Decoding the Structural Layer of Transcriptional Regulation

Computational Analyses of Chromatin and Chromosomal Aberrations

ROBIN ANDERSSON
Dissertation presented at Uppsala University to be publicly examined in C4:305, BMC, Husargatan 3, Uppsala, Tuesday, November 2, 2010 at 09:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract


Gene activity is regulated at two separate layers. Through structural and chemical properties of DNA – the primary layer of encoding – local signatures may enable, or disable, the binding of proteins or complexes of them with regulatory potential to the DNA. At a higher level – the structural layer of encoding – gene activity is regulated through the properties of higher order DNA structure, chromatin, and chromosome organization. Cells with abnormal chromosome compaction or organization, e.g. cancer cells, may thus have perturbed regulatory activities resulting in abnormal gene activity.

Hence, there is a great need to decode the transcriptional regulation encoded in both layers to further our understanding of the factors that control activity and life of a cell and, ultimately, an organism. Modern genome-wide studies with those aims rely on data-intense experiments requiring sophisticated computational and statistical methods for data handling and analyses. This thesis describes recent advances of analyzing experimental data from quantitative biological studies to decipher the structural layer of encoding in human cells.

Adopting an integrative approach when possible, combining multiple sources of data, allowed us to study the influences of chromatin (Papers I and II) and chromosomal aberrations (Paper IV) on transcription. Combining chromatin data with chromosomal aberration data allowed us to identify putative driver oncogenes and tumor-suppressor genes in cancer (Paper IV).

Bayesian approaches enabling the incorporation of background information in the models and the adaptability of such models to data have been very useful. Their usages yielded accurate and narrow detection of chromosomal breakpoints in cancer (Papers III and IV) and reliable positioning of nucleosomes and their dynamics during transcriptional regulation at functionally relevant regulatory elements (Paper II).

Using massively parallel sequencing data, we explored the chromatin landscapes of human cells (Papers I and II) and concluded that there is a preferential and evolutionary conserved positioning at internal exons nearly unaffected by the transcriptional level. We also observed a strong association between certain histone modifications and the inclusion or exclusion of an exon in the mature gene transcript, suggesting a functional role in splicing.

Keywords: Chromatin, Nucleosome positioning, Histone modifications, Chromosomal aberrations, Transcriptional regulation, Array-CGH, Next generation sequencing, ChIP-chip

Robin Andersson, The Linnaeus Centre for Bioinformatics, Box 598, Uppsala University, SE-75124 Uppsala, Sweden.

© Robin Andersson 2010

ISSN 1651-6214
ISBN 978-91-554-7897-1
urn:nbn:se:uu:diva-130999 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-130999)
To Emma
“And don’t it make you feel so sad  
Don’t the blood rush to your feet  
To think that everything you do today  
Tomorrow is obsolete”

Nick Cave – More news from nowhere

Cover: A Hilbert curve displaying locations of nucleosomes (dark grey) and exons (medium grey) as well as their co-localizations (light grey) on human chromosome 16.
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the respective publishers.

† These authors contributed equally to this work.


### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC</td>
<td>bacterial artificial chromosome</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CDS</td>
<td>(protein-) coding sequence</td>
</tr>
<tr>
<td>ChIP</td>
<td>chromatin immunoprecipitation</td>
</tr>
<tr>
<td>ChIP-chip</td>
<td>ChIP measured on a microarray</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>ChIP measured with massively parallel sequencing</td>
</tr>
<tr>
<td>CNV</td>
<td>copy number variation</td>
</tr>
<tr>
<td>CTD</td>
<td>carboxy terminal domain</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>Gb</td>
<td>Giga base pairs</td>
</tr>
<tr>
<td>GBM</td>
<td>glioblastoma multiforme</td>
</tr>
<tr>
<td>GTF</td>
<td>general transcription factor</td>
</tr>
<tr>
<td>HMM</td>
<td>hidden Markov model</td>
</tr>
<tr>
<td>IP</td>
<td>immunoprecipitation</td>
</tr>
<tr>
<td>kb</td>
<td>kilo base pairs</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov chain Monte Carlo</td>
</tr>
<tr>
<td>miRNA</td>
<td>micro RNA</td>
</tr>
<tr>
<td>MnaseI</td>
<td>micrococcal nuclease I</td>
</tr>
<tr>
<td>Mnase-seq</td>
<td>MnaseI digested DNA measured with massively parallel sequencing</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>ncRNA</td>
<td>non-coding RNA</td>
</tr>
<tr>
<td>NGS</td>
<td>next-generation sequencing</td>
</tr>
<tr>
<td>pdf</td>
<td>probability distribution function</td>
</tr>
<tr>
<td>PTM</td>
<td>post-translational modification</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNAPII</td>
<td>RNA polymerase II</td>
</tr>
<tr>
<td>SMAP</td>
<td>segmental maximum a posteriori</td>
</tr>
<tr>
<td>SNR</td>
<td>signal-to-noise ratio</td>
</tr>
<tr>
<td>ssDNA</td>
<td>single-stranded DNA</td>
</tr>
<tr>
<td>SuMMIt</td>
<td>Strand-based Mixture Modeling of protein-DNA Interactions</td>
</tr>
<tr>
<td>TF</td>
<td>transcription factor</td>
</tr>
<tr>
<td>TFBS</td>
<td>transcription factor binding site</td>
</tr>
</tbody>
</table>
Background

Modern biological studies rely on data-intense experiments requiring sophisticated computational and statistical methods for data handling, quality control, annotations, modeling, hypothesis testing and generation as well as experiment design. In parallel with the burst of recent technological advances in the field of molecular and medical biology, some of which are discussed in this thesis, there is an increasing effort to deal with the massive data generated by such platforms. The shift in technology from Sanger sequencing (Sanger et al. 1977) and qualitative measurements to microarrays (Heller 2002; Stoughton 2005) to massively parallel sequencing (Metzker 2010; Park 2009; Wang et al. 2009) is accompanied by a shift from qualitative to quantitative biology with qualitative follow-ups. Intertwined in this process of development is the incorporation of bioinformatics, which has transformed from a separate discipline, focused on specialized support and data storage, to an essential part, or even driver, of modern biological research. This thesis describes recent advances of computational and statistical approaches to decipher the control of activity in human cells using experimental data from quantitative biological studies.

The cell, about 10-100 μm in diameter, is the smallest unit of life. A single cell may constitute a whole organism, e.g. bacteria, or function as the building block of a multicellular organism, e.g. humans. The activity of cells determines the functions that control life of an organism. The human cell nucleus carries nearly all information in form of deoxyribonucleic acid (DNA) molecules that encode this activity. The total DNA of a human being, which in addition to cell nuclei DNA also includes the DNA of mitochondria, defines its genome.

Through certain combinations of nucleotides, stretches of DNA become functional entities, genes, which encode the functional products of a cell. Some human genes code for the make-up of proteins through combinations of nucleotide triplets that, after transcription and translation, collectively determine the construction of proteins from amino acids. Others carry similar information but are not translated into proteins after ribonucleic acid (RNA) generation. These non-coding RNAs (ncRNA) may, however, still have functional roles in the cell. Genes cover a large proportion of the human genome, but only fractions of these genes actually code for RNA. These sub-sequences, called exons, are recognized during transcription and separated from their counterparts, introns, in a process called splicing. The
The exact combination of exons into a gene transcript influences the formation of the end product and determines its function.

A human being comprises several trillions of cells that may be categorized into over 200 types. Some cell types have the same role in a variety of organs whereas others, the majority, are restricted to individual organs where they are organized into tissues. Likewise, some gene transcripts and their corresponding proteins are produced in all cell types whereas others are tissue specific and thus never produced in the majority of cells (Wang et al. 2008a). The difference in activity between cell types may be prominent, e.g. leukocytes and neurons. Still, with a few exceptions of cell types and somatically acquired variations, e.g. small-scale chromosomal rearrangements obtained after fertilization, the same genetic material is present in all healthy cells of an organism. Hence, although the control of gene activity is encoded in the DNA sequence itself, the regulation differs between cell types. The selective regulation of gene activity is also important when the cell needs to adapt to change in environmental conditions or when it enters a new phase in its life cycle.

Gene activity is regulated at two separable layers. The DNA molecule itself – the primary layer of encoding – is locally structurally and chemically influenced by its sequential combination of nucleotides (Garvie and Wolberger 2001; Parker et al. 2009; Rhodes et al. 1996). Furthermore, cytosine residues may be chemically modified through DNA methylation. As a consequence, the resulting local signatures may enable, or disable, the binding of proteins or complexes of them with regulatory potential to the DNA. Certain proteins, called transcription factors (TFs), recognize properties of short DNA sequences where they bind and possibly, directly or indirectly, recruit parts of the transcriptional machinery to genes. Single or complexes of proteins may also bind to enhancer or silencer regulatory sequences in the DNA where they may promote or repress gene activity, respectively (Farnham 2009). At a higher level – the structural layer of encoding – gene activity is regulated through the properties of higher order DNA structure and organization. To fit into the cell nucleus, the DNA molecules are compacted into chromatin together with proteins and organized into chromosomes. The level of compaction varies between regions of DNA and thus makes certain DNA loci more or less accessible for regulatory proteins or the transcriptional machinery. Moreover, at the chromosome level, the compaction may make gene-distal regulatory regions like enhancers and silencers gene-proximal in three-dimensional space, which may influence their regulatory potential (Gelato and Fischle 2008). Cells with abnormal chromosome compaction or organization, e.g. cancer cells, may thus have perturbed regulatory activities resulting in abnormal gene activity (Crans and Sakamoto 2001).

Hence, there is a great need to decode the transcriptional regulation encoded in both layers to further our understanding of the factors that control
activity and life of a cell and, ultimately, an organism. The computational and statistical approaches described in this thesis are developed for analyses of genome-wide experimental data in order to understand the regulatory code at the structural layer of encoding. The following sections aim to provide the reader with a sufficient background in biology and bioinformatics in order to understand the aims, methods and results of the included papers in this thesis.

The sequence layer of transcriptional encoding

DNA molecules are structurally organized into double helices of two complementary strands of nucleotides (either adenine (A), guanine (G), cytosine (C), or thymine (T)) in units of pairs connected with hydrogen bonds, called base pairs (bp) (Watson and Crick 1953) whereby A pairs with T and C with G. Human cell nuclei contain 22 pairs of tightly compacted DNA molecules in form of autosomal chromosomes plus two additional sex specific chromosomes, XX in women and XY in men. All together, the total count of base pairs in the haploid human genome, i.e. from one unit of each chromosome pair, is around 3.1 billion (3.1 Giga base pairs (Gb)) (Hubbard et al. 2009).

The human genome encodes one of nature’s most complex organisms. Nevertheless, neither the genome length, nor the number of contained genes (nearly 20,000 well-characterized protein-coding genes, Pruitt et al. 2009) needs to outnumber simpler organisms, e.g. Amoeba dubia (670 Gb, Parfrey et al. 2008) and Trichomonas vaginalis (approximately 60,000 protein-coding genes, Aurrecoechea et al. 2009). Less than 2% of the human genome is exonic, i.e. with known protein-coding potential (Gregory 2005; Taft et al. 2007). Put simply, the DNA encodes more than just proteins. While there is no obvious relationship between organismal complexity and either genome length or the total length of protein-coding sequences (CDS), there is a direct relationship between complexity and the number of exons of genes, the non-CDS proportion of the genome and the intron/exon length ratio (Keren et al. 2010; Taft et al. 2007). During transcription, introns are removed and exons are spliced together to form messenger RNA (mRNA), but all exons need not to be spliced back together. The transcription of a single gene can thus result in a number of different gene transcripts. This phenomenon is referred to as alternative splicing. Out of approximately 20,000 protein-coding genes, human cells can produce more than 100,000 protein variants, called isoforms (Hubbard et al. 2009). Hence, where the number of genes may limit the functional output of a cell, nature has found other ways to orchestrate its complexity.

Not only does the activity of cells from different organisms vary. Also, the gene activity of the same cell does change in response to environmental
changes, for instance when exposed to drugs (Weake and Workman 2010). In addition, the transcriptional output between cell types within the same organism may vary to a large extent. Less than half of all human protein-coding genes are ubiquitously expressed, i.e. transcribed in all cell types, while more than half of all protein-coding genes are expressed in a single cell type (Ramskold et al. 2009). Interestingly, more than 90% of human protein-coding genes undergo alternative splicing and as much as 60% of such events are tissue-specific (Wang et al. 2008a).

Which genes and gene transcripts to be expressed in a given cell is orchestrated by a plethora of regulatory factors. There are around 1,400 different human TFs, of which only a fraction are present in each cell type (Vaquerizas et al. 2009). Moreover, ncRNAs may affect the expression of genes, although their functions are not as well categorized. The majority of ncRNA are also expressed in a tissue-specific manner (Sasaki et al. 2007), which strengthen their importance in tissue-specific gene regulation. In addition, DNA methylation may perturb the regulatory potential of DNA sequences through, for instance, the change in chemical signatures of local DNA sequences that may impede the binding of TFs (Jones and Takai 2001).

Each human cell, out of trillions, contains a complex machinery of simultaneous interactions and interventions that directs expression of genes, whose products may also play an essential role in the next round of transcriptional regulation.

