Elucidating Mechanisms of Gene Regulation

Integration of High-Throughput Sequencing Data for Studying the Epigenome

Sonja Daniela Althammer

TESI DOCTORAL UPF / ANY 2012

DIRECTOR DE LA TESI
Eduardo Eyras
Departament Department of Experimental and Health Sciences
(CEXS)

This PhD thesis was carried out in the Regulatory Genomics Group of the Research Unit on Biomedical Informatics (GRIB), affiliated to the Department of Experimental and Health Sciences of UPF and to the “Institut Municipal d’Investigació Mèdica“ (IMIM), located at the Barcelona Biomedical Research Park (PRBB).

The underlying research was funded by the Generalitat de Catalunya by means of a fellowship (Modalidad B), as well as by the Plan Nacional I+D projects BIO2008-01091 and BIO2011-23920, and Consolider project CSD2009-00080 from the Spanish Government and project EURASNET-LSHG-CT-2005-518238 from the European Comission.

prbb Barcelona Biomedical Research Park

Generalitat de Catalunya
www.gencat.cat

GRIB RESEARCH PROGRAMME ON BIOMEDICAL INFORMATICS

UPF UNIVERSITAT POMPEU FABRA

Acknowledgements

I want to thank everyone who helped me with this thesis in any way, especially:

My supervisor,
Eduardo Eyras,

people that were directly involved in the projects,
Juan González-Vallinas, Cecilia Ballaré, Amadis Pages,
Nicolas Bellora, Eneritz Agirre,

other people from the lab that helped me a lot,
Alba Jené, Gunes Gundem, Sonja Hänzelmann,
Alfons Gonzalez and Miguel Sánchez, Carina Oliver, Khademul Islam,
Christian Pérez, Andre Corvelo, Michi Schroeder, Sophia Derdak, Emre Guney,
David Tamborero, Macarena Toll-Riera, Inmaculada Tur,

and people from other labs,
Sarah Bonnin, Julien Lagarde, Sarah Djebali, Christoforos Nikolaou,
Hagen Tilgner, Roderic Giugó, João Curado, Anna Bauer-Mehren,
Rory Johnson, Debayan Datta, Emre Guney, Anne Campagna, Erik Verschueren,
David Gonzalez, Angelika Merkel, Andrea Tanzer,
Pedro Ferreira, Romina Garrido, Colin Kingswood, Anaïs Bardet,
Guillaume Arras, Meromit Singer, Pau Carrio, Luciano Di Croce,
Nuria Lopez-Bigas, Tanya Vavouri, Francisco Câmara, Giancarlo Castellano,
Miguel Beato, Andy Pohl, Camilla Iannone, Nadine Richter, Oscar Gonzalez,
Micha Sammeth, Laura Quintana Rio, Laura Lopez, Filipe Pinto Teixeira Sousa,
Natàlia Ras, Derrien Thomas, Thomas Gingeras, Matt Ingham, Marco Mariotti,
Sylvain Foissac, Rosa Garcia-Verdugo,

more friends,
Veronika Oudova, Sara Soler and Fabio Ercole, Andrea Hildebrand,
Claudia Raba, Max Stöckle, Pamela Schricker, Ahmad Ahmadi, Robert Stöckle,
Thomas Hirschmann, Dorothée Saller, Aleix Navarro Sastre, Lise Andrieux,
Karin Wolfe, Karina Cerda Oñate, Daniela Espinosa, Luisa Lente

and, most importantly, my family.

Abstract

The recent advent of high-throughput sequencing (HTS) methods has triggered a revolution in gene regulation studies. Demand has never been higher to process the immense amount of emerging data to gain insight into the regulatory mechanisms of the cell.

We address this issue by describing methods to analyze, integrate and interpret HTS data from different sources. In particular, we developed and benchmarked Pyicos, a powerful toolkit that offers flexibility, versatility and efficient memory usage. We applied it to data from ChIP-Seq on progesterone receptor in breast cancer cells to gain insight into regulatory mechanisms of hormones. Moreover, we embedded Pyicos into a pipeline to integrate HTS data from different sources. In order to do so, we used data sets from ENCODE to systematically calculate signal changes between two cell lines. We thus created a model that accurately predicts the regulatory outcome of gene expression, based on epigenetic changes in a gene locus. Finally, we provide the processed data in a Biomart database to the scientific community.

Resumen

La llegada reciente de nuevos métodos de *High-Throughput Sequencing* (HTS) ha provocado una revolución en el estudio de la regulación génica. La necesidad de procesar la inmensa cantidad de datos generados, con el objectivo de estudiar los mecanismos regulatorios en la celula, nunca ha sido mayor.

En esta tesis abordamos este tema presentando métodos para analizar, integrar e interpretar datos HTS de diferentes fuentes. En particular, hemos desarollado Pyicos, un potente conjunto de herramientas que ofrece flexibilidad, versatilidad y un uso eficiente de la memoria. Lo hemos aplicado a datos de ChIP-Seq del receptor de progesterona en células de cáncer de mama con el fin de investigar los mecanismos de la regulación por hormonas. Además, hemos incorporado Pyicos en una *pipeline* para integrar los datos HTS de diferentes fuentes. Hemos usado los conjuntos de datos de ENCODE para calcular de forma sistemática los cambios de señal entre dos líneas celulares. De esta manera hemos logrado crear un modelo que predice con bastante precisión los cambios de la expresión génica, basándose en los cambios epigenéticos en el locus de un gen. Por último, hemos puesto los datos procesados a disposición de la comunidad científica en una base de datos Biomart.

Preface

Ever since the structure of the DNA was described, scientists have hoped to reveal its role in life. The excitement was even bigger when the genetic code was cracked. It discloses how a stretch of DNA is translated into a protein sequence. Despite significant progress in genetic research, we are still far away from understanding what happens exactly inside our cells.

The genome, located in the nucleus of each cell, is identical in every cell of our body. So why do cells have different appearances and functions? This is explained by differences in gene expression. Depending on the activity of the genes, and thus the concentration of proteins and other molecules in the cell, a certain cell state is expressed. This differentiates, for instance, a lung cell from a kidney cell. The regulation of gene expression can further determine if a cell is normal, or affected by a disease. Even though an individual genome can comprise a predisposition (e.g. a mutation) for a disease, the expression of a diseased state depends on gene regulation, which in turn relies on a complicated network of intracellular molecules and can further be affected by extracellular conditions.

In gene regulation studies, a typical task is to look for active binding sites of molecules that regulate a gene or a set of genes and to measure the corresponding change in expression. The recent development of High-Throughput Sequencing (HTS) methods has boosted the production of massive amounts of data, which gives unforeseen opportunities to tackle this task. By now, gene regulation studies are mainly based on HTS experiments, trying to reveal the complicated network of the different factors in a cell. The HTS technology is constantly gaining new applications, while computational methods to deal with the emerging data lag behind. This PhD addresses the development of methods to analyze, integrate and interpret HTS data, with the final goal of gaining insight into mechanisms of gene regulation.

Contents

List of Figures

Part I

Introduction

Chapter 1

INTRODUCTION

All cells of one individual accommodate identical genomes (genotypes) in the nucleus, however their appearances and functions (phenotypes) are different. This can be explained by the fact that in different cell types genes are differentially expressed. Genes code for proteins or non-coding RNAs. The abundance of these different gene products is crucial for the identity of the cell. They determine, for instance, whether a cell is normal or affected by a disease. Studying how genes are regulated is therefore fundamental for biomedical research.

1.1 Mechanisms of gene regulation

A gene is known as a stretch of DNA that is usually composed of exons and introns (see Figure 1.1). Within protein-coding genes, only exons code for the protein sequence. The rest of the genome is non-coding and used to be referred to as “junk DNA”. However by now, we know that non-coding DNA is anything but “junk”. In fact, as described in this thesis, it plays very important rules in gene regulation.

Genes can be regulated pre- or post-transcriptionally. Pre-transcriptionally, proteins bind to their cis-regulatory targets, the promoter or distal regulatory elements, where they affect the transcription of a gene. Transcription is initiated by various transcription factors that recruit RNA Polymerase II (RNAPII), an enzyme that transcribes DNA into RNA. At the post-transcriptional level, gene regulation occurs first through precursor messenger RNA (pre-mRNA) splicing and is controlled by splicing factors and small nuclear RNAs (snRNAs). The splicing procedure removes introns, and concatenates exons to each other. Finally, the mature mRNA is transported outside the nucleus and is translated into a protein (see Figure 1.2).

The two levels of gene regulation can not be completely separated. As

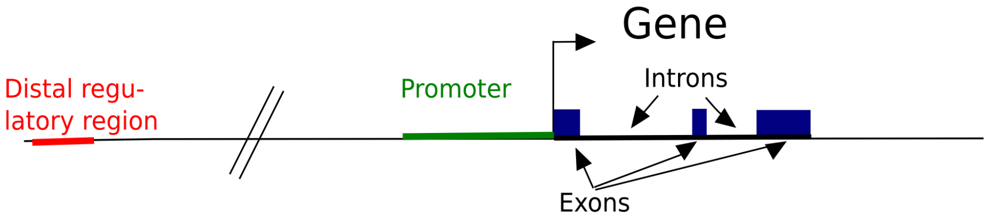


Figure 1.1: **What is a gene?** A gene is usually made up of exons and introns. In order to transcribe a gene, the transcription machinery binds to the promoter region, which lies upstream of that gene. Distal regulatory regions (enhancers or silencers) can further influence transcription.

the related factors are actually very close to each other, splicing can occur co-transcriptionally [Kornblihtt et al., 2004, Bentley, 2005].

Even though a substantial part of the genome is transcribed, only about 2% of it consists of protein coding genes. However the genome comprises also non-coding genes, whose transcripts operate in the cell either as structural RNAs (e.g. tRNAs) or as regulators for transcription [ENCODE Project Consortium, 2007]. The packaging of eukaryotic genomes into chromatin adds another layer of complexity to transcription regulation, which is explained in more detail in the section ”Chromatin and gene regulation”. Studying regulatory mechanisms of chromatin is a major challenge in epigenetics, the study of heritable variations in gene expression that cannot be explained by alterations in the DNA sequence [Waddington, 1953], but by epigenetic traits. Another important epigenetic trait is DNA methylation. It plays an important role in gene regulation, as hypermethylation of promoter regions has been shown to be involved in gene silencing and its deregulation is tightly related to cancer [Robertson, 2005]. Indeed, epigenetics has become very important for cancer research. Epigenetic changes can result into activation of oncogenes or silencing of tumor suppressor genes, and therefore initiate cell transformation.

1.2 Methods to study gene regulation

The study of gene regulation is nowadays dominated by techniques that are based on High-Throughput Sequencing (HTS), which overcome crucial limitations of the predecessor techniques.

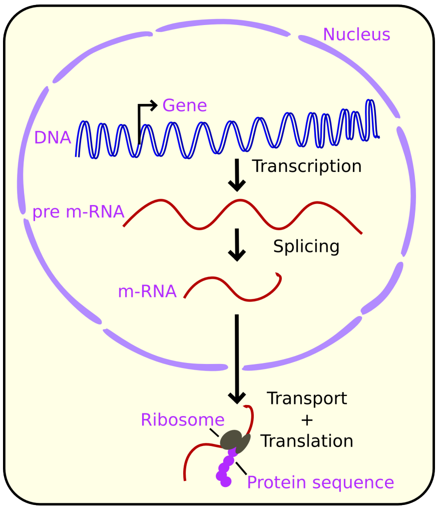


Figure 1.2: **From DNA to protein** This simplified overview shows how a gene is transcribed into pre-mRNA. The pre-mRNA is then spliced into mRNA. After leaving the nucleus, the mRNA is transcribed into a protein.

1.3 High-Throughput Sequencing (HTS)

The first draft of the human genome was completed in 2001 [Lander et al., 2001, Venter et al., 2001] and since then tremendous advances in sequencing technology have led to a new era of genome research. Back in those days Sanger sequencing used to be the method of choice, however since about six years it has been replaced by improved methods to sequence large amounts of DNA: The High-Throughput Sequencing (HTS) techniques. Currently Illumina/Solexa Genome Analyser (GA) and Hi-Seq instruments, Roche 454 and SOLiD from Life Technology, are the dominating HTS platforms that are commercially available [Pareek et al., 2011]. Most data analyzed in this thesis was generated on platforms from Illumina/Solexa, which produce short reads (typically as long as 75 or 100 base pairs [Metzker, 2010]). The arrival of the HTS technology has decreased the costs of sequencing significantly, which led to an increasing variety of applications. At the moment, many large-scale projects are based on HTS techniques, e.g. the Encyclopedia of DNA Elements (ENCODE) [ENCODE Project Consortium, 2007] project, funded by the US National Human Genome Research Institute (NHGRI). ENCODE aims to define the functional elements encoded in the human genome, such as genes, transcripts and transcriptional regulatory regions with their patterns of chromatin states and DNA methylation. The HTS technology has made this project feasible at a genome-wide scale. A major part of the present thesis relies on ENCODE data,

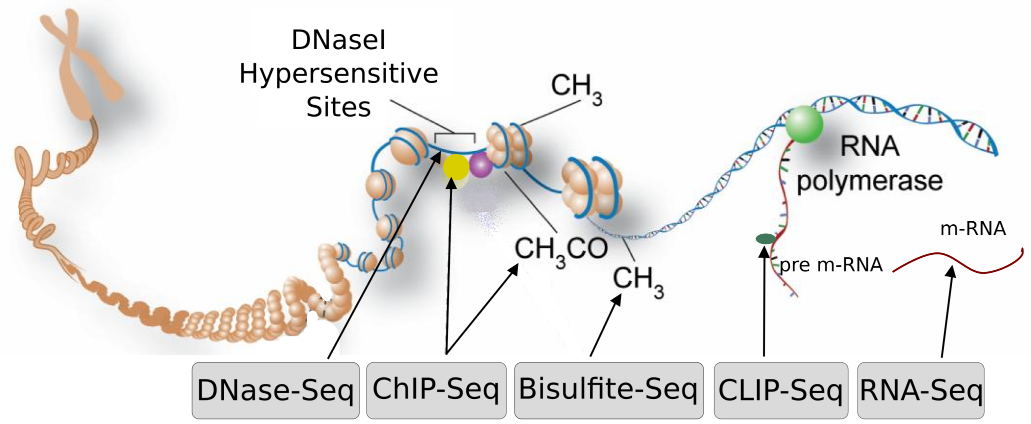


Figure 1.3: **HTS techniques and their applications.** DNase-Seq to determine regulatory sites. ChIP-Seq to detect protein-DNA binding sites. Bisulfite-Seq to identify DNA methylations. CLIP-Seq to detect protein-RNA binding sites. RNA-Seq for transcript quantification. Modified from [ENCODE Project Consortium, 2011]

since it provides a wide range of high quality HTS data sets in many different cell lines. The data is released to the public, thereby enabling the scientific community to interpret the human genome and apply it to medical research with the aim of improving health. Regulatory genomics, the study of mechanisms that regulate gene expression, is fundamental for medical research. In order to investigate regulatory genomics, it has become essential to define regulatory sites in the genome, to detect binding sites of specific proteins to DNA or RNA, to discover methylated regions of DNA and, last but not least, to measure the abundances of transcripts. Figure 1.3 shows different HTS techniques and their applications in gene regulation studies. This thesis deals with the analyzes and the integration of a great variety of HTS data whose generation involves different experimental techniques, as explained below.

1.4 ChIP-Seq

The ChIP-Seq method was developed in 2007 [Johnson et al., 2007, Robertson et al., 2007] and consists of Chromatin ImmunoPrecipitation (ChIP) followed by HTS (Seq). It has become the method of choice for detecting the binding sites of a specific protein to the DNA, offering higher sensitivity and specificity [Johnson et al., 2007, Robertson et al., 2007] than its predecessor ChIP-chip [Ren et al., 2000]. Moreover it solves limitations of the array, like the number of probes in the array and cross hybridization issues. In gene regulation

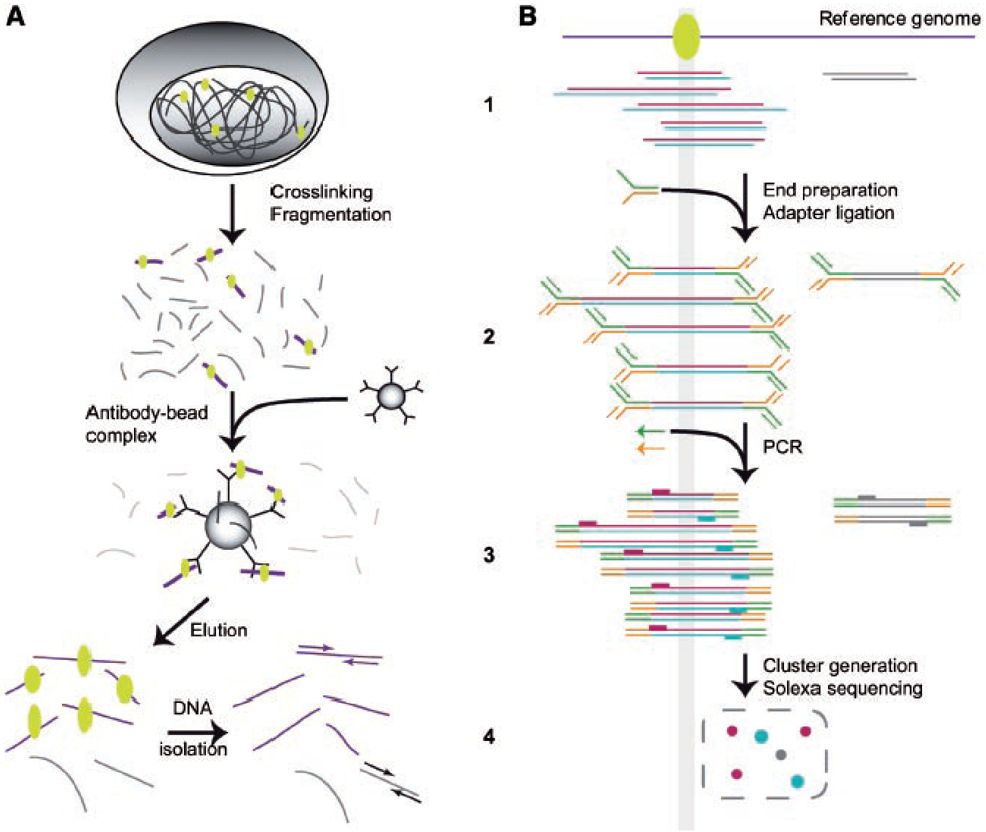


Figure 1.4: A: **Chromatin immunoprecipitation.** 1. DNA is bound by a protein of interest in the nucleus. 2. Cells are lysed and DNA is fragmented. 3. Fragments bound by protein of interest (purple) are bound by antibody bead complexes and precipitated. Some DNA fragments can be precipitated nonspecifically(gray). 4. ProteinDNA complexes are eluted and DNA is purified. Relative abundance of specific and non-specific fragments is analyzed by qPCR. B: **ChIP-Seq library construction.** 1. Specific (colored) non-specific (gray) immunoprecipitated fragments are shown mapped to genome. 2. DNA termini are polished, phosphorylated and adapters are ligated. 3. Library is PCR amplified. 4. DNA fragments are hybridized to flowcell, clusters are synthesized and sequenced. Adapted from Barski and Zhao [Barski and Zhao, 2009].

studies the DNA-binding protein of interest is often a transcription factor or a modified histone tail, whose presence or absence is typically associated with the transcription level of a gene. The protocol of ChIP-Seq starts by lysing the cells and fragmenting the chromatin, as illustrated in Figure 1.4. Bead complexes of specific antibodies are used to select those DNA fragments that are bound to the protein of interest. Next the protein-DNA complexes are eluted and the DNA is purified. After amplifying the DNA by several rounds of PCR, the short ends of the fragments are sequenced. The resulting sequenced tags, which we will refer to as reads, are then mapped back to the reference genome, resulting into clusters on the plus and minus strand, as illustrated in Figure 1.5. The binding site of the protein of interest is expected to be located at the plus to minus transition point between oppositionally directed clusters [Barski and Zhao, 2009].

There are several sources of biases to be aware of when analyzing ChIP-Seq data. The most important factor is the quality of the antibody. If it is not specific enough, fragments that are not actually bound to the protein of interest can be captured. Another bias is the under representation of AT-rich regions when low melting temperatures are used [Quail et al., 2008]. If a PCR amplification step is needed an additional bias is introduced: Areas with high GC content are not properly amplified [Barski and Zhao, 2009]. This leads to unequal distributions of read coverage among the targeted sequences and can result into duplicated reads, which have been preferentially synthesized due to their base composition. In general, the number of PCR cycles should be kept low or even avoided as in the amplification free library preparation [Kozarewa et al., 2009]. If PCR is indispensable, duplicated reads are usually excluded from further analyzes. A drawback of the short read length is that protein binding to repetitive regions might not be captured when uniquely mapping reads are used. Another factor is the preferred fragmentation of open chromatin, which can be compensated by comparing to a control experiment. Finally protein-protein interactions and looping of DNA can confound the interpretation of the data, as in this case there might not be a direct binding [Barski and Zhao, 2009].

1.5 CLIP-Seq

Cross-Linking ImmunoPrecipitation followed by HTS (CLIP-Seq) or HITS-CLIP is a powerful mean to detect protein-RNA interactions [Xue et al., 2009, Licatalosi et al., 2008]. Being able to determine the location of specific RNA-binding proteins can give insight into the regulation of alternative splicing. Relying on a method called CLIP [Ule et al., 2005], the protein-RNA complexes are crosslinked and purified, using a specific antibody. Next, the fragments of bound RNA are isolated and their ends are sequenced. When we map the reads

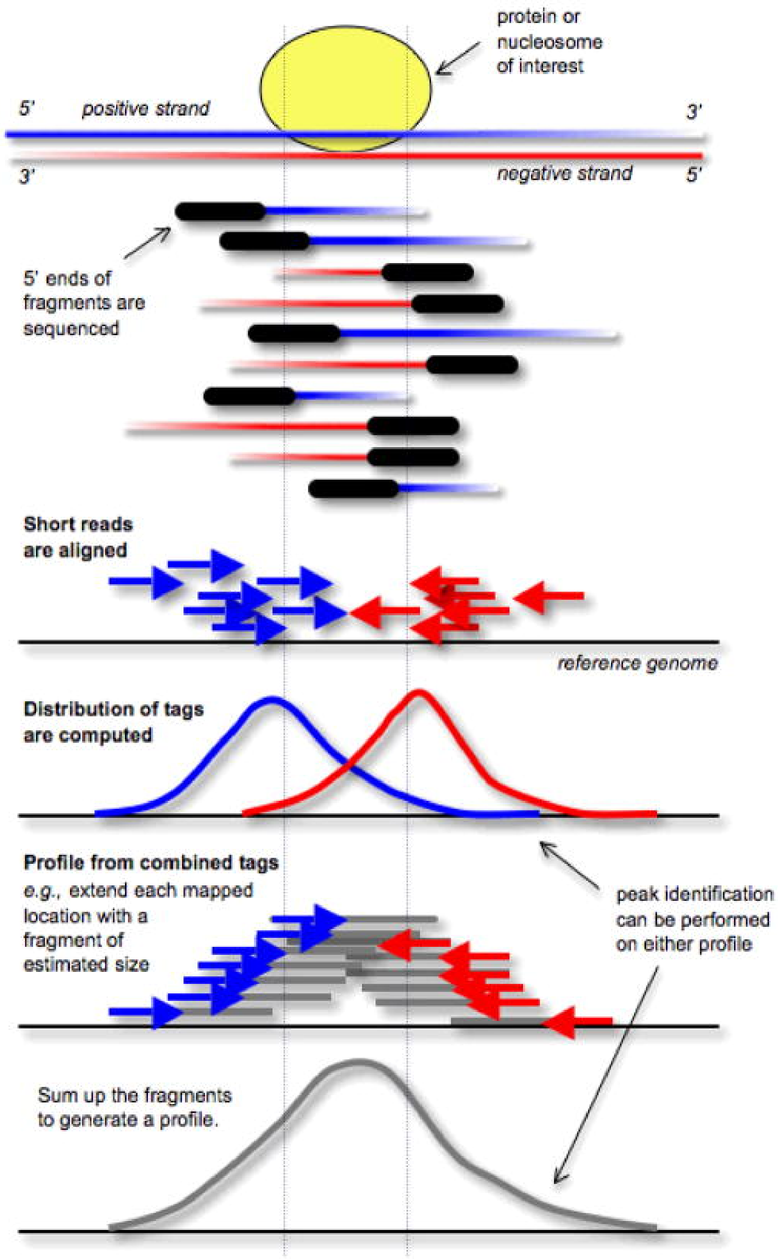


Figure 1.5: A: **Interpretation of ChIP-Seq data.** DNA fragments from a chromatin immunoprecipitation experiment are sequenced from the 5' end. Thus, alignment of these tags to the genome results into two peaks, one on each strand, flanking the location where the protein or nucleosome of interest was bound. Each tag can be extended by an estimated fragment size in the appropriate orientation resulting into a profile. The binding site is expected to be located close to the summit of this profile. Adapted from Park [Park, 2009].

to the genome or transcriptome we retrieve stranded clusters that represent the candidates of protein-RNA binding. Just as in the ChIP-Seq protocol, a crucial factor here is the specificity of the antibody.

1.6 RNA-Seq

RNA-Seq [Nagalakshmi et al., 2008] is increasingly replacing the DNA microarray technology [Schena et al., 1995] and is now the preferred method to study gene expression. It overcomes hybridization related limitations and makes possible to capture alternative splicing events [Pan et al., 2008], measure expression levels [Pepke et al., 2009], detect single nucleotide polymorphisms [Wang et al., 2008] or discover novel genes [Khalil et al., 2009].

The RNA-Seq experiment (Figure 1.6) starts typically by isolating PolyAdenylated (Poly-A+) RNA. In order to sequence the RNA, it has to be reverse transcribed into cDNA. Next, adapters are attached to one or both ends of the fragment. The cDNAs are amplified and finally sequenced [Roy et al., 2011]. After this, the bioinformatics workflow begins by mapping the reads to a reference genome or transcriptome. By counting the number of reads that fall within a region of interest we are basically able to quantify its expression. However, there are some issues one might want to consider beforehand. Additionally to the biases from sequencing and mapping, there are further challenges encountered with RNA-Seq data. First of all, it is not straightforward from which of the alternative transcripts a read emerged. Several approaches [Trapnell et al., 2010, Guttman et al., 2010] have been suggested to solve this problem. Furthermore, simply counting the reads that fall within a transcript, results into a so called length-bias: Longer transcripts contain more reads. To compensate for that, typically the counts are divided by the length of the transcript in which they were measured. Additionally one might want to divide by the library size (number of reads in a sample), in order to make different libraries comparable to each other. The combination of these two normalization methods has been introduced by Mortazavi et al., as the Reads Per Kilo base per Million mapped reads (RPKM) . [Mortazavi et al., 2008]. Another normalization approach is the trimmed mean of M-values normalization [Robinson and Oshlack, 2010], which targets the correction of biases from differences in sample sizes and expression patterns. It is based on the hypothesis that the majority of genes are expressed similarly between two samples. Even though several different normalization methods have been suggested, there is room for improvement to remove the biases that may confuse differential expression analysis.

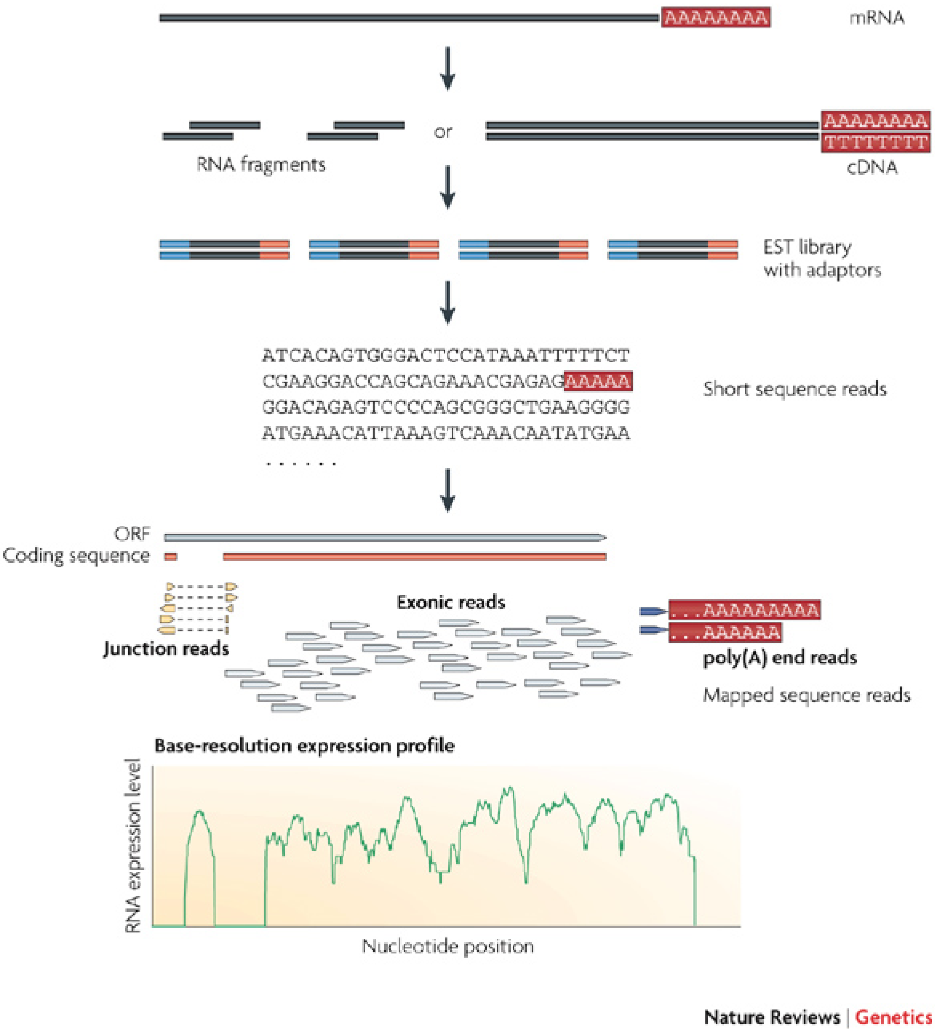


Figure 1.6: **A typical RNA-Seq experiment.** Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast open reading frame (ORF) with one intron is shown. Adapted from Wang et al. [Wang et al., 2009].

