# Computational tool for visualization, analysis and comparison of epigenomes

Óscar Reina García

TESI DOCTORAL UPF / ANY 2017

DIRECTOR DE LA TESI

Dra. Camille Stephan-Otto Attolini

Biostatistics and Bioinformatics, IRB Barcelona.

Dr. Fernando Azorín Marín

Chromatin Structure and Function, IRB Barcelona. Institut de Biologia Molecular de Barcelona.

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Universitat Pompeu Fabra
Barcelona

A mi familia, presente, pasada y futura.

# Agradecimientos

Antes de nada, quiero pedir perdón por adelantado a todos los que seguro me voy olvidar de mencionar, es lo que tiene acabar estas cosas de madrugada.

Quiero dar las gracias a mis directores, Camille y Ferran, por estar siempre ahí cuando ha hecho falta, sin vosotros esta tesis no hubiese sido posible, literalmente. Tampoco lo hubiese sido sin David Rossell ni Joan Font, que desde la estadística y la biología respectivamente supieron colocar la mosca adecuada en el anzuelo para hacerme picar y mantenerme pendiente del hilo y que además no dudaron en meterse en el río cuando hizo falta. A mis compañeros en el IRB, tanto los que han compartido horas y despacho (Evarist, Camille, Toni, Martina, José Luis y Adrià), como a todos demás los que habéis venido a buscar respuestas y seguramente os habéis llevado más preguntas, aunque por suerte muchas veces acompañadas de café o chocolate.

A Jordi Bernués y Lluïsa Espinàs y a toda la tropa del Azorin-lab pasada y presente, porque sin vosotros esta tesis tampoco hubiese sido lo mismo, de nuevo, literalmente. A las chicas-RIP, (hay vida más allá del ChIP). A Nuria Centeno, por toda la ayuda prestada como tutora durante estos años. A los compañeros de fatigas en el ICO, David, Raúl, Eli, (Laura)^3, Estherilla, Xavi, Raquel, Víctor, Xescu, MA, Núria, Susana Joan y Olga, a Mireia V. et al.,  Sergi, a Julio, por saber también lanzar la mosca al sitio adecuado. Sr. Dani y Sr. Abel, esto va para vosotros. A los amigos que me están esperando al otro lado, por favor no os enfadéis si os pido que esperéis un poco más. Luis, tú tampoco te vas a escapar, te prometo que me voy a poner con Magit en cuanto pueda. A la conexión Gijonenca, aunque seáis también de León o de Bilbao, gracias por adoptarme desde el primer día, se os echa de menos desde que se da el primer paso para volver. Queralt, a ver cuando volvemos a ponernos en marcha. Eva, gracias por tanto.

Gracias a Guillaume Filion por la inspiración de sus colores de la cromatina, y cómo no, también por esos datos ‘curados’ de K562, que no me ha venido nada mal. A Jan Graffelman, por sus siempre interesantes consejos. A la fotografía, por enseñarme el camino del ver y del ser. Al ciclismo, por darme tanto a cambio de tan poco. A todos los que habéis compartido un trozo del camino que me ha traído hasta aquí y con quienes espero compartir otros cuantos. Y por supuesto a Roman Kessler y Sergi Cuartero, mis compañeros en un viaje inesperado a través de la cromatina.

# Abstract

We developed a computational framework implemented as an R package for generation, visualization and functional and differential analysis of epigenome maps. Methods are provided for integrating and comparing data from different conditions or biological backgrounds, accounting and adjusting for systematic biases in order to provide an efficient and statistically robust base for differential analysis. We also provide methods for general data assessment and quality control, such as functions to study chromatin domain conservation between epigenomic backgrounds, to detect gross technical outliers and to help in the selection of candidate marks for de-novo epigenome mapping.

# Resum

Hem desenvolupat una metodologia computational implementada en forma de paquet pel llenguatge R per la generació i visualització de mapes epigenòmics, així com per dur a terme el seu anàlisi funcional i diferencial. Proporcionem mètodes per la integració i comparació de dades provinents de diferents condicions, identificant i eliminant biaixos sistemàtics per obtenir una base amb robustesa estadística per l'anàlisi diferencial. També proporcionem funcions per dur a terme un control de qualitat de les dades, per estudiar la conservació i integritat dels dominis de cromatina, per detectar errors tècnics i per ajudar en la selecció de factors epigenètics candidats per la generació de mapes epigenòmics 'de-novo'.

”The known is finite, the unknown infinite; intellectually we stand on an islet in the midst of an illimitable ocean of inexplicability.”

T. H. Huxley. 1886

# Contents

# List of Figures

# Chapter 1

# Introduction

## 1.1 Foreword

The human species does not possess the most complex visual system in the animal world. However, seeing has played a crucial role not only in our ability to survive, but in the development of a language, in the acquisition and the transmission of knowledge, in the generation of a culture, and therefore in discovering and understanding. From the infinitely small universe seen by Leuwenhoek to the Jovian moons of Galilei, our minds have walked the path of Science stepping on the bricks of observation. So, it is not surprising that in the era of data, with millions of invisible pieces that only exist as a binary state in an electronic nature, the ability to visualize what we know is more important than ever.

## 1.2 Prologue

The genome of eukaryotic organisms, that is, organisms whose cells contain a cell nucleus and other organelles surrounded by a membrane, is composed both of genetic information, encoded in the DNA sequence, and also epigenetic instructions that, contained in DNA-associated factors (such as regulatory RNAs, histone and non-histone proteins), regulate its expression. This complex of DNA, RNA and associated proteins is known as chromatin, from the Greek word 'khroma' (color), and is essential for DNA packaging within the cell nucleus, DNA damage prevention and DNA replication. Walther Flemming gave chromatin its name due to the ability of these structures to absorb color dyes and therefore become visible to the human eye under the microscope [Hardy et al., 2009]. Full understanding of genome function and regulation requires description of this epigenetic information. In other words it is necessary to describe the epigenome.

In the last decade, after sequencing the genomes of several model organisms, we have assisted to an unprecedented increase in availability of data collected on different aspects of genome functioning, ranging from gene expression and non-coding RNAs to genomic distribution of epigenetic factors. At the same time, there is also an increasingly large number of online databases and repositories related to gene functions and protein-protein interactions, and to their alterations in disease [Karnik and Meissner, 2013].

The exponential development of bioinformatics and biostatistics as necessary disciplines to process and understand the huge amounts of generated information have contributed with tools to analyze, visualize and integrate genomic data at a functional level. However, despite the fact that large efforts are being devoted to the description of the epigenome [Celniker et al., 2009; Bernstein et al., 2010; Dunham et al., 2012; Ernst and Kellis, 2010], the development of tools to integrate and visualize experimental results and databases on epigenetic factors and genetic elements remains a challenge Marx [2015]. In this context, we present a computational approach based on dimensionality reduction techniques, chroGPS (or *chromatin Global Positioning System*) to integrate, visualize, analyze and compare the associations between epigenetic factors and their relation to functional genetic elements in low dimensional epigenetic maps.

## 1.3 Epigenetics and chromatin

### 1.3.1 Beads on a string

In eukaryotic cells, DNA is found in the nucleus together with a complex of proteins that, by means of contributing to the generation of higher order structures (fig. 1.1), tightly condense it in the form of chromosomes, structures which are visible under the light microscope [Alberts et al., 2002]. This complex of DNA and proteins is known as chromatin. Chromatin can be classified in two main types. The first one, called euchromatin, presents a smaller degree of packaging and therefore it is reachable for transcriptional machinery. Heterochromatin on the other hand shows a much higher degree of condensation, and as such, it does not get usually transcribed into mRNA.

Nucleosomes are protein complexes composed by eight proteins known as core canonical histones

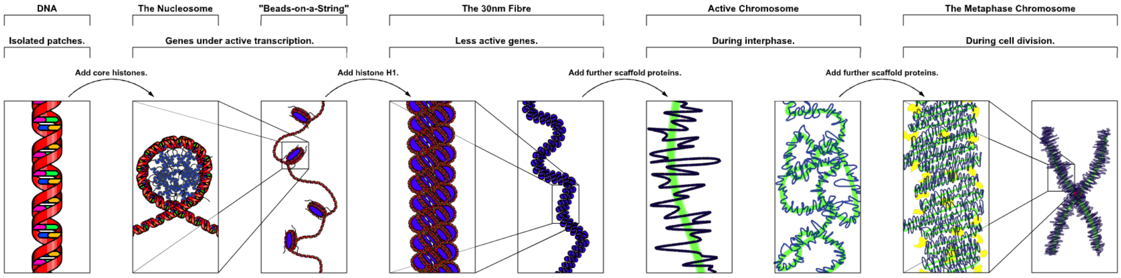


Figure 1.1: The major structures in DNA compaction: DNA, the nucleosome, the 10 nm "beads-on-a-string" fibre, the 30 nm chromatin fibre and the metaphase chromosome. Source: Richard Wheeler at en.wikipedia - Transferred from en.wikipedia to Commons by sevela.p., CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=4017531`

(specifically, 2 copies of H2A, H2B, H3 and H4 histones) that conform the basic DNA packaging unit in eukaryotic organisms (fig. 1.2). The nucleosome provides a structure for the DNA to wrap around and is essential in the conformation of higher order structures. Approximately, 146bp of DNA turn around the nucleosome in a left-handed turn. Together with around 20 and 60 base pairs of additional DNA on each side of the nucleosome, this conforms the 10nm fiber, or as is often described, the 'beads on a string' [Alberts et al., 2002; Weber and Henikoff, 2014]. Additionally, a special kind of histone, the linker histone H1 contacts the extremes of the DNA strand wrapped in the nucleosome, keeping it in place and conforming the chromatosome, which plays a crucial role in the conformation of higher order structures [Bayona-Feliu et al., 2017]. Nucleosomes are then wrapped to form the solenoid, a 30nm spiral structure which is supported by other histone proteins that contribute towards further DNA packaging within the cell nucleus.

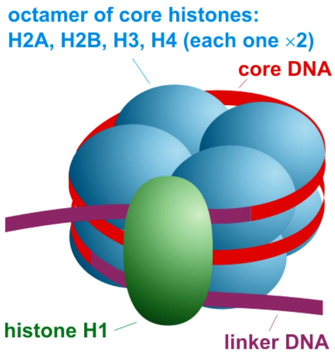


Figure 1.2: Schematic representation of the nucleosome. In blue, the eight core histones. Approximately 146bp of DNA (red) is wrapped around the nucleosome and fixed in place by linker histone H1 to help in the conformation of higher order structures. Source: Darekk2 - Own work, CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=21977693`

### Epigenetics of gene expression

Chromatin accesibility is essential in regulating transcription, and thus, gene expression [Li and Reinberg, 2011]. Histone proteins play a key role in this by carrying chemical modifications that are associated with activation or repression of transcription. The mechanistic model to explain how these dynamic modifications relate to effective transcription regulation range from direct alteration of chromatin packaging to recruitment of other chromatin binding protein complexes that indirectly initiate transcription (fig. 1.3). Chromatin modifications are often known with the general name of epigenetic marks [Salozhin et al., 2005]. As opposed to changes in DNA, changes in epigenetic marks are reversible, but it is still an open question how some chromatin epigenetic marks remain persistent after cell division [Berger, 2007]. Constitutive heterochromatin is mainly located at pericentromeric regions, that is, regions close to the nuclear chromosomal centromere, and is very poor in genes, as is basically composed of satellites (large arrays of tandemly repeating DNA) and other repeat elements. Pericentromeres are crucial structures during chromosome seggregation in mitosis yet show a very low degree of conservation between organisms, which suggest their functions could be regulated in an epigenetic manner [Saksouk et al., 2015].

