course, fika or as simple as 'hej'. It has been pleasure sharing this wonderful scientific environment with you. **Matti Nikkola**, many thanks for your enormous help over the years with all the confusing administrative issues. **Tiago Pinheiro**, thanks for organizing such

great parties and inviting me. Thank you **Lina Pettersson** for being enormously helpful on administrative matters. Thanks **Elvira, Nikola, Indira, Gonçalo, Jens, Giuseppe** and my precious friend **Alena "Alca" Salasova** for nice memories and sharing stories.

I'm very thankful to the entire **Ludwig Institute** for providing excellent research environment. Thank you **Charlotta Linderholm** for caring about our research and personal environment. How can I forget acknowledging **Eliza Joodmardi** and **Soheilla Rezaian** – the two people who are always there whenever we need any help, from very very early in the morning. **Eliza** I really enjoyed your motivational comments about Iranian and Azeri food with respect to whatever Omid and I were cooking. **Jorge Villarroel**, thanks for sharing your awesome life stories. Thank you **Thomas Perlmann** for maintaining such an excellent research institute that provided us cutting-edge research environment. Thanks **Johan Holmberg**, for always bringing the entertaining mood to Ludwig and sharing nice discussions over a cold beer. **Mats Anderling**, thank you for helping me and solving IT issues very quickly during all these years. My dear friend **Katarina Tiklova**, thank you for being there for me during difficult times, sharing our life stories, introducing me to your friends and playing innebandi. You are thoughtful friend, caring mother and very good scientist. Thank you **Danny Topcic, Maria Bergsland, Daniel Hagey, Danny Topcic, Cécile Zaouter, Susanne Klum, Nigel, Stuart, Andre Nobre, Linda Gillberg, Hilda Lundén Miguel, Bhumica Singla** and everyone else for sharing nice memories, chat over lunches and coffee breaks. I look forward to hearing your success stories.

Thanks to my friends from ITU: **Tuba Bucak, Bilge San, Kemal Sanli** and all the other friends and classmates for their friendship, helping out learning the depth of molecular biology and my bachelor degree supervisor Professor **Zeynep Petek Çakar** for being such a good role model scientist.

Dear doctor (real) **Parviz Mammadzada**, your friendship is extremely valuable to me. Thanks for sharing and enabling all those awesome memories, absolute support, personal development tips and being a progressive intellectual. I can always count on you. Thank you ustad **Rasim Ismayilov** for being an influential friend, a smart philosopher who introduced me so many beautiful things, I mean really so many. I can't describe how much I enjoy our high-level interaction. **Rustam Asgarov**, thank you for your deep and caring friendship. I'm happy for you moving in to Australia and having a time of your life. Well done bro! Thanks **Kamal Ismayilov**, for your friendship, smart interaction, multi-language jokes and unconventional memories. Patron **Ömer Saatcioglu**, thanks for always bringing awesome vibe, funny stories and being a caring friend. You have brought so many happy memories into my life. **Oguz Mehli**, although we have met not so long ago, you have become quite close friend and I'm happy to hang out with you. I promise I will make more time for gün oyunu. **Adem Björn Ergül**, thanks for being supportive and fun friend with a truly unique sense of humor. Sen tezi yazarken beni unuttun, cünkü onu yazan bir björn, bunu yazan tosun. Thank you **Oguzhan Erim**, for being such a cool and loyal friend whom I can always count on. I believe he will bring the best sustainable energy systems to whole Europe. My old and precious friend **Ömer**

Faruk Halicioglu, I have many deep nostalgic memories of playing Worms, watching Supernatural, going for a spontaneous lunch to eat "durum" in Istanbul and playing backgammon afterwards. Those were good times. Thank you **Vugar Aliyev, Hasan Alp, Ali Shirzad, Faradj Koliev, Bill Söderström, Tural Abdullayev** and entrepreneur friends **Leandro Agudelo** and **Errol Corcmaz,** and so many other friends whom we have shared great moments over the years. You guys are awesome! Thanks **Owais Mahmudi**, you are truly caring and kind friend. I wish you all the best in your new adventures with your family. Thank you my bioinformatician friends **Tojo James, Alejandro Fernandez, Hassan Foroughi** – you guys are really smart and wonderful! Warm thanks to my Polish friend **Magdalena Łopatowska** for being very thoughtful and kind with her complicated (de)motivational comics. Oh no! Furthermore, thanks **Aleksandra Laskawiec, Paulina Okulska** and Taiwanese friend **Tzupe Wang** for your lovely companionship and introducing me to your welcoming culture.