1.7 Bisulfite Sequencing

Bisulfite sequencing is a method to detect CpGs with a methylated C in the genome. The methylation of the C residues in CpG dinucleotides is the most common epigenetic modification of DNA in mammals [Bird, 2002]. It has been related to the silencing of the transcribed region located downstream and is therefore considered as a key regulator of gene expression. We have been using Reduced Representation Bisulfite Sequencing [Meissner et al., 2005] data from ENCODE in our analysis. The protocol starts by digesting the DNA with MspI, an enzyme that cuts at 5'-CCGG-3' independent of its methylation status [Waalwijk and Flavell, 1978]. A size selection step isolates fragments with an appropriate size for sequencing, resulting into a "reduced representation". Next, the DNA fragments are treated with bisulfite, which converts all unmethylated cytocines (Cs) to uracils, while methylated Cs are left intact. After PCR-amplification the DNA fragments are sequenced. When the reads are mapped to a reference, all Cs in the bisulfite sequence are suspect of methylation, while C to T transitions suggest the presence of unmethylated Cs. One problem of this method is the size selection step. If by chance the distance between two 5'-CCGG-3' is too large, the information about their methylation status might get lost.

1.8 DNase-Seq

DNase-Seq was developed [Boyle et al., 2008] to identify regulatory sites across the genome. First the DNA is digested with DNase I, an enzyme that preferentially digests nucleosome depleted regions, while it leaves the tightly wrapped heterochromatin intact. Then, the resulting fragments are sequenced and reads that have successfully been mapped to the reference genome are used to create a map of open chromatin. DNase I hypersensitive sites have been related to different types of regulatory elements, like promoters, enhancers, silencers and insulators [Song and Crawford, 2010].

1.9 Challenges of HTS data analyzes

The result of each of the experiments explained above are short sequenced reads. In order to assign the signal to genomic positions, the reads first need to be mapped to a reference. This task is out of the scope of this work and several tools have already been proposed for this purpose [Langmead et al., 2009, Li et al., 2008, Li and Durbin, 2009]. After the reads have been mapped to the reference, there

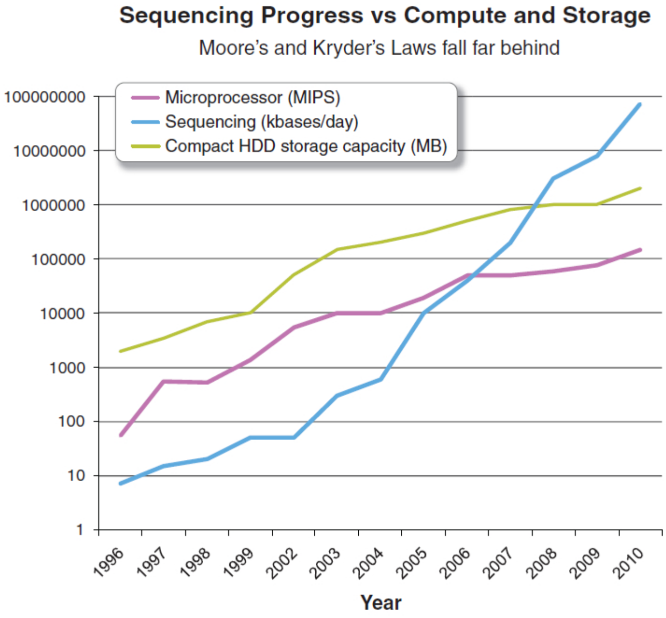


Figure 1.7: **Sequencing progress vs compute and storage.** A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields. Adapted from Kahn [Kahn, 2011].

are many challenges to overcome in order to detect significant signals. Depending on the underlying experiment, several sources of biases have to be taken into account and significance has to be calculated in an appropriate way. The final goal is to integrate data from different kinds of experiments in order to gain a better understanding of regulatory mechanisms and therefore provide a global, unified view of regulatory genomics. The study of regulatory mechanisms and how alterations can lead to diseases have an immense impact on biomedical research. With the constant development of novel types of HTS experiments during the last few years, the need for a unified tool to perform HTS analyzes and manipulation has been growing [Hawkins et al., 2010]. At present, many large collaborative projects are based on HTS and produce a massive quantity of data, such as ENCODE, The Cancer Genome Atlas [Collins and Barker, 2007] and the 1000 Genome Project [1000 Genomes Project Consortium, 2010].

As shown in Figure 1.7, the amount of HTS data is actually growing much faster than the capacity of disc storage and microprocessors [Kahn, 2011]. Therefore an important property of a HTS analysis tool is the ability of dealing with data in a large-scale. As it facilitates the integration of results, it is of great advantage to be able to analyze different data types with one tool. The development of such a tool is not straightforward, as different data types result into different signals which have to be interpreted in a specific manner. For example, to detect protein-DNA binding sites, the task consists of finding significant peaks on ChIP-seq data. Many so-called ”peak-callers“ have been suggested to solve this problem [Zhang et al., 2008, Fejes et al., 2008, Nix et al., 2008, Park, 2009, Pepke et al., 2009].

Discovering protein-RNA binding sites from CLIP-Seq data is a different task, since signals have to be detected within transcripts, rather than across the genome and as reads derive from one RNA strand instead of from a double DNA strand (like from ChIP-Seq). Yeo and colleagues have developed a method that deals with analyzing CLIP-Seq data [Yeo et al., 2009].

Another challenge in HTS data analysis is to measure the expression of transcripts, for which RNA-Seq data is used. To assess differential gene expression between two cell lines, typically the reads within a genic region are counted and compared. Several statistical models have been proposed to evaluate significant differences [Anders and Huber, 2010, Wang et al., 2010, Robinson et al., 2010]. The variety of experiments based on HTS is growing and so is the need for tools to analyze the emerging data.

1.10 Chromatin and gene regulation

Chromatin is formed by nucleosomes and DNA, units in which DNA is wrapped about 1.6 times around a histone octamer (a pair of each H2A, H2B, H3 and H4) [Felsenfeld and Groudine, 2003, Richmond and Davey, 2003]. These histone octamers bind approximately every 200 base pairs [Kornberg and Thomas, 1974] to the DNA while they compact it, as shown in Figure 1.8. The positioning of nucleosomes has been shown to be dependent on the underlying sequence. In fact, Segal and colleagues suggested a genomic code for nucleosome positioning based on the DNA sequence [Segal et al., 2006]. However, less than 10% of all nucleosomes are well-positioned but interestingly show properties which distinguish them from the bulk nucleosomes [Nikolaou et al., 2010]. Apart from compaction, nucleosomes have been found to have further functionalities. They influence pre-transcriptional regulation when they compete for binding sites with transcription factors and other regulatory elements [Jiang and Pugh, 2009]. Furthermore, histone tails can be modified and, according to their local concentration and combination, they can contribute to gene activation or silencing, resulting into a so called histone code [Jenuwein and Allis, 2001]. This code generally distinguishes between the epigenetic states, known as euchromatin (”on“) and heterochromatin (”off“). These epigenetic states, in turn, are created, kept and inherited during cell division, representing a so-called ”cellular

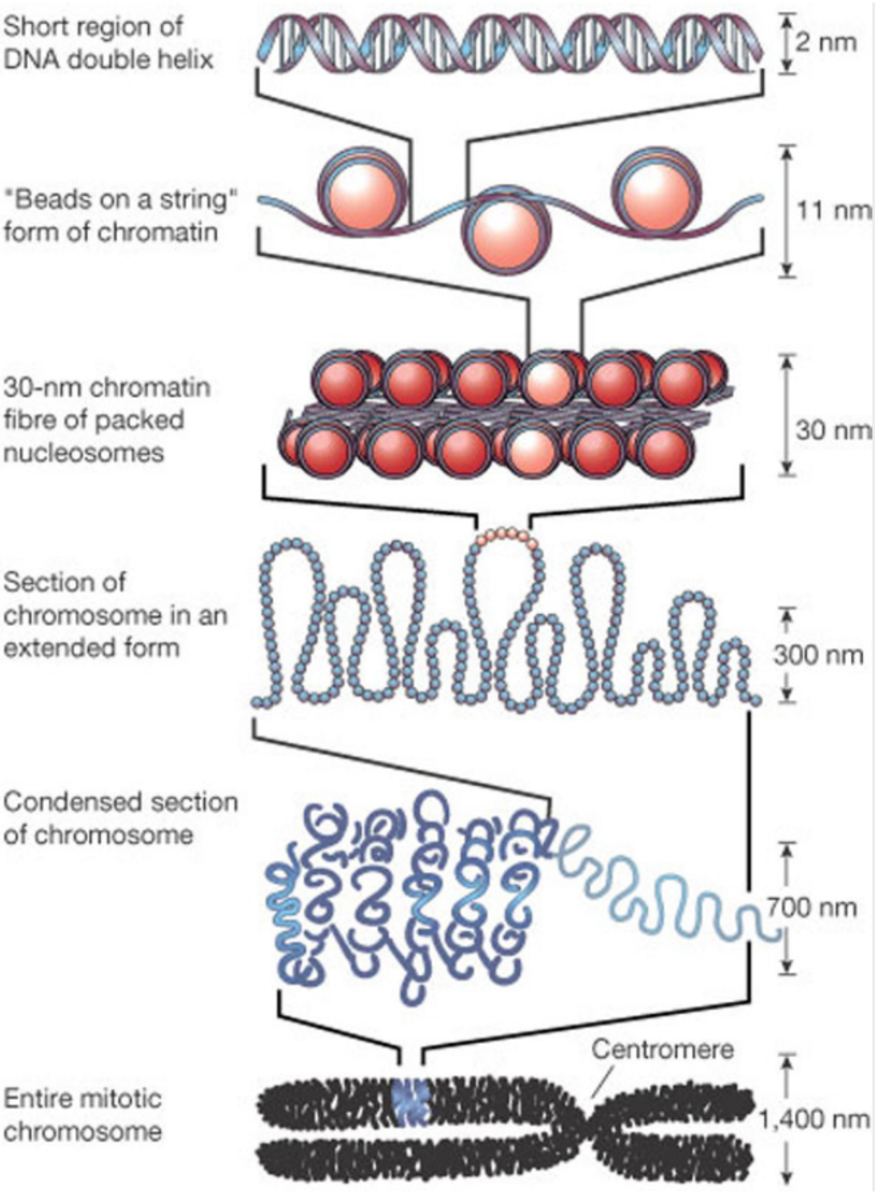

Short region of DNA double helix
2 nm
"Beads on a string" form of chromatin
11 nm
30-nm chromatin fibre of packed nucleosomes
30 nm
Section of chromosome in an extended form
300 nm
Condensed section of chromosome
700 nm
Centromere
Entire mitotic chromosome
1,400 nm

Figure 1.8: **The organization of DNA within the chromatin structure.** The lowest level of organization is the nucleosome, in which two superhelical turns of DNA are wound around the outside of a histone octamer. Nucleosomes are connected to one another by short stretches of linker DNA. At the next level of organization the string of nucleosomes is folded into a fiber about 30 nm in diameter, and these fibers are then further folded into higher-order structures. Adapted from Felsenfeld et al. [Felsenfeld and Groudine, 2003]

memory“.

Methylation and acethylation are the major types of histone modifications. They occur mostly on the lysine residues of the histone tail. In Figure 1.9 we show the densities of some well studied histone methylations and RNAPII around Transcription Start Sites (TSS) of genes with different expression states. For example RNAPII shows a significant peak at the TSS as well as higher densities downstream of TSS. Tri-methylation of H3 Lysine 27 (H3K27me3), a repressive mark, shows the highest densities around silent genes, while it is depleted around active genes. On the contrary, activating marks like methylations of H3 Lysine 4 (H3K4) are concentrated at or around the TSS of highly expressed genes and show very low occupancy in silent genes. The profile of tri-methylation of H3 Lysine 36 (H3K36me3), a transcription elongation mark, shows higher levels downstream of active genes, while it is depleted in silent genes. These associations of histone mark patterns with transcriptional activity motivated many groups to predict gene expression based on the combination of histone modifications around the TSS. Accordingly the application of different machine learning methods on histone patterns is now an active field of research. To give some examples, Yu et al. proposed a Bayesian network to derive relationships between histone modifications and gene expression [Yu et al., 2008]. At the same time Hon et al. applied an unsupervised learning method called ChromaSig to discover distinct clusters of chromatin signature [Hon et al., 2008]. Two years later a multivariate Hidden Markov Model that reveals discrete ”chromatin states“ was developed by Ernst and Kellis [Ernst and Kellis, 2010]. However, predictive methods so far have been focused on differences in gene expression considering only one cell line or tissue at a time, rather than on the relative differences between cell lines.

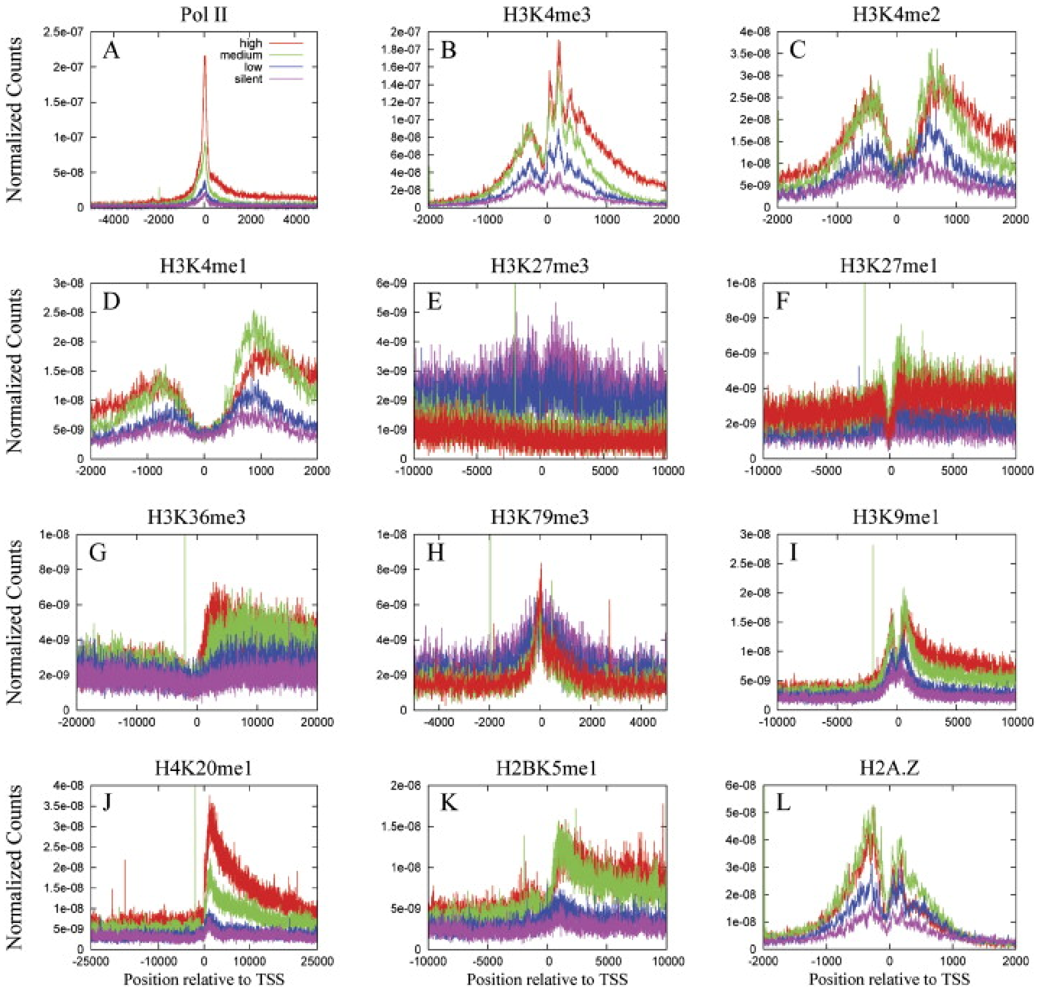


Figure 1.9: **Profile of ChIP-Seq data near TSS** (A)(L) Profiles of the methylated histones and RNA Polymerase II (PolII) indicated above each panel across the TSS for highly active, two stages of intermediately active and silent genes. Adapted from Barski et al. [Barski et al., 2007].

Part II

Objectives

Chapter 2

OBJECTIVES

The previous chapter aimed to give a general background about gene regulation studies. In the beginning, it described mechanisms of gene regulation. Next, it explained the experimental techniques for gene regulation studies, which are mostly based on High Throughput Sequencing (HTS) nowadays. It also pointed out the pitfalls of HTS data analysis before it finally explains what the integration of various data sets can achieve. The main objectives of this thesis can be summarized as follows:

1. The development of a versatile tool to analyze HTS data emerging from different types of experiments. Important properties of this tool are flexibility and memory efficiency, which are limitations of existing methods.

2. Applying the developed tool to elucidate the influence of progesterone on breast cancer cells.

3. Integrating data from different types of HTS experiments into a pipeline to gain insight into the epigenetic code that governs gene expression. The main goal here is to generate a predictive model that, based on signal changes from several epigenetic marks, predicts changes in gene expression between two conditions with high accuracy.

Part III

Results

Chapter 3

PYICOS: A VERSATILE TOOLKIT FOR THE ANALYSIS OF HIGH-THROUGHPUT SEQUENCING DATA

S. Althammer*, J. González-Vallinas*, C. Ballaré, M. Beato, and E. Eyras

* Both authors contributed equally to this work

Bioinformatics (Oxford, England), vol. 27, no. 24, pp. 3333-3340, Dec. 2011.

Althammer S, Gonzalez-Vallinas J, Ballare C, Beato M, Eyras E. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. Bioinformatics. 2011 Dec 15;27(24):3333-3340.

Chapter 4

STUDYING EFFECTS OF PROGESTIN IN BREAST CANCER CELLS

4.1 Studying effects of progestin in breast cancer cells

Within a collaborative project with the ”Chromatin and gene expression group” from the Center for Genomic Regulation in Barcelona, we applied Pyicos on ChIP-Seq data from the Progesterone Receptor (PR). Peaks of PR binding sites (PRbs) were obtained for samples with different times of progestin induction (0, 5, 30, 60 and 360 minutes) in T47D-MTVL breast cancer cells, as explained in methods. With ChIP followed by real time PCR, a number of regions were validated, and we obtained a set of 57 positive and 73 negative regions for PR binding. We considered these regions as a gold standard and compared the performance of Pyicos to those of MACS [Zhang et al., 2008] and MICSA [Boeva et al., 2010], two methods that have been specifically designed for peak calling on ChIP-Seq data. By calculating the receiver operating characteristics we show that Pyicos and MACS achieve very high accuracy, while they outcompete MICSA (Figure 4.1).

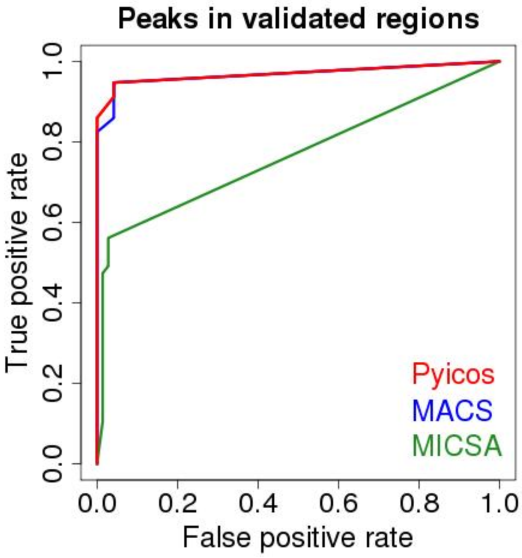


Figure 4.1: **Pyicos performance on regions validated for PR binding.** Comparison of the methods Pyicos, MACS and MICSA for peak calling on PR ChIP-Seq data. As a gold standard we used 57 positive and 73 negative binding regions, validated by ChIP-PCR.

The following analyzes were done using the induction time of 30 minutes. We calculated the peak densities in different regions related to genes and found that the region around the proximal promoter and the first exon were the most densely populated of all regions (see Figure 4.2 A). We further related the peaks to differential expression data from whole genome microarrays before and after progestin treatment. Figure 4.2 B demonstrates that genes that were up-regulated after stimulation showed a significantly higher density of PR binding sites (PRbs) in their surroundings, compared to down- or non-regulated genes. In fact, 66% of the up-regulated genes and 39% of down-regulated genes had at least one PRbs in the region starting 10 kilobases (kb) upstream and ending 5 kb downstream of a gene.

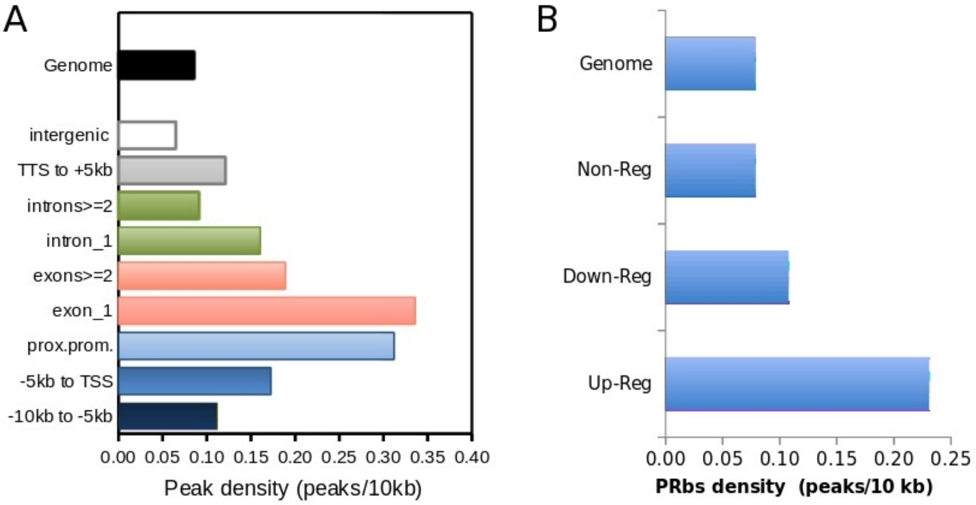


Figure 4.2: **Distribution of PRbs in breast cancer cells** A) We describe the distribution of PRbs in different regions related to genes. B) The density of PRbs is shown in the region starting 10kb upstream and ending 5 kb downstream of genes that are up-, down-, and non-regulated and along the genome.

In order to gain a better understanding of the regulatory mechanisms, we investigated the RNA-Polymerase II (RNAPII) binding in the same cell line. A ChIP-Seq experiment was performed on RNAPII, with different progestin induction times (0, 5 and 60 minutes). We explored how PR binding recruits RNAPII, by calculating the normalized number of reads from the different samples of RNAPII that overlap with PRbs on average. PRbs were separated according to the regulation of the nearest gene. In Figure 4.3 we see an elevation of RNAPII reads at PRbs related to up-, down- and non-regulated genes, even before hormone induction (0 minutes). The RNAPII peaks at PRbs related to up-regulated genes are highest at 5 minutes and reduce slightly at 60 minutes after hormone induction. Down-regulated genes show the highest RNAPII density after 5 minutes, and the lowest after 60 minutes. The RNAPII densities of 0 minutes and 60 minutes are very similar at PRbs related to non-regulated genes and slightly

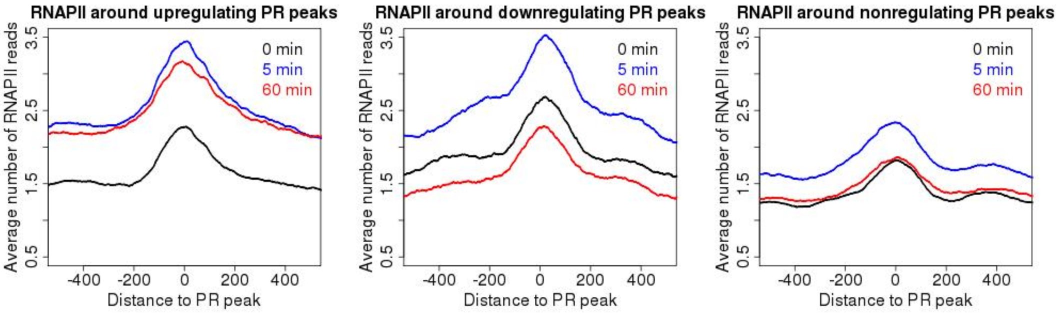


Figure 4.3: **RNAPII read density at PRbs** At PRbs that were related to up-, down- and non-regulated genes, we calculated read densities of RNAPII after different times of progestin induction.

elevated after 5 minutes of progestin induction. Overall we observe an increase of RNAPII after 5 minutes.

The PR binding in breast cancer cells described in this project, had an implication on further studies on gene regulation by hormones (see chapter 7.2 and Vicent et al. [Vicent et al., 2011]).

4.2 Methods

PR peak selection: PR ChIP-Seq reads were aligned to the reference genome (hg18) using GEM (http:// gemlibrary.sourceforge.net/) keeping only uniquely mapped reads with up to 2 mismatches. We used Pyicos to call significant peaks by performing the following steps:

1. We removed all reads falling in satellites or centromeric regions and we kept a maximum of 4 duplicates.

2. Reads were extended to 130 base pairs, in agreement with the fragment selection

3. The control (0 minutes) was subtracted from the normalized ChIP-Seq reads and the resulting effective reads were clustered.

4. Possible artifacts were removed. We considered two types of artifacts: peaks shorter than 100 base pairs, which may appear as an artifact after subtraction; and block-like clusters, which may appear from PCR artifacts. A block-like cluster was defined as such if the ratio between the length covered by the maximum of the peak and the length of the peak was greater than 0.25.

5. We further split the peaks if the read coverage goes below 5% of the peak summit, which is defined as the midpoint of the region of maximum coverage in the peak.

6. Poisson analysis was applied using peak-height on the resulting read-clusters and selected those that are significantly higher than the average height of the peaks in a chromosome (p-value of at most 0.001).

Chapter 5

PREDICTIVE MODELS OF GENE REGULATION FROM HIGH-THROUGHPUT EPIGENOMICS DATA

This article has been submitted for publication to the special issue ”Epigenetic Control of Reprogramming and Cellular Differentiation“ of ”Comparative and Functional Genomics“.

Predictive models of gene regulation from high-throughput epigenomics data

Sonja Althammer[1], Amadís Pagès[1], Eduardo Eyras[1,2,*]

[1]Computational Genomics Group, Universitat Pompeu Fabra, Dr. Aiguader 88, E08003 Barcelona, Spain
[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, E08010 Barcelona, Spain

[*]To whom the correspondence should be addressed: eduardo.eyras@upf.edu

Keywords: chromatin, machine learning, regulation, epigenetic code, ChIP-Seq, high-throughput sequencing

Abstract

The epigenetic regulation of gene expression involves variations in histone marks, DNA methylation, as well as changes in the protein-DNA binding activity of multiple factors. The synergistic or antagonistic action of these changes have suggested the existence of an epigenetic code for gene regulation. High-throughput Sequencing (HTS) datasets provide an opportunity without precedent to explore this code and to build quantitative models of gene regulation based on epigenetic changes between specific cellular conditions.

We describe a new computational framework that facilitates the systematic analysis of HTS epigenetic data for integrative studies. Our framework allows the storage of processed data for easy querying and for performing Machine Learning analyses. Our method relates epigenetic signals to expression by comparing the same genomic locus between two conditions, instead of comparing genomic loci to each other, thereby mitigating various biases from HTS experiments. We show the effectiveness of this methodology by building a model of gene regulation, using epigenetic data from the ENCODE project, which predicts with high accuracy whether a gene has a significant increase, decrease or no change in expression between two cell lines.

Our analyses indicates that the epigenetic code is quite degenerate and involve multiple regions of genes. We find that removing ambiguous signals from overlapping genes the prediction accuracy improves considerably, the attributes discriminate better between regulatory classes and the model built from one pair of cell lines perform with high accuracy in a second pair. Moreover, signal changes at the 1st exon, 1st intron and downstream of the polyadenylation site for specific histone marks are found to associate strongly with expression regulation. Our analyses also show a different epigenetic code for expression for intron-less and intron-containing genes, with more prominent difference for genes with low GC content in the promoter.

We are thus able to build a quantitative description of gene regulation based on epigenetic changes, which captures the relative importance of the location of epigenetic signals along the gene, thereby providing further evidence for a generic epigenetic code of expression regulation. Our computational framework provides a general methodology to do integrative analysis of epigenetic changes measured from HTS experiments, which can be applied to other studies, like cell differentiation or carcinogenesis.

INTRODUCTION

DNA associates with histone proteins to conform the chromatin [1]. Histones generally carry post-transcriptional modifications in cells and these modifications modulate the expression of genes [2], [3]. For instance, there is a genome-wide relation between the histone 3 Lysine 36 trimethylation (H3K36me3) and transcription activity [4], [5]. This and other epigenetic modifications are key to cellular differentiation [6] and their alterations have been associated to early stages of cellular transformation in tumors [7], [8]. The combinations of the histone modifications, which can have cooperative or opposed effects on the chromatin state, have been proposed to reflect a histone code that would determine the regulation of gene expression and the cell state [9]. High-throughput sequencing (HTS) technologies provide a very effective way to obtain information about the histone modification patterns at genome wide scale [10]. Efforts to integrate available genome-wide data sets about chromatin in various conditions are crucial towards improving our understanding of the role of epigenetics in gene regulation.