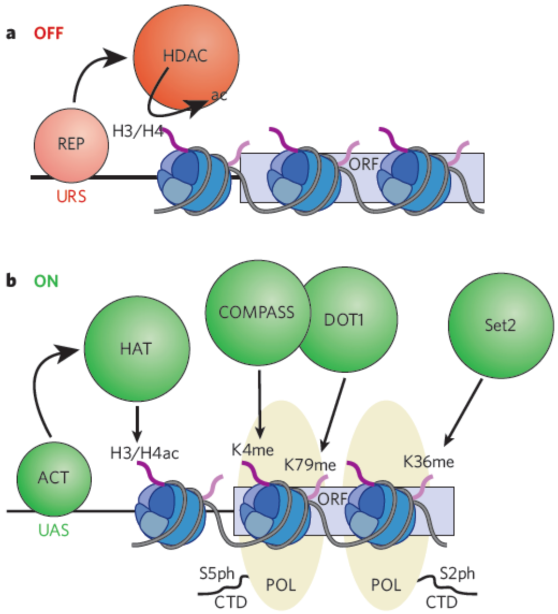


Figure 1.3: In the off state, the DNA-bound repressor (REP) at the upstream repressor site (URS) recruits negative modifiers, such as histone deacetylase (HDAC), which remove acetyl (ac) groups from histones. b, In the on state, DNA-bound activator (ACT) at the upstream activator site (UAS) recruits positive modifiers, such as histone acetylases (HAT), at the promoter, while DNA-bound RNA polymerase (POL) recruits histone methylases at the ORF. Early during elongation, the C-terminal domain (CTD) polymerase repeat is phosphorylated at serine 5 (S5ph), leading to recruitment of the COMPASS complex (Set1, part of the COMPASS complex, methylates H3K4) and DOT1 (which methylates H3K79). Later in elongation the CTD repeat is phosphorylated at serine 2 (S2ph), leading to recruitment of Set2 (which methylates H3K36). Source: [Berger, 2007]

The two main chemical modifications to histone proteins are Acetylation and Methylation, which are associated with activation and repression of transcriptional activity respectively. In Acetylation, special enzymes called histone Acetyl Transferases and Deacetylases effectively add or remove an acetyl group to N-terminal lysine residues out of the histone core of nucleosomes. Reciprocally, in histone Methylation, Methyl Transferases and Demethylases perform the analog task but this time in Histone cytosine residues. The possibility to add multiple acetyl or methyl groups, and to do that at different terminal residues gives as a result a large number of potential histone modifications, many of which have been deeply studied [Bannister and Kouzarides, 2011]. It is common to name histone modifications according to the type, number and position of their chemical modification, that is, H3k27me3 stands for a tri-methyl group in lysine 27 of histone H3. Both histone methylation and acetylation are key mechanisms in gene expression regulation, and their alteration has been associated with cancer and other diseases [Maze et al., 2014].

### 1.3.2 *Drosophila* epigenetics

*Drosophila melanogaster* is without doubt a model organism for genetic and epigenetic research. In the third decade of the last century, the relationship between genotype and phenotype was made through observation of changes in eye color in the fruit fly. By introducing mutations in the *white* gene researchers observed change in eye color from red (the normal fly eye color) to white. However, some cases of variegation, that is, partial coloration, were also observed. This was later found to be related to chromosomal inversions that somehow repressed expression of the white gene, due to the proximity to specific regions of the chromosome where certain proteins that acted as gene silencers were detected [D. et al., 2015]. These proteins had an effect on gene expression and silencing, and after being first discovered in *Drosophila* were later extensively studied in a wide range of species, from humans to yeast. The two main groups of proteins were classified in the Polycomb group (or PcG), and Trithorax (TrxG), that acted as general transcriptional gene repressors and activators respectively, through several mechanisms of histone modification and chromatin conformation remodeling [Schuettengruber et al., 2007]. With the advent of high-throughput Next Generation Sequencing techniques, the amount of epigenetic information available in *Drosophila* has grown exponentially [Kharchenko et al., 2011], and therefore it provides an ideal scenario for continued research in this field.

#### Transcription factors

Transcription factors (TF) are proteins that recognize and bind to specific DNA sequences and are involved in transcription of DNA to messenger RNA. Transcription factors recognize specific DNA-binding motifs in their target double or single-stranded DNA sequences thanks to DNA-binding domain (DBD), even though some of them present general binding affinity. These specific sequences of DNA are called enhancers and promoters. Whereas some TFs bind to DNA promoters regions near Transcriptional Start Sites (TSS), others bind to enhancer regulatory sequences that can contribute to regulation of transcription of their associated gene [Jones, 2012; Spitz and Furlong, 2012]. Enhancers can be located very far upstream from the gene whose transcription they affect. Among the different ways to control gene expression, regulation of transcription is the most common (fig. 1.4). Transcription factors directly control expression of genes along development and account for the differences in expression level in different cell types. The actual number of existing transcription factors is unknown, but in *Drosophila*, around 400 candidates from more than the evaluated 1000 were considered as site-specific TF according to the FlyTF database (`http://www.flytf.org`), whereas more than a thousand are usually considered for mammals [Vaquerizas et al., 2009].

#### Insulators

Conformation of chromatin in eukaryotic organisms leads to existence of adjacent genomic regions with completely different functions, with genes that need to be expressed in a certain tissue or

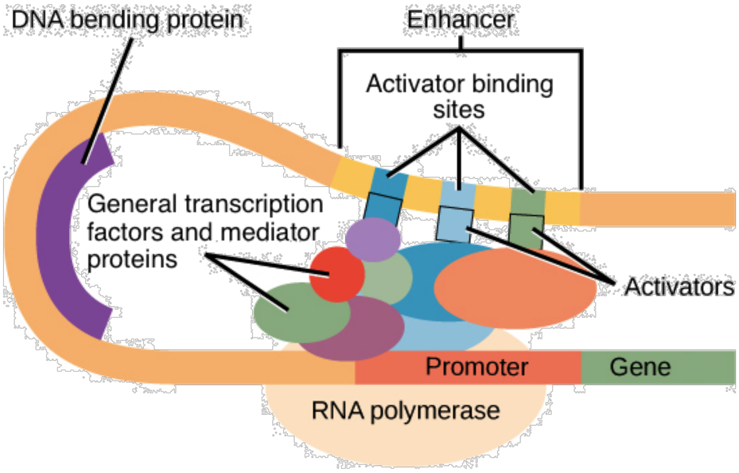


Figure 1.4: An enhancer is a DNA sequence that promotes transcription. Each enhancer is made up of short DNA sequences called distal control elements. Activators bound to the distal control elements interact with mediator proteins and transcription factors. Two different genes may have the same promoter but different distal control elements, enabling differential gene expression. Source: `https://archive.cnx.org/contents/53013107-747b-41b0-ad43-f4e97bd69ef1@2/gene-expression-eukaryotic-transcriptional-regulation-gpc`

developmental stage being close to genes that need to be silenced. It is therefore necessary to prevent inadequate interaction between these adjacent chromatin domains that could result in loss of expression regulation. Insulators are special DNA sequences recognized by insulator-binding proteins. Through this mechanism, insulator-binding proteins prevent inadequate transcriptional regulation between proximal enhancers and promoters, and also due to the spreading effect of nearby silencing heterochromatin (fig. 1.5) [Alberts et al., 2002; Gaszner and Felsenfeld, 2006; Van Bortle and Corces, 2013].

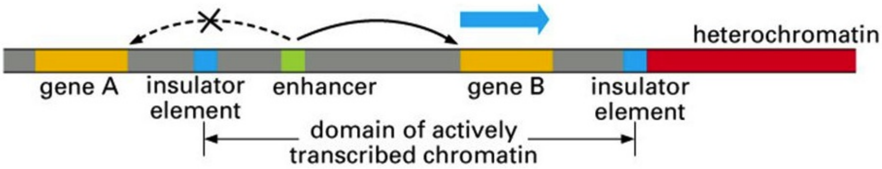


Figure 1.5: Insulators both prevent the spread of heterochromatin (right-hand side of diagram) and directionally block the action of enhancers (left-hand side). Thus gene B is properly regulated and gene B's enhancer is prevented from influencing the transcription of gene A. Source: Molecular biology of the Cell. 4th edition [Alberts et al., 2002].

Insulators are crucial for the maintenance of chromatin domains, acting as effective boundaries, and therefore in regulation of gene expression. They also have been found to be associated with the formation of chromatin loops through several molecular mechanisms [Gaszner and Felsenfeld, 2006]. Until recently, five insulators were described in *Drosophila*, Su(Hw), dCTCF (analog to mammal CTCF), Zw5, BEAF32 and in some cases GAF. However, this is a field of continued research as of today [Cuartero et al., 2014], as novel insulator/boundary related proteins are identified.

### 1.3.3 Chromatin domains

In recent years, the model of chromosomal organization and distribution within the cell nucleus has been widely studied [Dixon et al., 2016]. However, we are still far from understanding the internal structure and organization within chromosomes themselves. In this regard, the most important contribution to this field has been the discovery of topologically associated domains (TADs) using chromosome conformation capture technology (such as 3C, 4C, 5C and Hi-C) [Belton et al., 2012; Dekker et al., 2013], which allow studying the spatial organization of genomes at unprecedented resolution. TADs are self-interacting genomic regions, that is, regions that present a high level of physical interaction between themselves when compared with their degree of interaction with regions outside the TAD (fig. 1.6).

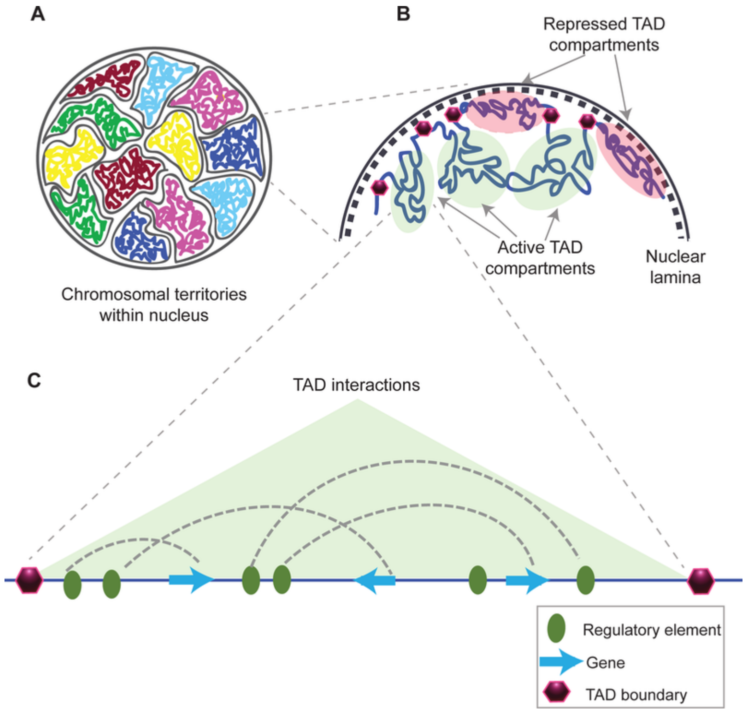


Figure 1.6: Structural organization of chromatin. (A) Chromosomes within an eukaryotic nucleus are found to occupy specific nuclear spaces, termed chromosomal territories. (B) Each chromosome is subdivided into topological associated domains (TAD) as found in Hi-C studies. TADs with repressed transcriptional activity tend to be associated with the nuclear lamina (dashed inner nuclear membrane and its associated structures), while active TADs tend to reside more in the nuclear interior. Each TAD is flanked by regions having low interaction frequencies, as determined by Hi-C, that are called TAD boundaries (purple hexagon). (C) An example of an active TAD with several interactions between distal regulatory elements and genes within it. Source: [Matharu and Ahituv, 2015]

We are just starting to understand possible functions of TADs, but nonetheless they have been found to be related to three dimensional organization of chromosomes and, as such, to alteration of gene regulation. TADs can range from hundreds of kb to actually several Mb [Dixon et al., 2016], and have been found to be stable even after many cell divisions. They also present a high degree of conservation in related species and are invariant across cell types, and they are considered to play a crucial role in chromosome folding and organization [Cremer and Cremer, 2010]. TADs have been found in *Drosophila*, human and mouse genomes, and although they have not been yet completely described in yeast, some very recent studies already report on some degree of genome compartmentalization [Tsochatzidou et al., 2017].

