Integrative epigenome analysis

Halfdan Rydbeck

Thesis for the degree of Philosophiae Doctor (PhD)

Department of Tumor Biology
Insitute of Cancer Research
The Norwegian Radium Hospital, Oslo University Hospital
Faculty of Medicine
University of Oslo, Norway
2013

Contents

List of Abbreviations

BI Broad Institute

CEPH Centre d'Etude du Polymorphisme Humain

CEU CEPH (Utah residents with ancestry from northern and western Europe)

ChIP-seq Chromatin Immuno Precipitation sequencing

CNP Copy Number Polymorphism

CNV Copy Number Variation

CpG C followed by a G on a single DNA strand

CTCF CCCTC-binding factor

DNA Deoxyribonucleic acid

DNase-seq DNase I hypersensitive sites sequencing

ENCODE Encyclopedia of DNA Elements

FAIRE-seq Formaldehyde-Assisted Isolation of Regulatory Elements sequencing

GUI Graphical User Interface

GWAS Genome Wide Association Study

H3K27me3 Histone-3 Lysine-27 ri-methylation

H3K36me3 Histone-3 Lysine-36 tri-methylation

H3K4me3 Histone-3 Lysine-4 tri-methylation

H3K9me2 Histone-3 Lysine-9 bi-methylation

HCP High CpG-content promoters

ICGC International Cancer Genome Consortium

LAD Lamina Associated Domains

LCP Low CpG-content promoters

MCD ultiple Concerted Disruption

NCI the National Cancer Institute

NDR Nucleosome Depleted Regions

NFR Nucleosome Free Regions

NHGRI US National Human Genome Research Institute

NHGRI the National Human Genome Research Institute

NIH National Institutes of Health

PcG polycomb group

RNA-seq RNA sequencing

RNA Ribonucleic acid

RNAPII RNA polymerase II

RRBS Reduced Representation Bisulphite Sequencing

sCNA somatic Copy Number Aberation

SNP Single Nucleotide Polymorphism

SNV Single Nucleotide Variation

TCGA The Cancer Genome Atlas

TFBS Transcription Factor Binding Site

TSS Transcription Start Site

UCSC University of California Santa Clara

UCSD University of California San Diego

UCSF University of California San Franscisco

UW University of Wisconsin

YRI Yoruba in Ibadan, Nigeria

List of Figures

Acknowledgements

Initially, I would like to express gratitude to *Professor Marcella Devoto* at Children's Hospital of Philadelphia, who through her insightful presentations made genetic epidemiology so interesting, and for introducing me to programming in R. I would like to thank *Professor Arnoldo Frigessi* at Statistics For Innovation (SFI), who was the first person at University of Oslo (UIO) that I got in contact with about the possibilities to study for a PhD degree at UIO. Except for being a co-supervisor to me he introduced me to *Professor Eivind Hovig*, who is involved in many projects bridging bioinformatics, genomics and tumor biology. Eivind Hovig became my main supervisor and through his network I ended up with two more supervisors, *Professor Ola Myklebost* at the Department of Tumor Biology and *Professor Knut Liestøl* at the Department of Informatics. I would like to express my very great appreciation to them all for offering me the opportunity to work within such an interesting field. Next I would like to thank all co-authors. I am particularly grateful for the assistance given by the developers of The Genomic HyperBrowser, *Professor Geir Kjetil Sandve*, *Sveinung Gundersen* and *Kai Trengereid*. I would like to offer my special thanks to *Dr. Leonardo Meza Zepeda* and *Dr. Stine Kresse* and *Marieke L. Kuijjer* for advice given in relation to copy number aberration and osteosarcoma treated in Papers I and II. My special thanks are extended to the *Tuesday meeting group* for all interesting presentations of bioinformatics topics. I would like to thank *Gro Nilsen* for sharing ideas and R scripts for copy number analysis and *Jonas Paulsen* for inspiringly paving the way in 3D analysis. Finally, I would like to thank my friends and closest family. A particular thanks goes to my dear *May-Helen* for all her love, support and patience.

List of Papers

Paper I Kresse SH, Rydbeck H, Skårn M, et al. Integrative Analysis Reveals Relationships of Genetic and Epigenetic Alterations in Osteosarcoma. PloS one 2012;7:e48262

Paper II Kuijjer ML, Rydbeck H, Kresse SH, et al. Identification of osteosarcoma driver genes by integrative analysis of copy number and gene expression data. Genes, chromosomes & cancer 2012;51:696–706

Paper III Sandve GK, Gundersen S, Rydbeck H, et al. The Genomic HyperBrowser: inferential genomics at the sequence level. Genome biology 2010;11:R121

Paper IV Rydbeck H, Sandve GK, Rye M, and Hovig E. ClusTrack: Defining distance and clustering for genomic element tracks to compare landscapes of occupancy. Submitted. 2012:1–19

Chapter 1

Introduction

The term epigenetics was coined by Conrad Waddington [1], for the purpose of having a concept linking the single version of a genome shared by all cell types of a multicellular organism to their varying phenotypes. The epigenome refers to all the epigenetic modifications across a genome. In Waddington's definition lies that the epigenome, unlike the genome, has an inherent plasticity across cell types enabling the epigenome to participate in the enactment of cellular change and differentiation. Today, it is known that the plasticity of the epigenome is mediated through reversible chemical modifications to DNA and histone modifications, which both alter gene expression. It is also known that the modifications, when needed to rigidly maintain cellular states, can be inherited across cell cycles. Many intriguing properties, like the ability to respond to environmental changes within a generation and to facilitate trait inheritance, have been reported for subsets of the underlying constituents of the epigenome. A molecular machinery, epigenetic remodelers and modifiers, has also been identified as responsible for the genomic positioning, maintenance and reading of epigenetic marks and is being increasingly well characterized and understood [2, 3]. The definitions of the epigenome, given in current reviews [4–8], tend to vary in their contents, reflecting that no universally accepted version exists, so far. A discrepancy in the included biological entities therefore also exists. There are a few circumstances that make it difficult to formulate a unified definition of the epigenome. First, the epigenome remains to be fully discovered and characterized. Second, most definitions rely on the epigenome mediating inheritance. Inheritance can, though, refer to two completely different events of the life cycle of the diploid organism, the transgenerational one and

the mitotic one. Also, somewhat contradictory, the epigenome mediates, in addition to phenotypic inheritance, phenotypic plasticity, depending on the mission of the cell. Epigenomic together with genomic properties will be put into the context of the diploid life cycle in Section 1.1. The potential for reproducible integrative analysis to characterize these properties in detail will hopefully be discerned. The terms transgenerational and mitotic inheritance and germline as well as somatic mutation will be defined and distinguished. In Section 1.2, five epigenomic components are described. Definitions of chromatin states and epigenomic landscapes are given. Examples of the roles that they play in chromatin biology, differentiation and disease are also given. DNA methylation and histone modifications, which are the subjects of analysis in this thesis, are two undisputed constituents of the epigenome. Together with the genome and other DNA interacting proteins, depending on definition, they make up the chromatin. The studies presented in this thesis integrate in various combinations the epigenomic data just mentioned, together with genomic aberration data and gene expression data. Biological data, that ends with -omic, is collectively referred to as omics data. Data where the genomic position is a central feature is often stored as genomic tracks. The format is essential for many of the analyses performed in this thesis and is described in Section 1.3.

The recently increased activity within epigenomic research has been fueled by two waves of emerging high throughput technologies, i.e. microarray and second generation sequencing, applied to mapping of DNA methylation and protein-DNA interactions. These technologies and some relevant applications of them will be discussed in Section 1.11. The analytical aspects of high throughput genomic and epigenomic studies of today consist of many sequential steps, referred to as pipelines or workflows. These steps include format customization, preprocessing, format transformation, normalization and finally primary and downstream analysis of the data. Due to the massive size and complexity of the input data, the results themselves, commonly in the form of size effects and p-values, are in such an abundance that visualization, for instance as heatmaps or genome browser views, is needed for comprehensibility. Thus, the intricacy of the analytic pipeline becomes an obstruction to its reproducibility, which is a requirement for scientific credibility. It also hampers the transparency of the analytical process. Even if a piece of software has an interface that is easy to use, running it as a blackbox prevents the detection of built-in errors and scrutiny of the analytical soundness. Many analyses, including some of the ones paving the way at the forefront of omics research and making the most interesting discoveries, suffer from the lack of such reproducibility and transparency. Simultane-

ously, at the forefront of bioinformatics software development, infrastructures/environments that facilitate development and usage of reproducible and transparent applications are created. The software Galaxy is one well known example based on a graphical user interphase, and R/Bioconductor is probably the most well known example based on a command line user interphase. After a description of individual components of the chromatin in Section 1.2 and how they make up the epigenomic landscape, the biological context in which the epigenome operates will be summarized and important studies contributing to related insights will be referenced. In Section 1.3 it is described how the epigenome through regulation of multiple aspects of gene expression contributes to determine cellular morphology and function. In Section 1.4 enzymes, that influence the epigenomic landscape and have become important targets for medication, are discussed. In Section 1.9 the epigenomic landscape is put into a context of circuits of gene expression and gene regulation. The production, modification and genomic positioning, in relation to genes, of epigenomic components regulate genes and drive cellular differentiation. Recent studies of the properties of components of the epigenome have, regardless of their reproducibility, already impacted the understanding of the epigenome remarkably, not least by suggesting a list of possible hypotheses to validate. The epigenome has been implicated in disease and especially in cancer development, some of these findings are treated in Section 1.10. As next generation sequencing technologies offer an ever increasing scope and resolution in characterizing components of the chromatin, one can foresee that future studies will be based on integration of many types of data to reveal mechanisms based on complex interactions. Given the large amount of possible integrative analyses, a given software system cannot likely, in its first version, be expected to cover them all in detail, but has to have the capacity to be adapted to the demand. It has to be scalable and extensible. Available software for integrative analysis of the epigenome are reviewed in Section 1.13. In Papers I-IV novel software tools are introduced. They are developed and utilized for reproducible integrative analysis of epigenomic, transcriptomic and genomic data using the R and Galaxy frameworks. These frameworks are further discussed in Sections 1.13.5 and 1.13.11.

Many of the methods for integrative epigenomic analysis tend to be developed by large consortia. The consortia have been formed during the latest ten years to collect epigenomic and other types of data. An important difference between their missions is the types of samples that they use. They have in common the declared priority to make data available through public databases for usage by the bioinformatics community. The data will eluci-

date processes in normal, disease and cancer development. These consortia are discussed in Section 1.14. Ease of access to such and other types of public data for integration with local data is an important determinant of the usefulness of a piece of bioinformatics software. References to data repositories are given in Section 1.15 .

1.1 Chromatin, replication and inheritance in the diploid life cycle

The genome and the epigenome, that together make up the chromatin, represent different capabilities of mediating inheritance in the diploid life cycle. Figure 1.1 on page 7 shows four generations of members of a family tree. In the figure mitotic inheritance occurs in cell lineages along the vertical bars, indicating the life spans of individuals. Transgenerational inheritance occurs along the horizontal colored lines indicating the conception.

The genome, despite being a rigid carrier of information, can occasionally, through mutation, fail to mediate inheritance. Such mutations occurring in the germ line will disrupt transgenerational inheritance and affect the genomes of every cell in the progeny and lead to genomic polymorphisms and disease predisposition. Mutations occurring in genomes that will not be passed on transgenerationally, so called somatic mutations, will affect subpopulation of cells within the bodies of organisms. When accelerated out of control such mutations lead to cancer. The epigenome is often described as governed by developmental programs (encoded in the genome) and therefore to have an inherent plasticity. It does, however, also need the capacity to be truthfully inherited as when mature fully differentiated cells are regenerated into identical daughter cells. Detected mechanisms for the copying of epigenomic marks in connection to DNA replication and a number of other suggested mechanisms for cellular or mitotic inheritance are described in [9]. Examples of manifestations of mitotic inheritance of the epigenome are imprinting and X-chromosome inactivation.

Any cell of an organism is connected to the zygote of the organism through a sequence of ancestral cells and their divisions. That connection is called a cell lineage. Most cells of the body of adult multicellular diploid organisms are naturally divided into two major types of cells. One type is the germ cell with a single, or haploid, set of genomic material. The other type is the somatic cell encompassing all cells, but the germ cell type, with a double, or diploid, set of genomic material. Exceptions, like multi nucleated cells

[10], do, however, exist. The cell divisions of somatic cell lineages are exclusively mitotic, which means that the mother cell splits into two daughter cells and provides each daughter cell with two of the four genomes available after replication, making them diploid. For the germ cell lineage, the sequence of mitotic cell divisions is ended by a meiotic one. In meiosis, the mother cell is instead divided into four daughter cells, and one of the four available genome copies after replication is distributed to each daughter cell, making them haploid.

Meiosis is also accompanied by an enzymatically administered shuffling of genomic segments between maternal and paternal homologues. This results in a recombination of genomic segments from these, so that each of the four haploid daughter cells carries a mix of maternal and paternal trait information. Recombination leads to the random segregation in pedigrees of variants of loci not located close to each other on a chromosome. Sets of such variants located close to each other on the genome, and therefore deviating from random segregation, are called haplotypes. The mapping of disease genes that have been performed during the last 20 years is dependent on that variants of proximal loci do not segregate randomly.

The life of an individual begins when haploid parental genetic materials are combined into an egg cell at conception, leaving it with two copies of the genome, one maternal and one paternal, and making it a diploid zygote. The prospect of epigenomic components being transferred and combined in the same event, so called transgenerational epigenetic inheritance [4], has generated great interest [11]. A few observations have been made that could reduce or obstruct the fulfilment of this prospect, like for example the erasure of methylation patterns in the germ line. An observation in support of transgenerational epigenomic inheritance is the transmission of non-coding RNA [4], from both sperm [12] and egg [13, 14] to zygote. Non coding RNA is gaining recognition as an epigenetic factor due to recent reports on its involvement in gene regulation and transfer across cell cycles [15]. After conception, the zygote will multiply through mitosis. Each division is preceded by the doubling of the genetic material through DNA replication, a process fundamental to the maintenance of information across generations of cells and organisms. Replication is an intricate activity of molecular interactions between proteins and DNA. In DNA replication, the existing DNA molecule is used as a template for the construction of a new one. This involves an unwinding and enzymatic cutting of the existing antiparallel double helix, which makes the process vulnerable to introduction of sequence errors into the daughter DNA molecules, or mutations. Any formed mutation will be a hazard to the fitness of the daughter cells of the division, and any of their cel-

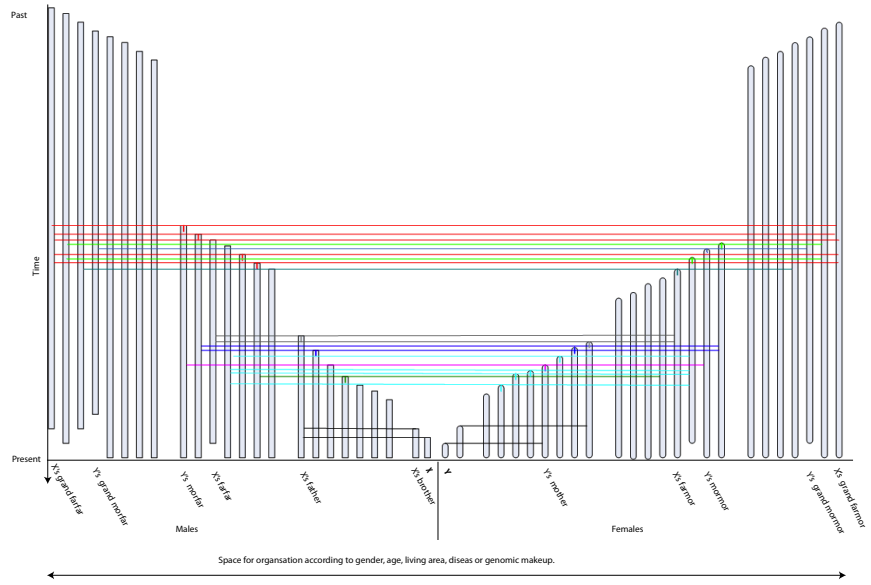
lular descendants inheriting the mutation. Elaborate molecular machinery for monitoring and regulating the outcome of replication has been shown to be present in many organisms [16]. This suggests that avoiding error introduction at replication is a highly prioritized task by the cell. Still, errors set off by replication are believed to be the major contributor to disease [17] and genome evolution [18].

One of the most central molecular units of the replication machinery, DNA polymerase, is also one of the most important tools applied in molecular biotechnology. Poly Cyclic Replication or Polymerase Chain Reaction (PCR) and DNA sequencing would not be possible without it. PCR and sequencing are fundamental to the high throughput technologies used to generate the data integrated in this thesis. The methods are described in Section 1.11.

If genomic mutations are introduced at replication, or during recombination, in the germ line, and are passed on to the haploid germ cells, and if the formed germ cells make it to conception, the mutations will be inherited at the level of the organism. Such mutations are referred to as germ line mutations. Mutations taking place in a somatic lineage, somatic mutations, will be less damaging, in the sense that they will only be inherited on a cellular level within the somatic lineage and within a single organism.

It has been shown in some cancers that homozygous disruptive mutations of tumor suppressor genes occur stepwise with the first disrupted allele being an inherited germ line mutation and the next one being a mutation occurring in the somatic lineage from which the cancer clone expanded [19]. This stepwise way of acquiring a homozygous gene disruption is commonly referred to as Knudson's two hit hypothesis [20]. One hypothesis regarding the nature of genomic and epigenomic interplay in cancer development is a two hit hypothesis involving them both [21]. A germline genomic alteration would thereby hit one allele of a locus and a somatic epigenetic alteration would silence the other.

Figure 1.1: The figure shows a family tree that allows for annotation of events, like births and deaths, along a time axis. A number of individuals are plotted along the x-axis. Time is represented along the y-axis with the x-axis intercept representing current time. A conception, or the transgenerational inheritance, is represented as a horizontal line connecting the three involved people. Conceptions involving the same parents have the same color. Benefits as compared with a regular pedigree are that individuals can be sorted in any order, for instance according to case control status, along the x-axis and that dates/time of birth and deaths, ages and ages at conception of individuals can be visually deduced. Interfamily generational shifts will also be seen/appear. It allows for illustration of cell lineages and the difference between cellular and transgenerational inheritance. Males are represented as rectangles with sharp edges while females have round edges. Extending this family tree to all life in the biosphere, and visualizing it in three dimensions, results in the ”Tree Of Life” or ”Mount Improbable” shown in Figure 1.3 at page 11.



1.2 Epigenomic components

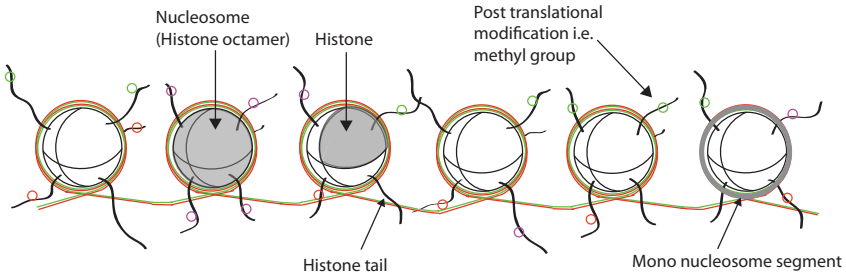
Recent discoveries and more inclusive definitions of the epigenome, like ”mechanisms other than changes in DNA sequence that perpetuate altered cellular activity states” [5], have made the epigenome to encompass:

1. Chemical modifications of the DNA
2. Histone proteins with various chemical modifications added to their amino acid tails
3. Non-protein coding RNA
4. Chromatin accessibility
5. Spatial organization of the chromatin.

The epigenome has been implicated in almost all genomic functional processes including transcription, recombination, DNA repair, replication, kinetochore and centromere formation by various studies [22]. Involvement of the epigenome in those processes can also be expected since its components seem to be present genome wide. Studies using the software reviewed in Section 1.13.10 shows that the combination of epigenomic components that occupy a given genomic region determines its current chromatin activity or state. The varying chromatin states along the genome have been called epigenomic landscapes. Such epigenomic landscapes should not be confused with the cellular development that Waddington meant when he introduced the term ”epigenetic landscapes”. The technologies used for acquiring data on the epigenomic components listed above have for decades undergone a constant development towards a higher genome coverage. It is only by the latest, second generation sequencing, technologies, discussed in Section 1.11, that it has become possible to collect genome wide data at (almost) base pair resolution. Also assays for probing the genomic locations of histones and DNA methylation on a genome wide scale are discussed. The data has revealed that chromatin accessibility and 3D organization are influenced by the genomic localization modifications of nucleosomes and DNA methylation through alterations of non-covalent interactions within and between nucleosomes. Most of our genome is normally packaged as transcriptionally repressive chromatin. This type of chromatin is heavily methylated and the DNA is packaged into compacted nucleosomes that contain deacetylated histones, a state referred to as heterochromatin. Heterochromatin is highly condensed, late to replicate, and contains primarily inactive genes. Another fraction of the genome is transcriptionally competent. It is called euchromatin. It has a relatively open configuration and contains most of the active

genes. The state of chromatin in these regions must be dynamic to meet the changing transcriptional requirements of a cell [23]. Methylation of CpG sites is the most common chemical modification of DNA. The CpG denotation is used to distinguish the C followed by a G on a single strand from the CG base pair. DNA methylation is primarily noted within centromeres, telomeres, inactive X-chromosomes, and repeat sequences [3]. CpG sites of eukaryotes are, with a varying frequency between cell types and stages, chemically modified by the addition of a methyl group. Histones are proteins that can interact with DNA to form the basic unit of chromatin, which is the nucleosome as depicted in Figure 1.2 on page 9. The resulting compaction of DNA makes the massive amount of genetic information stored in a genome fit into the limited space of a cell nucleus [24]. The nucleosome is made up by 147 bp of DNA wrapped twice around a histone octamer of four pairs of H2A, H2B, H3 and H4. The basic histone variants can be replaced with other ones, and chemical groups can be added to their amino acid tails changing their functional properties, see Figure 1.2 on page 9. How histone modifications are distributed across the genome varies between cell types and states , reflecting functional differences between these. The composition of histone modifications in a given site of chromatin has recently been shown to be associated with the activity of that genomic region [25, 26].

Figure 1.2: Chromatin is made up of DNA wound twice around histone octamers forming nucleosomes. Chemical modifications of the amino acid tails of the histones change their properties. The cell is equipped with a molecular machinery for the modification of histones.



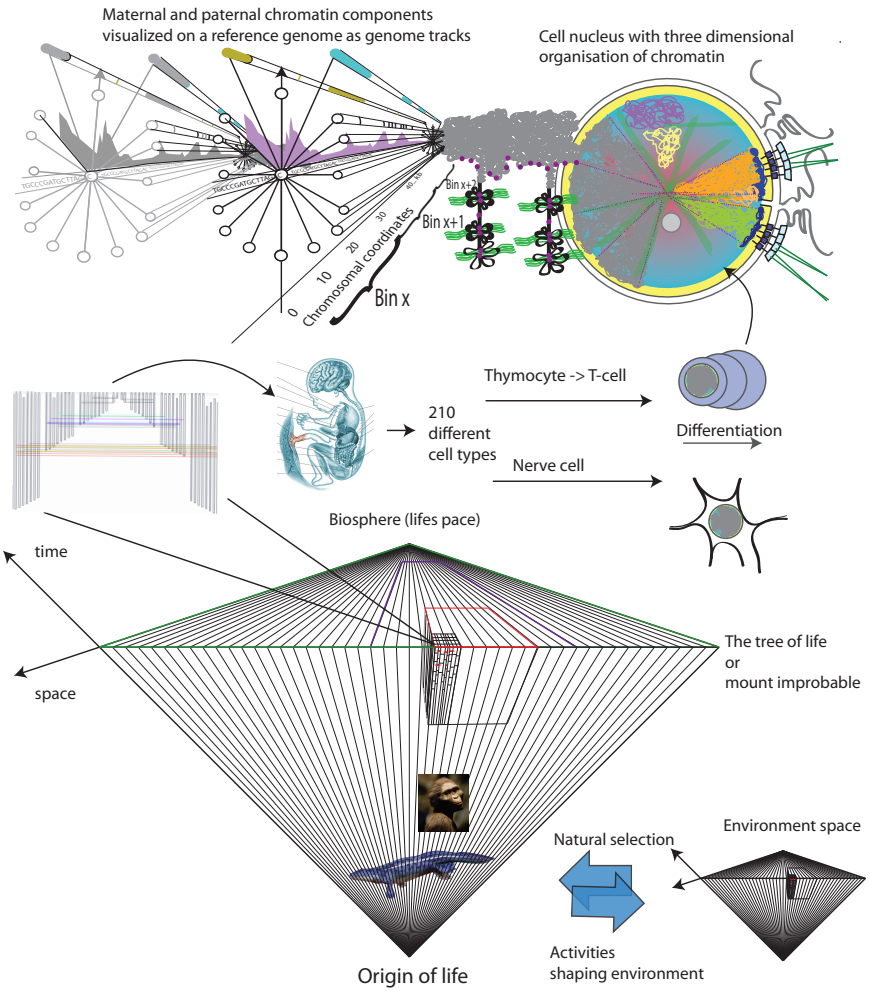
Interesting to note about non-protein coding RNA is that the majority of the human genome has been found to be transcribed into different types of such RNAs in at least one of the close to 200 human cell types [27]. A common definition of biological function is based on evolutionary conservation.

Many of the transcriptionally competent regions are not evolutionary conserved. A scientific debate has emerged about whether a transcriptionally competent genomic region has a function just because it is transcriptionally competent. Only a few percent of the human genome is evolutionary conserved. The functionality of, or the lack of functionality of, transcriptionally active non-conserved genomic regions are discussed in [28–30]. Anyhow, at least some ncRNAs are epigenetic factors with an important role in differentiation and disease.

1.3 Genomic tracks

An initial task of genomics research on an organism is to sequence and assemble a reference genome sequence from a single sample. The genome will then be annotated with functional information. To be able to efficiently indicate any genomic site, each reference chromosome is made into a one-dimensional coordinate system by incremental numbering of its bases, starting at the telomeric side of the short arm and ending at the telomeric side of the long arm. The chromosomes are then annotated using sets of coordinates, called genome annotation tracks or genomic tracks [31], describing the locations of related features. Genomic tracks are commonly stored as tab separated text files, where genomic feature locations, given as a chromosome name and start and stop positions, are given row wise. The basic genomic track format, just described, can serve as a data structure to store more information than just genomic locations. New columns of information are then added to the file. A frequently used genomic track is the definition of genome locations for all genes of an organism. It is frequently expanded to include more information. A simple example is the addition of gene expression values for each gene. Genomic tracks have become central for storing, manipulating and analyzing the reference genome alignment information of the sequence tags generated by next generation sequencing experiments. Genomic tracks can also be used for storing 3D interaction data [31]. The reference genome is a consensus sequence representing an average of a particular organism, meaning species. Figure 1.3 on page 11 shows all the biological subclasses or entities within a species, for which the reference genome can serve as a data structure. Genomic and epigenomic information on populations, families, individuals, cell populations and cell types can be stored as genomic tracks, that relate to a reference genome of the organism, from which the data is collected.

Figure 1.3: The biosphere is connected by a large family tree that could be called Mount Improbable in reference to the book "Climbing Mount Improbable" [32] by Richard Dawkins. The figure shows how Mount Improbable can be divided into smaller family trees and is ultimately made up of individuals. Individuals are made up of populations of cells. Cell nuclei contain chromatin folded in the three dimensional space. Unfolding chromosomes into a straight line forms a one-dimensional coordinate system onto which genomic and epigenomic features are positioned. Data on chromatin components is frequently stored in sets of pairs of genomic coordinates called genomic tracks. The environment space figure in the lower right corner suggests that for each individual there is a specific environmental niche that it interacts with.



1.4 Gene expression determines cellular morphology and function

A human is made up of about 200 different adult cell types, the number varies depending on the definition. Such cell types mature step by step out of embryonic cells by a process called differentiation by rounds of cell divisions, referred to as the lineage of the adult cell. The form and function of every cell type and state are determined by the varying presence of molecular structures and machineries. These functional units of the cell are made of proteins and RNAs. Proteins and RNAs are, through transcription and splicing, synthesized from templates encoded as subsequent blocks of DNA sequences, called exons, located in genomic regions, called genes. The level of transcriptional activity of a gene is tightly connected to the positionally related epigenome. The presence of epigenetic factors and transcription factors modulate the level of transcription of the gene. Transcription factors are generally small proteins, which bind to a specific sequence motif of less than ten bases, located either proximal to the gene, in the promoter or the gene body, or at more distal recognition sites called enhancers. Histones are larger proteins, which are less preferential in what sequences they bind to. They occupy 147 bases long DNA segments, by interacting with them as octamers. The chromatin composition is determined by cellular programs of cell divisions and differentiation, signaling from other cells of the body and environmental responses. A large scale study of the gene expression profiles in various human and mouse tissues is presented in [33] . The data is available from a database and web interface called BioGPS. All or a subset of the exons of the premature RNA are, after transcription, enzymatically cut out and pasted together. The exon cutting and pasting to form the final RNA product can usually be executed in alternative ways, through a process called RNA splicing. Functional RNAs are the end products of the expression of non-protein coding genes. Messenger RNAs, however, serve as intermediate information molecules between the genetic code and the protein alphabet. The expression of a protein-coding gene includes one further step of molecular conversion, where the messenger RNA is translated into a protein. Due to RNA splicing, believed to be epigenetically regulated, one gene can produce many RNA and protein products. Proteins can also be post translationally modified, increasing the possible number of functional products that the cell's repertoire of genes can produce. These and other processes make the number of protein structures that can possibly be generated out of the 25 thousand existing human genes, staggering [34]. The measuring of the transcriptional activity of genes is in that sense not cer-

tain to reflect the activity of its functional end product, which is usually the sought information. This type of experiment referred to as gene expression profiling, is, however, the most common way to analyze global gene expression, since a single experiment can capture information on the activity of all genes in the genome at once.

1.5 The promise of therapeutics through epigenomic modulators

A major motivation for studying the molecular mechanics of cellular change is to understand disease development, to be able to detect individual predisposition to disease at an early stage and to be able to apply customized therapeutics. It has turned out that histone modifiers and chromatin remodelers, enzymes responsible for shaping the epigenomic landscape are frequently aberrant in some cancers [2]. They are further described in Section 1.9.2. They have even been classified as driver genes in some tumors [3]. This has lead to the screening of drugs against malfunctioning histone modifiers and clinical trials are already on the way for some drugs. Pharmaceuticals have already been introduced as modulators of histones and other signaling proteins (oncogenes). Examples of targets for such small-molecule inhibitors for approved medicines are DNMTs, HDACs, and JAK2. A review of which histone modifiers have been found to be mutated and in what type of cancers is given in [2]. The review also covers recent findings of mutations in non-coding RNA and in histone genes. Genes of proteins responsible for the maintenance of DNA methylation, DNA methyl transferases (DNMTs), have also recently been shown to be frequently implicated in some malignancies. In [35] DNMT3A was reported deleted with a sample recurrence of up to 25% in patients with acute myeloid leukemia. Despite these therapeutic advances in cancer treatment it remains to determine why and how pharmaceuticals/inhibitors work. Revealing mechanisms of chromatin biology through integrative epigenome analysis can contribute to this effort.

1.6 Mechanistic inference from association of alterations

The overall purpose of integrating epigenomic data is pretty much summarized in the mission statement of the consortium The Encyclopedia of DNA

Elements (ENCODE). ENCODE is further discussed in Section 1.14. The mission of the consortium is to functionally annotate all parts of the genome. The function of a genomic region varies between cellular types and states, though. It is heavily debated to what extent all genomic regions are functional. By the current definition of biological function, a genomic region must have been selected for by natural selection to be functional [30]. It is, however, not trivial to establish whether a sequence has been selected by evolution or not.

Genomic regions are involved in different cellular processes at different points of time, just as genes are transcribed and replicated at different points of time. For the purpose of discussing inference of casual relations between genomic and epigenomic features, regulation of gene expression will here be used as an example. For such an analysis a genomic track of genes and their relative transcriptional activity in case versus control is then integrated with other alteration data between the same case and control that could explain the expression levels. There are many challenges to revealing any causal relation in such an approach:

- Genes are different in the way they are regulated.
- Due to biases it is not ideal to compare transcription levels between genes.
- Chromatin biology within a cell nucleus, including the transcription of genes, is enacted in three dimensions so that chromatin, distal in one dimensional space, or from separate chromosomes, can interact to determine the level of transcriptional activity.
- The same epigenomic component can have opposite effect on transcription depending on where it is located.
- Even though the integrated data has generally been collected from a single point of time it reflects events, that have taken place over time, possibly in different cellular processes across cell cycles and sometimes across generations. This allows for random events like mutations to have been compensated for by epigenetically mediated responses.
- Integrative analysis of genome wide data is commonly based on the selection of genomic segments, in which to look for the association. The specification of these segments must be done based on assumptions and generalizations. The promoter region of genes is, for instance, generally specified as 2kb upstream and 1 kb downstream of TSS.

Being familiar with the current understanding of the dynamics of chromatin

biology can, therefore, be of assistance in designing, analyzing and interpreting integrative analyses.

1.7 The history of the haploid genome representation

After about twenty years of genetic research being characterized by Sanger sequencing of human genes and small genomes as well as gene knockout and insertion studies [36, 37], the utility of a human reference genome surfaced. After collaborative efforts of dimensions never seen before in the field of biological research, a draft reference sequence was published in 2001 [38, 39]. Simultaneously with the ongoing projects of sequencing the human genome, a project for mapping the genetic basis of trait variation (and disease) was initiated. SNP discovery started when assembled genomic sequences were annotated at base pair positions of discrepancies between aligned reads. These heterozygous sites were reported as an SNVs (a Single Nucleotide Variation within the sample) and as a candidate SNPs (a Single Nucleotide Polymorphism, a variation existing with a frequency in a population).

Heterozygous sites indicated that different variants had been inherited paternally and maternally and, therefore, that the sites were polymorphic. The importance of genome wide polymorphism data for estimating genetic differences among humans was soon recognized. When the first human genome was assembled it was also annotated with single nucleotide variation.

In [38] the genome was presented as a haploid genome with sites of variation, while in [39] the genome was presented as diploid. Haploid presentations of human genomes have dominated since then, partly because of the large increase in complexity of storing and managing a diploid genome. The aligned sequence fragments had no information on whether they belonged to the maternal or the paternal chromosome of the homologous pair. The SNVs detected by alignment could, therefore, not be annotated with chromosome identity. Thus chromosome sharing, i.e. haplotypes, was not given directly from the raw data. The word haplotype is yet a biological term with dual meaning. Except for the definition used above, it can also refer to a block of SNPs on a chromosome that is in linkage disequilibrium (LD) with each other. Computational methods for estimating haplotype probabilities, referring to the LD-block definition, have later been developed. Such derived haplotype information does not correspond to diploid information, due to, among other things, the lack of gametic phase information. The ini-

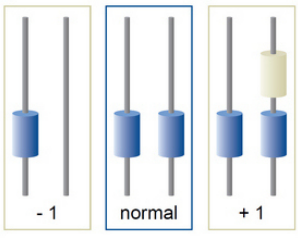
tial inability to capture diploid information has likely contributed to a reluctance in the development of storage formats and visualization tools for diploid genome information. At the time of the human genome project, it was known that variation in the number of copies of large regions of the genome could cause disease, as detected with cytogenetic methods in cases like trisomy 21 (Down's syndrome) [40]. Cytogenetic studies had revealed that some cancers and congenital disorders had genomic regions that deviated from having the regular one maternal and one paternal copy. This made their maternal and paternal genomes different in terms of length as depicted in Figure 1.4 on page 17. The chromosome pairs of genomes of healthy individuals were, however, at the sequencing of the first human genome, considered as being of the same length and to only vary between each other and between individuals in terms of sequence content in the form of, for instance, SNPs.

A large fraction of the SNPs, reported in the first large scale generation of human SNP data [41], was a result of the sequencing effort of the human genome. Follow up studies, of which the HapMap study [42, 43] is the most prominent, using many individuals from different populations, were performed to verify the reported SNPs. A possibly even more important contribution by that study was a description of genomic regions that are generally inherited together and not separated by recombination, the haplotypes. More recent studies, like the 1000 genomes project [44], are revealing even more polymorphisms. The existence of polymorphic markers, and the fact that maternal and paternal chromosomes recombine in the germline, opened up many possibilities for genome analysis. Important was the possibility to statistically associate a polymorphism to a disease/trait through linkage analysis, based on studying the similarity of their segregation pattern in family trees. Polymorphisms could also be associated with diseases/traits if they were found to be overrepresented in a case population as compared to a control population in GWAS. Haplotypes and SNPs identified in these studies, mainly those located outside genes and which remain to be functionally annotated, are now integrated with epigenomic information, primarily generated by the ENCODE consortium. The software HaploReg and RegulomeDB have been developed for that purpose and are described in Section 1.13.

It was first after the emergence of the microarray technology, that it became evident, that even healthy people could differ in (smaller size) copy number [45]. The use of microarrays allowed for detection of copy number variation at a higher resolution than previous methods had done, and led to the surprising discovery in 2004 that genomes, due to variation in the copy number

of shorter sequences, vary in length also among healthy individuals [46, 47]. Such variations, that occur in more than 5% of a population, are referred to as Copy Number Polymorphisms (CNPs). SNPs were for some time believed to affect more bases and to be more frequent than CNPs, but according to recent estimates of CNVs, the opposite is now known to be true [17]. Immediate questions of interest concerning CNPs are how they contribute to human phenotypic variation, where in the genome they occur, what the resulting distribution of human genome lengths is, when they were formed and how they were formed. Recent publications [48–50] have been able to approximate an answer to some of these questions. Most CNVs are relatively frequent in the human population and are believed to have been generated a long time ago. Most of them seem to follow the same haplotype pattern as SNPs and their contribution to disease would, therefore, have been revealed by association studies already performed. Hence, the conclusion is that new association studies using CNPs will lead to few new discoveries of disease risk loci [51]. The HapMap study samples, that originally were used for analyzing SNVs in different populations to identify SNPs and haplotypes, have now also been used to identify CNPs using microarray technology [48]. This study reports that in the sample two genomes on average differ by 1,098 CNV with a cumulative length of 24 Mb (0.78% of the genome). All differences found between the 41 studied samples encompassed 8,599 CNVs with a total coverage of 112,7 Mb (3.7%) of the genome. Some of these findings are reviewed in [52].

Figure 1.4: Somatic Copy Number Aberration, SCNA, Copy Number Polymorphism (CNP). Gain and loss of genomic sequences usually lead to the change in length of the maternal or paternal chromosome. Copy numbers are generally annotated as variation against a haploid reference genome. The location of the amplified or deleted sequence is annotated with a copy number deviating from the normal copy number of two. Sometimes a gain of copy is accompanied by a rearrangement so that the new copy locates to a new genomic site. Courtesy: http://www.imgm.com/index.php?id=5109



1.8 Separation of sample populations and diploidy

Most high throughput data collection methodologies, including next generation sequencing, require a relatively large amount of DNA, which brings along that the sample must be collected from a population of cells. The fact that the sample consists of many cells is not a problem as long as the cells homogeneously represent the feature, that is measured. A cell population sample from a healthy individual carries genomes, which are similar enough for sample heterogeneity not to be an issue. Regarding epigenomes, it is an open question to what extent they vary between cells in such a population [53]. For a cell population sample from a tumor, both genomic and epigenomic heterogeneity can be expected, since the sample will represent different stages of a clonal expansion [54]. The majority of published analyses has been performed without attempts to resolve the issue of sample heterogeneity. Analytical approaches have been applied to try to resolve epigenetic sample heterogeneity [55, 56]. Similarly, methods have been developed to resolve heterogeneity in the numbers of genomic copies in cancer samples [57, 58]. New methods are also developed to restrict the used sample size for DNA extraction [59, 60]. Taking diploidy into account is becoming increasingly feasible with next generation sequencing technologies, and increasingly important, because of the intent of integrative studies to infer mechanistic relations. For the same reason, the diploidy of the epigenome should be taken into account. As an example, an amplification of a heterozygously methylated locus can be considered. The amplification of an epigenetically silenced allele will be neutral to the expression of the gene, assuming that the epigenetic state is inherited to the gained copy. A dosage effect will occur, though, with the amplification of a transcriptionally active allele. Moreover, assuming the genome to be diploid in its entirety is generally incorrect. Many regions of the human genome have been reported to vary in its number of copies also among healthy people [61].

1.9 Chromatin dynamics

Cellular differentiation is driven by circuits of gene expression and gene regulation. The diploidy of the genome allows for parallel versions of these circuits, and possibly networking between the versions. Some circuit connections are pathways of transcription factor based gene regulation. Other

connections are pathways based on the production and genomic deposition of epigenomic components. Possible connections between genes expressed in the early embryo, the formation of epigenomic landscapes, and genes expressed in adult cells, is the subject of the following subsections. Starting with how the asymmetric expression of master regulators and developmental genes in the embryo and fetus, respectively, enforces polarity and the body plan. Continuing with their possible regulatory link to the expression of histone modifiers, and proteins with related functions, which in turn will mold the epigenomic landscape. Epigenomic occupancy in distinct genomic regions positionally associated to the transcription start sites of genes, either on the one dimensional DNA sequence, or through three-dimensional interactions will regulate the expression of genes. Detailed knowledge of global chromatin architecture, along with these *cis*-regulators, represents a crucial step towards understanding how genetic, epigenetic, and environmental or stochastic factors drive context-specific genome regulation [62].

