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

REGULATION OF THE VERTEBRATE TRANSCRIPTOME IN DEVELOPMENT AND DISEASE

Helena Storvall

Stockholm 2016

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Eprint AB 2015

ISBN 978-91-7676-188-5

Regulation of the Vertebrate Transcriptome in Development and Disease

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Helena Storvall

AKADEMISK AVHANDLING

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

Fredagen den 5:e Februari 2016, kl 13.00

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

Co-supervisor(s):
Elisabet Andersson
Karolinska Institutet
Department of Cell and Molecular Biology

Opponent:
Lars Feuk
Uppsala University
Department of Immunology, Genetics and Pathology

Examination Board:
Olof Emanuelsson
KTH Royal Institute of Technology
Division of Gene Technology

Marie Öhman
Stockholm University
Department of Molecular Biosciences

Gonçalo Castelo-Branco
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Molecular Neurobiology

To Mom and Dad, for always encouraging my curiousity.

”Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself.”

- Francis Crick

ABSTRACT

In the last decade we have seen a tremendous development in the omics area (genomics, transcriptomics, proteomics etc.), making high throughput methods increasingly cost-effective and available. The development in RNA-sequencing technology now enables us to sequence whole transcriptomes of hundreds or even thousands of samples or single cells simultaneously in only a few days. With the ability to quickly create millions of reads for thousands of genes in thousands of samples comes a computational challenge of how to make sense of the data. Due to the use of short sequencing reads, duplicate genes, biased base composition and repetitive regions in the genome, reads might not be uniquely assignable to a single gene. This problem can be solved either by computationally assigning multi-mapping reads to the most likely position, or excluding these reads and normalizing gene expression for the uniquely mappable positions in a gene. In **paper I**, we describe a software application for efficiently finding and storing the mappability data for every position in the genome, for subsequent use in normalization of RNA-seq data.

When the first drafts of the human genome were published in 2001, it became clear that the majority of our DNA does not consist of protein-coding genes. Since then, a multitude of new functional non-coding RNA species have been discovered, but also transcription of seemingly non-functional RNA from open chromatin regions, such as promoter upstream transcripts (PROMPTs). In **paper II**, we decipher the physical interactions between the exosome complex, the NEXT complex and the cap-binding complex, and the role each complex has in targeting PROMPTs for degradation.

In early embryonic development, having a mechanism for starting different developmental programs in a different set of cells is essential for multi-cellular organisms to develop. In the African clawed frog, *Xenopus laevis*, this mechanism involves sorting maternal RNA to different hemispheres of the oocyte, which will later be inherited asymmetrically to the cells in the developing embryo. The zygotic expression starts only after 12 cell divisions, and at the early stages the maternal RNA control the development. In **paper III**, we use *de novo* transcriptome assembly to get a good annotation of *X. laevis* in the absence of a fully assembled genome. We then use single cell RNA-sequencing to study the RNA sorting and search for sorting motifs in the 2-16 cell stage embryo.

An advantage of full length RNA-sequencing is the possibility to study alternative splicing alongside expression estimates. Spinal muscular atrophy (SMA) is genetic disease, characterized by progressive loss of somatic motor neurons. The disease is an effect of the loss of the *SMN1* gene, which is only partly compensated for by the orthologous *SMN2* gene since it is less efficient in producing full-length SMN protein. SMN is involved in spliceosome assembly, and even though it is ubiquitously expressed it specifically affects a subgroup of somatic motor neurons. In **paper IV**, we try to elucidate why some motor neurons are resistant and other vulnerable in the disease, by looking at both gene expression and splicing differences in a mouse model of SMA.

POPULAR SCIENCE SUMMARY IN SWEDISH

EN STUDIE I GENUTTRYCK

DNA är cellernas genetiska instruktionsbok som innehåller all information som behövs för att tillverka de komponenter som cellerna behöver för att fungera. Alla celler i vår kropp innehåller samma DNA (med några få undantag, som röda blodkroppar som inte innehåller DNA alls i sin mogna form) och informationen är skriven i en kod av fyra olika *nukleotider*, förkortade A, T, C och G. Trots det har olika celler vitt skilda funktioner i kroppen, som till exempel cellerna i ögats hornhinna som kan detektera ljus och sända informationen till hjärnan, nervceller som kan känna av omgivningen, lagra minnen, och kontrollera våra musklers rörelser, hjärtceller som kan kontrahera rytmiskt och simultant för att skapa våra hjärtslag, och tarmens celler som utsöndrar enzymer för att bryta ner näringsämnen i mat. Denna variation är möjlig eftersom alla instruktioner inte läses av, *uttrycks*, i alla celler samtidigt.

Informationen i DNA:t är uppdelat på mindre enheter, *gener*, ungefär som en kokbok är uppdelad i enskilda recept. När en gen läses av kopieras instruktionerna till en RNA molekyl (liknande DNA), som antingen kan ha en egen funktion i cellen, eller används som en instruktion för att skapa proteiner. Proteiner har många olika funktioner i cellerna, som att katalysera kemiska reaktioner, bygga upp strukturer som cellskelettet, och att kontrollera vilka gener som uttrycks. Proteiner består av långa kedjor av *aminosyror*. Protein-kodande RNA kallas ”messenger RNA”, eller mRNA, och innehåller en kod som motsvarar sekvensen av aminosyror i det blivande proteinet. RNA kan också ha funktioner i sig själv, till exempel mikroRNA (miRNA) som kan hindra uttryckta gener från att översättas till protein, medan transfer RNA (tRNA) används för att känna igen mRNA sekvens och översätta dem till rätt aminosyra. Det finns också många delar av vårt DNA som uttrycks och producerar RNA med en hittills okänd funktion.

I min avhandling har jag studerat identiteten och kvantiteten av RNA i olika typer av celler för att förstå mekanismer som kontrollerar genuttryck i embryo-utveckling, som respons på en sjukdom (spinal muskelatrofi, SMA), och som generella regulatoriska mekanismer i celler. Metoden jag använt kallas RNA-sekvensering, som innebär att man fångar upp och analyserar allt RNA i ett prov med celler. För att kunna läsa RNA-sekvensen omvandlar man den först till DNA (kallat cDNA), fragmenterar detta till kortare sekvenser, och kopierar upp dem i tillräckligt många kopior för att möjliggöra detektion. Sedan använder man en teknologi som genom kemiska reaktioner låter oss läsa koden av sekvenserna, och på så sätt kan vi avgöra vilka RNA-molekyler man hade i provet, och i vilka proportioner de fanns.

Min första artikel beskriver en beräkningsmetod för att förbättra analysen av RNA-sekvenseringsdata som jag implementerat till en mjukvara. Metoden kompenserar beräkningarna av genuttryck i situationer där man inte kan avgöra vilken gen en kort RNA-sekvens kommer ifrån för att sekvensen inte är unik för denna gen.

I artikel nummer två beskriver jag analysen av ett protein-komplex som bryter ner RNA. En viktig funktion hos komplexet tycks vara att bryta ner icke-kodande RNA som mest troligt inte har en funktion i cellen. Genom att slå ut olika proteiner i komplexet, antingen en och en eller i kombinationer av två och två, och studera hur det förhindrade nedbrytningen av RNA på olika sätt kunde vi avgöra de specifika funktionerna av de olika proteinerna i komplexet. På så sätt vet vi mer om hur nedbrytning av vissa typer av icke-funktionellt RNA kontrolleras.

I den tredje artikeln studerar jag hur RNA sorteras till olika positioner i ett grod-embryo i dess tidiga utvecklingsstadier. Embryot uttrycker inga egna gener under denna tid i utvecklingen, istället sorteras det RNA som fanns i äggcellen till olika positioner för att senare kontrollera vilken typ av celler som ska bildas i dessa positioner. Genom RNA-sekvensering kunde jag hitta ett antal sorterade RNA:n som inte var kända sedan tidigare, samt identifiera delar i sekvensen av de sorterade RNA-molekylerna som fungerar som sorterings-signaler. RNA-sortering är en intressant mekanism att studera eftersom den har viktiga funktioner i embryo-utveckling, och även har en funktion i nervceller i vuxna individer.

Den fjärde artikeln behandlar genuttrycks-skillnader i nervceller som en effekt av spinal muskelatrofi (SMA). SMA är en ärftlig sjukdom där man saknar en av generna (SMN1) för att skapa proteinet SMN. Människor har även en gen-kopia, SMN2, som skapar samma protein, men på grund av en skillnad i sekvensen är den mindre effektiv på att skapa det funktionella proteinet. Sjukdomen tar sig uttryck i nedbrytning av motoriska nervceller, vilket resulterar i muskelsvaghet och muskelatrofi. Motorneuronen som styr ögat tycks dock vara skyddade från nedbrytning. I detta projekt studerar vi en mus-modell av SMA för att förstå vad som gör att vissa motorneuron bryts ned, medan andra är skyddade. Vi har i nuläget hittat ett flertal gener som påverkas i unika mönster för varje celltyp, men projektet pågår fortfarande, och vi behöver både sekvensera mer prover och utföra mer analys innan vi kan dra övergripande slutsatser. Vi hoppas på att hitta genuttryck som har en skyddande mekanism i de motorneuron som inte påverkas av sjukdomen, och på så vis hitta nya potentiella terapier för SMA.

I mina projekt har jag till största del bedrivit grundforskning, det vill säga forskning som syftar till att utöka vår kunskap inom cellbiologi och hälsovetenskap utan att ha en specifik terapi eller medicin som mål. Upptäckter från grundforskningen i biologi utgör en viktig grund för vår förståelse av biologiska system, en kunskap som kan visa sig viktig för tillämpad forskning i framtiden.

LIST OF SCIENTIFIC PAPERS

I. **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

II. Andersen PR*, Domanski M*, Kristiansen MS, **Storvall H**, Ntini E, Verheggen C, et al. (2013). The human cap-binding complex is functionally connected to the nuclear RNA exosome. *Nature Structural & Molecular Biology*, *20*(12), 1367–1376. http://doi.org/10.1038/nsmb.2703

III. **Storvall H**, Reinius B, Yokota C, Björklund Å, Deng Q, Stenman J, Sandberg R (2015). Global identification of sorted RNAs in Xenopus laevis embryos using single-cell transcriptomics. Manuscript.

IV. Nichterwitz S, **Storvall H**, Comley LH, Nijssen J, Allodi I, Deng Q, Sandberg R and Hedlund E (2015). Resistant and vulnerable motor neurons display distinct transcriptional regulation in spinal muscular atrophy. Manuscript.

* These authors contributed equally

PUBLICATIONS NOT INCLUDED IN THE THESIS

Chivukula IV, Ramsköld D, **Storvall H**, Anderberg C, Jin S, Mamaeva V, Sahlgren C, Pietras K, Sandberg R, and Urban Lendahl (2015). Decoding breast cancer tissue-stroma interactions using species-specific sequencing. Breast cancer research : BCR, 17(1), p.109.

CONTENTS

LIST OF ABBREVIATIONS

ACMS	Affinity capture mass spectrometry
A	Adenine
ARS2	Arsenic resistance protein 2
bp	Base pairs
bDNA FISH	Branched DNA fluorescence *in situ* hybridization
C. elegans	*Caenorhabditis elegans*
CBC	Cap binding complex
CBP20	Cap binding protein 20
CBP80	Cap binding protein 80
CBCA	CBC-ARS2
CBCN	CBCA-NEXT
ChIP	Chromatin immunoprecipitation
CN10	Cranial nerve 10, dorsal vagus nerve
CN12	Cranial nerve 12, hypoglossal nucleus
CN3	Cranial nerve 3, oculomotor nucleus
CN4	Cranial nerve 4, trochlear nucleus
CN7	Cranial nerve 7, facial nucleus
C	Cytosine
dNTP	Deoxinucleotide triphosphate
DNA	Deoxyribonucleic acid
DEU	differential exon usage
D. melanogaster	*Drosophila melanogaster*
ER	Endoplasmatic reticulum
eRNA	Enhancer RNA
FACS	Fluorescence activated cell sorting
FPKM	Fragments per kilobase per million mapped reads
G	Guanine
iPSC	Induced pluripotent stem cells
LCM	Laser capture microdissection
LFC	log2 fold change
LINE	Long interspersed nuclear element
lncRNA	Long non-coding RNA
mRNA	Messenger RNA

miRNA	microRNA
NEXT complex	Nuclear exosome targeting complex
nt	Nucleotides
piRNA	piwi-interacting RNA
PCR	Polymerase chain reaction
P10	Postnatal day 10
P2	Postnatal day 2
P5	Postnatal day 5
PCA	Principal component analysis
PROMPT	Promoter upstream transcript
qPCR	Quantitative real time polymerase chain reaction
RPKM	Reads per kilobase per million mapped reads
RN	Red nucelus
RNA	Ribonucleic acid
RNP	Ribonucleoprotein
rRNA	Ribosomal RNA
RBP	RNA binding protein
RIP	RNA immunoprecipitation
RNAP II	RNA polymerase II
RNA-seq	RNA sequencing
SINE	Short interspersed nuclear element
smFISH	Single molecule *in situ* hybridization
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
snoRNA	Small nucleolar RNA
SC	Spinal cord
SMN	Survival Motor Neuron
T	Thymine
TF	Transcription factor
TSS	Transcription start site
tRNA	Transfer RNA
TMM	Trimmed mean of M values
U	Uracil
X. laevis	*Xenopus laevis*

1 THE CENTRAL DOGMA AND BEYOND

1.1 MACROMOLECULAR BUILDING BLOCKS OF THE CELL

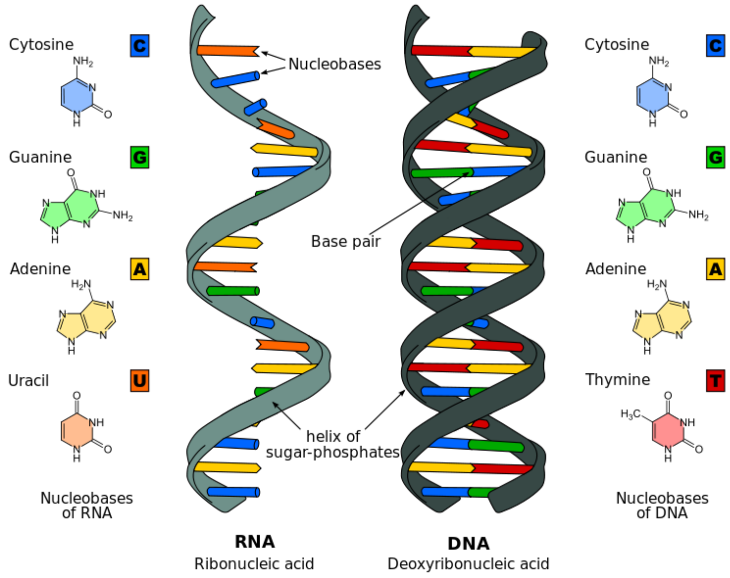

Figure 1. The structure and molecular building blocks of RNA and DNA (User:Sponk / Wikimedia Commons / CC-BY-SA-3.0).

Essentially all eukaryotic cells contain 3 important types of macromolecules that exert the function of the cell: DNA, RNA and proteins. DNA and RNA consist of chains of nitrogenous *bases* on a backbone of either deoxyribose (DNA) or ribose (RNA). The combination of a base, a ribose or deoxyribose molecule and a triphosphate molecule is called a *nucleotide.* In DNA, the four bases of the genetic code are Adenine (A), Thymine (T), Guanine (G) and Cytosine (C), while RNA has Uracil (U) instead of thymine. The bases have an interesting property of complementarity, where adenine can form hydrogen bonds with thymine or uracil, while guanine forms hydrogen bonds with cytosine (Fig 1). In 1953, the structure of DNA was published by James Watson and Francis Crick, based on the X-ray diffraction data from Rosalind Franklin (Watson & Crick 1953). They showed that this complementarity of bases caused DNA to form a double stranded helical structure, with one strand complementary to the other (Fig 1).

Just like RNA and DNA, proteins are also made up of long chains of smaller molecules, in this case amino acids, bound together by *covalent* peptide bonds (a fairly strong bond with shared electrons between the amino acids). Each protein is folded into a specific 3 dimensional structure to be able to exert its function. There are 20 amino acids with different inherent properties – 10 of them polar and thus *hydrophilic* (water-soluble), and the other 10 non-polar and *hydrophobic* (water in-soluble). The polar amino acids can have a positive or negative charge, or be uncharged. Due to these different properties, they can form weak, *non-covalent*, bonds to each other and support the 3D structure of the protein. For example, positively charged amino acids can form an ionic bond with negatively charged ones and uncharged polar amino acids can form hydrogen bonds to each other. The importance of the

3-dimensional structure can be exemplified by enzymes requiring certain amino acids to come together in a catalytic “pocket” where chemical processes can be aided, by the ability to make long fibres that provides the structural cytoskeleton of the cell, or by proteins that requires hydrophobic residues on their outside to allow them to stay in the cells fatty lipid membrane, while having a hydrophilic inside able to transport water soluble particles into or out of the cell. In other words, the structure is important for both activity, location, and physical properties of proteins (Alberts 2004).

1.2 THE SEQUENCE HYPOTHESIS AND THE CENTRAL DOGMA OF MICROBIOLOGY

In 1958, Francis Crick presented the idea of the Central Dogma in Microbiology as well as the Sequence Hypothesis (Crick 1958). The essence of the Central Dogma was the transfer of information from DNA to RNA and finally to protein, including also self-replicating potential of DNA and RNA. At this time, proteins were known as the most important functional units of the cells, providing diverse functions such as structure (e.g. keratin, providing structure to skin, hair and nails) and enzymatic activity for accelerating, *catalysing,* chemical processes (such as lipases breaking down fat), while the DNA was known as the carrier of genes, the heritable material, and thus containing instructions for making proteins. It was clear that the 3-dimensional structure of proteins was important for their function, but how proteins were synthesised and how the structure was achieved was still largely unknown and debated. Crick proposed that the 3-dimensional structure was in fact a consequence of the linear sequence of amino acids in proteins, and that this linear sequence was coded for in the DNA. This sequence hypothesis stood in contrast to a theory by Linus Pauling, that it was the *structure* rather than the sequence of protein that was encoded for in the genetic material (Strasser 2006), which had been a prevalent theory until that time. The role of RNA as a messenger, conferring information from DNA to protein was also discussed in Cricks review, and although the evidence for it was meagre, he also predicted the existence of transfer RNAs (which he called adapters) that helped translating the messenger RNA code into protein sequence (Crick 1958).