Recent publications have made progress in the definition of a histone code of gene expression by generating predictive models of transcriptional activity based on histone mark information [11–16]. They provide insights into possible mechanisms of regulation and a formal description of the postulated histone code [17], [18]. These methods generally relate the histone signals obtained from experiments of chromatin immunoprecipitation followed by HTS (ChIP-Seq) [19], with a read-out of the gene expression based on expression microarrays or HTS for mRNAs (RNA-Seq) [20]. In the present work, we include additional epigenetic data to extend this relation, namely, HTS of DNase I hypersensitive sites (DNase-Seq) [21], DNA methylation data [22], and ChIP-Seq data for CTCF and RNA-Polymerase II.

In previous approaches, a chromatin signal is generally represented in terms of read-counts or peak significance in the promoter and sometimes in the gene body of genes [23]. However, these methods are generally based on one single condition or cell line. Thus, they effectively compare the properties of different genes in a direct way; hence, the accuracy of the predictive model will be dependent on the accuracy of the estimation of the significance of the ChIP-Seq signals and, more importantly, they rely on the premise that signals in two different genes should be comparable. However, genes present many variable properties, like the number of introns or the presence of CpG islands in their promoter, that may affect these measurements. For instance, recent experiments show that the splicing machinery can recruit histone-modifying enzymes and influence the chromatin state, with the consequence that genes with introns tend to have higher levels of H3K36me3 signal [24]. Thus, the comparison of genes with and without introns is not straight forward. Additionally, various other factors may affect the local density of HTS signal [25]. For instance, the tag counts from a HTS experiment will be influenced by the chromatin structure of the DNA and by shearing effects [26–28], not all regions have the same mappability [29] and there is often a GC bias in the reads [30]. These issues will reflect on differences in coverage between regions, which will be even more exacerbated for the broad signals that are obtained for histone ChIP-Seq experiments. Control samples can partly alleviate this, but their effectiveness depends very much on the sequencing depth. Thus, HTS signals from two genes are not directly comparable in general,

We propose a new method to measure epigenetic signals and to relate them to expression based on the comparison between two conditions. In our approach, the same genomic locus is compared between two conditions; hence, the predictive model describes changes of gene expression in terms of changes in epigenetic mark densities between two conditions or cell types. Significance of these changes is calculated taking the read density into account, thereby mitigating the confounding effects mentioned earlier. Additionally, our method has the advantage of providing a continuous description of the changes, rather than an on-off state description.

To illustrate our method, we have built a model of expression regulation from epigeneticn changes using data from various ENCODE cell lines [31]. Our results show a different epigenetic code for expression for intron-less and intron-containing genes, being this difference more prominent in genes with low GC content around the transcription start site. Moreover, eliminating anti-sense transcription and overlapping promoters and tails from different genes, which has not been done before, the prediction accuracy improves considerably. Furthermore the predictive model built from one pair of cell lines performs with high accuracy in a different pair. Finally, we are able to generate a minimal code for expression regulation between two cell lines that is generic enough to correctly predict the regulatory outcome of up to 70% transcripts from a different pair of cell-lines.

MATERIAL AND METHODS

Genomic annotations

For our analyses we used the gene set from the 7th release of the GENCODE annotation (ftp://ftp.sanger.ac.uk/pub/gencode/release_7/gencode.v7.annotation.gtf.gz), which is based on the assembly GRCh37 (hg19) and is included in the Ensembl release 62 [32]. All transcripts encoded at each gene loci and the genomic region defined by them, which we name transcript loci, were considered initially. Those transcript loci from chromosome M and of biotype “pseudogene” were removed for the analysis.

We separated transcript loci into four groups; according to whether they were intron-containing (IC) or intron-less (IL), and according to whether they had a promoter with high CG (HCG) or low CG (LCG) content. We classified transcripts as HCG if the region of 4kb centred on the transcription start site (TSS) overlaps at least 200 bp with a CpG island, and LCG otherwise. CpG island annotations where obtained from the UCSC Table Browser (hg19) [33]. In order to obtain balanced sets for training and testing, an equal number of up- (Up) and down- (Dw) regulated transcripts were selected from each of the four groups. These groups were taken to be as large as possible, but such that the p-value of significance (Benjamini-Hochberg corrected) for the expression change for each transcript was smaller than 0.05. Furthermore, the same number of non-regulated (Nr) transcripts were selected. These are defined to have the highest p-values and sufficient expression, i.e. the density of reads measured in RPKM (reads per kilobase per million mapped reads as defined by [20]) was greater than 1 in a cell line from the pair. With this, we obtained four different sets(Table 1). As part of our analyses, we also filtered overlapping transcript loci that would make ambiguous the assignment of the marks with the correct expression change. We considered the following cases (Supplementary Figure 1):

- transcript loci that overlap in opposite strands.
- transcript loci whose promoters (2kb) overlap in opposite strands
- transcript loci whose tails (2kb) overlap in opposite strands

- transcript loci with overlapping promoter (2kb) and tail (2kb) on the same strand
- Overlapping transcript loci on the same strand but from different genes

Transcript-loci Set	Description	Pair1 - all	Pair1 - filtered	Pair2 -all	Pair2 - filtered
HCG IC	High CG promoter and intron-containing	6510	1959	2964	792
HCG IL	High CG promoter and intron-less	105	27	24	12
LCG IC	Low CG promoter and intron-containing	6705	1767	1980	585
LCG IL	Low CG promoter and intron-less	84	30	15	15

Table 1: Each of the four sets of transcript loci considered in our analysis. The numbers correspond to the loci before (all) or after (filtered) eliminating overlapping loci (Methods). From each set, we considered up-, down- or non-regulated transcript loci, each corresponding to 1/3 of the indicated numbers.

Datasets

We downloaded ChIP-Seq data for RNA Polymerase II (RNAPII), CCCTC-binding factor (CTCF) and various Histone marks (Table 2), data for DNase I hypersensitive sites (DNase-Seq), methylation data from Reduced Representation Bisulfite Sequencing (Methyl-RRBS) and RNA-Seq data from the Encode project (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/) for four cell lines: a chronic myelogenous leukemia line (K562), a lymphoblastoid line (GM12878), a human mammary epithelial line (HMEC) and a muscle myoblast line (HSMM) (Table 2). We considered two pairs of comparisons, P1: K562 vs GM12878 and P2: HSMM vs HMEC. We selected experiments that were available in at least four cell lines, except for RNAPII, which was only available in two of the selected cell lines. For all datasets, we used only reads that did not contain any uncalled bases (N). Moreover, for ChIP-Seq and DNase-Seq reads we kept only reads with mapping quality greater than 30. The Methyl-RRBS data was filtered for positions covered by at least 10 reads. The mean methylation of a region was defined to be the proportion of methylated sites over the total number of probed sites in that region. Further, we obtained the RPKMs for the RNA-Seq data for the individual transcript loci directly from Encode (http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/ /wgEncodeCshlLongRnaSeq/releaseLatest/).

For our analysis we considered for each transcript locus, a number of regions related to its exon-intron structure (Table 3). Subsequently, for each one of these regions and for each experimental data set, the z-score for the enrichment was calculated between a pair of cell lines using Pyicos [34]. The calculation was based on 2 replicas in one condition (K562 or HSMM) and 1 replica in the other condition (GM12878 or HMEC). Further, pseudocounts and RPKM normalization were used (details in Supplementary Material). These z-scores constitute the set of attributes that were used for Machine Learning (ML) analyses and corresponds to each region-experiment pair. As a control, random attributes were generated for each region by random sampling z-score values from all attributes for that region type.

Unless otherwise stated, accuracies of the models were measured calculating the average area under the receiver operating characteristic (ROC) curve (AUC) for a 10-fold cross validation. A ROC curve relates the rates of true positives (TPs) and false positives (FPs) produced by the model. The larger the area described by the ROC curve (AUC) the better the overall accuracy of the model. AUC = 1 indicates a model that predicts no false positives and all true cases correctly, and AUC = 0.5 indicates that the model is equivalent to random. In 10-fold cross validation, the data is split into 10 subsets and 10 evaluations are carried out iteratively, where in each iteration 9 subsets (nine-tenths of the instances) are used for training and one subset for testing. This method ensures that all instances are used for the evaluation and the overall accuracy is averaged over the ten iterations, so that it represents the mean behaviour of the model.

	Pair 1		Pair 2	
Cell lines	K562	GM12878	HSMM	HMEC
factor/mark				
CTCF	BROAD	BROAD	BROAD	BROAD
H3K27ac	BROAD	BROAD	BROAD	BROAD
H3K27me3	BROAD	BROAD	BROAD	BROAD
H3K36me3	BROAD	BROAD	BROAD	BROAD
H3K4me1	BROAD	BROAD	BROAD	BROAD
H3K4me2	BROAD	BROAD	BROAD	BROAD
H3K4me3	BROAD	BROAD	BROAD	BROAD
H3K9ac	BROAD	BROAD	BROAD	BROAD
H4K20me1	BROAD	BROAD	BROAD	BROAD
RNAPII	UT-A	UT-A	--	--
DNase-Seq	UW	UW	UW	UW
Methyl-RRBS	HA	HA	HA	HA
RNA-Seq	CSHL	CSHL	CSHL	CSHL

Table 2: ENCODE datasets and cell lines used for analysis: ChIP-Seq data for RNA Polymerase II (RNAPII), CTCF and various Histone marks, data for DNase I hypersensitive sites (DNase-Seq), methylation data from Reduced Representation Bisulfite Sequencing (Methyl-RRBS) and sequencing of long polyA+ whole cell RNA (RNA-Seq). For HMEC and HSMM cells RNAPII ChIP-Seq data was not available at the time of our analyses. Datasets were generated at the Broad Institute (BROAD), Cold Spring Harbor Laboratory (CSHL), University of Washington (UW), University of Texas at Austin (UT-A) and Hudson Alpha (HA).

Type	Region	Description
Fixed-length regions	Promoter 2kb	Region starting 2kb upstream of the transcription start site (TSS) and ending 1bp before the TSS;
	Promoter 5kb	Region starting 5kb upstream of the TSS and ending 1bp before the TSS;
	TSS +/- 2kb	Region starting 2kb upstream of the TSS and ending 2kb downstream
	TSS +/- 5kb	Region starting 5kb upstream of the TSS and ending 5kb downstream
	pA +/- 2kb	Region starting 2kb upstream of the pA and ending 2kb downstream
	tail	Region starting 1bp after the pA and ending 2kb downstream
Variable-length regions	First exon	Region corresponding to the first exon of the transcript locus
	First intron	Region corresponding to the first intron of the transcript locus
	GB	Gene body, i.e. region between the TSS and the poly-adenylation site (pA) of an annotated transcript locus
	GB3'ss	Region between the first 3' splice-site and the pA of an annotated transcript locus
	GB +/- 1kb	Gene body with additional 1kb stretches up- and downstream
	GB +/- 5kb	Gene body with additional 5kb stretches up- and downstream
	GB + 5kb	Gene body with an additional 5kb stretch downstream of the pA

Table 3: Regions considered per transcript locus for the calculation of the different attributes.

RESULTS AND DISCUSSION

A framework for integrative epigenetic studies

Our computational framework addresses three fundamental tasks in the process of acquiring knowledge: data mining, data manipulation, and data analysis, and is comprised of the following steps (i) an analysis pipeline to systematically identify the changes in expression and epigenetic signals between two conditions in multiple genomic regions, (ii) an automatic way to store the results in a Biomart system [35] for easy querying and filtering and (iii) a connectivity to the application WEKA [36], to allow the application of Machine Learning (ML) methods for creating predictive models of gene regulation.

In order to relate epigenetic signals to expression regulation, our method measures signal changes between two conditions rather than the signal level in one single condition. With this methodology, relative changes of the epigenetic state can be related to each other or to the relative change of expression. By considering relative signal changes, biases from HTS are mitigated. To verify this, we checked whether selecting significant regions according to RPKM densities or z-scores from our method would be biased by the GC content. We therefore considered the top 10% of genes in terms of the H3K4me3 RPKM (K562) in the gene body and found a Spearman correlation of 0.34 with GC content. However, selecting the top 10% of genes according to absolute z-scores for H3K4me3, given by the comparison between K562 and GM12878, resulted in no correlation with GC content (Spearman 0.02). Thus relating RPKM values to gene expression could result into false positives due to GC bias. When we repeated the same calculation on the 4kb region centered on the TSS, none of the two measures, RPKMs or z-scores, showed a GC bias (correlation coefficient of -0.02 and 0.05, respectively). As H3K4me3 is mostly distributed around the TSS [10] we deduce that in this case the real signal obscures the bias, while in the gene body, where no strong signal for H3K4me3 is present, the bias dominates over the signal.

We have developed an automatic pipeline that, given a set of regions and a number of high-throughput sequencing (HTS) datasets for two conditions, can systematically calculate the log-rate of change for each region and its significance in terms of a z-score (details in Supplementary File). Datasets used for this work are accessible in a Biomart database at http://regulatorygenomics.upf.edu/group/pages/software. We have modified Biomart so that datasets can also be exported as ARFF (attribute-relation file format), which can be uploaded directly to the WEKA system [36], a collection of open-source machine learning algorithms for data mining tasks, issued under the GNU General Public License. Our system thus provides the possibility of using own custom data to train models and evaluate different ML algorithms for the study of mechanisms of gene regulation.

In order to illustrate the potential of our framework we analyzed high-throughput sequencing (HTS) data from ENCODE [31] (Methods). We started by systematically calculating the changes between cell lines in pair P1 (K562 vs GM12878) and pair P2 (HSMM vs HMEC) for all the experiments in a variety of regions related to the transcript loci (Table 3). Most of the recently developed predictive methods use signals in the promoter region of genes or in a window around the transcription start site (TSS). We also included the gene body, as recent evidence suggests that the signal along this region will be informative as well [23]. Besides promoter, TSS and gene body regions, we also include a region for the 1st exon, the 1st intron, and the gene body downstream of the 1st intron, which have been shown to contain relevant chromatin signatures for transcriptional regulation [24], [37], [38], and have not been used before in a predictive model. We further considered additional windows around and beyond the poly-adenylation site (pA), resulting in a total of 13 different regions (Table 3) (Supplementary Figure 1). Accordingly, for the two pairs of cell lines P1 and P2, we had a total of 13 x 12 = 156 and 13 x 11 = 143 (as RNAPII was not available for P2) attributes per transcript locus, respectively, where each attribute is defined by the z-score of the enrichment value between the two cell lines for a region-experiment pair.

As classification value, we used expression information from RNA-Seq experiments from ENCODE in the corresponding cell lines. For each pair of cell lines we calculated the transcripts with significant increase (Up) or decrease (Dw) of expression. In order to build a predictive model of expression that can distinguish between either type of regulation (Up or Dw) and no change, we also considered non-regulated (Nr) transcripts, defined to have sufficient expression level and no significant change in expression between the same pair of cell lines (Methods).

Recent studies have shown that introns may influence the transcriptional regulation of genes [24], [38]. Therefore, we separated our transcripts sets according to whether they were intron-containing (IC) or intron-less (IL). Furthermore, several studies have highlighted that human promoters present different regulation according to their CG content [39–41]. Thus, we further split the sets according to whether a 4kb region centered on the TSS overlaps with a CpG island or not, resulting in high CG content (HCG) or low CG content (LCG) sets (Methods). Finally, in order to have a balanced set for training and testing, we selected from each type the same number of transcripts for each regulatory class (Table 1).

A generic epigenetic code for gene expression regulation

Using the datasets processed as above, we built a highly accurate and generic predictive model of gene expression changes based on epigenetic data. We tried various ML models to predict the three possible classes, up (Up), down (Dw) and non-regulated (Nr), and decided to use a Random Forest model [42], as it showed the best performance using 10-fold cross validation (data not shown). Table 4 shows the accuracies of this model tested on intron-containing sets for various training conditions. Remarkably, we obtain a higher accuracy for the LCG set than for the HCG set (Table 4). Incidentally, CpG-related genes are quite often housekeeping genes [43], which has been pointed out before as one of the reasons why predictions perform differently on each set [44]. According to this, LCG transcripts should be more frequently associated to genes with differential expression (Up or Dw). This is confirmed in our analysis, as we found that the performance was always higher for the prediction of Up and Dw loci than for non-regulated transcripts (Table 4). For intron-less (IL) loci we found the opposite behaviour, i.e. HCG-IL has higher accuracy than LCG-IL (Supplementary Table 1).

A)

Before filtering

	HCG - IC				***LCG – IC***			
Attributes	Up	Dw	Nr	Average	Up	Dw	Nr	Average
P1 (with RNAPII)	0.8	0.79	0.74	0.78	0.82	0.87	0.78	0.83
P1	0.79	0.79	0.74	0.77	0.83	0.86	0.76	0.82
P1 (CFS)	0.8	0.79	0.74	0.78	0.82	0.86	0.76	0.81
P2	0.85	0.83	0.81	0.83	0.9	0.88	0.83	0.87
P2 (CFS-P1)	0.85	0.83	0.8	0.83	0.9	0.88	0.83	0.87
P1-on-P2	0.83	0.77	0.63	0.74	0.88	0.83	0.71	0.81
P1(CFS)-on-P2	0.83	0.8	0.57	0.73	0.88	0.84	0.74	0.82

B)

After filtering

	HCG - IC				***LCG – IC***			
Attributes	Up	Dw	Nr	Average	Up	Dw	Nr	Average
P1 (with RNAPII)	0.79	0.84	0.76	0.8	0.85	0.9	0.81	0.86
P1	0.79	0.82	0.75	0.79	0.86	0.89	0.76	0.84
P1 (CFS)	0.79	0.81	0.73	0.78	0.84	0.9	0.77	0.84
P2	0.89	0.88	0.85	0.87	0.92	0.91	0.85	0.89
P2 (CFS-P1)	0.87	0.87	0.84	0.86	0.92	0.92	0.86	0.9
P1-on-P2	0.89	0.87	0.7	0.82	0.92	0.89	0.79	0.87
P1(CFS)-on-P2	0.85	0.82	0.68	0.78	0.91	0.89	0.81	0.87

Table 4: We show the accuracy in terms of the area under the ROC curve (AUC) for the 10-fold cross validation for the IC transcript sets for various training conditions. The results are shown for all the transcript loci before **A)** and after **B)** filtering for the overlaps in opposite strands and overlaps of promoters and tails (Methods). P1 (with RNAPII) corresponds to pair P1 with the additional RNAPII attribute, i.e. the same attributes as P2 plus RNAPII. P1 and P2 denote the models for each cell line pairs with all the attributes. P1(CFS) and P2(CFS) denote the models for P1 and P2, respectively, where the attributes used are those that have a score 80 or higher (maximum 100) using the CFS attribute selection method independently for P1 and P2. P2 (CFS-P1) indicates that the model was trained using the data from P2 but the attributes selected using CFS on P1. P1-on-P2 indicates that the model was trained with pair P1 with all attributes and tested on pair P2. P1(CFS)-on-P2 indicates that the model was trained with pair P1 with only selected attributes and tested on pair P2.

Interestingly, training a model for the first pair with (Table 4A, P1 (with RNAPII)) or without RNAPII data (Table 4A, P1) yields very similar accuracy for all sets, which suggests that the information provided by RNAPII is redundant with the histone data for prediction. Indeed, looking at the pairwise correlations of all marks for P1, separated per region and per transcript set (Figure 1 and Supplementary Figure 3), we observe a high correlation of the z-scores for RNAPII with most of the other signals (H3K36me3, DNase-Seq, CTCF, H3K4me2, H3K9ac, H3K27ac and H3K4me3).

With the aim of obtaining a minimal set of attributes that are sufficient to attain high prediction accuracy, we applied Correlation-based Feature Selection (CFS) [45]. This method works by iteratively testing subsets of attributes, retaining those that best correlate with the class values (Up, Dw or Nr) and removing those that have high redundancy. In this way, a minimal set of non-redundant attributes with optimal performance is selected. We applied CFS to the data from both pairs of cell lines and selected attributes that were selected in at least 80% of the validation rounds (Table 4A, P1(CFS) and P2(CFS)). Interestingly, CFS provided attributes related to all the regions (Supplementary Table 2A), indicating that histone marks along all regions of the transcript locus may be relevant for regulation. Additionally, the prediction accuracy did not suffer, while the model is simplified by removing redundant attributes (Table 4A, P1(CFS)).

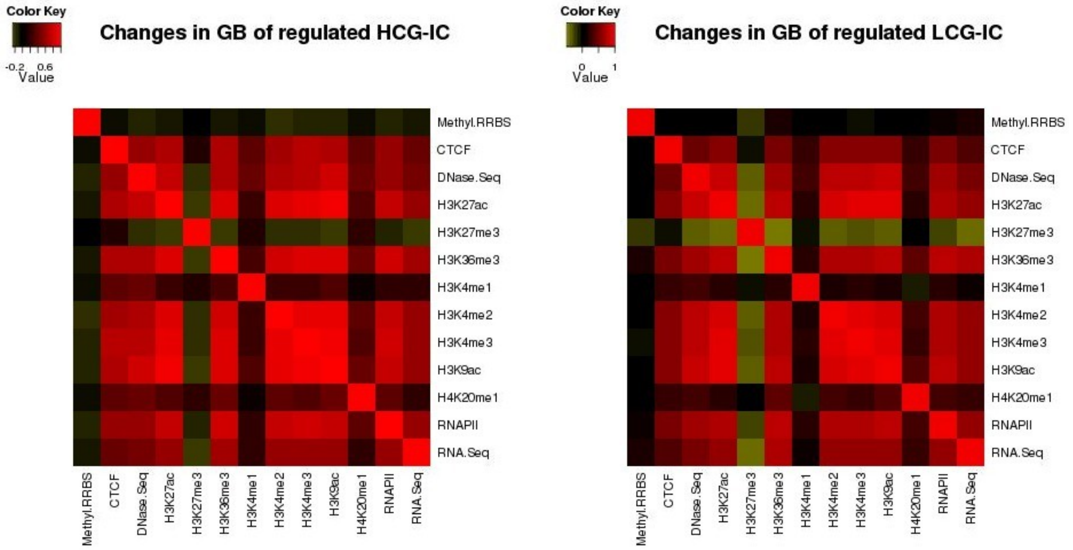


Figure 1: Pairwise correlations of marks and expression changes in gene bodies. Heatmaps are shown for regulated genes from the filtered intron-containing (IC) sets for high (HCG) and low (LCG) CpG promoters. The color represents the value of the Pearson correlation coefficient between the z-scores for every pair of attributes. For expression (RNA-Seq), the z-scores of the Up and Dw transcript loci were used to calculate the correlation.

With the aim of obtaining a generic epigenetic code of expression regulation, we decided to compare the attributes obtained from P1 with the attributes obtained for a second pair of cell lines (P2). Although CFS applied to both pairs, P1 and P2, yields a different set of optimal attributes, with only between 26% and 50% of coincidences between them (Supplementary Table 2), a model built on P2 with the attributes selected from P1 shows a high accuracy, which is comparable to the original model on P1 (Table 4A P2(CFS-P1)). That is, qualitatively, the attributes relevant for one pair of cell lines seem to be also relevant for the other one.

To test the generality of the model also in quantitative terms, i.e. in terms of the actual numerical model, we applied directly on P2 the model built from P1. However, this test across pairs did not achieve an accuracy as high as before (Table 4A, P1-on-P2 and P1(CFS)-on-P2). We hypothesized that the reduction of accuracy in the test across pairs comparison could be due to differences in the homogeneity of cell lines, which would produce a very variable pattern of signals. Alternatively, this lack of reproducibility could stem from the overlap of the gene body, promoters or tails from transcript loci from different genes, specially in the opposite strand, which would make ambiguous the association of the epigenetic signal change to a specific expression change. Accordingly, we removed from the training set those transcripts loci where the signal in one region could not be unambiguously assigned (Methods) (Supplementary Figure 4), thereby generating filtered sets for training and testing (Table 1). Interestingly, after removing these cases we observe a consistent increase in the accuracy of the prediction in all groups (Table 4B), with 60-78% of the instances correctly classified (Table 5).

We further explored whether the signals in one single region would be sufficient to predict the expression outcome. Accordingly, for each region we selected the common attributes from pairs P1 and P2 with CFS score ≥ 80% (Supplementary Table 3). Interestingly, the marks selected for a single region give a prediction accuracy that is comparable to that obtained with attributes from all regions (Supplementary Table 4). The highest accuracy was achieved using gene body +/-5kb, which is not surprising as it overlaps all the other regions. Interestingly, the 2kb region downstream of the pA, a region that has not been considered before, turns out to have a high predictive power, achieving an AUC of 0.89 for up-regulated IC-LCG transcripts based only on the signals for H3K27me3 and H3K36me3. Remarkably, one single mark in the region pA +/-2kb is enough to predict up-regulated genes with high accuracy (AUC = 0.85 and 0.81 for Up in IC-LCG and IC-HCG transcripts, respectively). As before, the models achieve higher AUCs for LCGs than for HCGs.

Attributes	Transcript loci set	Instances in total	Correctly classified instances
P1(CSF)	LCG-IC	1767	1185 (67.06 %)
	HCG-IC	1959	1182 (60.34 %)
P2(CSF-P1)	LCG-IC	585	454 (77.60 %)
	HCG-IC	792	577 (72.85 %)
P1(CSF)-on-P2	LCG-IC	585	410 (70.09 %)
	HCG-IC	792	445 (56.19 %)

Table 5: Correctly classified instances in each transcript subset. Sets are filtered to avoid overlapping gene bodies, promoters or tails from transcript loci from different genes in the same or opposite strands (Methods). Attribute selection has been applied to each pair: P1(CFS) and P2(CFS), for each of the subsets of intron-containing loci, high (HCG) or low (LCG) CG content promoter. The attribute sets correspond to the ones from Table 4B: P1(CFS) denotes the model for P1, where the attributes used are those that have a score 80 or higher (maximum 100) using the CFS attribute selection method. P2 (CFS-P1) indicates that the model was trained using the data from P2 but the attributes selected using CFS on P1. P1(CFS)-on-P2 indicates that the model was trained with pair P1 with only selected attributes and tested on pair P2.

The relative contribution of marks to the epigenetic code

With the aim to find the most relevant attributes that appear to determine the regulation of expression, we calculated the information gain (IG) [46] for all attributes in the subsets HCG-IC and LCG-IC on pair P1 for the unfiltered and the filtered sets (Table 1). The higher the IG value, the better the attribute can separate the three classes: Up, Dw and Nr. As a control, we generated random attributes for each region, obtained by random sampling z-score values from all attributes in that region. In Figure 2 and Supplementary Figure 5 we show how attributes rank in terms of IG within each region. Although the ranking is very similar before and after filtering transcript loci, we found an overall increase in IG values, indicating that the filtering step improves the specificity of the regulatory code. We found that for all subsets, H3K36me3 is the most informative attribute around the pA site and in gene body associated regions, whereas H3K27ac and H3K9ac are most informative in the promoter region, which agrees with previous analyses [47]. These two acetylation marks are in fact among the most informative marks in the promoter, around the TSS and in 1st intron and 1st exon regions. Interestingly, H3K36me3 is more informative in the 1st intron than in the 1st exon, which agrees with recent results relating H3K36me3 with splicing of the first intron [24]. Although methylation data shows anti-correlation with expression change in the promoter of HCG loci (Supplementary figure 2), we observe a modest contribution in the gene body to expression regulation (Figure 1 and 2).

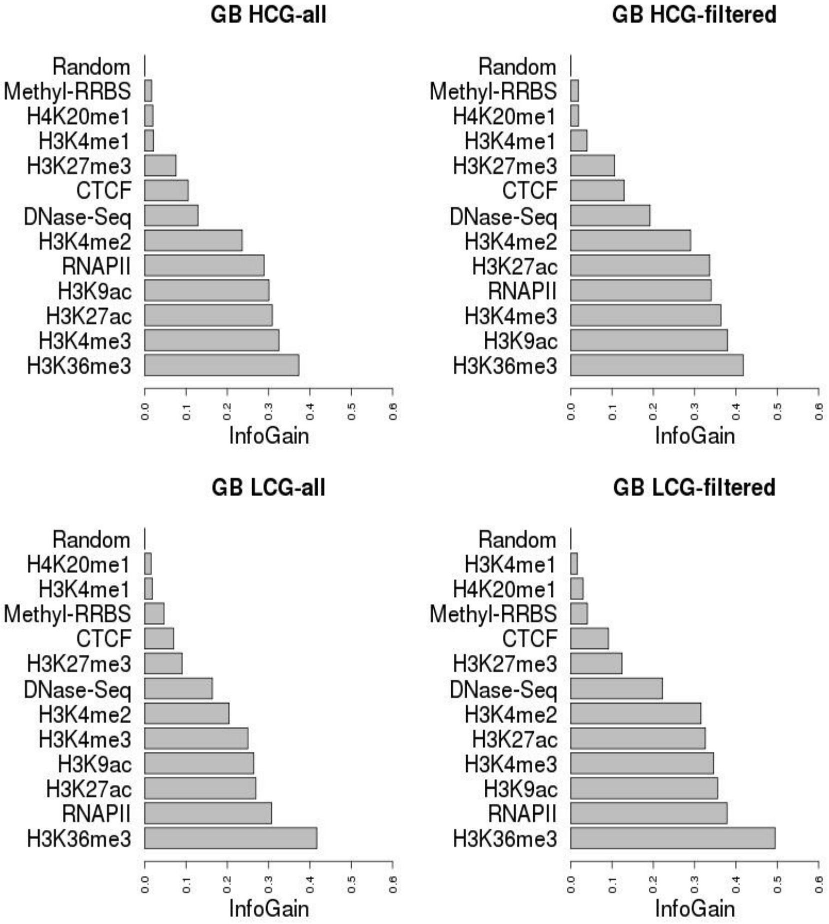


Figure 2: Information gain values measured for attributes in the gene body of intron-containing (IC) transcript loci, comparing before and after filtering loci according to overlap with transcripts from different genes (Methods). Data is shown for high (HCG) and low (LCG) CpG promoters. Random attributes generated by random sampling z-score values from all attributes in a given region are shown as a control.