Special thanks to my neighbor **Henry Fynde** for supporting me during difficult times and exploring the nice lunch & fika places around Stockholm. Thank you **Dai Lu** for making the illustration for this thesis, you are a very talented artist. Warm thanks to **Sevinj Ahmadova** and **Mehraj Abbasov** for introducing me to great research in my home country Azerbaijan, specifically Genetic Resources Institute of ANAS. I look forward to our future collaborations. I am also very grateful to **Ministry of Education of Azerbaijan** for financially supporting my bachelor studies.

Thanks to my dearest friends/brothers from high school. Thank you **Elvin Cabayev**, you have always been there for me, supported me through difficult times, and shared my happiest moments. I won't forget all the philosophical discussions we had at the "Nargiz restaurant", not to mention the delicious food and drinks. Thanks **Sadiq Mammadov** for somehow trusting me for all these years. I truly cherish our unorthodox friendship. **Said Misirli,** you are such a sensitive and caring friend, responsible father and eternal optimist. **Davud Davudov,** I'm grateful for your deep friendship that significantly changed my life. Thank you!! Many many thanks to **Shahin Aliyev, Eldar Khalilov, Vuqar Babayev, Zaur Azimov, Renat Azmammadov, Tural Abdurahimov, Ahmet Eyyublu, Bahtiyar Mammadov, Nizami Isayev, Dashqin Haciyev, Alim Nazarli, Resid Resulzade, Roman Ismayilov, Azad Selimov, Rauf Kerimov, Aqil Azakov, Aqil Mecidov, Rufet Zakirov** for your eternal friendship and all the sacrifices you made during our high school years. Specifically, teachers and personal from **Agdas Özel Türk Litseyi: Xalid Qedirov, Ramazan Shaban, Musatafa Vural, Fatih Koch, Adem Kilic, Kenan Güzel, Dünyamin Memmedov.** Truth to be told, I can't thank you guys enough. Qardaslarim sagolun, varolun! Also, thank you my teachers at **Ağdaş rayonu Pirkəkə kənd**, particularly **Fatma müellime**, who played incredibly crucial role in the early years of my primary education.

Thank you very much for those who found my thesis by searching online and learned something from it. For those who browsed to this section without checking the rest of the thesis, I got you ;) Thank you too. Apologies for all the friends and acquaintances whose name I forgot to mention here, or whom I have not met yet. Thank you!

I am deeply thankful and grateful to **Aysegul Coskun**, who was there for me for quite long, shared my happiest and saddest moments, supported me until she couldn't anymore. I wish you will be happier. My son, **Cansun**, I love you more than anything in the world. All my troubles used to disappear when you ran and hug me while screaming with excitement. I will always support you, be there for you and wish you will have a happy and fulfilling life.

Lastly, my family! Canlarım! Mamam **Rusxarə**, papam **Mahammad**, bacım **Gülnar**, qardaşım **Vüsal** sizə dərin minnətdarligimi bildirirəm. Məni həmişə sevən dəstəkləyen sizlər olmusunuz. Bütün ugurlarimi sizin sevginizə borcluyam. Zəhmətkeş mamam, tərəvəz əkib, toyuga-cücəyə baxib, qaynata qaynanaya qulluq edərkən eyni zamanda bizə de göz bəbəyi kimi baxdin. Urəyi tər-təmiz papam, sən səhər-axsam durmadan işlədin, günortaya kimi məktəbə gedib dərs deyirdin, sonra gəlib müxtəlif kənd təsərrüfatıyla məşgul olurdun. Bütün **qohum-eqrabalar**, **xalalarim**, **əmilərim**, **bibilərim**, **balacalar, qonsular** və **dost-tanislar** sizlərə də sonsuz təşəkkürümü bildirirəm. Sizlərin əməyinizi hec bir zaman unudmayacagam. Sizləri cox sevir və dəyər vərirəm. Xətiriniz daima əziz olacaq!