1.9.1 Master regulators in the early embryo

The lives of multicellular organisms start with conception and the formation of a single diploid zygote of two haploid germ cells. It is likely that chromatin states inherited from the germ cells contribute to determine the initial transcriptional activities of the zygote. The zygote will through asymmetric cell division, in which the mother cell forms one differentiated and one undifferentiated (stem) cell, give rise to all the cells types of the adult organisms. How asymmetric cell division is generated is not known, but models have been proposed [63]. Genes expressed in embryonic stem cells in the early development of an organism are called stemness factors. Examples of such factors are Sox2, Oct3/4 and Nanog. The expression of these is believed to maintain the pluripotent stem cell state. In differentiating cells another set of genes called differentiation factors, of which examples are GDF1-15 and TGF-B, contribute to cellular decisions on future differential programs or lineage commitment. Much of the regulatory circuitry connecting differentiation factors and other genes, expressed in the early embryo, through epigenomic modifiers and remodelers, and via epigenome organization to the expression profiles of adult cells, remains to be understood.

At later stages in the differentiation other developmental transcription factors, like the HOX gene family, become important for regulating transcription according to the body plan.

1.9.2 Epigenomic modifiers and remodelers

Mitotic epigenetic inheritance requires the epigenomic equivalent of genomic replication. At the same time different cell types have different epigenomic landscapes. Mechanisms for histone production and the deposition of histones genome wide are currently being mapped [64, 65]. Epigenomic modifiers and remodelers lay down, modify, reposition and remove the epigenome.

Histone modifiers are proteins that add and remove chemical groups, like the methyl, acetyl and phosphate ones, to various sites at the amino acid tale of histones. The modifiers recognize epigenomic signatures with protein units called chromatin readers. Many types of proteins interacting with the chromatin have chromatin reader subunits [66]. Nucleosome remodelers can move/translocate nucleosomes along the genome. Here follows a short description of the known categories of genes contributing to the maintenance of the epigenome. Chromatin modifiers are:

1. Enzymes adding methyl group to DNA: DNA (cytosine-5-)-methyltransferases (DNMT1, -3A and -3B)
2. Enzyme removing methyl groups from DNA: Ten-Eleven Translocation(TET)
3. Histone lysine acetyltransferases (KATs)
4. Histone deacetylases (HDACs)
5. Histone methylation: Histone lysine methyltransferases (KMT)
6. Histone demethylation: Jumonji demethylases
7. Histone phosphorylation: Kinases active in the nucleus
8. Histone dephosphorylation: Phosphatases active in the nucleus

Integrative analyses usually include correlating the activity of modifiers of the epigenome with the abundance or profiles of their target components of the epigenome. So is, for instance, the level of promoter methylation of the samples studied in **Paper I** correlated with the expression of the DNA (cytosine-5-)-methyltransferases (DNMTs). DNA methylation has a crucial role in differentiation and cancer. DNA methyl transferases DNMT3A and DNMT3B have been shown to add methyl groups *de novo* in relation to differentiation. DNMT1 has, on the contrary, been shown to maintain methylation patterns across cell divisions. An enzyme for active removal of DNA methylation, Ten-Eleven Translocation (TET), has also been identified. It is active in, for instance, the germline, when the methylation pattern of the

whole genome is known to be erased. The activities of these enzymes are reviewed in [67].

Histone lysine acetyltransferases (KATs) and Histone deacetylases (HDACs) are responsible for histone acetylation deacetylation, respectively.

Histone lysine methyltransferases (KMT) and, for instance, Jumonji demethylases are responsible for histone methylation and demethylation, respectively.

Kinases are enzymes that functionally modify other proteins by phosphorylation and they are frequently found to be aberrant in cancers. They are abundant in the cytoplasm of cells, but are also believed to be located in the nucleus. Phosphorylation is a common modification of histones. It could be that some of the kinases, that are frequently disrupted in cancers, are responsible for histone phosphorylation.

1.9.3 Epigenome organization and regulation of gene activity

The involvement of DNA methylation in gene and transposon silencing [68], imprinting [69] and X chromosome inactivation suggests that it is a tool that cells can use to regulate gene transcription. In [70] the plasticity of the methylome of different progenitor cells and its relation to gene expression is investigated. The relation of the epigenome to gene regulation has been the subject of many recent studies based on ChIP-Seq technology [25, 71–73]. These global studies validated many associations suggested from earlier locus specific ones and demonstrated some associations for the first time. The ChIP-Seq based studies further revealed the important functional consequences of chemical modification of the tails of histones. Associations indicate, but do not prove, the existence of causal relations. A few studies have been able to capture causation [74, 75]. In one of these studies the effects on DNA methylation and gene expression were monitored after the deletion of methyl transferases. CpG islands were subsequently depleted of methylation and linked genes were activated. Evidence for causation in the opposite direction has been demonstrated as well, proving that gene regulatory events can modify the methylation pattern at nearby sites [76] and that chromatin modifiers and nucleosome remodelers can be utilized in transcriptional processes [22]. Studies of gene regulation continuously reveal and define new types of genomic segments functionally associated to genes, for which epigenomic, together with other transacting factors like transcription

factors, occupancy regulates the transcriptional activity. The role of these regions, in epigenomic regulation of gene expression is reviewed in [62, 77]. A term for the collection of these regions, the Cistrome, was recently coined [78]. The regions are: the body of the gene itself, the promoter (+/- x kb of TSS) that can be populated by CPG islands, the transcription start site (TSS), distal enhancer sites, distal regions containing CpG shores [79], the genomic segments covering the whole (3D) environment of the gene. Here follows a review of important findings on their possible occupancy profiles of epigenetic factors and how the combination of these profiles switches the gene between three identified transcriptional states namely silent/inactive, poised and active.

Promoter Gene promoters are commonly divided into classes depending on their CpG content [73]. Most promoters (65%) tend to have associated regions with higher CpG content than the genome average called CpG islands. They are called High CpG content Promoters (HCP) and are believed to differ in the way they are regulated from Intermediate CpG content Promoters (ICP) and Low CpG content Promoters (LCP) [80, 81]. Most HCP genes are, in contrast to the other classes, occupied by H3K4me3 and have unmethylated CpG sites, or are hypomethylated, independent of their expression state. They are also believed to acquire an active state by default, while the other classes do not. HCP genes tend to be silenced by occupancy of H3K27me3 modification likely to be deposited by the Polycomb repressive complex. This is a repressive state that is easy to reverse into an active state. Promoters marked by both H3K4me3 and H3K27me3 are called bivalent indicating that the genes are poised for expression when needed. This type of repressive state is common in embryonic stem cells, targeting developmental genes that encode transcription factors and other regulators of cellular state, and is likely to contribute to the ability of these cells to reprogram and differentiate. Silencing of the other classes of genes is believed to be of the long-term kind, within heterochromatin, and identified by occupancy of H3K9me3 and hyper-methylation. LCP genes are believed to be expressed mostly in terminally differentiated cells.

CpG island It has been observed that the frequency of CpG sites in the human genome on average is less than expected from the frequencies of C's and T's [82]. A reason for the genomic depletion in CpG sites is that C in CpG tends to get methylated. Methylated Cs tend to spontaneously deaminate to form Ts. There are, however, regions of the genome where the CpG frequency rises to the expected one [83]. This

can be because CpG 's in these regions are rarely methylated and/or selection pressure prevents mutation in these regions. The regions are called CpG islands. Most genes have CpG islands coinciding with their promoter regions upstream their transcription start sites.

Transcription Start Site (TSS) The stability of the occupancy of nucleosomes, or how well positioned they are, varies along the gene. The stability is often visualized with aggregation plots, described in Section 1.13.3, using TSSs as anchor points [84]. Aggregation plots are also frequently used to demonstrate the distribution of individual histone marks around TSSs [25, 71]. The first nucleosome downstream of the TSS, the +1 nucleosome, is the most well positioned. In actively transcribed genes the site upstream of the +1 nucleosome is not occupied by any nucleosome and is referred to as either the nucleosome depleted regions (NDR) or the nucleosome free regions (NFR) [85–88]. Much of the research on NDRs/NFRs has been performed on yeast [89–91] although the patterns of aggregation plots are similar between organisms [90, 91].

CpG shore Genomic locations up to 2 kb proximal to CpG islands have been identified that vary to a large extent in their methylation states and seem to have great influence on gene expression [79].

Gene body The level of H3K36me3 occupancy in the body of a gene is associated with the level of transcriptional activity. The modification tends to occupy the bodies of transcribed genes [25]. H3K36me3 has also been shown to preferentially occupy expressed exons as compared to introns and exons not used and thereby demarcating the splicing of the gene. Methylation of CpG sites within the body of the gene has been demonstrated to promote transcription [92].

Enhancer Enhancers are genomic segments with regulatory influence on distal genes and are recognized by transcription factors and chromatin regulators [93]. Therefore, even though enhancers can be identified using a combination of chromatin marks, it remains a challenge to map each enhancer to the gene it regulates. Software has been developed to improve the mapping between genes and their enhancers [78]. Active enhancers have been shown to be enriched by H3K4me1 and depleted of H3K4me3 [72]. The chromatin pattern at enhancers is more variable across cell types than the pattern at promoters. This suggests that enhancers are more important in cell type specific gene regulation. How H3K4me1 is deposited at enhancers is yet not known.

Insulator Special motifs in the genome, like CCCTC, seem to be targets for DNA interacting proteins like CTCF. Through its ability to interact with many different proteins, CTCF is believed to have multiple diverse functions including transcription regulation and insulation of enhancer activity by forming long-range interactions and chromatin loops. Thus, CTCF and its binding sites contribute to the global higher order chromatin structure and to the formation of chromatin domains [94].

1.9.4 3D organization

The 3D conformation of chromatin is known to be dramatically different between cell cycle phases. During mitosis, the chromatin of each chromosome forms distinct shapes of either X shapes or rods. In interphase, the chromatin is known to be less dense in its structure, forming chromatin territories. The shape and positioning of chromatin territories are believed to vary between different cell types and, to some extent, reflect that actively expressed genomic regions tend to be kept closer to the center of the nucleus, while silent transcriptionally inactive ones are kept closer to the nuclear membrane. Objectives of integrative epigenome analysis are to map how the epigenome changes, and is influenced by the three dimensional organization of chromatin in the nucleus, its genome wide localizations and its interactions with other omics data during differentiation of healthy cells, and during cancer progression. It has only recently become possible to study the three dimensional organization of chromatin and epigenomic interactions. So, relatively few datasets, software and studies are available. The data requires the same considerations as other next generation sequencing data. 3D interaction data is collected from a population of cells and it represents averages within that population. The data is also collected as a haploid genome representing averages of interaction for paternal and maternal chromosomes. Heterogeneity between the cell population and between maternal and paternal chromatin might generate significant sources of noise. Pioneering studies have revealed a few interesting conditions. Here are some examples:

1. The time in which genomic regions replicate can be determined using nucleotide color labeling at replication [95]. Such studies have revealed that the genome can be segmented into time zones from early to late replication. The segments generally span multiple origins of replication. Specific histone marks as well as transcribed and poised genes

have been associated with early replication time zones, while other histone marks and long term silent genes have been associated with late replication timing [96–98].

2. A study of the locational association in the three dimensional nuclear space of sCNA break points and replication timing data from a tumor sample was published in [99]. The study reports that sCNA breakpoint locations share replication times and are close in 3D. The analysis, performed within the R language environment, suggests that many of the CNAs have arisen through interference between adjacent replication forks.

3. It was shown that genomic regions, with a low gene density and with little transcriptional activity in fibroblasts, are interacting with the nuclear lamina. Hence, they were termed Lamina Associated Domains(LADs). This indicated that genomic regions with low expression activity are located in the nuclear periphery [100]. In another related study the genomic locations of the H3K9me2, a mark of long term repression, were found to overlap significantly with LADs.

4. The polycomb group (PcG) proteins, known to be crucial to dynamic transcription silencing of large genomic regions in cellular differentiation, have been shown to form agglomerates, called polycomb bodies, in the three dimensional nuclear space [101]. Another silencing histone mark, H3K27me3, has been shown to co-locate with PcG complexes [102]. Functional dependencies between H3K27me3 members of the PcG complex have also been shown to exist [103].

5. Transcription tends to occur at discrete three dimensional sites, similar to the polycomb bodies, but are in the case of gene activation, called transcription factories. No particular histone mark has been associated with transcription factories, however. There is no evidence of transcription factories occuring either in center or in the periphery of the nucleus [104, 105].

1.10 Genomic and epigenomic alterations in disease

A genomic disorder is a disease that is caused by an inherited genomic rearrangement. A complex disease is caused by several genes that, together

with environmental and life style factors, affect the risk of getting the disease. Most genomic disorders are due to rearrangements spanning several genes, which makes it possible to classify them as complex diseases. For many complex diseases, such as autism and schizophrenia, genomic rearrangements have been identified as an underlying cause only in some of the patients. An aberrant copy number of a gene can have a disruptive effect for different reasons. It can be due to the gene dosage effect, so that the extra copy or the missing copy affects the expression of the gene. Not all genes are, however, dosage sensitive. The disruptive effect can also be due to that the added copy might be positioned in a region lacking regulatory sequences and the right 3D environment for transcription, the so called positional effect. Yet another reason is that deletion of a healthy allele can unmask a recessive mutation. Even though the list of reported onco and tumor suppressor genes disrupted by copy number aberrations is long [106], only a few studies have been devoted to thorough analysis of the existence of the dosage effect [107, 108].

DNA methylation has recently been implicated in a number of complex diseases that are related to loss of imprinting and to repeat instability [109]. Mutations in DNA methyl transferases have long been known to contribute to developmental abnormalities.

Cancer has for long been characterized as a disease caused by inherited (germ line) and sporadic (somatic) point mutations and genomic rearrangements. More recently, the evidence, not least in the form of mutations of epigenomic modifiers and remodelers, of epigenomic alterations being part of the cause of cancer are accumulating [110, 111]. Cancer is, however, generally not considered a genomic disorder because (most of) its rearrangements have been somatically acquired so that they are not germline inherited and limited to the cells of the cancer tissue. A few types of cancers are heritable and the primary cause is genetic. Most cancers, though, tend to occur sporadically and be influenced by lifestyle factors. As for these types there are many reasons to believe that the epigenome plays a central role in the etiology. Many cancers likely develop as an interaction between changes in the genome and the epigenome [3] with an accumulation of mutations and deregulations of a smaller set of driver genes and a larger set of passenger or hitchhiker genes [112]. Driver genes are defined as those changing the cellular phenotype and making it acquire the ”hallmarks of cancer” [113], of which examples are self sufficiency of growth signals and insensitivity to anti growth signals. Passenger genes are those that do not contribute to the tumor transformation. The epigenome influences the integrity of the genome. The majority of the human genome consists of repetitive sequences. In these regions most of

the methylated CpGs of the normal genome are found. It is believed that the methylation protects against the disruptive effect of repeat sequences. Methylation has also been shown to protect against non allelic homologous recombination, a frequent cause of copy number alteration [114]. Repetitive sequences tend to get hypo methylated in cancers. Some of the sequences, like LINEs and SINEs, are parasitic retrotransposons. It has been suggested that DNA methylation once evolved as protection against the disruptive effects of such parasitic sequences [115]. Other known effects of hypo methylation in cancer are oncogene up-regulation and loss of imprinting. Known effects of hyper methylation in cancer are silencing of tumor suppressor genes and non-coding RNAs by hyper methylation of CpG islands in promoters and of CpG shores. Many expressed genes have high levels of DNA methylation within the gene body, suggesting that the context and spatial distribution of DNA methylation are vital in transcriptional regulation. Recent observations might explain why and how different sets of genes tend to get methylated in different cancers. Genes, that are expressed in a cell type specific manner, and maintained in a poised state in the normal tissue by polycomb repression tend to also be the ones that become methylated in cancers of the same tissue [116, 117]. These observations have contributed to the formulation of the tumor stem cell hypothesis [115]. Changes in the normal patterns of occupancy of histone modifications have also been observed in cancer. Methylated repeat sequences are also occupied by H3K20me3 and H4K16Ac in normal cells. This occupancy pattern tends to vanish in cancer cells [118].

1.11 Technologies for data collection

Micro array and next generation sequencing technologies have been fundamental in the transition from locus specific studies to genome wide ones.

Second generation sequencing, a group of technologies that, compared to older technologies, allows for comparatively fast, easy and cheap retrieval of sequence information, from very long sequences like human genomes, without positional restrictions. They are based on simultaneous sequencing of massive amounts of about 100 base pairs long DNA fragments. There are many variations to constructing DNA fragment libraries for such sequencing. Some of them include steps where various technologies, like ChIP [119] or sodium bisulphite treatment [120], are used to select only the DNA fragments that are marked by, or interact with, a particular epigenetic factor or are transcribed.

Technologies for data collection in genomics are under constant development towards higher coverage, resolution, speed and lower costs. While contributing to a continuous improvement of the understanding of chromatin biology, this development also brings along a few challenges. Examples are the constant change in data formats so that old analytic pipelines cannot, without modification, be applied to new data and that the appreciated value of data tend to have a best before date, putting a pressure on labs to analyze the data fast. The best before date is also a challenge for maintainers of data repositories. For each type of data and sample, there might exist multiple versions acquired with technologies of different dates.

There are many variations to each of the technologies described below. This section is intended to describe the general principles of their application.

The data analyzed in **Papers I** and **II** in this thesis was acquired by oligo nucleotide micro array technology. Oligo nucleotides are short, single-stranded DNA. They can be computationally designed and printed onto a silicon surface/bead.

Microarray technologies can be used to determine which sequences of a DNA library that are present in a sample and to estimate their abundances in the sample. They are constructed by deposition of oligonucleotide probes, single stranded sequences, in a systematic and recorded pattern to a silicon plate. The purpose of the probes is to hybridize with and thereby immobilize its antisense complementary sequence if it is available in the sample. Common to most microarray experiments is that the analyzed signal represents the intensity difference between two samples. One of the samples is the control, normal or reference sample while the other is the case or tested sample.

In the case of gene expression profiling, the array should contain at least one probe for each transcript of the genes. To get a detectable signal probes must be printed in clusters of identical sequences. To perform an experiment (hybridize DNA sample to the array) the microarray is first washed with a library of DNA synthesized from, and complementary to RNA transcripts (cDNA), with each cDNA being marked by a reporter molecule. Unhybridized cDNA is subsequently rinsed off. A photo is taken of the array. Probe clusters, where hybridization occurs with labeled target cDNA, will appear as stained spots in the image. The abundance of the cDNA in the sample will affect the number of transcripts that bind to its corresponding probe cluster, which can be estimated by the intensity of its corresponding spot on the photo.

Designing probes for gene expression profiling is a bioinformatics task. Important to take into consideration is that a probe should match no other gene but the one that it is supposed to represent, and repeats and self complementarily sequences should be avoided. Finally, similar melting temperatures for all probes should be striven for to avoid problems at the hybridization step.

To determine the genotypes at sites of SNPs using array technology, two array probes must be constructed for each SNP, one probe that is the reverse complement of the major SNP variant and another that is the reverse complement of the minor SNP variant. The intensity of each hybridization reflects the degree of presence of the hybridized molecule. Hence the intensity can be used to assess the number of copies of the variants of a loci. For copy number estimation the sum of the intensity of the genotype at a SNP in the case sample is compared with the sum of the intensity of the genotype at that SNP in the matched control sample. To probe the genome for variation and copy number at the highest possible resolution microarrays for SNP detection tend to have a relatively large number of probes. The array used for copy number derivation in **Papers I** and **II**, Affymetrix Genome-Wide Human SNP Array 6.0 has about two million probes. About half of the probes on the array matches SNP loci, but the rest do not match polymorphic sites, and are only used for copy number estimation.

The principle of SNP microarrays can be used to determine the methylation states of CpG sites. It requires a conversion of the chemical modification of DNA that CpG methylation is into a variation in the DNA sequence [121, 122]. Methylation of C protects it from the mutating agent sodium bisulphite. By treating the DNA with sodium bisulphite unmethylated C will mutate into T. A variation in the DNA methylation state has thereby been converted into a single nucleotide variation (SNV). By designing a SNP array that distinguishes Cs and Ts at the site of C in CpGs of bisulphite treated DNA the methylation states of the untreated DNA can be determined. The array used to determine CpG site methylation states in **Paper I**, Illumina HumanMethylation27 BeadChip, targets 27000 CpG sites in CpG islands of promoters of about 14000 CCDS genes. There are many different types of pretreatment and analytical steps for determining DNA methylation. They are reviewed in [123].

The data analyzed in **Papers III** and **IV** in this thesis was acquired by the ChIP-seq technology, which is really a sequential application of chromatin immuno precipitation [124] and second generation sequencing [125]. Chromatin immuno precipitation is accompanied by many other technologies

that have turned high throughput when combined with sequencing. This has allowed for genome wide characterization of new aspects of chromatin biology. A variety of methods, DNase I–seq [126, 127], FAIRE–seq [128] and Sono–seq [129], have been developed for the mapping of open chromatin. Another application, where nucleosome positioning, and turnover is measured genome wide, is described in [130] .

Chromatin immuno precipitation is a method for making sections of crosslinked and sonicated chromatin ”fall out” as a solid in a solution by specific binding of an antibody with a given DNA interacting protein of interest. The method depends on the availability of an antibody that binds specifically to the DNA interacting protein. Chromatin immunoprecipitation was before the availability of next generation sequencing technology combined with microarray technology in a method called ChIP-chip. The genomic regions that can be queried by microarray technology are limited and decided at the time of the production of the array. This is why the transcriptional activity of many non protein coding genes has passed by undetected by microarray experiments. Arrays cannot be used for detection of somatic mutations in cancer since these somatic *de novo* mutations which are not identified by studying the normal variation within a human population.

By using next generation sequencing technology the activity of a component of the chromatin will be detected wherever in the genome it occurs. This is because next generation sequencing technology enables the mapping of the genome wide locations of a component of the chromatin as well as provides a measurement of the degree of activity or probability of occurrence at the locations. DNA for ChIP-Seq experiments is collected from millions of cells from a population that should represent a cell type and state. Genomic sequence fragments interacting with a given protein are mapped onto a reference genome. This determines the genomic locations of the interactions and allows for estimating the intensity or probability for an interaction at that locus within the population. The Illumina platform for next generation sequencing will here be used to explain general principles. It is one of the most, if not the most, widely used technologies and it is used for the majority of experiments behind the data in Papers III and IV.

For Illumina sequencing, for which the technological principles are schematized in Figure 1.5 on page 33, only two types of oligo nucleotide sequences are immobilized to the silicon plate/array/flow cell. The two types of probes are equally distributed with a fixed inter distance across the flow cell. They are used to capture library fragments, and later to act as PCR primers. The library of genomic fragments of interest is called a target library. A central

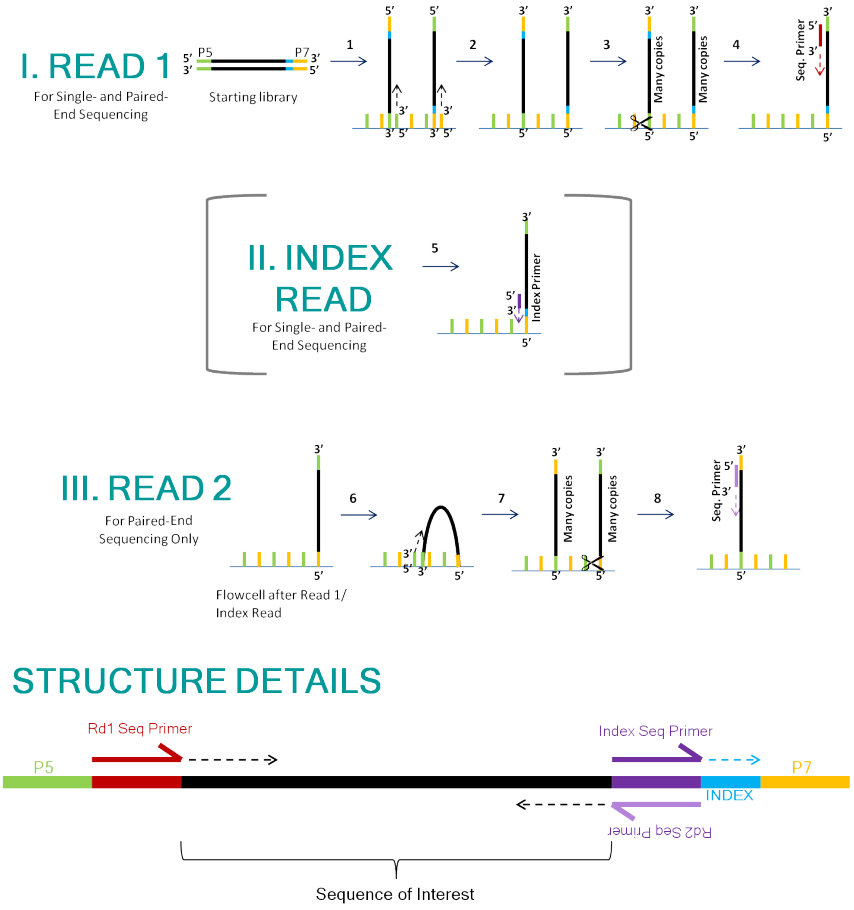
feature of the technology is the ligation of a pair of adaptor sequences to the flanks of each genomic fragment of the target library. One adaptor of the pair is ligated to one side of each fragment and the other adaptor is ligated to the other side of the fragment. The outermost parts of the adaptor constructs encode flow cell binding sequences complementary to one or the other of the oligos on the flow cell. The sequences of the adaptor constructs complementary to the array probes are called P5 and P7 respectively. They both also encode an enzyme cleavage site. A library is constructed. Single stranded library fragments are annealed to a fraction of the two types of probes of the flow cell. Many of the probes will remain free and will be used for hybridization when generating clusters of the fragments for signal amplification. This is done through an application of surface (bound) PCR amplification, called bridge amplification. The first cycle of bridge amplification uses the probes of the flow cell as primers and the annealed library fragments as templates. The flow cell probes will be elongated, upwards from the flow cell surface, to have a sequence complementary to the annealed library strand with its adapter constructs. The originally annealed strands are washed away. The flow cell will now consist of immobilized newly synthesized single stranded DNA with P5 and P7 sequences swaying at their tops as well as immobilized probes, still free for hybridization and complementary to P5 and P7. This will lead to annealing of the ends of the synthesized strands and the complementary probes. A new round of probe elongation will take place, this time using an immobilized strand as a template. Double strands will be denatured. Bridge amplification will in this way generate clusters of oligos. Each cluster will be a library fragment with flanking adapter sequences with either P5 or P7 attached to the flow cell probes. The oligos with P5 attached to the flow cell are enzymatically cut away so that all the remaining fragments of a cluster are attached by P7 and have the same direction. The adaptor constructs with P5 are also equipped with a primer sequence R1, for ”Read one”, next to the sequence of interest. Read one is sequencing from the P5 side of the fragments. The fragment strands of the clusters are subsequently sequenced by synthesis, from top to bottom, using primers with complementary sequences encoded in the P5 adaptor constructs adjacent to the fragment of interest. One by one solutions of one base are washed over the plate. Each base is dyed with a unique color. If the base is complementary to the base in the template and at the five prime end of the primer, it will be inserted. Incorporation of a nucleotide reversely terminates further base incorporation. The color of the incorporated base is monitored by a camera. The dye is removed and the replication un-terminated. A new base is incorporated. In each iteration laser is used to trigger fluorescence. The flow cell is imaged after each round of base addition generating 160 millions of simul-

taneous sequencing events reading ten billion bases per instrument run. It is possible to stop an experiment after sequencing each fragment from one end to the other, as described so far, in which case it is referred to as Single End Sequencing. It is also possible to perform Paired End Sequencing. In that case, the flow cell is stripped into its original condition by removing everything but the primer sequences. The procedure of bridge amplification is repeated, but this time sequences attached through P7 are cut away. The P7 constructs are also equipped with a primer R2, next to the sequence of interest, for ”Read two”. Read two is sequencing from R2 primer of the P7 side of the fragment and is used only for paired end sequencing. It is possible to simultaneously sequence DNA from many samples, so called multiplexing. This is achieved by using sets of adapter constructs equipped with an index code which allows for simultaneous sequencing of DNA from many samples, so called multiplexing. The antisense sequence of the R2 primer is then used as a primer for sequencing the index code positioned between R2 and P7.

In summary the adaptor sequences will enable the fragments to:

1. Anneal as a single strand to one of the oligos of the flow cell. This is what the sequences P5 and P7 at the far ends of each fragment are used for.
2. Allow for PCR enrichment of adapter ligated DNA fragments only. This is executed through bridge amplification.
3. Allow for indexing or “barcoding” of samples so multiple DNA libraries can be mixed together into one sequencing lane (known as multiplexing)

Figure 1.5: The Illumina sequencing platform can be used for both single and paired end sequencing. I.) shows how library fragments are equipped with P5 and P7 adapter constructs and hybridized to the matching probes on the flow cell. Fragments with the P5 adapter are enzymatically removed. Sequencing is performed using primers complementary to the P7 adaptor construct. II.) shows that indexes encoded in the P7 adapter construct can be used for the simultaneous sequencing libraries form multiple samples (multiplexing). III.) shows how paired end sequencing is achieved by a second round of sequencing. The procedure is the same as in I.), but fragments with the P7 adapter are instead removed. A detailed picture of the library fragments with adaptor constructs are shown at the bottom. Adapted from `http://nextgen.mgh.harvard.edu/IlluminaChemistry.html`



Next generation sequencing experiments are dependent on good algorithms for aligning the sequence reads. The contribution to recent advances in epigenomic by the development of such algorithm matches that of the development of sequencing technologies.

Second generation sequencing technologies appear likely to be replaced by new innovations of third generation sequencing. Nanopore sequencing promises, among other benefits, to provide personal genome information cheaper, faster and more accurately [131]. Workable solutions are not yet on the market, however, due to problems like the fact that several nucleotides contribute to the recorded signal [132].

1.12 Strategies for integration

One purpose of collecting genome and epigenome data is to look for indications of interactions between components of the chromatin. Despite the fact that chromatin biology takes place in three dimensions the available data has until recently been restricted to one dimensional genomic coordinates. Genomic tracks often contain information on biological states that could have causal relations. An obvious way to indicate causality between genomic tracks is to prove that they are locationally associated along the genome. The diploid nature of genomic information usually disappears when represented as genomic tracks which makes the data more difficult to analyze for true associations. Another purpose of collecting genome and epigenome data is to functionally annotate the reference genome. One could speculate that integrating more genomic tracks would lead to more detailed annotations at a higher resolution. Most integrative studies and software have up to now involved two, or at the most, three genomic tracks. The integration of histone data into chromatin states is an exception. Few guidelines are available, however, for how to integrate multiple tracks of other types of data. Section 1.13 describes software and divides them into categories based on their main analytic strategies. Pairs of genomic tracks can be statistically tested for genome wide and local positional dependency or association. This type of strategy will be referred to as **genome wide locational association**. Software, for which the main functionality is based on this strategy, is described in Section 1.13.5. Genome wide locational association resembles another type of statistical testing used in genomics, i.e. Genome Wide Association Studies (GWAS). GWAS establish associations between phenotype and genotype differences between two populations. The results of GWAS are association of variations at one or many genomic locations to a trait varia-

tion between two populations. They can be stored as genomic tracks and are of interest to the type of integrative analyses described here. GWAS results can be experimentally verified with locus specific assays.

The epigenomic composition of a genomic region reveals a lot of its functional activity. Multiple tracks of epigenomic data can be used for machine learning of chromatin states revealing whether a genomic segment is functionally active in a given cell type or not. This type of analysis will be referred to as **machine learning of chromatin states**. Commonly it is of interest to see how epigenomic data collected from specific cell types or states distributes in relation to genomic data. Data of higher resolution in terms of the hierarchy of biological entities is in other words analyzed in the context of data of lower resolution. Analyzing multiple tracks using the ”genome wide locational association” can be difficult because different similarity measures might be needed between tracks of different types. A solution to that might be to center the analysis to a genomic feature of particular interest. This type of analysis is referred to as **centric to genomic feature**. The activity of a gene is not only affected by the epigenetic state of its body, but also of that of proximal and distal associated regions outside it, likely due to their proximity and interaction in 3D. Most genome feature centric analyses are **gene centric** since gene regulation is a central theme in most integrative studies. Related studies defining the cistrome [78, 133] are centric to *cis*-regulatory regions. It is possible to center an integrative study on **sets of genomic features** as well, whether it be sets of genomic regions interacting in 3D or genes connected in a regulatory pathway. The recently arrived possibility to collect pairwise **3D interaction** data for genomic sites genome wide has made it possible to look for co-location in the three dimensional nuclear space between the analyzed tracks. Gene centric studies generally include steps in which genes are selected based on information derived from integrated data. Enrichment of the selected genes are then searched for by use of pathways, functional categories and gene ontologies. **Pathway centric** approaches have been developed where data have been integrated utilizing regulatory and signaling connections between genes. Independent of the chosen type of integrative analysis, it will generally be a challenge to compare tracks in terms of all aspects of similarities between the analyzed data. Genomic tracks with information beyond genome location have, for instance, both locational properties and amplitudes to consider. Open source software performing these types of analyses and some of their properties are also reviewed in the following subsections.

1.13 Software

No software available today for integrative analysis gets close to providing all features that could be wished for from such a tool. Many tools are mainly developed for single-track analysis and are often limited to the analysis of two tracks. Integration of more data generally requires a lot of assumptions and prior knowledge, and the analysis tends to be focused on individual loci or sets of a few loci that are known to share properties from before. Tractable features of software for integrative epigenomics:

1. Availability and ease of access of data
2. Preprocessing of microarray and next generation sequencing data
3. Dataformat conversions
4. Network and pathway perspective
5. Gene ontology perspective
6. Reproducibility
7. Transparency/Tools for Documentation
8. Scalability
9. Extendability
10. Workflows
11. Data exploration
12. Statistical testing
13. Machine learning for chromatin states
14. Peak calling
15. Motif discovery
16. Tools of analysis based on 3D interaction.
17. Secure access

1.13.1 Preprocessing

The key components of microarray analysis are study design, preprocessing, inference, classification and validation [134]. For each type of data that can be collected with oligonucleotides, there is significant room for variation in

the analytical result depending on the technology used for data acquisition, software, algorithms and parameter settings. (The room for variation in the result of next generation sequencing data is likely to be at least as large.) Preprocessing normally involves normalization [135–138] to remove systematic variation (bias), transformation to remove skewness of data and filtering. R and Bioconductor were at an early stage of the era of micro array analysis selected as an environment for the implementation of algorithms for both preprocessing and analysis. R is therefore the single programming environment offering the largest supply of tools for processing of microarray data. The preprocessing of oligonucleotide arrays for different types of data share some principles. Bioconductor has a special site for their analysis at `http://www.bioconductor.org/help/workflows/oligo-arrays/`. The analysis of copy number data requires special algorithms for inferring copy number segments out of adjacent probes with similar distribution of copy number signal intensity [139]. Preprocessing of next generation sequencing data includes quality control with software like FastQC [140] and sequence alignment with software like Tophat [141]. One analytic task is generally to infer differential signals based on the counts of aligned reads in a given region. R and Bioconductor continue to be a preferred environment for development and sharing of tools also for the analysis of next generation sequencing data. So is, for instance DESeq [142], a popular tool for inferring differential signals based on count data. Next generation sequencing data makes it possible to answer new biological questions, but also demands new algorithms [143, 144]. It can be used for understanding mechanisms underlying human gene expression variation through (eQTLs) [145] or for the characterization of alternative spicing [143]. An evaluation of methods for the analysis of differential expression using next generation sequencing is given in [146]. For many of the analytical procedures there are options for parameter settings, leading to significant possibilities, in similarity with the micro array technology, for variation in results between different analyses of the same data. When preprocessing ChIP-Seq it has to be decided whether to use fixed or variable read lengths and whether or not to include reads that map to multiple sites. Further, the ChIP-method generates a library of DNA fragments that are enriched for those bound by a given DNA interacting protein. The use of control files to remove the signals from tags not bound by the protein [147] is also a source of variation. Methods for genome wide profiling of copy number and structural rearrangements [148–150] and for analysis of DNA methylation [120, 151] are changing as well with the advent of new technologies.

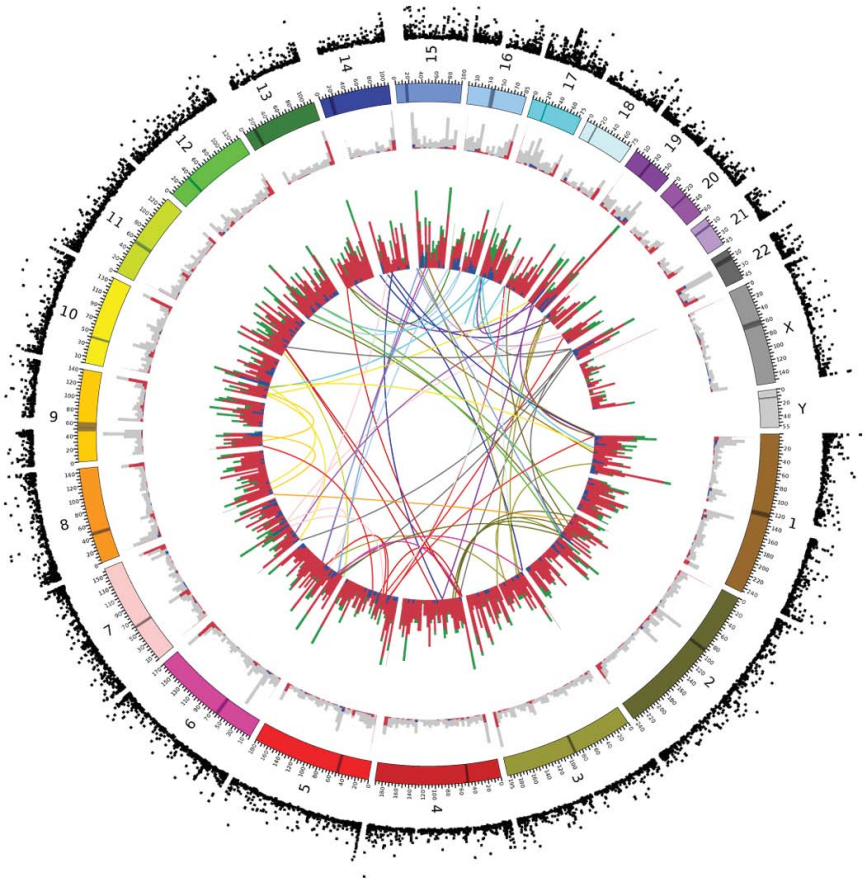
1.13.2 Visualization

Genome tracks were originally used for custom uploading and visualization of data in genome browsers. Recently visualizing genomic tracks in circular plots have become immensely popular. A review of tools for effective visualization of multidimensional cancer omics data, including genome browser, circular plots and heatmaps, is given in [61]. One of the more prominent tools for circular plots is Circos [152]. An example of such a plot is shown in Figure 1.6 on page 39. The plot that was published in [48] was used to show some properties of copy number variations in normal healthy people, and the distribution of likely underlying mechanistic causes. The 8599 CNVs displayed were identified in 40 HapMap samples. Many plots like this one, but with a much higher density of data points and with a higher number of tracks displayed have been published. Spotting inter track locational association in such plots is generally of high value. This is hard to do with the naked eye. Multiple levels of zooms would be needed. A solution to this is to calculate local measures of association and visualize these locus wide summarizes in a genome wide view.

1.13.3 Aggregation plots

Different forms of aggregation plots have, under different names, been in use for a while, but, to the best of our knowledge, the first time the term aggregation plot was used was in [153] a paper presenting the tool aggregation and correlation toolbox (ACT). To create an aggregation plot out of a genomic track of interest a set of equally sized genomic regions and their center points are needed. Subsets of genomic coordinates in the genomic track of interest located in the regions are selected. Each region is divided into bins so that the centre point of the regions separates the two central bins. A measure of occupancy is averaged across all equally sized regions for bins located similarly in relation to the center point. The averaged measure can be the overlap, or a feature count in the bin or feature signals. A line plot of the averages for each bin is called an aggregation plot. A non-uniform distribution will suggest that the track of interest prefers specific locations in relation to the center points. A second genomic track, like definitions of TSSs, is often used to define the center, or anchor, points.

Figure 1.6: In the center a plot of links is shown between origin and insertion site of inter-chromosomal duplications, colored according to their chromosome of origin. The indicated translocations suggest retrotransposons as a possible cause for these particular duplications. The first circle from the center is a frequency plot of duplications (green), deletions (red) and multiallelic (blue) loci. The second circle shows a stacked histogram of types of derived mechanistic causes. The third circle is a plot of individually colored chromosomes. The outermost circle shows a measure of population differentiation that indicates to what degree the variation contributes to the separation of populations (CEU and YRI). Courtesy [61].



1.13.4 Data exploration

To cluster objects is to store and represent objects in a way that reflects their relative similarity or dissimilarity with in respect to a set of features. The

definition of what a cluster is varies with its application [154]. One definition given in [154] is that: "A cluster is a set of data objects that are similar to each other, while data objects in different clusters are different from one another". Clustering, also called data exploration or unsupervised learning, is automated unsupervised classification of objects into clusters (categories, groups or subsets). The clustering of objects has been a method for learning, famously applied by Linnaeus and Darwin, for which documentation exists since the time of Aristotle. Central to the degree of reality reflected in the clustering is that it is based on measurements or observations of object features. The clustering procedure generally consists of four major steps:

1. Definition of a feature vector by selecting or extracting features representing an object/sample/genome track

2. Construction of a distance matrix by applying a proximity/distance measure to the feature vectors

3. Validation of clustering result (A clustering method will always produce a result irrespective of the nature of the data and it should, if possible, be cross-checked.)