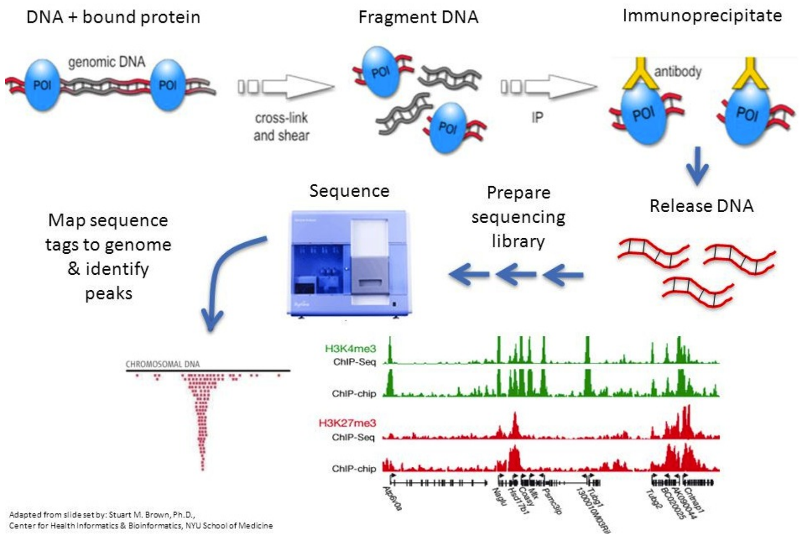
## 1.4 Chromatin analysis

### 1.4.1 Chromatin Immunoprecipitation

Chromatin Immunoprecipitation (ChIP) and related techniques are used to study the interaction between proteins and DNA. The objective of these techniques is to answer the question of which proteins are associated with specific genomic regions in a certain biological background, and to detect these protein-DNA binding sites as well as those of histone modifications and nucleosome positioning across the whole genome. In this regard, ChIP-seq, that is, chromatin immunoprecipitation experiments followed by high-throughput Next Generation Sequencing, has contributed largely not only to address a wide range of questions with a reduced cost in time and experimental resources but also to a vast improvement in resolution. ChIP-seq is often referred as the main technique addressing the question of high-troughput epigenetics, but there are actually many other techniques to approach the problem (See list below).

- ChIP-seq (Chromatin immunoprecipitation sequencing), aimed against identification of binding sites for DNA-binding proteins and histone modifications, can be used to identify chromatin states throughout the genome. Different modifications have been linked to various states of chromatin.
- DamID (DNA adenine methyltransferase identification) identifies binding sites by expressing the proposed DNA-binding protein as a fusion protein with DNA methyltransferase. Binding of the protein of interest to DNA localizes the methyltransferase in the region of the binding site.
- DNase-seq (DNase I hypersensitive sites Sequencing) uses the sensitivity of accessible regions in the genome to the DNase I enzyme to map open or accessible regions in the genome.
- FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements sequencing) uses the chemical properties of protein-bound DNA in a two-phase separation method to extract nucleosome depleted regions from the genome.
- ATAC-seq (Assay for Transposable Accessible Chromatin sequencing) uses the Tn5 transposase to integrate (synthetic) transposons into accessible regions of the genome consequentially highlighting the localisation of nucleosomes and transcription factors across the genome.
- DNA footprinting is a method aimed at identifying protein-bound DNA. It uses labeling and fragmentation coupled to gel electrophoresis to identify areas of the genome that have been bound by proteins.

- MNase-seq (Micrococcal Nuclease sequencing) uses the micrococcal nuclease enzyme to identify nucleosome positioning throughout the genome. Chromosome conformation capture determines the spatial organization of chromatin in the nucleus, by inferring genomic locations that physically interact.
- MACC profiling (Micrococcal nuclease ACCessibility profiling) uses titration series of chromatin digests with micrococcal nuclease to identify chromatin accessibility as well as to map nucleosomes and non-histone DNA-binding proteins in both open and closed regions of the genome.



Adapted from slide set by: Stuart M. Brown, Ph.D.,
Center for Health Informatics & Bioinformatics, NYU School of Medicine

Figure 1.7: A typical ChIP-seq experiment pipeline. DNA and their associated proteins are first crosslinked. This complex composed of DNA and protein as well as the rest of the cell components is subject to sonication or nuclease digestion. In order to separate protein-bound DNA fragments from the rest of material, an appropiate protein-specific antibody is used. The associated DNA fragments are purified and subject to high-throughput sequencing for further processing and bioinformatics analysis. Source: Adapted from Stuart M. Brown. Center for Health Informatics and Bioinformatics. NYU School of Medicine. `http://slideplayer.com/slide/3385783/`

In ChIP-seq (fig. 1.7), DNA and their associated proteins are first crosslinked (by treating cells with chemical exogenous or endogenous agents which as an effect fix DNA-protein interactions). Then, this complex composed of DNA and protein as well as the rest of the cell components is subject to sonication or nuclease digestion. The recommended fragment size for ChIP-seq ranges between 150 and 300bp [Kidder et al., 2011]. In order to separate protein-bound DNA fragments from the rest of material, an appropiate protein-specific antibody is used. This is often referred to as a potentially weak point in this type of immunoprecipitation technique [Landt et al., 2012], as sometimes it is difficult or nearly impossible to identify an antibody which binds properly to the protein of interest, and in many occasions non-specific binding results in the generation of experimental noise that masks proper DNA-protein interactions of interest. The associated DNA fragments are purified and subject to high-throughput sequencing for further processing and bioinformatics analysis.

### 1.4.2 Computational epigenetics

Epigenetics poses two important challenges from a computational perspective [Robinson and Pelizzola, 2015]. First, chromatin immunoprecipitation techniques based on Next Generation Sequencing deliver massive amounts of data in genome-wide distribution of several proteins and histone modifications. Even though bulk processing of these data has become easier both by the increase in raw computing power and memory, the spreading of parallel computing and also due to optimized algorithms and embedded sequencer processing that eliminates the most tedious parts of file processing such as image processing and basecalling, a bioinformatician still gets face to face with biological samples delivering individual files of several Gigabytes each that have to be stored, processed, checked for quality control, analyzed and of course interpreted. It is actually at the level of interpretation where the second challenge of these data becomes clearly visible, that is, the multiple faces of epigenetics and its combinatorial nature to control gene expression and other biological processes in a coordinated manner. These challenges have to be addressed with the combination of computer science and statistics, but also with deep biological knowledge of the intrinsic complexity of the mechanisms involved [Bock and Lengauer, 2008].

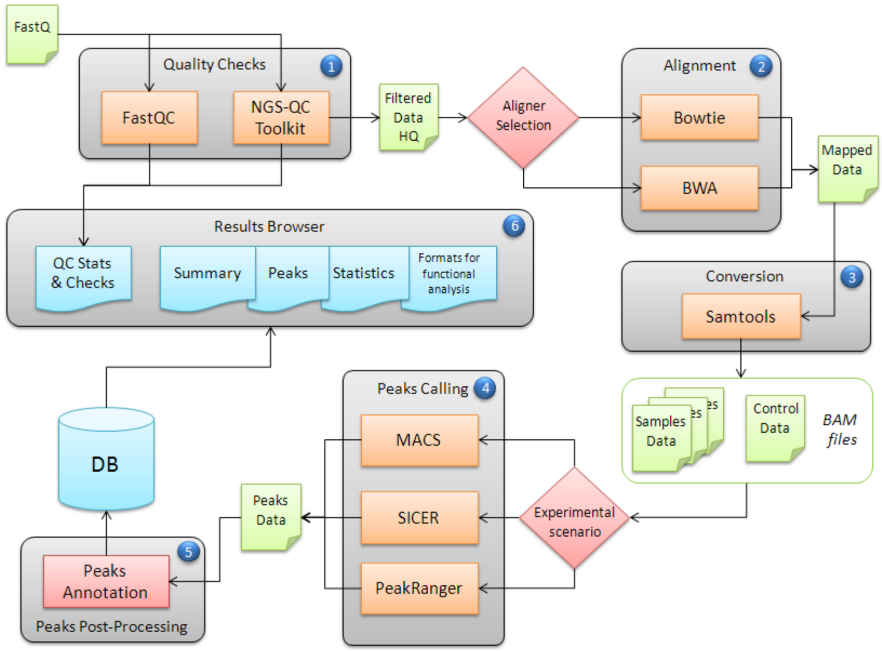


Figure 1.8: A typical ChIP-seq pipeline processing and analysis scenario comprising some of the most common steps and tools for quality control, sequencing adapter detection and removal, short-read alignment, putative binding site detection and further downstream analysis. Source: `https://bioinformatics.cineca.it/cast/workflow.php`

A typical ChIP-seq pipeline processing and analysis scenario [Nakato and Shirahige, 2017] (fig. 1.8) may start by checking the FastQ files (which contain raw sequence reads and base calling qualities from the ChIP-seq experiment itself as returned from the NGS sequencing machine), using quality control tools such as FastQC [Brown et al., 2017] to identify potential technical problems or sample contamination, and in some cases to proceed to filter the reads using read trimming or adapter cleaning softwares [Patel and Jain, 2012; Martin, 2011]. High quality data is then subject to alignment against the whole-genome DNA reference sequence of the corresponding organism or to custom built reference sequence versions. This is done using some of the commonly used short read aligner algorithms recommended for ChIP-seq data, namely Bowtie and Bowtie 2 [Langmead, 2010; Langmead and Salzberg, 2012], both based on the Burrows-Wheeler Algorithm [Canzar and Salzberg, 2017], accounting for a certain rate of sequencing errors and the potential repeated and repetitive nature of the obtained sequences and adjusting algorithm parameters as necessary. Unexpectedly low alignment rates to this reference genome can already point towards problems in sample sequencing, contamination or other issues such as abundance of repetitive sequences. After reads are aligned, tools such as sambamba or htSeqTools [Tarasov et al., 2015; Planet et al., 2012] are usually used to identify and remove putative PCR over-amplification artifacts. This is a somewhat common issue in this technique, as by its own nature immunoprecipitation followed by PCR amplification may lead to the same exact sequence being amplified many times, and it is therefore necessary to estimate the expected proportion of duplicated reads arising from true sequences under a certain False Discovery Rate (FDR) threshold [Planet et al., 2012], and exclude the rest from downstream analysis . At this step, it is normal to generate binary coverage whole-genome tracks of ChIP signal for using in genome browsers such as the Integrative Genomics Viewer [Thorvaldsdottir et al., 2013] (fig. 1.9). This is actually the first result of a ChIP-seq experiment that allows general visual inspection of binding signal across the genome of the investigated organism, and it is usually very useful for the biological researcher as a first general assessment of results.

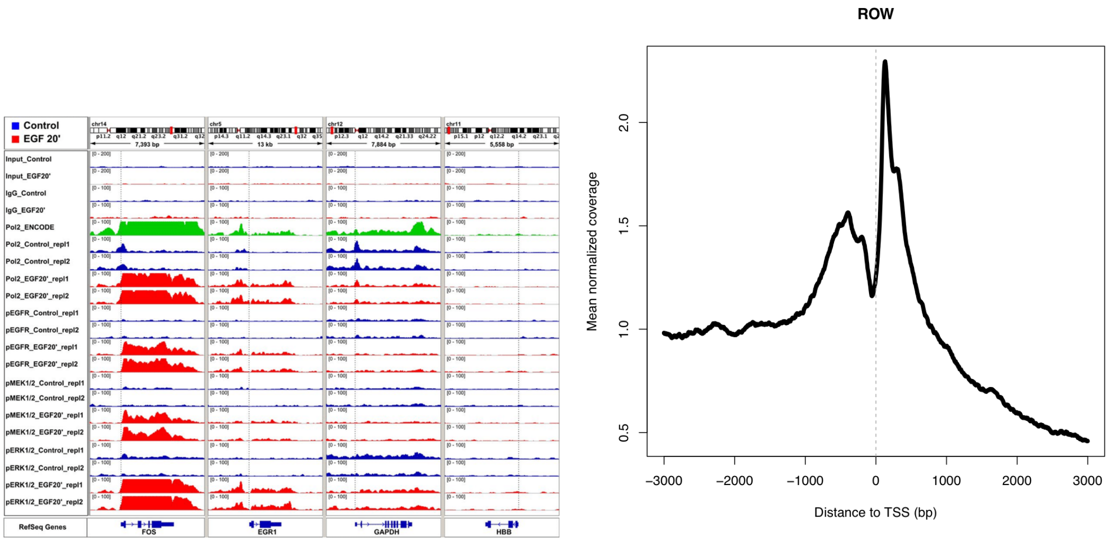


Figure 1.9: Left: ChIP-seq analysis of Pol2, EGFR and ERK kinases recruitment at selected loci in quiescent and EGF stimulated cells. BigWig input and ChIP-seq files for a given factor and condition were imported to IGV and data range was set for 100 to underscore differences between factors level at given locus. The relative enrichment of each factor at selected sites is inferred from the density of mapped fragments. Pol2 ENCODE ChIP-seq data for HeLa-S3 proliferating cells (ID: wgEncodeBroadHistoneHelas3Pol2bStdSig) were included to show similarities in transcriptional complex binding between datasets. Source: [Mikula et al., 2016]. Right: normalized binding profile around TSS of target genes for the WOC protein in *Drosophila melanogaster* S2 cells, showing both binding intensity and location [Kessler et al., 2015].