8 REFERENCES

Abdullayev, I., Kirkham, M., Björklund, Å. K., Simon, A., & Sandberg, R. (2013). A reference transcriptome and inferred proteome for the salamander Notophthalmus viridescens. *Experimental Cell Research*, *319*(8), 1187–1197. http://doi.org/10.1016/j.yexcr.2013.02.013

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2014a). Molecular Biology of the Cell. Garland Publishing.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2014b). Molecular Biology of the Cell, Sixth Edition. Garland Science.

Alwine, J. C., Kemp, D. J., & Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences*, *74*(12), 5350–5354.

Avilion, A. A., Nicolis, S. K., Pevny, L. H., Perez, L., Vivian, N., & Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes & Development*, *17*(1), 126–140. http://doi.org/10.1101/gad.224503

Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P., & Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes & Development*, *22*(20), 2773–2785. http://doi.org/10.1101/gad.1705308

Banerji, J., Olson, L., & Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, *33*(3), 729–740.

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2 Pt 1), 299–308.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–837. http://doi.org/10.1016/j.cell.2007.05.009

Bartman, C. R., Hsu, S. C., Hsiung, C. C.-S., Raj, A., & Blobel, G. A. (2016). Enhancer

Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Molecular Cell*, *62*(2), 237–247. http://doi.org/10.1016/j.molcel.2016.03.007
Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., et al. (2017). A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Reports*, *18*(3), 762–776. http://doi.org/10.1016/j.celrep.2016.12.063
Butler, J. E. F., & Kadonaga, J. T. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development*, *16*(20), 2583–2592. http://doi.org/10.1101/gad.1026202
Cheloufi, S., Santos, Dos, C. O., Chong, M. M. W., & Hannon, G. J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*, *465*(7298), 584–589. http://doi.org/10.1038/nature09092
Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (New York, N.Y.)*, *348*(6233), aaa6090–aaa6090. http://doi.org/10.1126/science.aaa6090
Chen, Q., Yan, M., Cao, Z., Li, X., Zhang, Y., Shi, J., et al. (2016). Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. *Science (New York, N.Y.)*, *351*(6271), 397–400. http://doi.org/10.1126/science.aad7977
Chepelev, I., Wei, G., Wangsa, D., Tang, Q., & Zhao, K. (2012). Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Research*, *22*(3), 490–503. http://doi.org/10.1038/cr.2012.15
Claverie, J.-M. (2005). Fewer genes, more noncoding RNA. *Science (New York, N.Y.)*, *309*(5740), 1529–1530. http://doi.org/10.1126/science.1116800
Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, *46*(12), 1311–1320. http://doi.org/10.1038/ng.3142
Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21931–21936. http://doi.org/10.1073/pnas.1016071107
Cullen, B. R. (2004). Transcription and processing of human microRNA precursors. *Molecular Cell*, *16*(6), 861–865. http://doi.org/10.1016/j.molcel.2004.12.002
Deng, Qiaolin, Ramsköld, D., Reinius, B., & Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science (New York, N.Y.)*, *343*(6167), 193–196. http://doi.org/10.1126/science.1245316
Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. http://doi.org/10.1093/bioinformatics/bts635
ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. http://doi.org/10.1038/nature11247
ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816. http://doi.org/10.1038/nature05874
Evans, M. J., & Kaufman, M. H. (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature*, *292*(5819), 154–156. http://doi.org/10.1038/292154a0
Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods*, *13*(3), 241–244.