4. Interpretation of the result.

The field within biology, in which clustering has been most frequently applied, is that of gene expression microarray analysis. There it has contributed to the transfer of functional knowledge between genes clustering together based on expression profiles across conditions and to the discovery of new disease sub types based on the expression profiles of disease samples across all genes.

1.13.5 Association by genomic localization

Two genomic tracks can be associated in multiple ways in terms of genomic positions. Two tracks describing genomic segments can, for instance, overlap with each other or just be in the proximity of each other. The number of alternative measures of associations increases when considering different track types. These alternatives are extensively covered in [155].

Galaxy

The core usage of Galaxy is for the documentation and construction of workflows for manipulation of genomics data. It is built around a Unix-like concept providing the user with a rich selection of small programs that each performs limited tasks. Galaxy is a web server with a Graphical User Interface (GUI) front end. The web server is freely available for local installation and can be extended by programmers proficient in Python. New tools developed can be shared through a Galaxy toolshed. The selection of tools available in any installation of Galaxy can be customized and tools from the toolshed can be installed. The programs or tools can be navigated through categories and selected from the GUI of Galaxy. Each execution of a program on a dataset, may it be the upload of one, is stored and made accessible as a history item from the GUI. Apart from being able to upload genomic tracks and other data formats, like formats for sequence data, from your desktop, data from all major genomic and epigenomic data bases can rapidly be accessed from the interface. History items can be piped into workflows that can be used for reproducing the data manipulation or analysis or for applying it to a different data set. Visualization of input or intermediary genome track data can be achieved in all major genome browsers. A tool for creating web pages, called Galaxy Pages, for documentation of, and with links to, history items and workflows can be created. Some of the more useful categories of tools for manipulation of genome tracks are:

Get and send data offers access to major databases and complementary software like Genome spaces.

Lift-Over converts from one version of a reference genome or genome freeze to another, like from hg18 to hg19.

Text Manipulation as well as Filter and Sort performs Excel types of operations so that for instance genomic regions with a value higher than a threshold can be selected.

Operate on Genomic Intervals allows for selecting regions that overlap between two genomic tracks. Alternatives to many of these operations are available as independently developed tools in BEDTools and in The Genomic HyperBrowser. BEDTools, which is a freestanding command line interface based software, is also included as a tool category in Galaxy.

The remaining tools in Galaxy handle tasks that are not immediately relevant to high-level analysis of epigenomic data, like simple plotting and

statistics, phylogenetics, motif prediction peak calling and population genetics. A few of these tools have within Galaxy been plagued with malfunction. From a perspective of integrative epigenome analysis Galaxy's weakest point is the lack of high level analytic tools. Galaxy is programmed in Python but also embeds tools written in other languages. Some tools use R libraries for statistical calculations. Others are merely embedded software like Plink, EMBOSS and The Genome Analysis Toolkit or GATK (from Broad Institute). The Galaxy method and software development have been published in [156–158]. The usage of a visualization tool in Galaxy for the analysis of Next Generation Sequencing data has been published in [159]. The Galaxy Tool Shed is described in [160] and available at `http://toolshed.g2.bx.psu.edu/`.

The Genomic HyperBrowser

The initial and core functionality of The Genomic HyperBrowser allowed for performing statistical analysis of association between pairs of genomic tracks and for calculating descriptive statistics, like average distance between elements, of individual tracks. Its initial development is described in Paper III [222]. In addition to focusing on development of new analysis methodology, a central focus was from the start that the developed functionality should be possible to use by biologists. It was thus decided to use a web-based interface. Also, it was decided that instead of requiring users to design custom workflows that combine smaller operations and statistical computations to suit their needs, the users should be presented with a list of possible questions that the system could answer based on the data a user was interested in. To simplify development, the Galaxy system was used as a web framework. The analysis functionality of the HyperBrowser was then offered as a single tool running on a local Galaxy instance. Later, a large number of tools were added to support a broader analysis setting, including initial processing of data, preparation of data for analysis, as well as several specialized forms of analysis beyond what is offered by the initial HyperBrowser tool [155]. The set includes tools for the generation of heatmaps for visual inspection and interaction with regulomes as described in [161]. It includes a tool for clustering of genomic tracks based on different distance measures as described in Paper IV [223]. Further, the tool set includes a tool for testing of association in 3D according to the method described in [162].

GenometriCorr

GenometriCorr [163] is an R package and it imports the Bioconductor packages IRanges and GRanges. It is also available as a downloadable software with a GUI or it can be integrated as a tool in Galaxy. It provides four tests for positional association of pairs of genomic tracks. The tests are:

Relative distance test It is used if the elements of two genomic tracks are close to each other using a normalized distance.

Absolute distance test It is also used if the elements of two genomic tracks are close to each other, using a distance that is not normalized.

Jaccard test is used to test if the elements of two genomic tracks overlap. The test is an equivalent to the ”segment overlap” test in the Genomic HyperBrowser.

Projection test It is used to test whether points of one track frequently fall within segments of another track. The test is an equivalent to the ”points inside segments test” of the Genomic HyperBrowser.

It is noted in the paper that the results of the tests are dependent on which of the two tracks is considered the reference track, which is fixed, and which is considered the query track, which is permuted. This is also true for some of the HyperBrowser tests but will be more evident due to the written formulation of the test as a readable sentence after selections have been made.

EpiExplorer

EpiExplorer [164] is one of the few web servers for genome track analysis developed independently of the Galaxy web server. This makes the provided solution unique and interesting but it suffers from the lack of the rich set of basic tools and of the features for transparency and reproducibility. One advantage of EpiExplorer is the speed with which results are generated. The speed is partly achieved by not calculating the exact measure association, however. This makes EpiExplorer suitable for initial exploration of many analytical alternatives and data sets. More rigorous and time-consuming statistical testing can be performed for validation. The principle of the analyses of EpiExplorer is that tracks of interest to the user, usually novel and possibly generated by a lab associated with the user, are compared to reference tracks stored in the database of EpiExplorer. The first step of an analysis is to upload these tracks. It is also possible to analyze any of the

preexisting tracks, which include chromatin and transcription factor binding data from the ENCODE project or epigenome data from the Roadmap Epigenomics Initiative or gene annotations from Gene Ontology and Online Mendelian Inheritance in Man (OMIM) and genome annotations from the UCSC Genome Browser. Five types of default reference genomic regions are available in EpiExplorer. They are:

1. CpG islands
2. Gene promoters
3. Transcription start sites
4. Predicted enhancer elements
5. 5-kb tiling regions, spanning the entire genome. Non-overlapping segments used as unbiased reference sequences in analysis.

Every EpiExplorer analysis is based on relating a genomic track, imported by the user in one of the provided ways, to the default tracks available within the system. For each default type of data a set of analyses deemed suitable is presented.

Genomic neighborhood plots are equivalent to aggregation plots described in Section 1.13.3. Any of the default tracks can be used as anchor points to create neighborhood plots of the track to be analyzed. Many tracks can be plotted simultaneously and are represented with different colors.

Bar charts visualizing the percentage of overlap of the track to be analyzed with a default reference track and a randomized control.

Frequency of methylation plot displaying the distribution of degree of overlap between the segments of the track to be analyzed and the segments of a default track. The distribution is compared with a randomized control track.

Enrichment table and word cloud illustrating the most highly enriched Gene Ontology (GO) terms among genes whose transcribed region is within 10 kb of a 5hmC hotspot.

1.13.6 Gene centric analysis

The genomic feature of main interest in most integrative studies has been, and probably will be, the gene. The usual workflow is that each type of data

is preprocessed separately using software designed for that particular type of data. Many such software solutions have features summarizing the data around genes, to make it gene centric. In **Papers I** and **II** we used such software for gene expression, DNA methylation and copy number data. R, Lumi and MethyLumi were used for analysis of gene expression and methylation data. For the gene expression data, intensities measured on probes representing exonic sequences were summarized into average gene intensities. For the methylation data probes representing CpG sites in promoter regions were made gene centric. Copy number data was also projected from genome covering segments onto regions covered by genes. Since all types of data have been summarized based on gene IDs instead of genome coordinates integration can be based on merging lists of gene IDs instead of selecting features with overlapping or positionally associated coordinates. Integration by merging lists of IDs is easier to accomplish, as it allows for the use of the regular data database operation language SQL, for example ”Select gene IDs that are ’hypo’, ’gain’ and ’over’ from ’sample X’”.

Sigma2

Sigma2 is a stand-alone software written in Java that requires the statistical package R and the database application MySQL. The intention of the software is first and foremost to be a platform for preparation of gene expression, copy number and DNA methylation data acquired by oligonucleotide micro array technology for integrative analysis. Sigma2 has a lot of functionality for processing and analysis of one type of data separately. It also has multiple tools for integrating two types of data both on the level of a case-control sample pair and on the level of multiple sample case control groups. For integration of more than two types of data it has a tool for analyzing case control sample pair data. In [21] the software was used to look at the frequency across samples of the result of Sigma2 integration. The term ”multiple concerted disruption” (MCD) was defined as the multiple aberration of a gene found in different types of data. Differentially expressed genes with MCD were selected for individual samples and then another threshold was put on sample recurrence. Variants of this strategy have also been implemented in [165] and **Paper I** and**II**.

The R script used in Papers I and II

The R script used in **Papers I** and **II** is described in the summary of these papers. It is used to select genes based on alterations in two types of data that

occur frequently across samples. It also looks for association of alteration in a gene centric way.

1.13.7 Analysis centric to cis regulator regions

The research effort invested into understanding *cis*-regulatory elements and functionally associating them with genes has recently increased. The production of ChIP-Seq data on genome wide binding sites of *trans*-acting factors has made this type of research possible.

Clustered AGgregation Tool

A typical application of aggregation plots (explained in 1.13.3) is to plot average histone modification occupancy around transcription start sites [25]. Nucleosome occupancy around TSSs has been shown to be asymmetric around TSSs reflecting transcription. For genomic regions, defined by anchor points with a sense/antisense or upstream/downstream polarity, a flipping of the antisense regions is necessary. This way the correct matching of antisense region bins with bins from sense regions when creating an aggregation plot. If the genomic track of interest is asymmetric around the anchor point, but occurs equally frequent in both genomic directions, reluctance of flipping the antisense, or the sense, segments will create aggregation plots that are (falsely) symmetric around the center points. This is pointed out in [166] and the option of segment flipping based on center point strand information is implemented in Clustered AGgregation Tool.

TFBSs and insulators like CTCF have frequently been used as anchor points [167]. These genomic tracks are not annotated with genomic direction. For insulators that are believed to separate active from inactive chromatin states, it is highly likely that the surrounding pattern of nucleosome occupancy is asymmetric when considering strand information. In [166], CTCFs used as anchor points are first clustered using Clustered AGgregation Tool based on the histone occupancy in flanking regions. Equally sized sub clusters of CTCFs with histone occupancy profiles being each others mirror images strongly indicate that strandedness is relevant. One of the sub-clusters in each mirror image pair is therefore flipped and the pairs are merged halving the number of sub-clusters. Aggregation plots are constructed for each sub cluster. The possible functional differences between the sub-clusters are investigated.

A typical application of aggregation plots is to plot average histone modification occupancy around Transcription Start Sites (TSSs) of genes, as is done in for instance [25]. An aggregation plot will reveal any pattern of occupancy of the histone modification that frequently occurs around the TSS. By clustering occupancy data around anchor sites subgroups of patterns will be suggested. Genome directionality might be the reason for pairs of subclusters that are mirror images. Genomic processes, like DNA synthesis, have direction => genomic polarity. When clustering aggregation plots of histone modifications around TSSs, utilizing the strand information using the Clustered AGgregation Tool in [166], patterns being the mirror images of each other were rare, indicating that the genomic events take place either on the sense strand or on the antisense one. In some cases subgroups being the mirror images of each other around the TSS (anchor point with strand information) might be suggested. These would indicate that at some TSSs the genomic activity, demarcated by the pattern of histone occupancy, occurs on the sense strand and at others it occurs on the minus strand. TSSs generally have strand information indicating the genomic directionality of their transcription activity. Genomic tracks defining many other genomic features, like enhancers, come without strand information. In [166] it is suggested that clustered aggregation plots of epigenetic data can help to infer the genomic direction of activity feature like *cis*-regulatory regions, that lack strand information.

ChIPseeqer

ChIPseeqer [168] is a software that permits combining tools for the analysis of ChIP-Seq data of histone modification into workflows. The used tools are reviewed in [168].

GREAT

The purpose of Genomic Regions Enrichment of Annotations Tool (GREAT) [133] is to enhance the functional classification of the gene targets of *trans*-regulatory factors. GREAT facilitates improving definitions of *cis*-regulatory regions for individual genes and the consideration of these definitions when performing functional enrichment analysis of *cis*-regulatory regions. The *cis*-regulatory regions are identified by localized measurements of DNA-binding events across an entire genome. By extending gene regulatory regions to not only include traditional promoter regions, but also to lots of remaining intergenic regions, GREAT enables the performing of functional

enrichment analysis, using also gene distal binding sites of a DNA interacting proteins.

Cistrome

Cistrome is an extension to Galaxy, devoted to the analysis of the chromatin of gene regulatory regions [78]. It offers tools primarily for preprocessing, analysis and integration (with expression data) of ChIP-Seq data. The cistrome is also a concept describing the research area in focus of the software, namely the ”set of *cis*-acting targets of a trans-acting factor” [78]. The cistrome is essentially the binding sites of transcription factors and nucleosomes (and ncRNAs). The software Cistrome harbors tools mainly devoted to peak calling of ChIP-Seq data like MAT, M2AC and MACs. Tools for standard gene expression and gene ontology analysis are also available. Tools for integration of the data mentioned above, like sitePro, Peak2Gene, CEAS, a tool for clustering of (aggregation heatmaps) and a binding motif discovery tool are available as well.

HaploReg and RegulomeDB

Many haplotypes identified as associated with a disease in GWAS have proven not to contain any functional SNPs within genes. This has puzzled researchers. The recently generated massive amount of data pointing out cell types specific enhancer regions has opened up for testing if functional SNPs are located within them. HaploReg [169] and RegulomeDB [170] are complementary web based tools and databases for integration of haplotypes identified as associated to disease through GWAS studies and genomic regions proven to have *cis*-regulatory activity in some cell lines. The *cis*-regulatory information has mainly been generated or assembled by the ENCODE study. HaploReg and RegulomeDB are both developed by the ENCODE group. The *cis*-regulatory regions are used to select candidate SNPs from the Haplotypes.

The databases contains manually curated regions, that have been experimentally characterized to be involved in regulation, containing ChIP-seq information for a variety of important regulatory factors across a diverse set of cell types, chromatin state information across over 100 cell types, and expression quantitative trait loci (eQTL) information connecting distal genomic regions with genes. Close to a thousand experimental data sets are included, covering over 100 tissues. Close to 400 million computational

predictions of transcription factor-binding sites are included. RegulomeDB contains literature-derived from information on enhancer regions as well.

1.13.8 Pathway/network centric analysis

Based on gene expression profiling studies of various designs, cellular networks of gene co-expression can be defined. A review of methodologies for this is given in [171]. Cellular networks can also be defined based on the interaction of proteins. These networks are generally visualized as nodes and edges forming a spheric structure. When the understanding of (parts/some of) these networks deepens to the degree that the chain of events of the molecular interactions can be established, and an end product can be specified, the structure is instead called a pathway. Most defined pathways of today describe metabolic conversion, gene regulation or signal transmission. Some of them describe processes, like cell proliferation, cell survival (apoptosis and necrosis) and angiogenesis, that when disrupted, contribute to cancer progression. Different cancer related pathways are reviewed in [172]. The use of high throughput technologies to construct cellular networks and biological pathways to understand the relationship between an organism's genome and its phenotypes is what defines the field of functional genomics. To study a cell in light of this holistic perspective is to perform "systems biology". Traditionally, analyses of gene expression and gene regulatory data have first been performed at the level of genes. Functional data in the form of gene ontology and pathway information is utilized at the end of the analytic pipeline commonly in the form of functional enrichment analysis. Some software tools of frequent usage for this type of analysis are listed and compared in [173]. More recently, analytic approaches where systems biology information is utilized at the outset of the analysis, have successfully been applied in the software Paradigm [174] and Pathifier [175]. When the mutations of a cancer subtype can be positioned on a specific pathway, which has been shown to be the case in many integrative studies of cancers [176][177], the knowledge of the structure of that pathway can help to identify bottleneck steps of molecular interaction. The proteins involved in such bottleneck steps are ideal targets for drugs. A drug can then be used for counteracting the effects of mutations of different proteins in the pathway [178].

Paradigm

The activity of a gene is not always reflected in its level of transcription. This can be due to silencing of the RNA or chemical modifications of the translated protein product. Just as the activity of a gene with only a slight increase in mRNA abundance, as compared to the reference, can turn out to be much higher than the reference, the activity of a gene with a much higher mRNA abundance can turn out to be no different from the reference with respect to its encoded protein. The idea of PARADIGM [174] is to use pathway information to integrate genetic and epigenetic data, and protein transcriptomic and proteomic data for inference of the activity of the involved genes and gene products in that pathway as a whole. This way, the activity of a transcription factor can be assessed not only by its gene expression, but also by the activity of its targets. The algorithm of PARADIGM is based on Probabilistic Graphical Models (PGMs) and factor graphs [179]. Although PARADIGM has proven to be one of the most successful software solutions in integrating cancer data, and has been used in most of the large scale projects by The Cancer Genome Atlas (TCGA) Research Network [180, 181], it falls short when measured in terms of reproducibility, transparency and accessibility. The central output of PARADIGM is a heatmap plot representing genes and their associated pathways. Columns correspond to samples and rows to entities. Red color indicates high calculated/inferred activity of a pathway entity and blue denotes low activity.

1.13.9 Analysis of proximity in three dimensions

No tools are yet available for this type of analysis. Theoretical guidelines on how to adjust for biases caused by structural dependencies in the 3D data when performing statistical testing are, however, starting to emerge [162].

1.13.10 Inference of chromatin states

One of the major contributions of ENCODE to the scientific community was to provide genome wide histone modification data. This type of data revealed the existence of chromatin states [72]. Chromatin modification data mostly generated by the ENCODE project has been used to train machine learning algorithms to divide the genomes into chromatin states. One of the first software packages to do that, while ENCODE was still in it first phase

mapping 1% of the human chromatin, was ChromaSig [182]. It uses a two-step algorithm to call chromatin states. First, it divides the genome into 2-kb bins and selects only those bins that are enriched in chromatin modifications. The bins are then clustered based on their similarities of chromatin signatures. The algorithms for inference of chromatin states were refined, resulting in new software. Two of them, ChromHMM [183] and Segway [184] were developed by groups within ENCODE. The algorithm of ChromHMM is based on Hidden Markov Models (HMM), while that of Segway employs a dynamic Bayesian network (DBN) approach. ChromHMM works on a resolution of 200-bp segments converting each track to Boolean values for each segment. Segway on the other hand operates on a base pair resolution. The algorithms also differ in how they assign labels to each chromatin state segment. Segway is well documented and is available at `http://noble.gs.washington.edu/proj/segway/`.

EpiGraph

EpiGRAPH [185] allows for integrating epigenomic data with genomic sequence information to, for instance, try to figure out whether tissue specific activity is encoded in the DNA sequence or whether the different genomic sequences have different probabilities of being sites for virus integration. EpiGRAPH can also be used to learn similar genomic regions.

1.13.11 Software environments for bioinformatics research

Taverna and myExperiment

Taverna [186–188] and myExperiment [189] are developed by an international diverse research group called myGrid. Taverna is software built to be a research environment, facilitating the collaboration within and between teams of software developers and researchers. The scope of the research process that Taverna is meant to be involved in is very large, and it has a wider scope than that of all other software mentioned in this thesis. The programming language and environment for statistical computing R, described later, might be an exception. The software is designed to address the complexity of bioinformatics and other modern research tasks through grid computing. Grid computing is a type of parallel computing connecting regular computers through a network interface. Grid computing is also

based on connecting web services and client applications, which require a piece of software, called a middle layer or middleware. Since the computers of a grid are connected over a network, their speed of intercommunication is at the low end. Grid computing is, therefore, most suited for parallel computing, where limited communication between processes is needed. Using grid computing for the analysis of massive amounts of data, available on the internet, is called e-science. Taverna addresses the demand for reproducibility and transparency by letting any complex analysis be composed as a workflow. A programming language devoted to the creation of workflows, the simple conceptual unified flow language (Scufl), is used for the purpose. One risk of embedding web services into work flows is that if the web service is discontinued, the work flow will not execute any longer. This type of problem is affecting at least a few Taverna workflows. Every analytic step is documented within the Taverna environment.

MyExperiment is an online environment developed by the same team as Taverna for sharing and exploring Taverna and other bioinformatics workflows. Since its release in 2007, myExperiment currently has over 3500 registered users and contains more than 1000 workflows.

GenomeSpace

GenomeSpace is online software that integrates databases and multitool software into one environment. The available databases and software at login are ArrayExpress, Cistrome, Cytoscape, Galaxy, GenePattern, Genomica, geWorkBench, GiTools, Integrative Genomics Viewer (IGV), InsilicoDB and UCSC Table Browser. The ones relevant to genome track analysis have been described in their separate sections in this thesis. Handling of large amounts of data is possible through storage in the Amazon cloud. Many tools for file format conversion are available from within GenomeSpace to facilitate the piping of data between different software environments. GenomeSpace is available at `http://www.genomespace.org`. The software integrated in GenePattern harbors many tools and provides features for reproducibility by themselves. GenePattern, for instance, provides access to more than 230 tools for gene expression analysis, proteomics, SNP analysis, flow cytometry, RNA-seq analysis, and common data processing tasks.

Spark and Genboree

Spark [190] clusters genomic regions in a similar fashion as the Clustered AGgregation Tool described in sub-section 1.13.7. Instead of basing the clustering of regions/anchor sites on one type of data, it allows for basing it on sets of genomic tracks. This makes it resemble the software that divides regions into chromatin states based on sets of epigenomic data. In contrast to the Clustered AGgregation Tool it does not include any option for flipping of aligned regions based on hypothesized strandedness. A Spark analysis starts by the user providing two types of input: 1) A genomic track specifying regions that will be used as anchor point, and 2) One or many genomic tracks whose distribution around the anchor points will be used to sort/cluster the anchor points. Spark uses the k-means algorithm for clustering for its speed performance. It allows for interactive manipulation of the clustering results, where clusters can be divided into further sub-clusters. This option is relevant, since the cluster solutions generated by k-means can be a non-optimal. Spark is written in the programming language Java. It is available at `http://www.sparkinsight.org`. It is also integrated as a service in Genboree at `http://www.genboree.org/java-bin/login.jsp`, which allows for parallel analyses. There is no publication about the Genboree web site. It appears from the web site, however, as if it will be used for a visualization and analysis of all spectra of next generation sequencing data. Geneboree already offers tools for variant calling from next generation sequencing and for ChIP-Seq analysis.

R and Bioconductor

R [191] is a programming language and an environment for statistical computing and graphics. It provides a wide variety of tools for statistical modeling, testing and clustering. It also provides many functions for plotting and image generation. Scripts performing useful functions that are not already available as packages in the R library can easily be turned into a package and added to the library with the consequence that the capabilities of R is constantly growing. An early response to the need of analytic reproducibility of high throughput data, especially microarray gene expression data about ten years ago, was the formation of the Bioconductor [192–196] project. In the pursuit of a suitable development and deployment environment for bioinformatics tasks, the team behind Bioconductor selected R as a platform. A reason for this was that R already was a popular programming language in the biostatistics and bioinformatics community and offered an interface to

people proficient in statistical computing and experts in the bioinformatics field. R and Bioconductor have been frequently used in especially gene expression analysis. It allows for easy access to annotation and experiment data through packages like BSGenome and Genomic Features. As for the analysis of next generation sequencing and other data in the genome track format its usability has been lagging. Concerns have also been raised that Bioconductor tools will not scale well with the ever-increasing data sizes. Some of these concerns are probably not justified. Many of the tools for handling next generation sequencing data like Biostrings and IRanges are written in C and work comparatively fast also on such large data sets. A tutorial for using these tools is available at `http://www.bioconductor.org/help/course-materials/2012/CSC2012/Bioconductor-tutorial.pdf`. Both R and Bioconductor have been used successfully for developing analytical pipelines in many of the initial analyses of The Cancer Genome Atlas data [177, 197]. Packages like sigaR `http://http://www.bioconductor.org/packages/2.11/bioc/vignettes/sigaR/inst/doc/sigaR.pdf` for integrative analysis are starting to appear in Bioconductor. A package for pathway centric data integration Pathifier [175] is awaiting approval for upload to Bioconductor.

R package cnaMet [198] selects genes that display a significantly higher expression when they are amplified and/or hypomethylated as compared to other copy number and methylation states and genes, that are significantly under-expressed when they are deleted and/or hypermethylated, as compared to other copy number and methylation states.

1.14 Consortia generating public data

For the purpose of promoting the understanding of chromatin biology in normal and disease development, a number of consortia have recently been formed. Their main missions are to collect genomic and epigenomic data from human cell lines and tissues and to provide public databases with the data. The consortia ENCODE, Roadmap of Epigenomics, The Cancer Genome Atlas Network and The International Cancer Genome Consortium will be described here. The samples from which data is collected for the respective consortium are listed, or can be browsed, at the following sites:

ENCODE `http://genome.ucsc.edu/ENCODE/cellTypes.html`

Roadmap of Epigenomics `http://www.roadmapepigenomics.org/data`

The Cancer Genome Atlas Network `https://tcga-data.nci.nih.gov/`

tcga/

The International Cancer Genome Consortium The ICGC data portal: http://dcc.icgc.org/web/

A more iclusive review of large scale epigenomics projects are given in [6].

1.14.1 ENCODE

Much of the research work within genomics and epigenomics has until recently been devoted to the analysis of genomic regions spanned by protein coding genes and their promoter regions. The ENCODE project has been performed in two phases. In the first phase, the ambition was to functionally annotate all elements in 1% of the genome. In the second phase, the whole genome was the target for functional annotation. In this effort, 1640 total experiments in 147 cell types [199] have been performed, and data acquisition assays, including ChIP-seq, DNase-seq, FAIRE-seq and RNA-seq have been developed. The importance of characterizing the human genome outside gene bodies is becoming apparent for many reasons. An important one is the failure of GWAS studies to map disease-associated loci to genes. Candidate regions are frequently located outside genes. The Encyclopedia of DNA Elements (ENCODE) project consortium [200] was launched in September 2003 by US National Human Genome Research Institute (NHGRI) one of the 27 Institutes and Centers of the National Institutes of Health, U.S. Department of Health and Human Services. In September 2012, results from the second phase were published in about 30 papers, with the main findings summarized in [200]. Its mission is to define all functional elements of the human genome. A similar initiative, modEncode, has the same mission, but using samples from the model organisms *Caenorhabditis elegans* (a type of flatworm) and *Drosophila melanogaster* (a type of fruit fly). The detection, by researchers both affiliated and not with ENCODE, that a large fraction of the genome outside protein coding genes is transcriptionally active, indicates that the human transcriptome and gene regulatory networks are far more complicated than previously appreciated. In their main publication on the human transcriptome [27], the consortium presented the remarkable observation that three quarters of the human genome is being transcribed in at least one cell type, and that most of that transcription belongs to non-protein coding genes. They also reported that the level of splicing used by genes is higher than previously anticipated. The functional annotation of intergenic regions by ENCODE data might contribute to the resolution of resolved the enigma of the evading functional validation

of GWAS SNPs. Initial analysis of the ENCODE data has lead to the development of data bases and web services for GWAS SNP functional annotation. The initial analysis has also lead to the development of tools using machine-learning algorithms to divide the genome into chromatin states based on epigenomic data. These tools are treated in Section 1.13.10. Large-scale applications of different mapping techniques for protein DNA interaction have lead to the mapping of many transcription factor binding sites in various tissues. Mapping techniques for the pairwise interaction of genomic sites in the three dimensional nuclear space have provided the infrastructure and data for analyzing genomic and epigenomic data for association in 3D.

1.14.2 Roadmap of Epigenomics

A sister project to ENCODE also launched by NIH is the Roadmap Of Epigenomics. The main objective of ENCODE is to map functional elements in healthy human cells. The data generated by ENCODE is, however, expected to assist in answering many questions related to disease. The missions of the two projects seem to be somewhat overlapping, and they also have a common data portal at `http://www.encode-roadmap.org/`. The Roadmap Of Epigenomics projects are primarily directed at providing data for the study of the deviation of mechanisms in disease, and how these deviations contribute to the disease. There are four "Roadmap of Epigenome Mapping Centers": BI, UCSD, UCSF, UW, and four "Data coordination and display centers", including UCSC. A joint presentation of ENCODE and the Roadmap Of Epigenome is available at : `http://www.genome.gov/Pages/Research/ENCODE/ASHG_2012_JStamatoyannopoulos_ENCODE_Roadmap.pdf`

1.14.3 The Cancer Genome Atlas Network

The Cancer Genome Atlas (TCGA) project was launched by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). Its objective is to characterize a selected set of tumor types in terms of gene expression, copy number variation, SNP genotyping, genome wide DNA methylation profiling, microRNA profiling and exome sequencing of at least 1200 genes. Integrative analysis of the data, to a large extent with a pathway centric view [176] through the usage of the software Paradigm, has been an important part of the published studies [181]. Integrative studies involving epigenomic data (DNA methylation) is hard to

find, however, and most integrative studies, so far, involve copy number aberrations and gene expression. These studies have revealed little about the interplay between epigenomic and genomic aberrations and how they interact to modify gene expression. Studies integrating the two types of data are reviewed in [106]. Epigenomic data collected by TCGA is currently limited to DNA methylation. A number of factors contribute in determining what tumor types get selected for analysis and which ones make it to publication first. Some of these factors are the incidence and survival statistics from the SEER Cancer Statistic website `http://seer.cancer.gov/`. Other factors are the extent to which good samples can be provided. This is determined by whether resection is performed before therapy for the type of tumor. Some of the tools used in the analysis of the Cancer Genome Atlas Data are available from the website of The Cancer Genome Analysis (TCGA) group at the Cancer Program of the Broad Institute of Harvard and MIT at `http://www.broadinstitute.org/cancer/cga/Home`. Significant publications on studies performed by TCGA on a few cancer types are [180, 181]

1.14.4 The International Cancer Genome Consortium

Parallel to the large projects launched by NIH there is an international collaborative project going on, that has grown out of national collaborative efforts similar to TCGA, on collecting and analyzing samples from cancer patients called The International Cancer Genome Consortium (ICGC). It was set off by a meeting in Canada in 2007 by 122 people from 22 countries involving experts in different fields of cancer research [201]. In 2008 goals and guidelines for the project were formulated and published in the document `http://icgc.org/icgc/goals-structure-policies-guidelines`. The goals include profiling of genomic, transcriptomic, and epigenomic changes in 50 different tumor types. Researchers from Asia, Australia, Europe, and North America have as of yet signed up for in total 47 project teams to study more than 21,000 tumor genomes. ICGC have in their document on goals and guidelines declared their preferred path from data generation to publication. Data generators are in the guidelines encouraged to release the data to a suitable public data base as soon as it has been verified and validated. The data is not considered as published, though, until a paper by the generators has been accepted by a peer review journal. Data users are asked not to submit any papers until the data has been published. As far as can be ascertained, no data have yet been published by ICGC. In a paper about the consortium [202] Mike Stratton writes, however, that: ”Sequencing of

a large number of cases for the major cancer subtypes will provide power to identify also rare mutation contributing to cancer".

1.15 Repositories

FactorBook A database of transcription factor binding sites: `http://www.factorbook.org`

Gene expresson omnibus A gene expression database: `http://www.ncbi.nlm.nih.gov/geo/`

Cosmic A database of somatic mutations in cancer: `http://www.sanger.ac.uk/genetics/CGP/cosmic/`

Epigenomc atlas `http://www.genboree.org/epigenomeatlas/index.rhtml`

PubMeth Important alterations in methylation in cancer has been coolected in: `http://www.pubmeth.org/`

The database of Genotypes and Phenotypes (dbGaP) Repository for primary sequence files of TCGC: `http://www.ncbi.nlm.nih.gov/gap`

Data Coordinating Center (DCC) All other types of data for TCGC: `http://cancergenome.nih.gov/`

TCGA data matix `http://tcga-data.nci.nih.gov/docs/publications/coadread_2012/`

Cancer genomic hub `http://tcga-data.nci.nih.gov/docs/publications/coadread_2012/` Database for sequences and alignments genrated by TCGA

ISB Regulome Explorer `http://explorer.cancerregulome.org/`

cBio Cancer Genomics Portal Next Generation Clustered Heat Maps: `http://bioinformatics.mdanderson.org/main/TCGA/Supplements/NGCHM-CRC`

cBio Cancer Genomics Portal `http://cbioportal.org`

cBio Cancer Genomics Portal Descriptions of the data can be found at`https://wiki.nci.nih.gov/x/j5dXAg`

The cancer Gene Census A list of genes an their involvement in cancer are maintained at: `http://www.sanger.ac.uk/genetics/CGP/Census/`

National Cancer Institute (NCI) Pathway Interaction Database(PID) `http://pid.nci.nih.gov/`

Kegg Pathway database: http://www.genome.jp/kegg/

Reactome Pathway databse: `http://www.reactome.org/ReactomeGWT/entrypoint.html`

IPA A commersial database of pathways: `http://www.ingenuity.com/products/ipa`

geneGO A commersial database of pathways: `https://portal.genego.com/`

Chapter 2

Aims of the study

The recently accelerating accumulation of old and new types of genetic and epigenetic information on a genome wide scale opens up for investigation of their interrelations on a global level. The new layers of information also allow for increased resolution in defining and selecting genomic features. There is a great need for new tools. The construction of them requires specific knowledge in many fields, particularly in biology, genomics, epigenomics, software development and statistics.

1. A main aim of the study was to provide methods for integrative analysis of disparate data types available in genomic and epigenomic studies. We wanted to explore avenues both to apply existing software, and to prototype dedicated solutions for a given project.

2. Having achieved viable solutions for integrative analysis, we aimed to apply the methods to datasets of sufficient genomic and epigenomic complexity available to us, with the further aim to obtain biological insights into the mechanistic aspects of the studies. Two such datasets consisted of osteosarcoma-derived data. We also wanted to enrich the epigenomic datatypes to include histone modification, as this is rapidly becoming a very important feature within genomic studies.

3. As all integrative analyses of genomic scale involve rich and complex data, there is a demand for ways to efficiently utilize and visualize such data. Therefore we wanted develop and try out efficient means of data exploration through visualization techniques.

Chapter 3

Summary of the papers

3.1 Paper I

Gene expression deregulation in cancer is frequently caused by both genomic and epigenomic aberrations. The phenotype of any cell is also a result of genomic and epigenomic interactions. The characterization of the nature of those interactions in cancer development is a fairly new, but very active area of research. Osteosarcoma is a type of cancer with extensive amounts of genomic aberrations and also of promoter hyper methylation. A data set, collected by a variety of oligonucleotide microarray technologies from 19 well-characterized osteosarcoma cell lines, was used in an attempt to perform a detailed characterization of the alterations. The dataset encompassed gene expression, promoter methylation and copy number data. Preprocessing of the expression and methylation data was performed in R while the copy number data was preprocessed in the Affymetrix Genotyping Console, and in Nexus. The data was initially clustered to identify sample sub-groupings based on a given type of data, followed by and to comparison of the types of data in terms of the generated cluster pattern of the samples. The hierarchical algorithm was used for clustering. The cell lines were clustered using the gene expression profiles, the methylation states of 27000 CpG sites and the copy number of about 2 million base pair locations, respectively, as features. The cell lines clustered in general differently, based on each type of data. The fact that methylation data came almost exclusively from CpG sites within gene promoter regions, and that the gene expression data is located to protein coding genes, and an early decision to focus on gene regulation, resulted in the integrative study being performed in a gene

centric way. Values of expression probes belonging to the same transcript, and those values of methylation probes carrying information on CpG sites from the same promoter region, were therefore averaged for their respective genes. Probe information was, however, maintained throughout the integration to enable the tracking of their individual contribution. The copy number data set, where segments covering every base pair of the genome were initially inferred, was projected to the bodies of genes. The three types of data could after such preprocessing be merged based on gene identifiers instead of using genomic coordinates. An R-script was developed for such a merge of the types of data and for further integrative analysis. Using a thresholding approach, it was decided whether genes were altered or not for the three types of data, respectively. Two-way dependencies between the types of data were estimated within each sample by calculating the odds ratio of a gene simultaneously being altered in two types of data. By conditioning odds ratio calculations on the state of the third type of data three-way dependencies were also evaluated. For the calculation of the odds ratio, which is a size effect statistic for the dependency, the construction of a contingency table is required. The same table was also used to calculate a measure of significance of dependency through a chi-square p-value.

On the genome wide level, there was a significant positive association between gain and over-expression, loss and under-expression, as well as hypermethylation and under-expression. No such global association existed between copy number and methylation data. A strengthening association between hyper-methylation and under-expression with increasing copy number, as revealed by the three-way analysis, suggests, however, that hypermethylation may oppose the effects of increased copy number for detrimental gene aberrations. The script also reports a list of genes with combinations of alterations that are frequent across the samples. A subset of that list contains genes that, frequently across samples, are over-expressed and have an additional aberration, hypo-methylation or gain, making regulatory sense or offering a potential regulatory explanation. Similarly, another subset would contain genes that are under expressed and have an additional aberration, hyper-methylation or loss. Genes in either of these two subsets were used for functional enrichment analysis. This identified the gene ontology terms ”embryonic skeletal system development” and ”morphogenesis”, as well as ”remodeling of extracellular matrix” as significantly enriched. A separate analysis of the genomic distribution of copy number aberration profiles in relation to gene density, performed in The Genomic HyperBrowser using a tool developed for the purpose, concluded that deletions tend to occur at gene poor locations, and that duplications tend to occur at gene rich loca-

tions.

3.2 Paper II

Clinical tumor samples and cultivated cell lines of tumors are different sources of information for characterizing a tumor type. Both of the sources have their analytical challenges. Cell lines tend, for instance, to accumulate genomic aberrations during cultivation, while clinical samples tend to constitute a heterogeneous population of cells. In this paper, the R-script developed for **Paper I** was applied to copy number aberration data and gene expression data from 29 clinical/patient samples in the analyses, referred to as paired. Two integrative analyses, identical in all aspects but for the use of mesenchymal stem cells (n = 12) and osteoblasts (n = 3) respectively, as references, were run in parallel in this study. They identified 445 and 138 genes, respectively. The method for selecting genes based on the two types of data available within the script was compared to the one available from within the commercial software Nexus, a tool which is primarily used for copy number derivation. That method relies on identifying genomic regions with simultaneous frequent copy number alterations and higher than expected numbers of differentially expressed genes. The paired method from the R-script found more than 90% of the genes that were found by the Nexus method and also an additional set of genes, including genes involved in cell cycle regulation. It was concluded that the paired method had higher sensitivity. The global significant positive association between gain and over-expression, loss and under-expression found using the cell line sample in **Paper I** was present also in this data set. A gene list for functional enrichment was assembled by merging the result of the two parallel analyses and also by requiring the genes to be differentially expressed. 31 genes were selected this way. 22 of the genes were associated to cancer. Fourteen of them were relevant to cell cycle regulation. The genes involved in cell cycle regulation could provide clues to the extensive genomic alteration characteristic of osteosarcoma.

3.3 Paper III

The methylation data used in **Papers I** and **II** was assayed by microarray technology. It only represents some few CpG sites from each promoter assayed, selected for their reported tendency to be methylated in cancers.

With next generation sequencing technology it is possible to acquire the methylation state of every CpG site of the genome. It has also become evident that the relationship between CpG methylation and gene expression is complex and depends on the location of the site in relation to the gene. The relationships between the occupancy of different histone modifications and gene expression are equally dependent on relative genomic location. Furthermore, many integrative analyses of genomic and epigenomic data do not have the gene regulatory aspect particularly in focus. Their objective might rather be to avoid any centricity at all. Generic solutions to these types of analyses are few. **Paper III** reports the construction of a web-based tool, The Genomic HyperBrowser, for flexible definitions of rules for integrative analysis of two types of data through statistical testing. It also defines five generic input formats into which any type of genomic or epigenomic data can be converted. The paper presents the set of statistical tests that was developed to be executed on these types of tracks, and on a pairwise combination of them at the time of writing the paper. In addition to performing the tests on a genome wide scale or globally, they can also be performed in genomic sub-regions, or locally in what is referred to as bins. The usability of the software is exemplified through four biological cases. One of the cases is part of this thesis project. In that example, the dependencies of the expression of a gene and the occupancy of different histone modifications in a regulatory region associated to the gene are determined. The data used was first published supporting a seminal publication on the functional organization of histone modifications in relation to gene regulation [25]. In the original paper, juxtaposed aggregation plots of histone occupancy for four expression classes of genes (from low to high), using their transcription start sites (TSS) as anchor points, were presented as results. The plots revealed an existing association between the occupancy of most of the histone modifications and the expression of the gene. For a description of aggregation plots, see [166]. The solution provided in Paper III is based on a statistical test that provides a size effect and a p-value. The genomic positions of histone modifications were extracted from raw data as generated from the ChIP-Seq technology, using the peak detection algorithm of the software (Nucleosome Positioning from Sequencing) [40]. These positions were treated as ”unmarked points”, one of the five generic formats. The gene expression values were used as in [25] and converted to ”marked segments”, another of the generic formats. An implementation of the Kendall’s rank correlation test was used to ask whether the number of unmarked points (histones), counted in a marked segment (gene bodies with expression values), correlates with the mark of the segment.