Aligned reads are then used as a source material to perform additional quality control analysis to assess immunoprecipition efficiency and general assessment of the experiment using PCA-like plots

[Planet et al., 2012; Diaz et al., 2012], and specially to identify putative protein binding sites or histone modifications using peak-calling algorithms such as MACS or MACS2, Sicer or htSeqTools [Feng et al., 2011; Planet et al., 2012; Xu et al., 2014]. Most current methods rely on modelling ChIP-seq reads based on the negative binomial distribution and use this information to detect enriched regions and candidate 'peak' sites in the immunoprecipitated samples, by comparing them with the signal obtained from control Input or IgG samples arising from performing ChIP with a non-specific antibody, usually belonging to a different species than the studied organism. These algorithms return a set of putative binding sites or histone modifications at a whole genome-level, together with a statistical assessment to control false-positive peak calls, and can already be used for result evaluation and to assess general and particular hypothesis at specific genome locations by the biological researcher.

Additionally, after read processing and binding it is usual to annotate the collection of provided binding sites to obtain the lists of genes over which they are located [Heinz et al., 2010; Zhu et al., 2010], to investigate their general location across the genome, their relative distribution and binding intensity relating to closest genomic features (fig. 1.9), and to perform functional analysis by means of Gene Ontology enrichment analysis [Bailey et al., 2013]. Individual binding site data can also be integrated with other in-house experimental information to perform further interpretative analysis, or compared with different conditions to investigate changes in overall binding site co-localization and differential binding events [Gel et al., 2016; Ross-Innes et al., 2012]. The identified collection of binding sites and their association with certain regions of the genome are also helpful as an additional quality control step to investigate potential issues due to sample contamination or to detect lack of immunoprecipitation efficiency [Carroll et al., 2014]. Finally, as previously mentioned, integration and comparison of the identified sets of binding sites for different epigenetic factors with the vast amount of available genetic and epigenetic information in the biological background of interest is crucial for further interpretation of the picture arising from this biological canvas and to obtain biologically meaningful conclusions.

### 1.4.3 The five colors of chromatin

High-throughput Next Generation Sequencing techniques allow obtaining genome-wide binding profiles for a large collection of epigenetic factors and associated proteins, and subsequently, opens the door to the application of computational methods to approach the question of chromatin analysis and classification. *Drosophila* is an ideal model organism to explore these hypothesis [Celniker et al., 2009]. After generation of Dam-ID whole-genome binding profile data for 53 chromatin associated proteins in *Drosophila melanogaster* Kc167 cells, [Filion et al., 2010] used Principal Component Analysis and Hidden-Markov models to analyze and classify the differential binding patterns obtained over fixed size bin genome segmentations. As a result, they presented a classification of chromatin on five different groups based on the combination of their unique identified binding sites. Thus, this approach introduced the now widely accepted concepts of BLUE, GREEN, BLACK, RED, and YELLOW chromatin (fig. 1.10). This classification delivered chromatin domains that ranged largely in their extension, with some extending up to hundreds of kilobases and surprisingly this low number of chromatin types reflected the known regulatory nature of their contained proteins.

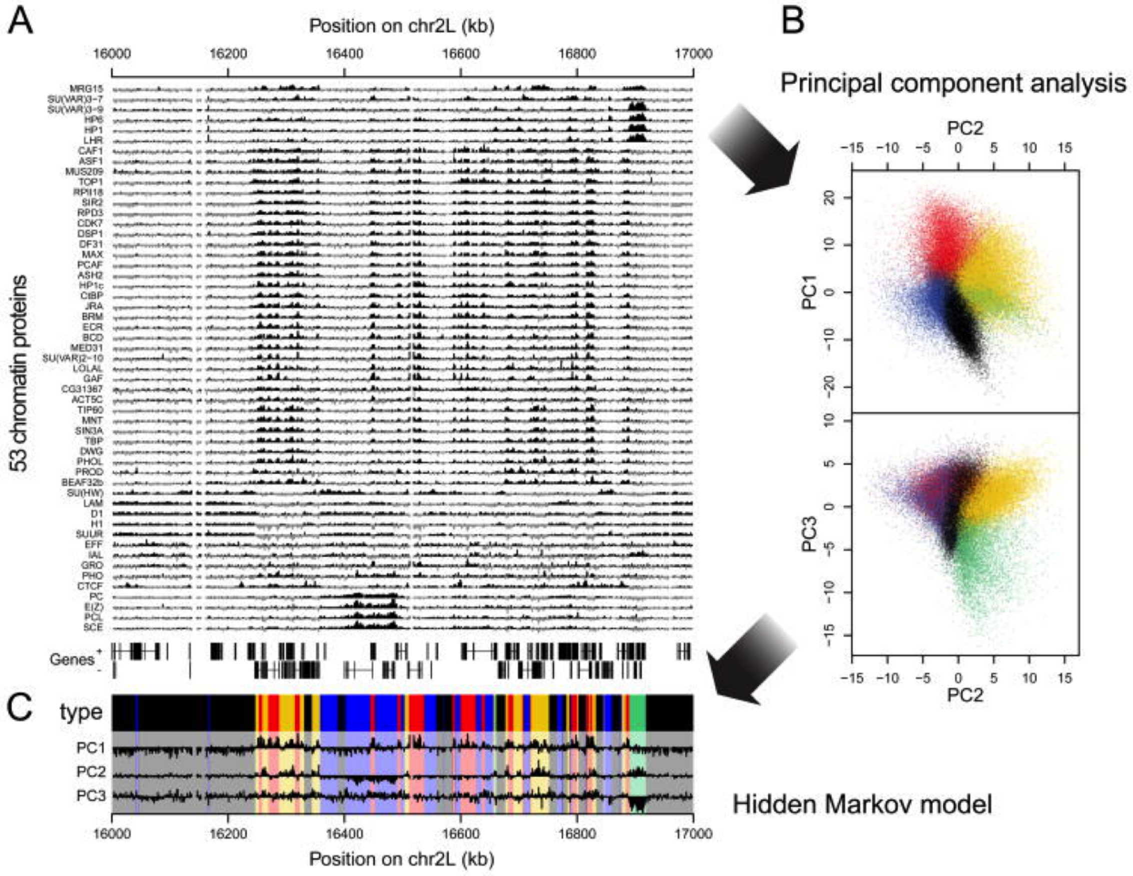


Figure 1.10: Overview of protein binding profiles and derivation of the 5-type chromatin segmentation. (A) Sample plot of all 53 DamID profiles (log2 enrichment over Dam-only control). Positive values are plotted in black, negative values in grey for contrast. Below the profiles, genes on both strands are depicted as lines with blocks indicating exons. (B) Two-dimensional projections of the data onto the first three principal components. Colored dots indicate the chromatin type of probed loci as inferred by a 5-state HMM. (C) Values of the first three principal components along the region shown in (A), with domains of the different chromatin types after segmentation by the 5-state HMM highlighted by the same colors as in (B). [Filion et al., 2010]

Authors found that BLUE and GREEN chromatin correspond to known repressed chromatin types. In detail, GREEN chromatin related to classic heterochromatin, characterized by the presence of SU(VAR)3-9, HP1, and HP1-interacting proteins, a type of chromatin typical of pericentromeric regions of chromosome 4 [Ebert et al., 2006; Greil et al., 2007]. BLUE chromatin is related to Polycomb, and as such is characterized by binding of Polycomb associated proteins PC, E(Z), PCL, and SCE. Interestingly, these findings were later verified by performing genome-wide ChIP analysis of histone modifications known to be associated with these heterochromatic types, but not used for classification using Hidden Markov Models.

In contrast, RED and YELLOW chromatin mainly represent active euchromatin. In most cases, genes belonging to these two chromatin domains presented important levels of mRNA and RNA polymerase, and were found to be enriched in H3K4me2, and H3K79me3, whereas presence of H3K9me2 and H3K27me3 was found to be low. Apart from these common features between both chromatin types, authors also found interesting differences. RED chromatin presented an abundance in several proteins that were almost totally absent in the other other chromatin types (namely SU(VAR)2-10 and MED31), and that were mostly related to nucleosome remodeling and regulation of chromosome structure, and present in various histone-modifying complexes [Martinez-Balbas et al., 1998; Tie et al., 2003]. Reciprocally, there is only one protein abundant in YELLOW chromatin but not in present in RED: MRG15, a chromodomain-containing protein reported to bind H3K36me3 [Zhang et al., 2006].

Finally, probably the most striking finding was BLACK chromatin. A chromatin domain that covers almost half of the queried genome and therefore is by far the most abundant type of chromatin in the tested biological background as well as the one presenting larger regions. Regions of BLACK chromatin have an average length of 17kb and many of the identified BLACK regions were larger than 100kb, and were characterized by presence, among others, of SU(HW) and also of Histone H1, often described as 'the big silencer of the genome' [Bayona-Feliu et al., 2017]. Even though BLACK chromatin is relatively empty of genes, it still contains more than 4000. Authors found that these genes presented very low or no transcriptional activity at all, and no presence of transcription associated marks was found. Authors conclude that while BLACK chromatin is indeed a silent type of chromatin covering large portions of the genome, some genes contained in this large domain were expressed, although only in certain tissues, suggesting that BLACK chromatin could present certain levels of dynamic regulation during development.

## 1.5 Analyzing high-dimensional data

### 1.5.1 Multidimensional Scaling (MDS)

Multidimensional scaling (MDS) [Kruskal, 1964a,b] is a statistical method originally from the field of psychometrics that represents measures of similarity (or dissimilarity) between pairs of elements as distances between points in a low-dimensional space, so that Euclidean distances between those points approximate the observed dissimilarities as much as possible [Borg and Groenen, 2003]. This graphical representation provided by MDS allows the user to literally 'look' at the data and to explore their structure from a visual point of view (which often shows regularities that remain hidden by the raw numbers), and to easily obtain a global judgment of the scale of their resemblances and differences. The four purposes of MDS are: a) to represent (dis)similarity data as distances in a low-dimensional space to make these data acessible to visual inspection and exploration; b) to test if and how certain criteria by which different objects can be differentiated are mirrored in corresponding empirical differences of these objects; c) to serve as a data analysis approach to discover the dimensions behind a set of (dis)similarities; d) to provide a psychological model to explain judgments of dissimilarity in terms of a rule that mimics a particular type of distance [Borg and Groenen, 2003].

It is important to note that the computed dissimilarities may reflect very complex characteristics of the observed set of elements in a high number of variables, and therefore, even though there are measures to estimate the accuracy of the obtained representation related to the observed dissimilarities, it is crucial to prioritize interpretability in a way that 'makes most sense' [Borg and Groenen, 2003]. Axes in MDS are arbitrary. This means that the obtained representation could be rotated in any way, or actually we could mirror the axis up to down, and left to right, and the

TABLE 1.1. Correlations of crime rates over 50 U.S. states.

| Crime | No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Murder | 1 | 1.00 | 0.52 | 0.34 | 0.81 | 0.28 | 0.06 | 0.11 |
| Rape | 2 | 0.52 | 1.00 | 0.55 | 0.70 | 0.68 | 0.60 | 0.44 |
| Robbery | 3 | 0.34 | 0.55 | 1.00 | 0.56 | 0.62 | 0.44 | 0.62 |
| Assault | 4 | 0.81 | 0.70 | 0.56 | 1.00 | 0.52 | 0.32 | 0.33 |
| Burglary | 5 | 0.28 | 0.68 | 0.62 | 0.52 | 1.00 | 0.80 | 0.70 |
| Larceny | 6 | 0.06 | 0.60 | 0.44 | 0.32 | 0.80 | 1.00 | 0.55 |
| Auto theft | 7 | 0.11 | 0.44 | 0.62 | 0.33 | 0.70 | 0.55 | 1.00 |

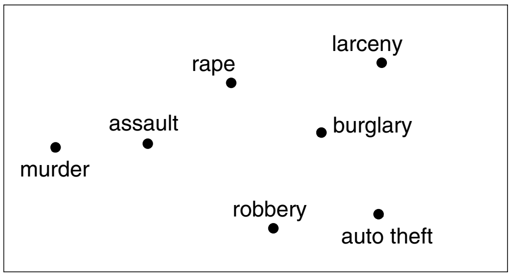


Figure 1.11: Table with crime rate correlations between the 50 U.S. States and 2-dimensional MDS with graphical representation of those correlations. Closely space points indicate crime rates which tend to be related to each other. Source: [Borg and Groenen, 2003]

distances between the represented elements would stay equal.