Transcription

The majority of eukaryotic gene transcription is performed by an enzymatic protein complex called RNA polymerase II (RNAPII). Transcription is directed from 5' to 3' of DNA with respect to the carbon atoms in the sugar backbone of DNA on any strand, sense or antisense, of the DNA double helix. Transcription undergoes three different stages; initiation, elongation and termination (Saunders et al. 2006).

Before transcriptional initiation, a pre-initiation complex (PIC) consisting of RNAPII and, so-called, general transcription factors (GTFs) forms at the promoter, often situated at the 5' end of the target gene (Figure 1c). Gene promoters are sometimes composed of certain DNA sequence elements, such as high content of A and T (TATA boxes) and compositions of certain nucleotides that form initiator sequences (Farnham 2009). These elements may facilitate the binding of GTFs that position and stabilize RNAPII near the transcription start site (TSS) of a gene. In the transition from initiation to elongation, RNAPII is chemically modified (Saunders et al. 2006; Weake and Workman 2010). Early in the transcription, a subunit (CDK7) of the GTF TFIIH remodels the PIC through phosphorylation of the amino acid serine at the 5th position (Ser5) of the carboxy-terminal domain (CTD) of
RNAPII (*Figure 1d*), while phosphorylation of Ser2 is associated with *productive elongation*.

*Figure 1.* From transcription initiation to elongation. (a) Promoter selection is achieved through binding of activators to DNA recognition sites. (b) Activators recruit co-activator protein complexes and nucleosome-remodellers, which reposition or eject histone octamers at the promoter. (c) Jointly, the bound factors cooperate in the recruitment of general transcription factors (GTFs) and RNA polymerase II (RNAPII) to form a pre-initiation complex (PIC). (d) Promoter clearance is accomplished through phosphorylation of serine 5 (Ser5) of the carboxy-terminal domain (CTD) of RNAPII and PIC remodeling through certain subunits of the GTFs. (e) RNAPII transcribes 20-40 bps into the gene and halts at a promoter-proximal pause site and proceeds with productive elongation after appropriate stimuli such as Ser2 phosphorylation of the RNAPII CTD. During elongation nucleosome remodellers facilitate effective passage of RNAPII. The histone octamers of nucleosomes are depicted as discs with units denoting the histones H2A, H2B, H3, and H4. Adapted by permission from Macmillan Publishers Ltd: Nat Rev Genet. (Weake, V.M. and Workman J.L., 11(6):426-37), copyright (2010).
In the elongation phase, RNAPII disassociates with some of the GTFs and promoter sequence elements and transcribes 20-40 bps where it pauses, a phenomenon known as transcriptional pausing (Weake and Workman 2010). Genes at this stage are referred to as being poised with RNAPII. Only after appropriate stimulation can RNAPII proceed with productive elongation (Figure 1e) and then continues to be elongated until it reaches some termination site where a completed transcript is released (Greive and von Hippel 2005; Saunders et al. 2006). Transcriptional pausing is suggested to be an efficient way of preparing for fast gene expression in response to stimuli, so called inducible gene expression (Weake and Workman 2010).

During splicing, introns are removed from the pre-mRNA and selected exons are joined to form mRNA. These events are likely to happen during transcription while the nascent RNA is still attached to RNAPII, in a co-transcriptional manner (Allemand et al. 2008; Pandit et al. 2008). At the sequence level, the selection of exons is determined by the recognition of splice sites by the spliceosome, a protein complex of nearly 200 subunits (Jurica and Moore 2003). The skipping of an exon, the most common of many possible alternative splicing events (Koscielny et al. 2009), produces a shorter mRNA that, eventually, may change the function of the protein resulting from mRNA translation. Certain features, apart from the regulatory machinery, are known to affect the inclusion or exclusion of an exon in the mature mRNA. These include the sequence of nucleotides at splice sites (Wang and Marin 2006) and the lengths of exons and introns. Large exons are more often alternatively spliced but constitutively spliced if flanked by short introns (Sterner et al. 1996).

From a gene’s point of view, transcription is performed between distinct genomic coordinates in a chromosome corresponding to a defined individual gene and in an active/inactive manner. However, life seems more complicated than that. Firstly, observations suggest a spatial organization of active RNAPII in the nucleus (Iborra et al. 1996), a feature that has implications for the regulatory machinery and the organization of genes in a genome (Sutherland and Bickmore 2009). These so-called transcription factories are, however, far from being fully characterized. Secondly, transcription of most genes is not stable in an on/off manner but rather discontinuous with pulses of activity and intermediate periods of inactivity (Chubb et al. 2006). Furthermore, it seems that a much larger portion than what can be accounted for by characterized protein-coding genes is actually being transcribed (The ENCODE Project Consortium 2007).

Transcriptional regulation at the sequence level

Gene transcription is regulated by a series of intertwined processes within the regulatory machinery. Diverse proteins are crucial for the correct and selective transcription of genes in a cell. DNA-binding of proteins is likely
achieved through both the recognition of chemical signatures of DNA bases and a sequence-dependent DNA shape (Rohs et al. 2010), although the former features are more extensively studied. Apart from the GTFs that are directly involved in RNAPII positioning, TFs that bind DNA sequences may regulate the transcription in various ways. Transcriptional activators recruit components of the transcriptional machinery and, possibly, co-activators to promoters of genes that, together with the GTFs, recruit and position RNAPII near the TSSs (Weake and Workman 2010) (Figure 1a-c). Alternatively, certain TFs, called repressors, may bind to silencer DNA regions where they can hinder binding of activators thus preventing the RNAPII recruitment to genes (Farnham 2009). At an even more complex stage, TF binding at insulator regions may prevent enhancer activity on promoters (Burgess-Beusse et al. 2002).

The estimated number of sequence-specific TFs is around 1,400, although very few have known functions or have been experimentally verified (Vaquerizas et al. 2009). They seem to have either general, non-tissue specific, roles or specific regulatory roles needed in only one or two cell types (Vaquerizas et al. 2009). Common to sequence-specific TFs is a sub-structure of the protein, a DNA binding domain, with potential of binding to transcription factor binding sites (TFBSs) along the DNA. Nevertheless, not all TFs with such a domain actually bind to DNA (Vaquerizas et al. 2009). TFs need not work individually, rather it seems that they often bind in clusters to cooperatively regulate gene activity (Farnham 2009). They may bind close to genes, cis, although many TFBSs (35%) suggest regulation in trans, i.e. at distal loci (The ENCODE Project Consortium 2007). In fact, silencer and enhancer regions are often located far from TSSs of genes (Farnham 2009), suggesting either an indirect regulatory role through a series of regulatory events or a direct role through close interaction in three-dimensional space due to DNA looping or favorable chromatin compaction. However, one cannot rule out that seemingly distal TFBSs are in fact proximal to uncharacterized promoters in the genome (The ENCODE Project Consortium 2007). Moreover, it is hard to, at a genome wide scale, determine whether a single TF binding has a functional regulatory role or not. The DNA sequence at a certain locus in the genome may solely have the desired nucleotide composition that favors TF binding but without any regulatory potential.

Genes coding for TFs are, themselves, regulated transcriptionally or post-transcriptionally. Small RNAs (20-30 nucleotides in size), e.g. micro RNAs (miRNA), may regulate transcription in a post-transcriptional manner, in which, for instance, they bind to mRNA through base-pairing and degrade the gene transcript or repress translation into protein (Kim et al. 2009). At least 15% of all genes have been found to undergo post-transcriptional regulation (Nikolaev et al. 2009), suggesting important roles of miRNA in cell division, apoptosis, i.e. programmed cell death, and differentiation (He
and Hannon 2004). Interestingly, at least in moss, miRNAs may also silence gene expression through interaction with DNA leading to DNA methylation (Khraiwesh et al. 2010).

DNA methylation adds to the complex machinery of transcriptional regulation. Through methylation of cytosine residues in *CpG* dinucleotides, i.e. cytosine followed by guanine in the DNA strand and attached together with phosphate, the chemically modified DNA may disfavor binding of TFs (Jones and Takai 2001; Watt and Molloy 1988). Additionally, DNA methylation of *CpG islands*, i.e. regions with high CG-content including sequences of *CpG* dinucleotides that are rarely methylated (Weber et al. 2007), in gene regulatory regions is associated with repression of gene activity (Cedar and Bergman 2009). Once a *CpG* dinucleotide is methylated it rarely ever gets demethylated. Rather, it likely stays methylated and even propagates its status to daughter cells after cell division, a phenomenon important during cell differentiation where developmental genes need to be silenced (Futscher et al. 2002; Wigler et al. 1981).

The structural layer of transcriptional encoding

Almost every cell in a human body contains more than 6 billion base pairs, summing up the two haploid genomes consisting of 23 chromosome pairs. Each bp is of length 3.4 Å (Watson and Crick 1953) yielding a total DNA length of around 2 m if stretched out. Since the cell nucleus is of limiting size, around 5 μm in diameter, the DNA molecules of eukaryotes are organized into various states of compaction, known as chromatin.

Apart from mere structural organization, the chromatin state serves as a transcriptional regulator at the structural layer. The magnitude and location of compaction will alter the structural conformation of DNA that affects the distance between genomic loci in three-dimensional space (Gelato and Fischle 2008). Indirectly, this may influence the regulatory capabilities of distal elements, such as enhancers, to target genes. Moreover, through dense or loose compaction the DNA becomes inaccessible or accessible, respectively, for regulators, e.g. TFs, and components of the transcriptional machinery, e.g. GTFs and RNAPII (Campos and Reinberg 2009; Farnham 2009).

At its highest level, chromatin is organized into chromosomes. However, the local level of DNA compaction varies between different chromosomal regions and is an indicator of loci with, for instance, transcribed genes (Gelato and Fischle 2008). The chromosomal organization will also affect the regulatory machinery through, for instance, chromatin-directed localization of chromosome territories in the cell nucleus (Bartova et al. 2008). The adjacent and correct localization of genes and their regulatory
elements in chromosomes and higher order chromatin is also essential for controlling gene activity.

When such organization is disturbed, e.g. through incapability of forming higher order chromatin structure (Gupta et al. 2008) or failure of repairing DNA breakage in dense chromatin regions (Cohn and D'Andrea 2008), the cell may either die or survive with altered transcriptional or regulatory activity leading to abnormal cell behavior. The latter event is common in various cancer forms.

DNA organization and compaction

Eukaryotic DNA compaction is, at its basic level, achieved through winding of DNA around an octamer of histone proteins, into units called nucleosomes (Figure 1 and Figure 2). Each histone octamer is, usually, composed of two H3-H4 dimers forming a tetramer with flanking H2A-H2B dimers around which the DNA wounds approximately 1.7 turns or 147 bp with a radius of 41.9 Å (Campos and Reinberg 2009; Richmond and Davey 2003). Histone variants do sometimes replace the canonical ones, which can result in altered compaction and structural changes and may be associated with functional properties (Talbert and Henikoff 2010). At this basic level of compaction, nucleosomes form repeating units like beads on a string. If coupled with linker histone H1, through surface properties of nucleosomes and inter-nucleosomal interactions, this primary structure of chromatin may conform into a fiber of 30 nm, the secondary structure, and into higher order chromatin structure, the tertiary structure (Tremethick 2007; Zhou et al. 2007) (Figure 2). The conformational nature of chromatin beyond the primary structure is, however, to date not fully characterized (Chien and van Noort 2009). Various factors will influence the organization of DNA into chromatin, as discussed below, making general consensus structures less likely. Chromatin is not uniformly organized in the cell nuclei. Rather, distinct chromatin regions with low condensation, euchromatin, and high condensation, heterochromatin, is present in the cell nucleus. The state of higher order chromatin is also regulated by incorporation of histone variants (Jin and Felsenfeld 2007; Talbert and Henikoff 2010) and chemical modifications of the histone tails (Campos and Reinberg 2009). The resulting chromatin state will subsequently influence DNA transcription, replication, recombination and repair (Gelato and Fischle 2008; Margueron and Reinberg 2010).

Hence, the nucleosome is a key player in regulation and cellular activity. The CTDs of histones, i.e. their “tails”, are rich in basic amino acid residues that are subject to different chemical modifications (Figure 2), such as acetylation, methylation, ubiquitination, sumoylation and phosphorylation (Kouzarides 2007). These post-translational modifications (PTMs), extensively studied during the last years, may serve as marks signaling for
inducing or preventing interaction with other partners (Taverna et al. 2007; Zhou et al. 2007), chromatin remodeling (Workman 2006), chromosomal relocation (Bartova et al. 2008) and gene activity (Campos and Reinberg 2009). Although attempts have been made to postulate a “histone code” to describe the functionality of individual histone modifications (Strahl and Allis 2000), or combinations of them (Wang et al. 2008b) their generality have been questioned (Sims and Reinberg 2008).

**Figure 2.** Factors affecting chromatin organization, localization, compaction and transcription. The structure of chromatin is determined through the effective wounding of DNA around histone octamers and formation into higher order structures. Nucleosome wounding and fiber formation are further affected by the incorporation of histone variants, posttranslational modifications (PTMs) of histone tails, such as phosphorylations (P), methylations (Me) and acetylations (Ac), methylation of CpG residues in DNA and intervening structural RNAs. The resulting chromatin organization, its nuclear localization, the histone modifications and chromatin-binding proteins compose a structural layer of transcriptional regulation. Adapted by permission from Macmillan Publishers Ltd: Nat Rev Mol Cell Biol. (Probst, A.V., et al., 10(3):192-206), copyright (2009).
Certain histone modifications affect the nuclear architecture of chromosomes and chromosome territories (Bartova et al. 2008), with territories rich in genes or with high gene activity located closer to the nuclear interior than territories with few or inactive genes (Croft et al. 1999; Williams et al. 2006). Nuclear interior localization, possibly at transcription factories, may be achieved through PTM-directed decondensation of chromatin and the formation of chromatin loops (Chambeyron and Bickmore 2004). DNA methylation is another factor that affects nuclear localization of chromatin (Bartova et al. 2008).