To a large extent, both the sequence hypothesis and the central dogma still holds true today. In the early 1960’s, the genetic code was deciphered by Marshall Nirenberg and his team, an effort that he was awarded the Nobel prize for in 1968 (Nirenberg 2004). Thanks to his work, we know now that proteins are encoded for in the DNA in the form of 3-nucleotide long *codons*. Out of the possible 64 combinations of trinucleotides, 61 corresponds to a specific amino acid, meaning that there is a degeneracy, and several codons code for the same amino acid in some cases. The remaining 3 codons code for transcription termination (Nirenberg et al. 1966). The genetic information from protein-coding genes is first copied, *transcribed,* from DNA into messenger RNA (mRNA). Transcription is controlled by transcription factors that recognize a *promoter* sequence upstream of the sequence of the gene itself. Transcription factors recruit the transcription machinery, including the RNA polymerase that copies the DNA sequence into the complementary RNA sequence. The mRNA is then transported to the cytoplasm where it is translated into protein by the ribosomes. The translation is mediated by transfer RNAs (tRNAs), which carry specific amino acids to the ribosome depending on the 3 nucleotides in their anticodon site, and builds the protein sequence by base pairing to the mRNA (Fig 2) (Alberts 2004). However, this is a simplified description, and the sequence hypothesis and central dogma does not tell the whole story.

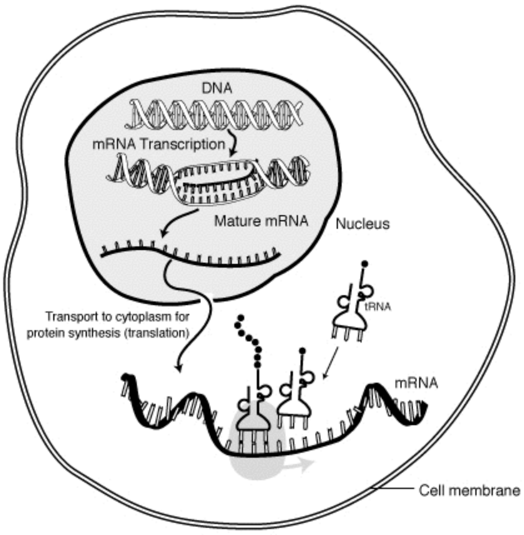

Figure 2. Transcription and translation of mRNA (User:Sverdrup / Wikimedia Commons / CC-BY-SA-3.0)

1.3 PRE-MRNA PROCESSING

The relationship between DNA sequence and protein sequence is not as simple as just a 3 nucleotide to one amino acid code. In fact, 95% of the sequence of human protein-coding genes does not contribute to producing amino acid sequence (Venter et al. 2001; Lander 2011). Before the mRNA is transported to ribosomes and translated, they go through a series of processing steps. The beginning of the mRNA that is transcribed first, called the 5′ end, is *capped*, which means that a guanine nucleotide modified with a methyl group is added to this end. The other end, the 3′ end, is *polyadenylated*, i.e. a tail of multiple adenine nucleotides is added. These modifications are thought to promote transport to the cytoplasm, promote translation, and also increase the stability and prevent degradation (Alberts 2004). Another processing step is *splicing* which essentially means that the splicing machinery, a complex consisting of both proteins and RNA, cut out chunks of the RNA and then join the ends of the remaining pieces back together. The spliced out regions are called introns, while the remaining regions are called exons (Gilbert 1978). Splicing occurs in essentially all *eukaryotes* (organisms with a cell nucleus), with the only known exception being a *nucleomorph*, an intracellular symbiont of algae, with a tiny genome (Lane et al. 2007). In *prokaryotes* (bacteria and archaea) on the other hand, only a few instances of splicing have been found with introns that are either self-splicing and excises themselves without the use of a splicing machinery (Edgell et al. 2000), or spliced with the splicing endonuclease that in eukaryotes is dedicated to splicing tRNA (Watanabe et al. 2002; Yokobori et al. 2009; Calvin & H. Li 2008). The prevalence and length of introns increase with the complexity of the organism, with unicellular eukaryotes exhibiting the shortest introns while *vertebrates* (organisms with a spine and skeleton) have the longest ones. How splicing evolved is not completely understood – some scientists subscribe to the idea that accumulation of introns occurred to allow for increased evolutionary complexity by providing alternative gene products from one gene, and the possibility of creating new genes from combinations of exons. Another theory is that introns occurred because of a lack of selective pressure due to the populations of multicellular organisms not being large enough to eliminate occurrence of introns (Rogozin et al. 2012). In either case, it is clear that introns do increase the organismal complexity by increasing the number of possible gene products. The reason for this is that the splicing pattern of a single gene is not necessarily constant. Many genes have several alternative splicing patterns, were different introns are included or excluded, resulting in different final versions, *isoforms*, of the mRNA being produced. In fact, over 90% of the human genes are subjected to alternative splicing (E. T. Wang et al. 2008). Using different exons can have a very direct effect by changing the sequence of the protein produced. For example, in the genes in the major histocompatibility complex (MHC), alternative splicing is important to generate diversity for antigen recognition in the immune system (Gamazon & Stranger 2014). Another example is the enzyme A2 beta which changes substrate specificity upon splicing (Ghosh et al. 2006). In many cases though, alternative splicing results in a loss of protein-coding potential. Interestingly, in fruit flies, alternative splicing of the *Sxl* gene is essential for sex determination, where the male version includes an exon that prevents the protein from being expressed (Salz 2011).

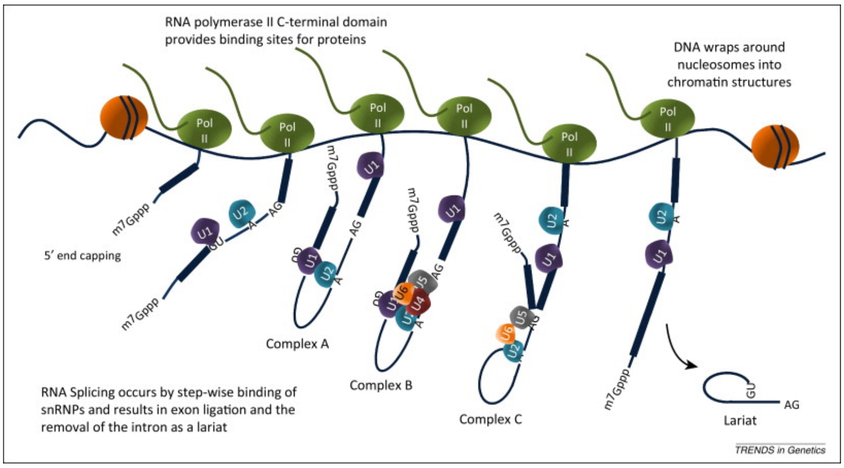
RNA polymerase II C-terminal domain provides binding sites for proteins
DNA wraps around nucleosomes into chromatin structures
Pol II
Pol II
Pol II
Pol II
Pol II
Pol II
m7Gppp
5′ end capping
m7Gppp
U1
GU
U2
A
AG
m7Gppp
U1
AG
GU
U1
U2
A
Complex A
m7Gppp
U1
AG
U5
U6
GU
U1
U4
U2
A
Complex B
m7Gppp
U2
A
U1
U5
AG
U6
U2
Complex C
U2
A
U1
m7Gppp
GU
AG
Lariat
RNA Splicing occurs by step-wise binding of snRNPs and results in exon ligation and the removal of the intron as a lariat
TRENDS in Genetics

Figure 3. Co-transcriptional mRNA processing. Capping occurs soon after transcription. Splicing factors bind to the emerging 5' splice site and branch point and the intron is excised before transcription terminates (Wong et al. 2014)**.**

1.4 ALTERNATIVE SPLICING

Regulation of alternative splicing is a complex process involving a multitude of different factors. The splicing reaction itself is performed by the *spliceosome*, a large complex consisting of multiple protein and RNA components. The spliceosome recognizes three splicing signals consisting of specific sequences patterns, *motifs*, in the pre-mRNA. Two of these, the 5′ and 3′ splice sites mark the exon-intron boundaries on both sides of an intron, while the third, the branch point is internal to the intron. The sequence motifs are quite loosely defined in *metazoans* (multicellular animals), i.e. there are many variations in the sequence of these splicing signals between exons in the same species (Y. Lee & Rio 2015). As a consequence, some "strong" splice sites have a greater affinity for the spliceosome and will more frequently result in splicing than "weak" sites (Koren et al. 2007). However, there are other factors determining the splicing pattern at a given time. Apart from the splicing signals, there are also motifs in both exons and introns that work as splicing enhancers or splicing repressors by attracting repressive or activating factors (Y. Lee & Rio 2015). More and more evidence suggests that most splicing in eukaryotes occurs co-transcriptionally, i.e. before transcription of the gene is completed (Fig 3), which also has functional implications on splicing. There is evidence showing physical interaction between RNA polymerase II (RNAP II) and splicing proteins, which can be altered by post-translational modifications to RNAPII, suggesting that it has a function in recruiting splicing factors. The rate of transcription has also been implicated in alternative splicing, where a slower transcription elongation can lead to more efficient splicing at weak splice sites. This complexity in splicing regulation allows for flexibility and increased complexity in the gene products produced. However, the system is also sensitive to perturbations, as mutations in the different RNA motifs can disrupt the normal splicing patterns (Merkhofer et al. 2014). One example is the mutation of a splicing enhancer in the SMN2 gene transforming it to a splicing repressor (Cartegni & Krainer 2002; Kashima & Manley 2003), which has important implications in Spinal Muscular Atrophy that we discuss in paper IV.

1.5 COUNTING GENES

The central dogma assumed that DNA was the carrier of genetic information, RNA was a means of transferring the information, and proteins were the functional units of the cell. This was still the predominant idea in the early 2000's, when the sequencing of the human genome was nearing completion. At this time a betting pool, Gene Sweepstake, was initiated at a Genome Sequencing and Biology meeting in Cold Spring Harbor Laboratory in 2000 where scientists entered their guess on the number of genes that would be found in the human genome (Choi 2003). The genome of the roundworm *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster* had already been sequenced, and the gene counts were around 18,500 and 13,500 respectively. Although many – including James Watson, one of the scientists behind the structure of DNA – were surprised that the fruit fly had fewer genes than the less complex roundworm, scientists were still expecting the gene count in humans to be vastly larger to reflect the complexity of human physiology (Watson 2001). Bets in the betting pool averaged around 61,000 genes, with many going up to over 100,000 (Choi 2003). However, to everyone's surprise, the protein-coding potential of the genome turned out to be much less than expected. At the end of the competition only 24,847 protein-coding genes had been identified, partly by applying prediction algorithms to genomic sequence. Today, only 21,990 of these have been validated as protein-coding (known protein-coding genes retrieved from Ensemble database 2015-12-04). The human genome consists of around

3 billion base pairs, and only ~25% of these are contained within protein-coding genes, and that is including the intron sequence – the protein-coding exonic sequence compose only about 3% of the genome (Venter et al. 2001; Hangauer et al. 2013). *C. elegans* on the other hand has 27% exonic and 26% intronic sequence in its 95.53 million base pair genome (The C elegans Sequencing Consortium 1998), while *D. melanogaster* has 20% exonic and 17% intronic sequence in a 120 base pair genome (Adams 2000). What is the remaining 75% of the human genome doing?

1.6 NON-CODING RNA - REDEFINING THE CONCEPT OF A GENE

Only a few years after Cricks predictions of RNA based "adapters", they were identified as transfer RNAs and their structure and sequence solved (Holley et al. 1965). Ribosomal RNA (rRNA) was also known at this point, although one was not certain whether it was translated or not. These were the first known RNAs that are not translated into proteins, but instead exert a function in the cell in RNA form. The small nuclear RNAs (snRNAs) involved in splicing were found in the late 1970's (Murray & Holliday 1979). Although a few other *non-coding* (i.e. not coding for protein) RNA species were found during the subsequent decades, not much attention was paid to this group until after the human genome sequence was solved in the early 2000's, and it stood clear that protein-coding genes only made up a minority of the sequence in the human genome (Eddy 2001). After this, it became evident that the view of genes mainly as protein-coding units needed to be revised. The research on non-coding RNAs has bloomed since then, and many new species has been found (Makarova & Kramerov 2007). In human, there are now over 23,000 annotated non-coding RNAs (ensemble database, not including pseudogenes), and they make up a diverse group regarding size, structure and function. For example, micro RNAs (miRNAs) are short (22 base pair) double stranded molecules that regulate mRNA expression, by base pairing to a complementary target mRNA and inhibit translation or induce degradation. Piwi-interacting RNAs (piRNAs) is another type of short, inhibitory RNA, but instead of targeting mRNA it silences *transposons*, mobile DNA elements that essentially parasitize on our genome (Makarova & Kramerov 2007). The small nucleolar RNAs (snoRNAs) are slightly longer, 60-300nt, and function by modifying other RNA (rRNA, snRNA and mRNA) post-transcriptionally (Mattick & Makunin 2006). The long non-coding RNAs (lncRNAs) are in itself a diverse class, with the common denominator that the transcripts are long (>200nt). Some of them are expressed at distinct locations distant from protein-coding genes, while others are *antisense* transcripts, transcribed from the complementary strand to the protein-coding DNA, and yet others originate from introns of protein-coding genes. One of the first identified lncRNA was *Xist*, which is essential for inactivating one of the X-chromosomes in females of placental mammals. Interestingly, *Xist* itself has an antisense transcript, *Tsix*, which negatively regulates the expression of *Xist.* In general, lncRNAs are often involved in altering the chromatin state, mostly in a repressive manner to inactivate transcription from particular regions of DNA (Kung et al. 2013).

In the last few years, several studies has shown that around 85% of the human genome is transcribed across a diverse collection of tissues (Djebali et al. 2012; Hangauer et al. 2013), although it has been debated whether some of the very low expression is actually functional (J. Wang et al. 2004; Pertea 2012). A large portion of this transcription constitutes intronic sequence or known non-coding genes, but there is still some dark matter in the RNA world that remains to be understood. One example of this is the transcription that occurs at the initiation site of transcription, the *promoter*, called promoter upstream transcripts (PROMPTs). PROMPTs can be transcribed in either sense or antisense direction with respect to the downstream gene, and they are fairly unstable and quickly degraded by the exosome complex (Preker et al. 2008). For most PROMPTs, no function has been found, except for a few cases where the PROMPTs down regulate expression of the downstream gene when their degradation is inhibited (Lloret-Llinares et al. 2015). Possibly, this transcription is only an inevitable consequence of recruiting the transcriptional machinery to the gene, and the machinery not being able to specifically transcribe only in the direction and strand of the actual gene. Another similar case is the enhancer RNAs (eRNAs). Enhancers are genomic elements that aid in transcription activation by binding to transcriptional activator proteins and bringing them to the promoter of a gene. They are often found at a distance from the gene itself, and come into physical contact with the promoter by looping the DNA (Alberts 2004). In 2010, Kim et al. discovered that DNA at active enhancer regions are being transcribed in a bi-directional manner, i.e. on both the forward and reverse strand but in opposite direction, starting from the centre of the enhancer (Kim et al. 2010). They found that the eRNA expression was correlated with the mRNA expression of the corresponding regulated gene and hypothesized that the transcription only occurs when enhancers are bound to active promoters. However, they could not deduce if this transcription had a biological importance, and if so, if it was the act of transcription that was important to keep the chromatin in an active state or if the transcripts themselves where functional. Recent studies have shown evidence of eRNAs serving a functional role in stabilizing the enhancer-promoter looping (W. Li et al. 2013; Hsieh et al. 2014) and that transcription of eRNAs precedes that of transcription at the target promoter upon cell transition intitation (Arner et al. 2015), suggesting eRNAs could have a functional role in transcription regulation.

A big portion of vertebrate genomes consists of *transposable elements*, or transposons, popularly called jumping DNA because of its inherent ability to move around in the genome. In humans, 45% of the DNA consists of remnants of these transposable elements, although most have lost the ability to “jump” due to mutations in their sequence (Pace & Feschotte 2007). There are two classes of transposons: *retrotransposons* that move by first being transcribed to RNA, and then reverse transcribed back into DNA and inserted at a new position in the genome, and DNA transposons that move by “cutting and pasting” its DNA sequence without an RNA intermediate. Two common types of retrotransposons are the Long

Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs). LINEs are transposable elements that also codes for the reverse transcriptase needed to transcribe it back to DNA (Alberts 2004), while SINEs lack the reverse transcriptase gene and utilizes that expressed by LINEs (Makarova & Kramerov 2007). These genes have been mainly considered as "selfish DNA", essentially parasites on the vertebrate genome. However, it has been found that the transcripts of a type of primate specific SINE, the Alu elements, are involved in regulating other genes both on a transcriptional and alternative splicing level (Dridi 2012). Thus, the vertebrate genome contains a very diverse repertoire of genetic material, and although some of it still seems like "junk" DNA, more and more functionality is continuously being discovered.

1.7 LOCALIZATION OF RNA

Adding to the complexity, RNA can also be localized to different sites or substructures of the cell. After transcription, mRNA is transported from the nucleus to the cytoplasm, where it is translated into protein. The transport through the nuclear membrane is aided by 5′ cap binding proteins (CBP20 and CBP80). However, the translation can take place at different locations in the cytoplasm. Many mRNAs are specifically transported to the *endoplasmic reticulum* (ER), an organelle consisting of a network of membranes, which is responsible for lipid synthesis and protein secretion. The mRNAs translated at the ER consists largely of secreted proteins, transmembrane proteins, and proteins serving a function in the ER itself (Chen et al. 2011). Other mRNAs localize to the *mitochondria*, the cells metabolic centres and energy factories, and these have been shown to encode proteins important for mitochondrial functions (Corral-Debrinski et al. 2000). Together this suggests that mRNA localization has a function in producing the protein product at the right location in the cell. Other types of RNAs, such as the non-coding RNA *Xist* never leaves the nucleus at all, since its function of transcriptional silencing is exerted in the nucleus.

RNA localization can also serve more cell type specific functions in certain cells. An example of this is RNA localization in *oocyte* (egg cell) formation in some organisms, like the fruit fly, *Drosophila melanogaster,* and the African clawed frog, *Xenopus laevis*. Here, RNA localize to different sides of the oocyte, which will later determine the axis of the embryo and the identity of the cells inheriting the localized RNA. Both protein-coding and non-coding RNAs have been shown to localize in the oocytes, and in *X. laevis* the non-coding Xlsirt RNAs that localize to the vegetal pole even have a role in directing localization of other RNAs (Kloc & Etkin 1994). Neurons also have a specific localization pattern, where some mRNAs are transported to and translated in the axon. The RNA of the CREB protein is one example, which has been shown to be translated in the axon upon extracellular stimulus and the protein is subsequently transported to the nucleus where it triggers a transcriptional response (Cox et al. 2008).