Classical Multidimensional Scaling [Torgerson, 1952] seeks to minimize a loss function called *stress*, which is simply the sum of squared differences $\sum_{i>j}(d_{ij} - \tilde{d}_{ij})^2$, where $d_{ij}$ is the input dissimilarity between objects $i$ and $j$ and $\tilde{d}_{ij}$ is the corresponding Euclidean distance in the low-dimensional representation. The stress is not a satisfactory measure of goodness-of-fit, as it depends on the scale of the original distances. Instead, [Kruskal, 1964a,b] proposed the *stress-1* function:

$$\sum_{i>j}(d_{ij} - \tilde{d}_{ij})^2 / \sum_{i<j} d_{ij}^2$$

which is normalized by the overall magnitude of the dissimilarities and is invariant to scale transformations of $d_{ij}$. Because the denominator in this formula depends only on the input distances $d_{ij}$, minimizing it is equivalent to minimizing *stress.* Sibson [Sibson, 1972] proposed using the $R^2$ coefficient instead, *i.e.* the squared Pearson correlation between $d_{ij}$ and $\tilde{d}_{ij}$. $R^2$ is invariant to location and scale transformations and guaranteed to be in $[0, 1]$.

isoMDS is a non-metric (or ordinal) MDS that chooses an initial k-dimensional (default k = 2) configuration (usually the one obtained from Classic MDS), to minimize the *stress* between the Euclidean distances in the obtained representation and a monotonic transformation of input dissimilarities that preserve the rank order. Starting from this initial configuration, an iterative algorithm is used to converge on an optimal solution, which mean the algorithm comes at a $O(n^2)$ cost, and therefore can become very slow for large datasets [Cox and Cox, 1994] and [Cox and Cox, 2000]. By default many non-metric MDS algorithm implementations are set up with a number of iterations that dates back to that times when computing was slow and expensive and that can actually end the process before it has converged on a local minimum, so it is important to check default settings and increase them when needed, as very small improvements in *stress* can nonetheless have a significant impact in the position of some points [Borg et al., 2012]. Several approaches have been made to address this computational problem using optimization and parallel computation [Strickert et al., 2005; Pawliczek and Dzwinel, 2009], and have been also adressed using Graphical Processing Units [Osipyan et al., 2015] with a great improvement in speed, however this is still a problem in which computational resources can limit practical usefulness.

### 1.5.2 Procrustes

In Greek mythology, Procrustes ('the one who stretches') was a rogue smith and host who offered hospitality to passing travelers with the promise of a pleasant meal and a night's rest on his special iron bed (fig. 1.12). This bed was supposed to have the unique property to match exactly the length of whoever used it. Essentially, after the meal was over and the guest was getting prepared for the night, the reality of this magical bed was unveiled, as Procrustes himself would take care to ensure this perfect match by stretching his shorter guests to fill the length of the bed, or by properly chopping those bits that came out on taller ones. However, this methodology proved unsuccessful when Theseus appeared by Procrustes house and actually turned the tables, with fatal consequences for the bed's owner.

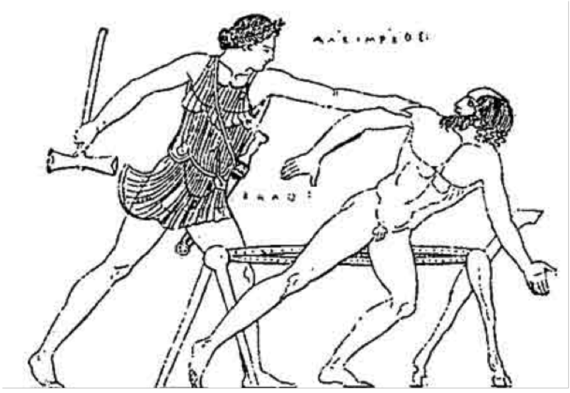


Figure 1.12: Procrustes and Theseus. Source: `http://www.mindwhirl.com/entrepreneurship/the-lure-of-the-procrustean-solution/`. Original drawing author unknown.

Procrustes analysis [Kendall, 1989] (fig. 1.13) combines two (or more) sets of points $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$ by estimating a location, scale and rotation transformation, so that the meaningful properties of the transformed set are conserved [Borg and Groenen, 2003]. The shift is estimated as the difference between the mean coordinates for $\mathbf{x}$ and $\mathbf{y}$, whereas the scale and rotation are determined from their covariance matrices. Procrustes is a general adjustment in the sense that it does not require careful consideration of what causes the systematic differences between sources, but it requires a minimal number of common points between $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_m)$. Specifically, estimating the shift requires at least one common point, while estimating the scale and rotation requires two, and in practice, it is recommended to have at least three or more points covering a moderately wide range of the point distribution to match to obtain a reliable solution. Further generalizations of this technique also allow linear distortion transformations, or different dimensionalities in the configurations [Borg and Groenen, 2003]. Procrustes solutions are important in practice to eliminate irrelevant (and often misleading) differences between two given MDS solutions, with practical applications in pattern recognition and image analysis problems [Duta et al., 1999]. Procrustes is also a widely used technique in analysis of high-dimensional data, with successful application in several biological scenarios [Wang et al., 2015].

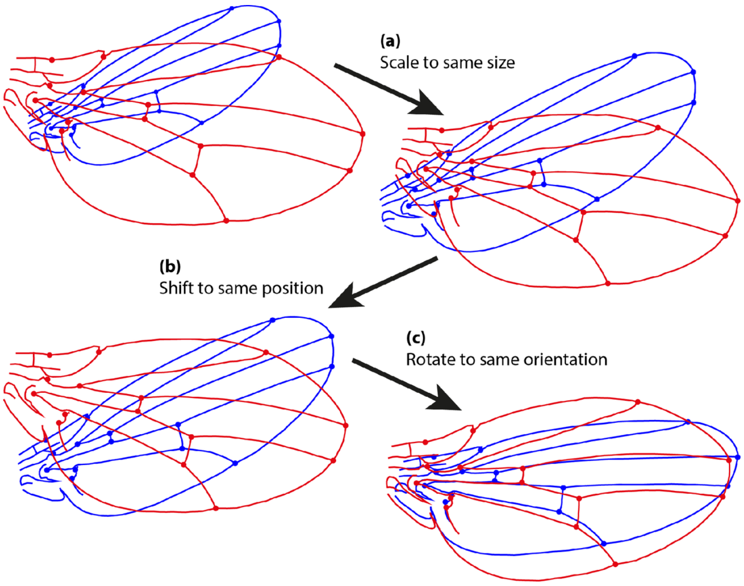
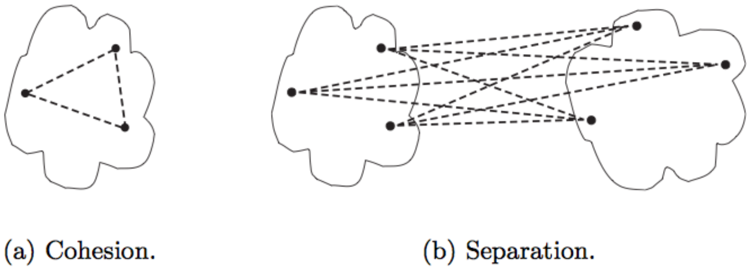


Figure 1.13: Procrustes superimposition. The figure shows the three transformation steps of an ordinary Procrustes fit for two configurations of landmarks. (a) Scaling of both configurations to the same size; (b) Transposition to the same position of the center of gravity; (c) Rotation to the orientation that provides the minimum sum of squared distances between corresponding landmarks. Source: [Klingenberg, 2015].

### 1.5.3 Cluster analysis

The aim of cluster analysis is to identify groups of objects that share similar characteristics. Elements within a group tend to present higher similarity between them than with those classified in other clusters and vice versa, that is, a desirable cluster classification presents a high level of cohesion and separation (fig. 1.14). It is important to note that clustering is not a specific algorithm per se, but more of a general methodology to perform data exploration and classification, and that the identified groups have to come together with an emphasis on usefulness and meaning, by identifying the natural structure underlying in the data. Later, this classification result can be used to perform further statistical analysis and to generate visual representations in order to be used for interpretation purposes. Cluster analysis has been used in a vast range of disciplines, from biology, ecology and genetics to psychology, marketing, and data mining. Human beings are particularly good at identifying groups of similar elements within a set of objects, and human vision and brain are capable of performing such a task even in small children in the most natural way. This makes cluster analysis a very intuitive and easy to interpret tool for performing exploratory analysis of new data [Tan et al., 2005].

A distinction can be made depending on wether cluster nesting is permitted, that is, if elements conforming major clusters can at the same time belong to smaller clusters within them. In contrast, a second group of clustering methods seek to perform a division of the data in order to

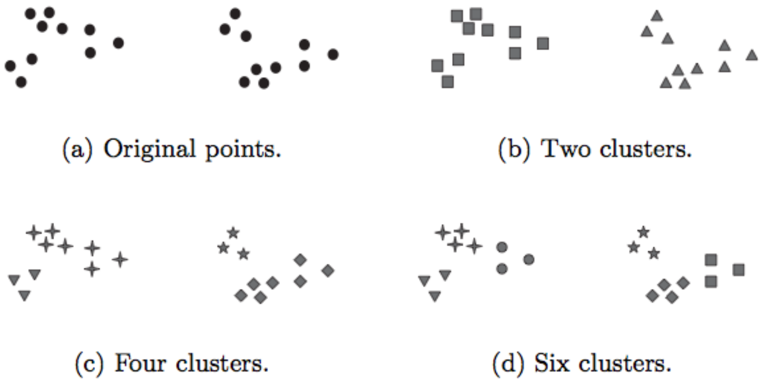


(a) Cohesion.

(b) Separation.

Figure 1.14: Cluster cohesion and separation. It is desirable for a given cluster solution to provide good levels of cohesion and separation, that is, that clusters are well-defined and well separated from the rest, something that can be estimated by measuring within and between-cluster distances. Source: [Tan et al., 2005]

classify each item in a unique non-overlapping cluster. In the first case, we talk of hierarchical clustering, whereas the second one is known as partitional clustering. Independently, we also differentiate between exclusive clustering methods, which assign each element to a unique cluster, and overlapping ones, in which a given element can belong to more than one cluster at the same time. Finally, in fuzzy clustering methods elements are classified in clusters with a certain probability (or uncertainty) value between 0 and 1 [Fraley and Raftery, 2006; Azzalini and Menardi, 2014]. It is also possible to use non-fuzzy hierarchical clustering methods and still obtain a probability 'score' for elements to be correctly classified within its assigned clusters, for instance by using Bayesian density estimation methods [Jara et al., 2011] with the obtained clustering classification results. As a final note, we can also differentiate between complete clustering methods, in which each and every element is classified in some cluster, and partial ones, that can return unclassified elements [Suthar et al., 2013].



(a) Original points.

(b) Two clusters.

(c) Four clusters.

(d) Six clusters.

Figure 1.15: Four possible clustering classifications for the same set of points. Source: [Tan et al., 2005].

In summary, cluster analysis is a discipline by itself and the vast amount of methodologies can cover an equal wide range of scenarios (fig. 1.15). However, it is not only important to ensure that the obtained classification is appropriate from a purely statistical point of view, for instance by ensuring that the identified clusters are robust (and therefore, reproducible) with bootstrap resampling or silhouette analysis [Rousseeuw, 1987], but also to verify that the provided classification is meaningful from the point of view of the original experimental question, interpretable, useful to obtain relevant conclusions and, as we mentioned in the case of Multidimensional Scaling, 'makes most sense'.