Not only is chromatin affected by the modifications on nucleosomes, the placement of nucleosomes along DNA may enable or disable higher order organization. Among many possible factors, the chemical and structural signature of local DNA (Garvie and Wolberger 2001; Parker et al. 2009; Rhodes et al. 1996; Rohs et al. 2010), the incorporation of histone variants (Jin and Felsenfeld 2007) and the competitive binding of TFs or other factors to DNA (Segal and Widom 2009) will affect their location (Figure 2). Moreover, during transcription and transcriptional regulation, nucleosomes are subject to remodeling, repositioning and eviction by chromatin remodelers (Figure 1) (Cairns 2009; Workman 2006).

Transcriptional regulation at the structural level
The fundamental concept of chromatin-directed gene silencing is that RNAPII cannot access heterochromatic regions. Inter-nucleosomal interactions are vital for formation of heterochromatin. Contacts between adjacent nucleosomes are made through direct interactions between lysine residues 16 to 20 on histone H4 (H4K16-20) and histone H2A on the interacting partner (Zhou et al. 2007). Acetylation of the lysine 16 residue (H4K16ac) impedes this interaction (Shogren-Knaak et al. 2006) which thus provides a mean to hinder chromatin compaction. Since females carry two copies of the X chromosome, only one is active in the cell nucleus. This is achieved through almost complete transcriptionally silent chromatin, indeed with the absence of acetylation on histone tails. In addition, it is marked by extensive DNA methylation, several PTMs of histones, such as trimethylation of lysine 27 and dimethylation of lysine 9 on histone H3 and monomethylation of lysine 20 on histone H4 (H3K27me3, H3K9me2, H4K20me1) (Brinkman et al. 2006; Reik and Lewis 2005). Chromatin silencing may also be mediated through intervening ncRNAs (Whitehead et al. 2009), e.g. XIST on the silent X chromosome.

In contrast, euchromatic regions are easier to transcribe. Nevertheless, such regions are not nucleosome-free. Rather, there is higher nucleosome occupancy within intragenic regions, i.e. within gene boundaries, than in intergenic regions, i.e. between genes (Campos and Reinberg 2009). Moreover, the observation that a large fraction of the human genome is
subject to transcription (The ENCODE Project Consortium 2007) suggests that much transcription is done through chromatin. However, there is a high association of less stable histone variants (Jin and Felsenfeld 2007) at genes (Barski et al. 2007), having less ability to interact with adjacent nucleosomes (Campos and Reinberg 2009) thus hindering the formation of higher order chromatin. In fact, the average nucleosomal landscape around and within protein-coding genes follows a characteristic pattern (Jiang and Pugh 2009; Schones et al. 2008). This pattern is characterized by one well-positioned nucleosome upstream, i.e. in the 5’ direction, of the TSS followed by a nucleosome-free region (NFR) and then the TSS. Immediately downstream, i.e. in the 3’ direction, of the TSS follows another well-positioned nucleosome. The transcription end site is also associated with a NFR. These nucleosome-associated loci are referred to as the -1 nucleosome, 5’ NFR, +1 nucleosome and 3’ NFR. The positions of the +2 to +5 nucleosomes are characterized by decreasing concordance among genes and cells. Hence, a gene is defined by more than just its genomic boundary.

The nucleosomal patterns of genes suggest possibilities of transcriptional regulation by nucleosomal placement apart from being a building block of chromatin structure. Firstly, histone PTMs around and within genes do relate with transcriptional activity (Barski et al. 2007; Campos and Reinberg 2009; Wang et al. 2008b). Individual histone modifications, such as H3K4me3 at TSSs, are associated with genes poised with RNAPII, whereas others, such as H3K36me3 and H3K27me3 at intragenic regions, are associated with transcribed and silent genes, respectively (Barski et al. 2007). However, single PTMs alone need not be predictive of transcriptional status and may in fact be associated with conflicting biological processes (Campos and Reinberg 2009) possibly suggesting multiple roles in a combinatorial manner (Wang et al. 2008b). Combinations of activating (H3K4me3) and inactivating (H3K27me3) PTMs, bivalency, have also been observed and associated with cell differentiation (Bernstein et al. 2006).

Secondly, histone modifications may enable or impede binding of chromatin-interacting proteins, which can have roles in chromatin formation or transcriptional regulation (Campos and Reinberg 2009; Gelato and Fischle 2008; Taverna et al. 2007). For instance, HP1 binding to H3K79me3 is suggested to promote condensation of chromatin (Thiru et al. 2004), whereas methylated H3 prevents DNA methylation through hindrance of binding by the DNA methyltransferase DNMT3L (Ooi et al. 2007). Thirdly, the mere placement of nucleosomes at genic loci can have a role in both transcription and transcriptional regulation (Jiang and Pugh 2009). During transcription, nucleosomes are temporarily displaced or ejected to enable efficient RNAPII progression (Figure 1) (Cairns 2009; Richmond and Davey 2003; Saunders et al. 2006; Weake and Workman 2010). A well-positioned +1 nucleosome may play a role in RNAPII pausing (Cairns 2009). Their occupancy within genes may also regulate transcription between cycles. Phosphorylation of
RNAPII Ser2 helps recruit the methyltransferase SET2 to RNAPII, which methylates H3K36 within transcribed genes (Kizer et al. 2005). H3K36me, in turn, has been suggested to promote deacetylation of histones, hindering transcription, between transcription cycles (Joshi and Struhl 2005). In the cell, nucleosomes and other DNA-binding proteins competitively bind to the DNA (Segal and Widom 2009). The binding of TFs is thus affected by the placement and stability of nucleosomes along DNA and enzymatic activities that modify, reposition, reconfigure or eject nucleosomes. Stably positioned nucleosomes will block the binding of some TFs without the involvement of other partners, such as histone modifiers or coactivators (Cairns 2009; Weake and Workman 2010). Since histone variants make nucleosomes less stable, such TFs are more likely to accomplish binding at important regions that often present those variants, e.g. enhancers and promoters (Barski et al. 2007). Other TFs, such as FOXA2, have been shown to directly interact with chromatin (Cirillo et al. 2002).

The regulation of transcription through chromatin adds to the already complex network of factors that competitively interact, promote and repress gene activity both directly and indirectly.

Abnormalities in transcriptional encoding

Abnormalities that perturb the complex regulatory or transcriptional machinery of a cell may eventually, if not hindered by one of many safety mechanisms, lead to genetic disorders such as cancer, which will be the focus of this section. Cancer is caused by alterations that disturb the balance in cell proliferation, survival and differentiation. It is associated with disrupted or abnormal effect of important genes controlling apoptosis, cell growth or DNA repair. Oncogenes are genes that, in one way or another, stimulate cell growth and persistence of the tumor through avoidance of apoptosis. These are derived from proto-oncogenes that are normal, non-cancer, genes that have gained abnormal behavior such as over-expression. Their counterparts, tumor-suppressor genes, do instead serve to impede cancerous behaviors through DNA damage repair, repression of proliferation and programmed cell death, i.e. apoptosis. Many of cancer-associated genes are altered across several cancer types (Futreal et al. 2004).

The aberrant behavior of oncogenes or tumor-suppressor genes may be due to a variety of reasons. Although the initiating factors are to a large degree unknown (Frohling and Dohner 2008), tumors are strongly associated with genetic alterations, such as large-scale chromosomal alterations (Albertson et al. 2003; Frohling and Dohner 2008) or point mutations (Futreal et al. 2004; Stratton et al. 2009), or chromatin abnormalities, such as disrupted DNA methylation landscapes or aberrant histone modifications (Esteller 2008). Oncogenes, for instance, may be acquired through point
mutations, i.e. substitution of one bp. More than 1% of human protein-coding genes are related to cancer via such *driver mutations*, among which 90% are acquired somatically and 20% are heritable (http://www.sanger.ac.uk/genetics/CGP/Census/) (Futreal et al. 2004). Furthermore, studies have indicated that the expression of many genes, more than 12%, is affected by chromosomal abnormalities in certain cancers (Pollack et al. 2002).

Cancer cells may have disrupted chromatin structure (Esteller 2008; Thorne et al. 2009). Abnormal patterns of PTMs may silence genes with tumor-suppressor like properties (Ke et al. 2009; Kondo et al. 2008; Richon et al. 2000). Loss of certain histone modifications at loci are associated with high risk of recurrence of prostate cancer (Seligson et al. 2005). The patterns may also differ at a more global level. Expression of histone modifying enzymes in cancer tissues is often different than in their normal counterpart and varies between cancers (Hamamoto et al. 2004; Ke et al. 2009; Ozdag et al. 2006; Simon and Lange 2008). The methyltransferase EZH2, that specifically trimethylates H3K27 (Kirmizis et al. 2004), is over-expressed in several cancers (Ke et al. 2009; Simon and Lange 2008). Likewise, over-expression of the H3K4-specific methyltransferase SMYD3 has been observed in various cancer cells (Hamamoto et al. 2004; Ke et al. 2009), suggesting H3K4 methylation at promoters of oncogenes. Interestingly, pre-marking by H3K27me3 of *de novo* DNA methylated genes has been suggested in colon cancer (Schlesinger et al. 2007).

The aberrant methylation patterns of DNA in cancer cells have been more extensively studied than their histone counterparts. Tumors are often associated with global DNA hypomethylation, i.e. low levels of DNA methylation (Esteller 2008), and the degree may increase in tumor progression from benign to malignant cancers (Fraga et al. 2004). In addition, cancers may present DNA hypermethylation, i.e. high levels of DNA methylation, of CpG islands in tumor-suppressor genes (Greger et al. 1989). Moreover, DNA methylation may inactivate the expression of miRNA genes (Lujambio et al. 2007). Although their causes are often unknown (Esteller 2008), their locations are often specific to the cancer type (Costello et al. 2000). DNA hypomethylation may mechanistically contribute to tumor development in various ways. DNA hypomethylation may promote chromosomal instabilities through, for instance, reactivation of *transposable elements*, i.e. DNA sequences that may relocate in the genome (Bestor 2005). In addition, since DNA methylation is associated with *imprinting*, i.e. selective expression of genes from one paternal allele, its loss may disrupt this control causing abnormal expression of growth-controlling genes (Feinberg 1999).
Chromosomal abnormalities are subdivided into balanced rearrangements and chromosomal imbalances. Balanced chromosomal rearrangements result in the formation of chimeric fusion genes with new or altered expression or in the deregulated expression of structurally normal genes. Chromosomal imbalances include gains and losses of genomic DNA, ranging from large-scale imbalances possibly affecting whole chromosomes (trisomies and monosomies) to small focal amplifications or deletions. Adapted from Fröhling, S. and Dohner, H., N Engl J Med. 359:722-734. Copyright © 2008 Massachusetts Medical Society. All rights reserved.

Chromosomal alterations are found in all major tumor types and may be associated with early events in tumorigenesis, i.e. tumor development (Albertson et al. 2003; Fröhling and Dohner 2008). More than 58,000 cases have, to date, been reported (Mitelman et al. 2010). The genome-wide patterns of alterations may be associated with specific tumor types, but a large majority of individual alterations and the aberrant behaviors of contained genes are present in several cancer types (Beroukhim et al. 2010). Although the initiating factor is often unknown, chromosomal alterations may result from perturbed regulation of damaged DNA caused by, for instance, environmental or occupational factors (Frohling and Dohner 2008). Chromosomal alterations are subdivided into balanced rearrangements and chromosomal imbalances (Figure 3).

A balanced chromosomal rearrangement may result in the formation of a chimeric fusion gene, when parts of two genes are fused together in the genome, with new or altered activity. It may also result in the juxtaposition of a gene regulatory element to a structurally normal gene leading to deregulated expression (Figure 3) (Frohling and Dohner 2008). The Philadelphia chromosome, for instance, is present in nearly all patients with
chronic myeloid leukemia and is a result of a translocation that forms a fusion gene with aberrant activity (Goldman and Melo 2003). Fusions may also perturb TF genes acquiring enhanced or aberrant transcriptional or regulatory activities. Juxtaposition of regulatory elements to proto-oncogenes may have critical consequences for tumorigenesis, causing, for instance, deregulated expression of oncogenes or over-expression of TFs (Frohling and Dohner 2008).

Unlike balanced chromosomal rearrangements, the functional consequences of chromosome imbalances are often unknown. They are categorized into genomic gains and losses and may be of varying sizes (Figure 3). Most cancers have many and large gains or losses. In fact, large alterations are approximately 30 times more common than focal alterations (Beroukhim et al. 2010), although small focal amplifications are, for instance, often observed in lung cancers (Weir et al. 2007). Losses probably contribute to tumor development by silenced or reduced function of contained genes, while gains may contribute to tumorigenesis by promoting the activity of them (Frohling and Dohner 2008). Changes in DNA copy number, i.e. divergence from the normal two-allele state of diploidy, may also affect contained miRNA genes, whose expressions may differ between cancer types (Lu et al. 2005).