The Kendall tau rank correlation coefficient captures the association between two measured quantities. In the particular case of this example the test translates to the following:

1. Comparing all pairs of genes with each other in terms of their gene expression values and histone counts in specified gene associated region.
2. Counting the number of concordant pairs, that is, pairs of genes satisfying that if the expression value is highest in the first gene then the number of histones is also highest. Similarly, the number of times the pairs are discordant is counted. If the two types of data are independent this is expected to occur equally frequent.
3. Calculating the Kendall tau rank correlation coefficient by subtracting the number of discordant from the number of concordant gene comparisons and dividing by the number of possible gene comparisons. This standardization makes the coefficient always vary between -1 and 1. The expected value assuming independence is 0, and a positive Kendall tau signifies positive association and a negative Kendall tau signifies a negative association.

Three more examples were developed to demonstrate the types of genomic research that can be performed with The GenomicHyperBrowser. The system still appears to be well on par with other existing solutions for statistical genomics, and is continuing to expand in functionality [155].

Here follows a short summary of the other three examples.

1. Finding genomic regions where the genomic location of integration sites of different retroviruses are similar. Virus integration sites are treated as unmarked points and promoter regions as unmarked segments. The question asked in every bin is whether viruses integrate into promoters more than expected by chance. In the example a number of hotspot loci is identified in which viruses prefer to integrate.
2. Overlap of H3K4me3 regions with SINE repeats. Here both types of data are treated as unmarked segments and it is tested globally and locally whether the overlap is higher than expected by chance.
3. Exon boundaries and melting forks. The tracks used are exon boundaries, DNA melting forks and GC-content. This example included in the paper is used to demonstrate a tool to find out if an association is due to a confounding third track. Probabilities of melting fork locations are treated as a function, exon boundaries as unmarked points and GC-content as function. It was shown that the existing correlation

between melting forks and exon boundary location could be explained by differences in GC content.

3.4 Paper IV

An introduction to data exploration, or clustering, was given in Section 1.13.4. Feature selection and summary as well as calculation of object distances are listed there as important components in the process of clustering objects. **Paper IV** presents a software for specification of how to select and summarize features, and how to calculate distances for genomic tracks. The distances are in the current version used for hierarchical clustering. In **Paper I**, the osteosarcoma cell lines were clustered using the three types of data as alternative inputs. To cluster based on copy number data the intensities of the about 2 million probes were used as feature vectors. This was computationally very demanding, and had to be performed on a supercomputer type of server. Replacing approximately 2 million probes with the segmented and copy number assigned data as input for clustering would imply a severe reduction of memory usage and of the scale of calculations. Such segment location data, for which there is no match between boundary locations across samples, is however, not trivially converted into feature vectors. **Paper IV** offers solutions to this and other similar types of problems by providing ways to infer primarily euclidian track distances for hierarchical clustering. Three ways to calculate inter-track distances are presented. The difference between the first two ways of deriving inter-track distance lies in the underlying extraction of features.

Solution 1 relies on dividing the genome into bins and determining the relative track coverage in each bin. The relative coverage, for the track to be clustered, of a bin is used as a feature. Clustering using this first definition of features is called ”Similarity of positional distribution along the genome”.

Solution 2 relies on looking at relative overlap with other annotation tracks in this setting, referred to as reference tracks. The relative coverage, for the track to be clustered, of a reference track is used as a feature. Clustering using this definition of features is called ”Similarity of relations to other sets of genomic features”.

Solution 3 differs in that it does not use feature vectors but derives inter track distance directly from the pairwise track overlap relative to

genome coverage. Clustering using this definition of features is called "Direct sequence-level similarity"

The three clustering methods were, together with methods for dendrogram and heat map plotting, implemented as a tool in The Genomic HyperBrowser. The tool was used for hierarchical clustering of cell types and states on the basis of histone modification data. More specifically, the data described genome wide occupancy of H3K4me1. In the example used in **Paper IV**, tracks of genes associated with GO terms, with one track per GO-term, were used as reference tracks. For validation of the clustering result, existing knowledge of the type and the differentiation stage of the clustered cells was used. Cells of similar types clustered together in all but one case. In that case, brain cells formed two separated sub clusters, which proved to represent fetal and adult brain cells.

In an attempt to find biological motivation for the large relative dissimilarity between fetal and adult brain cells, the genes that had a different occupancy between fetal and adult cells were extracted and used for functional enrichment analysis. Terms related to neuron development came up as the most significant, indicating that the signal is of biological nature.

Chapter 4

Discussion

In **Paper I**, two-way and three-way association between copy number, promoter methylation and gene expression data within a sample was established and quantified. The purpose of the three-way association was to identify any interactive influence of gene copy number and promoter methylation state on gene expression. The associations were investigated within each of the multiple samples. Associations across samples were used to select frequently deregulated genes. The analysis presented in **Paper II** was similar, but with input data restricted to copy number and gene expression data. In **Paper III** associations between histone modifications and gene expression were established and quantified within a sample from one cell type. In **Paper IV** data on genome wide occupancy of one histone mark acquired from multiple cell types was used to group the cell types based on similarities of occupancy of the mark. They were grouped based on measures of genome wide association.

4.1 Backtracking alterations in DNA methylation

The lack of association between the DNA copy number, DNA methylation and mRNA expression levels of the methyltransferases in **Paper I** indicates the existence of other mechanisms for genome wide change of methylation patterns. It is known that polycomb-mediated DNA methylation guided by silencing patterns of histone modifications is active in cancer [102, 103,

203]. Integrating such data as well could possibly have provided clues to the change of methylation patterns.

4.2 Establishing and quantifying association

A major objective of the study presented in **Paper I** was to look into the relationship between copy number and DNA methylation and their possible interactive influence on the gene expression. Reports on the nature of this relationship suggest that different mechanisms are active genome wide, and that the direction of influence goes two ways between the two types of data. Hypomethylation (outside genes) increases for instance genomic instability, and hypermethylation increases the base mutation rate [23]. A possible case of the opposite direction of influence is the co-occurance of hyper methylation of tumor suppressor genes with genomic deletions [23], suggesting the presence of a two-hit mechanism, where one allele is silenced by deletion and the other by promoter methylation. The relationship between gene expression and DNA methylation states seems to be equally complex. Methylation is known to promote gene silencing in, for instance, X-chromosome inactivation and imprinting [83], while gene regulation has also been reported to influence the methylation pattern [22]. Even though methylation and gene copy number alteration events across the whole genome, also outside genes, probably are involved in cancer progression, the promoter centricity of the methylation array, and the primary intent to integrate it with gene expression data, contributed to limiting the study to the promoter methylation states and gene copy number states. In any case where multiple probes were spanned by a gene promoter their values were averaged. As impressive as it is of the methylation array to measure the methylation states of 27 000 genomic sites it only reflects a small fraction of the genome wide CpG methylation states. The methylation array only represents selected CpG sites in gene promoters, known from previous studies to be methylated in cancer. Further, it has been suggested that the methylation status of CpG island shores, not represented on the used methylation array, is more relevant to gene regulation than to methylation of CpG islands [110]. Finally, the methylation status of the gene body is important to the expression of the gene as well. Hence, an analysis using DNA methylation data from next generation sequencing technologies, like whole genome [204] or Reduced Representation Bisulphite Sequencing (RRBS) [205], or even with the methylation status of every CpG site in the genome, has the potential to more meaningfully investigate its relation to gene expression. These larger

and more complex sets of data would, however, require extensions of existing software to handle, for instance in terms of opposite location dependent regulatory effects.

For the correct prediction of an association, given a certain mechanistic model, a number of considerations or assumptions have to be made. It is, as far as can be ascertained, not yet known whether methylation states are ”inherited” with copy number amplification. If such inheritance is a fact, the amplification of the methylated variant in a heterozygously methylated gene could automatically lead to the detection of hyper-methylation, since the ratio of methylated to unmethylated is shifted towards methylated. The diploidy of the genome, and heterozygous states, disregarded by most analyses, complicates many predictions based on any mechanism. As an example, predictions based on promoter methylation compensating a dosage effect, are not straight-forward. The gain of a methylated variant of a heterozygously methylated gene would, for instance, automatically due to ”inheritance” not lead to an increased expression. The gain of the other allele would, however, lead to an increased expression in the absence of a compensatory mechanism. In the study of **Paper I**, the pairwise relationships between the three types of data were initially measured by correlation within samples. They were similarly also estimated by odds ratios of having simultaneous alterations in the two types of data. Using the same contingency table as for the odds ratio, chi-square p-values were also calculated. The results suggest that gene expression sometimes, but surely not always, is affected by both copy number and methylation change. An explanation of the incomplete association between gene expression and the two other types of data could be that promoter methylation of genes can be used to maintain normal expression of a gene with an altered copy number. Many genes are also expressed in a tissue specific manner and those genes not being transcriptionally active in the investigated tissue will therefore not be affected by copy number changes. Possibly, the most interesting result of the two-way analysis was the lack of dependency between copy number and methylation data, suggesting that no such promoter methylation based mechanism for compensating for copy number aberration is in place. Another explanation of the lack of dependency on the genome wide level could be the drowning of the signal in the multitude of other signals from various active mechanisms. By conditioning on the expression state of the gene, mechanisms, not related to gene regulation, would be filtered out. Since the contingency table-based measure of association was easy to extend to look into three way interdependency of the data, that method, instead of the correlation-based one, was presented in the paper. An alternative analysis could have been

to perform linear regression with expression as the dependent variable and copy number and promoter methylation as fixed variables. The coefficient of the interaction term would have revealed any genome wide cooperativity, and competition for that matter, of influence of copy number and promoter methylation on gene expression. Instead, odds ratios and chi-square calculations were used. Odds ratios for expression and methylation levels calculated for only amplified genes showed that amplified genes tend to be normally-expressed or under-expressed when hyper methylated. The hypothesis that hyper-methylation adjusts the gene regulatory effects of copy number change is therefore not rejected by the results displayed in Figure 10.

The copy number and the methylation state data from the tumors are snapshots of the evolution of the tumor. Both states have likely been reached by multiple events spread out in time, where acquired methylation states may have triggered copy number events and copy number events may have triggered methylation of CpGs. The mapping of events of clonal expansion, at that level of detail, would benefit from the development of proper study designs, technology with higher genomic coverage and resolution. Also, analytical pipelines harboring the expertise of the field of statistics, computer science and molecular biology would contribute. Currently, no goldstandard software is available for this type of analysis, but ideas on how to find associations in these types of data have been put forth [206].

Paper I, also presents a separate study that was performed in The Genomic HyperBrowser, searching for association of frequency of copy number aberration and gene density. In that analysis chromosome arm level and focal copy number aberrations were not treated separately. It is reported that in tumors in general, such aberrations are either focal or span (close to) an entire chromosome arm [207]. The underlying mechanisms generating these types of aberrations are likely not the same, and the analysis using The Genomic HyperBrowser would have benefited from such a separation. The correlation of these subtypes of genomic aberrations to the other types of data might vary as well. A similar analysis could have been performed with the copy number data presented in **Paper II**, but it was not considered before the publication of that paper.

The study presented in **Paper II** only integrated copy number and gene expression data. Investigation into three-way dependencies was therefore not possible.

As stated in the summary, the contribution to **Paper III** is limited to the biological case example, which correlates histone occupancy and gene ex-

pression. More specifically the objective of the example case was to investigate how the occupancy of nucleosomes with various histone modifications in a region in proximity to the genes correlated with the expression of the gene. The used measure for correlation was the Kendall tau correlation and the significance was calculated with Kendall tau correlation test. The gene proximal region is defined to maximize the encompassed *cis*-acting regulatory factors. The positions of the nucleosomes with different histone modification were derived from ChIP-Seq data using the software NPS (Nucleosome Positioning from Sequencing) [208], which was not the same as the one used in the original publication of the data. A different algorithm was used partly because of a published comment pointing out weaknesses with the original one [209]. Another reason was that NPS did not, like other similar algorithms at that time, use sequence alignment data for individual histone modification separately to call peaks and infer nucleosome positions. Instead, it analyzed many histone modifications jointly utilizing the knowledge that they all should be positioned on shared nucleosomes, thereby increasing the positional resolution. Almost all peak calling algorithms are developed for, and applied with the best results to, small motif-specific DNA interacting proteins like transcription factors. Many of them are reviewed in [147]. The lack of similarity of the resulting peak calls between the different software solutions, especially when applied to histone data, remains unresolved. New software, like that published in [210] implementing many of the individual benefits of previous software, are still being developed and will hopefully resolve discrepancies. Peak calling algorithms work well for small motif-specific DNA interacting proteins that generally have distinct locations and are sparsely distributed across the genome. In contrast, larger non motif-specific DNA interacting proteins like nucleosomes are generally spanning larger genomic regions and can be densely populated. Some histone modifications are best identified by looking for patterns that span regions in the kilo base pair size range. In addition, histone modifications exert their functional impact in cooperation with other modifications shared by its nucleosome and by neighboring nucleosomes. Finally, the orientation of asymmetric histone patterns is believed to be functionally revealing [166]. The preprocessing and analysis of histone modification data have therefore departed from the traditional peak calling methodologies. An early attempt to capture wide distributions is the Sicer software and later examples include Chromasig, ChromHMM and Segway [182–184]. ChromHMM and Segway are tested and reviewed in [199]. These algorithms for extracting chromatin states do not necessarily make algorithms to exact histone positioning obsolete. Characterizing the genome distribution and the relative contribution to regulatory processes of individual histones will certainly still

be of interest. The challenges of analyzing histone positioning data properly is, however, not limited to the preprocessing step. It is commonly of interest to compare the regulatory influence of different histone modifications. To do that it is tempting to try to define gene-associated regions, that are commonly occupied by all histone modifications, and to use the frequency of occurrence in that region as a measure of involvement of the histone. This was done in **Paper III** and variations of this approach have been utilized in other important publications, for instance the ones using aggregation plots [25, 71, 211–214]. This way of defining the degree of involvement of histone neglects a few aspects of histone occupancy. It should be stated in this context that no analytical method as of today manages to capture all of these aspects. Different histones influence gene expression by occupying different regions associated to the gene. H3K4me3 tends to be highly concentrated in a focal region surrounding the transcription start sites of highly expressed genes, while H3K4me1 tends to occupy distal enhancers of such genes. The sparse occupancy of H3K27me3 in wide areas of silenced genes not only supports the idea that histones operate with different positional distributions but also suggests that they are detected with different signal amplitudes. Further, classes of genes vary in how they are regulated. In [215] it is reported that expressed housekeeping and cell-type specific genes have different H3K4me2 profiles in their promoter regions. It is also known that genes with a high promoter CpG content are regulated differently to those promoters having a low such content of CpG. Before any mechanistic model has been suggested for a functional interaction, the boundaries, within which a histone modification influences the expression of a gene, can only be roughly estimated. A possible improvement to the analytic strategy used in the histone example of **Paper III** would be to first determine where different types of histone modifications tend to occur and then to see if there are subclasses within histone modifications depending on gene class, and finally to correlate occupancy with gene expression. Future studies, in which it will be possible to define gene-associated regions with support from three-dimensional interaction data, will likely increase the resolution of these types of studies. In the histone example in **Paper III** two different pairs of boundaries are used for all genes and histones (flanking 2kb an 20kb of TSS). Also the gene expression data is afflicted with analytic challenges. Due to intensity biases between probes on the microarray chip, it is not recommended to infer relative transcriptional activity by comparing intensities between genes. This is why the change of expression within genes and between conditions constitutes the traditional output of gene expression analyses. It is believed, however, that the global nature of a correlation test will have an averaging effect on the biases.

In **Paper IV**, different ways of calculating distances/similarities between tracks are used for clustering tracks. These distances are similar to forms of measures of association.

4.2.1 The mutual exclusiveness of two-way aberrations

In the analysis of **Paper I**, we have shown that it is not common for different osteosarcoma cell lines to use copy number aberration and promoter methylation as alternative mechanisms between samples to deregulate a gene. No support is in other words found for a two-hit hypothesis of a combination of a genetic and epigenetic alteration leading to gene silencing. Integrating more omics data, like point mutations, genomic translocations, LOH, nucleosome occupancy, microRNAs or transcription factor regulation could possibly reveal other mechanisms working in parallel to silence both gene copies.

4.3 Finding genes with alterations in multiple types of data

The second major objective of **Paper I** was to use the three types of data to identify interesting genes and pathways. Two recurrence criteria were used to select genes:

Across data types meaning simultaneous different gene expression and at least an ”explaining” alteration in another data-type.

Across samples , meaning that to be selected, a two-way alteration has to recur in a minimum number of samples.

Integrative analysis performed this way not only selects deregulated genes but also annotates them with a possible cause of the deregulation. This type of information is valuable for many reasons. For example, tumor suppressor genes, that are silenced by promoter hyper-methylation, should, unlike those silenced by deletion, have the potential to be reactivated, making them potential targets for therapeutic drugs. A challenge in setting the cutoff limit for sample recurrence is to strike a balance between the high limit that filters out passenger genes, and the more forgiving one that allows for a degree of spread of alterations between the genes within a frequently altered pathway. Applying tests for significant differences on the gene level would exclude alterations, which are frequent on the pathway level, but not on the gene level.

Examples of pathways frequently altered in cancer are those regulating cell cycle progression and DNA repair.

The integrative study of **Paper II** is limited to datasets of copy number and gene expression. The lists of selected genes would probably have been twice as long, judging from transferring the size relations between lists from **Paper I**, if methylation data had been included. A few analytical choices were also made differently in **Paper II**. One of the more important ones was the decision to filter the gene list obtained from the paired (two way recurrence R-script) algorithm by requiring the genes also to be significantly differentially expressed. This implied a severe filtering of the original list and would not pick up genes sharing the mutation burden of a frequently altered pathway. In the paper, it is suggested that the sample recurrence threshold at 35% might have been too high, since important cancer genes, like CDKN2A and MDM2, would have been picked up with a slightly lower threshold. This emphasizes the importance of the pathway paradigm. This study complements that of **Paper I**, in that the data has been acquired from clinical samples instead of cell lines. The genomic and epigenomic states of neither clinical samples nor cell lines will truthfully represent the tumor state of interest. One of the advantages of using clinical samples is that the tumor cells within it are a better source of information on the clinical setting than the cell lines, since they have had no opportunity to accumulate further alterations during cultivation. One of the drawbacks of using clinical samples is that the pool of cells, used for extraction of the DNA, is commonly a mixture of normal stromal cells and tumor cells from different stages of the clonal expansion. It can be expected that the copy number data is less affected by the contamination than the gene expression, and that integrating the two datasets might filter out many non-cancer signals in the data. In this study a second integrative approach offered by the software Nexus was used to select genes as a comparison. The method was found to be less sensitive. It filters out genomic regions with a high frequency of gains or losses. Then a second filter is applied on the selected regions. Those, that contain either over-expressed or under-expressed genes exceeding the expected number, are selected. There is no obvious motivation for applying this filter. Whether a gene is a driver gene or not is not affected by whether its genomic neighbor genes are frequently deregulated or not. This method would also be blind to genes picked up by the pairwise method, developed in Paper I, that are located in genomic regions dominated by tissue specific genes not expressed in the analyzed cell types.

A number of approaches for integrating different types of microarray data from cancer have been published during the course of this study [180, 181,

216, 217] and a few software for integrateive analysis like PARADIGM [174], CNAmet [198] and Sigma2 [218]. Some of the principles behind these analytic tools have been reviewed elsewere [219].

4.4 Clustering of genomic tracks

In **Paper IV** three ways are suggested for calculating distance measures for clustering data in the simplest form of genome track format, which describes genomic locations of a set of segments. Solution 1, "Similarity of positional distribution along the genome", divides the genome into bins and uses the relative coverage in the bins as a feature vector. Epigenomic components are involved in multiple genomic activities dependent on the genomic location of their occupancy. In the example of **Paper IV**genomic tracks describing H3K4me1 occupancy were used. H3K4me1 is known to be involved in gene regulation through enhancer occupancy. Definitions of *cis*-regulatory regions, like enhancers, are now beginning to be available as genomic tracks. A modified version of the application of Solution 1, where distinct genomic regions, like enhancer regions were used as bins instead of binning the whole genome would have been an asset.

Most available genomic and epigenomic data, like nucleosome and histone modification data, is stored in the genome track format describing genomic location of sets of segments. There are, however, exemptions for which tools for clustering is needed. Point mutation spectra are stored in genome track format describing genomic location of sets of points. Solutions 1 to 3 in Paper IV could easily be applied to such tracks, but has not yet been implemented.

Some genomic tacks are annotated with more than just positional information. DNA methylation states are, for instance, stored in the genome track format describing the genomic location and the value of sets of points and copy number aberration profiles are stored in the format describing genomic location and the value of sets of segments. Distance calculations for such tracks becomes more complicated than for the unmarked types of tracks, since it has to reflect both positional and amplitude differences between two tracks. Clustering is a large field of research [154] and the opportunities for implementing more efficient and sensitive tools for data exploration, more variations of methods for feature extraction and summary and different proximity measures are plentiful and should be further explored.

Hierarchical clustering, along with K-means clustering, are the most fre-

quently applied clustering algorithms in medical biology. Contrary to hierarchical clustering, the K-means clustering algorithm partitions the objects into a defined set of clusters. K-means is more scalable than hierarchical clustering [154] and should be considered for tracks with many features that cannot be summarized in a justifiable way. In hierarchical clustering the object distances are generally visualized as dendrograms or binary trees. The reordering of the feature matrices is generally represented as sorted heat maps where the range and the intensity of the color represent the magnitude of the feature value.

These heat maps are especially useful in binary clustering where both the objects and the features are clustered, because they pinpoint the features that contribute to cluster separation. This is how the heat map of Additional file 2B in Paper IV reveals that the H3K4me1 level of occupancy in genes associated with ”glycerol metabolism” and those associated with ”xenobiotic metabolic process” or drug metabolism co vary across the analyzed cell lines and that the two genes sets are important in separating liver cells from other types of cells. This fits well with the knowledge that important parts of both glycerol and drug metabolism occur in the liver.

In the example case for Solution 3, the ”Similarity of relations to other sets of genomic features”, H3K4me1 tracks are clustered based on the occupancy within reference genome tracks defining GO term associated genes. An expected result is that reference tracks representing an ontology terms that are specific to given cells, like ”immune response”, will only be occupied by H3K4me3 tracks collected from that cell type. In fact many examples for this are found in the heat map of Additional file 2B in Paper IV. The used GO term tracks is based on gene body coordinates. H3K4me3’s proven tendency to regulate gene expression through enhancer occupancy suggests that GO term tracks using gene related enhancer region coordinates. Such tracks has , however not yet been defined.

Chapter 5

Conclusions

Integrative analysis of next generation sequencing data will contribute to the curation of many diseases and offers promises towards the revelation of laws of chromatin dynamics. The current investment into such analysis, with no match in history, manifested by the formation of consortia for assaying various omics disciplines, will increase the resolution of current biological understanding. The contribution from smaller labs, although not as visible, is probably at least as important. Many hurdles, exemplified by the point list below stands in the way, making it difficult to predict the path of discovery.

1. Efficient handling of massive and complex data sets
2. Generating data that is free of noise
3. Reducing false discoveries of non existing relationships
4. Reducing false rejections of existing relationships
5. Providing credibility to true findings.

The opportunities for software development are vast and environments like R/Bioconductor, Galaxy and The Genomic HyperBrowser constitute a promising foundation to further research and development within the field of Integrative Epigenomic Analysis.

References

1. Waddington CH. The Epigenotype. Endeavour 1942.
2. Dawson MA and Kouzarides T. Cancer epigenetics: from mechanism to therapy. Cell 2012;150:12–27.
3. Baylin SB and Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. Nature reviews Cancer 2011;11:726–734.
4. Daxinger L and Whitelaw E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. Nature reviews Genetics 2012;13:153–162.
5. Milosavljevic A. Emerging patterns of epigenomic variation. Trends in genetics : TIG 2011;27:242–250.
6. Satterlee JS, Schübeler D, and Ng HH. Tackling the epigenome: challenges and opportunities for collaboration. Nature Biotechnology 2010;28:1039–1044.
7. Bernstein BE, Meissner A, and Lander ES. The mammalian epigenome. Cell 2007.
8. Richards EJ. Inherited epigenetic variation–revisiting soft inheritance. Nature reviews Genetics 2006;7:395–401.
9. Margueron R and Reinberg D. Chromatin structure and the inheritance of epigenetic information. Nature reviews Genetics 2010;11:285–296.
10. Horsley V and Pavlath GK. Forming a multinucleated cell: molecules that regulate myoblast fusion. Cells, tissues, organs 2004;176:67–78.
11. Grentzinger T, Armenise C, Brun C, et al. piRNA-mediated transgenerational inheritance of an acquired trait. Genome Research 2012;22:1877–1888.
12. Zhao Y, Li Q, Yao C, et al. Characterization and quantification of mRNA transcripts in ejaculated spermatozoa of fertile men by serial analysis

of gene expression. Human reproduction (Oxford, England) 2006;21:1583–1590.
13. Watanabe T, Totoki Y, Toyoda A, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature 2008;453:539–543.
14. Tam OH, Aravin AA, Stein P, et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 2008;453:534–538.
15. Bourc'his D and Voinnet O. A small-RNA perspective on gametogenesis, fertilization, and early zygotic development. Science (New York, NY) 2010;330:617–622.
16. Ciccia A and Elledge SJ. The DNA damage response: making it safe to play with knives. Molecular cell 2010;40:179–204.
17. Stankiewicz P and Lupski JR. Structural variation in the human genome and its role in disease. Annual review of medicine 2010.
18. Liu P, Carvalho CM, Hastings P, and Lupski JR. Mechanisms for recurrent and complex human genomic rearrangements. Current opinion in genetics & development 2012.
19. Chial H. Tumor suppressor (TS) genes and the two-hit hypothesis. Nature Education 2008.
20. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America 1971;68:820–823.
21. Chari R, Coe BP, Vucic EA, Lockwood WW, and Lam WL. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. BMC systems biology 2010;4:67.
22. Li B, Carey M, and Workman JL. The role of chromatin during transcription. Cell 2007.
23. Jones PA and Baylin SB. The fundamental role of epigenetic events in cancer. Nature reviews Genetics 2002;3:415–428.
24. Grigoryev SA and Woodcock CL. Chromatin organization - the 30 nm fiber. Experimental cell research 2012;318:1448–1455.
25. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. Cell 2007;129:823–837.
26. Berbenetz NM, Nislow C, and Brown GW. Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. PLoS genetics 2010;6.
27. Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. Nature 2012;489:101–108.

28. Pheasant M and Mattick JS. Raising the estimate of functional human sequences. Genome Research 2007;17:1245–1253.
29. Mattick JS. A new paradigm for developmental biology. 2007.
30. Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, and Elhaik E. On the immortality of television sets: ”function” in the human genome according to the evolution-free gospel of ENCODE. Genome biology and evolution 2013.
31. Gundersen S, Kalas M, Abul O, Frigessi A, Hovig E, and Sandve GK. Identifying elemental genomic track types and representing them uniformly. BMC bioinformatics 2011;12:494.
32. Dawkins R. Climbing Mount Improbable. W. W. Norton, 1997.
33. Su AI, Wiltshire T, Batalov S, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proceedings of the National Academy of Sciences of the United States of America 2004;101:6062–6067.
34. Nilsen TW and Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature 2010;463:457–463.
35. Ley TJ, Ding L, Walter MJ, et al. DNMT3AMutations in Acute Myeloid Leukemia. The New England journal of medicine 2010;363:2424–2433.
36. Shendure J and Ji H. Next-generation DNA sequencing. Nature Biotechnology 2008;26:1135–1145.
37. Hutchison CA. DNA sequencing: bench to bedside and beyond. Nucleic acids research 2007;35:6227–6237.
38. Lander ES, Linton LM, Birren B, Nusbaum C, and Zody MC. Initial sequencing and analysis of the human genome. Nature 2001.
39. Venter JC, Adams MD, Myers EW, and Li PW. The sequence of the human genome. Science (New York, NY) 2001.
40. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nature genetics 2007.
41. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409:928–933.
42. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, and Yu F. The international HapMap project. Nature 2003.
43. Durbin RM, Altshuler DL, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–1073.
44. Kaiser J. DNA sequencing. A plan to capture human diversity in 1000 genomes. Science (New York, NY) 2008;319:395.

45. Feuk L, Carson AR, and Scherer SW. Structural variation in the human genome. Nature reviews Genetics 2006;7:85–97.
46. Sebat J, Lakshmi B, Troge J, Alexander J, and Young J. Large-scale copy number polymorphism in the human genome. Science (New York, NY) 2004.
47. Iafrate AJ, Feuk L, Rivera MN, and Listewnik ML. Detection of large-scale variation in the human genome. Nature 2004.
48. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. Nature 2010;464:704–712.
49. Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, and Kim S. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nature 2010.
50. Craddock N, Hurles ME, Cardin N, and Pearson RD. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 2010.
51. McCarroll SA, Kuruvilla FG, Korn JM, and Cawley S. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature 2008.
52. Alkan C, Coe BP, and Eichler EE. Genome structural variation discovery and genotyping. Nature reviews Genetics 2011;12:363–376.
53. Landan G, Cohen NM, Mukamel Z, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nature genetics 2012;44:1207–1214.
54. Greaves M and Maley CC. Clonal evolution in cancer. Nature 2012;481:306–313.
55. Siegmund KD, Marjoram P, Tavaré S, and Shibata D. High DNA methylation pattern intratumoral diversity implies weak selection in many human colorectal cancers. PloS one 2011;6:e21657.
56. Siegmund KD, Marjoram P, Woo YJ, Tavaré S, and Shibata D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. Proceedings of the National Academy of Sciences of the United States of America 2009;106:4828–4833.
57. Van Loo P and Campbell PJ. ABSOLUTE cancer genomics. Nature Biotechnology 2012;30:620–621.
58. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. Nature Biotechnology 2012;30:413–421.

59. Adli M, Zhu J, and Bernstein BE. Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nature Methods 2010.
60. Goren A, Ozsolak F, Shoresh N, et al. Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. Nature Methods 2009;7:47–49.
61. Schroeder MP, Gonzalez-Perez A, and Lopez-Bigas N. Visualizing multidimensional cancer genomics data. Genome medicine 2013;5:9.
62. Zhou VW, Goren A, and Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nature reviews Genetics 2011;12:7–18.
63. Fuchs E, Tumbar T, and Guasch G. Socializing with the neighbors: stem cells and their niche. Cell 2004;116:769–778.
64. De Koning L, Corpet A, Haber JE, and Almouzni G. Histone chaperones: an escort network regulating histone traffic. Nature structural & molecular biology 2007;14:997–1007.
65. Alabert C and Groth A. Chromatin replication and epigenome maintenance. Nature reviews. Molecular cell biology 2012;13:153–167.
66. Rosner M and Hengstschläger M. Targeting epigenetic readers in cancer. The New England journal of medicine 2012;367:1764–1765.
67. Law JA and Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature reviews Genetics 2010;11:204–220.
68. Oliver KR and Greene WK. Transposable elements: powerful facilitators of evolution. Bioessays 2009;31:703–714.
69. Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. Nature reviews Genetics 2011;12:565–575.
70. Ji H, Ehrlich LIR, Seita J, et al. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature 2010;467:338–342.
71. Wang Z, Zang C, Rosenfeld JA, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics 2008;40:897–903.
72. Heintzman ND, Stuart RK, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics 2007;39:311–318.
73. Mikkelsen TS, Ku M, Jaffe DB, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 2007;448:553–560.
74. Velasco G, Hubé F, Rollin J, et al. Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine

somatic tissues. Proceedings of the National Academy of Sciences of the United States of America 2010;107:9281–9286.
75. Borgel J, Guibert S, Li Y, et al. Targets and dynamics of promoter DNA methylation during early mouse development. Nature genetics 2010;42:1093–1100.
76. Stadler MB, Murr R, Burger L, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 2011;480:490–495.
77. Hawkins RD, Hon GC, and Ren B. Next-generation genomics: an integrative approach. Nature reviews Genetics 2010;11:476–486.
78. Liu T, Ortiz JA, Taing L, et al. Cistrome: an integrative platform for transcriptional regulation studies. Genome biology 2011;12:R83.
79. Irizarry RA, Ladd-Acosta C, Wen B, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature genetics 2009;41:178–186.
80. Saxonov S, Berg P, and Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America 2006;103:1412–1417.
81. Weber M, Hellmann I, Stadler MB, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature genetics 2007;39:457–466.
82. Bird AP. CpG-rich islands and the function of DNA methylation. Nature 1986.
83. Deaton AM and Bird A. CpG islands and the regulation of transcription. 2011.
84. Bai L and Morozov AV. Gene regulation by nucleosome positioning. Trends in genetics : TIG 2010.
85. Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, and Schreiber SL. Global nucleosome occupancy in yeast. Genome biology 2004;5:R62.
86. Lee CK, Shibata Y, Rao B, Strahl BD, and Lieb JD. Evidence for nucleosome depletion at active regulatory regions genome-wide. Nature genetics 2004;36:900–905.
87. Nishida H, Suzuki T, Kondo S, Miura H, Fujimura I, and Hayashizaki Y. Histone H3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology 2006;14:203–211.
88. Bargaje R, Alam MP, Patowary A, et al. Proximity of H2A.Z containing nucleosome to the transcription start site influences gene expres-

sion levels in the mammalian liver and brain. Nucleic acids research 2012;40:8965–8978.
89. Mavrich TN, Ioshikhes IP, Venters BJ, et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Research 2008;18:1073–1083.
90. Chodavarapu RK, Feng S, Bernatavichute YV, et al. Relationship between nucleosome positioning and DNA methylation. Nature 2010;466:388–392.
91. Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, and Iyer VR. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS biology 2008;6:e65.
92. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews Genetics 2012;13:484–492.
93. Visel A, Rubin EM, and Pennacchio LA. Genomic views of distant-acting enhancers. Nature 2009;461:199–205.
94. Phillips JE and Corces VG. CTCF: master weaver of the genome. Cell 2009.
95. Latt SA. Microfluorometric detection of deoxyribonucleic acid replication in human metaphase chromosomes. Proceedings of the National Academy of Sciences of the United States of America 1973;70:3395–3399.
96. Goren A and Cedar H. Replicating by the Clock. Nature reviews. Molecular cell biology 2003;4:25–32.
97. Ryba T, Hiratani I, Lu J, et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Research 2010;20:761–770.
98. Karnani N, Taylor C, Malhotra A, and Dutta A. Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. Genome Research 2007.
99. De S and Michor F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nature Biotechnology 2011;29:1103–1108.
100. Guelen L, Pagie L, Brasset E, Meuleman W, and Faza MB. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 2008.
101. Cmarko D, Verschure PJ, and Otte AP. Polycomb group gene silencing proteins are concentrated in the perichromatin compartment of the mammalian nucleus. Journal of cell ... 2003.
102. Lanzuolo C and Orlando V. Memories from the polycomb group proteins. Annual review of genetics 2012;46:561–589.

103. Simon JA and Kingston RE. Mechanisms of polycomb gene silencing: knowns and unknowns. Nature reviews. Molecular cell biology 2009;10:697–708.
104. Sexton T, Schober H, Fraser P, and Gasser SM. Gene regulation through nuclear organization. Nature structural & molecular biology 2007;14:1049–1055.
105. Schoenfelder S, Sexton T, Chakalova L, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nature genetics 2009;42:53–61.
106. Huang N, Shah PK, and Li C. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. Briefings in bioinformatics 2012;13:305–316.
107. Stranger BE, Forrest MS, Dunning M, et al. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. Science (New York, NY) 2007;315:848–853.
108. Roa BB, Garcia CA, Wise CA, et al. Gene dosage as a mechanism for a common autosomal dominant peripheral neuropathy: Charcot-Marie-Tooth disease type 1A. Progress in clinical and biological research 1993;384:187–205.
109. Robertson KD. DNA methylation and human disease. Nature reviews Genetics 2005;6:597–610.
110. Rodrı□guez-Paredes M and Esteller M. Cancer epigenetics reaches mainstream oncology. Nature medicine 2011;17:330–339.
111. Conte M and Altucci L. Molecular pathways: the complexity of the epigenome in cancer and recent clinical advances. Clinical cancer research : an official journal of the American Association for Cancer Research 2012;18:5526–5534.
112. Stratton MR, Campbell PJ, and Futreal PA. The cancer genome. Nature 2009;458:719–724.
113. Hanahan D and Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–674.
114. Putiri EL and Robertson KD. Epigenetic mechanisms and genome stability. Clinical epigenetics 2010;2:299–314.
115. Gopalakrishnan S, Van Emburgh BO, and Robertson KD. DNA methylation in development and human disease. Mutation research 2008;647:30–38.
116. Easwaran H, Johnstone SE, Van Neste L, et al. A DNA hypermethylation module for the stem/progenitor cell signature of cancer. Genome Research 2012;22:837–849.
117. Berman BP, Weisenberger DJ, Aman JF, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal can-

cer coincide with nuclear lamina-associated domains. Nature genetics 2012;44:40–46.
118. Fraga MF, Ballestar E, Villar-Garea A, et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. Nature genetics 2005;37:391–400.
119. Johnson DS, Mortazavi A, Myers RM, and Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science (New York, NY) 2007;316:1497–1502.
120. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, and Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic acids research 2005;33:5868–5877.
121. Harris RA, Wang T, Coarfa C, and Nagarajan RP. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nature 2010.
122. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009;462:315–322.
123. Laird PW. Principles and challenges of genomewide DNA methylation analysis. Nature reviews Genetics 2010;11:191–203.
124. Collas P. The current state of chromatin immunoprecipitation. Molecular biotechnology 2010;45:87–100.
125. Metzker ML. Sequencing technologies — the next generation. Nature reviews Genetics 2009;11:31–46.
126. Boyle AP, Davis S, Shulha HP, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 2008;132:311–322.
127. Hesselberth JR, Chen X, Zhang Z, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nature Methods 2009;6:283–289.
128. Giresi PG, Kim J, McDaniell RM, Iyer VR, and Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Research 2007;17:877–885.
129. Auerbach RK, Euskirchen G, Rozowsky J, et al. Mapping accessible chromatin regions using Sono-Seq. Proceedings of the National Academy of Sciences of the United States of America 2009;106:14926–14931.
130. Deal RB, Henikoff JG, and Henikoff S. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. Science (New York, NY) 2010;328:1161–1164.

131. Ku CS and Roukos DH. From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic medicine. Expert review of medical devices 2013;10:1–6.
132. Schneider GF and Dekker C. DNA sequencing with nanopores. Nature Biotechnology 2012;30:326–328.
133. McLean CY, Bristor D, Hiller M, et al. GREAT improves functional interpretation of cis-regulatory regions. Nature Biotechnology 2010;28:495–501.
134. Allison DB, Cui X, Page GP, and Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nature reviews Genetics 2006;7:55–65.
135. Wu Z, Irizarry RA, Gentleman R, Murillo FM, and Spencer F. A model based background adjustment for oligonucleotide expression arrays. 2004.
136. Bolstad BM, Irizarry RA, Åstrand M, and Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England) 2003;19:185–193.
137. Irizarry RA, Hobbs B, Collin F, and Barclay YB. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England) 2003.
138. Mecham BH, Nelson PS, and Storey JD. Supervised normalization of microarrays. Bioinformatics (Oxford, England) 2010;26:1308–1315.
139. Olshen AB, Venkatraman ES, Lucito R, and Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics (Oxford, England) 2004;5:557–572.
140. Andrews S. Andrews: FASTQC. A quality control tool for high... - Google Scholar. ... ://www bioinformatics babraham ac uk/projects/fastqc 2010.
141. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 2013;14:R36.
142. Anders S and Huber W. Differential expression analysis for sequence count data. Genome biology 2010;11:R106.
143. Katz Y, Wang ET, Airoldi EM, and Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods 2010;7:1009–1015.
144. Gresham D, Boer VM, Caudy A, et al. System-level analysis of genes and functions affecting survival during nutrient starvation in Saccharomyces cerevisiae. Genetics 2011;187:299–317.

145. Pickrell JK, Marioni JC, Pai AA, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 2010;464:768–772.
146. Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential expression analysis methods for RNA-seq data. arXiv.org 2013.
147. Pepke S, Wold B, and Mortazavi A. Computation for ChIP-seq and RNA-seq studies. Nature Methods 2009;6:S22–S32.
148. Medvedev P, Stanciu M, and Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nature Methods 2009;6:S13–20.
149. Pique-Regi R, Ortega A, Tewfik A, and Asgharzadeh S. Detecting Changes in DNA Copy Number: Reviewing signal processing techniques. Signal Processing Magazine, IEEE 2012;29:98–107.
150. Xi R, Kim TM, and Park PJ. Detecting structural variations in the human genome using next generation sequencing. Briefings in functional genomics 2010.
151. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nature reviews Genetics 2010.
152. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. Genome Research 2009;19:1639–1645.
153. Jee J, Rozowsky J, Yip KY, et al. ACT: aggregation and correlation toolbox for analyses of genome tracks. Bioinformatics (Oxford, England) 2011;27:1152–1154.
154. Xu R and Wunsch DC. Clustering Algorithms in Biomedical Research: A Review. IEEE Reviews in Biomedical Engineering 2010;3:120–154.
155. Sandve GK, Gundersen S, Johansen M, et al. The Genomic HyperBrowser: an analysis web server for genome-scale data. Nucleic acids research 2013.
156. Goecks J, Nekrutenko A, Taylor J, and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology 2010;11:R86.
157. Blankenberg D, Kuster GV, Coraor N, et al. Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2001.
158. Giardine B, Riemer C, Hardison RC, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Research 2005;15:1451–1455.