Although IG values determine how well an attribute separates the three sets, Up, Dw and Nr, we would expect that attributes that most directly associate with expression changes should show no change for the Nr set. That is, we should expect that the enrichment z-scores for Nr should distribute around zero. Accordingly, we defined an attribute to be optimal if the absolute vallue of the median for the Nr distribution is smaller than 0.1 and the IG is greater than 0.05. If more than one attribute accomplish these thresholds, we considered the one with the highest IG value. Interestingly, this analysis shows that the optimal attributes for H3K36me3 and H3K4me3 correspond to the 1st intron and 1st exon, respectively (Figure 3), which could be related to their role in the coupling between splicing and transcription [24], [38]. Moreover, for H3K9ac and H3K27ac the optimal attributes are the TSS-5kb and Promoter-5kb regions, respectively (Figure 3). DNase-Seq also presented the optimal distribution in the 1st exon, whereas CTCF and H3K4me2 were best in the GB-5kb region (Figure 3).

We did not find an optimal attribute for RNAPII, as the attribute for the gene body, the region with minimal median for the Nr distribution and largest IG (Supplementary Figure 6A), shows an enrichment for Nr similar to the Up subset, which could be due to an excess of RNAPII reads in one of the cell lines (Supplementary Figure 6B). Even though we could not find an optimal attribute H3K27me3, the z-score distributions for the 1st exon results into a clear trend that agrees with the anti-correlation of H3K27me3 and expression (Figure 3), despite the low IG (0.03): Up genes show almost no change, whereas Dw genes show the greatest enrichment (Figure 3), possibly indicating that there is an asymmetry in the pattern of this histone mark for silencing. We also did not find optimal attributes for Methyl-RRBS, H3K4me1 and H4K20me1. For Methyl-RRBS this is probably due to a large proportion of sites with reads but no methylation evidence (data not shown). The most informative region with minimal median for Nr for H3K4me1 indicates an enrichment of Up in GB +/- 5kb but a distribution for Dw and Nr centered on zero, indicating an asymmetry in transcriptional activation. Although H4K20me1 has been related to silent chromatin [48] the most informative of the attributes showed almost no difference between Up, Dw and Nr subsets. The absence of an optimal attribute for H3K4me1 in GB +/-5kb and for H4K20me1 in the 1st exon might be due to an unequal distribution of reads in K562 relative to GM12878, which does not occur for H3K27me3.

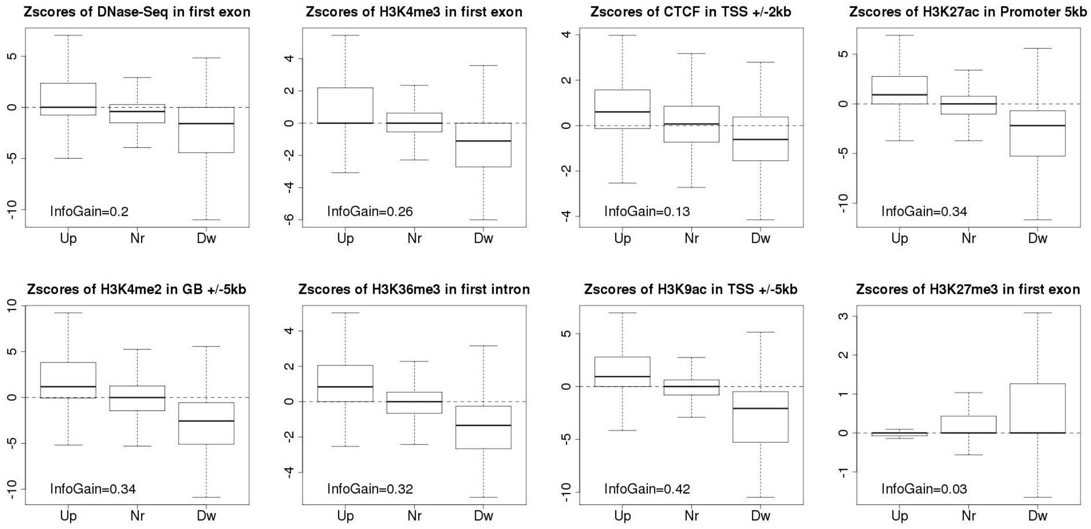


Figure 3: Distribution of z-scores for up- (Up), down- (Dw) and non- (Nr) regulated genes for the optimal attributes for each experiment, calculated by maximizing the Information Gain and minimizing the absolute value of the median for the z-score distribution of the Nr subset. The y-axis shows the z-score corresponding to the enrichment of the attribute. These distributions correspond to the set of LCG-IC loci of Pair1.

The effect of introns in the epigenetic code

A number of specific histone modifications have been related to the co-transcriptional splicing of introns [24], [38]. We therefore hypothesized that there should be relevant differences in the histone modifications between IC and IL loci. We thus compared the most informative attributes between intron-containing (IC) and intron-less (IL) loci (Figure 4 and Supplementary Figure 6). As there were many more IC than IL loci, we selected a subset of loci from IC of the same size as IL and compared the IG values for attributes related to fixed-length regions (Table 3). For HCG loci, although we found almost no differences when we ranked the attributes according to IG (Figure 4, Supplementary Figure 7), there is an overall reduction of the IG values in IL genes. Strikingly, we found that for LCG loci the IG becomes very small for most of the attributes. For instance, in the promoter region, most of the attributes that are informative for LCG-IC loci do not contribute at all in LCG-IL (Figure 4), and H3K36me3, which is considered most relevant downstream of the TSS, and H3K4me1, become the most informative attribute for LCG-IL loci (Figure 4). Similarly, in the tail regions most of the attributes that are informative for LCG-IC loci do not contribute for LCG-IL loci (Supplementary Figure 7), where the IG values are very low. In contrast, the tail region behaves more similarly for HCG-IC and HCG-IL, in terms of ranking and IG. This indicates that LCG-IL loci may be regulated through changes in different epigenetic signals not considered in this analysis.

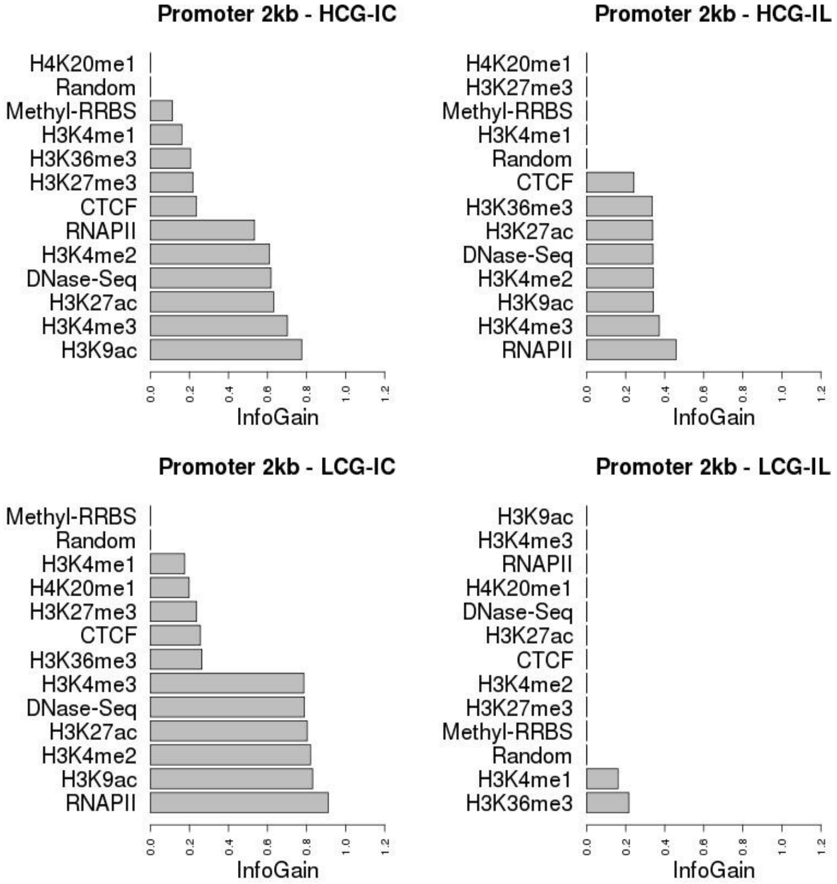


Figure 4: Information gain values measured for attributes in the 2kb promoter region, comparing intron-less (IL) genes with intron-containing (IC) genes before filtering transcripts (Methods). The compared sets were taken to be of the same size (105 transcripts for HCGs and 84 transcripts for LCGs).

CONCLUSIONS

A current challenge in epigenetics is how to extract biological knowledge from large volumes of data produced with new high-throughput technologies. Integrative tools and Machine Learning (ML) algorithms are crucial to this aim. In this article we have described a novel computational framework for the integration of high-throughput sequencing (HTS) epigenetic data that facilitates the generation and testing of quantitative models of gene regulation. Our methodology proposes a new way to relate epigenetic signals to expression using the comparison of the same locus between two conditions, instead of comparing loci to each other in a single condition, which can be affected by various biases. Three novel aspects of our methodology are that it 1) considers continuous values for the change in epigenetic signals, 2) it explores the enrichment of signals in multiple regions and 2) it can be applied to any HTS data type in two conditions.

We have shown the effectiveness of this methodology by building a predictive model of gene expression regulation based on epigenetic information for a pair of cell lines from the ENCODE project. The processed data used to build the models in this paper is available as a Biomart database at http://regulatorygenomics.upf.edu/software/. Our quantitative models can predict whether a gene shows expression differences (up or down) or no difference between two cell lines. The relevant attributes and the accuracy for each model vary according to whether transcript loci have high CpG-content promoters (HCG) or not (LCG) and whether they contain introns (IC) or not (IL). These differences indicate that the histone signals are very heterogeneous and that regulation depends strongly on the actual structural properties of promoters and genes. Our analyses also indicate that there is high redundancy in the histone code, as different groups of attributes from different regions can explain a similar number of regulatory events.

Additionally, we have taken into account a fact largely overlooked in previous publications, which is that a considerable number of gene loci overlap with each other [49] at promoter and tail regions, or over their gene bodies, either on the same or on opposite strands. Accordingly, previous models of expression based on histone marks have this confounding effect, since the strand-less ChIP-Seq signal cannot be unambiguously associated to the regulation of a specific gene. Interestingly, when we removed these overlapping genes, the prediction accuracy improves considerably and the predictive model built from one pair of cell lines perform with high accuracy in a second pair of different cell lines. We conclude that removing these overlapping loci allow us to build a more general epigenetic code for expression regulation. This is further confirmed by our analysis of the Information Gain (IG), which shows that attributes can separate better the three regulatory classes after the overlapping loci are removed. Notably, this filtering does not change the ranking of IG values, hence although we improve the quantitative description of the histone code, the qualitative description does not change. The IG analysis confirms the role of some of the histone marks, like H3K9ac and H3K27ac, in the promoter and around the transcription start site in expression regulation as described before in the literature; and uncovers new regions, like the first intron for H3K36me3, the first exon for H3K4me3 and downstream of the polyadenylation site for H3K36me3, where changes in these marks associate strongly with expression regulation. The role of these marks in the first exon and intron indicates a general role in the coupling between splicing and transcription, as recently shown in the literature. In this direction, we also explored the patterns of epigenetic changes between intron-containing (IC) and intron-less (IL) loci and found that IC loci contain more epigenetic information and can therefore be better characterised. These differences are more remarkable between high (HCG) and low CpG promoters (LCG), which suggests that the type of promoter might influence the epigenetic changes that take place in co-transcriptional splicing [24]. Alternatively, this could indicate that these loci have a distinct mode of regulation, possibly by other marks that have not been considered in this study.

The epigenetic signals analysed in this study provide a strong prediction power for expression regulation. However, the associations found do not necessarily imply causality or a direct functional effect, as the effect of a given histone mark may be context dependent and may occur through the action of other factors. Nevertheless, the models described reflect the complex network of gene regulation and provide some of the generic features of this network. Our methodology provides an effective way to integrate the continuous changes in epigenetic signals between different conditions. Applying this approach to datasets with more histone modifications and transcription factors will help expanding and characterizing further this complex regulatory network. In particular, the application of our approach to different developmental stages, disease states or treatments, will help uncovering the epigenetic mechanisms responsible for cellular differentiation and carcinogenesis.

Acknowledgements

The authors would like to thank Juan González-Vallinas, Eneritz Agirre, Nicolas Bellora, João Curado and Alba Jené for useful discussions and Julien Lagarde for the ENCODE data support.

Funding: Generalitat de Catalunya by FI grant (to S.A.); MICINN grant BIO2008-01091 (to E.E.); European Commission grant EURASNET-(LSHG-CT-2005-518238) (to E.E.).

References

[1] R. D. Kornberg and J. O. Thomas, “Chromatin structure; oligomers of the histones,” *Science (New York, N.Y.)*, vol. 184, no. 4139, pp. 865-868, May 1974.

[2] B. Li, M. Carey, and J. L. Workman, “The role of chromatin during transcription,” *Cell*, vol. 128, no. 4, pp. 707-719, Feb. 2007.

[3] J. Mellor, P. Dudek, and D. Clynes, “A glimpse into the epigenetic landscape of gene regulation,” *Current Opinion in Genetics & Development*, vol. 18, no. 2, pp. 116-122, Apr. 2008.

[4] D. K. Pokholok et al., “Genome-wide map of nucleosome acetylation and methylation in yeast,” *Cell*, vol. 122, no. 4, pp. 517-527, Aug. 2005.

[5] A. A. Joshi and K. Struhl, “Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation,” *Molecular Cell*, vol. 20, no. 6, pp. 971-978, Dec. 2005.

[6] R. Lister et al., “Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells,” *Nature*, vol. 471, no. 7336, pp. 68-73, Mar. 2011.

[7] L. Ellis, P. W. Atadja, and R. W. Johnstone, “Epigenetics in cancer: targeting chromatin modifications,” *Molecular Cancer Therapeutics*, vol. 8, no. 6, pp. 1409-1420, Jun. 2009.

[8] M. Kulis and M. Esteller, “DNA methylation and cancer,” *Advances in Genetics*, vol. 70, pp. 27-56, 2010.

[9] T. Jenuwein and C. D. Allis, “Translating the histone code,” *Science (New York, N.Y.)*, vol. 293, no. 5532, pp. 1074-1080, Aug. 2001.

[10] A. Barski et al., “High-resolution profiling of histone methylations in the human genome,” *Cell*, vol. 129, no. 4, pp. 823-837, May 2007.

[11] H. Yu, S. Zhu, B. Zhou, H. Xue, and J.-D. J. Han, “Inferring causal relationships among different histone modifications and gene expression,” *Genome Research*, vol. 18, no. 8, pp. 1314-1324, Aug. 2008.

[12] G. Hon, B. Ren, and W. Wang, “ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome,” *PLoS Computational Biology*, vol. 4, no. 10, p. e1000201, Oct. 2008.

[13] B. van Steensel, U. Braunschweig, G. J. Filion, M. Chen, J. G. van Bemmel, and T. Ideker, “Bayesian network analysis of targeting interactions in chromatin,” *Genome Research*, vol. 20, no. 2, pp. 190-200, Feb. 2010.

[14] R. Karlić, H.-R. Chung, J. Lasserre, K. Vlahovicek, and M. Vingron, “Histone modification levels are predictive for gene expression,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 7, pp. 2926-2931, Feb. 2010.

[15] J. Ernst and M. Kellis, “Discovery and characterization of chromatin states for systematic annotation of the human genome,” *Nature Biotechnology*, vol. 28, no. 8, pp. 817-825, Aug. 2010.

[16] C. Cheng and M. Gerstein, “Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells,” *Nucleic Acids Research*, vol. 40, no. 2, pp. 553-568, Jan. 2012.

[17] T. Jenuwein and C. D. Allis, “Translating the histone code,” *Science (New York, N.Y.)*, vol. 293, no. 5532, pp. 1074-1080, Aug. 2001.

[18] B. D. Strahl and C. D. Allis, “The language of covalent histone modifications,” *Nature*, vol. 403, no. 6765, pp. 41-45, Jan. 2000.

[19] G. Robertson et al., “Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing,” *Nature Methods*, vol. 4, no. 8, pp. 651-657, Aug. 2007.

[20] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nature Methods*, vol. 5, no. 7, pp. 621-628, Jul. 2008.

[21] A. P. Boyle et al., “High-resolution mapping and characterization of open chromatin across the genome,” *Cell*, vol. 132, no. 2, pp. 311-322, Jan. 2008.

[22] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch, “Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis,” *Nucleic Acids Research*, vol. 33, no. 18, pp. 5868-5877, 2005.

[23] S. A. Hoang, X. Xu, and S. Bekiranov, “Quantification of histone modification ChIP-seq enrichment for data mining and machine learning applications,” *BMC Research Notes*, vol. 4, p. 288, 2011.

[24] S. F. de Almeida et al., “Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36,” *Nature Structural & Molecular Biology*, vol. 18, no. 9, pp. 977-983, Sep. 2011.

[25] M.-S. Cheung, T. A. Down, I. Latorre, and J. Ahringer, “Systematic bias in high-throughput sequencing data and its correction by BEADS,” *Nucleic Acids Research*, vol. 39, no. 15, p. e103, Aug. 2011.

[26] L. Teytelman et al., “Impact of chromatin structures on DNA processing for genomic analyses,” *PloS One*, vol. 4, no. 8, p. e6700, 2009.

[27] R. K. Auerbach et al., “Mapping accessible chromatin regions using Sono-Seq,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 35, pp. 14926-14931, Sep. 2009.

[28] P. J. Park, “ChIP-seq: advantages and challenges of a maturing technology,” *Nature Reviews. Genetics*, vol. 10, no. 10, pp. 669-680, Oct. 2009.

[29] J. Rozowsky et al., “PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls,” *Nature Biotechnology*, vol. 27, no. 1, pp. 66-75, Jan. 2009.

[30] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, “Substantial biases in ultra-short read data sets from high-throughput DNA sequencing,” *Nucleic Acids Research*, vol. 36, no. 16, p. e105, Sep. 2008.

[31] R. M. Myers et al., “A user’s guide to the encyclopedia of DNA elements (ENCODE),” *PLoS Biology*, vol. 9, no. 4, p. e1001046, Apr. 2011.

[32] P. Flicek et al., “Ensembl 2012,” *Nucleic Acids Research*, vol. 40, no. Database issue, pp. D84-90, Jan. 2012.

[33] T. R. Dreszer et al., “The UCSC Genome Browser database: extensions and updates 2011,” *Nucleic Acids Research*, vol. 40, no. Database issue, pp. D918-923, Jan. 2012.

[34] S. Althammer, J. González-Vallinas, C. Ballaré, M. Beato, and E. Eyras, “Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data,” *Bioinformatics (Oxford, England)*, vol. 27, no. 24, pp. 3333-3340, Dec. 2011.

[35] A. Kasprzyk, “BioMart: driving a paradigm change in biological data management,” *Database: The Journal of Biological Databases and Curation*, vol. 2011, p. bar049, 2011.

[36] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, “Data mining in bioinformatics using Weka,” *Bioinformatics (Oxford, England)*, vol. 20, no. 15, pp. 2479-2481, Oct. 2004.

[37] J. T. Huff, A. M. Plocik, C. Guthrie, and K. R. Yamamoto, “Reciprocal intronic and exonic histone modification regions in humans,” *Nature Structural & Molecular Biology*, vol. 17, no. 12, pp. 1495-1499, Dec. 2010.

[38] R. J. Sims 3rd et al., “Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing,” *Molecular Cell*, vol. 28, no. 4, pp. 665-676, Nov. 2007.

[39] P. Carninci et al., “Genome-wide analysis of mammalian promoter architecture and evolution,” *Nature Genetics*, vol. 38, no. 6, pp. 626-635, Jun. 2006.

[40] S. Saxonov, P. Berg, and D. L. Brutlag, “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*, vol. 103, no. 5, pp. 1412-1417, Jan. 2006.

[41] E. Valen and A. Sandelin, “Genomic and chromatin signals underlying transcription start-site selection,” *Trends in Genetics: TIG*, vol. 27, no. 11, pp. 475-485, Nov. 2011.

[42] L. Breiman, “Random Forests,” vol. 45, no. 1, pp. 5-32, 2001.

[43] J. Schug, W.-P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert Jr, “Promoter features related to tissue specificity as measured by Shannon entropy,” *Genome Biology*, vol. 6, no. 4, p. R33, 2005.

[44] Z. Zhang and M. Q. Zhang, “Histone modification profiles are predictive for tissue/cell-type specific expression of both protein-coding and microRNA genes,” *BMC Bioinformatics*, vol. 12, p. 155, 2011.

[45] M. Hall, “Correlation-based feature selection for machine learning.,” *PhD Thesis. New Zealand: Department of Computer Science, Waikato University;*, 1999.

[46] T. Mitchell, “Machine Learning,” *The Mc-Graw-Hill Companies, Inc.*, 1997.

[47] Z. Wang et al., “Combinatorial patterns of histone acetylations and methylations in the human genome,” *Nature genetics*, vol. 40, no. 7, pp. 897-903, Jul. 2008.

[48] J. K. Sims, S. I. Houston, T. Magazinnik, and J. C. Rice, “A trans-tail histone code defined by monomethylated H4 Lys-20 and H3 Lys-9 demarcates distinct regions of silent chromatin,” *The Journal of Biological Chemistry*, vol. 281, no. 18, pp. 12760-12766, May 2006.

[49] S. Katayama et al., “Antisense transcription in the mammalian transcriptome,” *Science (New York, N.Y.)*, vol. 309, no. 5740, pp. 1564-1566, Sep. 2005.

Part IV

Discussion

Chapter 6

DISCUSSION

With the recent advent of High-Throughput Sequencing (HTS) methods, there is clearly a need for tools to assist downstream analyzes. Due to the constantly growing diversity of experiments based on HTS, tools to analyze the emerging data are required to be flexible. The mechanisms of gene regulation represent a complex network of many different factors. Thus we need to integrate data from various sources to gain insight into these mechanisms. Chapter 3 (a published article) and chapter 5 (a submitted article) address these issues, while chapter 4 shows an example of the efficiency of the tool Pyicos, described in chapter 3. In this chapter we point out the aims we have been pursuing, discuss critically the results of the articles and stress unsolved problems. Moreover, we describe limitations of the involved methods and suggest possibilities for improvement.

6.1 Pyicos: A versatile toolkit for the analysis of high-throughput sequencing data

With the aim of flexibly manipulating and analyzing HTS data from multiple sources, we developed the modular toolkit Pyicos. The choice of the word toolkit reflects the fact that one can use many basic operations of Pyicos separately, which provides the possibility to manipulate HTS reads from different sources. This gives a great advantage compared to other published methods, as it allows to explore the signals emerging from a newly developed HTS method. With the constant development of novel experiments that are based on HTS data, this represents a highly useful feature. However, in this article we focused on the typical applications, for which other methods have been available for comparison.

For protein-DNA binding site detection we developed a protocol called callpeaks, as a reference to the comparable methods, known as peak-callers. The compared methods and Pyicos perform similarly in terms of resulting peak quality and peak definition. Unlike other authors, we proved that the basic operations of Pyicos are useful, as they can improve the peak quality, measured as the number of peaks that contain the expected motif. In particular we showed how the peak calling can be improved when 1) duplicated reads are removed, 2) a control is subtracted or 3) the peaks are split. Removing duplicated reads is necessary to mitigate a bias from the amplification that is required for current sequencing techniques. However, with the amplification-free library preparation these artificial duplicates vanish. Another bias emerges from the preferred fragmentation of open chromatin, which can be diminished by subtracting a control. Finally, binding sites might be close to each other, and their corresponding peaks can thus appear to be merged. In order to detect these multiple binding sites, Pyicos provides an operation to split peaks. Unlike other methods, Pyicos is a user-friendly toolkit that allows customized analysis and can thus improve the results. In particular we demonstrated an improvement of the peak ranking on the progesterone receptor binding data when peak height instead of read-count was chosen for scoring. Moreover, the detection of CTCF binding sites was improved when peaks were split. ChIP-Seq data from different factors seem to vary in the properties of the resulting peaks, thus we need to be able to perform analysis in a flexible way.

Next, we applied the protocol for differential expression analysis on RNA-Seq data on liver and kidney samples, and compared the results to those of other methods, using the corresponding microarray data as a benchmarking set. While maintaining a similar accuracy, Pyicos shows some advantages compared to the methods DEGseq, DESeq and edgeR. With its accurate performance on simulated replicas, Pyicos provides the possibility to analyze numerous data sets that have been produced without replica. Moreover, Pyicos offers the choice to normalize the data by the number of reads in a sample or apply TMM-normalization. Additionally it provides a normalization by the length of the gene. However there are other normalization methods, which are not included into Pyicos. The issue of data normalization is not solved yet and we hope that further improvements in this area will mitigate the biases that might obscure biological meaningful findings.

We have further demonstrated Pyicos flexibility by defining protein-RNA binding sites from CLIP-Seq data and we have provided the corresponding protocol to the public. It showed a similar accuracy to the only method that was published at that time. Being aware of the constant growth of HTS data sets that is outpacing improvements of microprocessors, we developed Pyicos to minimize RAM usage, while it maintains reasonable running time.

6.2 Studying effects of progestin in breast cancer cells

In chapter 3 we already showed on ChIP-Seq data from PR that Pyicos can detect Progesterone Receptor binding sites (PRbs) with superior accuracy compared to other methods. Thus we used it to identify PRbs in breast cancer cells after different times of hormone induction. To evaluate Pyicos performance on this data set, we calculated overlaps of the resulting peaks with regions validated by ChIP-PCR.

We showed that the accuracy of Pyicos is similar or superior to those of other methods. In particular, its accuracy is comparable to that of MACS, a popular peak caller. The poor performance of MICSA on this data set might be explained by the integrated motif search that weights the resulting p-values of the peaks.

MICSA overvalues the importance of the motif occurrence and thus detects less true positive binding sites.

With the aim of determining the targets of the PR, we measured the density of predicted PRbs across the genome. When we analyzed the distribution of PRbs, we found that the density is the highest around Transcription Start Sites (TSS). Notably PRbs are more densely occupying the surrounding of the TSS of genes that are up-regulated upon hormone treatment. This suggests a relation of PR binding to the regulation of transcription.

In order to quantify the recruitment of RNA-PolymeraseII (RNAPII) at PRbs, we explored the corresponding ChIP-Seq data, on different incubation times of progestin. After 5 minutes of induction we see an overall increase of RNAPII at PRbs, surprisingly also at PRbs that are related to down-, and non-regulated genes. A possible explanation could be that the RNAPII is binding, but not active. In order to confirm this, one could do a ChIP-Seq experiment for the phosphorylation of serine 2 within the RNAPII C-terminal domain, as it is known to be active [B. Alberts and Raff, 2008]. Moreover, at 5 minutes we also observe the highest peaks in regulated genes, compared to other time points. Up-regulation seems to be taking effect already after 5 minutes, while down-regulation occurs later, as RNAPII density increases after 5 minutes and decreases only later.

This study required the integration of data emerging from multiple experiments, such as ChIP-Seq for PR and RNAPII, and gene expression microarray. Thus it is a good example of how data integration can provide a global view and elucidate regulatory mechanisms.

6.3 Predictive models of gene regulation from high-throughput epigenomics data

The main aim in this project was to explore the relationship between epigenetic changes and differential gene expression. Therefore we run enrichment analysis on various epigenetic data sets and RNA-Seq data from ENCODE in multiple regions related to a transcript loci. To do so, we embedded Pyicos into a pipeline that automatically runs enrichment analysis for all combination of data sets and regions . Providing results to other scientists is as important as producing them, hence we loaded the processed data into a Biomart database that can be publicly accessed. From there the data can be downloaded or directly uploaded to the WEKA system for data mining and machine learning purposes. The ENCODE HTS data sets are widely used, as they offer high quality and are openly available for the community. Thus we aimed to facilitate further integration of the processed data.

With the aim of exploring which mark or factor is related most strongly to expression change in each region, we calculated the corresponding information gains. As the targets of modifications are not thoroughly described yet, we tried to be as exhaustive as possible by taking into account multiple regions related to a transcript loci. This gave us the possibility to describe for all those different regions, the mark or factor with the strongest relation to changes in gene expression. For instance, in regions related to the genebody or to the polyadenylation site the tri-methylation of H3 Lysine 36 showed the highest information gain.

In order to explore the power of epigenetic variations to predict the regulatory outcomes of gene expression between two cell lines, we generated a predictive model. We further tested this model for accuracy and generality. After having shown a high accuracy of expression change predictions between the cell lines K562 and Gm12878 (training set), we could prove that the model trained on this data was generic enough to perform accurate predictions on Hsmm and Hmec (test set), a different pair of cell lines. Hence, we have shown that the created model is accurate and captures general features of the regulatory network.

In order to define a minimal code that is able to describe the complex network of gene regulation, we removed redundant and confounding features. In order to choose the most informative features, while minimizing redundancy, we applied correlation feature selection on the training set. The generality of the selected features was proven by applying the same selection on the features of the test set, and performing predictions with reasonable accuracy.