http://doi.org/10.1038/nmeth.3734
Faridani, O. R., Abdullayev, I., Hagemann-Jensen, M., Schell, J. P., Lanner, F., & Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. *Nature Biotechnology*, *34*(12), 1264–1266. http://doi.org/10.1038/nbt.3701
Femino, A. M., Fay, F. S., Fogarty, K., & Singer, R. H. (1998). Visualization of single RNA transcripts in situ. *Science (New York, N.Y.)*, *280*(5363), 585–590.
Freeman, L. A., & Garrard, W. T. (1992). DNA supercoiling in chromatin structure and gene expression. *Critical Reviews in Eukaryotic Gene Expression*, *2*(2), 165–209.
Freeman, W. M., Walker, S. J., & Vrana, K. E. (1999). Quantitative RT-PCR: pitfalls and potential. *BioTechniques*, *26*(1), 112–22– 124–5.
Friedman, R. C., Farh, K. K. H., Burge, C. B., & Bartel, D. P. (2008). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, *19*(1), 92–105. http://doi.org/10.1101/gr.082701.108
Fukaya, T., Lim, B., & Levine, M. (2016). Enhancer Control of Transcriptional Bursting. *Cell*, *166*(2), 358–368. http://doi.org/10.1016/j.cell.2016.05.025
Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Bin Mohamed, Y., et al. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, *461*(7269), 58–64. http://doi.org/10.1038/nature08497
Fullwood, M., Huang, P. Y. H., Han, Y., Handoko, L., Velkov, S., Wong, E., et al. (2010). Protocol: Sonication-based Circular Chromosome Conformation Capture with next-generation sequencing analysis for the detection of chromatin interactions. *Protocol Exchange*. http://doi.org/10.1038/protex.2010.207
García-González, E., Escamilla-Del-Arenal, M., Arzate-Mejía, R., & Recillas-Targa, F. (2016). Chromatin remodeling effects on enhancer activity. *Cellular and Molecular Life Sciences : CMLS*, *73*(15), 2897–2910. http://doi.org/10.1007/s00018-016-2184-3
Geng, J., Gates, P. B., Kumar, A., Guenther, S., Garza-Garcia, A., Kuenne, C., et al. (2015). Identification of the orphan gene Prod 1 in basal and other salamander families. *EvoDevo*, *6*(1), 9. http://doi.org/10.1186/s13227-015-0006-6
Gershenzon, N. I., & Ioshikhes, I. P. (2005). Promoter Classifier. *Applied Bioinformatics*, *4*(3), 205–209. http://doi.org/10.2165/00822942-200504030-00005
Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. http://doi.org/10.1038/nbt.1883
Graves, P., & Zeng, Y. (2012). Biogenesis of mammalian microRNAs: a global view. *Genomics, Proteomics & Bioinformatics*, *10*(5), 239–245. http://doi.org/10.1016/j.gpb.2012.06.004
Grishok, A., Pasquinelli, A. E., Conte, D., Li, N., Parrish, S., Ha, I., et al. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, *106*(1), 23–34.
Groff, A. F., Sanchez-Gomez, D. B., Soruco, M. M. L., Gerhardinger, C., Barutcu, A. R., Li, E., et al. (2016). In Vivo Characterization of Linc-p21 Reveals Functional cis-Regulatory DNA Elements. *Cell Reports*, *16*(8), 2178–2186. http://doi.org/10.1016/j.celrep.2016.07.050
Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, *525*(7568), 251–255. http://doi.org/10.1038/nature14966
Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Structural & Molecular Biology*, *11*(5), 394–403. http://doi.org/10.1038/nsmb763
Hammond, S. M., Boettcher, S., Caudy, A. A., Kobayashi, R., & Hannon, G. J. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science (New York, N.Y.)*, *293*(5532), 1146–1150. http://doi.org/10.1126/science.1064023

Han, J., Lee, Y., Yeom, K.-H., Kim, Y.-K., Jin, H., & Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes & Development*, *18*(24), 3016–3027. http://doi.org/10.1101/gad.1262504

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, *2*(3), 666–673. http://doi.org/10.1016/j.celrep.2012.08.003

Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A. Z., & Kay, M. A. (2010). Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA (New York, N.Y.)*, *16*(4), 673–695. http://doi.org/10.1261/rna.2000810

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, *39*(3), 311–318. http://doi.org/10.1038/ng1966

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., et al. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, *155*(4), 934–947. http://doi.org/10.1016/j.cell.2013.09.053