159. Goecks J, Coraor N, Galaxy Team, Nekrutenko A, and Taylor J. NGS analyses by visualization with Trackster. Nature Biotechnology 2012;30:1036–1039.
160. Lazarus R, Kaspi A, Ziemann M, and The Galaxy Team. Creating reusable tools from scripts: the Galaxy Tool Factory. Bioinformatics (Oxford, England) 2012;28:3139–3140.
161. Sandve GK, Gundersen S, Rydbeck H, et al. The differential disease regulome. BMC genomics 2011;12:353.
162. Paulsen J, Lien TG, Sandve GK, et al. Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. Nucleic acids research 2013.
163. Favorov A, Mularoni L, Cope LM, et al. Exploring massive, genome scale datasets with the GenometriCorr package. PLoS computational biology 2012;8:e1002529.
164. Halachev K, Bast H, Albrecht F, Lengauer T, and Bock C. EpiExplorer: live exploration and global analysis of large epigenomic datasets. Genome biology 2012;13:R96.
165. Sadikovic B, Yoshimoto M, Al-Romaih K, Maire G, Zielenska M, and Squire JA. In vitro analysis of integrated global high-resolution DNA methylation profiling with genomic imbalance and gene expression in osteosarcoma. PloS one 2008;3:e2834.
166. Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, et al. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Research 2012;22:1735–1747.
167. Wang J, Zhuang J, Iyer S, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. 2012.
168. Giannopoulou EG and Elemento O. An integrated ChIP-seq analysis platform with customizable workflows. BMC bioinformatics 2011;12:277.
169. Ward LD and Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research 2012;40:D930–4.
170. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Research 2012;22:1790–1797.
171. Friedman N. Inferring cellular networks using probabilistic graphical models. Science (New York, NY) 2004;303:799–805.
172. Furnari FB, Fenton T, Bachoo RM, et al. Malignant astrocytic glioma: genetics, biology, and paths to treatment. 2007.

173. Alvord G, Roayaei J, Stephens R, and Baseler MW. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome ... 2007.
174. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics (Oxford, England) 2010;26:i237–45.
175. Drier Y, Sheffer M, and Domany E. Pathway-based personalized analysis of cancer. Proceedings of the National Academy of Sciences of the United States of America 2013;110:6388–6393.
176. McLendon R, Friedman A, Bigner D, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–1068.
177. Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. Science (New York, NY) 2008;321:1807–1812.
178. Vogelstein B and Kinzler KW. Cancer genes and the pathways they control. Nature medicine 2004;10:789–799.
179. Kschischang FR and Frey BJ. Factor graphs and the sum-product algorithm. Information Theory 2001.
180. Network CGAR. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474:609–615.
181. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–337.
182. Hon G, Ren B, and Wang W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS computational biology 2008;4:e1000201.
183. Ernst J and Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature Biotechnology 2010;28:817–825.
184. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, and Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods 2012;9:473–476.
185. Bock C, Halachev K, Büch J, and Lengauer T. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi) genomic data. Genome biology 2009.
186. Oinn T, Addis M, Ferris J, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics (Oxford, England) 2004;20:3045–3054.
187. Hull D, Wolstencroft K, Stevens R, et al. Taverna: a tool for building and running workflows of services. 2006.

188. Li P, Castrillo JI, Velarde G, et al. Performing statistical analyses on quantitative data in Taverna workflows: an example using R and maxdBrowse to identify differentially-expressed genes from microarray data. BMC bioinformatics 2008;9:334.
189. Goble CA, Bhagat J, Aleksejevs S, et al. myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic acids research 2010;38:W677–82.
190. Nielsen CB, Younesy H, O'Geen H, et al. Spark: A navigational paradigm for genomic data exploration. Genome Research 2012;22:2262–2269.
191. R Core Team. R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2012. URL: `http://www.R-project.org`.
192. Du P, Kibbe WA, and Lin SM. nuID: a universal naming scheme of oligonucleotides for illumina, affymetrix, and other microarrays. Biology direct 2007;2:16.
193. Du P, Kibbe WA, and Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics (Oxford, England) 2008;24:1547–1548.
194. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology 2004;5:R80.
195. Du P, Zhang X, Huang CC, et al. Comparison of Beta-value and Mvalue methods for quantifying methylation levels by microarray analysis. BMC bioinformatics 2010;11:587.
196. Lin SM, Du P, Huber W, and Kibbe WA. Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic acids research 2008;36:e11.
197. Verhaak RGW, Hoadley KA, Purdom E, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell 2010;17:98–110.
198. Louhimo R and Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics (Oxford, England) 2011;27:887–888.
199. Hoffman MM, Ernst J, Wilder SP, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic acids research 2013;41:827–841.
200. ENCODE Project Consortium, Bernstein BE, Birney E, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.

201. Jennings J and Hudson TJ. Reflections on the founding of the international cancer genome consortium. Clinical chemistry 2013;59:18–21.
202. Stratton MR. Exploring the Genomes of Cancer Cells: Progress and Promise. Science Signaling 2011;331:1553.
203. Margueron R and Reinberg D. The Polycomb complex PRC2 and its mark in life. Nature 2011;469:343–349.
204. Johnson MD, Mueller M, and Game L. Single Nucleotide Analysis of Cytosine Methylation by Whole□Genome Shotgun Bisulfite Sequencing. Current Protocols in … 2012.
205. Smith ZD, Gu H, Bock C, Gnirke A, and Meissner A. High-throughput bisulfite sequencing in mammalian genomes. Methods 2009.
206. Reshef DN, Reshef YA, Finucane HK, et al. Detecting novel associations in large data sets. Science (New York, NY) 2011;334:1518–1524.
207. Beroukhim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. Nature 2010;463:899–905.
208. Zhang Y, Shin H, Song JS, Lei Y, and Liu XS. Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. BMC genomics 2008;9:537.
209. Schmid CD and Bucher P. ChIP-Seq data reveal nucleosome architecture of human promoters. Cell 2007;131:
210. Kornacker K, Rye M, Håndstad T, and Drabløs F. The Triform algorithm: improved sensitivity and specificity in ChIP-Seq peak finding. BMC bioinformatics 2012;13:176.
211. Lee W, Tillo D, Bray N, et al. A high-resolution atlas of nucleosome occupancy in yeast. Nature genetics 2007;39:1235–1244.
212. Mavrich TN, Jiang C, Ioshikhes IP, et al. Nucleosome organization in the Drosophila genome. Nature 2008;453:358–362.
213. Schones DE, Cui K, Cuddapah S, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell 2008;132:887–898.
214. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, and Sidow A. Determinants of nucleosome organization in primary human cells. Nature 2011;474:516–520.
215. Pekowska A, Benoukraf T, Ferrier P, and Spicuglia S. A unique H3K4me2 profile marks tissue-specific gene regulation. Genome Research 2010;20:1493–1502.
216. Christensen BC, Houseman EA, Poage GM, et al. Integrated profiling reveals a global correlation between epigenetic and genetic alterations in mesothelioma. Cancer research 2010;70:5686–5694.

217. Sandgren J, Andersson R, Rada-Iglesias A, et al. Integrative epigenomic and genomic analysis of malignant pheochromocytoma. Experimental & molecular medicine 2010;42:484–502.
218. Chari R, Coe BP, Wedseltoft C, et al. SIGMA2: A system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes, and transcriptomes. BMC bioinformatics 2008;9:422.
219. Chari R, Thu KL, Wilson IM, et al. Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer. Cancer and Metastasis ... 2010;29:73–93.

Paper I

OPEN ACCESS Freely available online

PLOS ONE

Integrative Analysis Reveals Relationships of Genetic and Epigenetic Alterations in Osteosarcoma

Stine H. Kresse[1☯], Halfdan Rydbeck[1,2☯], Magne Skårn[1], Heidi M. Namløs[1], Ana H. Barragan-Polania[1,3], Anne-Marie Cleton-Jansen[4], Massimo Serra[5], Knut Liestøl[2], Pancras C. W. Hogendoorn[4], Eivind Hovig[1,2], Ola Myklebost[1,3], Leonardo A. Meza-Zepeda[1,3]*

1 Department of Tumour Biology, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway, **2** Department of Informatics, University of Oslo, Oslo, Norway, **3** Norwegian Microarray Consortium, Department of Molecular Biosciences, University of Oslo, Oslo, Norway, **4** Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands, **5** Laboratory of Experimental Oncology, Istituto Ortopedico Rizzoli, Bologna, Italy

Abstract

Background: Osteosarcomas are the most common non-haematological primary malignant tumours of bone, and all conventional osteosarcomas are high-grade tumours showing complex genomic aberrations. We have integrated genome-wide genetic and epigenetic profiles from the EuroBoNeT panel of 19 human osteosarcoma cell lines based on microarray technologies.

Principal Findings: The cell lines showed complex patterns of DNA copy number changes, where genomic copy number gains were significantly associated with gene-rich regions and losses with gene-poor regions. By integrating the datasets, 350 genes were identified as having two types of aberrations (gain/over-expression, hypo-methylation/over-expression, loss/under-expression or hyper-methylation/under-expression) using a recurrence threshold of 6/19 (>30%) cell lines. The genes showed in general alterations in either DNA copy number or DNA methylation, both within individual samples and across the sample panel. These 350 genes are involved in embryonic skeletal system development and morphogenesis, as well as remodelling of extracellular matrix. The aberrations of three selected genes, *CXCL5*, *DLX5* and *RUNX2*, were validated in five cell lines and five tumour samples using PCR techniques. Several genes were hyper-methylated and under-expressed compared to normal osteoblasts, and expression could be reactivated by demethylation using 5-Aza-2′-deoxycytidine treatment for four genes tested; *AKAP12, CXCL5*, *EFEMP1* and *IL11RA*. Globally, there was as expected a significant positive association between gain and over-expression, loss and under-expression as well as hyper-methylation and under-expression, but gain was also associated with hyper-methylation and under-expression, suggesting that hyper-methylation may oppose the effects of increased copy number for detrimental genes.

Conclusions: Integrative analysis of genome-wide genetic and epigenetic alterations identified dependencies and relationships between DNA copy number, DNA methylation and mRNA expression in osteosarcomas, contributing to better understanding of osteosarcoma biology.

Citation: Kresse SH, Rydbeck H, Skårn M, Namløs HM, Barragan-Polania AH, et al. (2012) Integrative Analysis Reveals Relationships of Genetic and Epigenetic Alterations in Osteosarcoma. PLoS ONE 7(11): e48262. doi:10.1371/journal.pone.0048262

Editor: Qian Tao, The Chinese University of Hong Kong, Hong Kong

Received April 4, 2012; **Accepted** September 21, 2012; **Published** November 7, 2012

Funding: This work was supported by the European Network to Promote Research into Uncommon Cancers in Adults and Children: Pathology, Biology and Genetics of Bone Tumours (EuroBoNeT), the Norwegian Cancer Society (Ragnvarda F. Sørvik and Håkon Starheim's legacy) and the Norwegian Research Council. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: leonardm@rr-research.no

☯ These authors contributed equally to this work.

Introduction

Osteosarcoma is the most common non-haematological primary malignant tumour of bone, occurring most commonly in the metaphyseal regions of long bones in adolescents and young adults, but also in patients over 40 years of age [1]. Almost all conventional osteosarcomas are high-grade malignant tumours with poor prognosis, and 20–25% of the patients have detectable metastases at diagnosis [2,3]. The 5-year survival rate for patients diagnosed with osteosarcoma without presence of metastasis is 60–65% [3,4,5], whereas it is only 20–28% for osteosarcoma patients with metastases at diagnosis [3,6,7]. Even though the survival rate has improved considerably after the introduction of neoadjuvant chemotherapy, the need for advances in treatment regimens is still high.

Most conventional osteosarcomas have complex karyotypes with numerous and highly variable genomic aberrations. A vast number of DNA copy number changes have been identified using chromosome- and microarray-based comparative genomic hybridisation (CGH and array CGH), more recently also utilizing high-density single nucleotide polymorphism (SNP) microarrays [8,9,10]. Few, if any, consistent chromosomal aberrations have been recognized in osteosarcoma, mainly consisting of recurrent alterations in 6p, 8q, 13q and 17p [9,10,11,12]. Many genes

become deregulated due to genomic aberrations, and DNA copy number and gene expression data have been combined to identify oncogenes and tumour suppressor genes in osteosarcomas [10,11,13,14]. Another important mechanism for down-regulation of gene expression is DNA methylation, more specifically at CpG sites in the promoter region of genes. It has been speculated that epigenetic mechanisms may be more prevalent than mutation in childhood cancers like retinoblastoma [15]. Although a number of research groups have reported comparisons of alterations in DNA copy number, DNA methylation and mRNA expression for other types of cancers [16,17,18], only a few studies have examined the interdependence of these types of mechanisms in osteosarcoma [19,20]. The benefits of an integrative approach are that driver genes and their regulatory mechanisms may be identified, as well as relationships between mechanisms. The identification of molecular markers and pathways contributing to osteosarcoma development and progression may facilitate better diagnosis and prognostication, as well as the development of new treatment strategies.

As part of EuroBoNeT, a European Network of Excellence on bone tumours (http://www.eurobonet.eu), we have access to a large collection of clinical samples and resources for pre-clinical studies. One such resource is a collection of 19 osteosarcoma cell lines, which have been previously characterised in detail, including DNA fingerprinting to guarantee their identity [21]. Genetic, phenotypic and functional characterisation have shown that these cell lines robustly represent osteosarcoma clinical samples [21,22,23]. The EuroBoNeT osteosarcoma cell line panel will serve as a highly valuable, well-characterised model system for basic and pre-clinical studies.

By using various microarray technologies, genome-wide genetic and epigenetic changes were analysed in the EuroBoNeT osteosarcoma cell line panel. DNA copy number changes have been mapped at high resolution using the Affymetrix Genome-Wide Human SNP Array 6.0, DNA methylation status of approximately 27,000 CpG sites have been identified using the Illumina HumanMethylation27 BeadChip and global mRNA expression data have been obtained using the Illumina HumanWG-6 v2 Expression BeadChip. The different levels of genome-wide information have been analysed separately and integrated in order to identify recurrently altered genes showing more than one type of aberration, as well as the dependencies of the different types of aberrations in osteosarcomas.

Results

Genetic and Epigenetic Alterations in Osteosarcoma Cell Lines

DNA copy number changes in the EuroBoNeT panel of 19 human osteosarcoma cell lines [21] were mapped at high resolution using the Affymetrix Genome-Wide Human SNP Array 6.0, DNA methylation status of approximately 27,000 CpG sites was identified using the Illumina HumanMethylation27 BeadChip and global mRNA expression data were obtained using the Illumina HumanWG-6 v2 Expression BeadChip. For the two latter types of data, two normal osteoblast and four normal bone samples were included as controls. Clinical data for the osteosarcoma cell lines and normal samples are given in Table S1.

Unsupervised hierarchical clustering of the 19 cell lines and 6 normal samples was performed in R v.2.13.0 using the three types of microarray data, and the resulting cluster dendrograms are shown in Figure 1. The clustering was performed using the genome-wide probe intensities for the DNA copy number data, avgBeta (average ratio of signal from probe detecting methylation relative to the sum of both probes) probe values for the DNA methylation data and variance-stabilizing transformation (vst) and quantile normalised probe intensities for the mRNA expression data. Based on the distance of the dendrograms, the cell lines appeared more similar based on overall gene expression than copy number, with methylation in between. The cell lines clustered in general differently based on each type of data, although some similarities were seen, such as the co-clustering of IOR/OS9 and IOR/OS18 for all data types. The HOS cell line and its derivatives 143B and MNNG/HOS clustered together for all data types, with HOS and MNNG/HOS being more similar in terms of gene expression and methylation, and 143B and HOS in terms of copy number. The clustering patterns did not correlate with the clinical information associated with the sample of origin (Table S1), the cell line phenotypes, including the status of *CDKN2A*, *MDM2* and *TP53*, nor with the differentiation capacity or *in vivo* tumour formation capacity [21,22].

Furthermore, all the normal samples clustered together in one branch based on the methylation data, whereas the osteoblasts clustered together with the osteosarcoma cell lines for the expression data. Based on the distance of the dendrograms, the normal samples were more similar to each other than the osteosarcoma cell lines were, especially for the methylation data. Since the clustering pattern of osteoblasts and bone samples was markedly different for the expression data, the further comparisons of methylation and expression levels in the osteosarcoma cell lines were performed against only the osteoblasts. The osteosarcoma cell lines and osteoblasts are both *in vitro* grown samples, and would be expected to better separate cancer-associated properties.

For each cell line, genes with DNA copy number aberrations (gain and loss) were identified using the SNP rank segmentation algorithm in Nexus. Probes detecting variation in DNA methylation (hyper- and hypo-methylation) compared to the normal osteoblasts were identified using a cut-off of deltaBeta >0.4 and <-0.4, whereas probes detecting variation in mRNA expression (over- and under-expression) compared to the normal osteoblasts were identified using a cut-off of vst transformed and quantile normalised ratio >1 and <-1. The probes were collapsed to gene level for the analyses, keeping the probe level information. The number of genes with each type of aberration for all the cell lines are plotted in Figure 2 and listed in Table S2.

For the copy number changes, most cell lines showed more genes with gains than with losses (Figure 2A). The cell lines U-2 OS and MNNG/HOS had a different pattern, with almost similar numbers of genes gained and lost, whereas KPD diverged from all the other cell lines having a higher number of genes lost than gained. Excluding these three outliers, there was a correlation between the number of genes with gain or loss ($R^2 = 0.56$). The distribution of number of genes gained and lost did not correlate with the clustering pattern based on the copy number changes (Figure 1).

As expected, most cell lines showed more hyper-methylation than hypo-methylation, and there was an inverse correlation between the number of genes hyper- and hypo-methylated (Figure 2B, $R^2 = 0.35$). The cell line 143B had almost 20 times more genes hyper-methylated than hypo-methylated, whereas IOR/OS14 had slightly more genes hypo-methylated than hyper-methylated. The distribution of number of genes hyper- and hypo-methylated showed a trend to correlate with the clustering pattern (Figure 1). The two main subclusters showed different distributions with respect to the number of genes hyper- and hypo-methylated, with the exception of the cell lines IOR/OS15 and IOR/OS18. The DNA copy number, DNA methylation and mRNA expression levels of the methyltransfer-

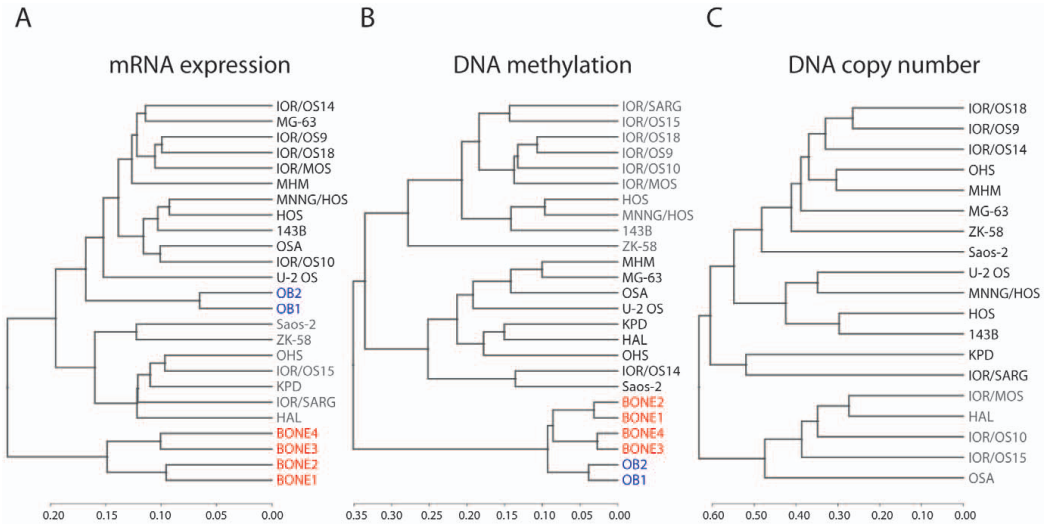


Figure 1. Hierarchical clustering of osteosarcoma cell lines and normal samples. Dendrograms from unsupervised hierarchical clustering of the osteosarcoma cell lines, normal bone and normal osteoblast samples based on genome-wide (A) mRNA expression (vst transformed and quantile normalised probe intensities), (B) DNA methylation (probe beta values) and (C) DNA copy number (probe intensities). The osteosarcoma cell lines have been colour-coded in gray and black, highlighting the two main subclusters, the normal bone samples in red and the normal osteoblast samples in blue. The clusters were made using Spearman correlation as distance measure and complete linkage.
doi:10.1371/journal.pone.0048262.g001

ase genes DNA (cytosine-5-)-methyltransferase 1, -3A and -3B (*DNMT1*, *-3A* and *-3B*) are shown in Figure S1. The genes were gained in several cell lines and hyper-methylated in some, but all cell lines except HAL showed similar mRNA expression levels as the normal osteoblasts. No correlations were found between the DNA copy number, DNA methylation and mRNA expression levels of *DNMT1*, *-3A* and *-3B*, and the number of hyper- and hypo-methylated genes.

The number of genes over- and under-expressed was more even, and there was a correlation between the number of genes over- and under-expressed (Figure 2C, $R^2 = 0.40$). The distribution of the number of over- and under-expressed genes reflected also partly the clustering pattern (Figure 1). The cell lines in one of the two main subclusters showed in general higher numbers of under-expressed genes.

A genome-wide frequency plot of alterations in DNA copy number is given in Figure S2. The cell lines showed more gains than losses, and an increased copy number of regions in almost every chromosome was present in more than 50% of the samples. The most frequent gains were regions in 2p, 14q, 20q and 8q, whereas the most frequent losses were regions in 13q, 3p and 6q. A genome-wide analysis using The Genomic HyperBrowser (http://hyperbrowser.uio.no/hb/) [24] identified an over-representation of gene-rich areas among frequently gained regions and gene-poor areas among frequently lost regions. The analysis was performed as two separate Monte Carlo-based hypothesis tests, for gain and loss respectively, giving p-value <0.001 in both cases. Tests were also performed separately for each chromosome arm (except the sex chromosomes), resulting in 27/39 significant arms for gain and all 39 arms significant for loss. Figure S3 shows the frequency plot of copy number aberrations and gene density for chromosome arms 2q, 8p, 19p and 19q, all with significant results for both gain and loss tests. The chromosome arms that were not significant for gain were 3p, 4p, 4q, 6q, 10p, 11p, 12p, 13q, 14q, 17q, 18p and 18q, and the frequency plot of copy number aberrations and gene density for these chromosome arms is shown in Figure S4.

The methylation data were analysed with the Bioconductor packages Limma and MethyLumi to identify differentially methylated genes compared to the normal osteoblasts. Using a cut-off of M-value ($\log_2$ ratio of intensity of probes detecting methylation and no methylation) >6, 328 significantly differentially methylated genes were identified, listed in Table S3. The gene list was analysed for functional enrichment in DAVID (Database for Annotation, Visualization and Integrated Discovery), and the top five terms in the top three clusters are listed in Table 1. The first cluster contained terms involving embryonic organ development and morphogenesis, as well as homeobox proteins and DNA binding, the second cluster contained terms involving thyroglobulin, whereas the third cluster contained terms involving potassium channel and ion transport. The top 10 clusters with all terms are listed in Table S4.

The expression data were analysed with the Bioconductor packages Limma and Lumi to identify differentially expressed genes compared to the normal osteoblasts. Using a cut-off of vst ratio >0.5, 283 significantly differentially expressed genes were identified, listed in Table S5. The gene list was analysed for functional enrichment in DAVID, and the top five terms in the top three clusters are listed in Table 2. The first cluster contained terms involving ribosome and translation, the second cluster terms involving fibrinogen, whereas the third cluster contained terms involving embryonic skeletal system and organ development and morphogenesis, as well as homeobox proteins and DNA binding. The top 10 clusters with all terms are listed in Table S6.

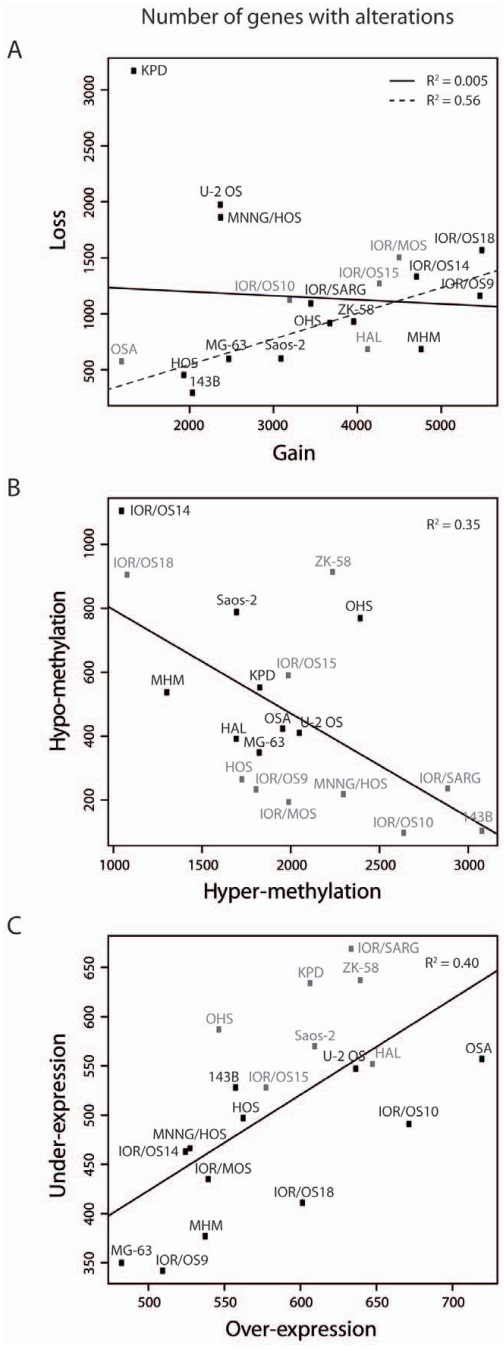


Figure 2. Number of genes with alterations. Plot of the number of genes with (A) gain and loss, (B) hyper- and hypo-methylation and (C) over- and under-expression for all the cell lines. The linear regression line is indicated in black. For the copy number (A), the linear regression line omitting the outlier samples U-2 OS, MNNG/HOS and KPD is indicated with a dashed line. The cell lines are colour-coded in gray and black according to the separation into two main subclusters from the respective unsupervised hierarchical clustering (Figure 1).
doi:10.1371/journal.pone.0048262.g002

Identification of Recurrently Altered Genes Using Genetic and Epigenetic Information

The lists of genes with alterations in DNA copy number, DNA methylation or mRNA expression level were combined for each cell line in order to identify genes showing more than one type of aberration. The 11,843 genes from chromosome 1–22 common to the three microarray platforms were included in the analyses. With two types of changes for each of the three data sets (gain and loss, hyper- and hypo-methylation, over- and under-expression), 12 two-way and 8 three-way combinations are possible. The combined lists with genes showing more than one type of aberration for the individual cell lines were subsequently compared in order to identify recurrently altered genes.

The number of genes showing more than one type of aberration is plotted as a function of number of samples with this occurrence for all two-way combinations in Figure 3. As can be seen, the combination gain and hyper-methylation had the highest frequency of occurrence, followed by gain and over-expression. In 6/19 cell lines (>30%), 546 genes were both gained and hyper-methylated, followed by 159 genes showing both gain and over-expression and 158 genes showing both hyper-methylation and under-expression.

The number of genes showing more than one type of aberration is plotted as a function of number of samples with this occurrence for all three-way combinations in Figure S5. In this case, 16 genes were gained, hyper-methylated and under-expressed in 6/19 cell lines (>30%), followed by 12 genes showing gain, hyper-methylation and over-expression. A recurrence plot for each data type is also given in Figure S5, showing that most genes with recurrent alterations were gained, followed by hyper-methylation.

To identify genes with altered expression level correlating with aberrations in DNA copy number or DNA methylation, the four two-way combinations gain/over-expression, hypo-methylation/over-expression, loss/under-expression and hyper-methylation/under-expression were considered. Using the sample recurrence threshold of six or more cell lines (>30%), these four combinations made up a total of 335 genes. The combinations gain/over-expression and hyper-methylation/under-expression gave the highest number of genes, 159 and 158, respectively. Of the 335 genes, only 11 showed simultaneous aberrations in both DNA copy number and DNA methylation. For genes with multiple probes, usually the same probes showed recurrent alterations.

Since changes in DNA copy number and DNA methylation may be alternative mechanisms for altering mRNA expression levels in the same direction, it was also investigated if genes with recurrent over-expression were either gained or hypo-methylated and if genes with recurrent under-expression were either lost or hyper-methylated in a total of six or more cell lines. However, only 15 additional genes were identified in this way, giving a total number of 350 recurrently altered genes. Thus, the majority of these genes showed alterations in either DNA copy number or DNA methylation, both within individual samples and across the sample panel.

This list of 350 genes, annotated with type of deviation and recurrence count, is given in Table S7.

The genomic locations of these 350 genes are visualised using Circos v0.52 in Figure 4. The genes were distributed rather evenly

Table 1. Enrichment analysis of differentially methylated genes using DAVID.

Cluster number	Enrichment score	Term	Counts	Population hits	FDR
1	5.13	Embryonic morphogenesis	25	307	4.7E-06
		Sequence-specific DNA binding	35	607	1.2E-05
		DNA-binding region:Homeobox	17	190	1.1E-04
		Embryonic organ development	17	172	2.2E-04
		Chordate embryonic development	22	331	1.6E-03
2*	3.47	Thyroglobulin type-1	5	17	0.24
		TY	5	17	0.27
		Thyroglobulin type-1	4	13	1.58
3	3.34	Voltage-dependent potassium channel	8	33	1.6E-03
		Ion transport	27	578	2.5E-03
		Potassium channel	10	78	5.3E-03
		Voltage-gated channel	12	150	0.04
		Potassium voltage-gated channel, alpha subunit, subfamilies A/C/D/F/G/S	6	20	0.03

The first five terms in the first three clusters are shown, with enrichment score. The counts and population hits are the number of genes in the gene list and background gene list, respectively, mapping to a specific term. FDR, false discovery rate.
*This cluster contained only three terms.
doi:10.1371/journal.pone.0048262.t001

over all chromosomes, but clusters of hyper-methylated and under-expressed genes were present in 3p, 11p and 19q. Clusters of gained and over-expressed genes were also present in 1q, 6p, 8q, 20q and 21q, whereas a cluster of lost and under-expressed genes was present in 9p. The homeobox genes were grouped in three gene clusters, located in 7p, 12q and 17q.

The gene list contained, among others, well-known oncogenes like cyclin-dependent kinase 4 (*CDK4*) and v-myc myelocytomatosis viral oncogene homolog (avian) (*MYC*), both gained and over-expressed, as well as transcription factors involved in normal bone development, like runt-related transcription factor 2 (*RUNX2*) and twist homolog 1 (Drosophila) (*TWIST1*). *RUNX2* was frequently gained and over-expressed, whereas *TWIST1* was frequently hyper-methylated and under-expressed in the cell lines compared to the osteoblasts. The list also contained a number of homeobox genes, 11 HOX family genes (*HOXA4*, *-A5*, *-A9*, *-B2*, *-B5*, *-B7*, *-B8*, *-B9*, *-C4*, *-C6* and *-C9*) and three other genes; distal-less homeobox 5 (*DLX5*), msh homeobox 1 (*MSX1*) and zinc fingers and homeoboxes 1 (*ZHX1*). All homeobox genes were frequently

Table 2. Enrichment analysis of differentially expressed genes using DAVID.

Cluster number	Enrichment score	Term	Counts	Population hits	FDR
1	3.28	Translational elongation	10	101	0.01
		Ribosome	10	87	0.01
		Ribosome	8	73	0.04
		Cytosolic ribosome	8	81	0.10
		Ribosomal protein	11	188	0.13
2	2.38	Fibrinogen, alpha/beta/gamma chain, C-terminal globular, subdomain 2	3	4	1.35
		Fibrinogen, alpha/beta/gamma chain, C-terminal globular, subdomain 1	4	23	4.19
		Fibrinogen C-terminal	4	32	9.72
		Fibrinogen, alpha/beta/gamma chain, C-terminal globular	4	32	10.5
		FBG	4	32	9.21
3	2.28	Embryonic skeletal system development	9	77	0.01
		Embryonic skeletal system morphogenesis	8	57	0.01
		Embryonic organ morphogenesis	10	133	0.09
		Skeletal system development	15	319	0.11
		Embryonic organ development	11	172	0.13

The first five terms in the first three clusters are shown, with enrichment score. The counts and population hits are the number of genes in the gene list and background gene list, respectively, mapping to a specific term. FDR, false discovery rate.
doi:10.1371/journal.pone.0048262.t002

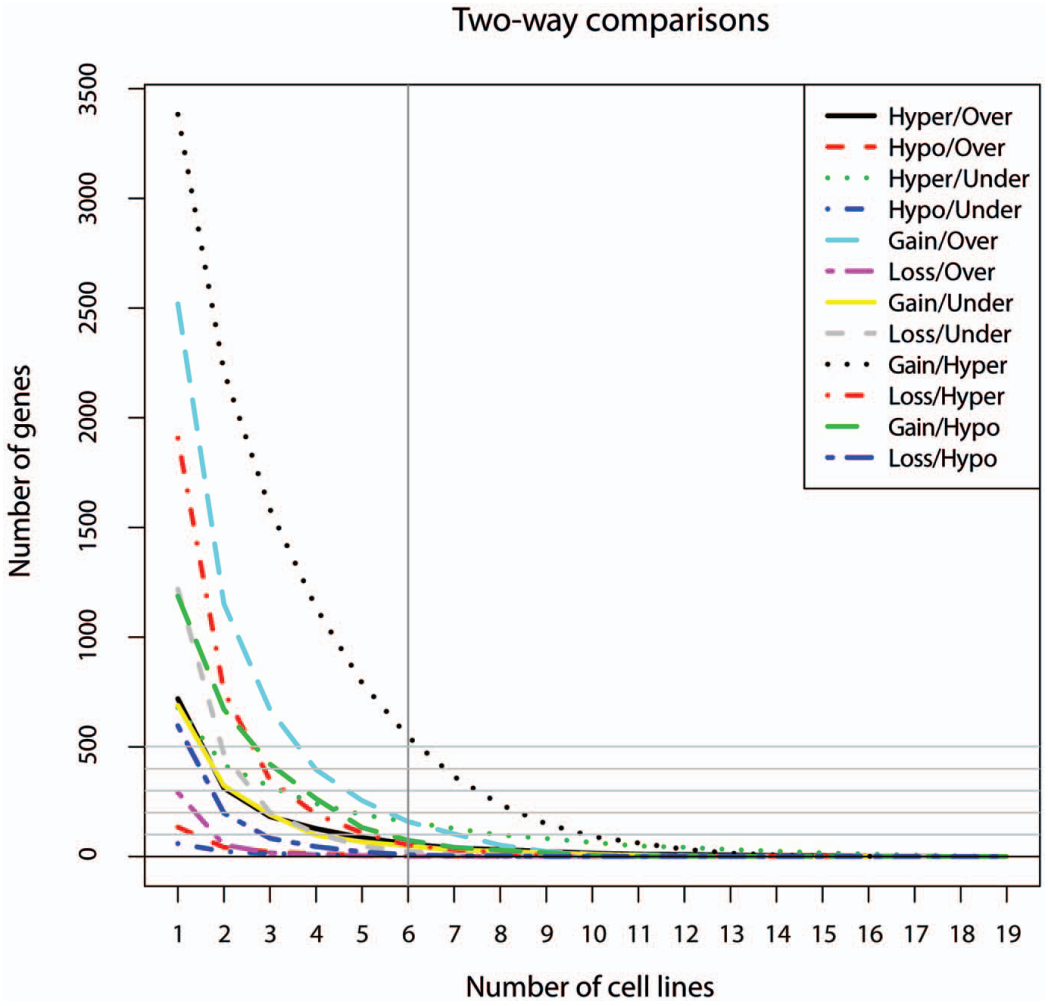


Figure 3. Number of genes with alterations for two-way combinations. Plot of the number of genes with alterations for all 12 two-way combinations at different sample recurrence thresholds. The recurrence threshold of 6/19 cell lines (>30%) is indicated with a black line.
doi:10.1371/journal.pone.0048262.g003

gained and over-expressed, except *MSX1* that was hyper-methylated and under-expressed.

DLX5 and *RUNX2* were gained and over-expressed in 6 and 7 of the 19 cell lines, respectively. The aberrations of these genes were validated in five of the cell lines and five osteosarcoma tumour samples using quantitative real-time PCR and RT-PCR, respectively. Clinical data for the tumour samples are given in Table S1. Figure 5 shows the DNA copy number and mRNA expression levels of *DLX5* and *RUNX2*. For the cell lines, the PCR and microarray data correlated well, except for the DNA copy number of *RUNX2* in IOR/OS14 and *DLX5* in KPD, where the PCR data showed normal copy number and not gain. All the tumour samples showed normal copy number of the genes, but showed increased expression of both *DLX5* and *RUNX2*, at similar levels as the cell lines showing over-expression.

The gene list was analysed for functional enrichment in DAVID, and the top five terms in the top three clusters are listed in Table 3. The first and third clusters both contained terms involving extracellular matrix, and terms involving signal peptide and collagen, respectively, whereas the second cluster contained terms involving embryonic skeletal system development and morphogenesis, as well as homeobox protein. The top 10 clusters with all terms are listed in Table S8.

Hierarchical clustering of the cell lines based on the expression level of these 350 genes is shown in Figure 6. Again, the clustering patterns correlated neither with the clinical information (Table S1) nor the known properties of the cell lines [21,22], but the cell lines

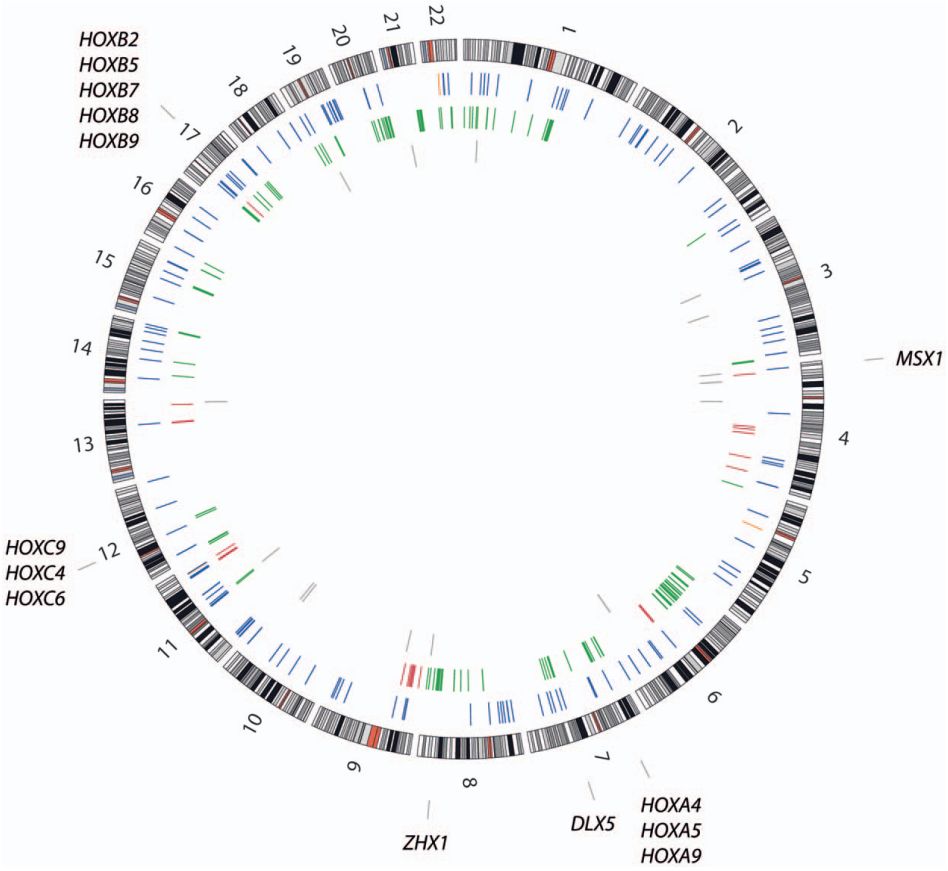


Figure 4. Genomic location of 350 genes that recurrently showed two types of aberrations. Circos plot showing the genomic location of the 350 genes that recurrently showed two types of aberrations. The locations are colour-coded according to the type of aberrations; orange, hypo-methylation/over-expression; blue, hyper-methylation/under-expression; green, gain/over-expression; red, loss/under-expression; gray, (gain or hypo-methylation)/over-expression and (loss or hyper-methylation)/under-expression. The genomic location of the 14 homeobox genes in the list is indicated in the outermost circle.
doi:10.1371/journal.pone.0048262.g004

separated in two main subgroups identical to the cluster based on the global gene expression data. The main terms from functional enrichment analysis using DAVID is indicated for each main subcluster of genes, showing mainly over-expression of genes associated with the terms skeletal development and homeodomain, whereas genes associated with the terms extracellular matrix, oxidative stress and collagen were mainly under-expressed. The same figure with all gene names shown is given in Figure S6.