Genomic variations are not restricted to genetic disorders such as cancers. In fact, a profound portion, up to 12 %, of the human genome is variable between individuals displaying common or rare copy number variations (CNVs) (Redon et al. 2006; Shaikh et al. 2009). CNVs can arise either meiotically, i.e. during sexual reproduction, or somatically indicated by CNV differences in monozygotic twin pairs (Bruder et al. 2008) and between tissue types (Piotrowski et al. 2008). CNVs, like large-scale chromosomal alterations, may greatly impact the expression of affected genes (Hastings et al. 2009). Such changes often have negative consequences indicated by some CNVs attributed to susceptibility to disease (Feuk et al. 2006; Redon et al. 2006; Sebat et al. 2004). Examples do, however, exist where CNVs might have positive consequences affecting resistance to malaria or susceptibility to HIV/AIDS (Hastings et al. 2009). Nevertheless, systematic cataloguing of non-disease associated normal variation through CNVs is essential to accurately avoid erroneously associated chromosomal alterations with cancers (de Stahl et al. 2008).

Decoding transcriptional regulation

To investigate the functionality of transcriptional regulation in cells, it is important to study regulators or regulatory factors at a whole-genome scale. More than 50% of human genes have alternative promoters (Kimura et al. 2006) and many TFBSs, 35%, are distal to them (The ENCODE Project
Consortium 2007). Hence, if focus is solely on well-defined promoters proximal to well established protein-coding genes one will most likely miss a large fraction of regulatory elements. Likewise, as discussed above, the functional consequences of a majority of tumor-associated alterations are unknown. It is likely that more than mere gene copy deviations will affect the transcriptional and regulatory activity of a cell. In the same manner, to readily annotate and associate chromatin landscapes, i.e. nucleosomal locations and their modifications, with regulation, the genome-wide approach is the only way to go. Ideally, regulatory landscapes and their interacting networks should be studied in an integrative manner, with multiple sources of measurements in the same cells, to fully assess the implications of individual regulators and responses to external stimuli or internal aberrations. Moreover, since a lot of regulatory and transcriptional activities are cell type specific, individual observations in one cell type cannot always be generally assumed in other types.

Genome-wide studies rely heavily on instrumentations that produce massive data sets requiring sophisticated computational and statistical methods for data handling, quality control, annotations, modeling, hypothesis testing and generation as well as experiment design. The following sections summarize state-of-the-art experimental techniques for studying transcriptional regulation, the data produced as well as approaches in bioinformatics to analyze such challenging data.

Regulatory landscapes

The regulation of gene transcription is determined by the locations and statuses of regulatory elements, e.g. TF binding to gene-proximal promoters or distal enhancers and histone modifications of TSS-proximal nucleosomes. How and where such events occur in the genome have implications for how cells respond to stimuli or how erroneous regulation lead to transcriptional defects. Hence, to decipher the regulatory networks of a cell it is crucial to first investigate the positional landscapes of regulators in its genome.

Measuring protein-DNA interactions

Chromatin immunoprecipitation (ChIP) (Buck and Lieb 2004) is a well-established protocol for identification of protein-DNA interactions, such as TF binding or positions of nucleosomes with certain PTMs, in the genome. In summary (Figure 4), cells are first fixed with formaldehyde that crosslinks, i.e. binds, proteins to each other and proteins to DNA. The DNA is then sheared by enzymatic cleavage or sonication, a process in which the DNA is fragmented by ultrasound at specific wave lengths, to lengths of a hundred to a couple of hundred bp. To gather the sheared DNA fragments of interest, i.e. those that are bound by a specific protein, the fragments are enriched by immunoprecipitation (IP) with a protein-specific antibody. The
crosslinks are then reversed and the enriched DNA fragments are purified. ChIP may be performed at targeted genomic loci or can be coupled to whole-genome experiments, in which the fragments are subsequently hybridized on, one or more, DNA microarrays (ChIP-chip) (Buck and Lieb 2004) or subjected to massively parallel sequencing (ChIP-seq) (Park 2009) (Figure 4).

Figure 4. Overview of ChIP-chip and ChIP-seq procedures. In summary, cells are fixed with formaldehyde from which chromatin is isolated and sheared, by sonication or enzymatic activity, to fragments of dedicated lengths. The fragments of interest are enriched by immunoprecipitation (IP) with a protein-specific antibody and purified. Subsequently the IP fragments and input (reference sample) are either labeled with fluorescent dyes and hybridized to microarrays (ChIP-chip) or directly sequenced on a massively parallel sequencer (ChIP-seq). Reprinted by permission from Macmillan Publishers Ltd: Nat Rev Genet. (Farnham, P.J., 10(9):605-16), copyright (2009).
When measuring nucleosome positioning in a genome, not caring about certain PTMs of histones, the IP step is not performed after digestion of fragments by micrococcal nuclease I (MnaseI). The subsequent experimental steps, microarray hybridization or massively parallel sequencing (Mnase-seq), are though similar. Below the procedures of ChIP-chip and ChIP-seq are summarized.

DNA microarrays are small glass plates or silicon chips that contain up to millions of probes, each composed of complementary DNA (cDNA) or an oligonucleotide designated to measure a certain DNA sequence in the genome (Johnson et al. 2008; Stoughton 2005). Depending on the provided density and resolution given by the length of each probe, which can vary from tens to hundreds of bp depending on the array type and manufacturer (Johnson et al. 2008; Park 2009), and the genomic probe-to-probe distance, one or multiple microarrays may be used to cover the whole genome. Alternatively, arrays can be used to measure certain features in the genome such as promoters. The resolution is further influenced by whether probes map individual regions in the genome that overlap or not. Individual regions can be covered by one probe or multiple probes in replicate. If only one probe is used for each genomic region, replicate arrays may be used to detect and avoid technical artifacts.

In a standard ChIP-chip experiment the IP fragments and the input, as control, are labeled with different fluorescent dyes and hybridized to microarrays either competitively or on different arrays (Buck and Lieb 2004; Kim and Ren 2006). During hybridization, single-stranded DNA (ssDNA) fragments are attached to matching probes on the microarray. The binding of fragments is measured by scanning the fluorescence of the used dyes with laser beams of appropriate wavelengths. Probes that correspond to the genomic sites where the studied proteins are bound are identified as those with stronger quantified fluorescence signal of the IP DNA than the control, indicated by their logarithmic ratio being greater than zero (Figure 5).

In a ChIP-seq experiment, the IP fragments of interests are sequenced directly instead of being hybridized on a microarray. Input is often sequenced as well and provides a background of non-specific enrichment, which can be used when detecting binding sites. Alternatively, mock ChIP experiments with unspecific antibodies may be used for control. Massively parallel sequencing, also referred to as next-generation sequencing (NGS) (Metzker 2010; Park 2009), where automated Sanger sequencing is the first-generation technology, is a recent alternative to microarray hybridization of ChIP fragments. Apart from sequencing of ChIP or MnaseI fragments, NGS has been applied in many areas, including whole-genome sequencing (Ley et al. 2008; Wheeler et al. 2008), mRNA expression profiling (RNA-seq) (Cloonan et al. 2008; Ramskold et al. 2009; Sultan et al. 2008), characterization of sequence and structural variation (Korbel et al. 2007; McKernan et al. 2009) and profiling of chromosomal rearrangements and breakpoints (Chen et al. 2010; Schweiger et al. 2009). Although the
chemistry and techniques differ between available NGS platforms, the basics are similar (Metzker 2010; Park 2009). Common adaptors are attached to the DNA fragments to be sequenced after which the fragments are converted to ssDNA, templates, placed on beads or a glass slide and then amplified to millions of fragments. Single bases or oligonucleotides are added to each template in parallel leading to the generation of a new strand through enzymatic activity. The identity of the base, or the first two bases, is determined through high-resolution imaging of incorporated fluorescent labels. A single experiment run can result in several hundred millions of reads, i.e. sequences, of tens to hundreds of nucleotides in length identified from terabytes of image data (Horner et al. 2010; Metzker 2010). These reads are subsequently aligned, i.e. mapped, to a reference genome yielding discrete positions for the majority of sequences. Hence, in contrast to microarrays, NGS results in non-continuous data.

ChIP-chip and ChIP-seq are currently the two main technologies for genome-wide identification of protein-DNA interactions. However, ChIP-seq has a number of advantages over ChIP-chip. Firstly, genome-wide studies require the capability to map features all over the genome. While sequencing relies on the ability to align reads to a reference genome that may diverge from the genome being studied, microarrays often represent only a fraction of the total genome. Secondly, the cost for whole-genome arrays with high resolution is huge and often greater than the cost of individual runs at sequencing centers (Hoffman and Jones 2009). Thirdly, ChIP-seq provides less obtrusive artifacts, lower signal-to-noise ratio (SNR) and larger dynamic range than ChIP on microarrays (Hoffman and Jones 2009; Park 2009). SNR is the amount of background, noise, in the data, while dynamic range refers to the ratio between the smallest and largest values in the resulting data. Finally and perhaps most importantly, NGS provides higher resolution than microarrays, an obvious advantage for accurate positioning of proteins along DNA. However, since NGS is a novel technology, ChIP-chip may be considered advantageous with more and well-developed methods for handling, pre-processing and post-processing of the resulting data. Despite these considerations the most limiting factor of both ChIP-chip and ChIP-seq is the low availability of specific antibodies (Hoffman and Jones 2009; Kim and Ren 2006).

Computational analyses of protein-DNA interactions

A typical microarray experiment of ChIP or MnaseI fragments results in a huge data set with continuous signals, the cardinality determined by the density, i.e. the number of probes on the microarray and the number of arrays used. Artifacts inherent in the use of microarrays, such as cross-hybridizations, i.e. partial bindings between multiple molecules to probes, sequence composition biases of oligonucleotides and spatial biases on the arrays need to be handled before further data processing is made. This is achieved through appropriate data normalization and filtering of probes with
suspicious signals (Royce et al. 2005; Smyth and Speed 2003; van de Wiel et al. 2010), details not covered here. The ultimate goal of ChIP-chip (and ChIP-seq) experiments is to, accurately, identify regions of DNA with bound proteins possibly with specific modifications. Computational and statistical methods with this aim are often referred to as *peak finders* or *peak callers*. The word “peak” in their name refers to the distribution of probe signals around the genomic coordinates of bound proteins forming a peak (*Figure 5*). Several methods have been proposed to this end. Approaches such as sliding windows, which averages the signals of tiling probes over genomic regions or a quantity of probes, and *hidden Markov models* (HMMs) are used (Ji and Wong 2005; Johnson et al. 2006; Li et al. 2005). Regions with statistically determined significant enrichment of DNA fragments, often in relation to control DNA, are subsequently inferred. Alternatively, bound regions may be inferred by statistical significance of individual probes, based on global signal ranking, and the requirement of sufficiently many called probes in its genomic neighborhood (Rada-Iglesias et al. 2005). Called regions of mock IP, if applied, may be used to filter suspicious regions.

*Figure 5.* H3K4me3 signal around the transcription start site of gene *DOK5*. Vertical staples indicate normalized and replicate-averaged log₂-ratios between H3K4me3 and input of probes (left vertical axis) at their corresponding genomic coordinates. Black staples depict probes in inferred regions of H3K4me3 after thresholding (above 6) of Z-scores (dashed line, right vertical axis) derived from a sliding window approach (Paper IV). The locations of gene transcripts associated through the UCSC genome browser database to the *DOK5* gene are shown.
The resulting data from ChIP-seq or Mnase-seq (Figure 6) is different in nature from that resulting from their microarray counterparts, hence requiring different approaches for identifying regions of protein-DNA interactions. The first step when analyzing such sequencing data is to map the sequenced reads from either ends of the DNA fragments to a reference genome (Figure 6a), e.g. a curated sequence of the human genome obtained from the National Center for Biotechnology Information (NCBI). This is actually one of the most computationally intensive and challenging tasks in NGS analysis and many tools exist to this end (Horner et al. 2010). Up to hundreds of millions of sequences need to be aligned to genome sizes of Gb in an accurate and flexible manner. Once mapping is performed, the following data analyses often include visualization of the data around genomic features (Enroth et al. 2010), such as TSSs, and systematic identification and modeling of bound regions, i.e. peak calling (Hoffman and Jones 2009; Pepke et al. 2009).

Like with DNA microarray data, peak calling using NGS data involves the identification of regions with high enrichment (Figure 6b) relative to a background using some criteria to separate true signals from noise. A profound fraction of sequenced reads may in fact be non-specific DNA fragments (Pepke et al. 2009), which need to be considered by peak finders. Control data may be used within this process or for subsequent filtering of called regions. Alternatively, background may be modeled directly in the data. Various peak callers have been proposed and range from simple user-specified to background-derived thresholding of aggregate reads (Schones et al. 2008), overlapping strand-directed extended reads (Figure 6c) (Robertson et al. 2007) to aggregations of user-specified or empirically derived strand-directed shifts (Figure 6d) (Boyle et al. 2008; Fejes et al. 2008; Johnson et al. 2007; Jothi et al. 2008; Kharchenko et al. 2008; Valouev et al. 2008b; Zhang et al. 2008). Combinations of these methods do exist (Rozowsky et al. 2009; Tuteja et al. 2009; Zang et al. 2009). The data may be further transformed using Gaussian kernel density estimations (Boyle et al. 2008; Valouev et al. 2008b). Strand-directed transformations are either done by extending reads aligned to the sense or antisense strands in sense or antisense direction, respectively, or by strand-directed shifts often determined by the average distance between sense and antisense reads. Regions with sufficient reads above a user-specified or background-guided threshold are then called. Although various techniques have been applied for positioning, the majority implies loose criteria for positioning, e.g. through sufficient aggregation of strand-directed extensions not requiring convincing support from both boundaries of a positioned protein or protein complex. To readily annotate regulatory landscapes, careful consideration of all information given in the data will be required. In Paper II, we proposed a statistical framework for positioning that calculates odds of true interactions.
against background noise (*Figure 6e*). In this approach, support from both ends of sequenced fragments was required for positioning.