Hsieh, T.-H. S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., & Rando, O. J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, *162*(1), 108–119. http://doi.org/10.1016/j.cell.2015.05.048

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., et al. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics*, *46*(2), 205–212. http://doi.org/10.1038/ng.2871

Huminiecki, Ł., & Horbańczuk, J. (2017). Can We Predict Gene Expression by Understanding Proximal Promoter Architecture? *Trends in Biotechnology*, *35*(6), 530–546. http://doi.org/10.1016/j.tibtech.2017.03.007

Hurst, L. D., Sachenkova, O., Daub, C., Forrest, A. R. R., Huminiecki, L., FANTOM consortium. (2014). A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators. *Genome Biology*, *15*(7), 413. http://doi.org/10.1186/s13059-014-0413-3

Hutvágner, G., & Zamore, P. D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. *Science (New York, N.Y.)*, *297*(5589), 2056–2060. http://doi.org/10.1126/science.1073827

Illingworth, C. M. (1974). Trapped fingers and amputated finger tips in children. *Journal of Pediatric Surgery*, *9*(6), 853–858.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, *21*(7), 1160–1167. http://doi.org/10.1101/gr.110882.110

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., & Linnarsson, S. (2012). Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nature Protocols*, *7*(5), 813–828. http://doi.org/10.1038/nprot.2012.022

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., et al. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, *11*(2), 163–166. http://doi.org/10.1038/nmeth.2772

Johnsson, P., Lipovich, L., Grandér, D., & Morris, K. V. (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochimica Et Biophysica Acta*, *1840*(3), 1063–1071. http://doi.org/10.1016/j.bbagen.2013.10.035

Ke, R., Mignardi, M., Hauling, T., & Nilsson, M. (2016). Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. *Human Mutation*, *37*(12), 1363–1367. http://doi.org/10.1002/humu.23051

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., & Nilsson, M.

(2013). In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, *10*(9), 857–860. http://doi.org/10.1038/nmeth.2563
Ketting, R. F., Fischer, S. E., Bernstein, E., Sijen, T., Hannon, G. J., & Plasterk, R. H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes & Development*, *15*(20), 2654–2659. http://doi.org/10.1101/gad.927801
Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, *11*(7), 740–742. http://doi.org/10.1038/nmeth.2967
Khvorova, A., Reynolds, A., & Jayasena, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell*, *115*(2), 209–216.
Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. http://doi.org/10.1038/nmeth.3317
Kim, H., Lee, G., Ganat, Y., Papapetrou, E. P., Lipchina, I., Socci, N. D., et al. (2011). miR-371-3 expression predicts neural differentiation propensity in human pluripotent stem cells. *Cell Stem Cell*, *8*(6), 695–706. http://doi.org/10.1016/j.stem.2011.04.002
Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187. http://doi.org/10.1038/nature09033
Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74. http://doi.org/10.1038/nmeth.1778
Ko, M. S., Nakauchi, H., & Takahashi, N. (1990). The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. *The EMBO Journal*, *9*(9), 2835–2842.
Kopp, J. L., Ormsbee, B. D., Desler, M., & Rizzino, A. (2008). Small Increases in the Level of Sox2 Trigger the Differentiation of Mouse Embryonic Stem Cells. *Stem Cells*, *26*(4), 903–911. http://doi.org/10.1634/stemcells.2007-0951
Krebs, A. R., Karmodiya, K., Lindahl-Allen, M., Struhl, K., & Tora, L. (2011). SAGA and ATAC histone acetyl transferase complexes regulate distinct sets of genes and ATAC defines a class of p300-independent enhancers. *Molecular Cell*, *44*(3), 410–423. http://doi.org/10.1016/j.molcel.2011.08.037
Kumar, A., Gates, P. B., Czarkwiani, A., & Brockes, J. P. (2015). An orphan gene is necessary for preaxial digit formation during salamander limb development. *Nature Communications*, *6*, 8684. http://doi.org/10.1038/ncomms9684
Lamond, A. I., & Earnshaw, W. C. (1998). Structure and function in the nucleus. *Science (New York, N.Y.)*, *280*(5363), 547–553.
Lee, D.-H., Gershenzon, N., Gupta, M., Ioshikhes, I. P., Reinberg, D., & Lewis, B. A. (2005). Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Molecular and Cellular Biology*, *25*(21), 9674–9686. http://doi.org/10.1128/MCB.25.21.9674-9686.2005
Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, *75*(5), 843–854.
Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., et al. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature*, *425*(6956), 415–419. http://doi.org/10.1038/nature01957
Lenhard, B., Sandelin, A., & Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews. Genetics*, *13*(4), 233–245. http://doi.org/10.1038/nrg3163
Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., et al. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin

and is associated with preaxial polydactyly. *Human Molecular Genetics*, *12*(14), 1725–1735.
Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., et al. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*, *7*(9), 709–715. http://doi.org/10.1038/nmeth.1491
Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. http://doi.org/10.1186/1471-2105-12-323
Lim, C. Y., Santoso, B., Boulay, T., Dong, E., Ohler, U., & Kadonaga, J. T. (2004). The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes & Development*, *18*(13), 1606–1617. http://doi.org/10.1101/gad.1193404
Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), 523–536. http://doi.org/10.1016/j.cell.2008.03.029
Looso, M., Preussner, J., Sousounis, K., Bruckskotten, M., Michel, C. S., Lignelli, E., et al. (2013). A de novo assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. *Genome Biology*, *14*(2), R16. http://doi.org/10.1186/gb-2013-14-2-r16
Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., et al. (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nature Methods*, *12*(1), 71–78. http://doi.org/10.1038/nmeth.3205
Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. http://doi.org/10.1016/j.cell.2015.05.002
Mayer, M., Schiffer, S., & Marchfelder, A. (2000). tRNA 3' processing in plants: nuclear and mitochondrial activities differ. *Biochemistry*, *39*(8), 2096–2105. http://doi.org/10.1021/bi992253e
Merika, M., Williams, A. J., Chen, G., Collins, T., & Thanos, D. (1998). Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription. *Molecular Cell*, *1*(2), 277–287.
Mohrs, M., Blankespoor, C. M., Wang, Z. E., Loots, G. G., Afzal, V., Hadeiba, H., et al. (2001). Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nature Immunology*, *2*(9), 842–847. http://doi.org/10.1038/ni0901-842
Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. http://doi.org/10.1038/nmeth.1226
Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, *320*(5881), 1344–1349. http://doi.org/10.1126/science.1158441
Okamura, K., Chung, W.-J., & Lai, E. C. (2008). The long and short of inverted repeat genes in animals: microRNAs, mirtrons and hairpin RNAs. *Cell Cycle (Georgetown, Tex.)*, *7*(18), 2840–2845. http://doi.org/10.4161/cc.7.18.6734
Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., et al. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Research*, *37*(18), e123–e123. http://doi.org/10.1093/nar/gkp596
Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., et al. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, *408*(6808), 86–89. http://doi.org/10.1038/35040556
Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)*, *344*(6190), 1396–1401.