A functional enrichment analysis was also performed for the 159 genes with gain and over-expression separately, generating a first cluster with terms involving embryonic skeletal system development and homeodomain. The top 10 clusters with all terms are listed in Table S9. A similar analysis of the 158 genes with hyper-methylation and under-expression generated a first cluster with the term extracellular matrix organisation. The top 10 clusters with all terms are listed in Table S10. The other two-way combinations hypo-methylation/over-expression and loss/under-expression did not generate any significant functional terms due to the low number of altered genes.

Relationships between Different Mechanisms for Alteration of Gene Regulation

The effects of alterations in DNA copy number and DNA methylation on mRNA expression were examined globally, as well as how the different types of aberrations (gain and loss, hyper- and hypo-methylation, over- and under-expression) related to each other. Heat maps visualising the odds ratios and significance of data dependencies for the 12 two-way combinations are shown in Figure 7. The odds ratio is a measure of effect size, describing the strength of association or non-independence between two binary data values. The significance was determined using Bonferroni-corrected chi-square p-values (p-value <0.05). The cell lines were clustered based on the odds ratios for the different categories of two-way combinations. No clear pattern between the clustering for

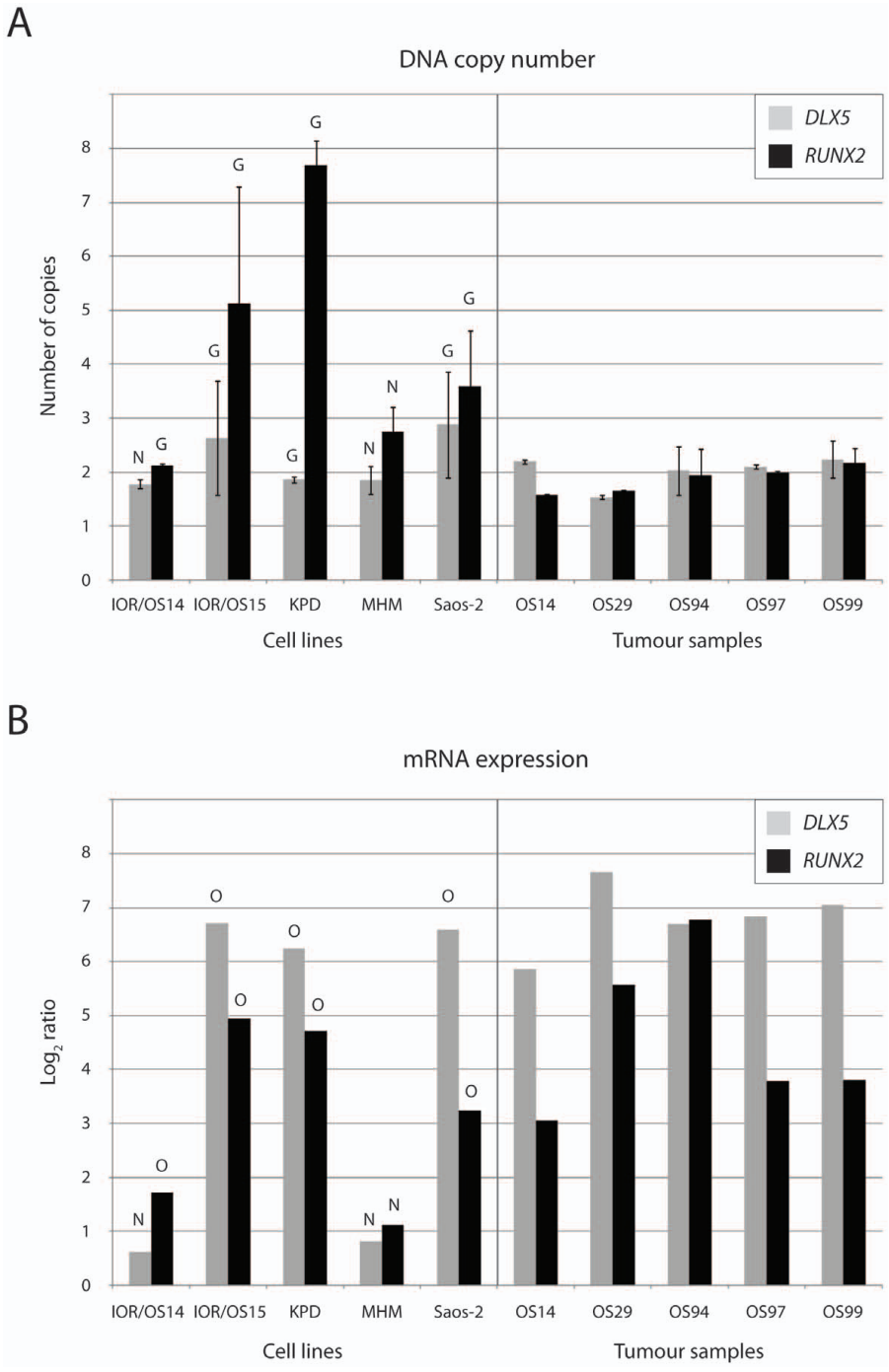

A
DNA copy number
DLX5
RUNX2
Number of copies
0
1
2
3
4
5
6
7
8
N
G
IOR/OS14
IOR/OS15
KPD
MHM
Saos-2
OS14
OS29
OS94
OS97
OS99
Cell lines
Tumour samples
B
mRNA expression
DLX5
RUNX2
Log2 ratio
0
1
2
3
4
5
6
7
8
N
O
IOR/OS14
IOR/OS15
KPD
MHM
Saos-2
OS14
OS29
OS94
OS97
OS99
Cell lines
Tumour samples

Figure 5. DNA copy number and mRNA expression of *DLX5* and *RUNX2*. Plot of (A) DNA copy number levels of *DLX5* and *RUNX2* based on quantitative real-time PCR and (B) mRNA expression levels of *DLX5* and *RUNX2* based on quantitative real-time RT-PCR, in five cell lines and five tumour samples. The DNA copy number levels have been normalised to the average copy number of two control genes, *EEF1G* and *FBXO11*, whereas the mRNA expression levels have been normalised to the expression of the house-keeping gene *GAPDH* and then to the average expression level of the two normal osteoblast samples. The DNA copy number and mRNA expression levels based on the microarray data are indicated for the cell lines; N, normal copy number/expression; G, gain; O, over-expression.
doi:10.1371/journal.pone.0048262.g005

the different categories was identified. Here, the related cell lines HOS, 143B and MNNG/HOS did not cluster together. Again, the clustering patterns correlated neither with the clinical information (Table S1) nor the known properties of the cell lines [21,22].

The dependencies between DNA copy number and mRNA expression were comparatively strong and significant for all cell lines. There was a positive association of gain/over-expression and loss/under-expression, and conversely gain/under-expression and loss/over-expression showed a negative association. For the DNA copy number and DNA methylation, there was in general either no association or a negative association for the different combinations, and only a few cell lines showed significant dependencies of some of the combinations.

The dependencies of DNA methylation and mRNA expression were significant in some of the cell lines, particularly for the combination hyper-methylation and under-expression that showed a positive association. The three other two-way combinations had in general a negative association. The two cell lines 143B and IOR/MOS differed from the other cell lines by having a positive association between hypo-methylation and over-expression.

Methylation of CpG islands in promoter regions may silence gene expression, and is one mechanism for inactivating genes. Of the 350 genes that showed two types of aberrations in at least 6/19 cell lines, 158 genes were hyper-methylated and under-expressed. One of the most frequently hyper-methylated and under-expressed genes was chemokine (C-X-C motif) ligand 5 (*CXCL5*), altered in 18/19 cell lines. Methylation-specific PCR and quantitative real-time RT-PCR were used to validate the promoter methylation status and the expression level of *CXCL5*, respectively, in five of the cell lines and five osteosarcoma tumour samples. Figure 8 shows the DNA methylation and mRNA expression levels of *CXCL5*, and gel pictures of the methylation-specific PCR products are shown in Figure S7. For the cell lines, the PCR and microarray data correlated well. The microarray data indicated that IOR/OS14 was the only cell line not hyper-methylated compared to the osteoblasts, but the PCR data showed that *CXCL5* was partially methylated also in this cell line, as well as in the osteoblasts (Figure S7). Partial or full methylation of the investigated CpG island in the promoter region was identified in all cell lines and tumour samples, and all samples except OS94 showed under-expression.

To validate a causal association between hyper-methylation and under-expression, DNA methylation was removed by culturing the cells in a medium containing 5-Aza-2′-deoxycytidine and the effect on gene expression levels investigated. Four genes being frequently hyper-methylated and under-expressed were selected; *CXCL5* (18/19 cell lines), A kinase (PRKA) anchor protein 12 (*AKAP12*) (14/19 cell lines), EGF containing fibulin-like extracellular matrix protein 1 (*EFEMP1*) (10/19 cell lines) and interleukin 11 receptor, alpha (*IL11RA*) (10/19 cell lines). Twelve of the cell lines were treated with 5-Aza-2′-deoxycytidine, and the gene expression level of these four genes with and without treatment was determined by quantitative real-time RT-PCR, as shown in Figure 9.

The extent of reactivation of gene expression varied between the cell lines. *CXCL5*, which showed most frequent hyper-methylation and under-expression (18/19 cell lines), was reactivated in all of the tested cell lines, with two cell lines showing

Table 3. Enrichment analysis of 350 genes that recurrently showed two types of aberrations using DAVID.

Cluster number	Enrichment score	Term	Counts	Population hits	FDR
1	4.14	Extracellular matrix	28	269	1.5E-04
		Secreted	66	1247	0.005
		Signal	105	2333	0.005
		Signal peptide	105	2333	0.006
		Extracellular region part	48	811	0.047
2	4.08	Skeletal system development	34	281	1.2E-07
		Embryonic morphogenesis	27	255	2.6E-04
		Embryonic skeletal system development	14	69	3.7E-04
		Short sequence motif:Antp-type hexapeptide	8	22	0.004
		Homeobox protein, antennapedia type, conserved site	8	23	0.005
3	3.45	Extracellular matrix	28	269	1.5E-04
		Trimer	9	23	2.6E-04
		Extracellular matrix	21	192	0.001
		Proteinaceous extracellular matrix	25	247	0.001
		Collagen	9	31	0.005

The first five terms in the first three clusters are shown, with enrichment score. The counts and population hits are the number of genes in the gene list and background gene list, respectively, mapping to a specific term. FDR, false discovery rate.
doi:10.1371/journal.pone.0048262.t003

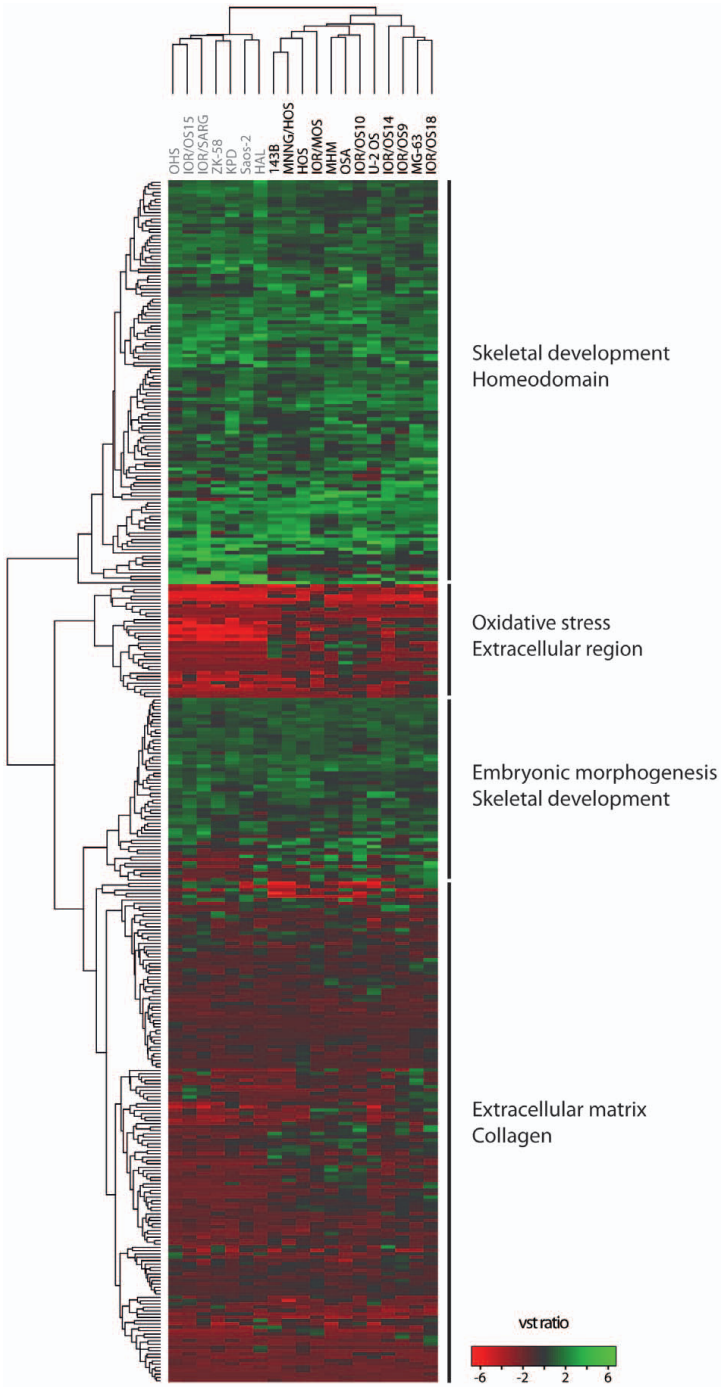

OHS
IOR/OS15
IOR/SARG
ZK-58
KPD
Saos-2
HAL
143B
MNNG/HOS
HOS
IOR/MOS
MHM
OSA
IOR/OS10
U-2 OS
IOR/OS14
IOR/OS9
MG-63
IOR/OS18
Skeletal development
Homeodomain
Oxidative stress
Extracellular region
Embryonic morphogenesis
Skeletal development
Extracellular matrix
Collagen
vst ratio
-6
-2
2
6

Figure 6. Hierarchical clustering based on 350 genes that recurrently showed two types of aberrations. Hierarchical clustering of the osteosarcoma cell lines based on the expression level of the 350 genes that recurrently showed two types of aberrations. The main terms from functional enrichment analysis using DAVID is indicated for each main subcluster of genes. The cell lines are colour-coded in gray and black according to the separation into two main subclusters from the unsupervised hierarchical clustering based on the global mRNA expression (Figure 1A). The cluster was made using Euclidian as distance measure and complete linkage. Green, increased gene expression; red, decreased gene expression.
doi:10.1371/journal.pone.0048262.g006

>100-fold increased expression level. The only cell line that was not hyper-methylated and under-expressed, IOR/OS14, showed the lowest level of increased expression (2–3 fold). For the genes *EFEMP1* and *AKAP12*, five and two cell lines showed >2-fold increased expression level, respectively, while only a low effect was observed for *IL11RA*. The genes that showed reactivation of expression were initially hyper-methylated in the affected cell lines, thus the demethylation treatment did not seem to affect the expression levels in general.

A heat map visualising the odds ratios and significance of data dependencies for different combinations of hyper-methylation and expression conditioning on the copy number state is shown in Figure 10. The significance was determined using Bonferroni-corrected chi-square p-values (p-value <0.05). The alterations were divided into three states for the DNA copy number (gain, normal and loss) and mRNA expression data (over-expression, normal and under-expression). Regardless of the DNA copy number state, no significant dependency was found, except for a few samples that showed dependencies between hyper-methylation and either normal or over-expression. However, there was a positive association of hyper-methylation and under-expression for genes with gain, and all cell lines, except two, showed significant dependencies of this combination. For some samples, there was also a positive association and significant dependency between hyper-methylation and under-expression for genes with normal copy number, but not for genes with loss.

Plots combining DNA copy number, DNA methylation and mRNA expression levels for the 16 genes that showed gain, hyper-methylation and under-expression in at least 6/19 cell lines are shown in Figure S8. The levels of methylation and expression anti-correlated in general quite well, but there were no clear differences in the pattern of methylation and expression levels between the cell lines with gain or normal copy number/loss for these genes.

Discussion

Cell lines are valuable model systems when studying cancer biology, especially for rare tumours like osteosarcomas, where material from clinical samples is scarce. The EuroBoNeT panel of 19 osteosarcoma cell lines used here has previously been characterised by many means, including genetic, phenotypic and functional characterisation, and the cell lines reflect well many properties of osteosarcoma tumours [21,22,23]. Thus, the cell line panel constitutes a highly valuable model system for analyses of genetic and epigenetic aberrations in osteosarcomas.

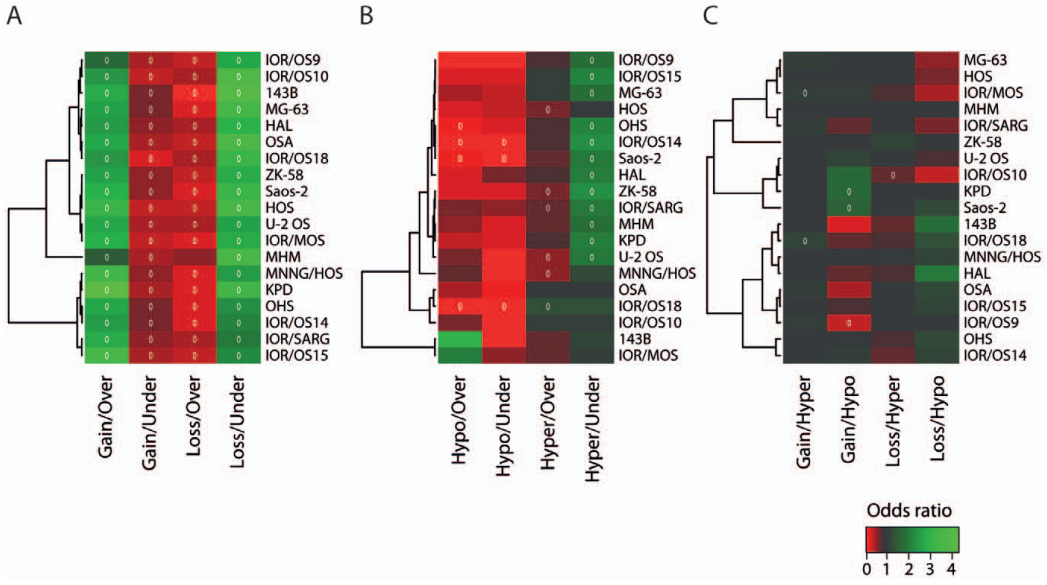


Figure 7. Data dependencies for two-way combinations. Heat map plots visualising the odds ratio and significance of data dependencies, with unsupervised hierarchical clustering of the cell lines, for two-way combinations of (A) DNA copy number and gene expression, (B) DNA methylation and mRNA expression and (C) DNA copy number and DNA methylation. The colours of the heat map plot represent the odds ratio for a gene of having one type of aberration given that it has another type of aberration. Green, positive association (odds ratio >1); black, no association (odds ratio = 1) and red, negative association (odds ratio <1). A white circle indicates significance (Benjamini & Hochberg-corrected chi-square p-value <0.05).
doi:10.1371/journal.pone.0048262.g007

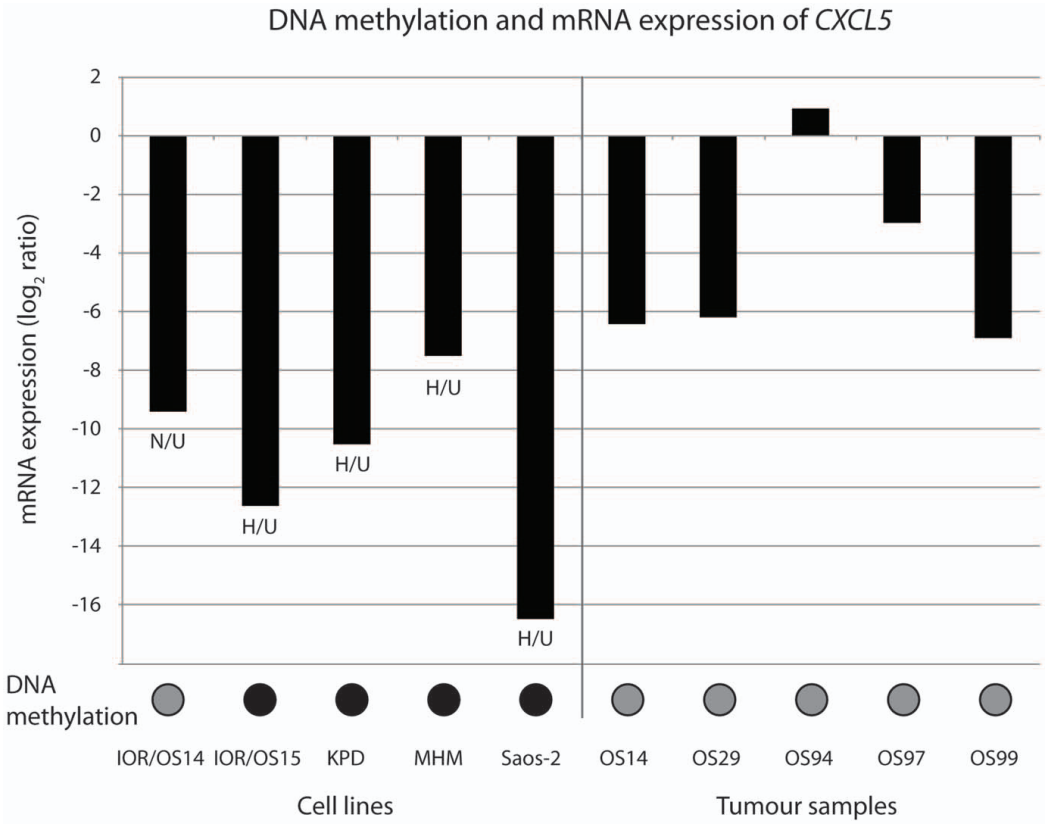


Figure 8. DNA methylation and mRNA expression of *CXCL5*. Plot of the DNA methylation status and mRNA expression level of *CXCL5* based on methylation-specific PCR and quantitative real-time RT-PCR, respectively, in five cell lines and five tumour samples. The mRNA expression levels have been normalised to the expression of the house-keeping gene *GAPDH* and then to the average expression level of the two normal osteoblast samples. The DNA methylation status is indicated with coloured circles; black, full methylation, grey, partial methylation. The DNA methylation and mRNA expression levels based on the microarray data are indicated for the cell lines; N, normal methylation; H, hyper-methylation; U, under-expression.
doi:10.1371/journal.pone.0048262.g008

In line with most conventional osteosarcomas, the cell lines showed a vast number of DNA copy number changes, reflecting the extreme genetic instability hallmarking high-grade osteosarcoma. Although recurrent alterations have been reported for almost every chromosome in osteosarcoma, gain of regions in 6p, 8q and 17p and loss of regions in 13q are most frequently reported [9,10,11,12]. These regions were recurrently altered in more than 50% of the cell lines (Figure S2). High-level amplification was found in 6p and 8q, as well as in 1q, which has also been previously reported [11,12,25,26]. When performing an unsupervised hierarchical clustering based on the DNA copy number profiles of the 19 cell lines and 32 osteosarcoma clinical samples [27], the cell lines were not systematically different from the clinical samples and all samples clustered intermingled (data not shown). All together, the results suggest that these cell lines are representative for osteosarcoma clinical samples in terms of DNA copy number changes.

Although the cell lines showed slightly more frequent regions of gain than loss, the number of genes with gain was far higher than the number of genes with loss for most cell lines (Figure 2A and Table S2). Genome-wide analysis using The Genomic Hyper-Browser showed that regions with high frequencies of gain were significantly associated with gene-rich regions of the genome and conversely regions with high frequencies of loss with gene-poor regions. There was a significant association both at the genome-wide level and for most individual chromosome arms (Figure S3 and S4). A similar analysis of 3,131 cancer specimens belonging to several histological types demonstrated that deletions showed a bias towards regions of low gene density, whereas no association was observed for amplifications [28]. This indicates that the association of gain and gene-rich areas observed here is special for osteosarcoma, or perhaps detectable because of the unusually high number of amplified regions. Since gaining one copy gives less relative change of gene dosage than losing one, and also does not remove functional germ-line or somatic gene variation, it

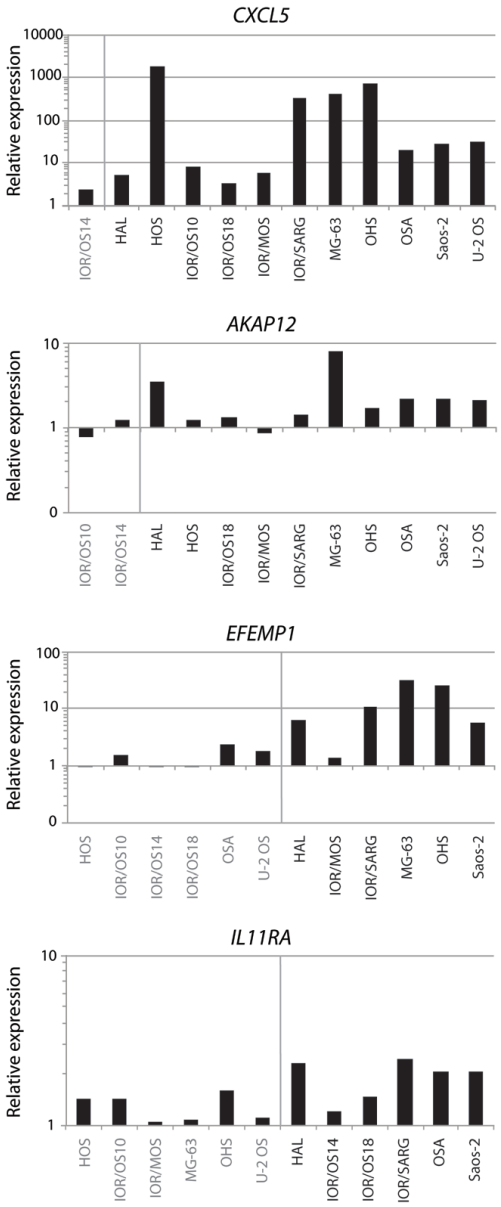


Figure 9. Gene expression after demethylation treatment. Relative gene expression levels of the frequently hyper-methylated and under-expressed genes *CXCL5, AKAP12, EFEMP1* and *IL11RA* after treatment with the demethylating agent 5-Aza-2′-deoxycytidine in 12 of the cell lines. The cell lines with hyper-methylation and under-expression of the genes are colour-coded in black, whereas gray colour indicates no aberrations in DNA methylation. The expression level without treatment has been set to 1 for each cell line.
doi:10.1371/journal.pone.0048262.g009

seems likely that initial loss of regions is on average more detrimental than gain. A general advantage of gains is that parts of an originally gained region that is disadvantageous may subsequently be lost, in this way generating the smaller, more focussed amplicons observed around some typical oncogenes, or the expression of passenger genes may be down-regulated by other mechanisms. For losses, on the other hand, regaining lost sequences is more complicated, as it requires additional rearrangement of the intact chromosome copy, and loss of heterozygosity would be maintained.

Since osteosarcomas have so many aberrations, a majority of these are most likely due to general instability of the genome. Recently, a new mechanism for genetic instability in cancer cells has been described, termed chromothripsis, in which a single chromosome is fragmented and then reassembled [29]. Chromothripsis has been suggested to occur in 2–3% of cancers, but the phenomenon has been observed in 25% of osteosarcoma and chordoma samples, affecting several chromosomes [29]. However, it seems likely that many of the genomic aberrations do not provide any advantage and may represent just genomic "noise". Such noise would be expected to be better tolerated in gene-poor regions, but cannot explain the enrichment of gains in gene-rich regions, which appears to be oncogenically more relevant.

In contrast to DNA copy number profiles, mRNA expression profiles are more dynamic and may be more influenced by cell culturing and growth conditions. However, the comparison with normal osteoblast cultures rather than bone tissue should cancel most of the effects of *in vitro* growth. In previous work, it was shown that mRNA expression profiles characteristic of the histological subtypes of primary high-grade osteosarcoma clinical samples are preserved in these cell lines [23], indicating that they are representative of the primary tumour from which they are derived.

Although the understanding of epigenetic regulation of specific genes in osteosarcomas is increasing [30,31], little is so far known about the global DNA methylation patterns. The cell lines showed in general more genes with hyper-methylation than hypo-methylation. Previous studies have shown that promoter-associated CpG islands are frequently hyper-methylated in cancer, whereas global hypo-methylation is often seen in gene-poor areas of the genome (reviewed in [32]). In line with the results here, previous investigations using Me-DIP-chips showed more hyper-methylation than hypo-methylation events in osteosarcoma tumours and cell lines compared to normal osteoblasts [19,20]. Interestingly, there was in general an inverse relationship between the number of genes with hyper- and hypo-methylation, as opposed to the DNA copy number and mRNA expression (Figure 2). In contrast to the DNA copy number, there seemed to be a relationship between the number of genes with hyper- and hypo-methylation and the clustering pattern. However, no associations were found between the DNA copy number, DNA methylation and mRNA expression levels of the methyltransferases *DNMT1, -3A* and *-3B* (Figure S1) and the number of hyper- and hypo-methylated genes for the individual cell lines (Figure 2 and Table S2).

Functional enrichment analyses of the differentially methylated and expressed genes, respectively (Table 1, 2, S5 and S6), showed common terms like embryonic organ development and morphogenesis. However, the overlap between the lists was limited, supporting the notion that DNA methylation is only one among several mechanisms influencing gene expression.

Although the genetic and epigenetic profiling data provide valuable information on their own, an integrative approach may facilitate the identification of key genes and regulatory mechan-

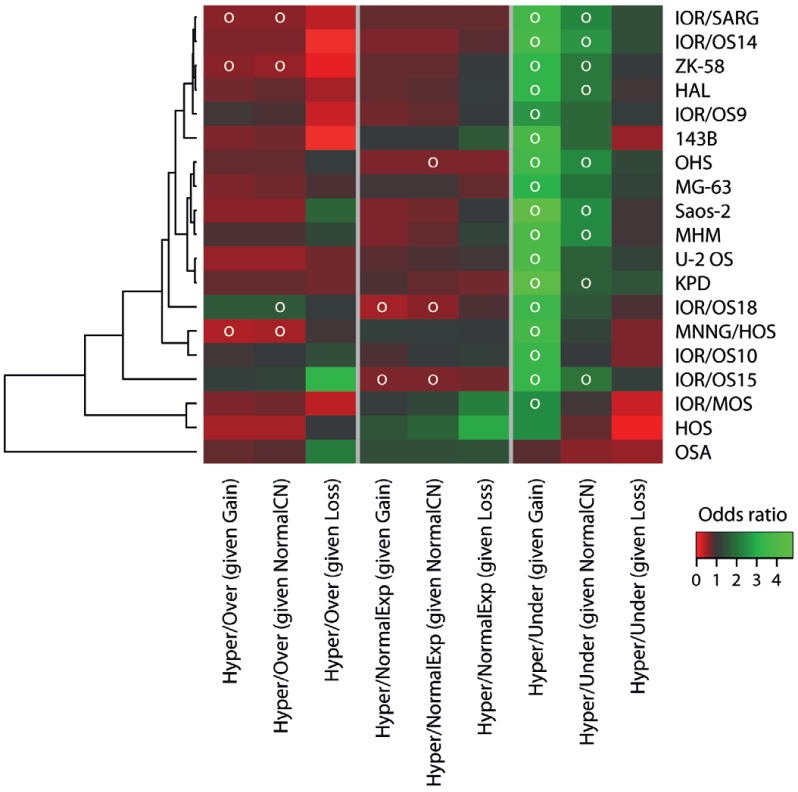


Figure 10. Data dependencies for three-way combinations. Heat map plots visualising the odds ratio and significance of data dependencies, with unsupervised hierarchical clustering of the cell lines, for combinations of hyper-methylation and mRNA expression, conditioning on the DNA copy number status. The colours of the heat map plot represent the odds ratio for a gene of having one type of aberration given that it has another type of aberration. Green, positive association (odds ratio >1); black, no association (odds ratio = 1) and red, negative association (odds ratio <1). A white circle indicates significance (Benjamini & Hochberg-corrected chi-square p-value <0.05). Exp, mRNA expression; CN, DNA copy number.
doi:10.1371/journal.pone.0048262.g010

isms involved in tumour development. A method originally used for the same type of data and for osteosarcoma cell lines [19] was adopted and further developed. In that study [19], cut-offs of intensity ratios were used to identify genes with aberrations in DNA copy number, DNA methylation and mRNA expression for individual samples, and Venn (two-way/three-way) analysis was used to select genes that showed alterations in more than one type of data. This approach was further improved by using the greater number of cell lines analysed here to filter the gene list obtained by Venn analysis based on recurrence. In this way, only genes that have the same two-way alteration in at least a certain number of the samples were identified. In addition, statistical tests for individual samples and heatmap visualisations were used to evaluate dependencies of types of data.

To examine how the different types of aberrations relate to each other, the odds ratios and significance of data dependencies for the 12 two-way combinations were identified for each cell line (Figure 7). There was a significant dependency between DNA copy number and mRNA expression, and to some extent between DNA methylation and mRNA expression, particularly for the combination hyper-methylation and under-expression. However, there was no dependency between the DNA copy number and DNA methylation, although the combination showing the highest number of recurrently altered genes was gain and hyper-methylation (Figure 3). Similar studies investigating five osteosarcoma tumours and two cell lines, respectively, showed strong correlation between gain and over-expression, loss and under-expression as well as gain and hypo-methylation [19,20]. In contrast to the results presented here, few genes showed gain and hyper-methylation. A reason why no dependency between gain and hypo-methylation was observed in this study may be the low number of hypo-methylated genes detected.

Genes for which an altered expression level was consistent with aberrations in DNA copy number or DNA methylation (gain/over-expression, hypo-methylation/over-expression, loss/under-expression and hyper-methylation/under-expression) were identified for each cell line. Since different mechanisms for alteration of a certain pathway may be involved in each cell line, a test looking for genes significantly altered in all cell lines might not detect samples with defects in the same pathway, because the specific genes affected may vary. On the other hand, a recurrence threshold is needed to filter out noise or sample-specific events and

to identify pathogenic alterations of general importance. A gene aberration frequency threshold of six or more cell lines (>30%) was chosen, giving a total of 335 recurrently altered genes. Interestingly, only 11 of these 335 genes showed simultaneous aberrations in both DNA copy number and DNA methylation. In addition, only 15 additional genes were identified when allowing genes with over-expression to be either gained or hypo-methylated and genes with under-expression to be either lost or hyper-methylated in a total of six or more cell lines, increasing the total number to 350 genes (Figure 6 and Table S7). This suggests that the expression levels of most genes with two types of aberrations, including aberrant mRNA expression, are regulated by alterations in either DNA copy number or DNA methylation, or conversely, that these mechanisms alter the activity of different subsets of genes. However, there will be other mechanisms like point mutations, loss of heterozygosity, nucleosome occupancy, micro-RNA (miRNA) or transcription factor regulation that also influence the mRNA expression levels.

By selecting genes with recurrent alterations in at least two of the three types of data, a list of genes involved in important biological functions and in a limited number of critical pathways was identified. Based on functional enrichment analysis, the most striking biological processes were development of the embryonic skeletal system and remodelling of the extracellular matrix (Table 3). A similar study of osteosarcoma showed that the most significant gene network, based on cumulative changes in DNA copy number, DNA methylation and mRNA expression, contained genes involved in organ and cellular development [20]. Although terms involving embryonic skeletal system were also identified using the list of differentially expressed genes from the comparison of the cell lines and the normal osteoblasts, the top cluster from that comparison contained general terms like translational elongation and ribosome (Table 2). The terms associated with the list of 350 genes that recurrently showed two types of aberrations seem highly relevant for osteosarcoma tumourigenesis, highlighting the significance of combining different types of data to identify important molecular markers and pathways involved.

Among the 159 recurrently gained and over-expressed genes, the most frequent were eukaryotic translation elongation factor 1 alpha 2 (*EEF1A2*, 13/19 cell lines), NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 9, 22 kDa (*NDUFB9*, 13/19), ribophorin II (*RPN2*, 12/19) and cystathionine-beta-synthase (*CBS*, 12/19). Fifty-one of these 159 genes were also gained and over-expressed in >6/29 osteosarcoma clinical samples based on identical types of microarray data [27]. Among these was *NDUFB9*, gained and over-expressed in 13 of the 29 clinical samples. *NDUFB9* is located in 8q24.13 and is an accessory subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase (Complex I), whereas *CBS* is a folate-metabolising gene located in 21q22.3. So far, little is known about a possible role for these genes in cancer development. *EEF1A2* and *RPN2* are both located in 20q, and *EEF1A2* has been suggested to be an oncogene and a diagnostic marker in various cancers [33], but has to our knowledge not previously been linked to sarcomas, whereas *RPN2* has been shown to confer drug resistance in breast cancer [34].

Another frequently gained and over-expressed gene was *RUNX2* (6/19 cell lines), which is a transcription factor essential for osteoblast maturation and bone development [35]. The aberrations of *RUNX2* were validated in five of the cell lines and five osteosarcoma tumour samples using quantitative real-time PCR and RT-PCR (Figure 5). None of the tumour samples showed gain of *RUNX2*, in contrast to the cell lines, but all showed increased expression at similar levels as the cell lines showing over-expression. *RUNX2* was also shown to be recurrently gained and over-expressed in 12/29 osteosarcoma clinical samples based on identical types of microarray data [27]. *RUNX2* is frequently amplified and over-expressed in osteosarcomas, and may play an important role in osteosarcoma tumourigenesis (reviewed in [35]).

Functional enrichment analysis of the 159 recurrently gained and over-expressed genes identified terms like embryonic skeletal system development and homeodomain in the first clusters (Table S9). HOX and other homeobox genes have crucial roles in development, and a number of these were gained and over-expressed, as has been frequently reported for other cancers (reviewed in [36]). Some of the gained HOX genes were also hyper-methylated, and did not show increased expression, whereas several others were hyper-methylated, but only one recurrently under-expressed (*MSX1*) (Table S3).

The homeobox gene *DLX5* was recurrently gained and over-expressed (7/19 cell lines), and this transcription factor interacts with bone morphogenetic protein (BMP) signalling and is involved in bone and cartilage development (reviewed in [37]). The aberrations of *DLX5* were validated in five of the cell lines and five osteosarcoma tumour samples using quantitative real-time PCR and RT-PCR (Figure 5). As for *RUNX2*, none of the tumour samples showed gain of *DLX5*, but all showed increased expression at similar levels as the cell lines showing over-expression. *DLX5* was also shown to be recurrently gained and over-expressed in 7/29 osteosarcoma clinical samples based on identical types of microarray data [27], and *DLX5* was part of a gene expression prediction profile that could distinguish different histological subtypes of osteosarcoma, being down-regulated in fibroblastic osteosarcoma [23]. *DLX5* has also been shown to be differentially methylated and under-expressed in enchondromas from patients with Ollier disease, which is a non-hereditary skeletal disorder [38].

Among the nine recurrently hypo-methylated and over-expressed genes was the gene "preferentially expressed antigen in melanoma" (*PRAME*, 11/19 cell lines), which is over-expressed and a prognostic marker for clinical outcome in various types of cancers [39]. Four of the cell lines showed simultaneously gain and two additional cell lines showed gain and over-expression. *PRAME* was also over-expressed in 12/29 osteosarcoma clinical samples based on identical types of microarray data (no methylation data was available) [27], and has also recently been shown by others to be over-expressed in osteosarcomas [40,41]. Hypo-methylation of *PRAME* has been demonstrated to be responsible for the increased expression in various types of cancer [42].

Among the 158 recurrently hyper-methylated and under-expressed genes, the most frequent were "mesoderm specific transcript homolog (mouse)" (*MEST*), neuronatin (*NNAT*) and *CXCL5*, all altered in 18/19 cell lines. *MEST* and *CXCL5* were also shown to be under-expressed in 29/29 osteosarcoma clinical samples, respectively, whereas *NNAT* was under-expressed in 27/29 samples, based on identical types of microarray data (no methylation data was available) [27]. Both *MEST* and *NNAT* are imprinted genes, and *MEST* has been shown to be down-regulated in a model of human osteosarcoma, suggesting a role in tumourigenesis [43]. Consistent with the results here, loss of expression of *NNAT* has been associated with promoter hyper-methylation in pituitary adenoma [44].

The frequent hyper-methylation and under-expression of *CXCL5* were validated in five of the cell lines and five osteosarcoma tumour samples using methylation-specific PCR and quantitative real-time RT-PCR, respectively (Figure 8 and S7). Partial or full methylation of the investigated CpG island in

the promoter region was identified in all cell lines, and partial methylation was also identified in all tumour samples and the normal osteoblasts. Although methylation-specific PCR is not a quantitatively accurate method, the amount of PCR products indicated a higher degree of methylation of *CXCL5* in all cell lines and two of the tumour samples compared to the normal osteoblasts (Figure S7). Furthermore, all samples except one tumour sample showed under-expression. For *CXCL5*, previous reports are more equivocal, showing up-regulation correlated to poor survival in colorectal and pancreatic cancer [45,46], whereas another study showed correlation with under-expression of *CXCL5* and poor survival for colorectal cancer [47]. Demethylation using 5-Aza-2′-deoxycytidine showed that *CXCL5* was reactivated in all cell lines tested, with two cell lines showing more than 100-fold increased expression level (Figure 9). Tumour-specific methylation of *CXCL5* has also been observed in 80% of primary lung adenocarcinomas and 65% of lung adenocarcinoma cell lines [48], and demethylation using 5-Aza-2′-deoxycytidine also restored the expression of *CXCL5* [48]. Similar results were observed here for the other genes tested, supporting that the low expression levels of these genes are indeed caused by promoter hyper-methylation.