In order to explore the heterogeneity of the epigenetic code across sets of loci with different structural properties, we created models while we distinguished between intron containing (IC) and intronless (IL) loci, as well as between loci with high (HCG) and low CpG-content (LCG) promoters. The predictive models came out to vary in the relevance of the attributes and the accuracies. These differences suggest a heterogeneity of the epigenetic code that depends on the actual structure of gene and promoter. In particular, we found higher prediction accuracies for IC-LCG than for IC-HCG. As HCG loci are generally related to housekeeping genes and LCG to tissue specific genes, we hypothesized that there would be stronger variations in LCG loci, which makes it easier to predict changes. However, the opposite behavior was found in IL genes: Accuracies for HCG are higher than those for LCG loci. IL genes seem to be regulated in a different manner. In spite of reasonable accuracies for LCG (0.71 or 0.72 on average, depending on the selected features), we observed very low values or 0 for information gain of the features. We hypothesized that IL-LCG might be regulated by other epigenetic signals that have not been captured by this study.

In order to adapt more accurately the behavior of the epigenetic changes, we filtered the loci for overlaps before generating the predictive models. Overlapping loci imply that some of the signals cannot be assigned unambiguously and thus obscure downstream analyzes, a fact that has been largely overlooked when models of gene expression were generated. In fact, by excluding these loci, we observe an over all increase in prediction accuracy. Especially for the test set the accuracy rises considerably, which confirms that the overlapping loci were confounding the models. With the filtered data the models achieved an accuracy of up to 0.92 (for up-regulated LCG-IC) on the test set.

In summary, we demonstrated the strong predictive power of epigenetic changes, however we cannot distinguish between causality and consequence in the found associations.

6.4 Limitations and future directions

At present, limitations are represented by the state of the technology and the methods. The current development of HTS instruments that produce longer reads, promises less false positive read mappings and can thus improve downstream analyzes, especially for repetitive parts of the genome.

Using our pipeline, we have associated epigenetic variations to expression changes on ENCODE data sets and provided the processed data to the public. By doing so, we greatly hope that it will be related to further genomic and epigenomic information produced by other labs. Embedding improved normalization methods into our pipeline, like the conditional quantile normalization [Hansen et al., 2012], could further improve results by mitigating the GC-bias of RNA-Seq reads.

As we recently found evidence for a relation between nucleosome positioning and exon recognition [Tilgner et al., 2009], we believe that a similar approach could be used to associate epigenetic changes to exon inclusion or skipping.

Another challenge is to relate epigenetic variations to the expression of a diseased state. Earlier, Altshuler and colleagues have successfully associated variations in the genomic sequence with diseases [Altshuler et al., 2008]. To complex association studies, the next step is to find epigenome-wide associations. This is a rather complicated study, as there is at least one different epigenome for each cell-type and condition. However several projects, e.g. the NIH Roadmap Epigenomics Mapping Consortium [Bernstein et al., 2010], the ENCODE project [ENCODE Project Consortium, 2007] and the International Cancer Genome Consortium ([Consortium et al., 2010], already aim to provide data from human epigenetic marks in a large scale. Finally the comparison of epigenetic information from normal and diseased states gives us great hope to identify disease mechanisms, which is the first step towards finding successful treatments.

Part V

Conclusions

Chapter 7

CONCLUSIONS

The work carried out during this PhD project can be summarized as follows:

1. The development of Pyicos, a powerful toolkit for the analysis of high-throughput sequencing (HTS) data. Pyicos offers versatility, flexibility and efficient memory usage. Its high accuracy was demonstrated by a comparison to published methods that have been specifically developed for certain analyzes. In particular, we showed that Pyicos calls peaks on ChIP-Seq data with a similar or better accuracy than MACS, USeq and FindPeaks. Moreover, it can be used to process CLIP-Seq data. We further applied Pyicos for differential expression analysis on RNA-Seq data, where it showed a comparable performance to those of DEGseq, DESeq and edgeR.

2. Pyicos has been applied to ChIP-seq data of progesterone receptor in breast cancer cells. We described the distribution of active progesterone binding sites upon hormone induction and related them to the corresponding microarray expression data and ChIP-Seq data from RNA Polymerase II, in order to gain insight into the regulation of gene expression by hormones. This work has further contributed to a manuscript that will be submitted for publication (see chapter 7.2).

3. Embedding Pyicos into a pipeline that integrates HTS data emerging from different experiments. In this context the enrichment analysis of Pyicos was applied to determine the significance of signal changes from various HTS experiments between two cell lines. For this analysis we used ENCODE data sets from four cell lines, doing pairwise comparisons of signals in K562 vs Gm12878 and Hsmm vs Hmec. We provided the processed data in a public Biomart database.

4. Generation of generic models that accurately predict changes in gene expression based on epigenetic changes. We further selected a minimal set of features to describe the epigenetic code. The predictive models have been generated taking into account the structural differences of loci and promoters, resulting into different models for intron containing and intronless loci, or high CpG-content and low CpG-content promoters. Finally, prediction accuracy was be improved by the elimination of ambiguous signals.

Part VI

Appendix

Chapter 8

APPENDIX

8.1 Supplementary material for chapter 3

Supplementary Material

Pyicos: A versatile toolkit for the analysis of high-throughput sequencing data

Sonja Althammer[1,‡], Juan González-Vallinas[1,‡], Cecilia Ballaré[2], Miguel Beato[1,2], Eduardo Eyras[1,3,*]

[1]Universitat Pompeu Fabra. Dr. Aiguader 88, E08003, Barcelona, Spain

[2]Centre for Genomic Regulation (CRG). Dr. Aiguader 88, E08003 Barcelona, Spain

[3]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, E08010, Barcelona, Spain

‡ These authors contributed equally

* Corresponding author

Althammer S, Gonzalez-Vallinas J, Ballare C, Beato M, Eyras E.Supplementary material: Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. Bioinformatics. 2011 Dec 15;27(24):3333-3340.

8.2 Nucleosome driven transcription factor binding and gene regulation

Cecilia Ballaré , Laura Gaveglia, Giancarlo Castellano, Sonja Althammer, Juan González-Vallinas, Eduardo Eyras, Francois LeDily, Roser Zaurin, Guillermo P. , Vicent, Miguel Beato1*

*Corresponding author: miguel.beato@crg.es (tel +34 933 161 119)

Abstract

Elucidating the global function of a transcription factor implies identification of its genomic binding sites. The role of chromatin in this context is unclear, but the dominant view is that factors bind preferentially to nucleosome-depleted regions. In contrast, we find that the progesterone receptor (PR) needs nucleosomes for optimal genome binding and function. In breast cancer cells we identified 4,000 hormone regulated genes and 25,000 genomic PR binding sites (PRbs), the majority encompassing several copies of the hexanucleotide TGTYCY (the progesterone responsive element, PRE). Strong functional PRbs are enriched around the 5-end of up-regulated genes, overlap with DNaseI hypersensitive sites, and exhibit high nucleosome occupancy. Hormone treatment results in remodeling of these nucleosomes and MNase cleavage. Conversely, weak PRbs and PREs that do not bind PR are not enriched in nucleosomes, suggesting a crucial role of nucleosomes for PR binding and hormonal gene regulation.

8.3 Supplementary material for chapter 5

Supplementary File

Predictive models of gene regulation from high-throughput epigenomics data

Sonja Althammer[1], Amadís Pagès[1], Eduardo Eyras[1,2,*]

[1]Computational Genomics Group, Universitat Pompeu Fabra, Dr. Aiguader 88, E08003 Barcelona, Spain

[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, E08010 Barcelona, Spain

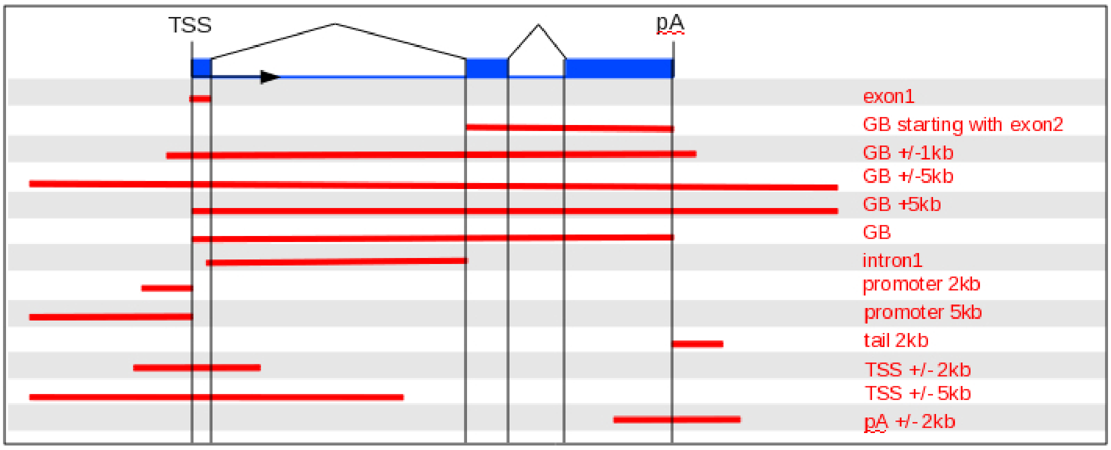
[*]To whom the correspondence should be addressed: eduardo.eyras@upf.edu

Keywords: chromatin, machine learning, regulation, histone code, ChIP-Seq, high-throughput sequencing, differential expression

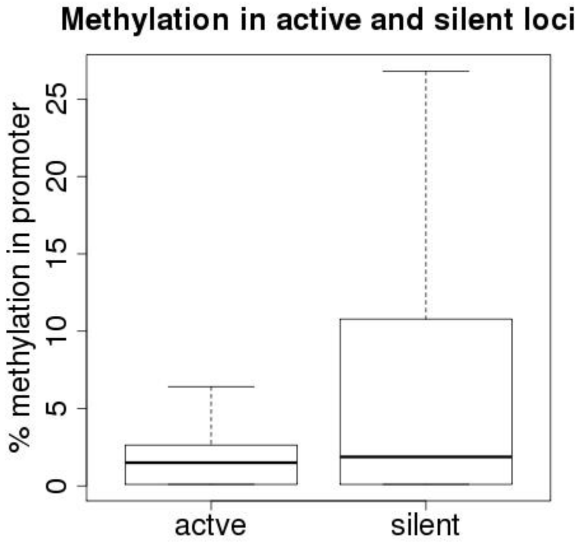
Supplementary Figures

Supplementary Figure 1 Graphical illustration of region definition

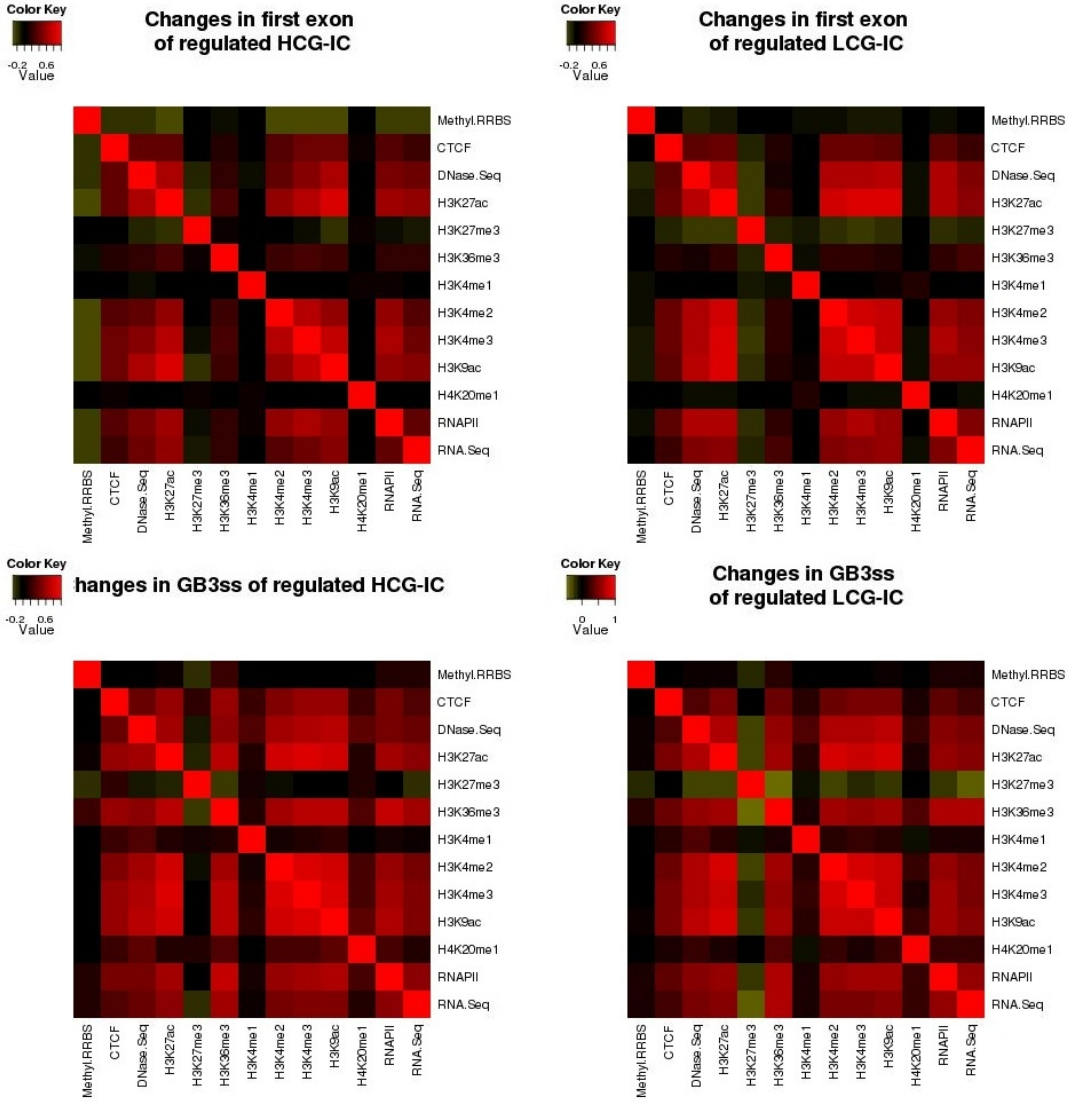
We defined 13 regions based on the gene annotations from Gencode version 7 (Ensembl 62).

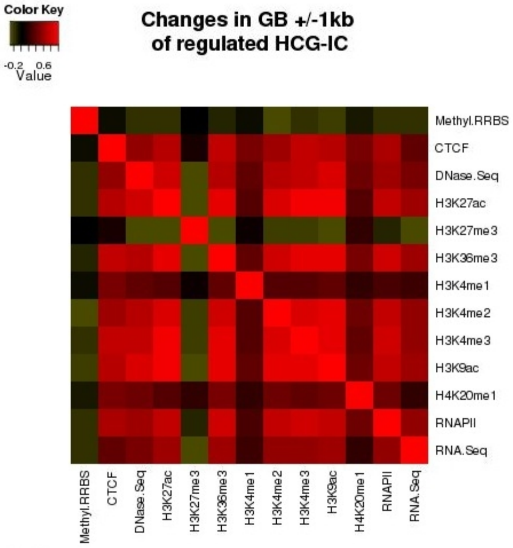


Supplementary Figure 2: DNA methylation measured in the promoter regions (2kb) of active and silent HCG transcript loci. The analysis was done in replica 1 of K562. Silent transcript loci are the 31,347 HCG transcript loci with an RPKM of 0 from RNA-Seq. Accordingly we selected the 31,347 top scoring HCG transcript loci in terms of RPKM from RNA-Seq as active transcript loci. Silent and active loci show significantly different DNA methylation in the promoter.

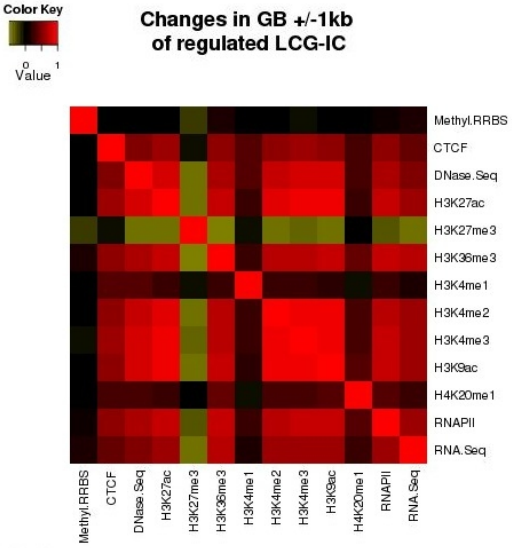


Supplementary Figure 3: Pairwise correlations of marks and expression changes at regulated loci

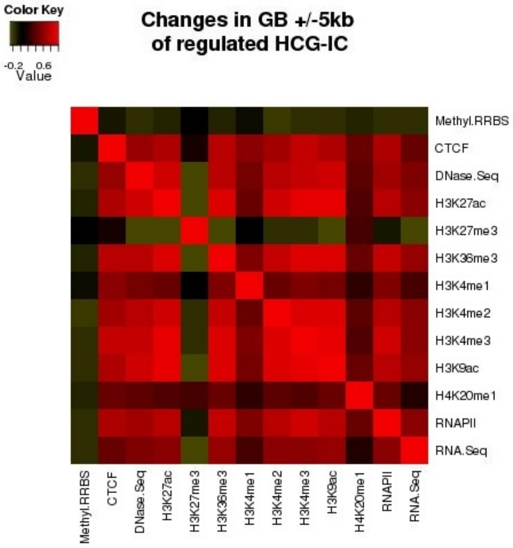




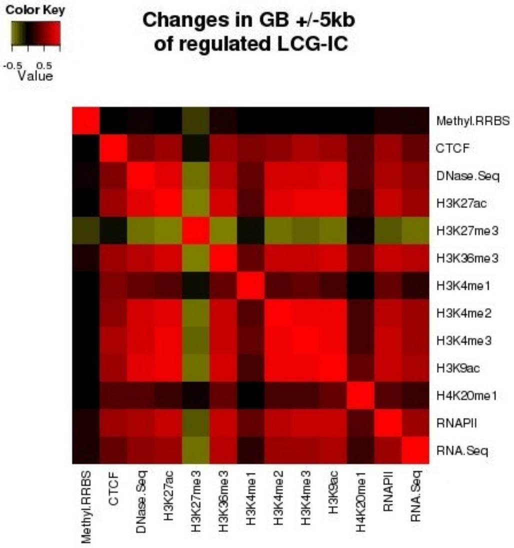
Changes in GB +/-1kb of regulated HCG-IC



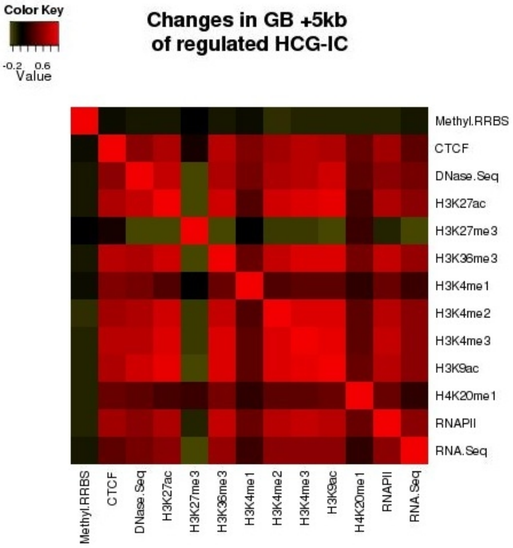
Changes in GB +/-1kb of regulated LCG-IC



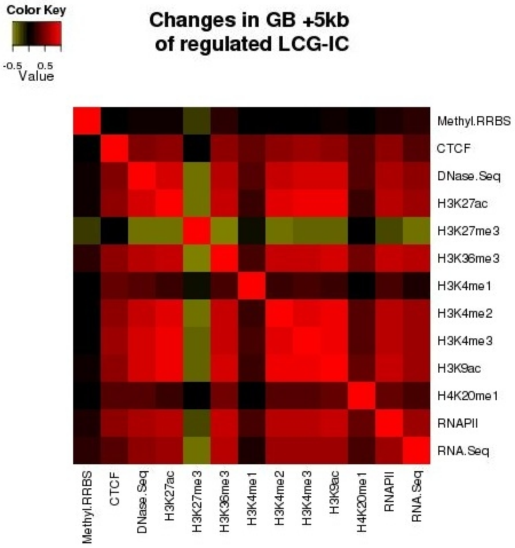
Changes in GB +/-5kb of regulated HCG-IC



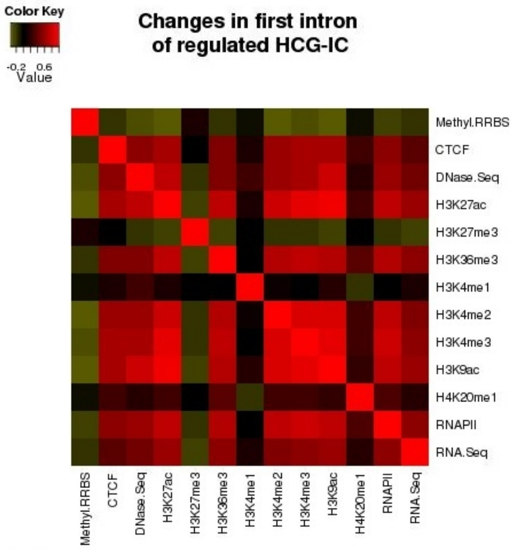
Changes in GB +/-5kb of regulated LCG-IC

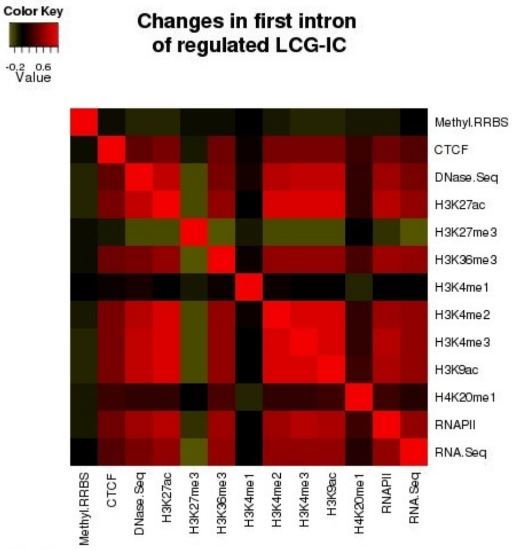


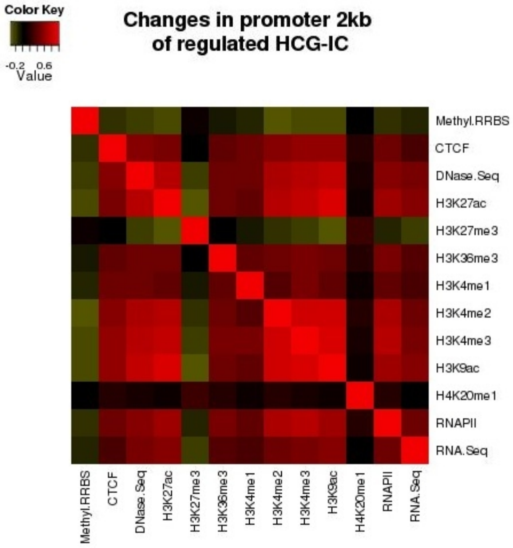
Changes in GB +5kb of regulated HCG-IC

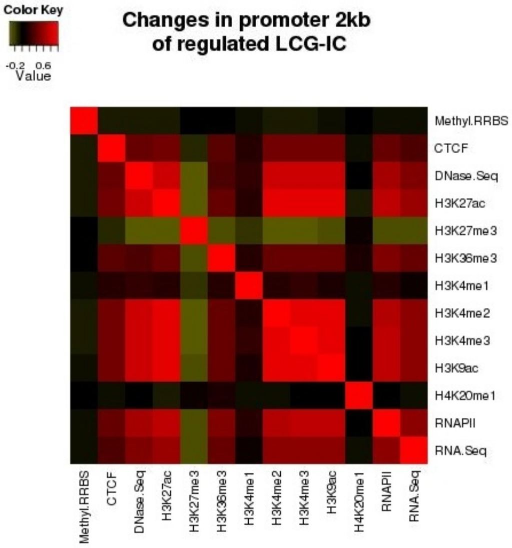


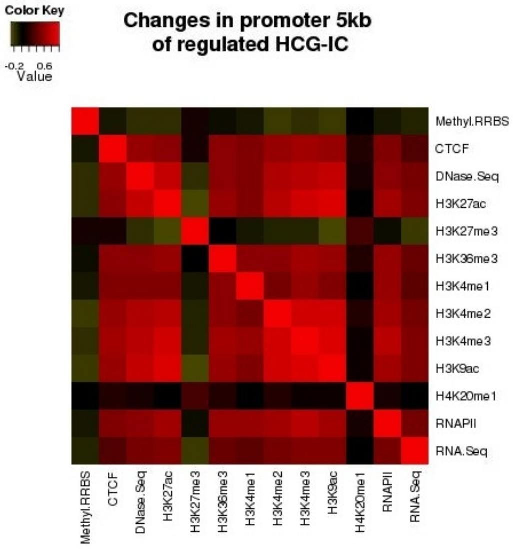
Changes in GB +5kb of regulated LCG-IC

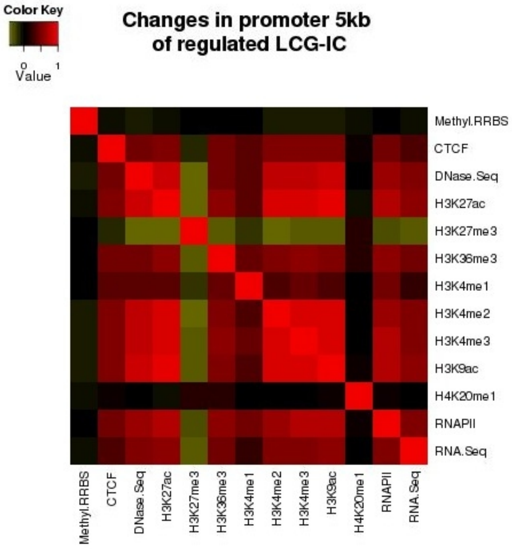

Changes in first intron of regulated HCG-IC

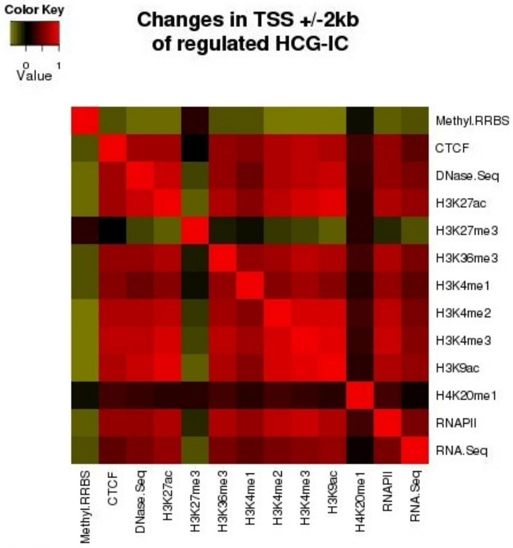

Changes in first intron of regulated LCG-IC


Changes in promoter 2kb of regulated HCG-IC

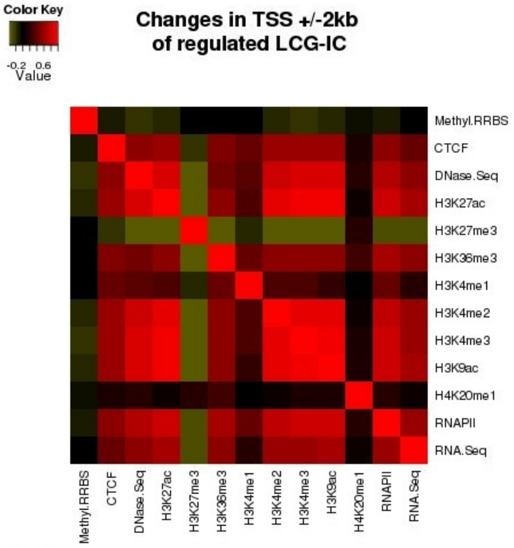

Changes in promoter 2kb of regulated LCG-IC


Changes in promoter 5kb of regulated HCG-IC

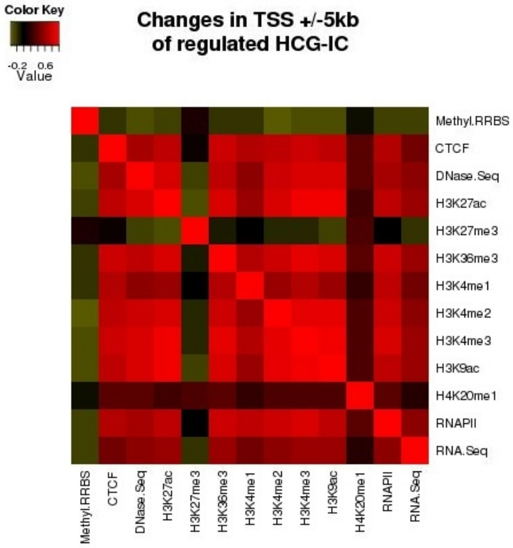

Changes in promoter 5kb of regulated LCG-IC



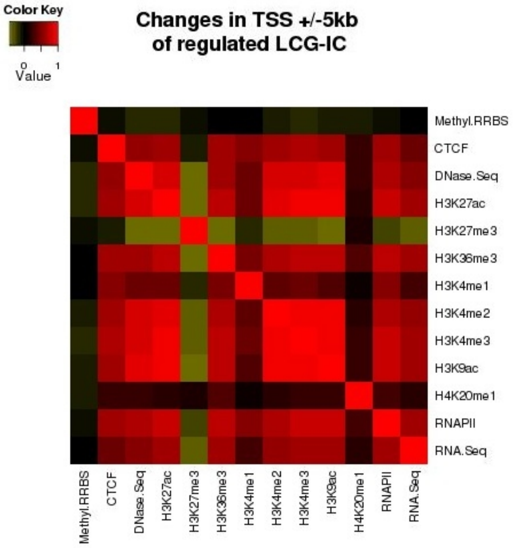
Changes in TSS +/-2kb of regulated HCG-IC