*Figure 6*. Characteristics of Mnase-seq data around nucleosomal regions. (a) Reads aligned to the sense (blue) and antisense (red) strands, depicted by line-extended dots indicating their genomic start positions and covered bases, define the genomic boundaries of a nucleosome. (b) Counts of read starts, i.e. the dots in (a). (c) Counts per bp of the number of covering reads aligned to sense (blue) and antisense (red) strands and from both strands after strand-directed extension from read starts (b) to 147 bp (black). (d) Counts per bp of the number of covering reads after a strand-directed shift of reads (black) by half the average distance between sense (blue) and antisense (red) aligned reads, indicated by sense- and antisense-directed arrows. (e) Log-odds of nucleosome positioning against background calculated using SuMMIIt (Paper II). Nucleosomal locations are derived from log-odds values above 0 (dashed red horizontal line). The distances between adjacent ticks on the horizontal axes are 200 bp.
Background information given by control data in form of input DNA or mock IPs can be accounted for through mere subtraction from ChIP data, relative enrichment fractions or post-filtering of inferred regions (Pepke et al. 2009). Alternatively, a possibly more powerful yet less explored approach could be to model the control data in the ChIP data and remove inferred noise through regression-based normalization (Enroth et al., manuscript).

Positional analysis of DNA microarray or NGS data have greatly added to the knowledge of transcriptional regulation with, for instance, complex regulatory landscapes of TFs (Farnham 2009), distinct nucleosomal patterns around genes (Jiang and Pugh 2009) and regulatory elements (Segal and Widom 2009), as well as association of histone marks with genes and their activity (Barski et al. 2007; Campos and Reinberg 2009; Kim and Ren 2006; Mikkelsen et al. 2007; The ENCODE Project Consortium 2007; Wang et al. 2008b), enhancers (Barski et al. 2007; Heintzman et al. 2009), ncRNAs (Guttman et al. 2009) and development (Bernstein et al. 2006). Furthermore, the characteristics of inferred nucleosome positions at individual genomic loci or general genomic features have been investigated. Studies have indicated loci containing phased or fuzzy positioned nucleosomes, reflecting high or low agreement of positioning among measured cells at these loci, respectively, and more or less well-positioned nucleosomes around certain genomic features, such as TSSs, reflecting the consistency of positioning among features (Jiang and Pugh 2009; Mavrich et al. 2008; Yuan et al. 2005).

Downstream analyses of positional data often include integration with gene expression data, measured on microarrays or with RNA-seq, aiming at coupling regulatory events with certain genes. Although different strategies have been proposed (Hoffman and Jones 2009), in nearly all approaches an event is coupled to its nearest gene, or genes within a predefined distance. However, as earlier discussed, it is hard to associate a regulatory element to a specific gene, the nearest gene may not be the regulated one and regulation may be performed in an indirect manner through several collaborating partners. High-throughput extensions to chromosome conformation capture (3C) (Dostie et al. 2006), a technique that map physical interactions between genomic elements, may help deciphering distal or combinatorial interactions.

Another area of post-calling analysis that has received a lot of attention is the investigation and subsequent modeling of sequence-directed protein-DNA binding. For TFs, this includes the screening of short TFBS sequences and their generalization into consensus motifs (Farnham 2009; Segal and Widom 2009). Investigations of sequence-directed positioning of nucleosomes have revealed periodicities of dinucleotides and other favoring sequence characteristics along and on the borders of histone-DNA interactions as well as disfavoring sequences, thoroughly summarized by Jiang et al. (2009) and Segal et al. (2009). Such characteristics are thought to reflect both the rotational settings, i.e. the local orientation of the DNA helix on the histone surface, and the translational settings, i.e. the DNA
midpoint position relative to a genomic coordinate, of nucleosomes (Jiang and Pugh 2009). Interestingly, it has been suggested that intron/exon junctions contain DNA sequences that promote nucleosome positioning (Baldi et al. 1996; Beckmann and Trifonov 1991), proposed to protect splice sites from mutations (Kogan and Trifonov 2005). Although the nucleotide composition of DNA will influence its stiffness and thus potential of wounding around the histones (Richmond and Davey 2003), nucleosome-DNA interactions are also affected by a plethora of other DNA binding proteins, such as TFs, their histone modifications, and the incorporation of histone variants.

Aberrational landscapes in cancer
Like with regulatory landscapes, e.g. nucleosomes and their histone modifications around genes, regulation through chromosomal aberrations apart from chimeric fusion genes cannot directly be inferred from their location. Solid cancers are heterogeneous, often displaying complex patterns of large-scale chromosomal alterations (Figure 7) (Frohling and Dohner 2008). Nevertheless, genome-wide DNA copy number studies often aim to locate driver oncogenes or tumor-suppressor genes within recurrent aberrant chromosomal regions (Cowin et al. 2010). Hence, systematic identification of the aberrational landscapes of cancers will help prioritizing the genes of interest. Inferred DNA copy number profiles may also aid in diagnosis of cancer type and prognosis of treatment outcome (Frohling and Dohner 2008). Sub-typing of cancers and their clinical outcome through gene expression profiling have effectively been performed in some cancers but have shown more difficult in others, often solid, tumors (Cowin et al. 2010) further calling for whole-genome aberrational profiling.

Measuring chromosomal rearrangements
Chromosomal imbalances can be measured genome-wide through comparative genomic hybridization (CGH) (Kallioniemi et al. 1992). In a typical CGH experiment coupled to DNA microarrays (array-CGH) (Mantripragada et al. 2004; Pinkel et al. 1998; Solinas-Toldo et al. 1997), total genomic DNA is isolated from test and reference samples, labeled differently and hybridized to a microarray after which the intensity of each fluorescent dye is measured. The intensity ratio between the test and reference signal for each spot on the microarray is, theoretically, proportional to the relative copy number of the corresponding genomic sequence (Figure 7). Deviation in ratio from expected values may be due to normal cell admixture in the cancer, i.e. test, sample. Like with microarrays for quantification of protein-DNA interactions, array probes for CGH measurements may be constructed in various ways yielding different characteristics (Diaz de Stahl et al. 2008; van de Wiel et al. 2010; Ylstra et
affecting resolution, SNR and specificity. These features as well as microarray-prone artifacts need to be considered and handled during array-CGH data analysis.

NGS-based approaches are recent alternatives to microarray hybridization of genomic DNA (Korbel et al. 2007) and may provide an advantage being able to also detect balanced chromosomal rearrangements (Chen et al. 2010). Such approaches use paired-end sequencing in which fragments are read from both ends instead of just one, common in for instance ChIP and nucleosome sequencing. Using reads from both ends of a fragment may pinpoint the breakpoints of translocations after separate alignment.

![Figure 7. Profile of copy number ratios along the genome of a glioblastoma multiforme tumor (sample G24460). The ratio of test DNA signals to a reference indicate DNA copy numbers in regions of homozygous deletions (zero copies, black), heterozygous deletions (one copy, red), diploidy (two copies, green), one-copy gains (three copies, dark blue), two-copy gains (four copies, light blue) and higher levels of amplifications (more than four copies, magenta). Individual chromosomes are collapsed sequentially and their boundaries indicated by alternating white and grey-shaded backgrounds. A few high level amplifications are outside the plotted range and indicated by triangles. The profiling of DNA copy number regions was done using a segmental maximum a posteriori approach (SMAP) (Paper III).](image)

**Computational analyses of chromosomal imbalances**

Recently developed commercial and custom-made genomic microarrays enable copy number analysis at a very high resolution, with several hundred thousands of measurement points. As a consequence of the large amount of data generated from such experiments, the use of automatic procedures for the identification of genomic aberrational breakpoints, segmentation, or the assignment of copy numbers, copy number profiling, to test DNA (Figure 7) has become an essential step in the analysis of array-CGH data. To this end, several methods have been proposed. A complete review of these methods is beyond the scope of this introductory text. Rather, focus will lie on methods that are relevant for and precedes our proposed method (Paper III) to
understand the underlying aims and challenges. A summary and a comparison of some early methods are provided by Lai et al. (2005).

The majority of methods assume a model with Gaussian distributions for which the means and, in some cases, the variances change at unknown breakpoints, which are inferred through optimal splitting of segments (Willenbrock and Fridlyand 2005). The common approach shared by such segmentation methods is to identify breakpoints in a manner that maximizes the likelihood probability distribution function, explained below (Hupe et al. 2004; Myers et al. 2004; Olshen et al. 2004; van de Wiel et al. 2007). The number of breakpoints may be controlled by a penalty that is extracted from the likelihood and increases with the number of breakpoints. A summary of segmentation methods is provided by Picard et al. (2005). Non-likelihood-based approaches include, for instance, smoothing methods (Eilers and de Menezes 2005; Hsu et al. 2005; Tibshirani and Wang 2008) and a clustering-based approach (Wang et al. 2005). Finally, HMMs have been adopted in many methods (Engler et al. 2006; Fridlyand et al. 2004; Marioni et al. 2006; Rueda and Diaz-Uriarte 2007; Shah et al. 2006; Stjernqvist et al. 2007).

Hidden Markov models are based on two stochastic processes (Rabiner 1989); one that produces the observations, e.g. the observed intensity ratios, and one that is hidden but can be observed through the former, e.g. the underlying DNA copy numbers. They are extensions of Markov chains; observable systems that at any time point, or probe in this context, are in any of a set of defined states \( Q \) and move between them according to associated transition probabilities (Figure 8). The probability of starting in a state at time point 1 is given by initial probabilities. For first-order Markov chains, the probability of a transition to a certain state \( S_i \) at time point \( t \) is only dependent on that and the current state, referred to as the Markov property, i.e.,

\[
p(q_t = S_i | q_{t-1} = S_j, q_{t-2} = S_k, \ldots) = p(q_t = S_i | q_{t-1} = S_j) = a_{ij}.
\]

In HMMs, the states are hidden and the observations, \( O \), are modeled through emission probabilities at each hidden state. Often, HMMs are considered homogeneous, meaning that neither the transitions nor the emissions do depend on the time index (Cappé et al. 2005).
Figure 8. A graphical representation of a three-state \( (Q = \{S_1, S_2, S_3\}) \) Markov chain with state transition probabilities \( a_{ij} \) between any two states \( S_i \) to \( S_j \), where \( 1 \leq i, j \leq 3 \). Initial state probabilities, denoted by \( \pi_i \) (\( 1 \leq i \leq 3 \)), originate from the artificial start state \( S_0 \).

Copy number changes between DNA segments are naturally modeled by transition events between hidden states in response to changed emission of intensity ratios. Given the observations, the goal is to infer the most probable corresponding sequence of hidden states describing the data. The probability of a certain sequence of copy number assignments given a sequence of observations and known parameters \( \lambda = (\Pi, A, \Omega) \) is given by the \textit{a posteriori} (posterior) probability distribution function (pdf) (Jaynes 2003):

\[
p(Q|O, \lambda) = \frac{p(Q|\lambda)p(O|Q, \lambda)}{p(O|\lambda)} = \frac{p(Q, O|\lambda)}{p(O|\lambda)}.
\]

\( \Pi, A \) and \( \Omega \) denotes initial probabilities, transition probabilities and parameters for emission probabilities, respectively. \( p(Q|\lambda) \) is called the \textit{a priori} (prior) pdf of the copy number states, \( p(O|Q, \lambda) \) the likelihood of the observed intensity ratios and \( p(O|\lambda) \) is a normalization constant. The optimal state realizations, i.e. path in the HMM, given the observations and HMM parameters may be derived using the Viterbi algorithm (Viterbi 1967) that maximizes \( p(Q,O|\lambda) \) or through sampling of the posterior pdf, e.g. through Markov Chain Monte Carlo (MCMC) simulations (Cappé et al. 2005). The former alternative is used by, for instance, Fridlyand et al. (2004), while the latter is adopted by others (Rueda and Diaz-Uriarte 2007; Stjernqvist et al. 2007). Often also the parameters of HMMs are unknown and need to be derived. Using a start solution for parameters, their optimal values may be re-estimated after state-sequence optimization using a \textit{backward–forward} algorithm or through \textit{expectation-maximization} approaches (Cappé et al. 2005; Rabiner 1989).

The majority of the above mentioned HMM-based methods use standard homogeneous discrete-index HMMs in which state transitions reflect the
underlying microarray probes. Hence the possible number of transitions and, as a consequence, the effective resolution of breakpoint detection is determined by the characteristics of the microarray used. When adjacent probes may be located genomically distant or, on the contrary, do overlap it is questionable whether a homogeneous HMM with strict Markov property is most beneficial. Non-homogeneously, state transition probabilities that depend on the distance between genomically adjacent probes may be incorporated to avoid such biases (Marioni et al. 2006; Rueda and Diaz-Uriarte 2007).