The significantly differentially methylated genes between the cell lines and the normal osteoblasts, which were all hyper-methylated, were enriched for terms like skeletal system development and homeodomain (Table 1), similar to the genes showing gain and over-expression (Table S9). Based on this, it seems like the methylation pattern reflects turning off a tissue-specific epigenetic program. However, the 158 genes that were both hyper-methylated and under-expressed were enriched for more general terms like signal peptide and extracellular matrix in the first clusters from the functional enrichment analysis (Table S10).

Genes with gain showed a positive association between hyper-methylation and under-expression (Figure 10), and the most common three-way combination was gain, hyper-methylation and under-expression (16 recurrently altered genes in >6/19 cell lines). This suggests that hyper-methylation of passenger genes in gained regions may be advantageous, conceivably because it counteracts the effect of over-expression of detrimental genes. Two of these 16 genes, S100 calcium binding protein A16 (*S100A16*) and maternally expressed 3 (non-protein coding) (*MEG3*), were also gained and under-expressed in 7 and 6 of 29 osteosarcoma clinical samples, respectively, based on identical types of microarray data (no methylation data was available) [27]. An integrative genomic analysis of familial breast tumours has also revealed frequent hyper-methylation of genes that showed copy number gain [49], and genes with copy number gains, low expression and high methylation levels have been identified in urothelial carcinomas by integrative analysis [50]. However, no consistent differences in the pattern of methylation and expression for the 16 recurrently altered genes were found when cell lines with gain were compared with those with normal copy number or loss (Figure S8). This suggests that methylation is not directly related to the amplification or deletion processes. In another study of osteosarcoma, gained and hyper-methylated genes showed far more over- than under-expression [20]. However, although a comparable number of genes showed recurrent gain, hyper-methylation and over-expression (12 recurrently altered genes in >6/19 cell lines), there was no significant dependency of this combination in these cell lines.

In summary, integrative analysis of genome-wide genetic and epigenetic alterations identified dependencies and relationships between DNA copy number, DNA methylation and mRNA expression in osteosarcomas. For the samples investigated, novel correlations between DNA copy number alterations and gene density were identified. The recurrently altered genes with two types of aberrations, including aberrant mRNA levels, showed in general alterations in either DNA copy number or DNA methylation, both within individual samples and across the sample panel. On the other hand, a positive association of gain with hyper-methylation and under-expression was observed, suggesting that hyper-methylation may oppose the effects of increased copy number for detrimental genes. This is especially an issue in osteosarcomas, which is highly genetically unstable, thereby suffering from many disadvantageous genomic aberrations that may be compensated for by other mechanisms. The analyses revealed a number of genes regulated by alterations in DNA copy number and DNA methylation, and additional experiments are needed to investigate their potential role in osteosarcoma development. The results show the importance of combining different types of molecular data to better comprehend the biology of osteosarcoma.

Materials and Methods

Osteosarcoma Cell Lines

Nineteen osteosarcoma cell lines collected within EuroBoNeT (http://www.eurobonet.eu) [21] were analysed. Four cell lines were established at the Norwegian Radium Hospital (HAL, KPD, MHM and OHS) and seven were established at the Istituto Ortopedico Rizzoli (IOR/OS9, IOR/OS10, IOR/OS14, IOR/OS15, IOR/OS18, IOR/MOS and IOR/SARG). The cell line ZK-58 [51] was kindly provided by Dr. Karl-Ludwig Schäfer, Düsseldorf, Germany. The cell lines 143B, HOS, MNNG/HOS, MG-63, OSA (SJSA-1), Saos-2 and U-2 OS were obtained from ATCC (http://www.lgcstandards-atcc.org). The cell lines 143B and MNNG/HOS are derived from the HOS cell line. Cell line authentication was performed by DNA profiling using short tandem repeats (STR) using Powerplex 16 (Promega, Madison, USA), and the data was validated using the profiles of the EuroBoNeT cell line bank [21] and ATCC. Data for all cell lines are given in Table S1.

The cells were grown in RPMI1640 (Lonza, Basel, Switzerland) or DMEM (Lonza) supplemented with 10% foetal calf serum (PAA Laboratories GmbH, Pasching, Austria), GlutaMAX (Life Technologies, California, USA) and penicillin/streptomycin (Lonza), at 37°C with 5% CO_2. All cells were split when reaching 80% confluency.

Osteosarcoma Tumour Samples

Five human sarcomas classified as conventional osteosarcomas were selected from a tumour collection at the Department of Tumor Biology at the Norwegian Radium Hospital. All tumors were diagnosed according to the current World Health Organization classification [1]. Tumour samples were collected immediately after surgery, cut into small pieces, frozen in liquid nitrogen and stored at −70°C until use. The clinical information was retrieved from the MEDinsight database at the Norwegian Radium Hospital. Data for all tumour samples are given in Table S1.

Normal Samples

Four normal bone samples and two osteoblast cultures were used as normal controls. Two normal bone samples were obtained from cancer patients (one with osteosarcoma and one with renal cell carcinoma) at the Norwegian Radium Hospital. The normal bone was collected as distant as possible from the tumour site, and SNP arrays confirmed that these samples had normal DNA copy

number. Two additional normal bone samples from different donors were purchased from Capital Biosciences (Maryland, USA). Two primary osteoblast cultures isolated from human calvaria of different donors were purchased from ScienCell Research Laboratories (California, USA). Data for all normal samples are given in Table S1.

The osteoblast cells were maintained in medium provided by the manufacturer, split when reaching 80% confluency, and harvested when enough cells for DNA and RNA isolation were obtained.

Ethics Statement

The information given to the patients, the written consent used, the collection of samples and the research project were approved by the ethical committee of Southern Norway (Project S-06133).

Array CGH

DNA was isolated using the Wizard Genomic DNA Purification Kit (Promega). High-resolution array CGH was performed using the Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, California, USA), containing more than 1.8 million SNPs, according to the manufacturer's protocol. Quality control was performed using the Genotyping Console v3.0.1 software (Affymetrix), applying the contrast quality control (CQC) algorithm with a minimal call rate of >86%. DNA copy number analysis was performed using the Nexus software (BioDiscovery, California, USA), with the SNPRank segmentation algorithm using default settings (threshold of 0.6 for high copy gain, 0.2 for gain, −0.2 for loss and −1.0 for homozygous loss). The categories high copy gain and gain were combined, as well as the categories loss and homozygous loss. For each cell line, tab separated text files with probe intensities, as well as copy number states for each gene, were exported for further analysis. The frequency plot of DNA copy number changes was made using Nexus. The SNP array dataset has been deposited in the Gene Expression Omnibus (GEO) data repository (www.ncbi.nlm.nih.gov/geo/, accession number GSE36003, SuperSeries number GSE36004).

DNA Methylation Profiling

DNA methylation profiling of approximately 27,000 CpG sites across the genome was performed using the Illumina HumanMethylation27 BeadChip (Illumina Inc., California, USA) according to the manufacturer's protocol. The array is used to estimate the level of methylation of the CpG sites. Data extraction and initial quality control of the bead summary raw data were performed using BeadStudio v3.1.0.0 and the Methylation module v1.9, both provided by Illumina. For each cell line and normal sample, tab separated text files with avgBeta (average ratio of signal from methylated probe relative to the sum of both methylated and unmethylated probes) values for each probe was exported for further analysis. The DNA methylation dataset has been deposited in the GEO data repository (www.ncbi.nlm.nih.gov/geo/, accession number GSE36002, SuperSeries number GSE36004).

mRNA Expression Profiling

RNA was isolated using the standard TRIzol procedure (Life Technologies), and further purified with an RNeasy mini column (QIAGEN GmbH, Hilden, Germany), according to the manufacturers' instructions. The purity and quantity of the extracted RNA were measured using the NanoDrop ND1000 spectrophotometer (Nanodrop Technologies, Delaware, USA), and the RNA integrity was evaluated using the Agilent 2100 Bioanalyzer and the RNA nano 6000 kit (Agilent Technologies Inc., California, USA).

mRNA expression profiling was performed using the Illumina HumanWG-6 v2 Expression BeadChip according to the manufacturer's protocol as previously described [52]. Data extraction and initial quality control of the bead summary raw data were performed using BeadStudio v3.1.0.0 from Illumina and the Gene Expression module v3.1.7. Variance-stabilizing transformation (vst) [53] and quantile normalisation were performed using the R package lumi, which is part of the Bioconductor project (http://www.R-project.org) [54]. The vst is almost identical to a $\log_2$ transformation, only differing at the lower end of intensities where the vst transformed values are slightly higher than the $\log_2$ transformed values. The data were annotated using the HumanWG-6_V2_R4_11223189_A annotation file from Illumina. For each cell line and normal sample, tab separated text files with vst transformed and quantile normalised intensities for each probe were exported for further analysis. The mRNA expression dataset has been deposited in the GEO data repository (www.ncbi.nlm.nih.gov/geo/, accession number GSE36001, SuperSeries number GSE36004).

Hierarchical Clustering

Unsupervised hierarchical clustering of all three data types was performed in R v.2.13.0, using the method complete linkage and Spearman correlation as distance measure. For the DNA copy number, the DNA methylation and the mRNA expression data, the probe intensities, avgBeta probe values and vst transformed and quantile normalised probe intensities, respectively, were used to calculate distances.

Identification of Alterations within Each Sample

The copy number, methylation and expression data were exported as tab separated text files from their respective native software. DNA copy number changes were identified using Nexus as previously described, assigning each gene with a copy number event (gain, normal or loss). Alterations in DNA methylation were identified by calculating the ratio between avgBeta probe values of the individual cell lines and the average of the controls (normal osteoblasts), deltaBeta. The thresholds used to define probes showing hyper-methylation and hypo-methylation were deltaBeta >0.4 and < −0.4, respectively. Alterations in mRNA expression were identified by calculating the ratio between vst transformed and quantile normalised probe intensities of the individual cell lines and the average of the controls (normal osteoblasts). The thresholds used to define probes showing over-expression and under-expression were vst ratio >1 and < −1, respectively. The probes were collapsed to gene level for the analyses, keeping the probe level information.

Six tab separated text files in total with binary scores (0 for no alteration and 1 for alteration) for the copy number (gain and loss), methylation (hyper-methylation and hypo-methylation) and expression (over-expression and under-expression) data for all genes were generated for each cell line using R scripts (available upon request). In cases where the probes for a gene showed different values and subsequently were assigned to different categories, the gene name was included in all the corresponding lists.

Comparison of Copy Number Frequency and Gene Distribution

A ".bgr" (bedgraph) file was exported from Nexus for use in genome browsers. DNA copy number gain frequency data were imported as a "marked.bed" data type into the The Genomic

Hyperbrowser (http://hyperbrowser.uio.no/hb/) [24], and analysed against UCSC known genes. Using these two tracks as input, the test "Higher values in segments" was used, testing whether gain frequency data (number of copy number gains for a given region) are higher in regions of genes than expected by chance. p-values were computed by Monte Carlo, using 1,000 MC samples. The underlying null hypothesis was that the gain value of a region, and the overlap with genes falling within the region, are uncorrelated. The test statistics used was the mean gain value inside regions covered by genes, and Monte Carlo estimates were computed by randomly permuting gain values (keeping the same segments as in the original gain track, but shuffling the gain values associated to these segments). The same analysis was performed on copy number loss frequency, except that the alternative hypothesis was that copy number values were lower instead of higher than expected.

Identification of Differentially Methylated and Expressed Genes

The Bioconductor packages Lumi, Limma and MethyLumi were used to perform t-tests between the osteosarcoma cell lines and normal osteoblasts to identify significantly differentially methylated and expressed genes, respectively. In MethyLumi, M-values ($\log_2$ ratio of methylated probe intensity and unmethylated probe intensity) were calculated and used to perform the t-tests. Separate lists with differentially methylated and expressed genes, with a Benjamini & Hochberg-corrected p-value <0.05 and absolute value of fold change >6 for the methylation data and >0.5 for the expression data, were used for functional enrichment analysis.

Functional Enrichment Analysis

The functional annotation tool of DAVID (Database for Annotation, Visualization and Integrated Discovery, developed by NIAID/NIH, http://david.abcc.ncifcrf.gov/home.jsp) [55,56] was used for functional enrichment analysis, with the DAVID default population background for *Homo sapiens* (for the gene lists from the integration analyses, the 11,843 genes from chromosome 1–22 common to the three microarray platforms were used as background). Genes were uploaded as Illumina probe IDs to avoid using official gene symbols that may be mapped ambiguously, as recommended by DAVID. For the methylation data, genes had to be uploaded using official gene symbols since DAVID does not permit Illumina methylation probe IDs for mapping. Default settings were used for the analyses.

Integration of All Data Types and Identification of Recurrently Altered Genes

The 11,843 genes from chromosome 1–22 common to the three microarray platforms were used for the integration of the data. The six text files with binary scores were combined in order to identify genes with alterations in two types of data and to create contingency tables for each cell line using R scripts (available upon request). With two types of changes for each of the three data sets (gain and loss, hyper- and hypo-methylation, over- and under-expression), 12 two-way combinations were possible, whereas 8 three-way combinations were possible. A recurrence threshold of 6/19 cell lines (>30%) was used to identify recurrently altered genes with two types of aberrations, considering the combinations gain/over-expression, hypo-methylation/over-expression, loss/under-expression and hyper-methylation/under-expression.

Data Type Dependencies

The contingency tables were used to evaluate data dependencies within each sample by calculating the odds ratio for the different two-way combinations of data, as well as the three-way combinations conditioning on the copy number state. The Bonferroni-corrected chi-square p-values of the combinations were also determined.

Quantitative Real-time PCR and RT-PCR

Quantitative real-time PCR was performed using the 7900HT Fast Real-Time PCR System (Life Technologies). The copy numbers of the genes distal-less homeobox 5 (*DLX5*) and runt-related transcription factor 2 (*RUNX2*) were determined using TaqMan Copy Number Assays (assay ID Hs01209848_cn and Hs00753612_cn, respectively). The genes eukaryotic translation elongation factor 1 gamma (*EEF1G*) and F-box protein 11 (*FBXO11*) (assay ID Hs03771595_cn and Hs02528370_cn, respectively) were used as endogenous controls for normalisation. These two genes are located in 11q12.3 and 2p16.3, respectively, and showed low level of DNA copy number changes in a large panel of osteosarcoma samples ([27] and Kresse et al, unpublished). The copy number levels were determined using the CopyCaller Software v2.0 program (Life Technologies) as described by the manufacturer, and the average copy number of *EEF1G* and *FBXO11* was used for normalisation.

The High Capacity RNA-to-cDNA Master Mix (Life Technologies) was used for cDNA synthesis, and quantitative real-time reverse-transcription PCR (qRT-PCR) was performed using the 7900HT Fast Real-Time PCR System (Life Technologies). The expression levels of the genes distal-less homeobox 5 (*DLX5*), runt-related transcription factor 2 (*RUNX2*), chemokine (C-X-C motif) ligand 5 (*CXCL5*), A kinase (PRKA) anchor protein 12 (*AKAP12*), EGF containing fibulin-like extracellular matrix protein 1 (*EFEMP1*) and interleukin 11 receptor, alpha (*IL11RA*) were determined using TaqMan Gene Expression Assays (assay ID Hs00193291_m1, Hs01047976_m1, Hs00171085_m1, Hs00374507_m1, Hs00244575_m1 and Hs00234415_m1, respectively). The gene glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*, assay ID Hs99999905_m1) was used as an endogenous control for normalisation. The relative expression levels were determined using the comparative C_T method as described by the manufacturer.

Methylation-specific PCR

Genomic DNA was sodium bisulphite-treated as previously described [57]. Twenty-two ng of converted DNA was used to assess the methylation status of the *CXCL5* promoter by methylation-specific PCR. The primers were designed using Methyl Primer Express v1.0 (Life Technologies). Each bisulphite-treated DNA was amplified in a 50 µl reaction volume using the following primer sets; CXCL5_M_F (5′-TTAGGAATTCGC-GATCGTTC-3′) and CXCL5_M_R (5′-CACCGCTAACGA-TAAACCCT-3′), as well as CXCL5_U_F (5′-AGTTTAG-GAATTTGTGATTGTTT-3′) and CXCL5_U_R (5′-CACCACTAACAATAAACCCTAAC-3′). The forward primer covers four CpGs and overlaps with one of the two probes for *CXCL5* on the Illumina HumanMethylation27 BeadChip (probe ID cg10088985). The primers cover a total of six CpGs located in the first exon of *CXCL5*.

PCR was carried out using the EpiTect MSP kit (QIAGEN) with the following PCR conditions; 95°C for 10 min, followed by 94°C for 15 s, 53°C for 15 s and 72°C for 30s, for 40 cycles and a final extension step at 72°C for 10 min. The PCR products were separated by electrophoresis using a 2% agarose gel. Each sample

was scored for the presence of PCR product for the methylated (M_F/M_R) and unmethylated (U_F/U_R) primer sets. The EpiTect PCR Control DNA Set (QIAGEN) was used for optimization and to confirm the specificity of the primers. DNase- and RNase-free water was used as a negative control.

Demethylation Treatment

Twelve of the osteosarcoma cell lines were seeded at a density of 5,000–7,500 cells/cm^2. The following day, the medium was replaced with a medium containing 1 μM of 5-Aza-2-deoxycytidine (Sigma Aldrich, Montana, USA), which again was replenished every 24 hours. After three days, total RNA was isolated using the QIAGEN miRNeasy Mini Kit (QIAGEN) according to the manufacturer's protocol, prior to treatment with amplification-grade DNase I (Life Technologies) to avoid amplification of contaminating genomic DNA.

Supporting Information

Figure S1 Plots of DNA copy number, DNA methylation and mRNA expression levels for DNA methyltransferase genes.
(PDF)

Figure S2 Genome-wide frequency plot of DNA copy number.
(PDF)

Figure S3 Frequency plot of copy number aberrations and gene density for chromosome arms 2q, 8p, 19p and 19q.
(PDF)

Figure S4 Frequency plot of copy number aberrations and gene density for chromosome arms not significant for gain.
(PDF)

Figure S5 Plots of the number of common genes for individual aberration types and three-way combinations at different sample recurrence thresholds.
(PDF)

Figure S6 Hierarchical clustering of osteosarcoma cell lines based on expression level of 350 genes that recurrently showed two types of aberrations.
(PDF)

Figure S7 Gel pictures of PCR products from methylation-specific PCR.
(PDF)

Figure S8 Plots of DNA copy number, DNA methylation and mRNA expression levels for 16 recurrent genes with gain, hyper-methylation and under-expression.
(PDF)

Table S1 Clinical data for osteosarcoma cell lines, osteosarcoma tumour samples and normal samples.
(PDF)

Table S2 Number of genes with each type of aberration (DNA copy number, DNA methylation and mRNA expression).
(PDF)

Table S3 List of 328 significantly differentially methylated genes.
(XLSX)

Table S4 Top 10 clusters from enrichment analysis of differentially methylated genes in DAVID.
(XLSX)

Table S5 List of 283 significantly differentially expressed genes.
(XLSX)

Table S6 Top 10 clusters from enrichment analysis of differentially expressed genes in DAVID.
(XLSX)

Table S7 List of 350 genes that recurrently showed two types of aberrations.
(XLSX)

Table S8 Top 10 clusters from enrichment analysis of 350 genes that recurrently showed two types of aberrations in DAVID.
(XLSX)

Table S9 Top 10 clusters from enrichment analysis of 159 genes with gain and over-expression in DAVID.
(XLSX)

Table S10 Top 10 clusters from enrichment analysis of 158 genes with hyper-methylation and under-expression in DAVID.
(XLSX)

Acknowledgments

We thank Russell Castro for technical assistance with cell culturing and the personnel at the Microarray Core Facility at The Norwegian Radium Hospital and Marcel Winter and Ronald Duim at the Leiden University Medical Center for technical assistance with microarray experiments.

Author Contributions

Conceived and designed the experiments: SHK OM LAMZ. Performed the experiments: SHK M. Skårn AHBP. Analyzed the data: SHK HR HMN KL EH OM LAMZ. Contributed reagents/materials/analysis tools: M. Serra. Wrote the paper: SHK HR OM LAMZ. Provided mRNA expression data: AMCJ PCWH. Organised the network that facilitated this study: PCWH.

References

1. Raymond AK, Ayala AG, Knuutila S (2002) Conventional osteosarcoma. In: Fletcher CDM, Unni KK, Mertens F, editors. World Health Organization Classification of Tumours Pathology and Genetics of Tumours of Soft Tissue and Bone. Lyon: IARC Press.
2. Aksnes LH, Hall KS, Folleraas G, Stenwig AE, Bjerkehagen B, et al. (2006) Management of high-grade bone sarcomas over two decades: the Norwegian Radium Hospital experience. Acta Oncol 45: 38–46.
3. PosthumaDeBoer J, Witlox MA, Kaspers GJ, van Royen BJ (2011) Molecular alterations as target for therapy in metastatic osteosarcoma: a review of literature. Clin Exp Metastasis 28: 493–503.
4. Hattinger CM, Pasello M, Ferrari S, Picci P, Serra M (2010) Emerging drugs for high-grade osteosarcoma. Expert Opin Emerg Drugs 15: 615–634.
5. Lewis IJ, Nooij MA, Whelan J, Sydes MR, Grimer R, et al. (2007) Improvement in histologic response but not survival in osteosarcoma patients treated with intensified chemotherapy: a randomized phase III trial of the European Osteosarcoma Intergroup. J Natl Cancer Inst 99: 112–128.
6. Bacci G, Briccoli A, Rocca M, Ferrari S, Donati D, et al. (2003) Neoadjuvant chemotherapy for osteosarcoma of the extremities with metastases at presentation: recent experience at the Rizzoli Institute in 57 patients treated with cisplatin, doxorubicin, and a high dose of methotrexate and ifosfamide. Ann Oncol 14: 1126–1134.
7. Kager L, Zoubek A, Potschger U, Kastner U, Flege S, et al. (2003) Primary metastatic osteosarcoma: presentation and outcome of patients treated on

neoadjuvant Cooperative Osteosarcoma Study Group protocols. J Clin Oncol 21: 2011–2018.
8. Kresse SH, Szuhai K, Barragan-Polania AH, Rydbeck H, Cleton-Jansen AM, et al. (2010) Evaluation of high-resolution microarray platforms for genomic profiling of bone tumours. BMC Res Notes 3: 223.
9. Smida J, Baumhoer D, Rosemann M, Walch A, Bielack S, et al. (2010) Genomic alterations and allelic imbalances are strong prognostic predictors in osteosarcoma. Clin Cancer Res 16: 4256–4267.
10. Yen CC, Chen WM, Chen TH, Chen WY, Chen PC, et al. (2009) Identification of chromosomal aberrations associated with disease progression and a novel 3q13.31 deletion involving LSAMP gene in osteosarcoma. Int J Oncol 35: 775–788.
11. Kresse SH, Ohnstad HO, Paulsen EB, Bjerkehagen B, Szuhai K, et al. (2009) LSAMP, a novel candidate tumor suppressor gene in human osteosarcomas, identified by array comparative genomic hybridization. Genes Chromosomes Cancer 48: 679–693.
12. Man TK, Lu XY, Jaeweon K, Perlaky L, Harris CP, et al. (2004) Genome-wide array comparative genomic hybridization analysis reveals distinct amplifications in osteosarcoma. BMC Cancer 4: 45.
13. Lockwood WW, Stack D, Morris T, Grehan D, O'Keane C, et al. (2011) Cyclin E1 is amplified and overexpressed in osteosarcoma. J Mol Diagn 13: 289–296.
14. Lu XY, Lu Y, Zhao YJ, Jaeweon K, Kang J, et al. (2008) Cell cycle regulator gene CDC5L, a potential target for 6p12-p21 amplicon in osteosarcoma. Mol Cancer Res 6: 937–946.
15. Zhang J, Benavente CA, McEvoy J, Flores-Otero J, Ding L, et al. (2012) A novel retinoblastoma therapy from genomic and epigenetic analyses. Nature 481: 329–334.
16. Chari R, Coe BP, Vucic EA, Lockwood WW, Lam WL (2010) An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. BMC Syst Biol 4.
17. Hogan LE, Meyer JA, Yang J, Wang J, Wong N, et al. (2011) Integrated genomic analysis of relapsed childhood acute lymphoblastic leukemia reveals therapeutic strategies. Blood 118: 5218–5226.
18. Kristensen VN, Vaske CJ, Ursini-Siegel J, Van Loo P, Nordgard SH, et al. (2011) Integrated molecular profiles of invasive breast tumors and ductal carcinoma in situ (DCIS) reveal differential vascular and interleukin signaling. Proc Natl Acad Sci U S A 109: 2802–2807.
19. Sadikovic B, Yoshimoto M, Al-Romaih K, Maire G, Zielenska M, et al. (2008) In vitro analysis of integrated global high-resolution DNA methylation profiling with genomic imbalance and gene expression in osteosarcoma. PloS ONE 3: e2834.
20. Sadikovic B, Yoshimoto M, Chilton-MacNeill S, Thorner P, Squire JA, et al. (2009) Identification of interactive networks of gene expression associated with osteosarcoma oncogenesis by integrated molecular profiling. Hum Mol Genet 18: 1962–1975.
21. Ottaviano L, Schaefer KL, Gajewski M, Huckenbeck W, Baldus S, et al. (2010) Molecular characterization of commonly used cell lines for bone tumor research: a trans-European EuroBoNet effort. Genes Chromosomes Cancer 49: 40–51.
22. Mohseny AB, Machado I, Cai Y, Schaefer KL, Serra M, et al. (2011) Functional characterization of osteosarcoma cell lines provides representative models to study the human disease. Lab Invest 91: 1195–1205.
23. Kuijjer ML, Namlos HM, Hauben EI, Machado I, Kresse SH, et al. (2011) mRNA expression profiles of primary high-grade central osteosarcoma are preserved in cell lines and xenografts. BMC Med Genomics 4: 66.
24. Sandve GK, Gundersen S, Rydbeck H, Glad IK, Holden L, et al. (2010) The Genomic HyperBrowser: inferential genomics at the sequence level. Genome Biol 11: R121.
25. Lau CC, Harris CP, Lu XY, Perlaky L, Gogineni S, et al. (2004) Frequent amplification and rearrangement of chromosomal bands 6p12–p21 and 17p11.2 in osteosarcoma. Genes Chromosomes Cancer 39: 11–21.
26. Ozaki T, Schaefer KL, Wai D, Buerger H, Flege S, et al. (2002) Genetic imbalances revealed by comparative genomic hybridization in osteosarcomas. Int J Cancer 102: 355–365.
27. Kuijjer ML, Rydbeck H, Kresse SH, Buddingh EP, Lid AB, et al. (2012) Identification of osteosarcoma driver genes by integrative analysis of copy number and gene expression data. Genes Chromosomes Cancer 51: 696–706.
28. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.
29. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, et al. (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144: 27–40.
30. Rao-Bindal K, Kleinerman ES (2011) Epigenetic regulation of apoptosis and cell cycle in osteosarcoma. Sarcoma 2011: 679457.
31. Cui J, Wang W, Li Z, Zhang Z, Wu B, et al. (2011) Epigenetic changes in osteosarcoma. Bull Cancer 98: E62–68.
32. Baylin SB, Jones PA (2011) A decade of exploring the cancer epigenome - biological and translational implications. Nat Rev Cancer 11: 726–734.
33. Lee MH, Surh YJ (2009) eEF1A2 as a putative oncogene. Ann N Y Acad Sci 1171: 87–93.
34. Honma K, Iwao-Koizumi K, Takeshita F, Yamamoto Y, Yoshida T, et al. (2008) RPN2 gene confers docetaxel resistance in breast cancer. Nat Med 14: 939–948.
35. Martin JW, Zielenska M, Stein GS, van Wijnen AJ, Squire JA (2011) The Role of RUNX2 in Osteosarcoma Oncogenesis. Sarcoma 2011: 282745.
36. Shah N, Sukumar S (2010) The Hox genes and their roles in oncogenesis. Nat Rev Cancer 10: 361–371.
37. Nishimura R, Hata K, Matsubara T, Wakabayashi M, Yoneda T (2012) Regulation of bone and cartilage development by network between BMP signalling and transcription factors. J Biochem 151: 247–254.
38. Pansuriya TC, van Eijk R, d'Adamo P, van Ruler MA, Kuijjer ML, et al. (2011) Somatic mosaic IDH1 and IDH2 mutations are associated with enchondroma and spindle cell hemangioma in Ollier disease and Maffucci syndrome. Nat Genet 43: 1256–1261.
39. Epping MT, Bernards R (2006) A causal role for the human tumor antigen preferentially expressed antigen of melanoma in cancer. Cancer Res 66: 10639–10642.
40. Zou C, Shen J, Tang Q, Yang Z, Yin J, et al. (2011) Cancer-testis antigens expressed in osteosarcoma identified by gene microarray correlate with a poor patient prognosis. Cancer 118: 1845–1855.
41. Toledo SR, Zago MA, Oliveira ID, Proto-Siqueira R, Okamoto OK, et al. (2011) Insights on PRAME and osteosarcoma by means of gene expression profiling. J Orthop Sci 16: 458–466.
42. Schenk T, Stengel S, Goellner S, Steinbach D, Saluz HP (2007) Hypomethylation of PRAME is responsible for its aberrant overexpression in human malignancies. Genes Chromosomes Cancer 46: 796–804.
43. Li Y, Meng G, Guo QN (2008) Changes in genomic imprinting and gene expression associated with transformation in a model of human osteosarcoma. Exp Mol Pathol 84: 234–239.
44. Revill K, Dudley KJ, Clayton RN, McNicol AM, Farrell WE (2009) Loss of neuronatin expression is associated with promoter hypermethylation in pituitary adenoma. Endocr Relat Cancer 16: 537–548.
45. Li A, King J, Moro A, Sugi MD, Dawson DW, et al. (2011) Overexpression of CXCL5 is associated with poor survival in patients with pancreatic cancer. Am J Pathol 178: 1340–1349.
46. Kawamura M, Toiyama Y, Tanaka K, Saigusa S, Okugawa Y, et al. (2011) CXCL5, a promoter of cell proliferation, migration and invasion, is a novel serum prognostic marker in patients with colorectal cancer. Eur J Cancer: In press.
47. Speetjens FM, Kuppen PJ, Sandel MH, Menon AG, Burg D, et al. (2008) Disrupted expression of CXCL5 in colorectal cancer is associated with rapid tumor formation in rats and poor prognosis in patients. Clin Cancer Res 14: 2276–2284.
48. Tessema M, Klinge DM, Yingling CM, Do K, Van Neste L, et al. (2010) Re-expression of CXCL14, a common target for epigenetic silencing in lung cancer, induces tumor necrosis. Oncogene 29: 5159–5170.
49. Flanagan JM, Cocciardi S, Waddell N, Johnstone CN, Marsh A, et al. (2010) DNA methylome of familial breast cancer identifies distinct profiles defined by mutation status. Am J Hum Genet 86: 420–433.
50. Lauss M, Aine M, Sjodahl G, Veerla S, Patschan O, et al. (2012) DNA methylation analyses of urothelial carcinoma reveal distinct epigenetic subtypes and an association between gene copy number and methylation status. Epigenetics 7.
51. Schulz A, Battmann A, Heinrichs CM, Kern A, Tiedemann A, et al. (1993) Properties and reactivity of a new human osteosarcoma cell line (HOS 58). Calcif Tissue Int 52: 30.
52. Buddingh EP, Kuijjer ML, Duim RA, Burger H, Agelopoulos K, et al. (2011) Tumor-infiltrating macrophages are associated with metastasis suppression in high-grade osteosarcoma: a rationale for treatment with macrophage activating agents. Clin Cancer Res 17: 2110–2119.
53. Lin SM, Du P, Huber W, Kibbe WA (2008) Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic Acids Res 36: e11.
54. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
55. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4: P3.
56. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.
57. Wu Q, Lothe RA, Ahlquist T, Silins I, Trope CG, et al. (2007) DNA methylation profiling of ovarian carcinomas and their in vitro models identifies HOXA9, HOXB5, SCGB3A1, and CRABP1 as novel targets. Mol Cancer 6: 45.

Paper III

Sandve et al. Genome Biology 2010, 11:R121
http://genomebiology.com/2010/11/12/R121

Genome Biology

SOFTWARE Open Access

The Genomic HyperBrowser: inferential genomics at the sequence level

Geir K Sandve[1], Sveinung Gundersen[2], Halfdan Rydbeck[1,3,5], Ingrid K Glad[4], Lars Holden[3], Marit Holden[3], Knut Liestøl[1,5], Trevor Clancy[2], Egil Ferkingstad[3], Morten Johansen[6], Vegard Nygaard[6], Eivind Tøstesen[6], Arnoldo Frigessi[3,7], Eivind Hovig[1,2,3,6*]

Abstract

The immense increase in the generation of genomic scale data poses an unmet analytical challenge, due to a lack of established methodology with the required flexibility and power. We propose a first principled approach to statistical analysis of sequence-level genomic information. We provide a growing collection of generic biological investigations that query pairwise relations between tracks, represented as mathematical objects, along the genome. The Genomic HyperBrowser implements the approach and is available at http://hyperbrowser.uio.no.

Rationale

The combination of high-throughput molecular techniques and deep DNA sequencing is now generating detailed genome-wide information at an unprecedented scale. As complete human genomic information at the detail of the ENCODE project [1] is being made available for the full genome, it is becoming possible to query relations between many organizational and informational elements embedded in the DNA code. These elements can often best be understood as acting in concert in a complex genomic setting, and research into functional information typically involves integrational aspects. The knowledge that may be derived from such analyses is, however, presently only harvested to a small degree. As is typical in the early phase of a new field, research is performed using a multitude of techniques and assumptions, without adhering to any established principled approaches. This makes it more difficult to compare, reproduce and realize the full implications of the various findings.

The available toolbox for generic genome scale annotation comparison is presently relatively small. Among the more prominent tools are those embedded within the genome browsers, or associated with them, such as Galaxy [2], BioMart [3], EpiGRAPH [4] and UCSC Cancer Genomics Browser [5]. BioMart at this point mostly offers flexible export of user-defined tracks and regions. Galaxy provides a richer, text-centric suite of operations. EpiGraph presents a solid set of statistical routines focused on analysis of user-defined case-control regions. The recently introduced UCSC Cancer Genomics Browser visualizes clinical omics data, as well as providing patient-centric statistical analyses.

We have developed novel statistical methodology and a robust software system for comparative analysis of sequence-level genomic data, enabling integrative systems biology, at the intersection of genomics, computational science and statistics. We focus on inferential investigations, where two genomic annotations, or tracks, are compared in order to find significant deviation from null-model behavior. Tracks may be defined by the researcher or extracted from the sizable library provided with the system. The system is open-ended, facilitating extensions by the user community.

Results

Overview

Our system is based on an abstract representation of generic genomic elements as mathematical objects. Hypotheses of interest are translated into mathematical relations. Concepts of randomization and track structure preservation are used to build complex problem-specific null models of the relation between two tracks. Formal inference is performed at a global or local scale, taking confounder tracks into account when necessary (Figure 1).

* Correspondence: ehovig@ifi.uio.no
[1]Department of Informatics, University of Oslo, Blindern, 0316 Oslo, Norway
Full list of author information is available at the end of the article

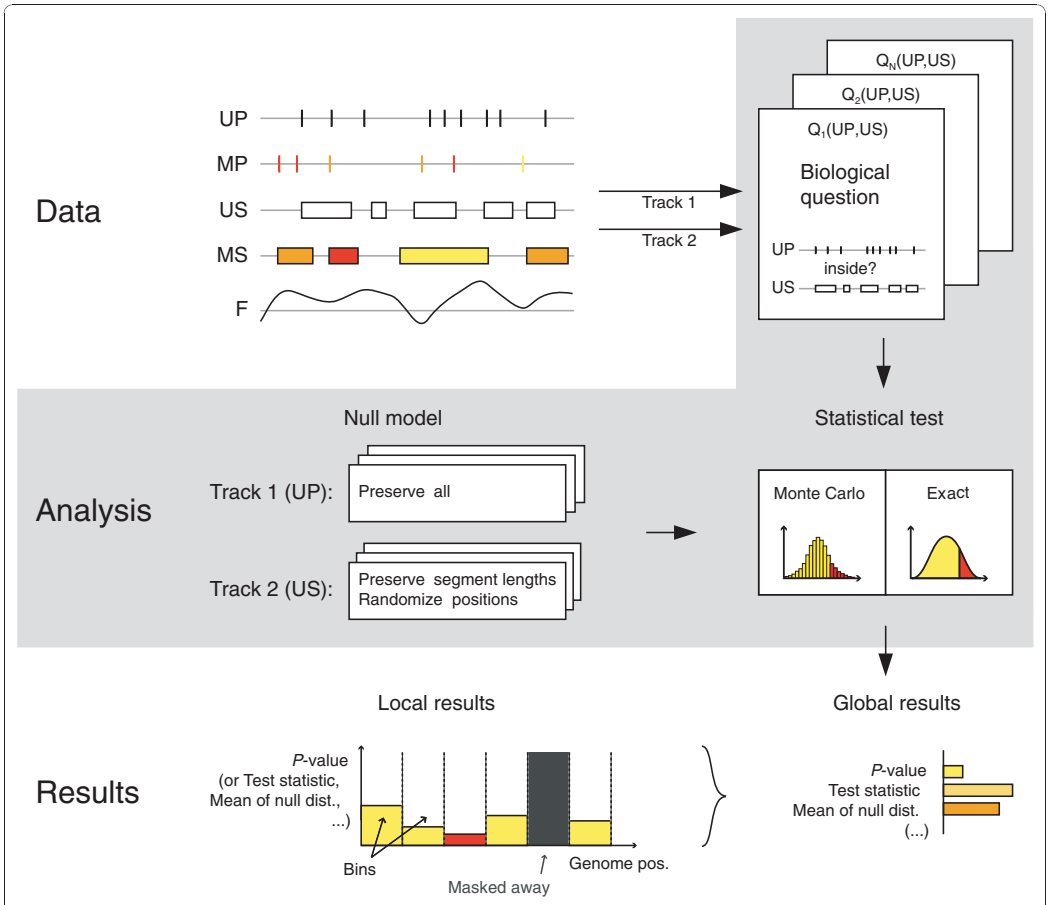


Figure 1 Flow diagram of the mathematics of genomic tracks. Genomic tracks are represented as geometric objects on the line defined by the base pairs of the genome sequence: (unmarked (UP) or marked (MP)) points, (unmarked (US) or marked (MS)) segments, and functions (F). The biologist identifies the two tracks to be compared, and the Genomic HyperBrowser detects their type. The biological question of interest is stated in terms of mathematical relations between the types of the two tracks. The relevant questions are proposed by the system. The biologist then selects the question and needs to specify the null hypothesis. For this purpose she is called to decide about what structures are preserved in each track, and how to randomize the rest. Thereafter, the Genomic HyperBrowser identifies the relevant test statistics, and computes actual *P*-values, either exactly or by Monte Carlo testing. Results are then reported, both for a global analysis, answering the question on the whole genome (or area of study), and for a local analysis. Here, the area is divided into bins, and the answer is given per bin. *P*-values, test-statistic, and effect sizes are reported, as tables and graphics. Significance is reported when found, after correction for multiple testing.

Abstract representation of genomic elements

A genome annotation track is a collection of objects of a specific genomic feature, such as genes, with base-pair-specific locations from the start of chromosome 1 to the end of chromosome Y. Tracks vary in biological content, but also in the form of the information they contain. A track representing genes contains positional information that can be reduced to 'segments' (intervals of base pairs) along the genome. A track of SNPs can be reduced to points (single base pairs) on the genome. The expression values of a gene, or the alleles of a SNP, are non-positional information parts and are attributed as 'marks' (numerical or categorical) to the corresponding positional objects, that is, segments or points. Finally, a track of DNA melting assigns a temperature to each base pair, describing a 'function' on the genome. We thus define five genomic types: unmarked points (UP), marked points (MP), unmarked segments (US),

marked segments (MS) and functions (F). These five types completely represent every one-dimensional geometry with marks.

Catalogue of investigations

We translate biological hypotheses of interest into a study of mathematical relations between genomic tracks, leading to a large collection of possible generic investigations.

Consider the relation between histone modifications and gene expression, as investigated by visual inspection in [6] (Figure S1 in Additional file 1). The question is whether the number of nucleosomes with a given histone modification (represented as type UP), counted in a region around the transcription start site (TSS) of a gene, correlates with the expression of the gene. The second track is represented as marked segments (MS). This study of histone modifications and gene expressions can then be phrased as a generic investigation between a pair of tracks (T1, T2) of type UP and MS: are the number of T1 points inside T2 segments correlated with T2 marks? Figure 2 shows the results when repeating this analysis for all histone modifications studied in [6], and different regions around the TSS. See Section 1 in Additional file 1 for a more detailed example investigation, analyzing the genome coverage by different gene definitions.

In the context of the catalogue of investigations, the genomic types are minimal models of information content. In the above example, nucleosome modifications are only used for counting, and thus considered unmarked points (UP), even though they are typically represented in the file system as marked or unmarked segments. As the gene-related properties of interest are the genome segments in which the nucleosomes are counted, as well as the corresponding gene expression values (marks), T2 is of the type marked segments (MS). The choice of genomic type clarifies the content of a track, and also restricts which analyses are appropriate. Investigations regarding the length of the elements of a track are, for instance, relevant for genes, but not for SNPs and DNA melting temperatures.