http://doi.org/10.1126/science.1254257
Pertea, M., & Salzberg, S. L. (2010). Between a chicken and a grape: estimating the number of human genes. *Genome Biology*, *11*(5), 206. http://doi.org/10.1186/gb-2010-11-5-206
Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181. http://doi.org/10.1038/nprot.2014.006
Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, *470*(7333), 279–283. http://doi.org/10.1038/nature09692
Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, *30*(8), 777–782. http://doi.org/10.1038/nbt.2282
Ramsköld, D., Wang, E. T., Burge, C. B., & Sandberg, R. (2009). An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLoS Computational Biology*, *5*(12), e1000598. http://doi.org/10.1371/journal.pcbi.1000598
Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2015). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *162*(3), 687–688. http://doi.org/10.1016/j.cell.2015.07.024
Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., et al. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, *403*(6772), 901–906. http://doi.org/10.1038/35002607
Rippe, K., Hippel, von, P. H., & Langowski, J. (1995). Action at a distance: DNA-looping and initiation of transcription. *Trends in Biochemical Sciences*, *20*(12), 500–506.
Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. http://doi.org/10.1038/nature14248
Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., & Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. *Genome Research*, *14*(10A), 1902–1910. http://doi.org/10.1101/gr.2722704
Ruby, J. G., Jan, C. H., & Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature*, *448*(7149), 83–86. http://doi.org/10.1038/nature05983
Sahlén, P., Abdullayev, I., Ramsköld, D., Matskova, L., Rilakovic, N., Lötstedt, B., et al. (2015). Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biology*, *16*(1), 156. http://doi.org/10.1186/s13059-015-0727-9
Saiz, L., Rubi, J. M., & Vilar, J. M. G. (2005). Inferring the in vivo looping properties of DNA. *Proceedings of the National Academy of Sciences*, *102*(49), 17642–17645. http://doi.org/10.1073/pnas.0505693102
Sandberg, R. (2014). Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, *11*(1), 22–24.
Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, *270*(5235), 467–470.
Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, *25*(4), 582–597. http://doi.org/10.1101/gr.185272.114
Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., et al. (2012). Three-Dimensional Folding and Functional Organization Principles of the

Drosophila Genome. *Cell*, 1–15. http://doi.org/10.1016/j.cell.2012.01.010
Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, *488*(7409), 116–120. http://doi.org/10.1038/nature11243
Sheng, K., Cao, W., Niu, Y., Deng, Q., & Zong, C. (2017). Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nature Methods*, *14*(3), 267–270. http://doi.org/10.1038/nmeth.4145
Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, *31*(19), 3210–3212. http://doi.org/10.1093/bioinformatics/btv351
Smale, S. T., & Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annual Review of Biochemistry*, *72*(1), 449–479. http://doi.org/10.1146/annurev.biochem.72.121801.161520
Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (New York, N.Y.)*, *353*(6294), 78–82. http://doi.org/10.1126/science.aaf2403
Storvall, H., Ramsköld, D., & Sandberg, R. (2013). Efficient and Comprehensive Representation of Uniqueness for Next-Generation Sequencing by Minimum Unique Length Analyses. *PloS One*, *8*(1), e53822. http://doi.org/10.1371/journal.pone.0053822
Stults, D. M., Killen, M. W., Pierce, H. H., & Pierce, A. J. (2007). Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Research*, *18*(1), 13–18. http://doi.org/10.1101/gr.6858507
Sugiura, T., Wang, H., Barsacchi, R., Simon, A., & Tanaka, E. M. (2016). MARCKS-like protein is an initiating molecule in axolotl appendage regeneration. *Nature*, *531*(7593), 237–240. http://doi.org/10.1038/nature16974
Suzuki, M., Yakushiji, N., Nakada, Y., Satoh, A., Ide, H., & Tamura, K. (2006). Limb regeneration in Xenopus laevis froglet. *TheScientificWorldJournal*, *6 Suppl 1*, 26–37. http://doi.org/10.1100/tsw.2006.325
Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., et al. (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Research*, *11*(5), 677–684. http://doi.org/10.1101/gr.164001
Taft, R. J., Glazov, E. A., Lassmann, T., Hayashizaki, Y., Carninci, P., & Mattick, J. S. (2009). Small RNAs derived from snoRNAs. *RNA (New York, N.Y.)*, *15*(7), 1233–1240. http://doi.org/10.1261/rna.1528909
Taft, R. J., Pheasant, M., & Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays*, *29*(3), 288–299. http://doi.org/10.1002/bies.20544
Tanaka, H. V., Ng, N. C. Y., Yang Yu, Z., Casco-Robles, M. M., Maruo, F., Tsonis, P. A., & Chiba, C. (2016). A developmentally regulated switch from stem cells to dedifferentiation for limb muscle regeneration in newts. *Nature Communications*, *7*, 11069. http://doi.org/10.1038/ncomms11069
Thomson, J. A., Itskovitz-Eldor, J., Shapiro, S. S., Waknitz, M. A., Swiergiel, J. J., Marshall, V. S., & Jones, J. M. (1998). Embryonic stem cell lines derived from human blastocysts. *Science (New York, N.Y.)*, *282*(5391), 1145–1147.
Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, *25*(9), 1105–1111. http://doi.org/10.1093/bioinformatics/btp120
Vadaie, N., & Morris, K. V. (2013). Long antisense non-coding RNAs and the epigenetic regulation of gene expression. *Biomolecular Concepts*, *4*(4), 411–415. http://doi.org/10.1515/bmc-2013-0014

van de Werken, H. J. G., Landan, G., Holwerda, S. J. B., Hoichman, M., Klous, P., Chachik, R., et al. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods*, *9*(10), 969–972. http://doi.org/10.1038/nmeth.2173