Many HMM-based methods infer the number of hidden states through model selection and perform copy number profiling/segmentation separately for each chromosome. Such approaches may easily overfit the model parameters to local effects in the chromosomes. Interpretation of results becomes questionable in cases in which inferred means and variances of the distributions associated with a certain DNA copy number differ between chromosomes. In some situations, however, chromosome-wise models may be preferable to genome-wide ones. Segmentation methods with chromosome-wise models are appropriate to detect relative copy number alterations between loci or mosaicism in the same chromosome when the actual copy number is not of interest (Rueda and Diaz-Uriarte 2007).

After effective profiling of DNA copy numbers, either through direct methods or through segmentation methods followed by subsequent computational interpretation of segment means, the profiles may be processed by further downstream analyses depending on the underlying aims. Various cluster-based approaches have been proposed for molecular sub-typing of cancers based on such data (Liu et al. 2006; Liu et al. 2007; Shah et al. 2009; Van Wieringen et al. 2008). The prognostic or diagnostic power of copy number data has also been investigated and tools for these purposes, mainly based on support vector machines, have been proposed, many of them summarized by van de Wiel et al. (2010). To effectively identify candidate driver oncogenes or tumor-suppressor genes, it is common to identify recurrent regions of chromosomal alterations across many samples (Rueda and Diaz-Uriarte 2010). Such a frequency-based approach may help identifying regions of importance in tumor development or initiation and relies on the assumption that such events are common between tumors of the same type.
Aims

Generally, the aims of the work presented in this thesis were to:

develop computational methods for accurate genome-wide profiling of chromatin and chromosome structure and to explore their characteristic landscapes and dynamics as well as involvement in transcriptional regulation.

Specifically, the individual aims of included papers were to:

I  use publicly available data from genome-wide studies to examine: (1) if nucleosomes are positioned at internal exons; (2) if certain histone modifications are preferentially found at internal exons compared to first exons; and (3) if exon positioning or histone modification is related to the transcriptional level,

II  (1) develop a method for precise and accurate mapping of nucleosomes or other protein–DNA interactions; (2) assess their positional characteristics in relation to both individual genomic coordinates and general genomic features; and to (3) investigate their dynamic nature in response to transcriptional regulation,

III (1) design and evaluate a non-homogeneous discrete-index hidden Markov model for genome-wide DNA copy number assignments that takes into account information about genomic position and overlap between probes on a microarray; and (2) to incorporate this model in a framework that enables the inclusion of user-controllable a priori information in the profiling process,

IV  (1) investigate the individual or combinatorial roles of chromosomal alterations and histone modifications on gene expression in cancer; and (2) to develop a strategy for integrative detection of candidate tumor-suppressor genes and oncogenes.
Chromatin regulation of transcription

Paper I

Methods

Genome-wide sequencing data for nucleosomes (Schones et al. 2008) and ChIP-seq data for 38 different histone methylations (Barski et al. 2007) and acetylations (Wang et al. 2008b) as well as gene (Su et al. 2004) and exon (Oberdoerffer et al. 2008) expression microarray data measured in CD4+ resting human T-cells are publicly available. Together, these data provide a unique resource for systematic exploration of transcription-related chromatin within and around exonic regions in a single cell type. For organismal comparisons, nucleosome Mnase-seq data measured in Caenorhabditis elegans (roundworm) (Valouev et al. 2008a) as well as histone modification ChIP-seq data and microarray gene expression data measured in Mus musculus (mouse) embryonic stem cells (Mikkelsen et al. 2007) were used.

All sequenced reads were prolonged according to their aligned strand to their average fragment lengths and the resulting aggregated count data were normalized to reflect the fraction over the expected count per bp.

Gene transcript and exon annotations for human, roundworm and mouse were extracted from the Ensembl (Hubbard et al. 2009) database releases 49 (NCBI 36), 50 (WS190) and 46 (NCBI 36), respectively. Using these annotations, the exons of gene transcripts were separated into first exons, internal exons and last exons. This categorization will most likely avoid influences of nucleosome and histone modification biases from 3’ and 5’ gene-end exons, possibly related to transcription initiation and termination. The genes were binned into three expression groups (low, medium and high) of equal cardinality. Similarly, the exons of highly expressed genes were binned according to their expression.

Using the SICTIN tools (Enroth et al. 2010) for binary storage and fast retrieval of NGS data we constructed average signal profiles, referred to as footprints, of nucleosome and histone modification data around first exons, internal exons and last exons across organisms to explore patterns of positioning. Similar footprints were made for exons and succeeding introns after grouping of exons according to expression level or genomic length. Selected human histone modification as well as nucleosome data were also
visualized at individual internal exons to inspect the generality suggested by the average footprinting around genomic features.

To evaluate possible human histone modification differences between exons and introns and their existence in relation to exon expression levels, we systematically compared the average aggregated and normalized count data within internal exons and their flanking introns in highly expressed genes. Firstly, we identified histone modifications with altered average signal (increase or decrease) in the high and low exon expression bins when compared to the average signal of the medium exon expression bin. Secondly, we required convincingly higher signal in exons compared to introns. Finally, we manually inspected the modifications to further avoid gene-boundary related influences. The relationships between selected histone modifications and expression levels were also explored using the Hoeffding’s D measure of dependence (Hollander and Wolfe 1999).

Results
Using NGS data for human, mouse and roundworm, we explored the positional landscapes of nucleosomes and their histone modifications in genes beyond the traditionally investigated regions around TSSs. Footprints of nucleosomal signals indicated apparent positioning patterns of nucleosomes within internal exons in both man and roundworm as well as the well-characterized patterns around first and last exons of genes. However, the signal was stronger at internal exons than at the +1 nucleosome downstream of TSSs, indicating higher concordance of positioning among internal exons than among first exons. In contrast, there were no indications of well-positioned nucleosomes in the introns flanking the internal exons. Furthermore, with the exception of a small fraction of the most highly transcribed genes, the positioning at internal exons was nearly unaffected by the transcriptional level indicating an important function of exonic nucleosomes as they are only temporarily ejected or displaced during transcription. Nucleosome positioning was apparent in nearly all exons except in those of lengths shorter than 50 bp. Interestingly, the median length of human internal exons suits very well with the expected length of DNA wound around the histone octamer in nucleosomes. Although gene lengths have changed during evolution, it is mainly due to the prolongation of introns and creation of new exons leaving the exon lengths intact (Taft et al. 2007). Hence, nucleosome positioning in internal exons seems evolutionary conserved given the length of exons and the concordance of nucleosome positioning between human and roundworm, subsequently also observed in *Plasmodium falciparum* (human malaria parasite) (Ponts et al. 2010) and *Oryzias latipes* (Japanese killifish) (Nahkuri et al. 2009).

In agreement with the findings of Kolasinska-Zwierz et al. (2009), we observed stronger indications of H3K36me3 on nucleosomes at internal
exons than at flanking introns in transcribed genes. H3K36me3 has previously been associated with active genes and suggested to accumulate over the whole transcribed region, to peak at the 3’ end of genes, and to dip at the very end (Bannister et al. 2005; Barski et al. 2007). Our results indicated that, in both human and mouse, the progressive increase in H3K36me3 toward 3’ end of transcribed genes was related to the exon distribution within genes since exons are more often located towards 3’ ends. These results led us to systematically explore human intragenic histone modifications and their relation to exon expression levels.

Histone modifications with higher occupancy in internal exons than in flanking introns of highly expressed genes and with no preference for TSSs were identified and grouped according to changed average signal between exon expression groups. As a result, four major classes were found. In the first class, we identified H3K36me3, H3K79me1, and H2BK5me1 characterized by progressively increasing signals from the low- to high-expressed exons. The opposite trend was observed in the second class, which comprised the histone marks H3K27me2 and H3K27me3, previously associated with gene silencing (Barski et al. 2007). Solely, H3K27me1 was contained in the third class and showed similar signal among high- and medium-expressed exons but a lower signal in those with low expression. The fourth class of histone modifications contained H3R2me1 and H3K36me1 and was characterized by a constant signal over all three levels of exon expression. Among these histone modifications, H3K36me3 and H3K79me1 stood out with the most convincing differences between internal exons and flanking introns and between exon expression groups. Furthermore, they were highly and more dependent on exon- rather than gene-expression levels as indicated by the Hoeffding’s D.

Our results of general nucleosome positioning and their expression-related histone modifications in internal exons were subsequently repeated and verified by other authors in human CD4+ T-cells as well as in other cell types and organisms (Hon et al. 2009; Luco et al. 2010; Nahkuri et al. 2009; Ponts et al. 2010; Schwartz et al. 2009; Spies et al. 2009; Tilgner et al. 2009).

Paper II
Methods
In order to accurately position nucleosomes, or other protein-DNA interactions, in a genome from aligned NGS reads we developed a Bayesian strategy called Strand-based Mixture Modeling of protein-DNA Interactions (SuMMIt). Since our aim was to identify reliable positions of interactions, we wanted to predict such locations only when support was given from both
ends of sequenced DNA fragments. With single-end sequencing of fragments, this requirement implicates a sufficient number of reads aligned to both sense and antisense strands at the flanks of interaction sites. Below, the SuMMIt approach for positioning of nucleosomes is summarized.

The distance between expected sense and antisense read start positions defining a nucleosome is known \textit{a priori} to be 147 bp (Richmond and Davey 2003). Therefore, in theory, the genomic mid-position of a nucleosome is determined by sense and antisense reads starting at positions \((147-1) / 2 = 73\) bp upstream and downstream of the interaction center position, respectively. In practice, the data is heterogeneous due to, for instance, variability in nucleosome positioning between cells, biased DNA sequence directed cleavage during enzymatic digestion of chromatin, and alignment problems. To cope with this, we considered reads within flanking windows when defining nucleosome mid-positions. We used counts of start positions for sense \((W^+\)) and antisense \((W^-)\) reads within windows at \([i-\theta_{f,1}, i-\theta_{f,2}]\) and \([i+\theta_{r,1}, i+\theta_{r,2}]\) of a putative nucleosome mid-position \(i\), respectively \((\text{Figure 9})\), where \(\theta_{f,1}\), \(\theta_{f,2}\), \(\theta_{r,1}\) and \(\theta_{r,2}\) were derived from the size range of selected DNA fragments subjected to sequencing.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure9.png}
\caption{(a) Sense (blue) and antisense (red) reads and (b) the counts of their start positions per bp as indicated by dots in (a). Sense and antisense reads supporting a nucleosome mid-position at position \(i\) are those that are located within windows at \([i-\theta_{f,1}, i-\theta_{f,2}]\) and \([i+\theta_{r,1}, i+\theta_{r,2}]\), respectively.}
\end{figure}
Since a large proportion of windows will not define nucleosome mid-positions, we considered $W^+$ and $W^-$ to be distributed as mixtures of Poisson distributions (Frühwirth-Schnatter 2006), capturing data defining either mid-positions or background noise:

$$W^+ \sim p_s \text{Pois} (\lambda_s) + p_{\neg s} \text{Pois}(\lambda_{\neg s}),$$

$$W^- \sim p_s \text{Pois} (\lambda_s) + p_{\neg s} \text{Pois}(\lambda_{\neg s}),$$

The mixture proportions, $p$’s, and the Poisson parameters, $\lambda$’s, were unknown and estimated from the data through Gibbs sampling (Frühwirth-Schnatter 2006), a MCMC technique in which the unknowns are sampled from the posterior.

After parameter-estimation of Poisson mixtures, we calculated log-odds of nucleosome mid-positions against the background component of the mixture using $W^+$ and $W^-$ data separately. A nucleosome was predicted when both ends of sequenced fragments suggested an interaction, i.e. when both $LO^+ > 0$ and $LO^- > 0$ (Figure 6e):

$$LO_i^+ = \log \left( \frac{p(\text{Nucleosome start flanking } i \mid W^+ , \Omega_1 )}{p(\text{No nucleosome start flanking } i \mid W^+ , \Omega_1 )} \right) = \log \left( \frac{p(w_i^+ \mid \lambda_s)p_s}{p(w_i^+ \mid \lambda_{\neg s})p_{\neg s}} \right),$$

$$LO_i^- = \log \left( \frac{p(\text{Nucleosome end flanking } i \mid W^- , \Omega_2 )}{p(\text{No nucleosome end flanking } i \mid W^- , \Omega_2 )} \right) = \log \left( \frac{p(w_i^- \mid \lambda_e)p_e}{p(w_i^- \mid \lambda_{\neg e})p_{\neg e}} \right),$$

where $\Omega_1 = \{p_s, p_{\neg s}, \lambda_s, \lambda_{\neg s}\}$ and $\Omega_2 = \{p_e, p_{\neg e}, \lambda_e, \lambda_{\neg e}\}$.

The fuzzyness, i.e. level of concordance among cells, of each called nucleosome position was calculated as the positional spread (standard deviation) of sense and antisense reads falling into windows $[i-(\theta_{f,1}+\theta_{r,1}), i-1]$ and $[i+1, i+\theta_{r,1}+\theta_{r,2}]$ of a nucleosome mid-position $i$, respectively. Using these windows, no reads defining adjacent non-overlapping nucleosomes will affect the fuzzyness score.