The mechanism of targeting RNAs to different locations in the cell is complex and not fully understood. The transport can be either active and involving motor protein transporting the RNA along the microtubule cytoskeleton, or take place by passive diffusion and entrapment at the target site. In either case, there needs to be some kind of "zipcode", *motif*, in the RNA to target it to different sites. For a few RNAs, a region required for sorting has been determined, however in most cases it is not a case of one simple motif, but rather a cluster of several short motifs. How the region is recognized for localization, i.e. possible importance of the spacing of motifs or secondary structure, is not fully understood.

There are RNA binding proteins (RBPs) that can recognize and bind to specific RNA localization motifs, though they are usually not exerting the function alone but rather in complexes with other proteins and sometimes also non-coding RNAs. RNA can also be localized to granular structures containing proteins and several different RNA molecules in multiple copies, which are then localized in a coordinated fashion. Although a few cases have been thoroughly examined and mapped out, in most cases the full picture of which combination of proteins coordinates localization and how they recognize their target RNAs is not known (Blower 2013).

1.8 CREATING CELL DIVERSITY FROM ONE GENOME

Assuming that the sequence of A, T, G and C, and subsequently sequence of amino acids were the only thing coding for the cells functionalities would make it hard to explain the diversity of different cell types in multi-cellular organisms such as ourselves. The genetic code that is inherited from our parents is also inherited from cell to cell as our body develops, and each cell will contain the same DNA (with the exception of rare novel mutations, specialized cell types such as red blood cells that do not contain DNA, or germ cells that only contains half of the genetic material). Still, different types of cells in our body have very different functions – from the light sensing cells of the retina to rhythmically contracting cells of our hearts, to excretory cells of the gut producing enzymes digesting food, to the diverse set of nerve cells in our brains capable of relaying information about the environment, controlling our movements and storing information.

The reason why this differences can occur is that while the DNA sequence is the same across cells, the RNA and protein populations are not. Each cell has its specific sets of genes being expressed, i.e. translated into RNA and protein. The master regulators of expression are the transcription factors, a set of proteins with DNA binding domain. The DNA binding domains of different transcription factors (TFs) recognize different sequence motifs controlling transcription, both at promoters close to the transcription start site of genes, and at enhancers that can occur up to 1 million base pairs (bp) away from the promoter. Transcription factors usually have a set of genes that they activate, and can thus start and maintain whole programs of transcription. However, each cell must know what transcription factors to express in the first place to give the cell its specific identity. Moreover, to enable renewal of the cell population in a tissue or its growth during embryogenesis, the cells need a way of inheriting cellular identities without altering the DNA sequence. This form of inheritance is called *epigenetic* inheritance, and is defined as a self sustained mark that can be transferred to daughter cells after division, and has an effect on gene expression. Epigenetic mechanisms can be trans-acting, i.e. originate from one site on the genome and exert its function at another site. The most common example of this mechanism is transcription factors that are self-propagating, i.e. regulates itself by a positive feed back loop, as well as regulating other genes controlling the cell state. Alternatively, cis-acting epigenetic mechanisms act at to regulate the site of the inherited trait. One example of a cis-acting epigenetic mark is the silencing of one of the X-chromosomes in female mammals by *Xist* expression from the same chromosome (Bonasio et al. 2010). Another example is DNA methylation of repetitive regions that serve to keep the region silenced (Jones 2012). The DNA is wound around protein complexes called *histones*, and different covalent modifications of histones have also been shown to be associated with different transcriptional states, although a stable inheritance has not been shown, and there is no evidence of the modification being causal to activation or silencing (Schübeler 2015).

The first epigenetic marks are established during embryonic development to initiate differentiation of cells towards specific fates. The initial establishment of an epigenetic state is usually exerted by transient expression of differentiation factors that initiates expression of other self-propagating factors (Bonasio et al. 2010). However, although epigenetic marks are inheritable from a cell to its progeny, they are not static, and as development proceeds, marks are lost and gained to increasingly specialize cells for certain functions. It is even possible to reverse the whole epigenetic state of a fully differentiated cell and revert it back into a stem cell state by adding a specific set of transcription factors, creating so called induced pluripotent stem cells (iPSCs) (Takahashi & Yamanaka 2006), a discovery that was awarded with the Nobel prize to Shinya Yamanaka and John Gurdon in 2012.

1.9 POST-TRANSCRIPTIONAL AND POST-TRANSLATIONAL MODIFICATIONS

Apart from the complexity in regulating expression explained above, there are more layers regulating the function of the final gene products. There are post-transcriptional modifications exerted on both protein-coding and non-coding RNA, some of them by trimming the RNA to create a functional product, such as for miRNAs and tRNAs. Another type of modification, RNA editing, change the RNA sequence itself, and can in case of mRNA alter the composition of the final protein (Gott & Emeson 2000). There is also control exerted after translation – chemical modifications of proteins can also alter the function and/or structure of the protein. Phosphorylation for example is a common way of switching proteins enzymatic function on or off (Johnson 2009), while lipidation (covalent attachment of a lipid) allows the protein to locate to and interact with the cell membrane (Hentschel et al. 2015).

1.10 SUMMARY

Taken together, the regulatory principles discussed in this chapter shows that the complexity of gene regulation in vertebrate cells are much more complex than could be imagined at the time the central dogma was postulated. There are still many aspects that are not fully understood. With the advent of new high throughput technologies, studying genomics, gene expression and protein abundances at a global level is becoming increasingly affordable and available, although we still have a long way to go before we completely understand the relationship between all the regulatory mechanisms controlling cell identity and behaviour.

In this thesis, three out of four papers are global studies of RNA expression in different biological contexts, while the fourth (**paper I**) describes a useful tool for analysing high throughput transcriptomics data. Why is it interesting then to study RNA expression, when these molecules are often intermediate steps to a final functional product? One simple reason that RNA is more widely studied than proteins is simply that the technology for high throughput transcriptomics measurements has advanced quicker than that of proteomics. As I will discuss in the next section, it is possible to determine the sequence and quantity of nearly the whole RNA population in a biological sample. Although mRNA levels and protein levels are not perfectly correlated (Vogel & Marcotte 2012), studying mRNA abundance will still give an idea of the protein population in a cell. However, there are also questions that can only be answered with RNA level data – those regarding regulatory principles of expression. In **paper II** we studied the principles of expression and degradation of PROMPTs, which can of course only be determined on RNA-level, since they have no protein product. In **paper III** we study differential localization of RNA in the *Xenopus laevis* early embryo, a crucial mechanism for patterning the embryo at later stages. Finally, in **paper IV** we study both transcriptional and splicing effects in a disease model to understand the disease progression. Splicing effects might not be detectable at all at the protein level, since alternative splicing can lead to transcripts incapable of producing protein.

2 SEQUENCING

2.1 READING AND DECIPHERING THE CELLS INSTRUCTION BOOK

After the structure and composition of DNA was solved in 1953, and the key to deciphering the genomic code was identified in the early 1960's, the next big question was – how do we read the code written in the DNA?

Already in the early 50's, a method for determining protein sequence was developed, where the N-terminal amino acids were sequentially cleaved off and identified by their chemical properties (Edman et al. 1950). However, sequencing DNA was more difficult. The protein sequencing method was already laborious and time consuming, and could not be employed on peptides longer than 50-60 amino acids. DNA is generally much longer, and the nucleotides are much more chemically similar to each other, so the same methods as for amino acids could not be used. In addition, no enzymes that cleaved DNA at base-specific sites were known at the time (Hutchison 2007). It took until 1965 before the first nucleotide sequence was solved, which was that of tRNA (Holley et al. 1965). Sequencing at this time was done using a digestion-based method, where the RNA was partially digested from one end, and the sequence was determined by studying the size and composition of the shorter fragments produced using a method called 2D-diffraction (Sanger et al. 1965). The first complete genome, that of the bacteriophage MS2, was sequenced using this technique in 1976 (Fiers et al. 1976).

In 1977, Fred Sanger developed a technique that was a major breakthrough for the sequencing field. His method utilized the DNA polymerase, and instead of sequencing by digestion the method employed a sequencing-by-synthesis approach (Sanger et al. 1977). The method was based on providing the DNA polymerase with both normal nucleotides (deoxynucleotide triphosphates, dNTPs) and nucleotides blocked at the 5′ end, making them unable to bind additional nucleotides and they would therefore terminate the DNA synthesis after being incorporated. By adding the modified nucleotide at a much lower concentration compared to the un-modified ones, short DNA sequences, *oligonucleotides*, of different lengths would be synthesized before randomly incorporating a modified base. The synthesized DNA fragments were separated by size using gel electrophoresis, a method based on loading molecules on a polyacrylamide gel and applying current, which causes the molecules to move at speeds negatively proportional to their size. In the original protocol, four different reactions were performed, with one of the four nucleotide modified in each, and each loaded separately on the gel to enable identification of which nucleotide each fragment ended with (Sanger et al. 1977). This method, popularly called Sanger sequencing, became widely used, and subsequently improved by using modified nucleotides labelled with a different fluorescent dye for each base to enable performing the whole reaction in only one tube (Fig 4a). The fluorescent label allowed for automatic detection, and machines were built that would run many samples in parallel and perform automatic fluorescence detection.

When the human genome project started, Sanger sequencing was still the predominant method. To enable sequencing of the vast amount and length of the human DNA sequence, “shotgun” sequencing approaches were used. These approaches involved fragmenting the DNA into shorter pieces, cloning them into bacterial artificial chromosomes (BACs), which are were inserted into bacteria where they could be amplified by the bacteria’s replication system. This approach was necessary since the sequencing machines had a limitation of how long reads it could sequence, and amplification was needed to have enough input material to enable detection. The human genome project started as an academic initiative, with many labs across the world dividing the work between them (Lander et al. 2001). However, a competitive effort to sequence the human genome was initiated by Craig Venter and his company Celera Genomics, which employed a less laborious but more computationally challenging method to assemble the genome. The threat of Celera getting there first and patenting the human genome sequence sped up the effort by the academic non-profit organization, and in 2001 both groups published their first drafts of the human genome in Science and Nature respectively (Lander et al. 2001; Venter et al. 2001).

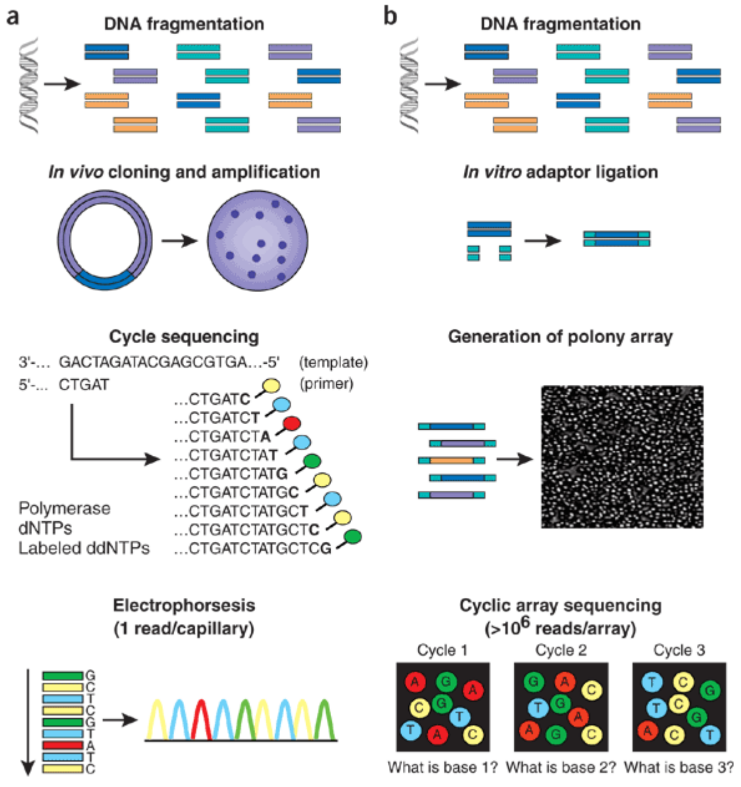

Figure 4. Workflow of (a) Sanger sequencing and (b) second generation sequencing. (Shendure & Ji 2008)

2.2 SECOND GENERATION SEQUENCERS

Even with the approach of Craig Venter, the process of assembling the human genome was long and resource demanding. With the increasing interest in genomic data, multiple new techniques were developed in the early 2000's, and the second generation of sequencing started to emerge. The new technologies provided a much higher throughput by massively parallelizing the sequencing. This was achieved by eliminating the gel electrophoresis step, and instead keeping the DNA fragments stationary while decoding the sequence step by step, allowing many sequencing reactions to take place in parallel on a small surface (Fig 4b). The first publicly available second-generation sequencing method was the 454 pyrosequencing. With this method, the DNA is fragmented and each short fragment is bound to a micro-bead where it is amplified in place by polymerase chain reaction (PCR), and each bead will in the end be covered with multiple copies of the same single stranded DNA sequence (Fig 5a). The beads are then transferred to a plate containing micro-wells, holding only one bead each. A sequencing-by-synthesis approach is then employed, where the plate is flooded with one type of base at a time. As bases are incorporated ATP is produced from pyrophosphate, which acts as a substrate for an enzyme that produces a luminescence signal, which in turn can be detected to deduce where the base was incorporated.

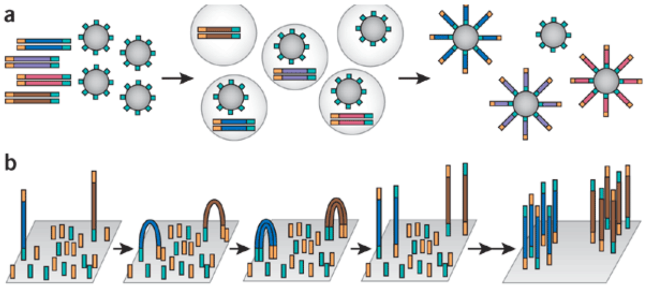

Figure 5. Clonal amplification of next genereration sequencing libraries. (a) Emulsion PCR used in 454 and SOLiD sequencing. (b) Bridge amplification used in Illumina sequencing. (Shendure & Ji 2008)

Solexa, now acquired by Illumina, developed another massively parallel sequencing-by synthesis method. In this method, the DNA fragments are attached to a glass slide where they are locally amplified into clusters of the same fragment (Fig 5b). By adding all four fluorescently labelled bases simultaneously, which are reversibly blocked from elongation at the 5´ end, one base is incorporated at each cluster at a time. After identifying the base, both the blocking and the fluorescent label are removed to enable next cycle to proceed. The SOLiD technology (Sequencing by Oligonucleotide Ligation and Detection) also appeared around the same time, and uses a ligation based technology combined with a micro bead PCR method (Heather & Chain 2015). These new technologies dramatically increased the throughput of sequencing – where the Sanger sequencing machines could sequence hundreds of fragments at a time, the new methods sequenced millions. The major drawback compared to the Sanger sequencing is that these methods generally produce shorter read lengths, typically between 35-400bp, which is compensated for with more sophisticated and efficient computational tools to enable assembly of whole genomes. Today, sequencing of a human genome at a 30x depth can be performed in only a few days and at a cost of only $1000 on Illuminas HiSeq X Ten system, compared to the $2.7 billion dollars and decade spent to sequence the first human genome, opening up for potentials to use the method for clinical applications.

2.3 THIRD GENERATION SEQUENCERS

There is not a clear consensus on what defines the third generation sequencers. Some define it as single molecule sequencing, where the SMRT platform from Pacific Biosciences is included. In this method, an array of nanostructures called zero mode waveguides (ZMW) is used, where one DNA polymerase molecule is attached to the bottom of each ZMW. As the polymerase copies a DNA strand and incorporates fluorophore tagged bases, a signal can be read at the bottom of the well. The ZMWs are designed so that their diameter is smaller than the wavelength of the laser used for detection, which causes the light to decay exponentially as it passes through, and detection only occurs at the very bottom (Fig 6a). This technique makes it possible to detect signal from a single molecule, instead of having to amplify it into a cluster of identical sequences to enable detection (Heather & Chain 2015).

Others count the SMRT platform to the 2nd generation sequencers, because it utilizes a DNA polymerase enzyme that copies the sequence in order to read it, while a true third generation sequencer should read the original DNA molecule itself. The third generation sequencers (according to this definition) that are under development today all utilizes either biological nanopores (i.e. naturally occurring and protein based) or solid-state nanopores (synthetically manufactured by materials such as silicone and graphene). The pores are embedded into a lipid layer, similar to the cell membrane, and by applying a current to the membrane ions are transported through the pore. The current over the membrane can be measured, and when a DNA molecule is threaded through the pore each nucleotide will disrupt the ion transport to a different degree, and can be measured by a change in current (McGinn & Gut 2013). Oxford Nanopore Technologies (Fig 6b) were the first to have a nanopore sequencer on the market, one of which is a USB sized portable device. The quality of the nanopore sequencing is not yet at the same level as the second generation sequencers, but the technology is still being developed and holds a great promise for the future (Heather & Chain 2015).

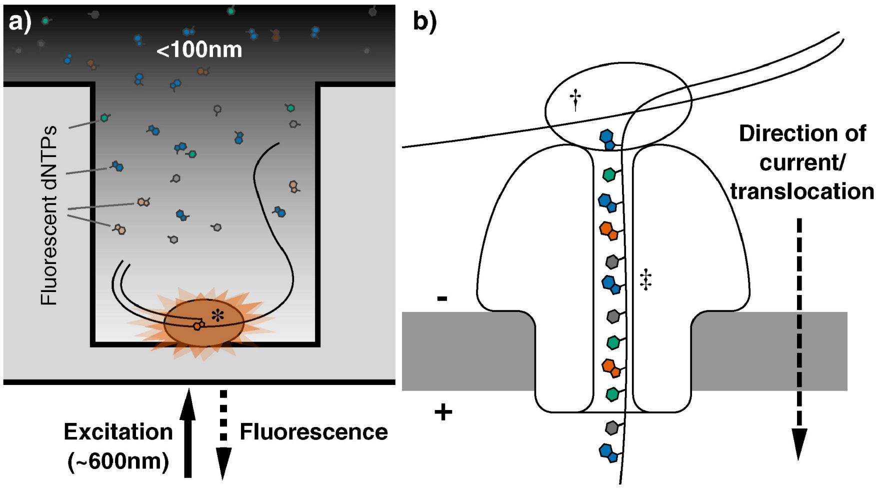

Figure 6. Third generation sequencing technology. (a) SMRT platform from Pacific Biosciences. (b) Nanopore DNA sequencing. (Heather & Chain 2015)

2.4 GENE EXPRESSION QUANTIFICATION

The simplest method for measuring RNA abundance is northern blot, a gel electrophoresis method with quantification measured by strength of staining has been around since 1977 (Alwine et al. 1977). In the 1990's, more sophisticated gene expression methods started emerging, with the first one being quantitative real time PCR (qPCR), an extension of the PCR (polymerase chain reaction) method. PCR is a method for amplifying DNA in a test tube without inserting the DNA into a living organism (such as bacteria). It is done by iterative cycles of denaturing DNA into single strands, annealing of a short complementary *primer* sequence to prime it for transcription, and transcribing the DNA using a bacterial DNA polymerase, thereby generating multiple copies of a specific original sequence determined by the primers. In qPCR, a fluorescent signal proportional to the amount of DNA produced at each cycle is generated, and by determining at which cycle it reaches a threshold level an estimation of the amount of starting material can be made (Higuchi et al. 1992; Higuchi et al. 1993; Wittwer et al. 1997). The same method can be performed on RNA by first enzymatically *reverse transcribing* it into single stranded DNA (called cDNA, for coding DNA), and it is then called reverse transcription qPCR, or RT-qPCR. This method is quite sensitive at detecting relative abundances, but it is not very high throughput (Mutz et al. 2013).

