Summary

The body of all humans is built of small building blocks called cells, that carry out the functions necessary for life. The genetic material is contained within the nucleus of the cells. In the nucleus of the cell two copies of each of the 22 autosomal chromosomes and 2 sex chromosomes are stored. These chromosomes are composed of twisted ladder shaped deoxyribonucleic acid (DNA) containing 4 nucleotide bases called adenine (A), thymine (T), guanine (G), and cytosine (C). Genes are a specific combination of these nucleotides which together comprise a complete set of instructions for the cell to be able to function, grow and survive. In a multi-step process the information coded in the DNA is converted into a functional protein molecule that can be used by the cell. The first step in this process is transcription which uses information from the DNA to make messenger RNA (mRNA). In the second step, information from the mRNA is translated into a protein. In addition, several types of so-called non-coding RNA molecules are encoded by the DNA. These molecules are transcribed into RNA but not translated into protein. An example of these non-coding molecules are microRNAs (miRNAs). Instead of becoming protein, miRNAs bind to mRNAs and inhibit their translation into protein. In this thesis we investigated changes on DNA, mRNA and miRNA molecule levels.

Cancer is caused by a variety of factors that lead to abnormalities in DNA sequence and/or amount that change the normal functioning of the cell. Accumulation of these abnormalities and deregulations can ultimately lead to cancer. During this process of carcinogenesis, the genes that control the normal function of the cell (tumor suppressor genes) become inactive, while genes that initiate proliferation (oncogenes), start the process of fast and uncontrolled cell division. The process of carcinogenesis can be caused by various carcinogens like smoking, UV light, hereditary genetic alterations and viral infection. The most well-known case of viral infection caused carcinogenesis is cervical cancer, induced by human papillomavirus (HPV). Although more than 100 types of HPV exist, more than 70% of cervical carcinomas are caused by the high-risk types HPV16 and HPV18. Worldwide, cervical cancer is the fourth most diagnosed cancer in females, with the highest incidence in developing countries, due to the lack of population based screening programs. It is now widely accepted that involvement of high-risk types of HPV followed by genetic alterations can ultimately lead to cervical cancer.

The development and progression of cancer is a dynamic and complex biological process. To properly address the complexity of carcinogenesis, we need to simulate this process using experimental conditions we can control that allow us to follow this process over time. In this thesis we used a model consisting of normal human cells in which we introduced HPV that we interrogated at multiple moments in time at 3 molecular levels. In our cervical cancer experimental model two cell lines are infected with HPV16 and two with HPV18. The obtained cell line experimental model is shown to faithfully mimic cervical cancer development morphologically and (epi)genetically, which is called transformation. During this transformation process we monitored all cell lines at different stages which allowed us to gain insights in genes consistently altered over time, as well as deregulation of a group of genes (pathways) that work together.

The goal of this thesis was to conduct comprehensive investigation of HPV-induced carcinogenesis using the above described model. For this we needed to develop and apply novel statistical methodology. The cell lines were profiled at 8 consecutive time points for DNA copy number, mRNA, and miRNA gene expression. Although many statistical models exist for analysis of time course data, all of them are focused on a single molecular level. However, we strongly believe that to be able to truly understand cancer, alterations in both DNA and RNA need to be investigated. So due to the fact that our experiment comprises data from different molecular levels, we needed to develop novel integrative statistical methodology to address this problem.

In this thesis we developed the integrative longitudinal statistical methodology for analyzing the data, both for single genes and groups of related genes (called pathways). In Chapter 2, which aims to perform temporal differential expression analysis, genes with the highest variation over time caused by abnormalities in DNA copy number were identified. On the other hand, analysis of well-defined groups of related genes (pathways) requires a methodology for network reconstruction presented in Chapter 3 and Chapter 4. Network reconstruction aims to identify gene-gene interactions between mRNAs, as well as relations between different molecular levels (e.g. how does a DNA copy number change affect this gene's mRNA expression levels). Employing this methodology to our experiment allows us to zoom-into the data and identify key genes within pathways that may be potential biomarkers or therapeutic targets.

Subsequent interpretation of the analysis result is divided into several steps, to ultimately reduce the number of potential gene candidates for an improved understanding of the underlying carcinogenic process. First, temporal differential mRNA and miRNA gene expression analysis were performed, where 106 miRNAs and 3642 mRNAs are identified with significant variation over time. Out of these number of genes, it is found that approximately 33% of differently expressed mRNA and miRNA are associated with chromosome abnormalities. All these genes are either up- or down-regulated in at least three out of four cell lines over time due to the presence of more or less DNA encoding for these genes. In order to further understand the effect of the observed changes for the functioning of the cell, the analysis is moved to the pathway level.

We have selected a number of pathways that warrant further investigation based on the fact that they were overrepresented among the previously identified altered genes. For these pathways, temporal gene regulatory network reconstruction was performed both within the mRNA level, as well as the effect of DNA copy number on mRNAs. Here we aimed to identify genes with the highest number of interactions within the pathway over time, which we called regulators as we believe they represent key altered genes for this pathway during cancer development. ID1 and PITX2 were identified as main regulators of the altered TGF-beta signaling pathway, BRWD3, and NF2 for mTOR signaling, while PIGT and DAPP1 for focal adhesion pathway. Functional validation experiments in the cell lines support the validity of our approach and the relevance of the identified genes.

This thesis sets the basis of integrative temporal statistical methodology in the direction of the differential expression analysis, gene regulatory network reconstruction

and identification of miRNA targets. The validity and applicability of our approaches were illustrated in the comprehensive investigation of molecular mechanisms underlying HPV-induced transformation in a cell line model. Our developed statistical methodology are important for an upcoming area in cancer research. Although currently in the literature multi-level time course experiments and corresponding datasets are rare, the expected introduction of blood-based tests for early cancer detection and treatment monitoring will generate molecular time course data in the near future.