## 1.6 Visualizing the epigenome.

*Seeing comes before words. The child looks and recognizes before it can speak.*
*It is seeing which establishes our place in the surrounding world; we explain that world with words, but words can never undo the fact that we are surrounded by it.*

John Berger. Ways of seeing. 1972.

Despite being principles dating from more than 50 years ago, the four major influences acting on data analysis in 1965 are equally valid today: 1) The formal theories of statistics; 2) Accelerating developments in computers and display devices; 3) The challenge, in many fields, of more and larger bodies of data; 4) The emphasis on quantification in a wider variety of disciplines [Cooley and Tukey, 1965; Herr, 2014]. Since then, exponential advances in technology, computer science and telecommunications, as well as our own fast adaptation to them, have boosted these four principles, and in some way have added a fifth one: the instant access to unprecedented amounts of visual, interactive data. In his book *Semiologie Graphique* (Semiology of Graphics, [Bertin, 1967]), Jacques Bertin, a French cartographer and theorist, introduced and developed the theory of Visual Language as a Sign System where images are perceived as a set of signs in which a sender encodes information that is received and decoded by the Receiver (fig. 1.16). This monumental work, based on his experience as a cartographer and geographer, represents the first and widest intent to provide a theoretical foundation to Information Visualization, and also lead to the establishment of Visualization Design Principles, that can be summarized in two main statements:

- Tell the truth and nothing but the truth (don't lie, and don't lie by omission).
- Use encodings that people decode better (where better = faster and/or more accurate)

Development of efficient tools for data visualization as well as functional and differential analysis in the field of epigenetics is a field of continued research. Together with the generation of databases containing high-throughput data on whole genome distribution of DNA associated proteins and other factors in human and model organisms, several approaches have been made with a scope on visualization of this information and its integration with existing genomics and biomedical data. Not surprisingly, some of the most important advances in this field came from the main public consortiums generating this data such as ENCODE, modENCODE and Roadmap epigenomics ([Dunham et al., 2012; Celniker et al., 2009; Bernstein et al., 2010; Kharchenko et al., 2011; Riddle et al., 2011]), as a way to offer a channel of generating public repositories to store and share the generated results with the scientific community. Also, the unprecedented availability of these

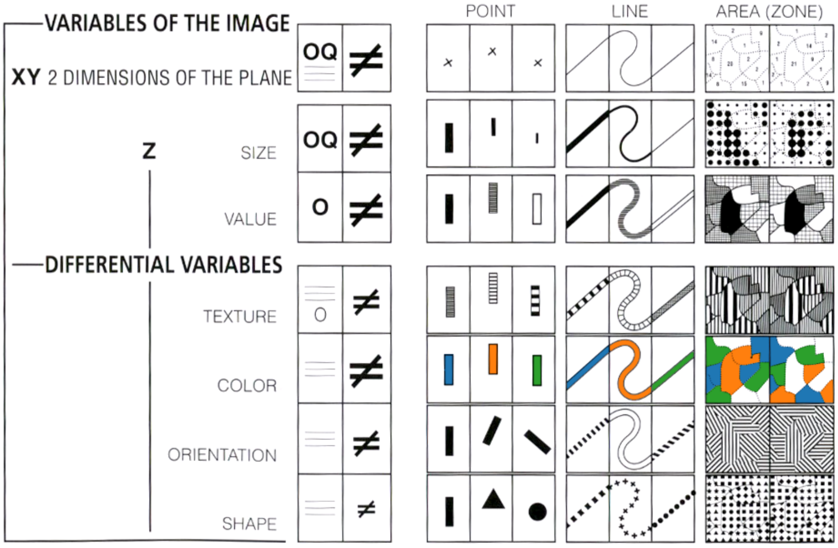


Figure 1.16: Within the plane a mark can be at the top or the bottom, to the right or the left. The eye perceives two independent dimensions along X and Y, which are distinguished orthogonally. A variation in light energy produces a third dimension in Z, which is independent of X and Y. The eye is sensitive, along the Z dimension, to 6 independent visual variables, which can be superimposed on the planar figures: the size of the marks, their value, texture, color, orientation, and shape. They can represent differences (), similarities (=), a quantified order (Q), or a nonquantified order (O), and can express groups, hierarchies, or vertical movements. Source: Jacques Bertin. Semiology of Graphics. 1967 [Bertin, 1967] and adapted by `https://medium.com/@mbostock/introducing-d3-scale-61980c51545f`

data and its quick adoption by the biological and biomedical fields as a way to explore and test functional hypothesis brought as a result some very interesting contributions to the question [Zhou et al., 2011, 2014].

The range of methodologies and techniques to face this challenge started with the classic 1 dimensional linear genome browsers [Yates et al., 2016; Zhou et al., 2014] (fig. 1.17), on which epigenetic information is often represented as a track or combination of tracks containing epigenetic information of the observed genomic coordinates. Other applications opted for more complex 2 dimensional representations of the genome using Hilbert-matrix based approaches to generate a two-dimensional picture of the genome based on a mathematical representation of euclidean space [Kharchenko et al., 2011; Riddle et al., 2011] (fig. 1.18), using color to represent genomic and epigenomic information. Finally, dimensionality reduction techniques have been used to generate two and three-dimensional representations based on whole-genome segmentation and computation of epigenetic states [Filion et al., 2010; van Bemmel et al., 2013]. In this regard, dimensionality reduction approaches together with clustering and machine-learning analysis and other statistical methods provide an interesting set of tools to approach the challenge of epigenomics data visualization and offers the opportunity to introduce additional visual information based on classic visualization principles [Bertin, 1967].

Functional analysis of the obtained epigenomic landscapes in several biological scenarios have been faced using Hidden Markov Models and Bayesian networks [Juan et al., 2016]. However, in many

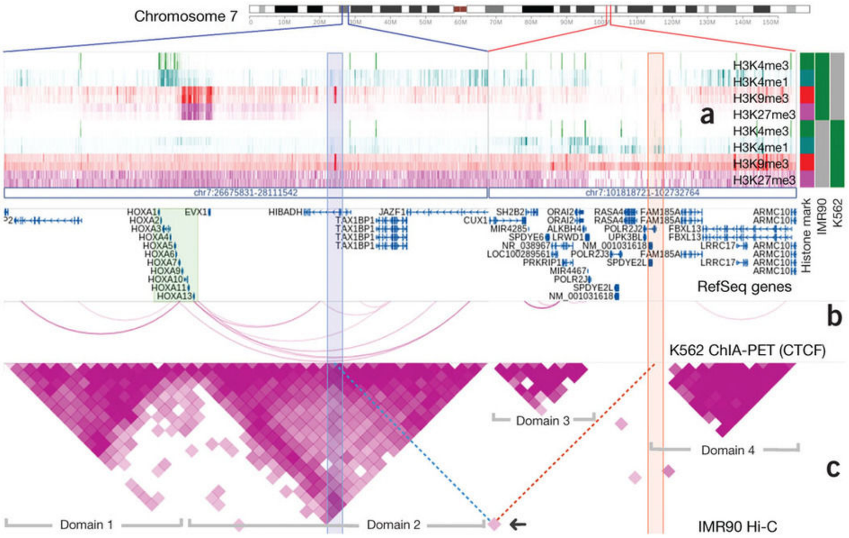


Figure 1.17: A capture of the Washu Epigenome browser. a) Heat-map view of histone modification profiles of IMR90 (top) and K562 cells (bottom; red and fuchsia for repressive histone marks, green and teal for active histone marks, identified by the color blocks in the metadata color map on the right). (b) ChIA-PET track from K562 cells obtained using DNA-binding factor CTCF (ENCODE data). (c) Hi-C track from IMR90 cells5. The triangle shapes in the Hi-C track depict chromatin domains in IMR90 cells (labeled as domains 1), and the arcs in the ChIA-PET tracks indicate similar domain structure in K562 cells. Two interacting loci in either genomic region are highlighted by the two semitransparent columns, and the Hi-C cell is indicated by an arrow. Source: [Zhou et al., 2014]

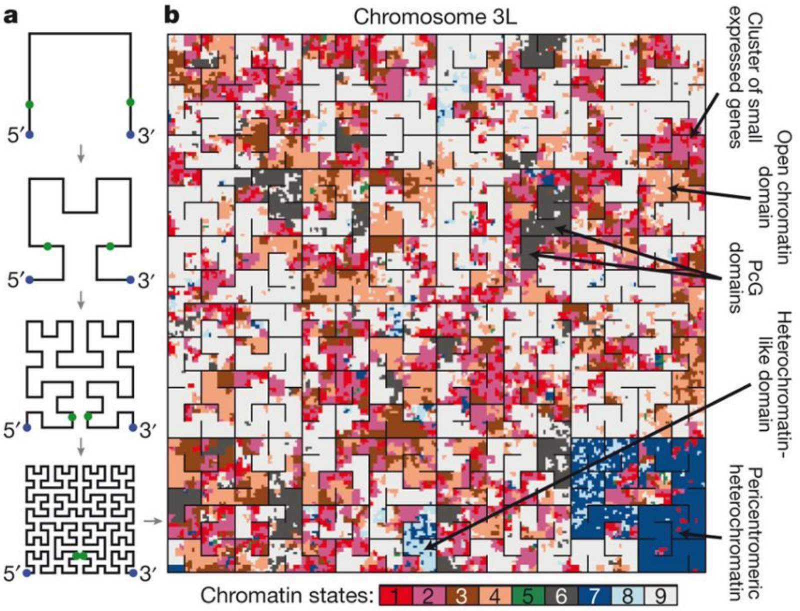

a
5′
3′
5′
3′
5′
3′
5′
3′
b
Chromosome 3L
Cluster of small expressed genes
Open chromatin domain
PcG domains
Heterochromatin-like domain
Pericentromeric heterochromatin
Chromatin states:
1
2
3
4
5
6
7
8
9

Figure 1.18: a, The chromosome is folded using a geometric pattern (Hilbert space-filling curve) that maintains spatial proximity of nearby regions. An illustration of the first four folding steps is shown. Source: [Kharchenko et al., 2011].

cases these methods were specially developed and tuned to cope precisely with the vast generated amounts of data. Thus, sometimes they lack the required flexibility to be adapted as a more general tool for the non-specialist researcher with a broader question or willing to use them for exploratory purposes: to investigate other biological backgrounds of interest, to obtain some functional relationship information about in-house epigenetics data, or to easily add custom annotation of gene expression or gene classification of interest to a given dataset, and their results can be difficult to interpret from a visual point of view.

Regarding the field of comparative epigenomics, several methods have been developed to approach the issue of differential methylation, and also for whole epigenome comparison [Zhang et al., 2013; Cao and Zhong, 2013; He and Wang, 2017], which usually rely on the simultaneous visualization and pseudo-automatic analysis of genome-browser 1-dimensional epigenetic tracks. As in the field of data visualization, the main consortiums behind the generation of big epigenomics data repositories have contributed with very interesting approaches to perform differential analysis of epigenomics data across multiple cell types and datasets [Yen and Kellis, 2015], and others have also approached this question analyzing precomputed information of chromatin states data [Li et al., 2016]. Also, in order to overcome the drawback of the high number of epigenetic marks needed to generate whole-genome epigenetic datasets, methods such as ChromImpute [Ernst and Kellis, 2015] take advantage of the combinatiorial nature of epigenetic marks and how experimentally obtained information on certain groups of marks can be successfully used to impute others.

Finally, without doubt the development of bioinformatics and biostatistics as necessary tools for modern Data Science and the appearance of Big Data brought us the Golden Age of data visualization [Friendly, 2008], when finally many of the very advanced concepts developed in the first years of computational statistics found a field in which they could fulfil their potential almost with the only limit of our imagination [Cooley and Tukey, 1965; Bertin, 1967]. We also have to admit the crucial contribution of the internet to this, as we owe to the World Wide Web for the instant imaging era in which we currently are.