A more high throughput method came out in the mid 90's, where quantification was done on a microarray format with short oligonucleotides representing sequences from different genes were bound at specific coordinates of the array. The cDNA was tagged with fluorescent labels, and by measuring the strength of fluorescence in different spots one could estimate the initial abundance of RNA. One microarray approach, cDNA microarrays, used cDNA as probes on the microarray chip, and was used for differential expression by using two different colours of fluorophores and comparing the intensities (Schena et al. 1995). The cDNA arrays were usually customized by research labs for their specific application, causing large variability between array experiments (Woo et al. 2004). Oligonucleotide arrays on the other hand use short oligonucleotides (~25nt) that are synthesized in place on the array, with several oligos representing each gene (Allison et al. 2006). This approach was commercialized by Affymetrix, creating an array with less array-to-array variability and can thus be used to compare expression levels between samples although only one sample is hybridized to each chip (Woo et al. 2004). The early versions of oligonucleotide arrays were limited to known annotated genes, and usually only detected the 3′ of mRNA. However, the complexity of arrays have evolved over the years and now includes possibilities to detect expression across the whole transcript using exon arrays (Kapur et al. 2007) and even un-annotated genes using tiling arrays with oligos covering the whole genome (C. Lee & Roy 2004; Moore & Silver 2008; Bertone et al. 2005). The latest versions of transcriptome arrays also contain oligos spanning splice junctions (fact sheet for Affymetrix GeneChip® Human Transcriptome Array 2.0).

During the first generation sequencing era, sequencing the whole RNA population of cells for quantification measures was not feasible. There was a method called Serial Analysis of Gene Expression (SAGE) for extracting short fragments (around 10bp) from the 3′ end of cDNA, ligating them together and quantify by sequencing. However, this method was very limited by the short read length. With the advent of 2nd generation sequencing technology, whole transcriptome sequencing became a viable option for gene expression studies, and in 2007 and 2008 the first whole transcriptome sequencing studies were emerging using the Illumina or 454 pyrosequencing technology (Weber et al. 2007; Torres et al. 2008; Sugarbaker et al. 2008; Marioni et al. 2008; Morin et al. 2008).

The greatest advantage that RNA-sequencing has over microarrays is the ability to produce reads across the full length of transcripts, and allow for detection of previously unknown genes or splice forms. Although tiling arrays can also be used for this purpose, they might have problems with achieving full genome coverage and still avoiding cross-hybridization artefacts (Malone & Oliver 2011). Microarrays can also be sensitive to single nucleotide polymorphisms (SNPs) i.e. differences between the reference genome and the one used for expression studies. A mismatch between the oligonucleotide and the gene sequence can weaken the hybridization to the chip, and give a misleading low expression value (Hutchison 2007). Using RNA-sequencing there will be no loss in signal due to SNPs, but possibilities of SNPs will instead need to be accounted for in the alignment strategy to identify RNA-seq reads. Also, RNA-sequencing allows for studies of transcriptomes of organisms without an annotated genome.

There is a problem faced by whole transcriptome RNA-sequencing that is not encountered by microarrays or genome sequencing – that of the highly abundant ribosomal RNA. Ribosomal RNAs are highly expressed, especially in rapidly dividing cells, and can make up around 80% of the total RNA content of a cell (Lodish et al. 2000). Since the expression levels of rRNA is usually not of much interest, strategies needs to be employed to avoid rRNAs taking up most of the sequencing power, and leaving very poor resolution for the rest of the transcriptome. To avoid this, two different approaches are commonly used. Ribosomal RNAs are not polyadenylated, i.e. they lack the tail of adenines at its end, while mRNA and many lncRNAs are. This fact can be exploited to “fish” out only the polyA+ fraction of the transcriptome using complementary oligo dT primers as bait, and discarding the rest. Another approach is to use oligos matching ribosomal sequence to bind and remove the rRNA. Both these methods are employed after RNA purification from the cells, and usually use oligos bound to small magnetic beads. By using a magnet to capture the beads, the bound RNA can be separated from the unbound RNA that is still in solution (O'Neil et al. 2013).

The library preparation protocol for RNA-sequencing is similar to that of DNA sequencing, except for the initial rRNA removal or polyA+ enrichment step mentioned above, and the fact that the RNA needs to be reverse transcribed into cDNA. However, the sequencing machines were designed with genome sequencing and genome assembly in mind, where a coverage of 30x billions of nucleotides is a desirable output. For RNA-sequencing, the number of reads produced by a unit in the sequencing machines, such as a lane in an Illumina flow cell, is usually vastly more than what is needed for gene expression measurements. To allow for higher efficiency and reduced cost of RNA-sequencing, it is therefore common to multiplex by giving each sample a unique nucleotide barcode that will be present in every fragment of the library, and let several samples run on the same lane. The DNA barcode is also sequenced, and is subsequently used to computationally assign the sequenced read to the sample of origin.

2.5 SINGLE CELL RNA-SEQUENCING

A classical RNA-sequencing protocol requires fairly high amount of input RNA, 100 nanograms in the case of Illumina sequencing. To obtain this mass of RNA, tens of thousands of cells need to be used for input. Due to this constraint, RNA-seq has traditionally been performed on bulk tissue samples, or on cells grown in cell culture to increase the volumes. The problem with this approach is that the gene expression obtained will be an average over the cells in the input sample, and in both tissue and cell culture the cell population is often not homogeneous. Therefore, this approach risks drowning out signals from rare cell populations, or creating false co-expression patterns between genes that are actually expressed in different cell types. In some cases it is possible to sort different cell types from each other to get a more homogeneous sample, but this requires that each cell type one wants to capture has known genes expressed that distinguishes it from the others and that it can be sorted by, which is not always possible.

In the recent years, several new methods have been developed to circumvent this problem and enable sequencing from single cells. Since single cells contain so little RNA, the development of single cell sequencing has been a quest of assuring as little loss of RNA as possible in the library preparation, and robustly amplifying the cDNA to a level where it can be detected with the 2nd generation technology. To prevent loss of RNA in the library preparation, the single cell methods aim to perform the reaction in the same test tubes as much as possible, since there is a risk of loosing some RNA to the test tube or the pipette with every transfer. For this reason, the RNA purification step is omitted as well as the polyA enrichment and ribosomal depletion, and the RNA is directly reverse transcribed from the cell lysate. To avoid sequencing rRNA, PCR primers containing an oligo dT sequence is used to amplify only polyadenylated RNA. Unfortunately, this means that all non-polyadenylated RNAs are lost, and there is not yet any single cell method that can capture these RNAs.

To achieve an RNA amount high enough for sequencing, different methods uses different approaches for amplifying the library. The most common approach is to use PCR, which is straight forward, fast and can be used on low input material. Amplification with PCR can bias the sample for RNA with higher content of G and C nucleotides and shorter transcript lengths, but this effect can be minimized by choosing an appropriate enzyme for the reaction (Dabney & Meyer 2012). However, PCR amplification is exponential since it can use the newly synthesized strand as a template for further amplification. Another method is *in vitro* transcription, where RNA is transcribed from cDNA. This method has the advantage that the amplification is linear, but the problem is that it is slow and requires a large amount of starting material. CEL-seq is based on in vitro transcription and solves the input problem by barcoding the samples at the 3´ end before pooling them and amplifying them together (Hashimshony et al. 2012). Since the barcode is added to the full-length transcript before amplification and fragmentation, it is only available in the 3´ most fragments, so this method is highly 3´ biased.

The single tagged reverse transcription (STRT) method also adds the sample barcode at the reverse transcriptase step but to the 5´ end instead of the 3´ end, and will thus yield a 5´ biased library. An advantage of adding the barcode this early is that the remaining reactions can be performed on a pool of multiple samples, which makes it easier to produce many libraries simultaneously (Islam et al. 2012). Also, STRT has been coupled with the use of *unique molecular barcodes* (UMIs), which essentially means that each primer will contain a random 5 bases long sequence, *5-mer*, in addition to the sample specific barcode. Since it is unlikely that two reads originating from two different molecules of the same mRNA will by chance have the same UMI, the UMIs can be used to estimate an original molecule count and compensate for PCR amplification bias (Islam et al. 2014).

The disadvantage of STRT and CEL-seq is that they are limited to either 5′ or 3′ coverage. They also have rather low sensitivity – according to a recent study STRT captures only 12.8% of the expressed genes (Macosko et al. 2015). A method published from our lab, Smart-seq2, enables whole length mRNA sequencing by amplifying and fragmenting the cDNA before addition of barcodes. In this way, fragments along the whole length of transcripts are sequenced. Having the sequence of full-length transcripts is valuable, since it enables detection of different splice variants, and also possibilities to detect SNPs and mutations. This method also has greater sensitivity, achieving a gene detection rate around 50% (Picelli et al. 2013; Picelli et al. 2014).

Single-cell sequencing has led to many new discoveries that was not possible using bulk RNA sequencing, such as cellular heterogeneity within tumours which might be important for treatment strategies (Patel et al. 2014) and discovery of a rare cell type in the intestine (Grün et al. 2015).

2.6 SPATIALLY RESOLVED SEQUENCING

The next challenge in sequencing is to enable spatial resolution of the information, i.e. to determine the location of RNA molecules within a tissue or even within a cell. This field is currently developing fast, event though it has yet to reach the throughput of conventional sequencing methods. The current approaches for spatial RNA measurements are either imaging based or sequencing based. The imaging based methods use fluorescently labelled DNA probes complementary to a target sequence. The challenges with imaging is that a single fluorescent probe bound to a single RNA molecule does not give enough signal to allow for detection, and that multiplexing is limited to the number of fluorophores available with non-overlapping wavelength spectra. A way to amplify the signal of an RNA is to hybridize multiple probes to each RNA molecule as in single molecule *in situ* hybridization (smFISH) (Femino et al. 1998), or even hybridizing a scaffold probe that itself allows for hybridization of several labelled probes as in branched DNA fluorescence *in situ* hybridization (bDNA FISH) (Battich et al. 2013). Another approach is to first reverse transcribe the RNA, and then use a so called pad-lock probe (Nilsson et al. 1994), which will become circularized when binding to the right target and can then be amplified by rolling circle amplification into a long chain of copies that can be detected by labelled probes. To enable greater multiplexing, smFISH can utilize multiple fluorophores for one RNA molecule, either by looking for a specific signature of colour combinations, or by performing several sequential hybridizations with specific colour signatures for one RNA in each cycle (Lubeck & Cai 2012; Lubeck et al. 2014).

Sequencing based spatial transcriptomics can be achieved by capturing RNA from individual regions or cells while recording their original position in the tissue, create a sequencing library per region and perform conventional second generation sequencing on the samples. The transcriptome profiles can then be mapped back to create a spatial map of expression. In **paper IV** we utilize laser capture microdissection to retrieve RNA profiles from specific positions in mouse brain. Another method for retrieving spatially resolved samples, microtomy sequencing, is based on sectioning replicates along different axis, sequencing whole sections, and then create a 3D map of the tissue (Combs & Eisen 2013; Junker et al. 2014; Okamura-Oho et al. 2012). Transcriptome *in vivo* analysis (TIVA) uses an approach were mRNA capture is photoactivated in a controlled way in an individual cell (Lovatt et al. 2014).

Another approach is to actually perform the sequencing *in situ,* in the tissue itself, as done in *in situ* RNA-seq (Ke et al. 2013) and fluorescent *in situ* RNA sequencing (FISSEQ) (J. H. Lee et al. 2014). These methods involve a rolling circle amplification step to get sufficient template for detection, and then employs a sequencing-by-ligation approach similar to SOLiD sequencing (Crosetto et al. 2015).

3 ANALYSIS OF RNA-SEQUENCING DATA

3.1 WHAT IS EXPRESSED?

Today, using the 2nd generation sequencing methods, the sequencing itself is the fastest and easiest step of the workflow. Performing RNA sequencing, or other types of sequencing, will generate large text files with millions of short sequence reads, and you are then left with the challenge of identifying the expressed genes, quantifying the expression, and finding meaningful information among the thousands of expressed genes.

The first step of analysis is finding out where each read came from, which is traditionally done by aligning the reads to the reference genome, if there is one available. There are several different sequence aligners that have been developed for the purpose of short read alignment. However, some of these aligners were built as an extension of aligners designed to align DNA to a genome, and as a result were not very efficient in mapping spliced RNA-seq reads (Trapnell et al. 2009). The aligner STAR, which stands for Spliced Transcript Alignment to a Reference, was developed to improve both speed and accuracy of alignment, especially for spliced reads (Dobin et al. 2013).This is the most tested and efficient RNA-seq mapper to date, which we have chosen as basis for our RNA-seq alignments.

3.2 HOW MUCH IS IT EXPRESSED?

When the reads have been mapped to the genome, the next step is to determine the expression level of the expressed genes. This problem turns out to be more complex that it sounds. First of all, we have to decide what we want to estimate with our gene expression calculations – are we looking for the absolute number of molecules, or is it sufficient to estimate the relative proportions of RNA within the sample? Each cell will have a certain number of RNA molecules at the point in time when we choose to sequence it, which are produced from a subset of genes, and the number of RNA molecules from each gene will be different. The rationale behind studying RNA expression in the first place is the assumption that the level of expression of a certain gene tells us something about how important the gene product is for the cell at this time. So, ideally, we might want to know the exact number of RNA molecules of each type that were present in this cell. At least when performing single cell sequencing this is desirable, while for bulk sequencing it has less meaning since even if we knew the exact number of cells in the sample, the measure would still be an average across cells.

However, achieving an absolute molecule count is not trivial. As I discussed in the single cell sequencing section, the STRT method (Islam et al. 2012; Islam et al. 2014) attempts to do this by using unique molecular barcodes (UMIs) to weed out PCR duplicates from actual copies of transcripts. If the method had a perfect sensitivity, this would work, but since it only captures 15% of all reads the sampling will create a lot of noise that is hard to control for, and lowly expressed genes can be lost.

An alternative metric to estimate is the proportional abundance of each transcript in the cell, i.e. the percentage of the cells RNA molecules that comes from each gene. A measure of the proportional abundance is more attainable, since the unintentional subsampling of RNA that occurs due to loss of RNA in sequencing protocols should still conserve the initial proportion between molecules, at least for the medium and high expressed genes that do not risk getting lost altogether in the subsampling. The true proportional abundance would thus be the number of molecules of a transcript, divided by the total number RNA molecules in the cell.

To estimate this abundance, we have to remember that in case of full-length RNA sequencing, the RNA is fragmented into shorter pieces and a long RNA molecule will therefore yield more reads than a short one. The raw read counts for a gene will therefore need to be normalized for the length of the transcript. The next issue is that of read depth. Since the sequencing library is amplified by PCR, the number of reads we sequence is not related to the original number of molecules but rather arbitrary and depends on experimental set up and technical variation. This means that the read depth also has to be accounted for when estimating transcript abundance.

A commonly used metric for RNA abundance is RPKM (Reads Per Kilobase per Million mapped reads), suggested by Mortazavi et al. (Mortazavi et al. 2008). It is calculated as follows:

$$RPKM_g = \frac{C_g \times 10^9}{L_g \times N}$$

where C_g is the read count in the gene, L_g is the length of the gene, and N is the total read count in the sample, thus normalizing for both transcript length and read depth. A similar version of this is FPKM, Fragments Per Kilobase and Million mapped reads. The use of "fragments" instead of reads allows the use of paired end data, where one fragment is represented by one or two read "mates" depending on if they were both mappable to the genome. However, the N that is used for normalization in these methods does not accurately reflect the total transcript abundance, but is confounded by the length of the expressed transcripts. Another metric, Transcripts Per Million (TPM), which takes this into account, is now gaining in popularity. For TPM, the count of each transcript is first divided by the transcript length, just as for RPKM. These length-normalized values are then summarized for all genes to generate a better estimator of total transcript abundance in the sample, and the normalized value for each gene is divided by this factor. It is calculated as follows:

$$TPM_g = \frac{\frac{C_g}{L_g} \times 10^6}{\sum_{i \in G} \frac{C_i}{L_i}}$$

where C and L are as before the counts and length for the genes, and G is the set of genes analysed. TPM is a better estimator for the true relative abundance of a transcript, and also has the advantage that it gives a more intuitive comparison between samples since the sum of all TPMs will always be 1 million, and as long as the same genes are studied across samples it will also give the same average TPM per sample (B. Li et al. 2010; Wagner et al. 2012).

Another challenge with calculating expression levels is that even the C and L in these equations are not absolute, due to the different splice forms of a gene. Since the same gene can be expressed including and excluding different exons and terminating at different sites, and often exists in several splice forms in the same cell, the length will depend on which exons are actually expressed, and to what level. For the same reason, the counts are not straight forward either, since a read mapping to a specific place in the genome can still belong to several different isoforms of the gene. We do not want these type of reads to be counted multiple times (once for each isoform), so computational strategies need to be employed to determine which read to assign to which isoform, and which exons to use for length normalization. Also, when using short reads there might be regions in genes which are not uniquely mappable, which could be adjusted for either by eliminating multimapping reads and reducing the effective transcript length by the unmappable region or adjusting the count value to a computationally estimated value. I will discuss this further in the results for **paper I**.