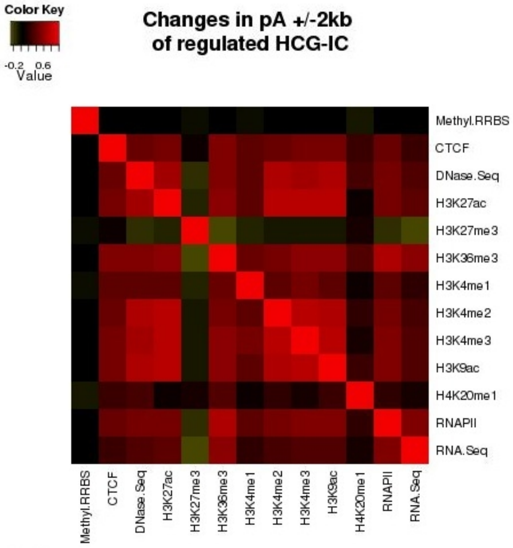
Changes in TSS +/-2kb of regulated LCG-IC



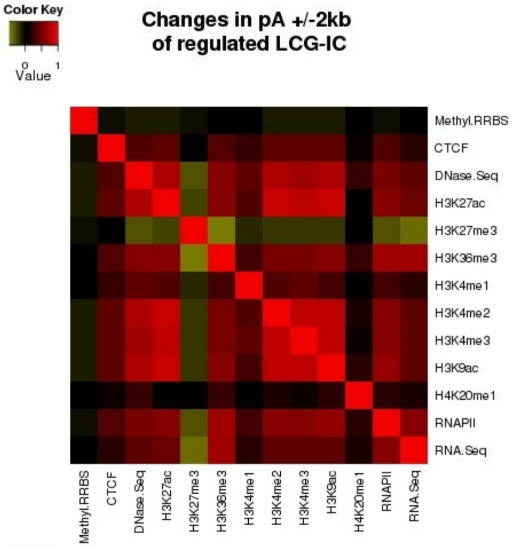
Changes in TSS +/-5kb of regulated HCG-IC



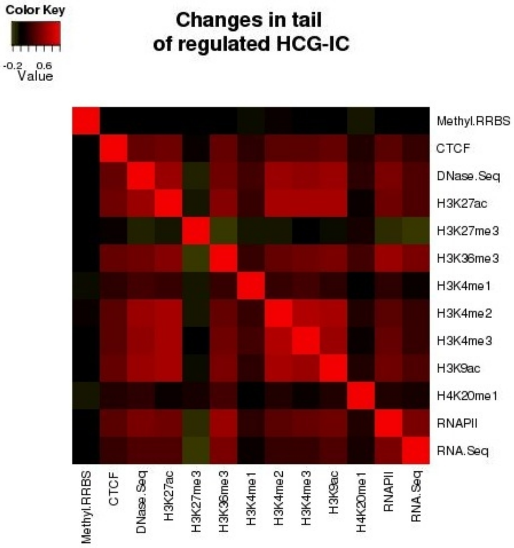
Changes in TSS +/-5kb of regulated LCG-IC



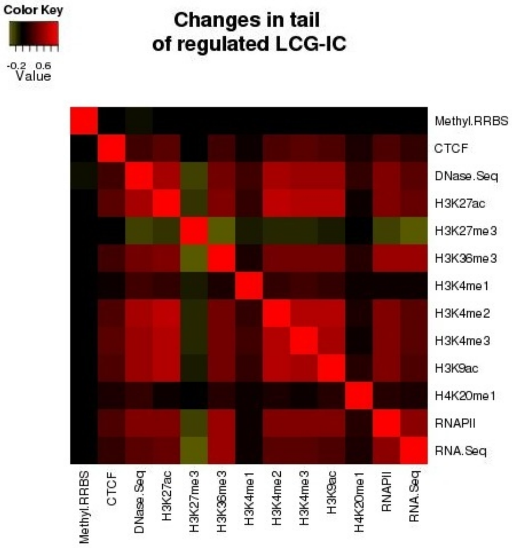
Changes in pA +/-2kb of regulated HCG-IC



Changes in pA +/-2kb of regulated LCG-IC

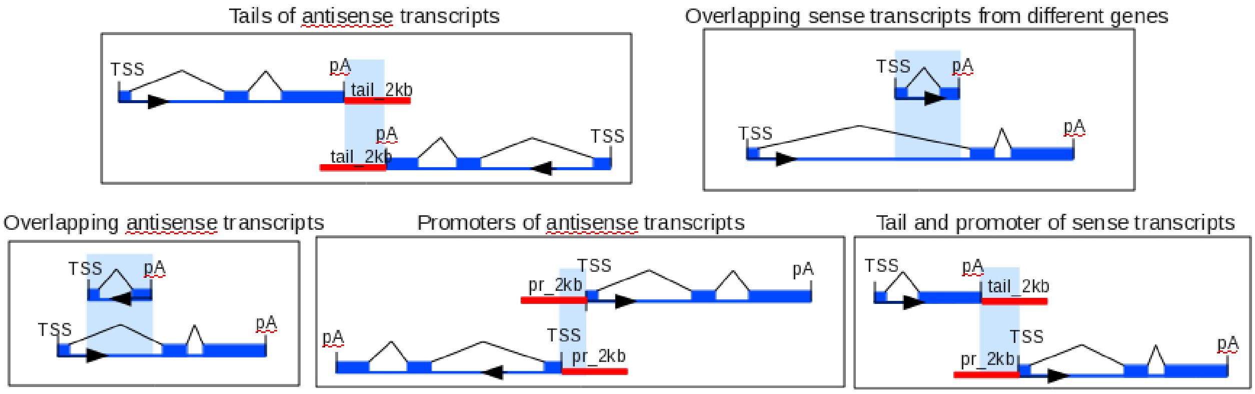


Changes in tail of regulated HCG-IC

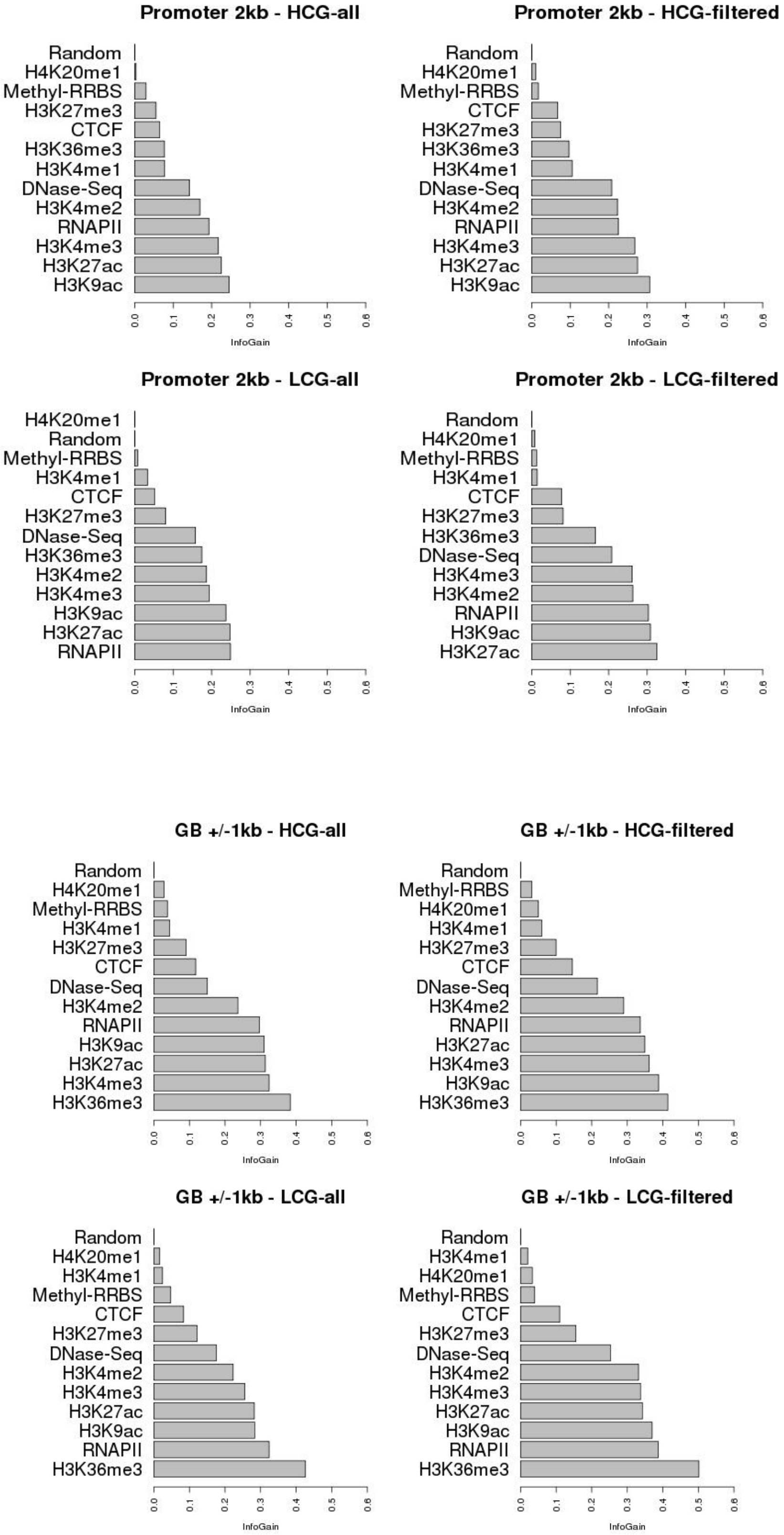


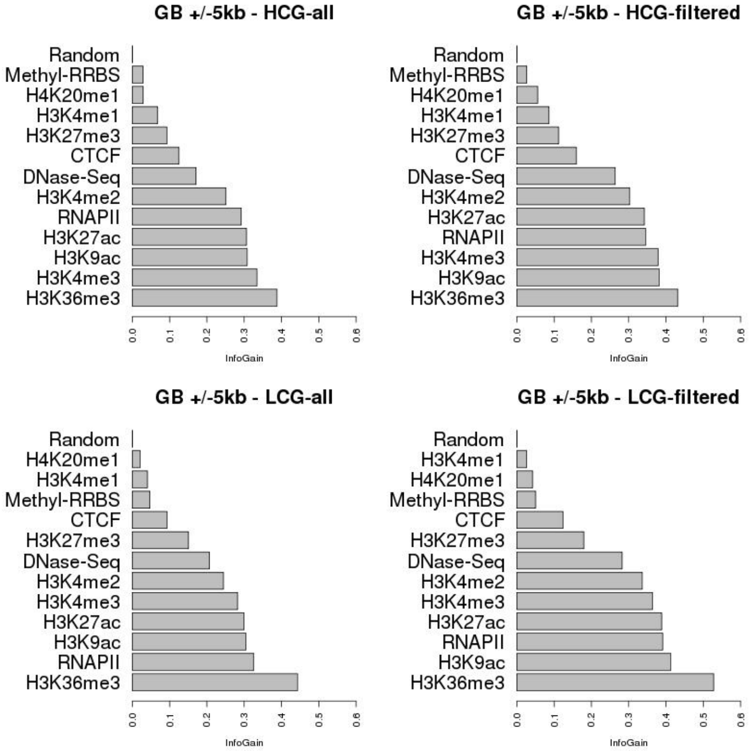
Changes in tail of regulated LCG-IC

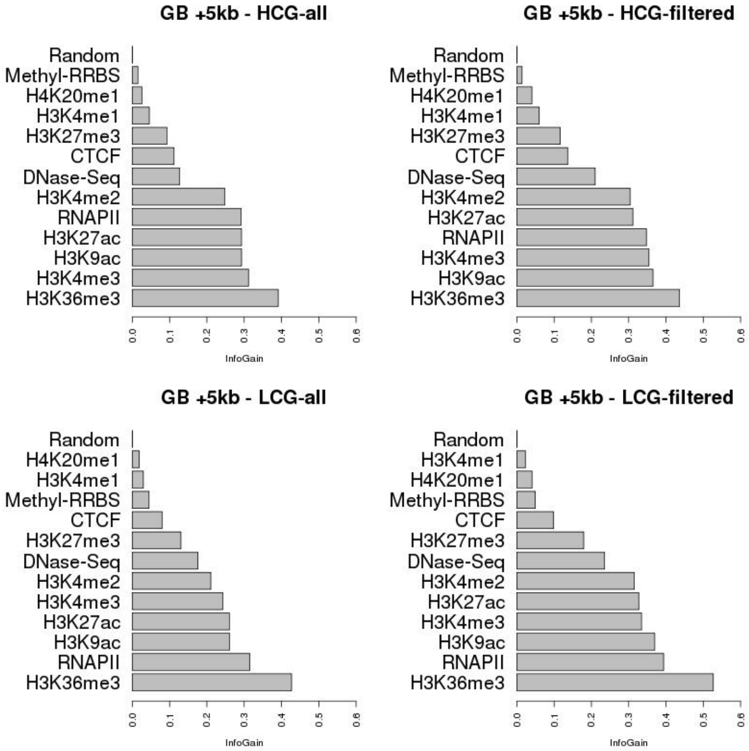
Supplementary Figure 4: Different configurations of transcripts that were overlapping in a way that made the epigenetic signals ambiguous. Filtering the transcripts by excluding these cases resulted into a cleaner predictive model and thus into an improved prediction accuracy.

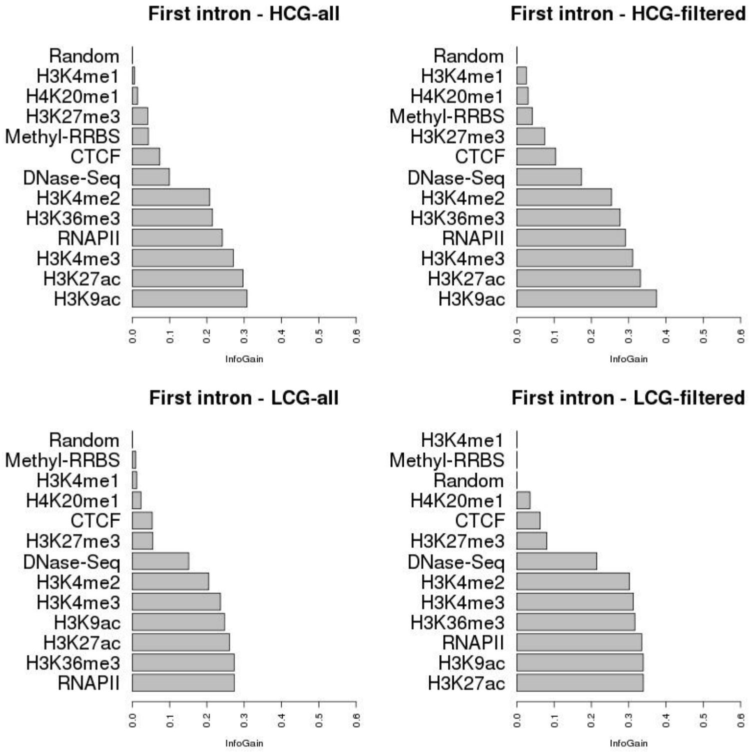

Tails of antisense transcripts
TSS
pA
tail_2kb
pA
tail_2kb
TSS
Overlapping sense transcripts from different genes
TSS
pA
TSS
pA
Overlapping antisense transcripts
TSS
pA
TSS
pA
Promoters of antisense transcripts
TSS
pr_2kb
pA
pA
TSS
pr_2kb
Tail and promoter of sense transcripts
TSS
pA
tail_2kb
TSS
pr_2kb
pA

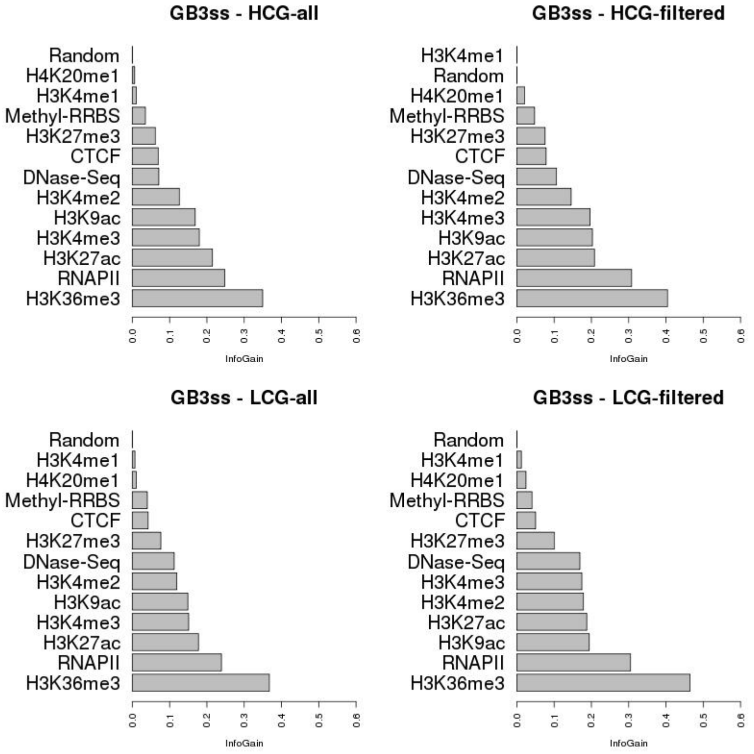
Supplementary Figure 5: Information Gain before and after removing ambiguous signals (pair1).

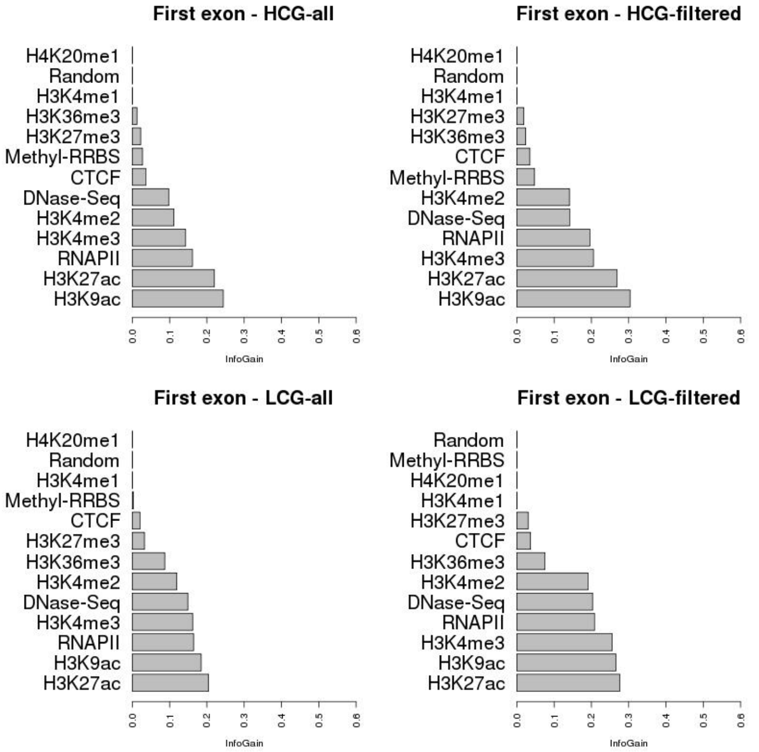



GB +/-5kb - HCG-all
Random
Methyl-RRBS
H4K20me1
H3K4me1
H3K27me3
CTCF
DNase-Seq
H3K4me2
RNAPII
H3K27ac
H3K9ac
H3K4me3
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
GB +/-5kb - HCG-filtered
Random
Methyl-RRBS
H4K20me1
H3K4me1
H3K27me3
CTCF
DNase-Seq
H3K4me2
H3K27ac
RNAPII
H3K4me3
H3K9ac
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
GB +/-5kb - LCG-all
Random
H4K20me1
H3K4me1
Methyl-RRBS
CTCF
H3K27me3
DNase-Seq
H3K4me2
H3K4me3
H3K27ac
H3K9ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
GB +/-5kb - LCG-filtered
Random
H3K4me1
H4K20me1
Methyl-RRBS
CTCF
H3K27me3
DNase-Seq
H3K4me2
H3K4me3
H3K27ac
RNAPII
H3K9ac
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain


GB +5kb - HCG-all
Random
Methyl-RRBS
H4K20me1
H3K4me1
H3K27me3
CTCF
DNase-Seq
H3K4me2
RNAPII
H3K27ac
H3K9ac
H3K4me3
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
GB +5kb - HCG-filtered
Random
Methyl-RRBS
H4K20me1
H3K4me1
H3K27me3
CTCF
DNase-Seq
H3K4me2
H3K27ac
RNAPII
H3K4me3
H3K9ac
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
GB +5kb - LCG-all
Random
H4K20me1
H3K4me1
Methyl-RRBS
CTCF
H3K27me3
DNase-Seq
H3K4me2
H3K4me3
H3K27ac
H3K9ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
GB +5kb - LCG-filtered
Random
H3K4me1
H4K20me1
Methyl-RRBS
CTCF
H3K27me3
DNase-Seq
H3K4me2
H3K27ac
H3K4me3
H3K9ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain


First intron - HCG-all
Random
H3K4me1
H4K20me1
H3K27me3
Methyl-RRBS
CTCF
DNase-Seq
H3K4me2
H3K36me3
RNAPII
H3K4me3
H3K27ac
H3K9ac
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
First intron - HCG-filtered
Random
H3K4me1
H4K20me1
Methyl-RRBS
H3K27me3
CTCF
DNase-Seq
H3K4me2
H3K36me3
RNAPII
H3K4me3
H3K27ac
H3K9ac
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
First intron - LCG-all
Random
Methyl-RRBS
H3K4me1
H4K20me1
CTCF
H3K27me3
DNase-Seq
H3K4me2
H3K4me3
H3K9ac
H3K27ac
H3K36me3
RNAPII
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
First intron - LCG-filtered
H3K4me1
Methyl-RRBS
Random
H4K20me1
CTCF
H3K27me3
DNase-Seq
H3K4me2
H3K4me3
H3K36me3
RNAPII
H3K9ac
H3K27ac
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain

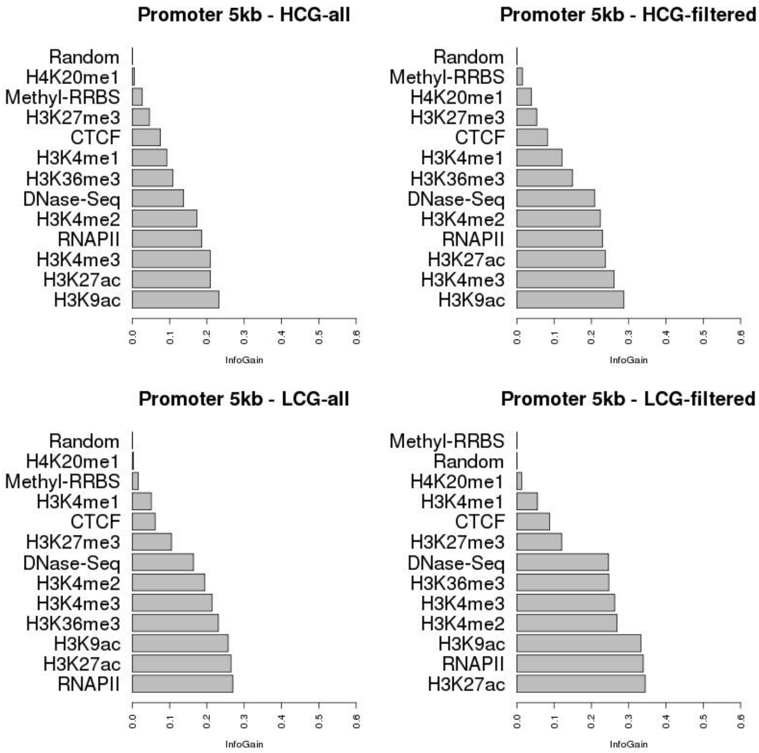

GB3ss - HCG-all
Random
H4K20me1
H3K4me1
Methyl-RRBS
H3K27me3
CTCF
DNase-Seq
H3K4me2
H3K9ac
H3K4me3
H3K27ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
GB3ss - HCG-filtered
H3K4me1
Random
H4K20me1
Methyl-RRBS
H3K27me3
CTCF
DNase-Seq
H3K4me2
H3K4me3
H3K9ac
H3K27ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
GB3ss - LCG-all
Random
H3K4me1
H4K20me1
Methyl-RRBS
CTCF
H3K27me3
DNase-Seq
H3K4me2
H3K9ac
H3K4me3
H3K27ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
GB3ss - LCG-filtered
Random
H3K4me1
H4K20me1
Methyl-RRBS
CTCF
H3K27me3
DNase-Seq
H3K4me3
H3K4me2
H3K27ac
H3K9ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain


First exon - HCG-all

First exon - HCG-filtered

First exon - LCG-all

First exon - LCG-filtered

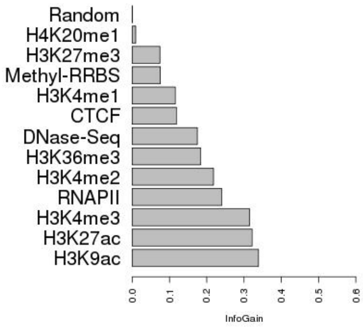

Promoter 5kb - HCG-all

Promoter 5kb - HCG-filtered

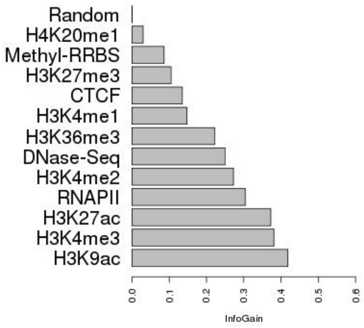
Promoter 5kb - LCG-all

Promoter 5kb - LCG-filtered

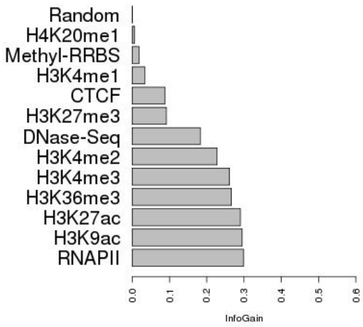
TSS 2kb - HCG-all



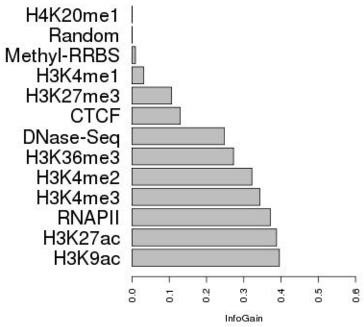
TSS 2kb - HCG-filtered



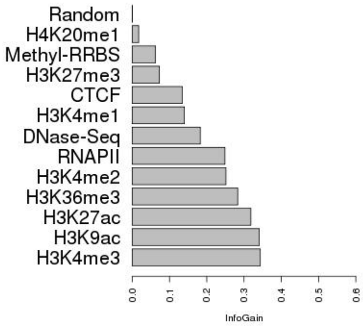
TSS 2kb - LCG-all



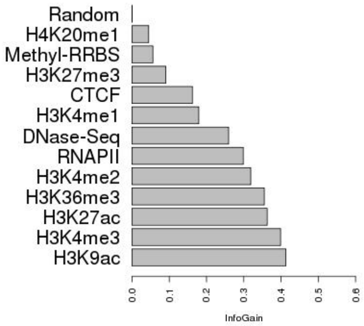
TSS 2kb - LCG-filtered



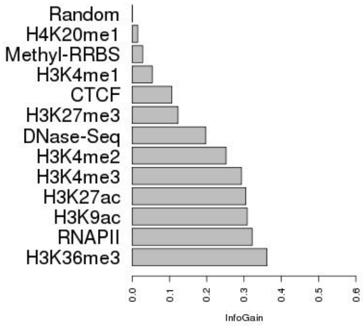
TSS 5kb - HCG-all



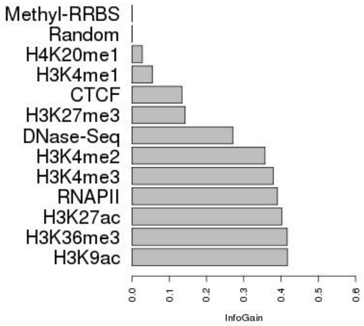
TSS 5kb - HCG-filtered

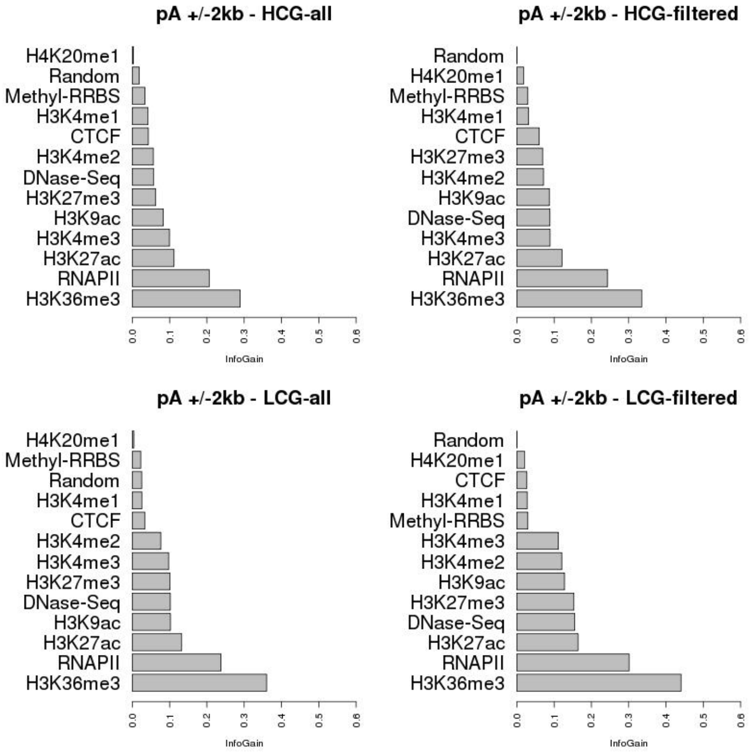


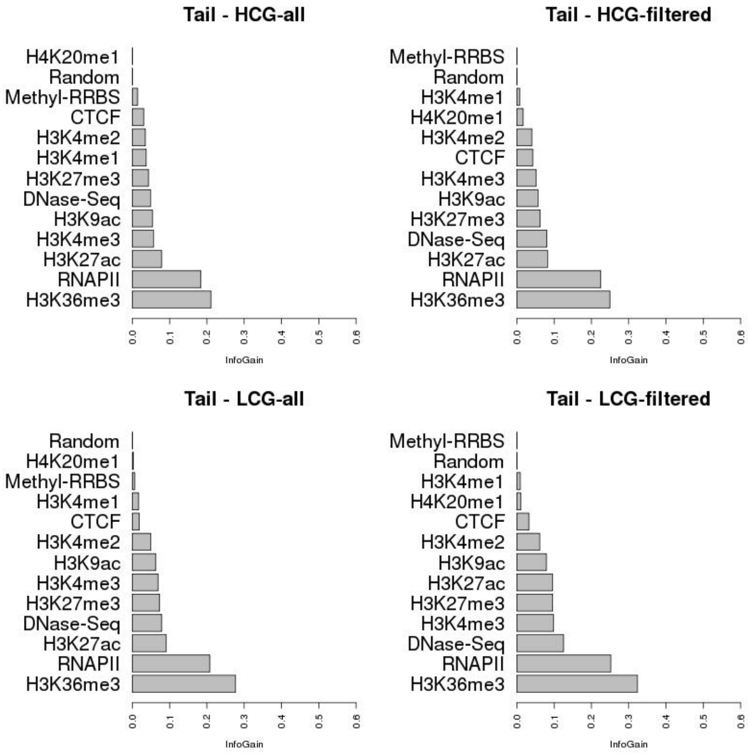
TSS 5kb - LCG-all



TSS 5kb - LCG-filtered

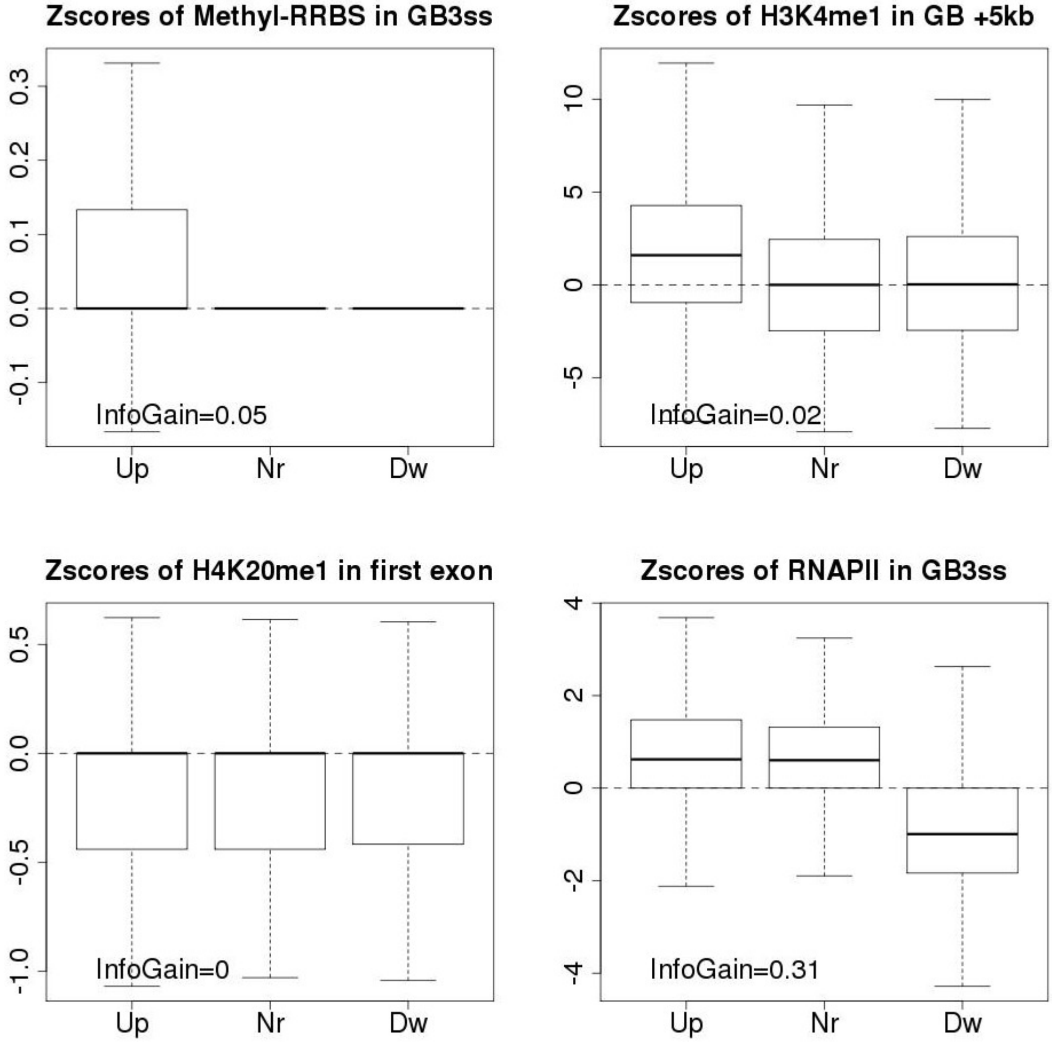



pA +/-2kb - HCG-all
H4K20me1
Random
Methyl-RRBS
H3K4me1
CTCF
H3K4me2
DNase-Seq
H3K27me3
H3K9ac
H3K4me3
H3K27ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
pA +/-2kb - HCG-filtered
Random
H4K20me1
Methyl-RRBS
H3K4me1
CTCF
H3K27me3
H3K4me2
H3K9ac
DNase-Seq
H3K4me3
H3K27ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
pA +/-2kb - LCG-all
H4K20me1
Methyl-RRBS
Random
H3K4me1
CTCF
H3K4me2
H3K4me3
H3K27me3
DNase-Seq
H3K9ac
H3K27ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
pA +/-2kb - LCG-filtered
Random
H4K20me1
CTCF
H3K4me1
Methyl-RRBS
H3K4me3
H3K4me2
H3K9ac
H3K27me3
DNase-Seq
H3K27ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain

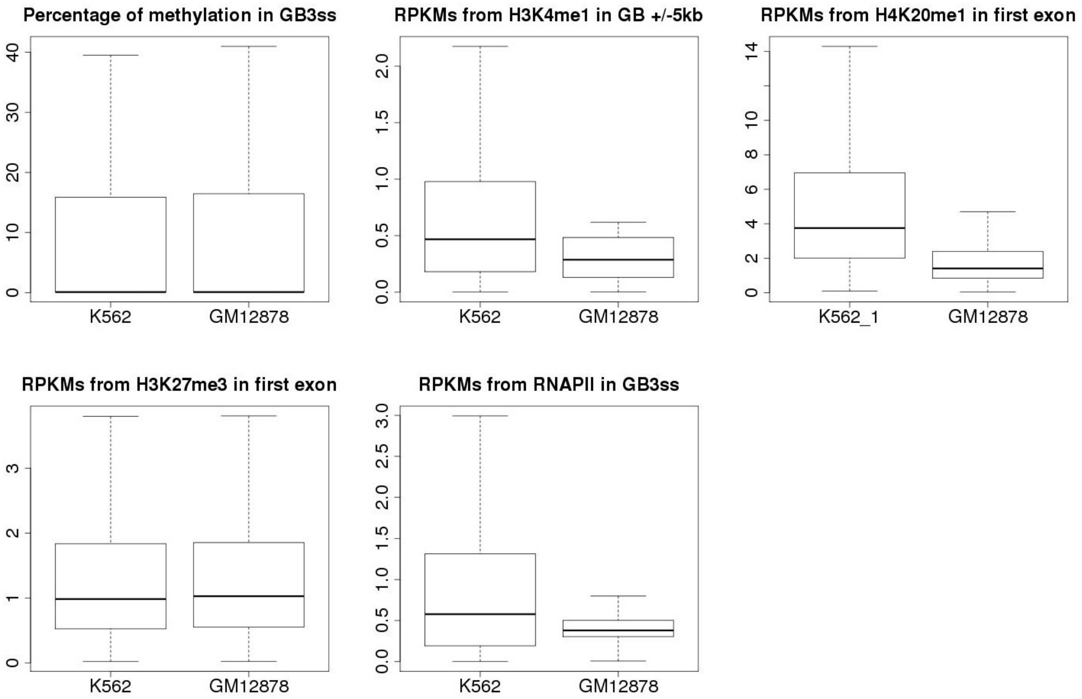

Tail - HCG-all
H4K20me1
Random
Methyl-RRBS
CTCF
H3K4me2
H3K4me1
H3K27me3
DNase-Seq
H3K9ac
H3K4me3
H3K27ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
Tail - HCG-filtered
Methyl-RRBS
Random
H3K4me1
H4K20me1
H3K4me2
CTCF
H3K4me3
H3K9ac
H3K27me3
DNase-Seq
H3K27ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
Tail - LCG-all
Random
H4K20me1
Methyl-RRBS
H3K4me1
CTCF
H3K4me2
H3K9ac
H3K4me3
H3K27me3
DNase-Seq
H3K27ac
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain
Tail - LCG-filtered
Methyl-RRBS
Random
H3K4me1
H4K20me1
CTCF
H3K4me2
H3K9ac
H3K27ac
H3K27me3
H3K4me3
DNase-Seq
RNAPII
H3K36me3
0.0
0.1
0.2
0.3
0.4
0.5
0.6
InfoGain

Supplementary Figure 6:

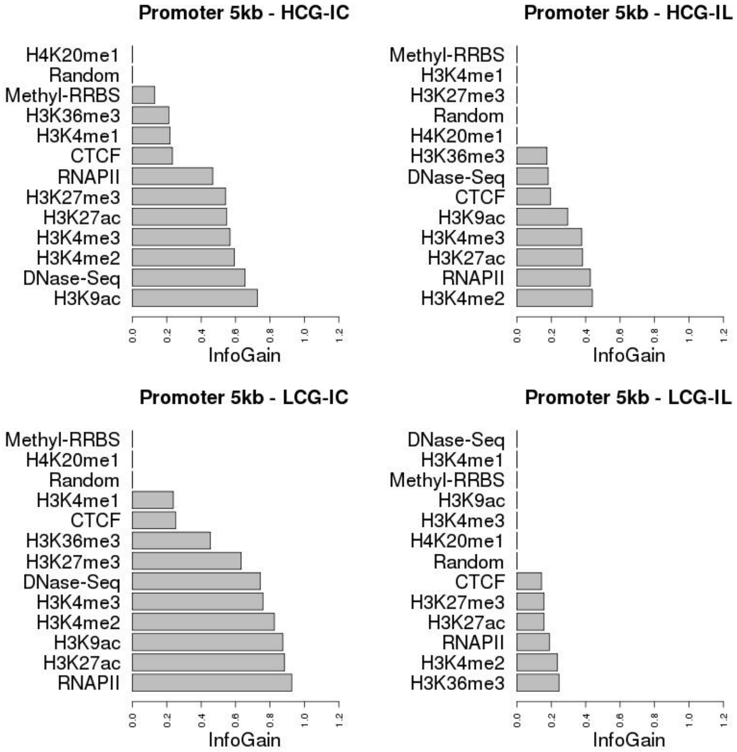
A) Distribution of z-scores for up- (Up), down- (Dw) and non- (Nr) regulated genes for the non-optimal attributes for each experiment, calculated by maximizing the Information Gain and minimizing the absolute value of the median for the z-score distribution of the Nr subset. The y-axis shows the z-score corresponding to the enrichment of the attribute. These distributions correspond to the set of LCG-IC loci of Pair1.

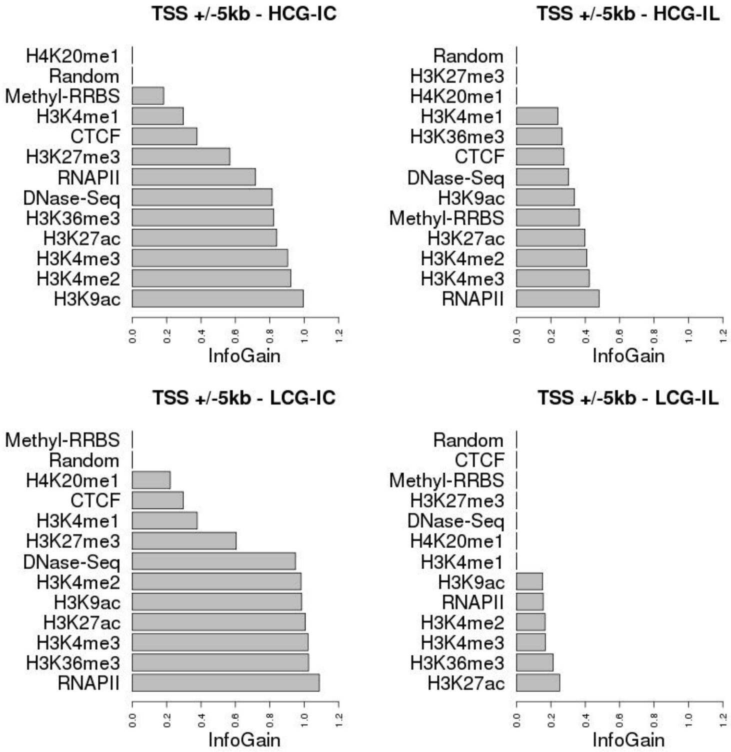


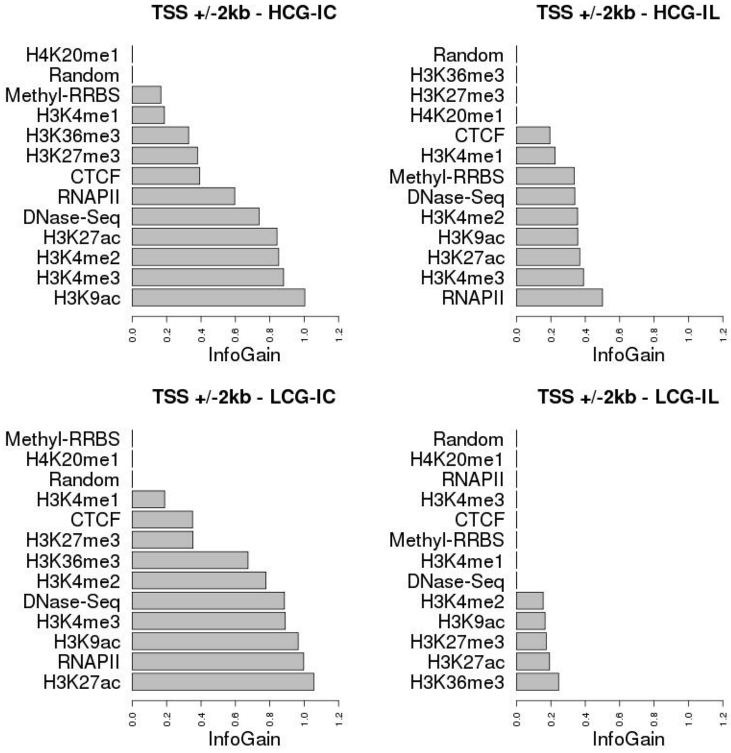
B) Overall distribution of signal from K562 and GM12878 in best regions.

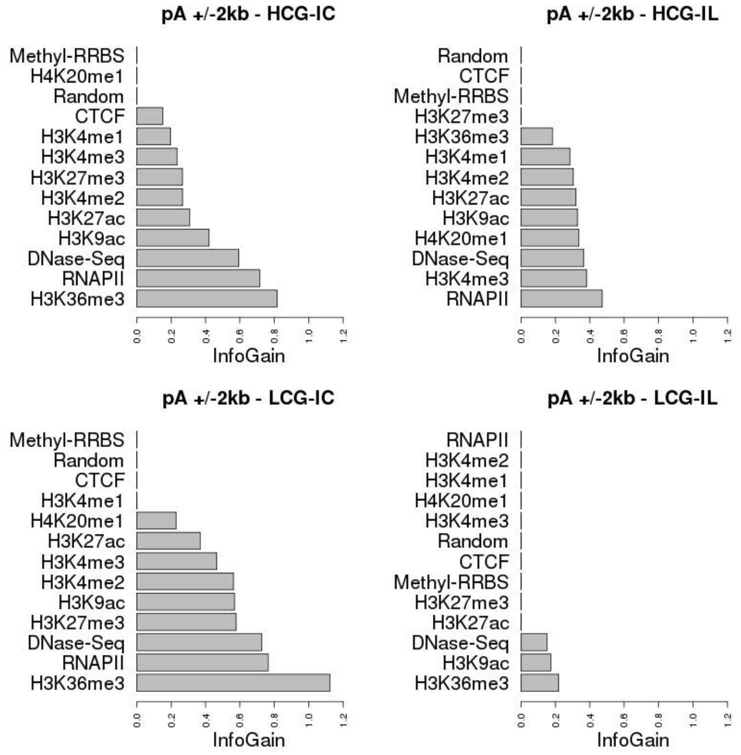


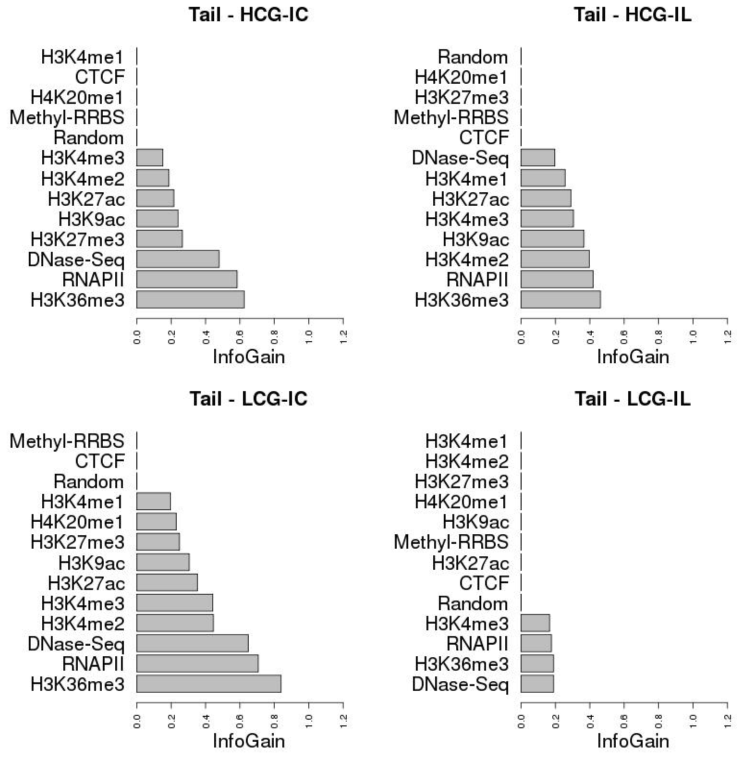
Supplementary Figure 7: Information Gain in intron-containing and intron-less genes (pair1). For this analysis we did not remove confounding effects as this would lead to very small groups. Sets from Pair1 and Pair2 have comparable sized and regions (avoiding length-bias.)




TSS +/-5kb - HCG-IC
H4K20me1
Random
Methyl-RRBS
H3K4me1
CTCF
H3K27me3
RNAPII
DNase-Seq
H3K36me3
H3K27ac
H3K4me3
H3K4me2
H3K9ac
0.0 0.2 0.4 0.6 0.8 1.0 1.2
InfoGain
TSS +/-5kb - HCG-IL
Random
H3K27me3
H4K20me1
H3K4me1
H3K36me3
CTCF
DNase-Seq
H3K9ac
Methyl-RRBS
H3K27ac
H3K4me2
H3K4me3
RNAPII
0.0 0.2 0.4 0.6 0.8 1.0 1.2
InfoGain
TSS +/-5kb - LCG-IC
Methyl-RRBS
Random
H4K20me1
CTCF
H3K4me1
H3K27me3
DNase-Seq
H3K4me2
H3K9ac
H3K27ac
H3K4me3
H3K36me3
RNAPII
0.0 0.2 0.4 0.6 0.8 1.0 1.2
InfoGain
TSS +/-5kb - LCG-IL
Random
CTCF
Methyl-RRBS
H3K27me3
DNase-Seq
H4K20me1
H3K4me1
H3K9ac
RNAPII
H3K4me2
H3K4me3
H3K36me3
H3K27ac
0.0 0.2 0.4 0.6 0.8 1.0 1.2
InfoGain


TSS +/-2kb - HCG-IC
H4K20me1
Random
Methyl-RRBS
H3K4me1
H3K36me3
H3K27me3
CTCF
RNAPII
DNase-Seq
H3K27ac
H3K4me2
H3K4me3
H3K9ac
0.0 0.2 0.4 0.6 0.8 1.0 1.2
InfoGain
TSS +/-2kb - HCG-IL
Random
H3K36me3
H3K27me3
H4K20me1
CTCF
H3K4me1
Methyl-RRBS
DNase-Seq
H3K4me2
H3K9ac
H3K27ac
H3K4me3
RNAPII
0.0 0.2 0.4 0.6 0.8 1.0 1.2
InfoGain
TSS +/-2kb - LCG-IC
Methyl-RRBS
H4K20me1
Random
H3K4me1
CTCF
H3K27me3
H3K36me3
H3K4me2
DNase-Seq
H3K4me3
H3K9ac
RNAPII
H3K27ac
0.0 0.2 0.4 0.6 0.8 1.0 1.2
InfoGain
TSS +/-2kb - LCG-IL
Random
H4K20me1
RNAPII
H3K4me3
CTCF
Methyl-RRBS
H3K4me1
DNase-Seq
H3K4me2
H3K9ac
H3K27me3
H3K27ac
H3K36me3
0.0 0.2 0.4 0.6 0.8 1.0 1.2
InfoGain


pA +/-2kb - HCG-IC
Methyl-RRBS
H4K20me1
Random
CTCF
H3K4me1
H3K4me3
H3K27me3
H3K4me2
H3K27ac
H3K9ac
DNase-Seq
RNAPII
H3K36me3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
InfoGain
pA +/-2kb - HCG-IL
Random
CTCF
Methyl-RRBS
H3K27me3
H3K36me3
H3K4me1
H3K4me2
H3K27ac
H3K9ac
H4K20me1
DNase-Seq
H3K4me3
RNAPII
0.0
0.2
0.4
0.6
0.8
1.0
1.2
InfoGain
pA +/-2kb - LCG-IC
Methyl-RRBS
Random
CTCF
H3K4me1
H4K20me1
H3K27ac
H3K4me3
H3K4me2
H3K9ac
H3K27me3
DNase-Seq
RNAPII
H3K36me3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
InfoGain
pA +/-2kb - LCG-IL
RNAPII
H3K4me2
H3K4me1
H4K20me1
H3K4me3
Random
CTCF
Methyl-RRBS
H3K27me3
H3K27ac
DNase-Seq
H3K9ac
H3K36me3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
InfoGain


Tail - HCG-IC
H3K4me1
CTCF
H4K20me1
Methyl-RRBS
Random
H3K4me3
H3K4me2
H3K27ac
H3K9ac
H3K27me3
DNase-Seq
RNAPII
H3K36me3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
InfoGain
Tail - HCG-IL
Random
H4K20me1
H3K27me3
Methyl-RRBS
CTCF
DNase-Seq
H3K4me1
H3K27ac
H3K4me3
H3K9ac
H3K4me2
RNAPII
H3K36me3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
InfoGain
Tail - LCG-IC
Methyl-RRBS
CTCF
Random
H3K4me1
H4K20me1
H3K27me3
H3K9ac
H3K27ac
H3K4me3
H3K4me2
DNase-Seq
RNAPII
H3K36me3
0.0
0.2
0.4
0.6
0.8
1.0
1.2
InfoGain
Tail - LCG-IL
H3K4me1
H3K4me2
H3K27me3
H4K20me1
H3K9ac
Methyl-RRBS
H3K27ac
CTCF
Random
H3K4me3
RNAPII
H3K36me3
DNase-Seq
0.0
0.2
0.4
0.6
0.8
1.0
1.2
InfoGain

Supplementary tables:

Supplementary table 1: Accuracy in terms of the area under the ROC curve (AUC) for the 10-fold cross validation for the IL transcript sets for various training conditions**.** P1 (with RNAPII) corresponds to pair P1 with the additional RNAPII feature, i.e. the same features as P2 plus RNAPII. P1 and P2 denote the models for each cell line pairs with all the features. P1(CFS) and P2(CFS) denote the models for P1 and P2, respectively, where the features used are those that have a score of 80 or higher (maximum 100) using the CFS feature selection method independently for P1 and P2. P2 (CFS-P1) indicates that the model was trained using the data from P2 but the features selected using CFS on P1.