# Chapter 2

# Objectives

The five main objectives of this thesis are:

- To develop and implement a computational and statistical approach for the generation and visualization of epigenetic factor maps.
- To develop and implement a computational and statistical approach for the generation, visualization and functional analysis of whole-genome epigenetic maps.
- To develop and and implement a computational approach for differential analysis of factors and gene maps.
- To explore the possibility to generate and analyze epigenetic maps with combined information from two different species.
- To develop and implement a computational strategy to select the minimum set of factors that are required to generated biologically informative factor and gene epigenetic maps.

# Chapter 3

# Results

Font-Burgada J, Reina O, Rossell D, Azorín F. chroGPS, a global chromatin positioning system for the functional analysis and visualization of the epigenome. Nucleic Acids Res. 2014 Feb;42(4):2126–37. DOI: 10.1093/nar/gkt1186

Kessler R, Tisserand J, Font-Burgada J, Reina O, Coch L, Attolini CS-O, et al. dDsk2 regulates H2Bub1 and RNA polymerase II pausing at dHP1c complex target genes. Nat Commun. 2015 Apr 28;6(1):7049. DOI: 10.1038/ncomms8049

Reina, O., Azorin, F., and Stephan-Otto, C. (2017). **chroGPS2, differential analysis of epigenome maps in R**. Unpublished manuscript.

# chroGPS2: differential analysis of epigenome maps in R.

Oscar Reina, Fernando Azorin and Camille Stephan-Otto Attolini.

November 2, 2017

## 1 Abstract

We present chroGPS2, a computational framework for differential analysis of epigenomes. Methods are provided for efficient integration and comparison of data from different conditions or biological backgrounds, accounting and adjusting for systematic biases in order to provide an efficient and statistically robust base for differential analysis. We also include functionalities for general data assessment and quality control prior to comparing maps, such as functions to study chromatin domain conservation between epigenomic backgrounds, to detect gross technical outliers and also to help in the selection of candidate marks for de-novo epigenome mapping.

- Availability: `https://github.com/singlecoated/chroGPS2`
- Contact: oscar.reina@irbbarcelona.org

## 2 Introduction

In recent years, we have assisted to an unprecedented increase in availability of epigenomics data related to whole-genome distribution of transcription factors, histone modifications and other DNA binding proteins [Celniker et al., 2009; Dunham et al., 2012; Bernstein et al., 2010], which helped to establish a deeper knowledge about chromatin states and topologically associated domains [Filion et al., 2010; Serra et al., 2016], and to offer a broad scope of genomic regulation from both a functional and structural point of view. However, there remains a need for efficient tools for visualization, functional analysis and comparison of epigenomics data [Marx, 2015]. Previously, we developed chroGPS [Font-Burgada et al., 2014], an R package based on dimensionality reduction techniques (namely Multidimensional Scaling, MDS [Borg and Groenen, 2003]) to measure and visualize genome-wide associations between epigenetic factors and their relationship to functional genetic elements in low dimensional maps. Now we extend this software and introduce novel features to perform differential analysis of epigenome maps using Procrustes [Lisboa et al., 2014], hierarchical clustering and Bayesian density estimation methods [Jara et al., 2011]. Additionally, we provide functions for general data assessment, quality control and to help in the selection of epigenetic factors for de-novo epigenome mapping. ChroGPS2 is integrated in Bioconductor [Huber et al., 2015], an open-source collection of R packages for computational analysis of omics data. We illustrate our approach using two publicly available datasets containing extensively mapped human and fruit fly epigenomes (https://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html). Supplementary material includes detailed data accession information, workflows and additional examples in several biological scenarios.

## 3 Data assessment and Quality Control

We offer functions to perform a first general assessment of functional relationships between epigenetic factors and conformation of chromatin domains across datasets, and to highlight major differences between them. After measuring pairwise similarities between epigenetic factors based on genome-wide overlap of their binding profiles with the *distGPS* function, the *domainDist* one evaluates and compares these results at factor and domain level using Pearson correlation and Mantel-Hazel tests, pointing towards major changes that may suggest strong biological effects but also gross technical problems or annotation errors (i.e. sample mislabeling, truncated or empty data, inefficient immunoprecipitation, see Supplementary section 3.1).

## 4 Select candidates for de-novo epigenome experiments

Selection of candidate factors when designing high-throughput ChIP-Seq, Dam-ID or Hi-C experiments in order to compare a known scenario with an novel one poses important experimental and economical challenges. The *rankFactors* function effectively ranks a set of factors based on how much they contribute towards conservation of their respective chromatin domain identity. Additionally, if domain identity is unknown or incomplete we rely on relative conservation of functional relationships between individual epigenomic marks and genetic elements (i.e. genes, promoters, etc) and how information from experimentally mapped factors can be used to successfully impute presence/absence of others [Ernst and Kellis, 2015]. In detail, we offer methods based on linear and logistic regression to rank marks based on their ability to predict others, helping the researcher to refine selection of potential candidates for experimental mapping (See Supplementary section 3.2).

## 5 Comparing epigenomic factor maps.

chroGPS factor maps use MDS to represent genome-wide epigenetic factor co-localization based on similarities between their binding profile overlaps according to a metric of choice. This core methodology is also used as a starting point to identify differences between epigenomic factor associations at different conditions, time points, cell lines, or between different species.

- **Assess differences between conditions:** Comparing two epigenomic factor maps from the same species with a rich number of common factors is relatively straightforward, as they share a genomic background over which pairwise similarities *s(i,j)* between their genome-wide binding profile overlaps can be computed using the *distGPS* function, and represented together in a low dimensional space using MDS. This already produces a joint map which accurately represents putative functional relationships between all elements from both datasets [Font-Burgada et al., 2014]. Further adjustment and differential analysis is performed with the function *diffGPS.factors*, which uses

Procrustes to identify potential biological differences between both backgrounds after adjusting for possible technical and biological biases. Procrustes matches two sets of points represented in a low-dimensional space by using the information from common elements (landmarks) to compute a transformation involving scaling, translation and rotation to minimize euclidean distances between landmark points while preserving relative spatial configuration within each set. General goodness of fit and magnitude of the observed changes between common factors is measured via Procrustes sum of squares, and statistical assessment can be performed via permutation tests [Gel et al., 2016]. This approach has been successfully applied to integrate and compare the epigenomic landscape obtained at different time points during genome replication in *Drosophila* S2 cells (See Figure 1 and supplementary section 4.3).

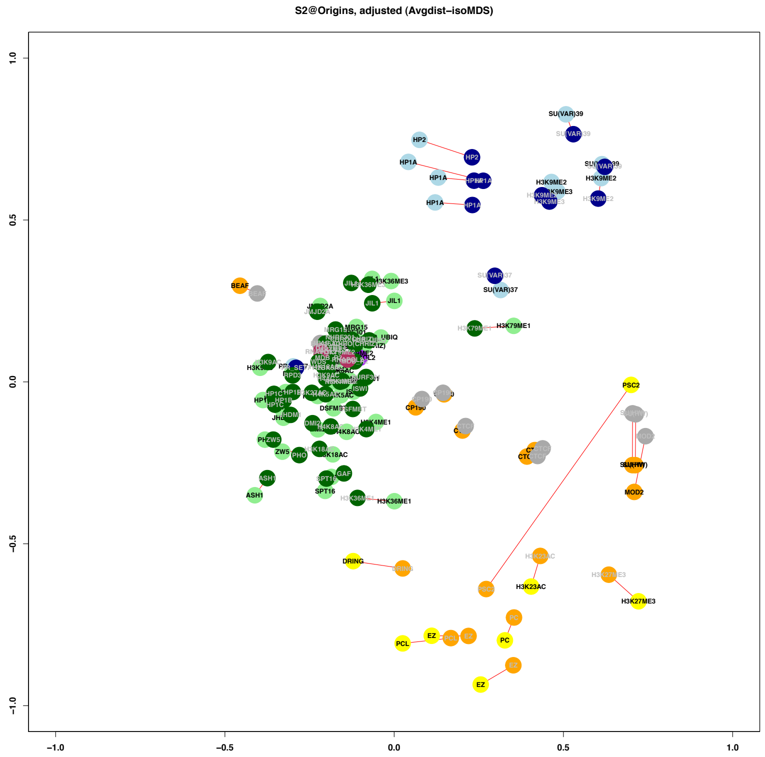


Figure 1: Differential chroGPS-factors map showing the epigenomic landscape differences at *Drosophila melanogaster* Early (light) and Late (dark) time points during genome replication.

- **Investigate differences between species:** Another interesting hypothesis is to explore potential differences between homolog epigenomic factors in two distant species over which common whole-genome co-localization cannot be assessed, but where conservation of their respective relative distances between common factors can still be compared. For this purpose, after generating individual maps for each background *G* and *G'*, position of common elements *i, i'* is used to match, adjust and compare them using Procrustes. We successfully used this approach to study potential functional differences between homolog boundary/insulator elements in *Drosophila* S2 and human

K562 cell lines, and to assess their relationship with spatial configuration of transcriptional activity surrounding CTCF binding sites (See supplementary section 4.4.2).

## 6 Comparing epigenomic gene maps

ChroGPS-genes maps represent similarities between genomic features by computing pairwise distances between the binary vectors defined by their epigenetic footprint, that is, their collection of nearby epigenetic marks, and projecting them in a low dimensional space via MDS [Font-Burgada et al., 2014]. Downstream functional analysis can be performed by using hierarchical clustering with the *clusGPS* function or using the original and approximated distances with other machine learning or class discovery methods of choice.

- **Detecting epigenomic footprint differences:** The *diffGPS.genes* function uses essentially this strategy to compare two sets of mapped factors from the same or different species using epigenomic footprint information of both datasets simultaneously. First, common factors between both sets are selected. Then, pairwise similarities between their whole-genome epigenetic footprints are computed to generate a map comprising all common genes from both datasets, relying on inter-species gene homology information if necessary. Identification of genes presenting changes in their epigenomic footprints is then straightforward, as they will be indicated by those genes being represented by different points for each set.

- **Interpreting loss of epigenomic footprint identity:** A first approach to study those differences is to rank them using Procrustes to compute the residual sum of squares between points representing differentially located genes, but this offers little insight on the observed changes from a functional perspective. We use unsupervised hierarchical clustering followed by computation of Bayesian posterior probability of cluster classification using a Dirichlet Process mixture of normals [Jara et al., 2011] to identify and characterize clusters of genes with similar epigenetic patterns, and to report genes found to be strongly classified under two distinct clusters between both conditions. Akin with results obtained from other high-throughput technologies, candidate gene lists can be used for further downstream functional analysis, such as Gene Ontology enrichment or Gene Set Enrichment Analysis, as well as compared with other genomics data such as changes in gene expression or regulation to further explore interesting hypothesis. We illustrate the use of this strategy to identify and characterize genes potentially involved in epigenomic changes between third instar larvae *Drosophila* S2 cells and BG3 neuronal tissue. (See Figure 2 and Supplementary section 5.2).

## 7 Visualization

We facilitate the visualization of results by offering functions to export graphical outputs into open formats for using them with R network/graph visualization packages, or with external softwares such as Cytoscape with the function *gps2xgmml*. Additionally, supplementary code includes examples on to export results as interactive HTML5 outputs using the shinyRGL and plotLy packages.