3.3 WHAT IF THERE IS NO SEQUENCED GENOME?

If the genome sequence is not known, but there is an available set of transcript sequences, these can be used for expression quantification. RSEM (B. Li & Dewey 2011), Sailfish (Patro et al. 2014), Salmon (Patro et al. 2015) and Kallisto (Bray et al. 2015) are specifically designed to use transcriptome sequence as a basis. When mapping to a transcriptome, multimapping is an even bigger issue than in genomic mapping, since isoforms of the same gene will inevitably contain some overlapping regions. These programs determine the likelihood of a read coming from a certain transcript, and estimate the expression value of transcripts based on this. Salmon and Kallisto use k-mer based approaches to speed up the calculations, meaning they find the possible origins of shorter subsequences (length k) of the read and combines the k-mer information to calculate the probability of origins. Sailfish uses a lightweight alignment approach to find a chain of super-maximal exact matches (SMEMs) or maximal exact matches (MEMs) for a read, meaning there is no pre-defined k-mer length, instead the algorithm finds the maximal matches, i.e. a match such that adding another base to the k-mer will cause a mismatch (Patro et al. 2015; Bray et al. 2015; Patro et al. 2014; B. Li & Dewey 2011).

If the available transcriptome is inadequate, another option is to assemble a transcriptome from RNA-seq data. Creating a transcriptome from polyA+ enriched RNA-seq data requires much less sequencing depth and computational power than assembling a whole genome, since only the spliced coding sequence is present. It presents a different challenge though in resolving the sequence of different splice forms, and unlike genome data the sequencing depth of individual genes will vary depending on expression level. Transcriptome aligners have been developed specifically for this purpose, and in **paper III** we assemble a transcriptome from paired-end RNA-seq data from *Xenopus laevis* early development using Trinity (Grabherr et al. 2011).

3.4 WHAT IS THE DIFFERENCE IN EXPRESSION?

The goal of gene expression analysis is often to determine the differences in expression between different conditions, for example comparing healthy to disease. To enable comparison one first has to ensure that the two libraries are scaled properly to be comparable. As described above, the read abundance for a transcript depends on the length of the transcript, the sequencing depth and the general composition and quantity of RNA molecules in the sample. The sampling of a transcript depends not only on its own properties, but also on the abundance of other transcripts, and expression changes in highly expressed genes can change the sampling of lowly expressed genes. Generally, RPKM or TPM values are not used in differential expression analysis, since transcript length does not matter when the comparison is made for the same gene between samples. Instead, differential expression packages use other normalization strategies to account for library depth and composition. Here I will discuss three common programs that I have some experience with, DESeq, edgeR and Cuffdiff.

DESeq calculates the read count ratio for each gene compared to the total read count, and, assuming most genes are not differentially expressed, they use the median-of-ratios as a scaling factor (Anders et al. 2012; Love et al. 2014). edgeR instead uses weighted trimmed means of M values (TMM), where the M-values are log ratios of library size normalized expression values between the samples. The genes with highest M-values and highest average read counts are trimmed away before calculating a weighted average of M values, which is then used as a normalization factor (Robinson & Smyth 2008; Robinson et al. 2010). Cuffdiff uses a similar method to DESeq, but performs a scaling first within conditions and then between conditions (Rapaport et al. 2013).

After normalization, the task is to perform a statistical test to determine for each gene if the expression difference is significant or not. Both DESeq and edgeR assumes that the data follows a negative binomial distribution, where the variance ν is defined as $\nu = \mu + \alpha\mu^2$, where μ is the mean and α is the dispersion factor. Since the sample size is usually small (typically 3-5 replicates) in RNA-seq experiments, estimating a gene-wise dispersion value is quite unreliable. Therefore, edgeR uses either a common dispersion value that is the same for all genes, or a moderate tag-based dispersion, where the dispersion is calculated per gene and then “squeezed” toward the common dispersion value (Robinson & Smyth 2007). DESeq on the other hand calculate a dispersion estimate across genes with similar expression values. In DESeq2, this method is replaced by calculating gene-wise maximum likelihood estimations (MLE) of dispersions, fitting a curve to these values, and shrinking the gene-wise dispersions estimates toward the line (Love et al. 2014). Cuffdiff calculates the variance similar to DESeq for single isoform genes, while a mixture model of negative binomial is used for multi-isoform genes. DESeq and edgeR use a Fisher exact test adapted for negative binominal distributions to test for differential expression, while Cuffdiff uses a t-test.

Since the differential expression tests are done once per gene, i.e. thousands of times, using the p-value for significance would by chance give us a high number of false positives. Therefore, a correction needs to be employed to account for the multiple hypothesis testing. The methods mentioned above includes correction methods such as Benjamini-Hochberg correction to control this so called type I error (Rapaport et al. 2013).

Single-cell sequencing is often used for an unbiased analysis and *de novo* cell type discovery in tissue. The nature of this type of data is that there are no known biological conditions assigned to each sample, and the identity must instead be derived from the RNA-seq data. Recently, Brennecke et al. presented a method to enable discovery of differential expression patterns without prior knowledge of cell identities. It is based on modelling the technical and biological variation, and reporting the genes with a variance exceeding the expected biological variation (Brennecke et al. 2013). This method has proven very useful when *de novo* classifying cells, as these highly variable genes can subsequently be used for clustering the single cell samples and discovering subgroups within the population.

3.5 VISUALIZING DIFFERENCES

Due to the many variables (i.e. genes) describing each data point (sample or cell), visualizing the differences requires specialized techniques. One common way of visualizing differences between groups of cells is principal component analysis (PCA). It is based on reducing the dimensions of the multivariate dataset to 2 or 3 dimensions, which can easily be visualized. The basic principle of PCA is to find the direction of greatest variability within the multidimensional data, and using this direction as the first "component" to be plotted on the first axis. The next component is found as the most variable directions perpendicular to the first, and the next perpendicular to both, and so on. In addition to PCA, there are other dimensionality reduction methods specifically designed for visualization of omics data. One of them is t-distributed stochastic neighbour embedding (t-SNE), which is a non-linear dimensionality reduction creating a visualization better able to distinguish between sample clusters (Van der Maaten & Hinton 2008). Another method, monocle, provides a dimensionality reduction specialized on visualizing an axis of differentiation. It finds the longest path through a minimum spanning tree in the data set and thereby creates a trajectory of pseudotime in the data (Trapnell et al. 2014).

3.6 WHICH ISOFORM IS EXPRESSED?

The splicing pattern of a gene can make a difference between creating a functional protein or a non-functional one, or even between the RNA being translated or degraded. Thus, splicing patterns are important to study to fully understand the relationship between what is transcribed and how the cell functions. To determine which exons are excluded or included in a transcript, special types of reads are important to study:

1. The splice junction reads that map across two exons, providing information about which exons are physically connected.
2. The reads mapping across exon-intron borders, signalling that an intron is included.

These types of reads are essential to determine how exons are connected to each other. Therefore it is important that the aligner used performs well with spliced reads. However, the reads mapping within exons are also important for estimating the levels of inclusion/exclusion of specific exons in comparison to the surroundings. Splicing analysis can be done with an exon centric, isoform centric or splicing event centric approach. Isoform centric refers to estimating the levels of known annotated isoforms of a gene, while the exon centric approach focuses expression levels of each exon separately, and can therefore detect novel splice forms. Event centric methods focuses on known splicing events, e.g. detecting inclusion levels of an exon known to be skipped in some isoforms. There is a multitude of splicing software programs, but not all of them are designed to detect splicing differences between conditions, a feature that is useful to understand splicing effects of disease states or cell type differences. Some commonly used software with the ability to detect differential splicing are Cufflinks/Cuffdiff, DEXSeq and MISO. DEXSeq (Anders et al. 2012) use an exon centric approach, which estimates the reads mapping to "counting bins" consisting of whole or part of exons (in case an exon has alternative splice sites, it is split into several counting bins). A generalized linear model is then calculated for each gene, including factors representing baseline expression, as well as factors representing the effect of the sample condition on both gene expression and coverage of the counting bin. DEXSeq is based on a similar model as DESeq, assuming a negative binomial distribution and calculating dispersion estimates by sharing information between similarly "expressed" counting bins. Differential exon usage is tested with the null hypothesis that none of the conditions affects exon usage, and one test is performed per counting bin. This method provides information of which exons are excluded or included, but cannot tell anything about how the exons are connected or the specific splice forms present in the sample. Junction-mapping reads are not used to connect exons, instead they are split between their corresponding counting bins.

MISO (Katz et al. 2010) on the other hand uses both an isoform centric and event centric approach. In the event centric approach, each event represents two alternative outcomes of splicing, such as exon exclusion/inclusion, intron exclusion/inclusion, mutually exclusive exons and differential 3´ or 5´ splice sites. One of the alternative outcomes will be denoted as the "exclusion" event, and the other as the "inclusion" event. The number of reads exclusively supporting either the exclusion or inclusion event, such as junction reads or reads falling within an alternative exon, are used to estimate an inclusion level of the specific event. The advantage of this method is that it provides information on which splicing donors and acceptors are connected. However, the downside is that it uses a Bayesian statistical approach for alternative splicing measurements without providing a means to use biological replicates. Another limitation is that the comparison of events is always binary, i.e. there is no room for comparing for example 3 different alternative splice sites in the same exon simultaneously.

Cuffdiff estimates isoform level differential expression by default, and can thus be used for both differential expression and differential splicing analysis. Isoform level expression is calculated by using both reads exclusive to an isoform and those common between isoforms to calculate create a probability score of the expression level for each of the possible isoforms (Trapnell et al. 2010; Trapnell et al. 2013).

4 AIMS

The over all aim of my thesis work was to increase the understanding of regulatory principles at the RNA level in development and disease, and to improve upon the methodologies used for analysis of RNA sequencing.

The specific aims for each study were:

Paper I:

- To provide an efficient method for normalizing for lack of mappability in genes, flexible enough to allow to work across the span of commonly used read length.
- Analysing the mappability of different genomic regions, to guide appropriate read length choices for different sequencing applications.

Paper II:

- Decipher the roles of the exosome complex, NEXT complex and cap-binding complex in degradation of PROMPTs and other assumingly non-functional transcripts

Paper III:

- Map the sorting patterns of maternal RNA in the early *Xenopus laevis* development and finding motifs in the RNA sequence that signals for sorting to specific locations

Paper IV:

- Deciphering the protective effect in somatic motor neurons resistant to degradation in Spinal Muscular Atrophy (SMA).
- Gaining knowledge on the mechanism behind SMA progression by studying both expression and splicing effects of the disease.

5 RESULTS AND DISCUSSION

5.1 PAPER I: MINIMUM UNIQUE LENGTH ANALYSES WITH MULTO

Typical RNA sequencing experiments uses read lengths of 25-150 basepairs, and the length used for a particular experiment can depend on several factors. First, the choice of sequencing technology will dictate what lengths are possible, and usually there is a standard length that is commonly used with that technology. Second, choosing to multiplex your sample by adding barcode indexes will also have an effect, since some of the reagents that would have been used for sequencing your input fragments will instead be used for reading index sequence. And finally, it is a question about cost versus benefit of sequencing longer reads, since longer reads will be more expensive.

Since our genomes consists of a code of only 4 different base pairs and 3.3 billion bases, it is impossible for all short segments of the genome to be unique. Each substring, *k-mer,* of sequence, with k being the length of the sequence, can occur in 4^k different combinations. Theoretically, this could mean that a 16 base pair sequence would be sufficient to create a genome unique at every position (4^16 ≈ 4.3 billion bases) However, this is not the case, because of several different effects creating biases in the genome; most genomes have a bias for more G-C or more A-T content, and there are duplicated genes and repetitive regions which all further increases the chances of the same k-mer occurring multiple times. This means that sequencing short reads will pose a challenge when trying to identify the gene of origin for some reads.

I was working on a project where the lack of mappability of some reads could potentially cause a very problematic bias in the analysis, because I was looking at read coverage at small regions overlapping the exon-intron junction. We had sequenced RNA from nuclear and cytoplasmic RNA separately, and were aiming to study splicing dynamics using the nuclear fractions were intronic sequence could still be present in pre-mRNAs. To elucidate if the splicing occurred co-transcriptionally, we studied read coverage at regions overlapping the splice-junction (i.e. were reads went across the exon/intron borders) of both the 3´ and 5´ end of introns, and in nearby intronic regions of the same size (Fig 7). When looking at such limited regions, and trying to compare the expression between regions of the same gene, it was crucial that the reads within them were correctly mapped and that un-mappable regions were not considered as empty in the analysis.

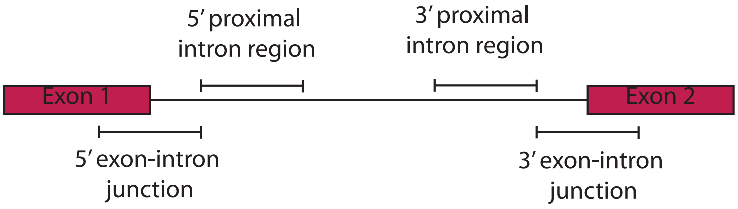

Figure 7. Short regions overlapping exon-intron junctions and intronic regions of the same size were used to study splicing kinetics.

Of course, this is also an important issue when estimating expression levels on whole genes from RNA-seq data with short reads. Different software for gene expression quantification has different approaches to solving this problem. ERANGE and Cufflinks both use computational methods to predict the origin of multimapping reads from all its possible positions. ERANGE bases the prediction on the amount of uniquely mapping reads to the same genes (Mortazavi et al. 2008). The problem with this methods is that it assigns reads to a region based on the mappability of other reads, and when the number of mappable positions in that region is low to start with, this kind of prediction does not help much. Cufflinks employs a similar method, but utilizes both uniquely mapping and multimapping reads to create a probability score of the expression level for each gene (Trapnell et al. 2010). Another way to solve the problem is to only normalize by the number of mappable positions in the gene (E. T. Wang et al. 2008; Pan et al. 2008; S. Lee et al. 2011). However, as stated above, the number of mappable positions will vary depending on the read length used. In my case there were available files containing mappability information for a few specific read lengths, but the length I was using was not represented, and I did not want to use an approximation in case it affected the data.

At this time it was not really known how much the non-unique reads affected the outcome of gene expression analysis. I set out to create a tool for producing uniqueness information across all useful read lengths that could subsequently be used for normalizing RNA-seq data, both for whole genes and smaller regions, and to study how the mappability of short reads affected RNA-seq data analysis.

In **paper I**, we describe the software MULTo (Minimum Unique Length Tool) that I created to improve uniqueness normalization and analysis. Essentially, the program makes use of the range of numerical values a single byte in a binary file can take on, i.e. 0 to 255. MULTo will go through every position in a genome and iteratively analyse k-mers of different length until the shortest unique k-mer is found, and store this number k as a byte in a binary file. One file per chromosome will be created, and the position of the byte in the file translates to the starting position of the k-mer in the chromosome. In this way, we could efficiently store detailed mappability information across a wide range of possible read lengths in a compact format. The MULTo suite also contains a program for querying specific regions for uniqueness, either at a range of read lengths or at one particular read length. The program was also incorporated into our in house expression estimation software, rpkmforgenes (Ramsköld et al. 2009), to enable automatic normalization of only the unique length of genes.

When analysing the uniqueness information across different read lengths, we found a dramatic increase in mappability as read lengths for single read sequencing were increased from less than 70% of genes having 90% unique positions at 20bp to 83% having 90% unique positions at 50bp. Increasing the read length to 100 only increased this figure by a few percentage points, while using paired end sequencing gave a much better over all mappability regardless of length. We could also show that different genomic regions had different degree of mappability, with enhancers and promoter regions being among the most unique, while intergenic regions (without identified genes) showed the least uniqueness. Together, this information can serve as a useful guide when choosing read length for different sequencing applications.

When comparing my normalization method to ERANGE and Cufflinks, we found that both ERANGE and Cufflinks produced a less robust compensation than ours for non-unique positions when there was a large proportion of the gene that was not mappable. We could also identify a problem with Cufflinks overestimating the expression of short genes.

Taken together, I have in this project provided both a useful tool for uniqueness analysis and uniqueness compensation in sequencing data, as well as an in depth analysis of uniqueness that can be used to guide choices on read length for different sequencing approaches.

5.2 PAPER II: THE EXOSOME COMPLEX AND RNA DEGRADATION

RNA degradation is an important part of gene regulation at many levels. The lifetime of an mRNAs will determine how long the protein is produced, and how quick the cell can change the protein level in response to extrinsic or intrinsic signals. Degradation is also important for removal of intron sequence, and for quality control by removing or processing defective RNAs. There are several different RNA degradation machineries, some degrading from the 3′ or 5′ end (exonucleases), others that cut RNA internally (endonucleases) (Houseley & Tollervey 2009). The exosome complex is an important degradation machinery containing 3′-5′ exonucleases. It is present in both the nucleus and cytoplasm, and targets several different RNA species for degradation, including degradation of mRNA in the cytoplasm (Raijmakers et al. 2004).

The promoter upstream transcripts (PROMPTs) are a species of RNA thought to be non-functional. Only a few instances has been shown where PROMPTs are associated with gene regulation (Lloret-Llinares et al. 2015), though causality has not been proven. They are fairly quickly degraded by the exosome complex, suggesting that the reason for expression might be simply because the transcription machinery has been recruited and is not specific enough to transcribe only the gene. In humans, the trimeric NEXT complex, consisting of hMTR4, RBM7 and ZCCHC8, has been shown to aid in exosomal degradation of PROMPTs. The NEXT complex had in turn been shown to associate with the arsenic resistance protein 2 (ARS2) and the cap binding complex (CBC) consisting of cap-binding proteins CBP20 and CBP80, which we together denote the CBCA complex.

In **paper II**, we examine the physical interaction between the exosome, the NEXT complex and the CBCA complex by affinity capture mass spectrometry (ACMS), study their RNA binding profile by RNA immunoprecipitation (RIP) and elucidate the function of the subunits by studying effects of single or combinatorial knockdowns of the constituent proteins and analysing the transcriptional effects.

At the time when I got involved in this project, Jensen lab had already performed the ACMS and RIP analysis. The ACMS analysis proved a physical interaction between CBCA and NEXT complexes together with the zinc-finger protein ZC3H18 (this combined complex is hereafter called CBCN) and of CBCN to the exosome, and by RIP they had found that CBCN bound specifically to PROMPTs and U1 snRNA. To assay the function of each constituent protein, a single and combinatorial knockdown approach using siRNA was employed. Using qPCR for known PROMPTs and snRNAs, a synergistic effect on PROMPT accumulation between some of the combinatorial knockdowns compared to the single knockdowns could be detected, and through chromatin immunoprecipitation (ChIP) assay of RNA polymerase II in these regions they found evidence of a read-through-transcription effect.

They now wanted to test the whole transcriptome effects of the knockdowns, and performed ribosome-depleted, strand specific RNA-seq on these samples. My contribution to this product was the analysis of this RNA-seq dataset, by building a dynamic program to analyse the expression profile in different regions for the different knockdown samples. This proved to be an interesting task, with quite specific challenges to solve.