The five genomic types lead to 15 unordered pairs (T1, T2) of track type combinations, with each combination defining a specific set of relevant analyses. For instance, the UP-US combination defines several investigations of potential interest: are the T1 points falling inside the T2 segments more than expected by chance? Do the points accumulate more at the borders of the segments, instead of being spread evenly within? Do the points fall closer to the segments than expected? A growing collection of abstract mathematical versions of biological questions is provided. We have currently implemented 13 different analyses, filling 8 of the 15 possible combinations of track types (see Additional file 2 for mathematical details). Note that information reduction of a track to a simpler type (for example, segments to points) may open up additional analytical opportunities, and are handled dynamically by the system - for example, by treating segments as their middle points.

Global and local inference

A global analysis investigates if a certain relation between two tracks is found in a domain as a whole. A local analysis is based on partitioning the domain into smaller units, called bins, and performing the analysis in each unit separately. Local analysis can be used to investigate if and where two tracks display significant concordant or discordant behavior, and thus be used to generate hypotheses on the existence of biological mechanisms explaining such perturbations. Local investigations may also be used to examine global results in more detail. The length of each bin defines the scale of the analysis. Inference is then based on the computation of *P*-values, locally in each bin, or globally, under the null model.

To illustrate the value of local analysis, we consider viral integration events in the human genome. These may result in disease and may also be a consequence of retroviral gene therapy. Derse *et al.* [7] examined integration for six types of retroviruses, with different viral integrases, thus having different integration sites (type UP). Using these data, we asked whether there are hotspots of integration inside 2-kb flanking regions of predicted promoters (type US), that is, whether and where the points are falling inside the segments more than expected by chance. Figure 3 displays the hotspots as calculated *P*-values in bins across the genome, using the subset of murine leukemia virus (MLV) sites. We find locations of increased integration, thus generating hypotheses on the role of integration site sequences and their context.

Local analysis may be used to avoid drawing incorrect conclusions from global investigations. Consider the repressive histone modification H3K27me3 as studied in [8]. Data from ChIP-chip experiments on mouse chromosome 17 were analyzed, finding that H3K27me3 falls in domains that are enriched in short interspersed nuclear element (SINE) and depleted in long interspersed nuclear element (LINE) repeats. Using the line of enquiry raised in [8], we asked whether H3K27me3 regions (type US) significantly overlap with SINE repeats (type US), but here using formal statistical testing at the base pair level. The chosen null model only allows local rearrangements of genomic elements (for more detail, see next section). This preserves local biological structure, but allows for some controlled level of randomness.

Performing this test globally on the whole chromosome 17 leads to rejection of the null hypothesis ($P = 10^{-4}$),

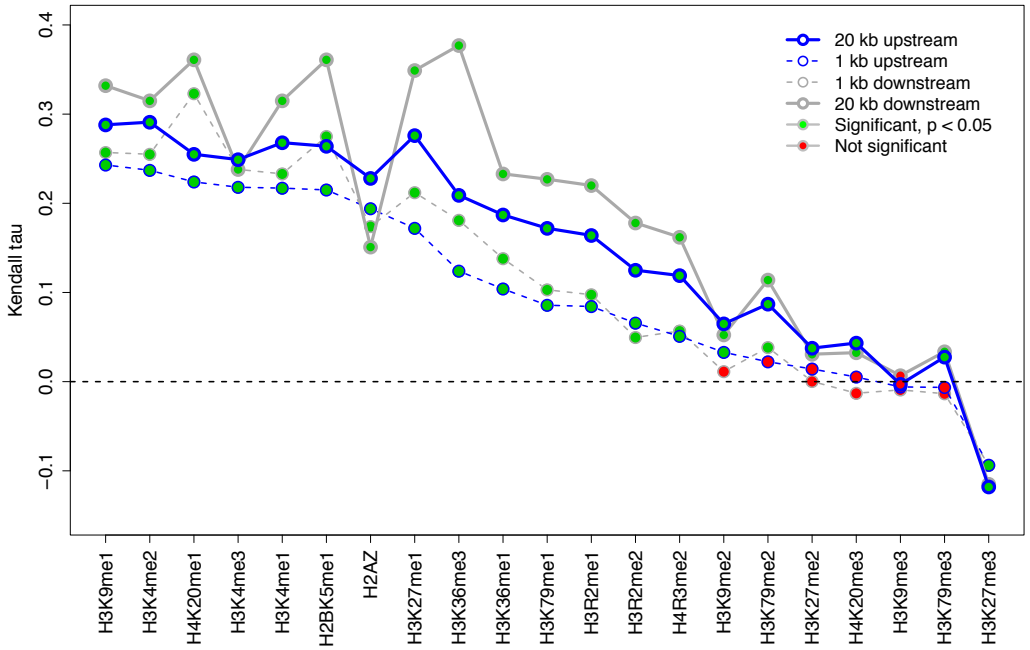


Figure 2 Gene regulation by histone modifications. The correlation between occupancy of 21 different histone modifications and gene expression within 4 different regions around the TSS (up- and downstream, 1 and 20 kb), sorted by correlation in 1-kb upstream regions. Sixteen of 21 histone modifications show significant correlation in 1-kb upstream regions, while inspection of the actual value of Kendall's tau (Table S1 in Additional file 1) shows very little effect size for 6 of these 16 (<0.1).

in line with [8]. However, a local analysis leads to a deeper understanding. At a 5-Mbp scale, no significant findings were obtained in any of the 19 bins (10% false discovery rate (FDR)-corrected). The frequency of H3K27me3 segments varies considerably along chromosome 17 (Figure S2 in Additional file 1), which may cause the observed discrepancy between local and global results.

Precise specification of null models

A crucial aspect of an investigation is the precise formalization of the null model, which should reflect the combination of stochastic and selective events that constitutes the evolution behind the observed genomic feature.

Consider again the example of H3K27me3 versus repeating elements. In the chosen null model, we preserved the repeat segments exactly, but permuted the positions of the H3K27me3 segments, while preserving segment and intersegment lengths. We then computed the total overlap between the segments, and used a Monte Carlo test to quantify the departure from the null model. The effect of using alternative null models is shown in Table 1. The null model examined in the first column, which does not preserve the dependency between neighboring base pairs, produces lower *P*-values. Unrealistically simple null models may thus lead to false positives. In fact, two simulated independent tracks may appear to have a significant association if their individual characteristics are not appropriately modeled (Section 2 in Additional file 1). In this example, the choice between the biologically more reasonable null models is difficult. The two other columns of Table 1 include models that preserve more of the biological structure. The fact that these models do not lead to clear rejection of the null hypotheses suggests that we in this case lack strong evidence against the null hypothesis. Thus, examining the results obtained for a set of different null models may often contribute important information. The null model should reflect biological realism, but also allow sufficient variation to permit the construction of tests. A set of simulated synthetic tracks is provided as an aid for assessing appropriate null models (Additional file 3).

The Genomic HyperBrowser allows the user to define an appropriate null model by specifying (a) a preservation

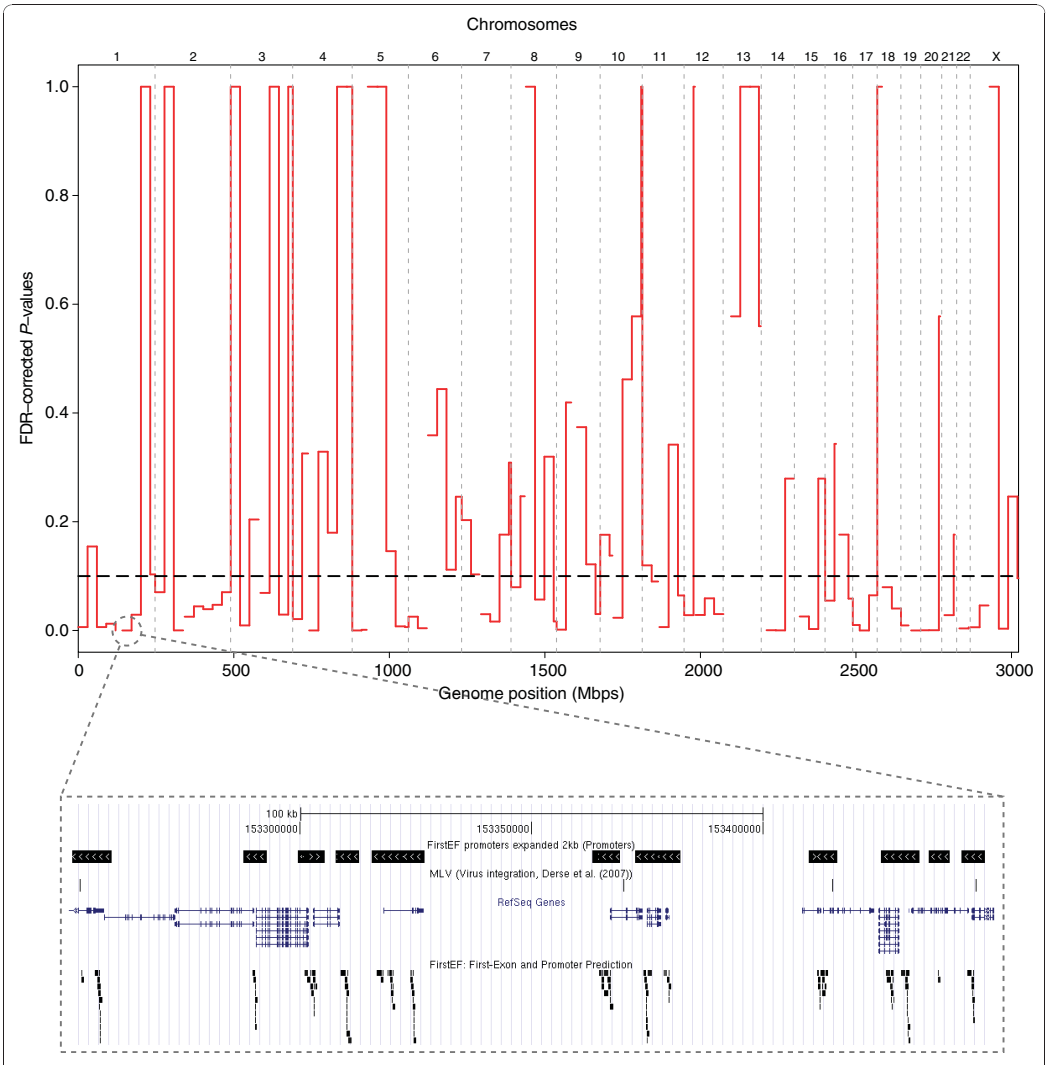


Figure 3 Viral integration sites. Plot of false discovery rate (FDR)-adjusted *P*-values along the genome, in 30-Mbp bins. Small *P*-values indicate regions where murine leukemia virus (MLV) integrates inside 2-kb regions around FirstEF promoters more frequently than by chance. The FDR cutoff at 10% is shown as a dashed line. The inset of a local area (chromosome 1:153,250,001-153,450,000) indicates FirstEF promoters expanded by 2 kb in both directions, MLV integration sites, RefSeq genes, and unflanked FirstEF sites.

rule for each track, and (b) a stochastic process, describing how the non-preserved elements should be randomized. Preservation fixes elements or characteristics of a track as present in the data. For each genomic type, we have developed a hierarchy of less and less strict preservation rules, starting from preserving the entire track exactly (Section 3 in Additional file 1). For example, these preservation options for unmarked segments can be assumed: (i) preserve all, as in data; (ii) preserve segments and intervals between segments, in number and length, but not their ordering; (iii) preserve only the segments, in number and length, but not their position; (iv) preserve only the number of base pairs in segments, not segment position or number. Depending

Table 1 Significant bins of the overlap test between H3K27me3 segments and SINE repeats under various null models

Tracks to randomize	Preserve total number of base pairs covered	Preserve segment lengths, but randomize position	Preserve segment and intersegment lengths, but randomize positions
H3K27me3	10/19	1/19	0/19
SINE	10/19	5/19	4/19
H3K27me3 and SINE	10/19	5/19	4/19

The number of significant bins of the overlap test between H3K27me3 segments and SINE repeats under different preservation and randomization rules for the null model. The test was performed in 19 bins on mouse chromosome 17, with the MEFB1 cell line. (Use of the MEFF cell line gave similar results; Table S2 in Additional file 1). In this case, less preservation of biological structure leads to smaller *P*-values. Also, randomizing the SINE track gave smaller *P*-values than randomizing the H3K27me3 track (or both).

on the test statistic T, the level of preservation and the chosen randomization, *P*-values are computed exactly, asymptotically or by standard or sequential Monte Carlo [9,10].

Confounder tracks

The relation between two tracks of interest may often be modulated by a third track. Such a third track may act as a confounder, leading, if ignored, to dubious conclusions on the relation between the two tracks of interest.

Consider the relation of coding regions to the melting stability of the DNA double helix. Melting forks have been found to coincide with exon boundaries [11-15]. Although few studies have reported statistical measures of such correlation [11], the correlation is confirmed by a straightforward investigation. Tracks (type F) representing the probabilities of melting fork locations [16] in *Saccharomyces cerevisiae*, were compared to tracks containing all exon boundaries (Figure 4). We asked if the melting fork probabilities (P) were higher than expected at the exon boundaries (E) than elsewhere. In the null model, the function was conserved, while points were uniformly randomized in each chromosome. Monte Carlo testing was carried out on the chromosomes separately, giving *P*-values <0.0005 (Table S3 in Additional file 1). In the absence of a confounder, it is thus tempting to conclude that there is an interesting relation between DNA melting and coding regions, for which functional implications have been previously discussed [15,17,18].

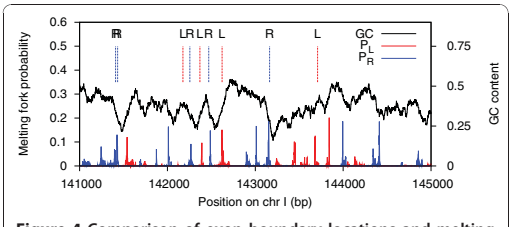


Figure 4 Comparison of exon boundary locations and melting fork probability peaks. Independent analyses were carried out on left and right exon boundaries as compared to left- and right-facing melting forks, respectively. In the upper part, dashed vertical lines indicate left (L, red) and right (R, blue) exon boundaries. In the lower part, probabilities of left- and right-facing melting forks appear as red and blue peaks, respectively. The black curve shows the GC content in a 100-bp sliding window (values on right axis).

An alternative view is that the GC content, being higher inside exons than outside, contains information about exon location that is simply carried over, or decoded, by a melting analysis, thus acting as a confounder. We have developed a methodology to investigate such situations further. Non-preserved elements of a null model can be randomized according to a non-homogeneous Poisson process with a base-pair-varying intensity, which can depend on a third (or several) modulating genomic tracks [19,20]. We have defined an algebra for the construction of intensities, where tracks are combined, to allow rich and flexible constructions of randomness (see Materials and methods).

To investigate the influence of GC content on the exon-melting relation, we first generated a pair of custom tracks (type F), assigning to each base the value given by the GC content in the 100-bp left and right flanking regions, respectively, weighted by a linearly decreasing function. These two functions were used, together with the exon boundary track, to create an intensity curve proportional to the probability of exon points, given GC content (see Materials and methods). When performing the same analysis as before, but now using the null model based on this intensity curve (rather than assuming uniformity), a significant relationship was found in only one yeast chromosome (Table S3 in Additional file 1). In conclusion, there is a melting-exon relationship in yeast, but it may simply be a consequence of differences in GC content at the exon boundaries (high GC inside, low GC outside), which may exist for biological reasons not involving melting fork locations.

Resolving complexity: system architecture

The Genomic HyperBrowser is an integrated, open-source system for genome analysis. It is continually evolving, supporting 28 different analyses for significance testing, as well as 62 different descriptive

statistics. The system currently hosts 184,500 tracks. Most of these represent literature-based information, previously mostly utilized in network-based approaches [21]. As natural language based text mining allows for the identification of a wide variety of biological entities, we have generated tracks representing genomic locations associated with terms for the complete gene ontology tree, all Medical Subject Heading (MeSH) terms, chemicals, and anatomy.

The system is implemented in Python [22], a high-level programming language that allows fast and robust software development. A main weakness of Python compared to languages like C++ is its slower performance. Thus, a two-level architecture has been designed. At the highest level, Python objects and logic have been used extensively to provide the required flexibility. At the base-pair level, data are handled as low-level vectors, combining near-optimal storage with efficient indexing, allowing the use of vector operations to ensure speed. Interoperability with standard file formats in the field [23] is provided by parallel storage of original file formats and preprocessed vector representations. To reduce the memory footprint of analyses on genome-wide data, an iterative divide-and-conquer algorithm is automatically carried out when applicable. A further speedup is achieved by memoizing intermediate results to disk, automatically retrieving them when needed for the same or different analyses on the same track(s) at any subsequent time, by any user.

The system provides a web-based user interface with a low entry point. However, the complex interdependencies between the large body of available tracks, a number of syntactically different analyses, and a range of choices for constructing null models, all pose challenges to the concepts of simplicity and ease of use. In order to simplify the task of making choices, a step-wise approach has been implemented, displaying only the relevant options at each stage. This guided approach hides unnecessary complexities from the researcher, while confronting her with important design choices as needed. We rely on a dynamic system to infer appropriate options, aiding maintenance. The list of selectable tracks is based on scans of available files on disk. The list of relevant questions is based on short runs of all implemented analyses, using a minimal part of the actual data from the selected tracks. For each analysis, a set of relevant options is defined. The dynamics of the system also provides automatic removal of analyses that fail to run, enhancing system robustness.

Allowing extensibility along with efficiency and system dynamics is a challenge. The complexities of the software solutions are hidden in the backbone of the system, simplifying coding of statistical modules. Each module declares the data types it supports and which results are needed from other modules. The backbone automatically checks whether the selected tracks meet the requirements, and if so, makes sure the intermediate computations are carried out in correct order. Redundant computations are avoided through the use of a RAM-based memoization scheme. The system also provides a component-based framework for Monte Carlo tests, where any test statistic can be combined with any relevant randomization algorithm, simplifying development. In addition, a framework for writing unit and integration tests [24] is included. Further details on the system architecture are provided in Section 4 in Additional file 1.

Step-by-step guide to HyperBrowser analysis

One of the main goals of the Genomic HyperBrowser is to facilitate sophisticated statistical analyses. A range of textual guides and screencasts are available in the help section at the web page, demonstrating execution of various analyses, how to work with private data, and more. To give an impression of the user experience, we here provide a step-by-step guide to the analysis of broad local enrichment (BLOC) segments versus SINE repeats, as discussed in the section on 'Precise specification of null models'.

First, we open 'hyperbrowser.uio.no' in a web browser and we select the 'Perform analysis' tool under 'The Genomic HyperBrowser' in the left-hand menu. We select the mouse genome (mm8) and continue to select tracks of interest. As the first track, we select 'Chromatin'-'Histone modifications'-'BLOC segments'-'MEFB1'. These are the BLOC segments according to the algorithm of Pauler *et al.* [8] for the MEFB1 cell line. As the second track, we select 'Sequence'-'Repeating elements'-'SINE'. Now that both tracks have been selected, a list of relevant investigations is presented in the interface (that is, investigations that are compatible with the genomic types of the two tracks: US versus US). We select the question of 'Overlap?' in the 'Hypothesis testing' category, and the options relevant for this analysis are subsequently displayed in the interface. The different choices for 'Null model' will produce the various numbers in Table 1 (six different choices are directly available from the list. The other variants can be achieved by reversing the selection order of the tracks). The original BLOC paper [8] focused on chromosome 17. We want to perform a local analysis along this chromosome, avoiding the first three megabases that are centromeric. Under 'Region and scale' we thus choose to 'Compare in' a custom specified region, writing 'chr17:3m-' as 'Region of the genome' and writing '5 m' (5 megabases) as 'Bin size'. Clicking the 'Start analysis' button will then perform an appropriate statistical test according to the selected null model assumption, and output textual and graphical

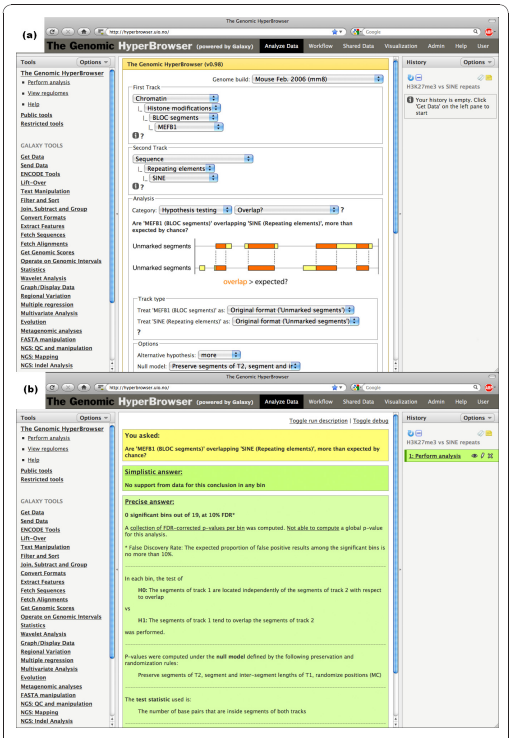


Figure 5 Screenshots of the Genomic HyperBrowser. **(a)** Screenshot of the main interface for selecting analysis options. The selections for the example relating H3K27me3 BLOCs to SINE repeats have been pre-selected. In the interface, the user selects a genome build followed by two tracks. A list of relevant investigations is then presented, based on the genomic types of the two tracks. After selecting an investigation, the interface presents the user with a choice of null models, alternative hypotheses and other relevant options. **(b)** Screenshot of the results of the analysis. The question asked by the user is presented at the top, in this case: 'Are 'MEFB1 (BLOC segments)' overlapping 'SINE (Repeating elements)' more than expected by chance?' A first, simplistic answer is then presented: 'No support from data for this conclusion in any bin'. A more precise answer follows, detailing any global *P*-values, a summary of local FDR-corrected *P*-values, the particular set of null and alternative hypotheses tested, in addition to a legend of the test statistic that has been used. Further links to a PDF file containing the statistical details of the test, and to more detailed tables of relevant statistics for both the global and the local analysis are also included. The global result table also includes links to plots and export opportunities for the individual statistics.

results to a new Galaxy history element. Figure 5a shows the user interface covering all selections above and Figure 5b shows the answer page that results from this analysis.

This example assumed the BLOC segments were already in the system. If not, they could simply be uploaded to the Galaxy history and then selected in the first track menu as '– From history (bed, wig) –'-'[your BLOC history element]'. For information on how to use the Galaxy system, we refer to the Galaxy web site [25].

Discussion

The current leap in high-throughput sequencing technology is opening the way for a range of genome-wide annotations beyond the presently abundant gene-centric data. Not least, chromatin-related data are becoming increasingly important for understanding higher-level organization and regulation of the genome [26].

As is typical for a subfield that has not reached maturation, analysis of new massive sequence-level data is performed on a per-project basis. For instance, a paper on the ENCODE project describes how inference can be done by Monte Carlo testing, sampling bins for one of the real tracks at random genome locations under the null hypothesis [1]. Independently, a newer study of histone modifications instead permuted bins of data for one of the tracks [27]. Although genomic visualization tools have been available for several years, few generic tools exist for inference at the sequence level.

The following aspects distinguish our work from currently available systems. First, we focus on genomic information of a sequential nature, that is, with specific base-pair locations on a genome, and thus not restricted to only genes. Second, it focuses on the comparison of pairs of genomic tracks, possibly taking others into account through the concept of intensity tracks. Third, all comparisons are performed using formal statistical testing. Fourth, we provide analyses on any scale, from genome-wide studies to miniature investigations on particular loci. Fifth, we offer flexible choices of null models for exploration and choice where relevant. Finally, we provide a user interface where the user describes the data and the null models, while the system based on this chooses the appropriate statistical test. Comparing this to the EpiGRAPH and Galaxy frameworks, which we believe are the closest existing systems, we find that both require substantial technical expertise when choosing the correct analysis and options. EpiGRAPH is focused on a specific type of scenario that, according to our cataloguing, amounts to the comparison of unmarked points or segments versus categorically marked segments (with mark being case or control). Galaxy provides a simple user interface, is rich in tools for manipulating and analyzing datasets of diverse formats, but has little support for formal statistical testing. Note also that our system is tightly connected to Galaxy and can make use of all the tools provided within Galaxy.

We provide tools for abstraction and cataloguing of what we believe are typical questions of broad interest.

The abstractions of genomic data, the proposing of prototype investigations, and the careful attention given to null models simplifies statistical inference for a range of possible research topics. Our approach invites researchers to build relevant null models in a controlled manner, so that specific biological assumptions can be realistically represented by preservation, randomness and intensity based confounders. In addition, time used for repetitive tasks like file parsing and calculation of descriptive statistics may be significantly reduced.

Our system is highly extensible. The software is open source, inviting the community to add new investigations and tools. Attention has been given to component-based coding and simple interfaces, facilitating extensions of the system.

The highly specialized nature of many research investigations poses a major challenge for a generic system such as the one presented here. Even though a range of analyses and options are provided, chances are that at a given level of complexity, functionality beyond what is provided by a generic system will be needed. Still, the time and effort used to reach such a point may be shortened considerably, and it should in many cases be possible to meet demands through custom extensions.

Genomic mechanisms commonly involve more than two tracks, and the current focus on pair-wise interrogations is limiting. Our methodology allows the incorporation of additional tracks through the concept of an intensity track that modulates the null hypothesis, acting as a confounder. However, the investigation of genuine multi-track interactions is not yet possible within the system, as complex modeling and testing of multiple dependencies will be required.

Attention should be given to the trade-off between fine resolution and lack of precision. When large bins are considered, there may be too little homogeneity, while small bins may contain too little data. There is also an unresolved trade-off relating to preservation of tracks in null-hypotheses construction: too little preservation may give unrealistically small *P*-values, while too strong preservation may give too limited randomness.

On a more specific note, a set of tissue-specific analytical options would be beneficial with respect to many types of experimental data - for example, chromatin, expression and also gene subset tracks. Such options are now under development.

Novel sequencing technologies are instrumental in realizing the personalized genomes [28], and with them the task of identifying phenotype-associated information contained in each genome. An imminent challenge in understanding cellular organization is that of the three dimensions of the genome. While a number of genomes have been sequenced, and a number of important cellular elements have been mapped on a linear scale, the mapping of the three-dimensional organization of the DNA and chromatin in the nucleus is still only in its beginnings. Consequently, the impact of this organization on cell regulation is still largely unresolved. However, the advent of methods like Hi-C [29] permits detailed maps of three-dimensional DNA interactions to be combined with coarser methods of mapping of other elements. It appears that looking simultaneously at multiple scales seems important for understanding the dynamics of different functional aspects, from chromosomal domains down to the nucleosome scale. The need for taking multiple scales into account has recently been emphasized in both theoretical and analytical settings [30,31]. Consequently, statistical genomics needs to consider several scales when proper analytical routines are developed. Our approach is open to three-dimensional extensions, where the bins, which are flexibly selected in the system, will become three-dimensional volumes, and local comparison will be within each volume. What appears much more complex is the level of dependence of such volumes. But as the three-dimensional organization of the genome will become increasingly known, appropriate volume topologies will be possible, so that neighboring volumes representing three-dimensional contiguity may be used as a basis for statistical tests.

Conclusions

By introducing a generic methodology to genome analysis, we find that a range of genomic data sets can be represented by the same mathematical objects, and that a small set of such objects suffice to describe the bulk of current data sets. Similarly, a range of biological investigations can be reduced to similar statistical analyses. The need for precise control of assumptions and other parameters can furthermore be met by generic concepts such as preservation and randomization, local analysis (binning) and confounder tracks.

Applying these ideas on a sample set of genomic investigations underlines that the generic concepts fit naturally to concrete analyses, and that such a generic treatment may expose vagueness of biological conclusions or expose unforeseen issues. A re-analysis of the relation between BLOC segments of histone modification and SINE repeats shows that conclusions regarding direct overlap at the base-pair level depends on the randomizations used in the significance analysis. Using biologically reasonable null models, the correspondence between BLOC segments and SINE repeats appears not to be due to overlap at the base-pair level, but rather seems to be due to local variation in intensities of both tracks. This does not directly oppose the original conclusions, but brings further insight into the nature of the relation. Similarly, an analysis of the relation between DNA melting and exon location confirms the

conclusion from previous studies that exon boundaries coincide with gradients of melting temperature. However, taking GC content into account as a possible confounder, the analysis does not suggest a direct functional relation between melting and exons. Instead, it suggests that the association is due to the relationship of both exons and melting tracks to GC content.

We believe the generic concepts and challenges identified by our work will trigger community efforts to improve genome analysis methodology. The Genomic HyperBrowser demonstrates the feasibility of applying our approach to large-scale genomic datasets, providing a concrete basis for further research and development in inferential genomics. We thus consider the solutions presented here more like a start than an end of this important endeavor.

Materials and methods

Statistical methods

A track is defined over the whole genome or only in parts of it, masking away the rest. In a local analysis, statistical tests are performed in each bin with sufficient sample size. Resizing of bins allows for localization of events (similarities, differences, and so on, between the two tracks) with flexible precision. Preservation rules leads to conditional *P*-values that are not necessarily ordered, even if the preservation mechanism is incremental. Statistical tests have been tried on simulated data, also when model assumptions are not completely fulfilled. Standard Monte Carlo requires deciding on the number of Monte Carlo samples. We suggest at least two to five times the number of tests, in order to allow for FDR adjustment. Additionally, we adopt sequential Monte Carlo, where the algorithm continues sampling until the observed statistic has been exceeded a given number of times (say 20) [9]. This gives better estimates of small *P*-values with overall reduced computations. Intensity tracks are used to define non-standard null hypothesis. Several strategies for building intensity curves are described in Section 3 in Additional file 1. Intensity curves allow performing randomizations that mimic another track (or a combination of tracks), useful to account for confounding effects. For unmarked points, the intensity curve can be any regular function $\lambda_0(b)$ where b is the position along, say, a chromosome. If $\lambda_0(b) = c$ (constant), points are uniformly distributed. As another example, $\lambda_0(b)$ can be a kernel density estimate based on the track of observed points. In general, the intensity $\lambda_0(b)$ may depend on several different tracks $g_1, g_2, ..., g_k$, through a function s, so that $\lambda_0(b) = s(g_1(b), g_2(b), ..., g_k(b))$, for example, $\lambda_0(b) = c + \Sigma\beta_i g_i(b)$. An important case that requires a special choice of intensity track is when the comparison between two tracks T_1 and T_2 might be confounded by a third, confounder, track T_3. This is discussed in further detail in Section 5 in Additional file 1 for the melting-exon example, where each track depends on a function of the GC content.

Software system

The Genomic HyperBrowser [30] is implemented in Python [22], version 2.7. It runs as a stand-alone application tightly connected to the Galaxy framework [2], using the version dated 2010-10-04. The user interface is based on Mako templates for Python [32], version 0.2.5, and Javascript library Jquery [33], version 1.4.2. The software uses NumPy [34], version 1.5.1rc1, for disk based vector mapping and fast vector operations. R [35], version 2.10.1, is used for plotting and basic statistical routines, using the RPy API [36], version 1.0.3. The software is open source and freely available, using GPL [37] version 3, and can be downloaded from [30]. The Genomic HyperBrowser runs on a dedicated Linux server, with large computations offloaded to the Titan cluster [38].

Biological example: histone modifications versus gene expression

Raw histone modification data [39] were preprocessed using the NPS (Nucleosome Positioning from Sequencing) software [40], using peak detection, leading to nucleosome positioning information as short segments, treated as unmarked points (UP). Raw microarray expression values [41] were used to represent gene expression, in line with [6]. Direct comparison of the expression levels of individual probes is not generally justified. As Barski *et al*. [6] compares sets of 1,000 genes each, the direct comparison of values between groups may be justified by noise averaging (although not discussed in [6]). Using Kendall's rank correlation test, a similar reduction of error is obtained. Detailed correlation values for the different histone modifications are given in Table S1 in Additional file 1. The distribution of histone modifications relative to TSS is given for two different modifications in Figure S4 in Additional file 1.

Biological example: histone modifications versus repeating elements

ChIP-seq data on histone modification [39,42] were preprocessed using the SICER software [43], which returns clusters of neighboring nucleosomes as islands unlikely to have appeared by chance, using an appropriate random background model. These clusters are treated as unmarked segments (US). The ChIP-chip data of H3K27me3 positions were obtained directly from Pauler *et al*. [8], and were preprocessed by them using their BLOCs software, which returns broad local enrichments, also treated as unmarked segments (US). Detailed overlap results between repeats and different histone modification sources are given in Table S2 in Additional file 1.

Biological example: exons versus DNA melting

The melting fork probability tracks $P_L(x)$ and $P_R(x)$ used in this study were obtained using the Poland-Scheraga model [44]. To make the correction for GC content, a pair of GC-based function tracks, L(x) and R(x), were created using a moving window approach. Let E_L (E_R) be the left (right) exon boundaries. For testing the melting-exon relation in tracks (E_L, P_L), an intensity track was created based on L(x), R(x) and E_L(and similarly for tracks (E_R, P_R)). See Section 5 in Additional file 1 for more details.

Additional material

Additional file 1: Supplementary material. Miscellaneous supplementary material: gene coverage example. On the importance of realistic null models. On mathematics of genomic tracks. On system architecture. On Exon DNA melting example. Supplementary figures and tables.

Additional file 2: Statistical tests. Detailed description of the statistical tests implemented in the software system.

Additional file 3: Supplementary note on simulation. Description of basic algorithms for simulating synthetic tracks, used to assess statistical tests.

Abbreviations

BLOC: broad local enrichment; bp: base pair; F: function; FDR: false discovery rate; kb: kilo base pairs; LINE: long interspersed nuclear element; Mbp: mega base pairs; MP: marked point; MS: marked segment; SINE: short interspersed nuclear element; SNP: single-nucleotide polymorphism; TSS: transcription start site; UP: unmarked point; US: unmarked segment.

Acknowledgements

We gratefully acknowledge ChIP-chip data provision from Florian M Pauler, and helpful comments on the manuscript from Magnus Lie Hetland, Sylvia Richardson and Håvard Rue. Gro Nilsen is acknowledged for some plotting functions, and Peter Wiedswang for administrative assistance. We thank the Scientific Computing Group at USIT for providing friendly and helpful assistance on system administration. We also thank PubGene, Inc. for kind assistance in the development of literature tracks. Additional funding was kindly provided by EMBIO, UiO and Helse Sør-Øst. This work was performed in association with 'Statistics for Innovation', a Centre for Research-Based Innovation funded by the Research Council of Norway.

Author details

[1]Department of Informatics, University of Oslo, Blindern, 0316 Oslo, Norway. [2]Department of Tumor Biology, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, 0310 Oslo, Norway. [3]Statistics For Innovation, Norwegian Computing Center, 0314 Oslo, Norway. [4]Department of Mathematics, University of Oslo, Blindern, 0316 Oslo, Norway. [5]Centre for Cancer Biomedicine, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, 0310 Oslo, Norway. [6]Institute for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, 0310 Oslo, Norway. [7]Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Blindern, 0317 Oslo, Norway.

Authors' contributions

GKS, AF and EH conceived the approach, GKS, SG and MJ developed the software, GKS, SG, HR, TC, VN and EH developed novel track types, IKG, LH, MH, KL, EF and AF developed the statistical concepts, GKS, SG and HR tested and validated the system, and GKS, SG, HR, ET and EH developed the biological examples. All authors participated in the manuscript development, and read and approved the final manuscript.

Competing interests

Eivind Hovig is a shareholder of PubGene, Inc. All other authors declare that they have no competing interests.

Received: 27 August 2010 Revised: 8 December 2010
Accepted: 23 December 2010 Published: 23 December 2010

References

1. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004, **306**:636-640.
2. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A: **Galaxy: a platform for interactive large-scale genome analysis.** *Genome Res* 2005, **15**:1451-1455.
3. Pruess M, Kersey P, Apweiler R: **The Integr8 project–a resource for genomic and proteomic data.** *In Silico Biol* 2005, **5**:179-185.
4. Bock C, Halachev K, Buch J, Lengauer T: **EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data.** *Genome Biol* 2009, **10**:R14.
5. Zhu J, Sanborn JZ, Benz S, Szeto C, Hsu F, Kuhn RM, Karolchik D, Archie J, Lenburg ME, Esserman LJ, Kent WJ, Haussler D, Wang T: **The UCSC Cancer Genomics Browser.** *Nat Methods* 2009, **6**:239-240.
6. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129**:823-837.
7. Derse D, Crise B, Li Y, Princler G, Lum N, Stewart C, McGrath CF, Hughes SH, Munroe DJ, Wu X: **Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses.** *J Virol* 2007, **81**:6731-6741.
8. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, Sommer A, Aszodi A, Jenuwein T, Barlow DP: **H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome.** *Genome Res* 2009, **19**:221-233.
9. Besag J, Clifford P: **Sequential Monte Carlo p-values.** *Biometrika* 1991, **78**:301-304.
10. Manly BFJ: *Randomization, Bootstrap and Monte Carlo Methods in Biology* Boca Raton, FL: Chapman and Hall; 2007.
11. Jost D, Everaers R: **Genome wide application of DNA melting analysis.** *J Phys Condensed Matter* 2009, **21**:034108.
12. King GJ: **Stability, structure and complexity of yeast chromosome III.** *Nucleic Acids Res* 1993, **21**:4239-4245.
13. Liu F, Tostesen E, Sundet JK, Jenssen TK, Bock C, Jerstad GI, Thilly WG, Hovig E: **The human genomic melting map.** *PLoS Comput Biol* 2007, **3**:e93.
14. Suyama A, Wada A: **Correlation between thermal stability maps and genetic maps of double-stranded DNAs.** *J Theor Biol* 1983, **105**:133-145.
15. Yeramian E: **Genes and the physics of the DNA double-helix.** *Gene* 2000, **255**:139-150.
16. Tøstesen E, Sandve GK, Liu F, Hovig E: **Segmentation of DNA sequences into twostate regions and melting fork regions.** *J Phys Condensed Matter* 2009, **21**:034109.
17. Carlon E, Malki ML, Blossey R: **Exons, introns, and DNA thermodynamics.** *Phys Rev Lett* 2005, **94**:178101.
18. Hanai R, Suyama A, Wada A: **Vestiges of lost introns in the thermal stability map of DNA.** *FEBS Lett* 1988, **226**:247-249.
19. Cox DR, Isham V: *Point Processes* Boca Raton, FL: Chapman and Hall; 1980.
20. Grandell J: *Mixed Poisson Processes* Boca Raton, FL: Chapman and Hall; 1997.
21. Jenssen TK, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression.** *Nat Genet* 2001, **28**:21-28.
22. **Python Reference Manual.** [http://docs.python.org/release/2.5.2/ref/ref.html].
23. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12**:996-1006.
24. Beck K: *Test Driven Development* London: Addison-Wesley Profession; 2002.
25. **Galaxy.** [http://main.g2.bx.psu.edu/].
26. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range**

interactions reveals folding principles of the human genome. *Science* 2009, **326**:289-293.
27. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K: **Combinatorial patterns of histone acetylations and methylations in the human genome.** *Nat Genet* 2008, **40**:897-903.
28. **1000Genomes.** [http://www.1000genomes.org/].
29. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326**:289-293.
30. Naumova N, Dekker J: **Integrating one-dimensional and three-dimensional maps of genomes.** *J Cell Sci* **123**:1979-1988.
31. Knoch TA, Goker M, Lohner R, Abuseiris A, Grosveld FG: **Fine-structured multi-scaling long-range correlations in completely sequenced genomes - features, origin, and classification.** *Eur Biophys J* 2009, **38**:757-779.
32. **Mako.** [http://www.makotemplates.org].
33. **JQuery.** [http://jquery.com].
34. Oliphant TE: In *Guide to NumPy*. Edited by: Spanish Fork UT. Trelgol Publishing; 2006:.
35. Team R: *R: A Language and Environment for Statistical Computing* Vienna: Austria; R Foundation for Statistical Computing; 2006.
36. **RPy a robust Python interface to the R Programming Language.** [http://rpy.sf.net].
37. **GPL.** [http://www.gnu.org/copyleft/gpl.html].
38. **Titan.** [http://www.notur.no/hardware/titan/].
39. Barski A, Zhao K: **Genomic location analysis by ChIP-Seq.** *J Cell Biochem* 2009, **107**:11-18.
40. Zhang Y, Shin H, Song JS, Lei Y, Liu XS: **Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq.** *BMC Genomics* 2008, **9**:537.
41. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101**:6062-6067.
42. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448**:553-560.
43. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W: **A clustering approach for identification of enriched domains from histone modification ChIP-Seq data.** *Bioinformatics* 2009, **25**:1952-1958.
44. Poland D, Scheraga HA: *Theory of Helix-Coil Transitions in Biopolymers* New York: Academic Press; 1970.
45. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, *et al*: **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.** *Genome Res* 2009, **19**:1316-1323.
46. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives.** *Nucleic Acids Res* 2009, **37**:D32-36.
47. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, *et al*: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30**:38-41.
48. Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL: **The vertebrate genome annotation (Vega) database.** *Nucleic Acids Res* 2008, **36**:D753-760.
49. Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, Takeda J, Taniya T, Sakate R, Kikugawa S, Shimada M, Tanino M, Koyanagi KO, Barrero RA, Gough C, Chun HW, Habara T, Hanaoka H, Hayakawa Y, Hilton PB, Kaneko Y, Kanno M, Kawahara Y, Kawamura T, Matsuya A, Nagata N, Nishikata K, Noda AO, Nurimoto S, Saichi N, Sakai H, *et al*: **The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts.** *Nucleic Acids Res* 2008, **36**:D793-799.

doi:10.1186/gb-2010-11-12-r121
Cite this article as: Sandve *et al.*: The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biology* 2010 **11**:R121.