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., & Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature Reviews. Genetics*, *10*(4), 252–263. http://doi.org/10.1038/nrg2538

Venter, J. C. (2001). The Sequence of the Human Genome. *Science (New York, N.Y.)*, *291*(5507), 1304–1351. http://doi.org/10.1126/science.1058040

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., et al. (2009a). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, *457*(7231), 854–858. http://doi.org/10.1038/nature07730

Visel, A., Rubin, E. M., & Pennacchio, L. A. (2009b). Genomic views of distant-acting enhancers. *Nature*, *461*(7261), 199–205. http://doi.org/10.1038/nature08451

Walters, M. C., Fiering, S., Eidemiller, J., Magis, W., Groudine, M., & Martin, D. I. (1995). Enhancers increase the probability but not the level of gene expression. *Proceedings of the National Academy of Sciences*, *92*(15), 7125–7129.

Wang, H., & Simon, A. (2016). Skeletal muscle dedifferentiation during salamander limb regeneration. *Current Opinion in Genetics & Development*, *40*, 108–112. http://doi.org/10.1016/j.gde.2016.06.013

Whited, J. L., & Tabin, C. J. (2009). Limb regeneration revisited. *Journal of Biology*, *8*(1), 5. http://doi.org/10.1186/jbiol105

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., et al. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2), 307–319. http://doi.org/10.1016/j.cell.2013.03.035

Wong, J. M., & Bateman, E. (1994). TBP-DNA interactions in the minor groove discriminate between A:T and T:A base pairs. *Nucleic Acids Research*, *22*(10), 1890–1896.

Wu, C.-H., Huang, T.-Y., Chen, B.-S., Chiou, L.-L., & Lee, H.-S. (2015). Long-duration muscle dedifferentiation during limb regeneration in axolotls. *PloS One*, *10*(2), e0116068. http://doi.org/10.1371/journal.pone.0116068

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, *26*(7), 873–881. http://doi.org/10.1093/bioinformatics/btq057

Xi, H., Yu, Y., Fu, Y., Foley, J., Halees, A., & Weng, Z. (2007). Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Research*, *17*(6), 798–806. http://doi.org/10.1101/gr.5754707

Yang, C., Bolotin, E., Jiang, T., Sladek, F. M., & Martinez, E. (2007). Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, *389*(1), 52–65. http://doi.org/10.1016/j.gene.2006.09.029

Yao, T. P., Oh, S. P., Fuchs, M., Zhou, N. D., Ch'ng, L. E., Newsome, D., et al. (1998). Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. *Cell*, *93*(3), 361–372.

Yi, R., Qin, Y., Macara, I. G., & Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & Development*, *17*(24), 3011–3016. http://doi.org/10.1101/gad.1158803

Young, H. E., Bailey, C. F., & Dalley, B. K. (1983). Gross morphological analysis of limb regeneration in postmetamorphic adult Ambystoma. *The Anatomical Record*, *206*(3), 295–306. http://doi.org/10.1002/ar.1092060308

Zhang, H., Kolb, F. A., Brondani, V., Billy, E., & Filipowicz, W. (2002). Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *The EMBO Journal*, *21*(21), 5875–5885. http://doi.org/10.1093/emboj/cdf582

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, *38*(11), 1341–1347. http://doi.org/10.1038/ng1891

Zwaka, T. P., & Thomson, J. A. (2003). *Homologous recombination in human embryonic stem cells. Nature biotechnology* (Vol. 21, pp. 319–321). Nature Publishing Group.