	HCG – IL				***LCG – IL***			
	Up	Dw	Nr	Average	Up	Dw	Nr	Average
P1 (with RNAPII)	0.88	0.91	0.87	0.88	0.72	0.7	0.71	0.71
P1	0.87	0.87	0.82	0.85	0.78	0.76	0.62	0.72
P1 (CFS)	0.87	0.91	0.84	0.87	0.81	0.7	0.65	0.72

Supplementary Table2: Features selected by Correlation Feature Selection (appearing in at least 80% of validations)

A) 29 features from P1 HCG-IC before filtering ambiguous signal (bold ones also for P2)

9(90 %)	first exon Dnase
9(90 %)	first exon H3k4me2
10(100 %)	first exon H3k9ac
8(80 %)	GB3ss H3k27ac
10(100 %)	**GB3ss H3k36me3**
8(80 %)	GB +/-1kb H3k27me3
10(100 %)	GB +/-1kb H3k36me3
10(100 %)	**GB +/-5kb H3k27ac**
9(90 %)	**GB +/-5kb H3k36me3**
9(90 %)	GB +/-5kb H3k4me2
9(90 %)	**GB +5kb H3k4me3**
9(90 %)	**GB H3k27ac**
10(100 %)	**GB H3k36me3**
10(100 %)	GB H3k9ac
9(90 %)	**first intron Methyl**
9(90 %)	first intron Ctcf
8(80 %)	first intron H3k36me3
8(80 %)	**first intron H3k4me3**
8(80 %)	first intron H3k9ac
9(90 %)	Promoter 2kb H3k4me2
9(90 %)	**tail H3k36me3**
8(80 %)	TSS +/-2kb H3k4me2
9(90 %)	**TSS +/-2kb H3k9ac**
10(100 %)	**TSS +/-5kb H3k27ac**
8(80 %)	**TSS +/-5kb H3k36me3**
8(80 %)	TSS +/-5kb H3k9ac
9(90 %)	TSS +/-5kb H4K20me1
9(90 %)	**pA +/-2kb H3k27me3**
10(100 %)	**pA +/-2kb H3k36me3**

B) 33 features from P2 HCG-IC before filtering ambiguous signal (bold ones also for P1):

9(90 %)	first exon H3k4me3
9(90 %)	**GB3ss H3k36me3**
8(80 %)	GB +/-1kb H3k27ac
10(100 %)	GB +/-1kb H3k36me3
8(80 %)	GB +/-1kb H3k4me3
8(80 %)	GB +/-5kb Dnase
10(100 %)	**GB +/-5kb H3k27ac**
10(100 %)	**GB +/-5kb H3k36me3**
8(80 %)	GB +/-5kb H4K20me1
8(80 %)	GB +5kb Ctcf
8(80 %)	GB +5kb H3k36me3
9(90 %)	**GB +5kb H3k4me3**
9(90 %)	GB +5kb H4K20me1
8(80 %)	**GB H3k27ac**
10(100 %	**GB H3k36me3**
10(100 %)	GB H3k4me3
8(80 %)	**first intron Methyl**
8(80 %)	first intron H3k4me2
10(100 %)	**first intron H3k4me3**
9(90 %)	Promoter 2kb Ctcf
9(90 %)	Promoter 5kb Dnase
8(80 %)	Promoter 5kb H3k27ac
9(90 %)	**tail H3k36me3**
10(100 %)	TSS +/-2kb _H3k27ac

10(100 %)	TSS +/-2kb H3k4me3
9(90 %)	**TSS +/-2kb H3k9ac**
9(90 %)	**TSS +/-5kb H3k27ac**
10(100 %)	TSS +/-5kb H3k27me3
9(90 %)	**TSS +/-5kb H3k36me3**
10(100 %)	TSS +/-5kb H3k4me2
10(100 %)	TSS +/-5kb H3k4me3
9(90 %)	**pA +/-2kb H3k27me3**
10(100 %)	**pA +/-2kb H3k36me3**

C) 25 features from P1 LCG-IC before filtering ambiguous signal (bold ones also for P2)

9(90 %)	first exon Dnase
8(80 %)	**first exon H3k4me2**
8(80 %)	first exon H3k9ac
10(100 %)	**GB3ss H3k36me3**
8(80 %)	GB +/-1kb H3k27ac
8(80 %)	GB +/-1kb H3k27me3
10(100 %)	**GB +/-1kb H3k36me3**
10(100 %)	**GB +/-5kb H3k36me3**
8(80 %)	GB +5kb H3k9ac
10(100 %)	GB Methyl
10(100 %)	**GB H3k36me3**
8(80 %)	GB H3k9ac
8(80 %)	first intron Methyl
8(80 %)	**first intron H3k36me3**
9(90 %)	first intron H4K20me1
9(90 %)	Promoter 2kb H3k27me3
8(80 %)	Promoter 2kb H3k36me3
8(80 %)	Promoter 5kb Ctcf
9(90 %)	Promoter 5kb H3k36me3
8(80 %)	tail H3k27me3
9(90 %)	**tail H3k36me3**
8(80 %)	**TSS +/-2kb H3k36me3**
9(90 %)	TSS +/-2kb H3k9ac
10(100 %)	**TSS +/-5kb H3k36me3**
10(100 %)	**pA +/-5kb H3k36me3**

D) 20 features from P2 LCG-IC before filtering ambiguous signal (bold ones also for P1):

8(80 %)	first exon H3k36me3
8(80 %)	**first exon H3k4me2**
8(80 %)	first exon H3k4me3
10(100 %)	**GB3ss H3k36me3**
10(100 %)	**GB +/-1kb H3k36me3**
8(80 %)	**GB +/-5kb H3k36me3**
8(80 %)	GB H3k27ac
10(100 %)	**GB H3k36me3**
10(100 %)	**first intron H3k36me3**
8(80 %)	first intron H3k4me2
8(80 %)	Promoter 2kb Ctcf
8(80 %)	Promoter 5kb H3k27ac
10(100 %)	**tail H3k36me3**
8(80 %)	tail H3k9ac
10(100 %)	**TSS +/-2kb H3k36me3**
10(100 %)	TSS +/-2kb H3k4me3
9(90 %)	**TSS +/- 5kb H3k36me3**
8(80 %)	TSS +/-5kb H3k4me2
9(90 %)	pA +/-2kb H3k27me3
10(100 %)	**pA +/-2kb H3k36me3**

E) 16 features from P1 HCG-IC after filtering ambiguous signal (bold ones also appear for P2)

10(100 %)	first exon H3k4me2
8(80 %)	first exon H3k9ac
8(80 %)	**GB3ss H3k36me3**
8(80 %)	**GB +/-1kb H3k36me3**
9(90 %)	**GB +/-5kb H3k36me3**
8(80 %)	GB +/-5kb_onlyTTS H3k4me1
8(80 %)	first intron Methyl
8(80 %)	**first intron H3k27ac**
10(100 %)	Promoter 2kb H3k4me3
9(90 %)	tail H3k36me3
8(80 %)	TSS +/- 2kb Methyl
8(80 %)	TSS +/- 2kb H3k36me3
9(90 %)	**TSS +/- 5kb H3k27ac**
10(100 %)	TSS +/- 5kb H3k36me3
8(80 %)	**TSS +/- 5kb H3k4me3**
9(90 %)	pA +/- 2kb H3k36me3

F) 16 features from P2 HCG-IC after filtering ambiguous signal (bold ones also appear for P1)

8(80 %)	**GB3ss H3k36me3**
8(80 %)	GB +/-1kb H3k27a
9(90 %)	**GB +/-1kb H3k36me3**
9(90 %)	GB +/-1kb H3k4me3
8(80 %)	GB +/-5kb H3k27ac
9(90 %)	**GB +/-5kb H3k36me3**
8(80 %)	GB +/-5kb_onlyTTS H3k36me3
8(80 %)	genebody H3k36me3
9(90 %)	genebody H3k9ac
8(80 %)	**first intron H3k27ac**
10(100 %)	first intron H3k4me3
8(80 %)	first intron H3k9ac
9(90 %)	Promoter 5kb H3k27ac
10(100 %)	TSS +/- 2kb H3k4me3
8(80 %)	**TSS +/- 5kb H3k27ac**
10(100 %)	**TSS +/- 5kb H3k4me3**

G) 23 features from P1 LCG-IC after filtering ambiguous signal (bold ones also appear for P2)

10(100 %)	first exon Dnase
10(100 %)	first exon H3k4me2
10(100 %)	first exon H3k4me3
9(90 %)	first exon H3k9ac
10(100 %)	**GB3ss H3k36me3**
8(80 %)	GB3ss H3k4me3
8(80 %)	GB +/-1kb Methyl
8(80 %)	GB +/-1kb Dnase
8(80 %)	GB +/-1kb H3k27me3
9(90 %)	**GB +/-1kb H3k36me3**
8(80 %)	GB +/-1kb H3k4me2
8(80 %)	GB +/-5kb Ctcf
10(100 %)	**GB +/-5kb H3k36me3**
8(80 %)	GB +/-5kb_onlyTTS Methyl
8(80 %)	GB +/-5kb_onlyTTS Dnase
10(100 %)	**GB +/-5kb_onlyTTS H3k36me3**
9(90 %)	genebody Methyl
9(90 %)	first intron H3k27ac
8(80 %)	Promoter 2kb Ctcf
8(80 %)	Promoter 5kb H3k36me3
8(80 %)	**tail H3k36me3**
9(90 %)	**TSS +/- 5kb H3k36me3**

10(100 %)	pA +/- 2kb H3k36me3

H) 13 features from P2 LCG-IC after filtering ambiguous signal (bold ones also appear for P1)

10(100 %)	**GB3ss H3k36me3**
9(90 %)	**GB +/-1kb H3k36me3**
9(90 %)	**GB +/-5kb H3k36me3**
9(90 %)	**GB +/-5kb_onlyTTS H3k36me3**
8(80 %)	genebody Ctcf
9(90 %)	Promoter 2kb H3k27me3
10(100 %)	Promoter 2kb H3k4me3
8(80 %)	**tail H3k36me3**
9(90 %)	**TSS +/- 5kb H3k36me3**

Supplementary Table 3: Selected features (>=80%) pair1 and pair2 and both.

Feature selection is done on intron containing genes for each region separately.

LCG – IC

	exon1	**GB3ss**	**GB+/-1kb**	**GB+/-5kb**	**GB+5kb**	**GB**	**intron1**	**Prom 2kb**	**Prom 5kb**	**Tail**	**TSS+/-2kb**	**TSS+/-5kb**	**pA+/-2kb**
Methyl		Methyl	Methyl		Methyl	Methyl	Methyl	Methyl					
Ctcf													CTCF
Dnase	**Dnase**	Dnase		Dnase	Dnase	Dnase		**Dnase**	**Dnase**	Dnase	Dnase		
H3K27ac	**H3K27ac**	H3K27ac	**H3K27ac**	**H3K27ac**	**H3K27ac**	H3K27ac	**H3K27ac**	**H3K27ac**	**H3K27ac**	H3K27ac	**H3K27ac**	**H3K27ac**	H3K27ac
H3K27me3	H3K27me3	H3K27me3	**H3K27me3**	**H3K27me3**	**H3K27me3**	H3K27me3	**H3K27me3**	**H3K27me3**	**H3K27me3**	**H3K27me3**	**H3K27me3**	**H3K27me3**	H3K27me3
H3K36me3	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**
H3K4me1								H3K4me1	H3K4me1	H3K4me1			
H3K4me2	**H3K4me2**		**H3K4me2**			**H3K4me2**	**H3K4me2**	**H3K4me2**	**H3K4me2**		**H3K4me2**	**H3K4me2**	
H3K4me3	**H3K4me3**	H3K4me3	**H3K4me3**	**H3K4me3**	H3K4me3		**H3K4me3**	**H3K4me3**	**H3K4me3**		**H3K4me3**	**H3K4me3**	
H3K9ac	**H3K9ac**	H3K9ac	**H3K9ac**	**H3K9ac**	H3K9ac	H3K9ac	**H3K9ac**	**H3K9ac**	**H3K9ac**		**H3K9ac**	H3K9ac	
H4K20me1													

HCG - IC

Methyl	**exon1**	**GB3ss**	**GB+/-1kb**	**GB+/-5kb**	**GB+5kb**	**GB**	**intron1**	**Prom 2kb**	**Prom 5kb**	**Tail**	**TSS+/-2kb**	**TSS+/-5kb**	**pA+/-2kb**
Ctcf													
Dnase	Dnase		Dnase	Dnase		Dnase	Dnase	Dnase	Dnase	Dnase	Dnase	Dnase	
H3K27ac	**H3K27ac**	H3K27ac	**H3K27ac**	**H3K27ac**	**H3K27ac**	**H3K27ac**	**H3K27ac**	**H3K27ac**	**H3K27ac**	H3K27ac	**H3K27ac**	**H3K27ac**	
H3K27me3		H3K27me3		**H3K27me3**	H3K27me3			H3K27me3	H3K27me3	**H3K27me3**	H3K27me3	**H3K27me3**	
H3K36me3		**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**	**H3K36me3**		H3K36me3	**H3K36me3**	H3K36me3	**H3K36me3**	**H3K36me3**
H3K4me1	H3K4me1							H3K4me1	H3K4me1			**H3K4me1**	
H3K4me2	H3K4me2		H3K4me2	**H3K4me2**	**H3K4me2**	H3K4me2	H3K4me2	H3K4me2	H3K4me2		H3K4me2	H3K4me2	
H3K4me3	**H3K4me3**	H3K4me3	**H3K4me3**	**H3K4me3**	**H3K4me3**	**H3K4me3**	**H3K4me3**	**H3K4me3**	**H3K4me3**		**H3K4me3**	**H3K4me3**	
H3K9ac	H3K9ac	H3K9ac	H3K9ac	**H3K9ac**	H3K9ac	H3K9ac	H3K9ac	H3K9ac	H3K9ac		**H3K9ac**	**H3K9ac**	
H4K20me1													

Supplementary Table 4: AUC for prediction on selected features that overlap from pair 1 and pair 2.

Predictions are done on pair 2 in intron containing genes for each region separately. Beforehand features have been selected that have a CFS score >=80% (black and bold in Supplementary Table 2).

LCG-IC -using only intersecting attributes (black)

	exon1	GB3ss	GB+/-1kb	GB+/-5kb	GB+5kb	GB	intron1	Prom 2kb	Prom 5kb	Tail	TSS+/-2kb	TSS+/-5kb	pA+/-2kb	all
UP	0.84	0.88	0.92	0.92	0.91	0.91	0.89	0.88	0.91	0.89	0.93	0.92	0.85	0.86
DW	0.82	0.79	0.91	0.91	0.88	0.88	0.86	0.86	0.89	0.82	0.92	0.89	0.78	0.89
nonReg	0.73	0.73	0.86	0.86	0.84	0.82	0.81	0.76	0.81	0.74	0.83	0.84	0.69	0.76
average	0.8	0.8	0.9	0.9	0.88	0.87	0.85	0.83	0.87	0.82	0.89	0.88	0.78	0.84

HCG-IC -using only intersecting attributes (black)

	exon1	GB3ss	GB+/-1kb	GB+/-5kb	GB+5kb	GB	intron1	Prom 2kb	Prom 5kb	Tail	TSS+/-2kb	TSS+/-5kb	pA+/-2kb	all
UP	0.79	0.82	0.87	0.88	0.88	0.86	0.85	0.77	0.78	0.76	0.86	0.88	0.81	0.79
DW	0.79	0.76	0.85	0.86	0.86	0.86	0.82	0.77	0.77	0.75	0.86	0.88	0.72	0.82
nonReg	0.76	0.7	0.81	0.8	0.82	0.81	0.79	0.67	0.7	0.66	0.82	0.83	0.69	0.75
average	0.78	0.76	0.84	0.85	0.85	0.84	0.82	0.73	0.75	0.72	0.85	0.86	0.74	0.79

Supplementary methods:

Command lines to run pipeline based on XML-files:

First download files from: http://regulatorygenomics.upf.edu/XML_ENCODE/

# run enrichment analysis

```
sh rgmartutils/enrichment -c config_K562_Gm12878.xml
sh rgmartutils/enrichment -c config_Hsmm_Hmec.xml
```

(in stdout you will see the name of result_folder1 and result_folder2)

# recover files:

```
sh rgmartutils/recover config_K562_Gm12878.xml result_folder1
sh rgmartutils/recover config_Hsmm_Hmec.xml result_folder2
```

# to upload in biomart:

```
sh rgmartutils/dbimport -c config_K562_Gm12878.xml
sh rgmartutils/dbimport -c config_Hsmm_Hmec.xml
```

Command lines for Pyicos (version 1.0.3) enrichment analysis:

For ChIP-Seq and Dnase-Seq:

```
pyicos enrichment wgEncodeBroadHistoneK562H3k9acStdAlnRep1.filtered.sam
wgEncodeBroadHistoneGm12878H3k9acStdAlnRep1.filtered.sam RESULT_pA +/-
2kb__k562_Gm12878__H3k9ac.enrichment -o -f sam --replica-a
wgEncodeBroadHistoneK562H3k9acStdAlnRep2.filtered.sam -o --region pA +/- 2kb.bed
--region-format bed --n-norm --len-norm --binstep 1000 --pseudocount
```

For Methylation data (mean methylation with 0.1 as pseudocount):

```
pyicos enrichcount K562_Gm_meth_genebody.mean.pc RESULT_K562_Gm_meth_genebody.mean.enr
--total-reads-a 10000000 --total-reads-b 10000000 # total reads are made up and do not
matter in this case as we do not normalize
```

For RNA-Seq (RPKMs from ENCODE with 3.5e-5 as pseudocount (half of minimum in K562_1)):

```
pyicos enrichcount RPKM.K562_Gm12878 RESULT_RPKM_K562_Gm12878.enr --total-reads-a
10000000 --total-reads-b 10000000 # total reads are made up and do not matter in this case as
we do not normalize
```

Biomart-powered database

We used Biomart [1] as the platform for deploying a set of databases with enrichment data between different cell lines for several chromatin marks, DNA-interacting proteins and chromatin-interacting proteins, over a wide range of annotated regions in the human genome. Each database may include a number of datasets, one per each pair of cell lines compared in terms of enrichment. We populated the Biomart database with enrichment z-scores for the datasets from Table 2. These are stored as a set of feature tables, where each feature is a pair (signal,region), .e.g. RNAPII-genebody or H3K27me3-TSS +/- 2kb, built from the datasets from Table 2 and the regions described in Table 3. Together with this Biomart database, we installed a local mirror of Ensembl Biomart (Release 54) [2] and modified the Biomart platform in order to make crossed queries possible between our set of databases and the Ensembl Release 54 Mart Database. Data contained in this Biomart-powered set of databases can also be accessed independently of our ML framework through the website http://regulatorygenomics.upf.edu/group/pages/software .

Bibliography

[1] A. Kasprzyk, “BioMart: driving a paradigm change in biological data management,” *Database: The Journal of Biological Databases and Curation*, vol. 2011, p. bar049, 2011.

[2] P. Flicek et al., “Ensembl 2012,” *Nucleic Acids Research*, vol. 40, no. Database issue, pp. D84-90, Jan. 2012.

Bibliography

[1000 Genomes Project Consortium, 2010] 1000 Genomes Project Consortium, e. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

[Altshuler et al., 2008] Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322(5903):881–888.

[Anders and Huber, 2010] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106.

[B. Alberts and Raff, 2008] B. Alberts, A.Johnson, J. L. and Raff, M. (2008). Molecular biology of the cell. *Garland Science*, 5:341.

[Barski et al., 2007] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837.

[Barski and Zhao, 2009] Barski, A. and Zhao, K. (2009). Genomic location analysis by chip-seq. *J Cell Biochem*, 107(1):11–18.

[Bentley, 2005] Bentley, D. L. (2005). Rules of engagement: co-transcriptional recruitment of pre-mrna processing factors. *Curr Opin Cell Biol*, 17(3):251–256.

[Bernstein et al., 2010] Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., and Thomson, J. A. (2010). The nih roadmap epigenomics mapping consortium. *Nat Biotechnol*, 28(10):1045–1048.

[Bird, 2002] Bird, A. (2002). Dna methylation patterns and epigenetic memory. *Genes Dev*, 16(1):6–21.

[Boeva et al., 2010] Boeva, V., Surdez, D., Guillon, N., Tirode, F., Fejes, A. P., Delattre, O., and Barillot, E. (2010). De novo motif identification improves the accuracy of predicting transcription factor binding sites in chip-seq data analysis. *Nucleic Acids Res*, 38(11):e126.

[Boyle et al., 2008] Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322.

[Collins and Barker, 2007] Collins, F. S. and Barker, A. D. (2007). Mapping the cancer genome. pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am*, 296(3):50–57.

[Consortium et al., 2010] Consortium, I. C. G., Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernab, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D. P., Laplace, F., Youyong, L., Nettekoven, G., Ozenberger, B., Peterson, J., Rao, T. S., Remacle, J., Schafer, A. J., Shibata, T., Stratton, M. R., Vockley, J. G., Watanabe, K., Yang, H., Yuen, M. M. F., Knoppers, B. M., Bobrow, M., Cambon-Thomsen, A., Dressler, L. G., Dyke, S. O. M., Joly, Y., Kato, K., Kennedy, K. L., Nicols, P., Parker, M. J., Rial-Sebbag, E., Romeo-Casabona, C. M., Shaw, K. M., Wallace, S., Wiesner, G. L., Zeps, N., Lichter, P., Biankin, A. V., Chabannon, C., Chin, L., Clment, B., de Alava, E., Degos, F., Ferguson, M. L., Geary, P., Hayes, D. N., Hudson, T. J., Johns, A. L., Kasprzyk, A., Nakagawa, H., Penny, R., Piris, M. A., Sarin, R., Scarpa, A., Shibata, T., van de Vijver, M., Futreal, P. A., Aburatani, H., Bays, M., Botwell, D. D. L., Campbell, P. J., Estivill, X., Gerhard, D. S., Grimmond, S. M., Gut, I., Hirst, M., Lpez-Otn, C., Majumder, P., Marra, M., McPherson, J. D., Nakagawa, H., Ning, Z., Puente, X. S., Ruan, Y., Shibata, T., Stratton, M. R., Stunnenberg, H. G., Swerdlow, H., Velculescu, V. E., Wilson, R. K., Xue, H. H., Yang, L., Spellman, P. T., Bader, G. D., Boutros, P. C., Campbell, P. J., Flicek, P., Getz, G., Guig, R., Guo, G., Haussler, D., Heath, S., Hubbard, T. J., Jiang, T., Jones, S. M., Li, Q., Lpez-Bigas, N., Luo, R., Muthuswamy, L., Ouellette, B. F. F., Pearson, J. V., Puente, X. S., Quesada, V., Raphael, B. J., Sander, C., Shibata, T., Speed, T. P., Stein, L. D., Stuart, J. M., Teague, J. W., Totoki, Y., Tsunoda, T., Valencia, A., Wheeler, D. A., Wu, H., Zhao, S., Zhou, G., Stein, L. D., Guig, R., Hubbard, T. J., Joly, Y., Jones, S. M., Kasprzyk, A., Lathrop, M., Lpez-Bigas, N., Ouellette, B. F. F., Spellman, P. T., Teague, J. W., Thomas, G., Valencia, A., Yoshida, T., Kennedy, K. L., Axton, M., Dyke, S. O. M., Futreal, P. A., Gerhard, D. S., Gunter, C., Guyer, M., Hudson, T. J., McPherson, J. D., Miller, L. J., Ozenberger, B., Shaw, K. M., Kasprzyk, A., Stein, L. D., Zhang, J., Haider, S. A., Wang, J., Yung, C. K., Cros, A., Cross, A., Liang, Y., Gnaneshan, S., Guberman, J., Hsu, J., Bobrow, M., Chalmers, D. R. C., Hasel, K. W., Joly, Y., Kaan, T. S. H., Kennedy, K. L., Knoppers, B. M., Lowrance, W. W., Masui, T., Nicols, P., Rial-Sebbag, E., Rodriguez, L. L., Vergely, C., Yoshida, T., Grimmond, S. M., Biankin, A. V., Bowtell, D. D. L., Cloonan, N., deFazio, A., Eshleman, J. R., Etemadmoghadam, D., Gardiner, B. B., Gardiner, B. A., Kench, J. G., Scarpa, A., Sutherland, R. L., Tempero, M. A., Waddell, N. J., Wilson, P. J., McPherson, J. D., Gallinger, S., Tsao, M.-S., Shaw, P. A., Petersen, G. M., Mukhopadhyay, D., Chin, L., DePinho, R. A., Thayer, S., Muthuswamy, L., Shazand, K., Beck, T., Sam, M., Timms, L., Ballin, V., Lu, Y., Ji, J., Zhang, X., Chen, F., Hu, X., Zhou, G., Yang, Q., Tian, G., Zhang, L., Xing, X., Li, X., Zhu, Z., Yu, Y., Yu, J., Yang, H., Lathrop, M., Tost, J., Brennan, P., Holcatova, I., Zaridze, D., Brazma, A., Egevard, L., Prokhortchouk, E., Banks, R. E., Uhln, M., Cambon-Thomsen, A., Viksna, J., Ponten, F., Skryabin, K., Stratton, M. R., Futreal, P. A., Birney, E., Borg, A., Brresen-Dale, A.-L., Caldas, C., Foekens, J. A., Martin, S., Reis-Filho, J. S., Richardson, A. L., Sotiriou, C., Stunnenberg, H. G., Thoms, G., van de Vijver, M., van't Veer, L., Calvo, F., Birnbaum, D., Blanche, H., Boucher, P., Boyault, S., Chabannon, C., Gut, I., Masson-Jacquemier, J. D., Lathrop, M., Pauport, I., Pivot, X., Vincent-Salomon, A., Tabone, E., Theillet, C., Thomas, G., Tost, J., Treilleux, I., Calvo, F., Bioulac-Sage, P., Clment, B., Decaens, T., Degos, F., Franco, D., Gut, I., Gut, M., Heath, S., Lathrop, M., Samuel, D., Thomas, G., Zucman-Rossi, J., Lichter, P., Eils, R., Brors, B., Korbel, J. O., Korshunov, A., Landgraf, P., Lehrach, H., Pfister, S., Radlwimmer, B., Reifenberger, G., Taylor, M. D., von Kalle, C., Majumder, P. P., Sarin, R., Rao, T. S., Bhan, M. K., Scarpa, A., Pederzoli, P., Lawlor, R. A., Delledonne, M., Bardelli, A., Biankin, A. V., Grimmond, S. M., Gress, T., Klimstra, D., Zamboni, G., Shibata, T., Nakamura, Y., Nakagawa, H., Kusada, J., Tsunoda, T., Miyano, S., Aburatani, H., Kato, K., Fujimoto, A., Yoshida, T., Campo, E., Lpez-Otn, C., Estivill, X., Guig, R., de Sanjos, S., Piris, M. A., Montserrat, E., Gonzlez-Daz, M., Puente, X. S., Jares, P., Valencia, A., Himmelbauer, H., Himmelbaue, H., Quesada, V., Bea, S., Stratton, M. R., Futreal, P. A., Campbell, P. J., Vincent-Salomon, A., Richardson, A. L., Reis-Filho, J. S., van de Vijver, M., Thomas, G., Masson-Jacquemier, J. D., Aparicio, S., Borg, A., Brresen-Dale, A.-L., Caldas, C., Foekens, J. A., Stunnenberg, H. G., van't Veer, L., Easton, D. F., Spellm (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998.

[ENCODE Project Consortium, 2007] ENCODE Project Consortium, e. (2007).

Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816.

[ENCODE Project Consortium, 2011] ENCODE Project Consortium, e. (2011). A user's guide to the encyclopedia of dna elements (encode). *PLoS Biol*, 9(4):e1001046.

[Ernst and Kellis, 2010] Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, 28(8):817–825.

[Fejes et al., 2008] Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., and Jones, S. J. M. (2008). Findpeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730.

[Felsenfeld and Groudine, 2003] Felsenfeld, G. and Groudine, M. (2003). Controlling the double helix. *Nature*, 421(6921):448–453.

[Guttman et al., 2010] Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat Biotechnol*, 28(5):503–510.

[Hansen et al., 2012] Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in rna-seq data using conditional quantile normalization. *Biostatistics*.

[Hawkins et al., 2010] Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat Rev Genet*, 11(7):476–486.

[Hon et al., 2008] Hon, G., Ren, B., and Wang, W. (2008). Chromasig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput Biol*, 4(10):e1000201.

[Jenuwein and Allis, 2001] Jenuwein, T. and Allis, C. D. (2001). Translating the histone code. *Science*, 293(5532):1074–1080.

[Jiang and Pugh, 2009] Jiang, C. and Pugh, B. F. (2009). Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*, 10(3):161–172.

[Johnson et al., 2007] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502.

[Kahn, 2011] Kahn, S. D. (2011). On the future of genomic data. *Science*, 331(6018):728–729.

[Khalil et al., 2009] Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., Regev, A., Lander, E. S., and Rinn, J. L. (2009). Many human large intergenic noncoding rnas associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A*, 106(28):11667–11672.

[Kornberg and Thomas, 1974] Kornberg, R. D. and Thomas, J. O. (1974). Chromatin structure; oligomers of the histones. *Science*, 184(139):865–868.

[Kornblihtt et al., 2004] Kornblihtt, A. R., de la Mata, M., Fededa, J. P., Munoz, M. J., and Nogues, G. (2004). Multiple links between transcription and splicing. *RNA*, 10(10):1489–1498.

[Kozarewa et al., 2009] Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. (2009). Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (g+c)-biased genomes. *Nat Methods*, 6(4):291–295.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and , I. H. G. S. C. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Langmead et al., 2009] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[Li et al., 2008] Li, H., Ruan, J., and Durbin, R. (2008). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858.

[Licatalosi et al., 2008] Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008). Hits-clip yields genome-wide insights into brain alternative rna processing. *Nature*, 456(7221):464–469.

[Meissner et al., 2005] Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic Acids Res*, 33(18):5868–5877.

[Metzker, 2010] Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46.

[Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–628.

[Nagalakshmi et al., 2008] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349.

[Nikolaou et al., 2010] Nikolaou, C., Althammer, S., Beato, M., and Guig, R. (2010). Structural constraints revealed in consistent nucleosome positions in the genome of s. cerevisiae. *Epigenetics Chromatin*, 3(1):20.

[Nix et al., 2008] Nix, D. A., Courdy, S. J., and Boucher, K. M. (2008). Empirical methods for controlling false positives and estimating confidence in chip-seq peaks. *BMC Bioinformatics*, 9:523.

[Pan et al., 2008] Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415.

[Pareek et al., 2011] Pareek, C. S., Smoczynski, R., and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *J Appl Genet*, 52(4):413–435.

[Park, 2009] Park, P. J. (2009). Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680.

[Pepke et al., 2009] Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for chip-seq and rna-seq studies. *Nat Methods*, 6(11 Suppl):S22–S32.

[Quail et al., 2008] Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., and Turner, D. J. (2008). A large genome center's improvements to the illumina sequencing system. *Nat Methods*, 5(12):1005–1010.

[Ren et al., 2000] Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000). Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–2309.

[Richmond and Davey, 2003] Richmond, T. J. and Davey, C. A. (2003). The structure of dna in the nucleosome core. *Nature*, 423(6936):145–150.

[Robertson et al., 2007] Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8):651–657.

[Robertson, 2005] Robertson, K. D. (2005). Dna methylation and human disease. *Nat Rev Genet*, 6(8):597–610.

[Robinson et al., 2010] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[Robinson and Oshlack, 2010] Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25.

[Roy et al., 2011] Roy, N. C., Altermann, E., Park, Z. A., and McNabb, W. C. (2011). A comparison of analog and next-generation transcriptomic tools for mammalian studies. *Brief Funct Genomics*, 10(3):135–150.

[Schena et al., 1995] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470.

[Segal et al., 2006] Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thstrm, A., Field, Y., Moore, I. K., Wang, J.-P. Z., and Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778.

[Song and Crawford, 2010] Song, L. and Crawford, G. E. (2010). Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*, 2010(2):pdb.prot5384.

[Tilgner et al., 2009] Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcrcel, J., and Guig, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol*, 16(9):996–1001.

[Trapnell et al., 2010] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515.

[Ule et al., 2005] Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005). Clip: a method for identifying protein-rna interaction sites in living cells. *Methods*, 37(4):376–386.

[Venter et al., 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guig, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

[Vicent et al., 2011] Vicent, G. P., Nacht, A. S., Font-Mateu, J., Castellano, G., Gaveglia, L., Ballar, C., and Beato, M. (2011). Four enzymes cooperate to displace histone h1 during the first minute of hormonal gene activation. *Genes Dev*, 25(8):845–862.

[Waalwijk and Flavell, 1978] Waalwijk, C. and Flavell, R. A. (1978). Mspi, an isoschizomer of hpaii which cleaves both unmethylated and methylated hpaii sites. *Nucleic Acids Res*, 5(9):3231–3236.

[Waddington, 1953] Waddington, C. H. (1953). Principles of embryology. *Sym. Soc. Exp. Biol.*, 7:186.

[Wang et al., 2010] Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). Degseq: an r package for identifying differentially expressed genes from rna-seq data. *Bioinformatics*, 26(1):136–138.

[Wang et al., 2008] Wang, X., Sun, Q., McGrath, S. D., Mardis, E. R., Soloway, P. D., and Clark, A. G. (2008). Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One*, 3(12):e3839.

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63.

[Xue et al., 2009] Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D. L., Sun, H., Fu, X.-D., and Zhang, Y. (2009). Genome-wide analysis of ptb-rna interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell*, 36(6):996–1006.

[Yeo et al., 2009] Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X.-D., and Gage, F. H. (2009). An rna code for the fox2 splicing regulator revealed by mapping rna-protein interactions in stem cells. *Nat Struct Mol Biol*, 16(2):130–137.

[Yu et al., 2008] Yu, H., Zhu, S., Zhou, B., Xue, H., and Han, J.-D. J. (2008). Inferring causal relationships among different histone modifications and gene expression. *Genome Res*, 18(8):1314–1324.

[Zhang et al., 2008] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137.