The Poisson mixture model of SuMMIt allowed for predicting changed nucleosome occupancy between samples. We aimed to locate loci where nucleosomes were depleted after treatment with transforming growth factor beta 1 (TGFβ1). To identify such events we applied strict filtering of inferred nucleosomal locations in TGFβ- cells that did not match any identified nucleosome in TGFβ+ cells. We required no overlap with positive positioning log-odds ($LO^+ \text{ or } LO^-)$ in TGFβ+ cells and, in addition, a log-odds of change greater than 10. Using clover (Frith et al. 2004), the resulting loci of nucleosome depletions were searched for overrepresented TF binding motifs defined by the JASPAR database (Portales-Casamar et al. 2010).

Micrococcal nuclease I was used to digest chromatin isolated from the human HepG2 cell-line before and after treatment with TGFβ1 to leave only nucleosome-bound DNA fragments. Fragments of appropriate sizes (130-180 bp) were subsequently selected and sequenced on the SOLiD NGS platform.
Results

In total, ~396 and ~301 million reads from TGFB- and TGFB+ cells were uniquely aligned to the reference genome (hg18, NCBI36). These figures correspond to an average read coverage of ~6.4X and ~4.9X, respectively.

The log-odds of SuMMIt provided clear indications of which positions had good support from both strands and revealed details that can be hidden using mere counts of read alignments. SuMMIt proved superior in detecting differences among cells within the same sample as well as in terms of resolution of positioning when compared with a simpler procedure using aggregations of strand-directed extensions of reads to expected fragment length (Paper I). Averaging of nucleosome log-odds values around TF binding sites revealed a striking number of located nucleosome peaks. Inspection of individual binding sites did, in fact, show that only a small minority of sites had flanking nucleosomes at the same distance from the mid-position of TF-DNA interaction demonstrating that important, although weak, patterns that are largely hidden due to background could be detected using SuMMIt. Deviations from nucleosomal patterns suggested by averaging over loci at individual locations were also detected around TSSs of highly transcribed genes. Only a small minority of TSSs mimicked the average pattern and surprisingly many lacked flanking nucleosomes at all.

With SuMMIt, we identified around 3 million reliable locations of nucleosomes in each sample. Of these, the majority was located too close to adjacent nucleosomes to be considered distinct indicating that many nucleosomes had discrepancies in localization between cells in the same sample. Examination of those with non-conflicting localization revealed sparser nucleosome occupancy in introns when compared to exons, in agreement with the results of Paper I. However, assessment of nucleosome phasing indicated that a higher fraction of fuzzy nucleosomes than phased ones was located at exonic regions. This was not the case for nucleosomes in intronic regions. We further profiled the AT-content around nucleosome mid-positions and found that fuzzy nucleosomes lacked boundary regions with high AT-content. Since exons often have high GC-content and AT-content has nucleosome disfavoring properties (Segal and Widom 2009), this finding explains a lower agreement in positioning between cells at exonic sequences.

Treatment of HepG2 cells with TGFB1 for one hour did not affect the large majority of nucleosomes. Still, surprisingly many nucleosomes were depleted after TGFB1 treatment suggesting that chromatin remodeling is a major player in TGFB-related transcriptional regulation. Many of these depletions were also associated with expression changes. Mining the depleted loci for transcription factor motifs revealed many putative binding sites for TFs relevant for TGFB-signaling that were also related to change in gene or exon expression.
Chromosomal aberrations and transcription

Paper III

Methods

Copy number profiling aims at classifying microarray probes to discrete copy number classes based on their intensity ratios and chromosomal positions. For a given sequence, of length \( T \), of genomically ordered probes and the corresponding sequence of observed intensity ratios, \( O \), start positions, \( S_p \), end positions, \( E_p \) and chromosome identifiers, \( Ch \), the goal is to infer the most plausible sequence of copy number assignments, \( Q \), describing the observed data.

We modeled the copy number assignments using a six-state HMM with parameters \( \lambda \) describing initial probabilities, \( \Pi \), transition probabilities, \( A \) and parameters, \( \Omega \), of Gaussian distributions, i.e. means and standard deviations, associated with each state. Although the number of states is flexible, the used six states allowed for capturing double deletion, single deletion, normal, gain, double gain and amplification. Distance-based transition probabilities (Colella et al. 2007; Marioni et al. 2006; Rueda and Diaz-Uriarte 2007), \( A^d \), which take into account the genomic distance, \( d \), between any two consecutive probes \( t \) and \( t + 1 \) were modeled using a length parameter, \( L \), that controls the convergence of transition probabilities towards equality. In detail,

\[
a_{ij}^d = a_{ij} - \rho \cdot \left( \frac{a_{ij}}{6} \right), \text{ where}
\]

\[
\rho = \begin{cases} 
1 - \exp \left( -\frac{d}{L} \right), & \text{if } d > 0 \\
0, & \text{otherwise}
\end{cases}
\]

\[
d = s_{p_{t+1}} - e_p, \text{ for } \]

\[
a_{ij} \in A, s_{p_{t+1}} \in S_p, e_p \in E_p, \]

\((1 \leq i, j \leq 6), (1 \leq t < T)\).

To model the dependency between probes that may arise when they overlap in terms of genomic position, we defined the emission probability of an observation in a given HMM state to be distributed as a mixture of Gaussian distributions that incorporates information from preceding probes, \( o_{lap} \),
with assigned states that overlaps with the current. \( o(r,t) \) denotes the fraction of overlap between probes \( r \) and \( t \). The emission probability was defined as

\[
b_{s_j}(o_t) = p_{(t)} \left\{ q_r \in Q : r \in o_{lap}, q_t = s_j, \Omega \right\} \\
= \alpha b_{s_j}(o_t) + \sum_{r \in o_{lap}} \beta_r b_{d_r}(o_t),
\]

where

\[
\alpha = \left( 1 + \sum_{r \in o_{lap}} o(r,t) \right)^{-1}, \quad \beta = \alpha o(r,t), \quad \text{and}
\]

\[
b_{s_j}(o_t) = p_{(t)} \left\{ q_t = s_j, \Omega \right\}, \quad \text{such that}
\]

\[
o_t q_t = s_j, \Omega \sim N(o_t, \omega_j), \quad \text{for}
\]

\[
o_t \in O, \quad s_j \in Q, \quad \omega_j = (\mu_j, \sigma_j) \in \Omega,
\]

\[
1 \leq r \leq t-1, \quad 1 \leq t \leq T, \quad 1 \leq i, j \leq 6.
\]

The parameters, \( \lambda \), that describe the HMM were unknown and needed to be inferred. We choose to find the most plausible state sequence by maximizing the joint posterior probability of \( Q \) and \( \lambda \) given \( O \) and \( P = \{Sp, Ep, Ch\} \) through alternate maximization over \( Q \) and \( \lambda \) in a segmental maximum a posteriori manner (Gauvain and Lee 1992), i.e. at iteration \( m \),

\[
Q^{(m)} = \arg \max_Q \left[ p(Q, O | P, \lambda^{(m)}) \right]
\]

\[
\lambda^{(m+1)} = \arg \max_\lambda \left[ p(Q^{(m)}, O | P, \lambda) p(\lambda) \right].
\]

The former maximization was achieved through chromosome-wise runs of a modified Viterbi algorithm in which the characteristics of our model, in terms of genomic distance and overlap between probes, were taken into account. A gradient descent method with individual learning rate adaptations (Bagos et al. 2004) was applied for local maximization of the HMM parameters. Given a start solution for the parameters and prior information, such as expected means of state distributions, \( p(\lambda) \), alternation of the two maximization steps until no further improvements can be made defines the SMAP procedure (Figure 10). The SMAP procedure was implemented in R (R Development Core Team 2010), relying on implementations of selected parts in C for efficiency, and was made available as an open source package in Bioconductor (Gentleman et al. 2004).

We evaluated the predictive performance of SMAP on synthetic data with varying inter-probe distances and overlap-affected intensity ratios as well as on array-CGH data from three glioblastoma multiforme (GBM) samples (Nord et al. 2009). The GBM samples were measured on a bacterial artificial chromosome (BAC) array with more than 32,000 probes together covering
more than 99% of the sequenced human genome (de Stahl et al. 2008). The predictions were subsequently compared with those made by the segmentation methods BioHMM (Marioni et al. 2006) and DNAcopy (Olshen et al. 2004). Eight different configurations of SMAP were tested with the following characteristics: with or without overlap consideration, standard or distance-based transition probabilities and genome-wide or chromosome-wise HMMs. We compared sensitivity and specificity for gained and deleted segments both overall and in windows around true breakpoints in the synthetic data. Sensitivity for gained segments is the proportion of truly gained segments which were correctly identified as such, whereas specificity for gain is the proportion of truly non-gained segments predicted to be non-gained.

Figure 10. A segmental maximum a posteriori (SMAP) approach for inferring the optimal state sequence of an HMM describing DNA copy numbers. The SMAP procedure takes as input a start solution of parameters and prior information. Alternating optimization of the copy number assignments, with fixed parameters, and optimization of the parameters, with fixed assignments, until no significant improvements can be made does the process of joint posterior maximization.

Results

In contrast to BioHMM and DNACopy, which are segmentation methods, SMAP does not require post-processing and pruning of made predictions. Since the fundamental task of segmentation approaches is to find the optimal splitting of segments, both methods relied on accurate ways to deal with distinct segments having Gaussian means too close to each other and to transform these segment means into copy numbers. To enable a fair comparison with SMAP, we dealt with these issues as recommended by Willenbrock and Fridlyand (2005).
Copy number profiling on synthetic data indicated superior performances of SMAP compared to the other two methods. Although the predictive differences between SMAP configurations were not prominent, the SMAP configuration with overlap-based emission pdfs, distance-based transition probabilities and genome-wide HMM showed superior performances both overall and in small windows centered on true copy number breakpoints. In particular, overlap-based emission probabilities yielded great improvements in sensitivity within these windows still having the highest specificities among the compared methods.

On the GBM data, SMAP was able to identify both large-scale regions and changes affecting only single probes on the array with aberrant copy number. However, single probes with intensity ratios in conflict with overlapping probes were discarded and not called due to the built-in overlap consideration in the model. Furthermore, using prior information about expected state means derived from the expected fraction of normal cell admixture in the test samples and an iterative adaptation of parameters to the data, SMAP identified intuitive distributions associated with each copy number state.

Paper IV

Methods

Pheochromocytoma is a rare endocrine tumor of the adrenal medulla, situated at the center of adrenal glands on top of kidneys, which mostly occurs sporadic with unknown causes. Most previous studies had focused on benign cases and profiled DNA copy number alterations or DNA methylations. Hence, little was known about the role of histone modifications in the perturbation of gene expression in malignant pheochromocytoma. To investigate the influence of histone modifications and DNA copy number aberrations on gene expression changes in this tumor type, whole-genome arrays were used to measure H3K4me3, H3K27me3 and chromosomal aberrations in one tumor sample. H3K4me3 is associated with RNAPII-recruited genes, whereas H3K27me3 is related to gene silencing (Barski et al. 2007; Mikkelsen et al. 2007; Wang et al. 2008b). ChIP performed on tumor tissue as well as ChIP input were hybridized separately in two replicates on Affymetrix GeneChip Human Tiling 2.0R Array sets, each set containing seven arrays with, all together, more than 40 million oligonucleotide probes. The probes are of length 25 bp and are on average genomically spaced every 35 bp. The chromosomal aberrations were measured on a 32K BAC array (Sandgren et al. 2010) and subsequent copy number profiling was performed using SMAP (Paper III). Gene expression were measured using Affymetrix U133 Plus 2.0 arrays in the same tumor.
sample and in healthy tissue obtained from the surrounding adrenal medulla of a patient suffering from benign pheochromocytoma. Unfortunately, no matching healthy tissue from the adrenal medulla of the studied patient could be obtained for gene expression or histone modification profiling.

The Affymetrix Tiling Analysis Software Developers Kit Revision 4 (Affymetrix Inc.) was used to preprocess the ChIP-chip data. Quantile-normalization was performed on each replicate of ChIP measurements together with the input measurements. This kind of normalization forces both distributions to have the same properties through rank-based averaging of individual signals. Since the measurement of short oligonucleotides results in high variance and thus are unreliable on their own (Ylstra et al. 2006), the normalized data was smoothened using a sliding window of 301 bp assigning the median value of the window to the center probe. Subsequently, the two replicates of log2-transformed ratios between ChIP and input signals were normalized together, by quantile normalization, and then averaged.

Genomic regions with histone modifications were identified using a sliding window approach implemented in R (R Development Core Team 2010). Firstly, for each probe on the array, a standard score was calculated from the average of log2-transformed ratios between ChIP and input signals of probes within a window of 150 bp. The scores were obtained against the global mean and standard deviation of the considered histone modification log2-ratios. A standard score, or Z-score, indicates the divergence in standard deviations from the mean of a population whose distribution and parameters are known. The population distribution was unknown and instead approximated by the characteristics of the log2-ratios of the array set. Assuming that the distribution of sufficiently many draws will be approximately normally distributed, a randomly selected probe from the approximated population with Z-score above 6 occurs with a probability less than 0.05 after Bonferroni correction for multiple testing. Regions of enrichment were defined as those containing at least two enriched probes (Z-score ≥ 6) flanked by existing measurements within 110 bp, i.e. the length of three averagely spaced probes, and with no contained gap longer than that length. Regions were subsequently prolonged with probes within flanking 75 bp and merged with neighboring regions if located within a distance of 1 kb, i.e. 1000 bp.