First of all we needed to determine which regions to look for PROMPT expression. This might seem straightforward – as can be understood by the name this transcripts are located upstream of promoters of expressed genes. The problem is that many genes have several alternative promoters and alternative transcription start sites. By choosing a site downstream of the true active promoter would mean that the "PROMPT" region we are looking at overlaps with the expressed gene itself. Choosing one too far upstream could mean that we are actually missing the first part of the PROMPT. Another challenge was cases when upstream genes were close enough to the promoter to fall into the presumptive PROMPT region.

I solved both problems by simply filtering out transcription start sites (TSS) with annotated exons within the upstream 5000bp. This would remove both those with upstream TSS of the same gene, or overlap with other genes. This means that we always used the furthest upstream TSS. To determine if this TSS was actually active, we used transcript level RPKM calculations and filtered for the TSS that had an accompanying transcript that reached an expression threshold (RPKM >= 4 was used in the paper).

We wanted to get a coverage profile around the transcription start sites that could be comparable across samples. Since our RNA-seq data was sequenced by paired-end sequencing, it was not straight forward which regions to count as covered. We could potentially have filled in the sequence between the two mates and counted that also to the coverage. However, we decided that the comparison would be fairer if we only used the actual reads to not bias for fragment length differences or artificially introduce coverage in introns. When the fragment was so short that the two mates overlapped, we only counted the overlapping region once.

To be able to compare the coverage profiles from different samples to each other we also had to normalize for read depth. This also proved to be non-trivial for reasons that were very specific to this dataset. By disturbing exosome function and impairing RNA degradation we actually affected the transcriptome profile as a whole. While mRNAs seemed unaffected, many non-coding RNAs were present at higher abundance in the knockdowns. Normally, we would normalize to all reads that mapped to any exon, but for this dataset this could not be done. Instead we normalized only to the reads mapped to mRNA exons to avoid biasing the values (Fig 8).

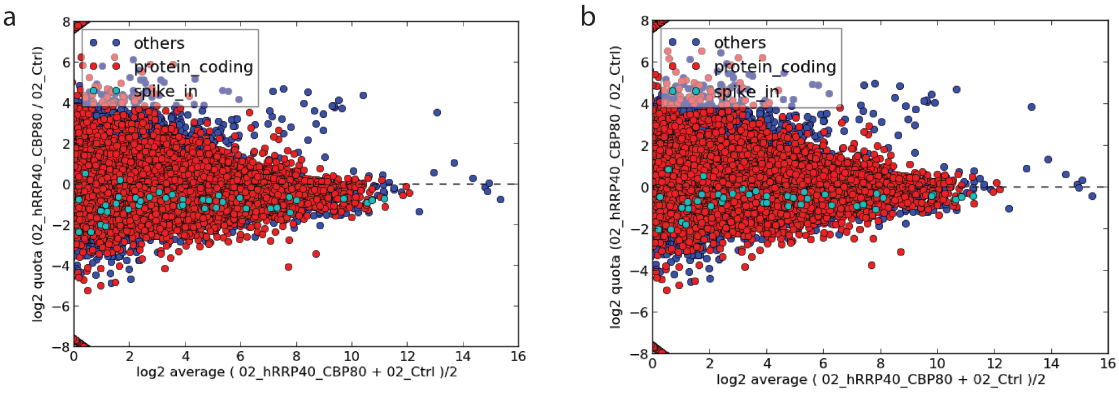

Figure 8. MA-plots of a double knockdown sample (hRRP40 + CBP80) vs control. (a) When normalizing to all mapped reads, the RPKM values are underestimated for protein-coding genes in the knock-down sample. (b) Normalizing by only reads mapped to mRNA makes the RPKM values more comparable between the samples.

In the end, the analysis resulted in the coverage plots that can be seen in figure 7 and supplementary figure 4 and 5 in paper II. We could show that the read-through-transcription profile was general for PROMPTs when knocking out the CBCA component CBP80 or ARS2, suggesting that CBCA has a role in transcription termination of PROMPTs. To get accurate read-through profiles for histones we filtered away histones with overlapping annotations downstream of the 3′ end.

Initially, the transcriptome project was intended to become a stand-alone publication. Therefore, we planned to do more in depth studies of the RNA-seq data and study how enhancer RNA and other interesting regions in the genome was affected by the perturbations of the exosome and its interaction partners. Therefore I implemented functions to also study bi-directional transcription and performing the appropriate filters and normalizations. The script also contains option for plotting different RNA biotypes separately (one could for example choose to study PROMPTs only at lncRNAs). However, the transcriptome story was in the end combined with the ACMS, RIP and ChIP data to create a more comprehensive study, and the remaining analysis is not yet used.

5.3 PAPER III: MRNA SORTING IN EARLY STAGE *X. LAEVIS* EMBRYOS

Although all vertebrates go through stages of embryonic development that are similar in many aspects, such as having the same three germ layers that further on develop into the same kinds of tissues across all organisms, the very early events after fertilization can be quite different. The African clawed frog *Xenopus laevis* has a clear polarization already in the oocyte, in which some specific RNAs and proteins are sorted to the animal pole or vegetal pole. There is evidence that many of these sorted RNAs are kept in their respective pole as the oocyte is fertilized and the embryo starts to divide (Freeman et al. 2008; Horvay et al. 2006; MacArthur et al. 2000). In the early stages there is no or little transcription from the embryonic genome, so lingering maternal RNAs constitute all the instructions to these cells up until the embryo has gone through around 12 divisions (Langley et al. 2014).

In this project we wanted to study the sorting of maternal RNA and if possible identify potential sorting signals at single cell resolution by performing single cell RNA sequencing on individual blastomeres at the 2-16 cell stages. Since there is no genome assembled for *Xenopus laevis*, we first assembled a transcriptome to use as basis for subsequent RNA abundance estimations.

To reconstruct the *X. laevis* transcriptome, we performed paired-end sequencing on pooled whole embryos from 3 different stages (1-cell stage, 512-cell stage and stage 10.5). I performed transcriptome assembly using both the genome guided and *de-novo* assembly functions in the Trinity suite (Grabherr et al. 2011) as well as cufflinks (Trapnell et al. 2010). Downstream of this analysis, I used “Program to Assemble Spliced Alignments” (PASA), which is designed to create comprehensive gene annotations from cDNA sequences (Haas et al. 2003). I combined the two Trinity transcriptomes with the cufflinks transcriptome and publicly available *X. laevis* cDNA sequences from UniGene when running PASA. This approach was especially helpful in creating longer transcript assemblies using the different inputs, however I found that when it came to grouping the transcripts into genes it often made mistakes, perhaps because of the fragmented state of the genome. When identifying the transcripts by aligning them to known annotations from different species using BLAST (Camacho et al. 2009), two transcripts from the same “PASA gene” would sometimes align to different annotated genes. I decided to primarily trust the BLAST alignments, and the assemblies with sufficient alignment scores to the same gene or a homologue were considered transcripts of the same gene. I translated the gene names of all BLAST hits to the official name of the *Xenopus laevis* or the closely related *Xenopus tropicalis* homologue to have a consistent annotation.

Xenopus laevis is pseudotetraploid, i.e. it has 4 copies of each chromosome originating from two different ancestral species. This constitutes an extra challenge when trying to assemble and annotate its genome and transcriptome. In this study, I chose to disregard the possibility of genes from different ancestral chromosomes having different sequences as to simplify the challenge of annotation. When using my assembled transcriptome as a basis for gene expression analysis for single cell RNA-seq data, I only used the transcripts that I could assign a known gene name with high confidence. My transcriptome reconstruction improved the length of most previously annotated genes, and added 1,711 genes that were not previously named in the UniGene database, which makes this carefully annotated transcriptome a useful resource for the *Xenopus laevis* research community.

To deepen our understanding of mRNA sorting in early embryonic development, we sequenced single cells from 2-16 cell stage that were each carefully annotated by their exact position in the embryo. Since the *Xenopus laevis* embryo has a clear dorsal and ventral side already from fertilization, we were expecting to see some differences in mRNA sorting along the dorsal ventral axis. There are also proteins known to localize to the dorsal side, like dishevelled (Dsh) and beta-catenin (Houston 2012; Miller et al. 1999). In our single cell RNA-seq data, we observed a slight preference for known vegetal RNAs to be more concentrated in the vegetal-dorsal side at 4 cell and 8 cell stages, but this pattern did not hold true in the 16 cell stage were we could find no significant difference at all between vegetal-ventral and vegetal-dorsal cells. This is in line with the results of recent single cell studies using qPCR in *Xenopus laevis* (Flachsova et al. 2013) and RNA-seq in *Xenopus tropicalis* (De Domenico et al. 2015), which could not see a difference at 8-16 cell stage. However, when using a more unbiased approach to find the top variable genes within embryos, we did find a group of genes that clustered more tightly together in subsets of vegetal pole cell, although this subset was not specifically localized to ventral or dorsal side in the embryos. Surprisingly, considering its lack of specific localization, the set contains RNA for dorsal determinants such as *Sybu* and *Wnt11*. The group is generally enriched in germ plasm RNAs, suggesting the tight clustering will subsequently lead to a high concentration of these RNA in the future germ cells.

We wanted to know more about the mechanisms behind RNA sorting in the frog embryo, and if there was a common sorting signal among the vegetal or animal pole RNAs. There are some well-known short motifs that appear in the 3´UTRs of known vegetal genes, such as the R1 motif (UGCAC) in *Nanos1* E2 (WYCAC) and VM1 (YYUCU) in *VegT* and *Gdf1* and these generally appear in clusters of multiple instances (W = A or C and Y = U or C). Betley et al. hypothesized that it was this clustering of short repeats that was important for localization, and with their own software REPFIND, they found that vegetally localized transcripts were enriched for CAC-containing repeats. In our study, we used both REPFIND and a short motif discovery program, Weeder (Pavesi et al. 2001), to search for motifs in the 3´UTRs of the vegetally enriched RNAs in our study. We could confirm both CAC-containing motifs and instances of the R1, E2 and VM1 in the vegetal RNA. However, not all vegetally sorted RNA had CAC-containing clusters, and not all RNAs with CAC-clusters were sorted to the vegetal pole. Therefore there are clearly additional sorting signals, which are not detectable with these methods. One possibility could be that it is the secondary structure rather than the sequence itself that determines the location. Most studies have found that RNA binding proteins require single stranded RNA for motif recognition (Ellis et al. 2007; Ray et al. 2013), however the secondary structure could serve another function by hiding motifs in double stranded sequence to prevent some RNAs from being sorted (X. Li et al. 2010). The secondary structure is hard to predict completely from whole transcripts, or even 3´UTRs, since the sequence is so long that there will be multiple ways it can base pair to form secondary structures. One possible approach, suggested by Li et al., is to instead take all possible secondary structures into account and calculate an overall predicted accessibility value at each segment of the sequence (X. Li et al. 2010). Another challenge is to solve the case of RNAs that are sorted but do not have the motif. One reason for the lack of consensus motifs could be that these RNAs require binding by several RNPs with different motifs, which are on their own not strong enough to be detected. Understanding the sorting mechanisms in such cases would probably require some more experimental work, such as performing a pull-down assay of the specific RNA and identify the proteins bound to the RNA with western blot or mass spectrometry (Marín-Béjar & Huarte 2015). Although the laborious methods in the latter approach is beyond the scope of this study, testing for motif accessibility due to secondary structure could provide a good complement to our study. We are also planning to validate the localization of some of the novel sorted genes we discovered by RNA FISH to strengthen our conclusions.

5.4 PAPER IV: TRANSCRIPTOME PROFILING OF SPINAL MUSCULAR ATROPHY (SMA)

Spinal Muscular Atrophy (SMA) is a motor neuron disease, characterized by a progressive loss of somatic motor neurons, and the most common genetic cause of infant death, with a prevalence of 1-6 out of 10,000 live births. It is a monogenic disease, caused by a loss or loss-of-function mutation of the Survival Motor Neuron 1 gene (*SMN1*). The protein product of *SMN1*, the SMN protein, is involved in assembly of the splicing machinery. In mouse models, it has been shown that a complete loss of SMN protein is embryonic lethal, but in humans a gene duplication has resulted in a paralogous gene, *SMN2*, with almost identical sequence to *SMN1*, which partly compensates for the loss of *SMN1*. The two genes differ by 5 nucleotides (Monani et al. 1999), one of which is a C to T nucleotide transition in exon 7. Even though this transition does not affect the amino acid sequence, it still has a substantial effect since it disrupts a splicing enhancer and causes alternative splicing of the gene (Cartegni & Krainer 2002; Kashima & Manley 2003). In *SMN2*, exon 7 is skipped in over 90% of the processed mRNAs, leading to the production of a truncated and unstable SMN protein (SMNΔ7) (Ruggiu et al. 2012). The disease is graded from I to IV, with grade I onset occurring before 6 months of age with a life expectancy of 2 years, while grade IV presents in adulthood with only mild muscle weakness symptoms (Munsat & Davies 1992). The severity of the disease is negatively correlated with the copy number of *SMN2* in the patient (McAndrew et al. 1997).

Although *SMN1* is ubiquitously expressed across all human tissues, the pathology is restricted to the motor system. Somatic motor neurons of the spinal cord are especially vulnerable to low SMN levels, and there is also evidence that satellite cells, a muscle progenitor, may contribute to the muscle pathology (Iascone et al. 2015). However, motor neurons are most likely driver of the disease, as muscle does not require high levels of SMN protein (Iyer et al. 2015) while SMN is required for motor neuron function (McGovern et al. 2015). Somatic motor neurons innervating different structures in the eye on the other hand are resistant to the disease (Comley et al. 2015; Kubota et al. 2000), which enables the use of eye-tracking devices as a communication aid for advanced stage patients (Comley et al. 2015; Kubota et al. 2000).

In this study, we used a mouse model of SMA I to study RNA expression and splicing effects during disease progression in different motor neuron groups. The aim of the project was to identify protective gene expression changes within resistant motor neurons and potentially damaging gene responses in the vulnerable neurons to better understand disease progression and cell type specificity in SMA. Future experiment aim to modulate possible protective and/or damaging gene responses *in vitro* and *in vivo* to test their therapeutic potential.

Since mice only have one SMN gene, the mouse SMA model carries a transgene for two copies of human *SMN2* and of SMNΔ7, while the endogenous *Smn1* gene is knocked out (Le et al. 2005). In this mouse model, the pups do not appear symptomatic until 5 days after birth, but the therapeutic window to rescue the animals from disease by restoring SMA levels falls within the first postnatal week (Le et al. 2011; Lutz et al. 2011), and the life span is about 13 days (Le et al. 2005). Therefore, we chose to study vulnerable and resistant somatic motor neurons from postnatal day 2 and 5, to elucidate the mechanistic of the disease onset in the mouse model.

We used cells from lumbar spinal cord and facial nucleus (CN7) to represent the vulnerable population, and the oculomotor (CN3) and trochlear (CN4) nucleus of the eye as well as the hypoglossal nucleus (CN12) innervating the tongue as a resistant population (Comley et al. 2015). As controls, we also used visceral (non-somatic) motor neurons from the dorsal vagus nerve (CN10), and non-cholinergic red nucleus neurons (RN). Since not all of these cells have known unique marker genes, it was not possible to get a pure cell population using fluorescence activated cell sorting (FACS). Instead, we used laser capture microdissection (LCM) to outline and extract these anatomically distinct cellular populations of interest from fixed tissue slides. We picked and pooled around 150 cells of each type from each animal, and performed RNA-seq using Smart-seq2. A disadvantage of using LCM for sample acquisition is that it is hard to avoid some degree of RNA degradation in this process. Since Smart-Seq2 captures polyadenylated RNA, only the 3' most fragments of degraded mRNA will be included in the sequencing data. However, despite having a clear 3' bias in our samples, we still managed to get read coverage across the whole gene body of mRNAs.

I decided to use DESeq2 (Love et al. 2014) for the differential splicing analysis. I had previously seen that edgeR was prone to give too high significance to lowly expressed genes, when the expression difference was low in magnitude but high in fold change. DESeq on the other hand had shown too poor sensitivity in the past, which I hoped would have been improved in the new release. However, I found that the independent filtering and Cook's distance filtering options added to DESeq2 caused problems with detection of our positive control gene, *Smn1*. The independent filtering is a way for DESeq2 to enrich for genes that are more likely to be differentially expressed by ranking them on mean expression, and then iteratively estimating the number of significant genes obtained after multiple testing correction when removing differently sized quantiles of genes from the lowest ranked set. This approach seems to backfire in our data and we are left with only the 90th quantile of genes in some samples, which effectively meant a threshold base mean read count of 250 reads. I think the approach is generally problematic – although it does filter in an unbiased way, it assumes that getting a higher number of significant genes is always better, which is not the case if some of these are false positives.

The Cook's distance filter in DESeq2 is intended to remove strong outliers if the sample has more than 6 replicates, or remove the whole gene from analysis when there are 6 or less replicates. DESeq2 also uses a log2 fold change (LFC) shrinkage algorithm to reduce the problem of getting high fold changes and consequently high significance for lowly expressed genes. The samples from SMA mice would sometimes have one replicate with 3-10 reads mapping to *Smn1*, possibly from mismapping of *SMN2* reads, while the remaining samples had 0 reads. These cases would both cause the Cook's distance filter to remove the gene due to outliers, and to cause very strong fold change shrinkage. The latter seems to be a problem that occurs specifically when one condition has zero expression in most replicates, and adding a pseudo-count of one read to all genes eliminated the problem.

Hence, we customized our use of DESeq2 to run without independent filtering and without Cook's distance filtering. Instead, we removed all genes with zero count across all samples and added a pseudo-count of 1 read to the remaining genes before running the differential expression algorithm. With this approach, down regulation of our positive control gene *Smn1* in the SMA mice was highly significant in all cell types, suggesting that the method is performing well.

We could distinguish distinct transcriptional profiles for each neuronal subtype and using principal component analysis we discovered that CN7, CN12 and spinal motor neurons clustered together, while resistant CN3/4 clustered separately. Comparing the expression between healthy and SMA neurons produced a surprisingly low number of genes commonly regulated across cell types in response to the disease. We could not find any candidate "protective" genes that were common across the resistant cells but not the vulnerable ones, and neither could we find any "risk" genes in common to only vulnerable cells. To understand which genes are involved in disease vulnerability and disease we will need to do a more in depth analysis taking both basal gene expression levels and expression changes upon disease into account. We did find *Cdkn1a*, a cell cycle factor also implicated in apoptosis, up-regulated in disease across all somatic neurons, and a subcomponent of the spliceosome, *Snrpa1*, up-regulated in most somatic neurons.

