Chapter 7

Summary in English

Bioinformatics, first introduced in 1970 by Paulien Hogeweg and Ben Hesper, has changed and evolved drastically. Bioinformatics was formed as a field to get information about various processes in biological systems (hence the term *bio-informatics*). Initially, bioinformatics focused on analyzing a handful of gene and protein sequences, protein structures and organism evolution based on protein sequences and structural domains. However, with years it shifted to the exploration of whole genomes and transcriptomes and the study of biological processes at a genome- (and transcriptome-) wide scale in model systems, living cells and organisms. After the introduction of one of the greatest scientific and technological breakthroughs of the past two decades – massively parallel Next Generation Sequencing (NGS) – sequencing of genomes and transcriptomes of whole organisms became possible. Bioinformatics faced a new challenge of analyzing high throughput genomic and transcriptomic data. The main challenge introduced by complex NGS data is in its nature – DNA has to be shredded before sequencing, therefore NGS produces millions of very short sequenced DNA fragments (*reads*). Despite the recent increase in read length, finding the origin of reads on the genome is still a challenge, as the length still does not exceed hundreds of nucleotides. Pacific Biosciences' SMRT technology and Oxford Nanopore, recently introduced high-throughtput sequencing technologies, are the only technoloies available on the market capable of sequencing thousands of bases.

An additional challenge in analyzing NGS data emerges when working with transcriptomic material – reads originating from RNA. Multiple RNA transcripts from one gene often share a specific region, and each transcript can be present multiple times. Therefore, the origin of reads – the transcript they are coming from – is hard to find. This introduces numerous biases and limits our understanding of RNA biology and molecular processes happening at the RNA level. As this thesis was written, the analysis of an RNA sequencing sample was limited due to the limited amount of programs for the bioinformatic analysis of RNA. Available RNA analysis tools only focussed on finding differences in the expression of certain genes between different samples (i.e., healthy individuals versus diseased patients), discover alternative splicing events, identify alternative polyadenylation events.

The focus of this thesis was to explore the possibilities of current bioinformatic analysis of transciptomic data and complement such tools by newly developed programs. The latter was particularly important for atypical RNA analyses revealing yet unexplored molecular mechanisms in RNA biogenesis. This is covered in **Chapter 2**, where we analyzed nuclear RNA – not commonly used for RNA-seq. The majority of established RNA-seq data sets are derived from total cellular or cytoplasmic RNA. In the analysis of total RNA, the focus is on the result of transcription and RNA processing. Studying nuclear RNA facilitates a more thorough exploration of ongoing mRNA processing events like splicing.

Splicing is a process during which the regions of a gene called introns are removed and the remaining regions – exons – are joined. As simple as splicing might sound at first, it consists of numerous steps, and each of those steps deserves a thorough exploration. Analyzing reads from nuclear RNA in **Chapters 2** and **3** of this thesis, we discovered that introns in human genes are not always excised sequentially, from

the beginning towards the end of the gene. Also, some introns are not removed at once, they are rather excised in multiple steps, piece by piece. In fact, using the example of the gene coding for dystrophin (*DMD*), the longest gene within the human genome, we show that more than half of *DMD* introns are removed in multiple steps.

In **Chapter 4**, we complemented publicly available programs with in-house developed scripts and performed a study of age related changes in RNA processing, namely splicing, in over 600 healthy Dutch individuals. The analyses were conducted on adult individuals with ages spanning more than 60 years. We found alternative splicing events such as exon skipping – when during the process of splicing not only introns but also an exon is excised – occurring more often in older individuals. We noticed the same trend for another type of alternative splicing event, intron retention, an event when an intron is not excised and is included in the nascent transcript. Both of these events were occurring more often in older individuals transcriptome-wide, without any preference for a specific pathway or functional group of genes. These findings suggested that such age related changes might have a fundamental nature and can potentially help to understand how and why we age.

Next to the bioinformatic analyses of RNA sequencing samples from healthy individuals, in **Chapter 5** we analyzed samples from patients with breast cancer. In this study we were particularly interested in allele-specific expression, a biological event potentially important for the pathophysiology of cancer. A healthy human genome is diploid, which means that every chromosome occurs in pairs and that there are two copies of each gene in a cell (the two copies are the two *alleles*). For each position of a gene sequence two different nucleotides can be present (unless we are facing an insertion or a deletion, which results in a missing nucleotide). Chromosomal aberrations – duplication or deletion of (parts of) a chromosome (allele) – are frequent in tumors. Such events may result in an unequal expression of the two alleles. This unequal allelic expression may also be present in healthy individuals and can be directed via biological mechanisms different from DNA chromosomal aberrations (e.g. parental imprinting). In this study we compared the differences between imbalanced allelic expression between healthy and tumor tissue and explored the potential role of allelic imbalance in the pathogenesis of breast cancer. We provided a detailed classification of allele-specific expression cases, compared them in control versus tumor tissues of the same patients and explained each pattern on RNA expression (e.g., a downregulation of an allele's expression) by the underlying DNA alterations.

In summary, this thesis focuses on the uncovering of previously unexplored information in RNA sequencing data. We show that use of unconventional types of RNA for sequencing and new tools for the analysis RNA-seq can reveal novel insights in the transcriptional and post-transcriptional regulation of gene expression, relevant for the understanding of normal physiological processes such as aging and pathophysiological processes such as Duchenne muscular dystrophy and cancer.