After careful inspection of the locations of histone modifications in areas around transcription start sites of genes, histone modifications were associated with gene transcripts, as defined by the University of California, Santa Cruz (UCSC) Genome Browser database (Rhead et al. 2010), if located within 2.5 kb from their TSS (Figure 5). These gene transcripts were also associated with the containing DNA copy number and their transcriptional level determined from gene expression arrays.
To systematically identify potential driver genes, we investigated the gene transcripts and their associations. Potential tumor-suppressor genes were identified as those with decreased expression and within regions of chromosomal deletions or associated with H3K27me3. Prospective oncogenes were scanned for in the group of gene transcripts with increased expression and within regions of chromosomal gains or marked with H3K4me3.

Results

DNA copy number profiling using SMAP revealed a complex pattern of chromosomal aberrations in the malignant pheochromocytoma. The total fraction of altered genomic sequence represented nearly 30% of the genome in which gains were nearly five times more common than deletions. Histone modification profiling identified around 16,000 and 28,000 enriched regions of H3K4me3 and H3K27me3, respectively. The majority of H3K4me3 marks were located in close proximity of TSSs, while H3K27me3 regions were more evenly distributed around the 20 kb region surrounding the TSS. As a consequence, a higher fraction of H3K4me3 regions were associated with gene transcripts than were H3K27me3 enriched regions. Of these transcripts, H3K4me3 association was more frequently coupled to housekeeping and metabolic process than what was expected by chance, while H3K27me3 associated gene transcripts showed a tendency for developmental processes assumed silenced.

Coupling with gene expression data indicated that high expression of genes in the tumor had a clear tendency of H3K4me3 enrichment or association with chromosomal gain. In contrast, low expressed genes showed higher association with chromosomal deletions or H3K27me3. Comparing the transcriptional levels in the tumor with the healthy tissue indicated the same associations for genes with decreased expression as for those with low expression and the lack of associations coupled with high expressed genes. Increased expression was strongly associated with chromosomal gains but not with H3K4me3. These results, despite the lack of matching healthy adrenal medulla tissue, indicated that either histone modifications or chromosomal aberrations, or both, constituted the underlying cause responsible for altered expression of a substantial fraction of genes in the investigated malignant pheochromocytoma.

Systematic screening of genes according to their associated histone modifications, DNA copy number statuses and expression levels revealed many putative oncogenes and tumor-suppressor genes, many previously associated with other tumor types. Among those identified, potential oncogenes were GNAS, INSM1, DOK5 (Figure 5), ETV1, RET, NTRK1, IGF2, and EZH2. Putative tumor-suppressor genes identified were, for instance, TGIF1, DSC3, TNFRSF10B, RASSF2, HOXA9, PTPRE, and CDH11.
Conclusions

Whole-genome studies with aims to decipher the regulatory code of chromatin and chromosome organization rely on efficient and reliable methods for handling and modeling of large data sets as well as accurate strategies for hypothesis testing and generation. Recent technological advancements have posed new challenges for bioinformatics research but also given opportunities to test and challenge existing hypotheses within fundamental cell biology. The awash of data that has followed allows for integrative studies coupling chromatin landscapes with, for instance, transcriptional statuses or chromosomal aberrations with PTMs of histones.

The common focus of all papers included in this thesis is the analyses of data from whole genome studies aiming at deciphering the regulatory implications of the structural properties of chromatin or chromosome organization. Using multiple data sources measuring different cellular activities or statuses we have approached the data in an integrative manner (Papers I, II and IV). This has allowed us to test existing hypothesis, such as the nucleosome positioning potentials of exons and how such localization may affect transcription (Papers I and II), and generate new ones (Papers I, II and IV) that require further studies. With new experimental data we have also tested the generality of well-accepted models under scrutiny (Paper II) and concluded that nucleosomal positioning patterns observed through averaging over many TSSs are rarely present at individual genes. Bayesian approaches enabling the incorporation of a priori information and the adaptation of statistical models to the data have proven very useful yielding accurate and narrow detection of chromosomal breakpoints in cancer (Papers III and IV) and reliable positioning of nucleosomes and their dynamics during transcriptional regulation (Paper II).

Our results showed that chromatin plays a major role in transcriptional regulation. In addition to the well-studied regions around TSSs of genes, nucleosomes have high preference for internal exons and the occupancy is nearly unaffected by the transcriptional status of genes (Papers I and II). These findings indicate that the placement of nucleosomes in exons, although in a non-phased way (Paper II), plays important roles in a cell since the positional preference has been preserved during evolution (Paper I). In fact, certain PTMs of histones were strongly related with exonic nucleosomes and indicative of exon inclusion or exclusion in a gene transcript (Paper I) suggesting a role in splicing. Hence, chromatin
regulation of transcription also seems to take place beyond the regions surrounding the TSS of a gene. In addition, the process of nucleosome positioning and repositioning at regulatory elements impeding or enabling access to DNA for TFs, respectively, adds to the regulatory machinery of chromatin. Eviction of nucleosomes for access of regulatory factors participating in signaling pathways responding to external stimuli seems to be a common phenomenon (Paper II). PTMs of histones may also play a causative role of perturbed transcriptional activity. Our results suggest that either histone modifications or chromosomal aberrations, or both, constitute the underlying cause responsible for altered expression of a substantial fraction of genes in cancer (Paper IV). Using an integrative approach we were able to pinpoint putative driver oncogenes and tumor-suppressor genes in cancer.

Taken together, our approaches to analyze whole genome data to decipher the structural layer of encoding and the results we have generated using them give strong indications of the power of integrative studies and Bayesian methods.

Our work also shows the new role of bioinformatics as an essential partner and indeed driver in fundamental biological research. Further technological advancements will push the throughput of data and the resolution requiring sophisticated computational approaches for processing and downstream analyses of data. The awash of data produced from NGS platforms is only at its initial state. The regulatory activities of many more genomes will be profiled, in various cell types, during varying conditions and at single molecule level. The multitude of data expected will allow us to study the control of transcription and its dynamics with higher granularity and precision leading to new conclusions in fundamental biology and revisions of current models of transcriptional regulation.

Forthcoming technological developments will determine the new era of quantitative biology, which in no doubt will rely heavily on bioinformatics.
Livet hos varje organism, må det vara en bakterie eller en multicellulär organism såsom människan, kontrolleras uteslutande av aktiviteten hos celler. Cellkärnan innehåller nästan all information i form av DNA, genomet, som beskriver denna aktivitet. DNA-molekyler är strukturellt organiserade i två komplementära strängar av nukleotider i enheter av baspar. Alla dubbelsträngar av DNA i en cell, normalt sett 46 stycken hos människor, organiseras och packas i 22 par av kromosomer samt två könskromosomer. Genom kombinationer av nukleotider defineras funktionella enheter, gener, som beskriver sammansättningen av RNA vilket i sin tur kan beskriva hur proteiner konstrueras. Gener i sig innehåller funktionella element, exoner, som beskriver dess slutprodukt och kan kombineras på olika sätt genom så kallad splitsning. En gen har i sig ingen funktionell egenskap förutom som informationsbärare. Aktiviteten i celler styrs emellertid av dess produkter; RNA eller protein.

En människa består av flera miljarder celler vilka kan grupperas i över 200 olika celltyper. Vissa celltyper existerar och har samma roll i flera organ medan andra har specifika egenskaper och är begränsade till enstaka organ. På samma sätt avläses informationen hos vissa gener i alla celltyper medan andra gener endast används i ett fåtal celltyper. Ändå, med vissa undantag av celltyp och variation i DNA, så är genomet detsamma i alla celler hos en individ. Även om kontrollen av genaktivitet är kodat i DNA så skiljer sig dess reglering åt mellan celltyper och under olika faser av en cells livscykel.

Genaktivitet styrs i två separerbara nivåer. Den strukturella och kemiska signaturen hos regioner i DNA påverkas av dess kombination av nukleotider och eventuella kemiska modifieringar, den första nivån av genreglering. Följaktligen kan dessa egenskaper påverka möjligheten för proteiner med genreglerande egenskaper att binda till DNA. Transkriptionsfaktorer innefattas i denna typ av DNA-bindande proteiner, vilka kan styra rekryteringen av det maskineri av proteiner som behövs för avläsning, transkription, av gener.

I den andra nivån, den strukturella nivån av genreglering, styrs generens aktivitet av högre organisation och packning av DNA i kromosomer. För att få plats i cellkärnan måste DNA packas extremt väl. Nukleosomer, komplex av histonproteiner med DNA virat kring sig, utgör grunden för sådan packning. Packningen påverkas även av kemiska modifieringar, så kallade histonmodifieringar, och internukleosomala interaktioner vilket resulterar i

För att förstå hur genreglering påverkas av kromatiska eller onormala kromosomala tillstånd krävs storskaliga biologiska experiment samt anmärkningsvärda bioinformatiska metoder vilket möjliggör mätningen av sådana egenskaper i hela genomet. Analys av genreglering genom kromatin innefattar kartläggning av nukleosomers positioner, dess modifieringar samt hur position eller status påverkar geners uttryck eller hur dessa egenskaper i sig påverkas av faktorer, t.ex. externa stimuli, vilka ändrar cellers tillstånd och aktivitet. Kromosomala avvikelse kan vara gemensamma eller skilja sig väsentligt mellan olika cancertyper. För att förstå vilka gener vars uttryck särskiljer en cancercell från en frisk sådan krävs kartläggning av deras omslutande kromosomala avvikelse.


Våra resultat har visat att kromatin spelar en viktig roll i regleringen av gener transkription. Vi har visat att nukleosomer har en stark preferens för exoner och att deras position i sådana regioner ej påverkas nämnvärt av geners aktivitet (Paper I & Paper II). Detta antyder att positioneringen av nukleosomer över exoner, dock med marginella skillnader mellan enskilda celler (Paper II), har en viktig roll eftersom denna preferens bevarats genom evolutionen (Paper I). Faktum är att vissa histonmodifieringar visade sig starkt kopplade till exoniska nukleosomer och indikerade användandet av
enskilda exoner i transkription av en gen (Paper I) vilket tyder på en roll i splitsning. Dessutom verkade nukleosomes positionering vara viktig och deras förflyttning vara frekvent förekommande i regleringen av transkriptionsfaktorers åtkomst till DNA (Paper II).

Acknowledgements

To state that others have influenced the work presented in this thesis would be a great underestimation. I am in great dept to numerous people for invaluable knowledge, expertise, opinions, support, guidance and friendship making this thesis a reality.

Many thanks are due to

Professor Jan Komorowski, my main advisor. Thank you for scientific guidance and for encouraging me to (almost unrestrictedly) explore the world of biology and medicine and mine their wealth in data. I am very grateful for you trusting my capabilities.

Professor Jan Dumanski, my co-advisor. Our collaborative projects have been challenging but fun. I really appreciate your scientific visions and eagerness to conduct rigorous research.

Professor Claes Wadelius, my co-advisor. Thank you for introducing me to the field of chromatin, for numerous research ideas, support, guidance and for putting me on the right track when I think I understand something but obviously do not.

Professor emeritus Jan Maluszynski, former master thesis supervisor at Linköping University, for being the initiating factor for me starting my PhD studies. I will always remember your encouragements and care.

Former (Torgeir, Alice, Claes, Helena, Jakub, Eva, Adam, Marcin, Aleksejs, Alvaro, Alvaro (yes there were two of them) and Olle) and present (Stefan, Marcin (not the former one), Susanne and Henric) colleagues at the Linnaeus Centre for Bioinformatics. I am especially grateful for scientific and non-scientific discussions as well as explorations in (sometimes rather obscure) music history held within the Aquarium (Stefan, Jakub and Adam).

Former and present members of the Jan Dumanski/Teresita Diaz de Ståhl group. Thank you (especially Johanna, Helena, Tere, Calle, Arek and
Uwe) for good collaborations and for providing me with a lot of challenging work during my first years as a PhD student.

Former and present members (Madhu, Mehdi, Ola and Alvaro (again)) of the Claes Wadelius group as well as the participants of the Chromatin Journal club (especially Chandrasekhar Kanduri and his gang in addition to the ones mentioned above) for vivid (and rather detailed) discussions regarding chromatin biology.

The administrative staff (Gunilla, Sigrid and Christin) for invaluable support.

The folks at the BMC Computer Department (Nils-Einar, Emil, Gustavo, Anders and Jerker) for maintenance of servers and supporting our needs.

Inge, Elna, Jim, Maria and Povel. Thank you for everlasting support and encouragements during these years and for letting me try over and over again to explain what I really do in plain Swedish.

Bosse, Lotta, Sanna, Kjetil and Jenna for great support during this process. I really appreciate you helping me and my family out when the schedules were too heavy.

Emma, my love. Without your support and encouragements not a word in this thesis would have been written. A simple thank you is not enough. I am very grateful for you being at my side during all those years. A special “tack så mycket” (or “takto” in his own words) to my wonderful son Lowe for constantly reminding me of what is important in life.

Robin Andersson
Uppsala, September 2010


Schlesinger, Y., Strausman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff,


Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 768*

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-130999