Since SMN is involved in splicing, we also wanted to investigate whether we could detect splicing changes as an effect of the disease. Due to the degraded state of RNA in our samples, we were not sure if this was at all possible. I first tried using DEXSeq (Anders et al. 2012), which is based on an exon level analysis that they call differential exon usage (DEU). In this method they calculate the exon usage by comparing the read count in each exon to the read count in the whole transcript, and by comparing these relative numbers across conditions they determine the differential exon usage. However, it soon turned out that this method was not feasible for this data set, because of its 3' biased profile. DEXSeq had a tendency to overestimate DEU when the coverage of an exon is low, which in our dataset created a false positive DEU bias toward the 5' end. Instead I used MISO (Mixtures of Isoforms), which is focused on splicing events rather than individual exons, and uses reads that exclusively support a specific splice form (both exon reads and and spliced reads) to calculate a probable inclusion level of the two alternative splice forms in the event (Katz et al. 2010). I found that the sequenced read depth coverage in individual replicates was not enough to get a satisfying amount of these exclusive type of reads for splicing analysis. We therefore pooled all replicates from the same condition to get more power in the analysis.

From the splicing analysis, we could confirm a previously known splicing change in SMA – an increased inclusion of an alternative exon in *Uspl1*. A previous study reported a reduced inclusion level of *SMN2* exon 7 in vulnerable spinal MNs compared to resistant non-MNs in the spinal cord, and hypothesized that this could be the cause of their differential sensitivity (Ruggiu et al. 2012). However, our data, comparing multiple somatic MN groups with non-MNs, indicates that this is not the cause of differential vulnerability among MN groups or between MNs and non-MNs. The inclusion levels were roughly the same (1-4%) across all cell types. In general, the majority of alternative splicing events due to SMA were cell type specific, just as the differential expression response.

This study is still on going, and with a third time point (P10) we hope to be able to say more about the progression of the disease and when the most important events occur. There is need for a more in depth splicing analysis to determine if there are functionally important changes in splicing that might contribute to the disease. It would be valuable to find protective factors downstream of SMN in the resistant motor neurons, as this could open up for potential new therapies.

6 SUMMARY AND FUTURE PERSPECTIVES

The high throughput sequencing technology has opened up for a new world of discoveries, where global gene expression, mutations, DNA or RNA binding profiles, and chromatin conformation can be studied. The new studies are reinforcing a very complex picture of vertebrate gene regulation, where transcription, splicing, RNA degradation, DNA methylation and chromatin modifications seem to be functionally closely intertwined, although cause and effect is not trivial to elucidate. Evidence from RNA-sequencing studies has shown that although a very low proportion of the genome is protein coding, most of it is still transcribed, in some cases to create functionally important non-coding RNAs, and in other cases most likely as a consequence of RNA polymerase recruitment to the region (such as PROMPTs). A large proportion of the transcribed sequence also consists of intronic sequence, and although the introns are not functional themselves, the exon-intron structure is important for providing an increased layer of complexity by means of alternative splicing. In **paper IV**, we used RNA-sequencing to understand how a splicing difference in *SMN2* compared to *SMN1* affects gene expression and splicing to create pathological phenotypes in somatic motor neurons in Spinal Muscular Atrophy, where *SMN1* is lost. Due to technological advancements in RNA-sequencing, this could be done in small samples of only a single cell type per sample even when the cell type was rare and did not have known unique marker genes, and on full-length RNA rather than 5' or 3' tags. Thus, we could examine both expression and splicing differences between vulnerable and resistant cell types.

There are several promising therapies for SMA under development with different approaches to treating the disease. One approach involves using antisense oligonucleotides that target the splicing silencer in *SMN2* exon 7, to increase its efficiency at producing full length SMN. ISIS-SMN$_{Rx}$ is a drug using this approach that is currently in phase III clinical trials. A problem with the SMN-dependent therapies might not be only be effective when used early in the disease as suggested by the early therapeutic window in the mouse model. Therefore, studies like ours to determine the downstream effects can be important for finding therapeutic targets that can be addressed later in the progression. There are also promising SMN-independent drugs under development like Olesoxime, a small molecule that may act to protect mitochondria from cellular stress, which has been shown to have a protective effect in SMA. This drug is also in phase III clinical trials. A good approach might be to combine SMN-dependent and -independent approaches for an effective therapy (Tisdale & Pellizzoni 2015).

With the study in **paper II** we have increased the understanding of the connection between transcription and degradation for non-functional RNA, by combining several high throughput technologies like mass spectrometry, RNA-sequencing and chromatin- and RNA immunoprecipitation. These kind of combinatorial studies are important to get a systematic view of how regulation is exerted, and I think this is something that is desirable to pursue in the field, in order to ultimately understand functionality at both the RNA and protein level.

It is also evident that there are nucleotides sequences in both DNA and RNA that provides recognition signals for DNA and RNA binding proteins rather than coding for the final product. These signals regulate functions such as expression initiation of DNA, and localization, splicing and other processing steps of RNA. However, the identity of localization signals in RNA has proven more difficult to study than expected, as we discuss in **paper III**, probably reflecting that specificity of contact surfaces of RNA are formed from both the sequence itself, possibly base modifications and its secondary structure. To decode the rules for this regulation would require a combination of RNA-sequencing with pull-down assays for RNA binding proteins, and predictions of availability based on secondary structures.

Although RNA sequencing has allowed for tremendous progress in the understanding of gene regulation, it has its limitations. One such limitation is that of short read lengths, which complicate the identification of the origin of reads. In **paper I** we presented a method for handling this problem to avoid bias in gene expression estimation. Although this method is useful for assuring a robust normalization, it has its limitations when it comes to highly repetitive genes, or genes with highly similar paralogues, which might not be possible to estimate based on only uniquely mapping reads. The uniqueness information rendered by the program and presented in the paper can also be useful for guiding choices of read length in different sequencing applications.

Novel sequencing technologies are enabling spatial resolution for RNA-sequencing, either by barcoding reads by position and conventional sequencing, or by performing the sequencing reaction in situ. However, there is still a limitation in the current technologies regarding the temporal aspect of gene expression. This information is not available since RNA is sampled at a distinct time point, which kills the cells and prevents further investigation of the same cell. Therefore, dynamics of gene expression can only be indirectly inferred from looking at the range of states in a collection of states instead of directly observed in a single cell. To my knowledge, there are currently no technologies that enable high throughput live imaging of gene expression. The next big challenge in the omics field is to develop technologies that allow for getting high throughput data on protein, RNA and DNA level simultaneously from single cells, and preferably while preserving spatial and temporal information.

7 ACKNOWLEDGEMENTS

First of all, thank you **Rickard Sandberg** for being a great supervisor. Thank you for encouraging me to pursue my own ideas, and letting me choose my own path, but also recognizing when I'm stuck and give valuable help and suggestions. Thank you for giving me the opportunity to go to world-class conferences and courses, for teaching me Python and for being the kind of researcher and person I look up to.

To my co-supervisor **Elisabet "Lizzy" Andersson**, thank you for your help and support during my PhD, and also for just being a nice person to chat with.

A big thank you to all **my collaborators**, without you this work would not have been possible. Thank you **Peter Refsing Andersen** and **Torben Heick Jensen** for a fun and fruitful collaboration and introducing me to the world of PROMPTs and exosome degradation. I learnt a lot, and thought it was a fun and challenging project. Thank you **Chika Yokota** and **Jan Stenman** for the collaboration in the *Xenopus laevis* project. Without Chika's expertise in frog embryos this project could never have happened and I'm very grateful for all your efforts and help. Thank you **Eva Hedlund** and **Susanne Nichterwitz** for our collaboration on the SMA project. It is such an interesting project, and the two of you are great to work with. I think we are learning a lot from each other, and I'm as excited as you to see the final time point and where the data finally takes us! Thank you both for putting this manuscript together so fast, and to Eva for proof reading the SMA text for my thesis. Thank you also to **Indira Chivukula** and **Nigel Kee** for including me in your projects.

Thank you to the present and past members of the **Sandberg lab**, and the new **Deng lab** for being really great people both to work with and hang out with. I love the atmosphere in our lab, and I will miss you all when I go on to new adventures. **Björn Reinius**, thank you for being a friend, and a tremendous help with everything from library construction to text editing and general advice. I'm sure you will soon be heading a group of your own, and I'm sure you will do it well. **Per Johnsson**, congratulations to your new baby! And thank you for taking time from being a new dad to give feedback on my thesis. **Åsa Segerstolpe**, thank you for being so nice and relieving me of the sequencing delivery duties during the hectic thesis-writing times. **Daniel Edsgärd**, thank you for being the bioinformatics guru of the lab, and trying to teach the rest of us some proper statistics (and also some Salsa moves). **Qiaolin Deng**, I'm so happy for you that you have your own group now! Well deserved, you will go far. And thank you for bringing me in on the SMA project with Eva. **Åsa Björklund** and **Athanasia "Sissy" Palasantza**, thank you for being great colleagues, and for showing the world that cool girls do bioinformatics. **Sophie Petropoulos** and **Michael Hagemann-Jensen**, thanks for providing both hard work and party spirit to the lab. And **Gösta Winberg**

too, I applaud both your skills in the lab and in drink mixing. **Marlene Yilmaz**, we had some good times as room mates in Holland, learning words of wisdom both from big scientists and random DJs. **Omid Faridani, Mtakai Ngara** and **Geng Chen**, thank you for all the interesting lunch discussions on culture, religion, history and politics. Especially those long sunny ones on the roof terrace. **Ilgar Abdullayev**, thank you for being a very good friend and a fellow space nerd. Let's keep in touch and have more discussions on life, astronomy, physics, metaphysics and movies. Soon it is your turn to defend! All the best, I know you will do well. **Daniel Ramsköld**, thank you for always being helpful and for sharing all that knowledge you have. **Sven Sagasser,** you are always a lot of fun, and always the gentleman. I wish you the best with deciphering the biology of your sea creatures, and with your future adventures in Holland. **Ersen Kavak,** thank you for all the good times we had, dancing in Herräng, New Years in Boden, Valborgs in Uppsala and you showing us your Istanbul at the lab retreat. I wish you all the best in life.

Thank you to **Jens Mittag, Susi Gralla** and **Milica Uhde** in the Vennström lab for "adopting" me in your group when I was new and didn't know anyone. Those lunch breaks grew in to a great friendship! **Jens**, I miss your signature sarcasm and plans of world domination at the lunch table. All the best to you and **Henriette**, and little Martha. **Milica,** I'm so happy for you that you found an exciting new job! It sounds like a great fit for you. I'm sure you will be in a fabulous flat in a big city with your family in a few years, making time for both a great career, your lovely family, and having those champagne lunches you dreamed about. **Chris,** I admire your ambition and drive, and I'm sure you will reach those career goals you are aiming for. Thanks for the fun times during these years. All the best to little Sophie and Alex. **Susi,** you did such a great job on your defence, and with your smarts, sweet personality and all your positive energy I'm sure you will be great at whatever you do next. We both were never really sure what we wanted to do when we "grow up", but it seems like you have a good idea and a plan now. All the best of luck to you! **Micha,** thank you for your friendship, and all the fun times at courses and CMB pubs. Best of luck to you with the rest of your PhD! And of course all the best to Kalle and Nuka. **Amy Warner,** thank you for your friendship, for good times together and for sharing my vintage obsession. I hope you will have a great time at your new job in England, and I hope to come visit you there some time. My best to Daniel and Indie.

Thanks you **Anna-Klara Wicksén Blanchard** & **Albert "Blanchi" Blanchard**, **Emma Andersson** & **Nicola Fritz**, **Michalina Lewicka-Yammine** & **Samer Yammine** for many nice birthday parties and Christmas dinners together with you and your little ones.

I'm also happy that I got to know **Bodil Karlsson** and **Pegah Rouhi**. Thanks both of you for good times at parties, and to **Bodil** for all those vintage tips.

Thank you to everyone at **Ludwig**. It is really a great place to be working at, small enough to know everyone and filled with nice and helpful people. Thanks to **Nigel Kee, Daniel Hagey, Danny Topcic, Stuart Fell, Maria Bergsland, Nick Volakakis, Vilma Rraklli, Cécile Zaouter, Susanne Klum, Isabelle Westerlund** and everyone else for nice chats at lunches, coffee breaks and pubs, and also for hanging out at fun courses and conferences. And the best of luck to all you PhDs of my "generation", I'm sure the next few years will be filled with some excellent dissertations at Ludwig.

Thank you to all the past and present cool people at **CMB** that I always bump into at the best courses, trips and parties, like **Simona Hankeova, Pedro Réu, Niko Vojnovic, Anders Mutvei, Jens Magnusson, Gonçalo Brito, Tiago Pinheiro, Bettina Reichenbach, Indira Chivukula** and everyone else. I have enjoyed your company in lecture halls, parties, crowded cabins and ski slopes! A special thanks to **Tiago** for all the AWESOME parties. Thank you **Tanya Henshall** for your friendship, for fun times dancing together and for showing me your Adelaide.

Thank you also to the administrative staff at CMB and Ludwig, especially **Matti Nikkola** who helped us made the career course in Solvik happen, and made sure that my dissertation application got in on time!

Thank you **Katarina Tiklova, Bhumica Singla, Nina Kaukua, Natalija Gerasimcik, Vanessa von Hofer, Leela Karlstein** and all the other cool girls in the book club for our lovely dinners and discussions.

Thank you **Per Lönroth**, **Taraneh Foroughi** and **Marcus Meurling** for game nights and dinners and for being my friends outside the science bubble. **Per**, I hope you read everything up to this point like you promised. ;-)

Tack **Carolin Lindqvist** för alla bra länkar om SMA, och för att du är så intresserad av vårt projekt. Jag kommer att hålla dig uppdaterad om vi hittar något spännande.

Thank you to the **Blond Mafia** from Cold Spring Harbor. **Britta Will** and **Sjoerd Huisman**, I miss our early morning swims. **Evi Pashos** and **Juliane Schmidt**, that short sweet trip to Greece was amazing. I hope we all meet up again soon.

Thank you **Maria Johansson**, whom I have known so long I basically consider you my sister. I love how we can meet after a year, and pick up where we were as if no time passed. All the best to you and your family, and I hope we will stay friends many more years.

To my family. Thank you to my brother **Johan** and my sister **Linda,** for all those happy memories we share, for growing up to be such cool people, for being my best friends, and for being the only ones who really get and share my weird sense of humour. My life would have been so boring without you. **Mamma** och **Pappa.** Tack för att ni alltid uppmuntrat min nyfikenhet, och gett mig självförtroendet att känna att jag kan uppnå vilka mål jag än siktar på. Tack för er kärlek och ert aldrig sviktande stöd. Älskar er.

Tack till tidigare generationer som inspirerat mig. Min **Mormor Astrid** som alla säger att jag tar efter, när jag spenderar timmar i blåbärsskogen och oroar mig om alla fått tillräckligt med mat och bullar. Min snälla **Morfar Folke** som köpte mig min finaste sammetsklänning, och alltid var stiligt klädd om han så bara skulle till affären. Att vara hos er har alltid varit att vara mer än hemma för mig. Till min **Farmor Göta**, som alltid var så intresserad och mån om alla sina barnbarns skolgång. Jag vet att hon skulle varit stolt nu, och visat den här boken för alla sina grannar. Till **Farfar Folke**, som jag mest minns att han var snäll och lång, men som säkert också hade varit stolt. Till morfars syster **Linnéa**, en av mina idoler, som var ingenjör och byggde ett eget hus och sydde sina egna kläder och var allmänt fylld med jävlar anamma.

Joakim. Thank you for standing by me, and loving me all these years, and especially during these last months of crazy workload. Thank you for comforting me in my low points, for pushing me to do better, and for putting me first even at a time when you often came second to my work. Thank you for listening to me complain and panic at times, and for proof reading my thesis text even though it is way out of your field. I don’t know how I would have managed without you. I love you.

8 REFERENCES

Adams, M.D., 2000. The Genome Sequence of Drosophila melanogaster. *Science (New York, N.Y.)*, 287(5461), pp.2185–2195.

Alberts, B., 2004. *Essential Cell Biology*, Taylor & Francis.

Allison, D.B. et al., 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nature reviews. Genetics*, 7(1), pp.55–65.

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 of the United States of America*, 74(12), pp.5350–5354.

Anders, S., Reyes, A. & Huber, W., 2012. Detecting differential usage of exons from RNA-seq data. *Genome research*, 22(10), pp.2008–2017.

Arner, E. et al., 2015. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science (New York, N.Y.)*, 347(6225), pp.1010–1014.

Battich, N., Stoeger, T. & Pelkmans, L., 2013. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nature methods*, 10(11), pp.1127–1133.

Bertone, P., Gerstein, M. & Snyder, M., 2005. Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 13(3), pp.259–274.

Blower, M.D., 2013. Molecular insights into intracellular RNA localization. *International review of cell and molecular biology*, 302, pp.1–39.

Bonasio, R., Tu, S. & Reinberg, D., 2010. Molecular signals of epigenetic states. *Science (New York, N.Y.)*, 330(6004), pp.612–616.

Bray, N. et al., 2015. Near-optimal RNA-Seq quantification. *arXiv.org*, q-bio.QM.

Brennecke, P. et al., 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods*, 10(11), pp.1093–1095.

Calvin, K. & Li, H., 2008. RNA-splicing endonuclease structure and function. *Cellular and Molecular Life Sciences*, 65(7-8), pp.1176–1185.

Camacho, C. et al., 2009. BLAST+: architecture and applications. *BMC bioinformatics*, 10(1), p.421.

Cartegni, L. & Krainer, A.R., 2002. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nature genetics*, 30(4), pp.377–384.

Chen, Q. et al., 2011. Hierarchical regulation of mRNA partitioning between the cytoplasm and the endoplasmic reticulum of mammalian cells. *Molecular biology* *of the cell*, 22(14), pp.2646–2658.

Choi, C.Q., 2003. Who'll sweep the Gene Sweepstake? *Genome biology*, 4(4), pp.spotlight–20030430–01.

Combs, P.A. & Eisen, M.B., 2013. Sequencing mRNA from cryo-sliced Drosophila embryos to determine genome-wide spatial patterns of gene expression. B. Jennings, ed. *PloS one*, 8(8), p.e71820.

Comley, L.H. et al., 2015. Cross-disease comparison of amyotrophic lateral sclerosis and spinal muscular atrophy reveals conservation of selective vulnerability but differential neuromuscular junction pathology. *The Journal of comparative neurology*, pp.n/a–n/a.

Corral-Debrinski, M., Blugeon, C. & Jacq, C., 2000. In yeast, the 3' untranslated region or the presequence of ATM1 is required for the exclusive localization of its mRNA to the vicinity of mitochondria. *Molecular and Cellular Biology*, 20(21), pp.7881–7892.

Cox, L.J. et al., 2008. Intra-axonal translation and retrograde trafficking of CREB promotes neuronal survival. *Nature cell biology*, 10(2), pp.149–159.

Crick, F.H., 1958. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12, pp.138–163.

Crosetto, N., Bienko, M. & van Oudenaarden, A., 2015. Spatially resolved transcriptomics and beyond. *Nature reviews. Genetics*, 16(1), pp.57–66.

Dabney, J. & Meyer, M., 2012. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, 52(2), pp.87–94.

De Domenico, E. et al., 2015. Molecular asymmetry in the 8-cell stage Xenopus tropicalis embryo described by single blastomere transcript sequencing. *Developmental biology*, 408(2), pp.252–268.

Djebali, S. et al., 2012. Landscape of transcription in human cells. *Nature*, 489(7414), pp.101–108.

Dobin, A. et al., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1), pp.15–21.

Dridi, S., 2012. Alu mobile elements: from junk DNA to genomic gems. *Scientifica*, 2012(5039), pp.545328–11.

Eddy, S.R., 2001. Non-coding RNA genes and the modern RNA world. *Nature reviews. Genetics*, 2(12), pp.919–929.

Edgell, D.R., Belfort, M. & Shub, D.A., 2000. Barriers to Intron Promiscuity in Bacteria. *Journal of Bacteriology*, 182(19), pp.5281–5289.

Edman, P. et al., 1950. Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chemica Scandinavica*, 4, pp.283–293.

Ellis, J.J., Broom, M. & Jones, S., 2007. Protein-RNA interactions: structural analysis and functional classes. *Proteins*, 66(4), pp.903–911.

Femino, A.M. et al., 1998. Visualization of single RNA transcripts in situ. *Science (New York, N.Y.)*, 280(5363), pp.585–590.

Fiers, W. et al., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260(5551), pp.500–507.

Flachsova, M., Sindelka, R. & Kubista, M., 2013. Single blastomere expression profiling of Xenopus laevis embryos of 8 to 32-cells reveals developmental asymmetry. *Scientific Reports*, 3, p.2278.

Freeman, S.D. et al., 2008. Extracellular regulation of developmental cell signaling by XtSulf1. *Developmental biology*, 320(2), pp.436–445.

Gamazon, E.R. & Stranger, B.E., 2014. Genomics of alternative splicing: evolution, development and pathophysiology. *Human Genetics*, 133(6), pp.679–687.

Ghosh, M. et al., 2006. Identification of the expressed form of human cytosolic phospholipase A2beta (cPLA2beta): cPLA2beta3 is a novel variant localized to mitochondria and early endosomes. *Journal of Biological Chemistry*, 281(24), pp.16615–16624.

Gilbert, W., 1978. Why genes in pieces? *, Published online: 09 February 1978; | doi:10.1038/271501a0*, 271(5645), pp.501–501.

Gott, J.M. & Emeson, R.B., 2000. Functions and mechanisms of RNA editing. *Annual review of genetics*, 34(1), pp.499–531.

Grabherr, M.G. et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7), pp.644–652.

Grün, D. et al., 2015. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568), pp.251–255.

Haas, B.J. et al., 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, 31(19), pp.5654–5666.

Hangauer, M.J., Vaughn, I.W. & McManus, M.T., 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. J. L. Rinn, ed. *PLoS genetics*, 9(6), p.e1003569.

Hashimshony, T. et al., 2012. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports*, 2(3), pp.666–673.

Heather, J.M. & Chain, B., 2015. The sequence of sequencers: The history of sequencing DNA. *Genomics*.

Hentschel, A., Zahedi, R.P. & Ahrends, R., 2015. Protein lipid modifications - more than just a greasy ballast. *Proteomics*, pp.n/a–n/a.

Higuchi, R. et al., 1993. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Bio/technology (Nature Publishing Company)*, 11(9), pp.1026–1030.

Higuchi, R. et al., 1992. Simultaneous amplification and detection of specific DNA sequences. *Bio/technology (Nature Publishing Company)*, 10(4), pp.413–417.

Holley, R.W. et al., 1965. STRUCTURE OF A RIBONUCLEIC ACID. *Science (New York, N.Y.)*, 147(3664), pp.1462–1465.

Horvay, K. et al., 2006. Xenopus Dead end mRNA is a localized maternal determinant that serves a conserved function in germ cell development. *Developmental biology*, 291(1), pp.1–11.

Houseley, J. & Tollervey, D., 2009. The many pathways of RNA degradation. *Cell*, 136(4), pp.763–776.

Houston, D.W., 2012. Cortical rotation and messenger RNA localization in Xenopus axis formation. *Wiley interdisciplinary reviews. Developmental biology*, 1(3), pp.371–388.

Hsieh, C.-L. et al., 2014. Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proceedings of the National Academy of Sciences of the United States of America*, 111(20), pp.7319–7324.

Hutchison, C.A., 2007. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, 35(18), pp.6227–6237.

Iascone, D.M., Henderson, C.E. & Lee, J.C., 2015. Spinal muscular atrophy: from tissue specificity to therapeutic strategies. *F1000Prime Rep*, 7(4), p.04.

Islam, S. et al., 2012. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nature protocols*, 7(5), pp.813–828.

Islam, S. et al., 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods*, 11(2), pp.163–166.

Iyer, C.C. et al., 2015. Low levels of Survival Motor Neuron protein are sufficient for normal muscle function in the SMNΔ7 mouse model of SMA. *Human molecular genetics*, 24(21), pp.6160–6173.

Johnson, L.N., 2009. The regulation of protein phosphorylation. *Biochemical Society transactions*, 37(Pt 4), pp.627–641.

Jones, P.A., 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13(7), pp.484–492.

Junker, J.P. et al., 2014. Genome-wide RNA Tomography in the zebrafish embryo. *Cell*, 159(3), pp.662–675.

Kapur, K. et al., 2007. Exon arrays provide accurate assessments of gene expression. *Genome biology*, 8(5), p.R82.

Kashima, T. & Manley, J.L., 2003. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nature genetics*, 34(4), pp.460–463.

Katz, Y. et al., 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12), pp.1009–1015.

Ke, R. et al., 2013. In situ sequencing for RNA analysis in preserved tissue and cells. *Nature methods*, 10(9), pp.857–860.

Kim, T.-K. et al., 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295), pp.182–187.

Kloc, M. & Etkin, L.D., 1994. Delocalization of Vg1 mRNA from the vegetal cortex in Xenopus oocytes after destruction of Xlsirt RNA. *Science (New York, N.Y.)*, 265(5175), pp.1101–1103.

Koren, E., Lev-Maor, G. & Ast, G., 2007. The Emergence of Alternative 3′ and 5′ Splice Site Exons from Constitutive Exons. *PLoS computational biology*, 3(5), p.e95.

Kubota, M. et al., 2000. New ocular movement detector system as a communication tool in ventilator-assisted Werdnig–Hoffmann disease. *Developmental Medicine & Child Neurology*, 42(01), p.61.

Kung, J.T.Y., Colognori, D. & Lee, J.T., 2013. Long noncoding RNAs: past, present, and future. *Genetics*, 193(3), pp.651–669.

Lander, E.S., 2011. Initial impact of the sequencing of the human genome. *Nature*, 470(7333), pp.187–197.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Lane, C.E. et al., 2007. Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50), pp.19908–19913.

Langley, A.R. et al., 2014. New insights into the maternal to zygotic transition. *Development (Cambridge, England)*, 141(20), pp.3834–3841.

Le, T.T. et al., 2005. SMNDelta7, the major product of the centromeric survival motor neuron (SMN2) gene, extends survival in mice with spinal muscular atrophy and associates with full-length SMN. *Human molecular genetics*, 14(6), pp.845–857.

Le, T.T. et al., 2011. Temporal requirement for high SMN expression in SMA mice. *Human molecular genetics*, 20(18), pp.3578–3591.

Lee, C. & Roy, M., 2004. Analysis of alternative splicing with microarrays: successes and challenges. *Genome biology*, 5(7), p.231.

Lee, J.H. et al., 2014. Highly multiplexed subcellular RNA sequencing in situ. *Science (New York, N.Y.)*, 343(6177), pp.1360–1363.

Lee, S. et al., 2011. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Research*, 39(2), pp.e9–e9.

Lee, Y. & Rio, D.C., 2015. Mechanisms and Regulation of Alternative Pre-mRNA

Splicing. *Annual Review of Biochemistry*, 84(1), pp.291–323.

Li, B. & Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1), p.323.

Li, B. et al., 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics (Oxford, England)*, 26(4), pp.493–500.

Li, W. et al., 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455), pp.516–520.

Li, X. et al., 2010. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA (New York, N.Y.)*, 16(6), pp.1096–1107.

Lloret-Llinares, M. et al., 2015. Relationships between PROMPT and gene expression. *RNA biology*, pp.0–00.

Lodish, H. et al., 2000. *Molecular Cell Biology*,

Lovatt, D. et al., 2014. Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nature methods*, 11(2), pp.190–196.

Love, M.I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12), p.550.

Lubeck, E. & Cai, L., 2012. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature methods*, 9(7), pp.743–748.

Lubeck, E. et al., 2014. Single-cell in situ RNA profiling by sequential hybridization. *Nature methods*, 11(4), pp.360–361.

Lutz, C.M. et al., 2011. Postsymptomatic restoration of SMN rescues the disease phenotype in a mouse model of severe spinal muscular atrophy. *The Journal of clinical investigation*, 121(8), pp.3029–3041.

MacArthur, H. et al., 2000. DEADSouth is a germ plasm specific DEAD-box RNA helicase in Xenopus related to eIF4A. *Mechanisms of development*, 95(1-2), pp.291–295.

Macosko, E.Z. et al., 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5), pp.1202–1214.

Makarova, J.A. & Kramerov, D.A., 2007. Noncoding RNAs. *Biochemistry. Biokhimii͡a*, 72(11), pp.1161–1178.

Malone, J.H. & Oliver, B., 2011. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9(1), p.34.

Marioni, J.C. et al., 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9), pp.1509–1517.

Marín-Béjar, O. & Huarte, M., 2015. RNA pulldown protocol for in vitro detection and identification of RNA-associated proteins. *Methods in molecular biology (Clifton, N.J.)*, 1206(Chapter 8), pp.87–95.

Mattick, J.S. & Makunin, I.V., 2006. Non-coding RNA. *Human molecular genetics*, 15 Spec No 1(90001), pp.R17–29.

McAndrew, P.E. et al., 1997. Identification of Proximal Spinal Muscular Atrophy Carriers and Patients by Analysis of SMNT and SMNC Gene Copy Number. *The American Journal of Human Genetics*, 60(6), pp.1411–1422.

McGinn, S. & Gut, I.G., 2013. DNA sequencing - spanning the generations. *New biotechnology*, 30(4), pp.366–372.

McGovern, V.L. et al., 2015. SMN expression is required in motor neurons to rescue electrophysiological deficits in the SMNΔ7 mouse model of SMA. *Human molecular genetics*, 24(19), pp.5524–5541.

Merkhofer, E.C., Hu, P. & Johnson, T.L., 2014. Introduction to Cotranscriptional RNA Splicing. In *Spliceosomal Pre-mRNA Splicing*. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 83–96.

Miller, J.R. et al., 1999. Establishment of the dorsal-ventral axis in Xenopus embryos coincides with the dorsal enrichment of dishevelled that is dependent on cortical rotation. *The Journal of cell biology*, 146(2), pp.427–437.

Monani, U.R. et al., 1999. A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2. *Human molecular genetics*, 8(7), pp.1177–1183.

Moore, M.J. & Silver, P.A., 2008. Global analysis of mRNA splicing. *RNA (New York, N.Y.)*, 14(2), pp.197–203.

Morin, R. et al., 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, 45(1), pp.81–94.

Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7), pp.621–628.

Munsat, T.L. & Davies, K.E., 1992. International SMA consortium meeting. (26-28 June 1992, Bonn, Germany). In Neuromuscular disorders : NMD. pp. 423–428.

Murray, V. & Holliday, R., 1979. Mechanism for RNA splicing of gene transcripts. *FEBS Letters*, 106(1), pp.5–7.

Mutz, K.-O. et al., 2013. Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology*, 24(1), pp.22–30.

Nilsson, M. et al., 1994. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science (New York, N.Y.)*, 265(5181), pp.2085–2088.

Nirenberg, M., 2004. Historical review: Deciphering the genetic code – a personal account. *Trends in Biochemical Sciences*, 29(1), pp.46–54.

Nirenberg, M. et al., 1966. The RNA code and protein synthesis. *Cold Spring Harbor symposia on quantitative biology*, 31, pp.11–24.

O'Neil, D., Glowatz, H. & Schlumpberger, M., 2013. Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 4, pp.Unit 4.19–4.19.8.

Okamura-Oho, Y. et al., 2012. Transcriptome tomography for brain analysis in the web-accessible anatomical space. S. Hayasaka, ed. *PloS one*, 7(9), p.e45373.

Pace, J.K. & Feschotte, C., 2007. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome research*, 17(4), pp.422–432.

Pan, Q. et al., 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12), pp.1413–1415.

Patel, A.P. et al., 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)*, 344(6190), pp.1396–1401.

Patro, R., Duggal, G. & Kingsford, C., 2015. *Accurate, fast, and model-aware transcript expression quantification with Salmon*, Cold Spring Harbor Labs Journals.

Patro, R., Mount, S.M. & Kingsford, C., 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5), pp.462–464.

Pavesi, G., Mauri, G. & Pesole, G., 2001. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics (Oxford, England)*, 17 Suppl 1, pp.S207–14.

Pertea, M., 2012. The human transcriptome: an unfinished story. *Genes*, 3(3), pp.344–360.

Picelli, S. et al., 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols*, 9(1), pp.171–181.

Picelli, S. et al., 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11), pp.1096–1098.

Preker, P. et al., 2008. RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. *Science (New York, N.Y.)*, 322(5909), pp.1851–1854.

Raijmakers, R., Schilders, G. & Pruijn, G.J.M., 2004. The exosome, a molecular machine for controlled RNA degradation in both nucleus and cytoplasm. *European journal of cell biology*, 83(5), pp.175–183.

Ramsköld, D. et al., 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. L. J. Jensen, ed. *PLoS computational biology*, 5(12), p.e1000598.

Rapaport, F. et al., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology*, 14(9), p.R95.

Ray, D. et al., 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457), pp.172–177.

Robinson, M.D. & Smyth, G.K., 2007. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics (Oxford, England)*, 23(21), pp.2881–2887.

Robinson, M.D. & Smyth, G.K., 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics (Oxford, England)*, 9(2), pp.321–332.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), pp.139–140.

Rogozin, I.B. et al., 2012. Origin and evolution of spliceosomal introns. *Biology direct*, 7(1), p.11.

Ruggiu, M. et al., 2012. A role for SMN exon 7 splicing in the selective vulnerability of motor neurons in spinal muscular atrophy. *Molecular and Cellular Biology*, 32(1), pp.126–138.

Salz, H.K., 2011. Sex determination in insects: a binary decision based on alternative splicing. *Current Opinion in Genetics & Development*, 21(4), pp.395–400.

Sanger, F., Brownlee, G.G. & Barrell, B.G., 1965. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of Molecular Biology*, 13(2), pp.373–398.

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–5467.

Schena, M. et al., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235), pp.467–470.

Schübeler, D., 2015. Function and information content of DNA methylation. *Nature*, 517(7534), pp.321–326.

Shendure, J. & Ji, H., 2008. Next-generation DNA sequencing. *Nature biotechnology*, 26(10), pp.1135–1145.

Strasser, B.J., 2006. *A world in one dimension: Linus Pauling, Francis Crick and the central dogma of molecular biology*,

Sugarbaker, D.J. et al., 2008. Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9), pp.3521–3526.

Takahashi, K. & Yamanaka, S., 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4), pp.663–676.

The C elegans Sequencing Consortium, 1998. Genome Sequence of the Nematode C.&nbsp;elegans: A Platform for Investigating Biology. *Science (New York, N.Y.)*, 282(5396), pp.2012–2018.

Tisdale, S. & Pellizzoni, L., 2015. Disease mechanisms and therapeutic approaches in spinal muscular atrophy. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35(23), pp.8691–8700.

Torres, T.T. et al., 2008. Gene expression profiling by massively parallel sequencing. *Genome research*, 18(1), pp.172–177.

Trapnell, C. et al., 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1), pp.46–53.

Trapnell, C. et al., 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4), pp.381–386.

Trapnell, C. et al., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5), pp.511–515.

Trapnell, C., Pachter, L. & Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9), pp.1105–1111.

Van der Maaten, L. & Hinton, G., 2008. Visualizing data using t-SNE. *Journal of Machine Learning ...*.

Venter, J.C. et al., 2001. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp.1304–1351.

Vogel, C. & Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13(4), pp.227–232.

Wagner, G.P., Kin, K. & Lynch, V.J., 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences = Theorie in den Biowissenschaften*, 131(4), pp.281–285.

Wang, E.T. et al., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), pp.470–476.

Wang, J. et al., 2004. Mouse transcriptome: Neutral evolution of “non-coding” complementary DNAs. *Nature*, 431(7010).

Watanabe, Y.-I. et al., 2002. Introns in protein-coding genes in Archaea. *FEBS Letters*, 510(1-2), pp.27–30.

Watson, J.D., 2001. The human genome revealed. *Genome research*, 11(11), pp.1803–1804.

Watson, J.D. & Crick, F.H., 1953. The structure of DNA. *Cold Spring Harbor symposia on quantitative biology*, 18, pp.123–131.

Weber, A.P.M. et al., 2007. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant physiology*, 144(1), pp.32–42.

Wittwer, C.T. et al., 1997. Continuous fluorescence monitoring of rapid cycle DNA amplification. *BioTechniques*, 22(1), pp.130–1– 134–8.

Wong, M.S., Wright, W.E. & Shay, J.W., 2014. Alternative splicing regulation of telomerase: a new paradigm? *Trends in genetics : TIG*, 30(10), pp.430–438.

Woo, Y. et al., 2004. A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *Journal of biomolecular techniques : JBT*, 15(4), pp.276–284.

Yokobori, S.-I. et al., 2009. Gain and loss of an intron in a protein-coding gene in Archaea: the case of an archaeal RNA pseudouridine synthase gene. *BMC Evolutionary Biology*, 9(1), p.198.