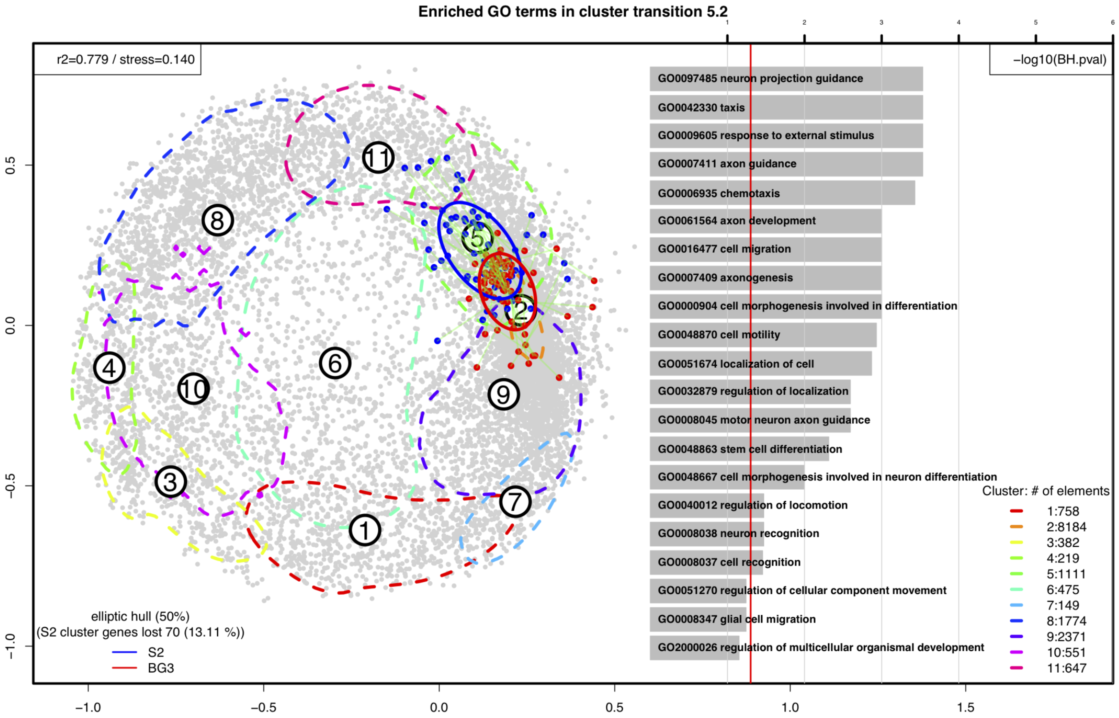


Figure 2: Differential chroGPS-genes map of *Drosophila melanogaster* S2 and BG3 cell lines. Focus is put on genes changing from cluster 5 in S2 (blue, moderate HP1 repression) to cluster 2 in BG3 (red, transcriptionally active). Dashed lines indicate probability contours containing 50 percent of genes for each cluster. On the right, the top 20 enriched Gene Ontology terms for genes involved in the analyzed cluster transition are shown. Selected genes are specially enriched in nervous system and cell differentiation categories.

# References

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol., 28(10):1045–1048.

Borg, I. and Groenen, P. (2003). Modern multidimensional scaling: Theory and applications. Journal of Educational Measurement, 40(3):277–280.

Celniker, S. E., Dillon, L. A., Gerstein, M. B., Gunsalus, K. C., et al. (2009). Unlocking the secrets of the genome. Nature, 459(7249):927–930.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414):57–74.

Ernst, J. and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol., 33(4):364–376.

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., et al. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell, 143(2):212–224.

Font-Burgada, J., Reina, O., Rossell, D., and Azorin, F. (2014). chroGPS, a global chromatin positioning system for the functional analysis and visualization of the epigenome. Nucleic Acids Res., 42(4):2126–2137.

Gel, B., Diez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., and Malinverni, R. (2016). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics, 32(2):289–291.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods, 12(2):115–121.

Jara, A., Hanson, T. E., Quintana, F. A., Muller, P., and Rosner, G. L. (2011). DPpackage: Bayesian Non- and Semi-parametric Modelling in R. J Stat Softw, 40(5):1–30.

Lisboa, F. J. G., Peres-Neto, P. R., Chaer, G. M., Jesus, E. d. C., et al. (2014). Much beyond mantel: Bringing procrustes association metric to the plant and soil ecologist's toolbox. PLOS ONE, 9(6):1–9.

Marx, V. (2015). Visualizing epigenomic data. Nat. Methods, 12(6):499–502.

Serra, F., Baù, D., Filion, G., and Marti-Renom, M. A. (2016). Structural features of the fly chromatin colors revealed by automatic three-dimensional modeling. bioRxiv.

# chroGPS: differential analysis of epigenome maps in R.

Supplementary material

Oscar Reina, Fernando Azorin and Camille Stephan-Otto Attolini.

November 2, 2017

## 1 DATA ACQUISITION AND FORMATTING

The source material for generating and comparing epigenomic maps with chroGPS is a list of genome-wide predicted binding sites for each condition, formatted as a *GRangesList* object, that is, a list of Genomic Ranges. These can simply be collections of plain-text BED or GFF files downloaded from ENCODE, modENCODE, Roadmap Epigenomics [Bernstein et al., 2010; Celniker et al., 2009; Dunham et al., 2012], or any other public data repository, or in-house generated data coming from any of the widely used Peak Calling algorithms such as MACS, Sicer, etc. Such files can be easily imported into R using custom code or readily available functions from many other Bioconductor packages [Gentleman et al., 2004]. Alternatively, some specific packages already provide routines to efficiently query, filter and download such kind of data directly from the official repositories in GRanges format. See below for a comprehensive list of online databases and tools taken from the `Epigenie` website.). Additionally, supplementary code offers some guidelines on importing genomic-interval based data from several sources.

Supplementary R code and data available at:
`https://github.com/singlecoated/chroGPS2/tree/master/examples/getData'`

- IHEC Data Portal: The International Human Epigenome Consortium (IHEC) brings forth reference epigenomes relevant to health and disease. View, search, and download all the data.
- ROADMAP Epigenomics: The NIH Roadmap Epigenomics Mapping Consortium offers maps of histone modifications, chromatin accessibility, DNA methylation, and mRNA expression across 100s of human cell types and tissues.
- CEEHRC Platform: A reference epigenome project for human cells and not the typical stem cell lines.
- DeepBlue: Store and work with genomic and epigenomic data from a number of international consortiums.
- Epigenome Browser: For the UCSC genome browser fans.
- WashU Epigenome Browser: A web browser that fffers tracks from ENCODE and Roadmap Epigenomics projects.
- Ensembl: Featuring ENCODE.
- GenExp: A web-based visualization tool to interactively explore a genomic database.

- The Epigenome Atlas: Human reference epigenomes.
- Classification of Human Transcription Factors: The mother list of transcription factors and their binding sites.

# 2 DATA EXPLORATION AND QUALITY CONTROL

Before performing differential analysis some useful information can be obtained from original data to determine goodness of the selected datasets and help interpreting the observed differences afterwards. In detail, we offer functions to study the degree of conservation between chromatin domains defined by groups of factors under two given conditions, and to assess presence of gross technical problems or annotation errors present in the data that could affect subsequent integration and comparison. Additionally, we implement methods to help researchers with selection of candidate marks when designing de-novo epigenome mapping experiments in order to compare a novel background or condition with already existing ones, a scenario posing important technical and economical challenges.

## 2.1 Exploring factors and domains

The core methodology of chroGPS, that is, computation of pairwise similarities (and thus, dis-similarities that can be interpreted as distances) between epigenomic factors based on their binding profile overlaps, already provides some useful insight on relative configuration of epigenomic factor domains present in the data. In detail, this information can be compared across multiple datasets from which a rich number of common mapped factors is available, to assess correlation in vectors of similarities between pairs of analog factors or domains, in order to identify potentially strong biological or technical differences between them. We make use of this functionality using the *domainDist* function to assess the already observed general domain conservation between *Drosophila melanogaster* S2 and BG3 cell lines [Celniker et al., 2009; Font-Burgada et al., 2014], and to see what happens when we artificially introduce a 'wrong' instance of factor EZ in the S2 dataset (See Supplementary Figure 1).

We will start by loading an object with downloaded Binding Sites for several *Drosophila melanogaster* S2 and BG3 epigenomic factors annotated for closest and overlapping dm3 genes using the *AnnotatePeakInBatch* function from the *ChIPpeakAnno* package [Zhu et al., 2010], and will unify experimental replicates by simply joining all reported binding sites for each factor. Additional methods of replicate management are provided in the function *mergeReplicates.*

Afterwards, we will use the color information provided in the *s2names* data frame object, which will effectively serve as aliases for our chromatin domains of interest. And then, we will take care of computing pairwise similarities / distances between S2 and BG3 epigenomic factors based on their whole genome binding profile overlaps [Font-Burgada et al., 2014], and will use the *domainDist* function to calculate both within and between-domain distances, which can be described as the collection of computed mathematical distances measured between all pairs of factors belonging to a certain chromatin domain (intra), or between all possible pairs of factors from two different domains (inter). This information gives a valuable information regarding cohesion and separation of the different domains observed in the map, and can be used to start assessing putative conservation between factors on different biological backgrounds.

The returned intra and inter-domain distance objects can be used for generating custom plots and performing further downstream analysis, and the observed differences between datasets can be assessed statistically via Mantel-Hazel tests or permutation tests [Gel et al., 2016]. We can also see the effects of introducing an artificial outlier EZ sample in the S2 cell line, by randomly selecting binding sites from other factors.

Supplementary R code and data available at:
`https://github.com/singlecoated/chroGPS2/tree/master/examples/domainDist'`

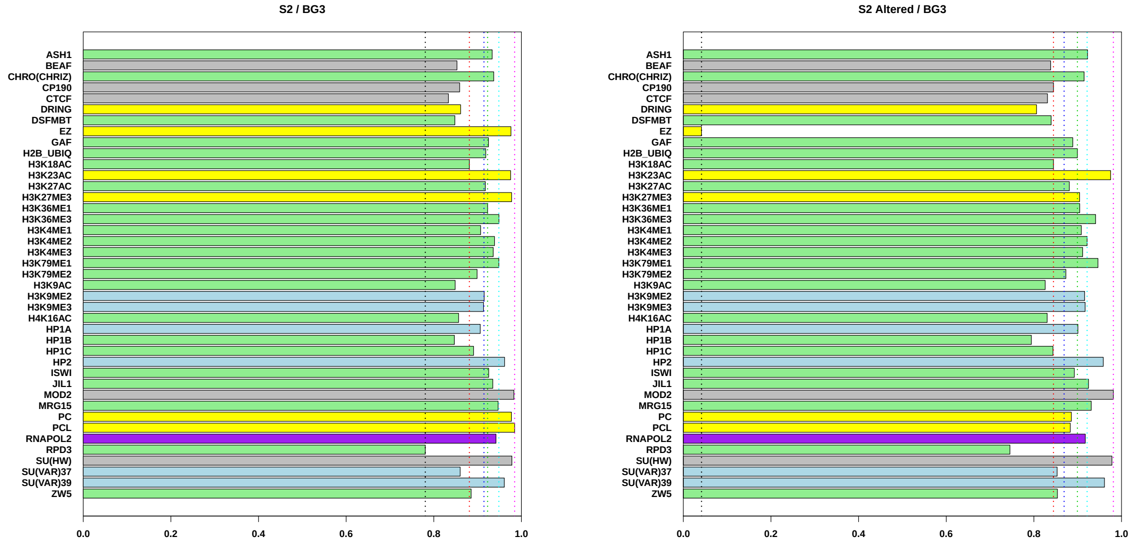


Figure 1: Detection of potential technical artifacts by checking conservation of relative distances between common factors. Left: Conservation for *Drosophila* S2 and BG3 epigenomic datasets. Right: The same plot where artificially modified EZ factor is introduced in the S2 dataset.

## 2.2 Selecting factors for de-novo epigenome mapping

**Maximize chromatin domain identity:** Chromatin domains offer an insightful and intuitive way to interpret epigenomic map conformation, by providing a biological context to factors based on functional relationships between them. When such information is available, a straightforward approach to select candidate factors for performing a de-novo epigenome mapping is to select those ones giving maximum robustness to their corresponding domain. This is easily achieved using the *rankFactors* function. We can use our chromatin color values as an alias to define chromatin domains. In this example we see how to perform a domain distance based selection for the HP1a repression considering a subset of 4 different factors (See 2).

**Ranking factors based on functional relationship with genetic elements**. Alternatively, or whenever domain information is not available, selection of candidate factors can be based upon conservation of functional relationships between epigenetic factors and genetic elements and how information on experimentally mapped factors has been successfully used to impute unknown ones [Ernst and Kellis, 2015]. The data for performing this analysis is that one used to generate chroGPS-genes maps (See Supplementary section 4), that is, a binary matrix of $N$ genes (rows) per $M$ columns (factors), where each cell equals 1 if that gene has an assigned binding site for that factor, and 0 otherwise. The *rankFactors* function provides methods based on linear and logistic regression to rank factors based on how accurately they can be predicted by others. At each iteration, the factor which can be best predicted by the rest is removed and prediction accuracies are recomputed (see 2). On the downside, these methods can be computationally intensive and we recommend using parallel computation or reducing the number of iterations if computing power is limited.

Supplementary R code and data available at:
`https://github.com/singlecoated/chroGPS2/tree/master/examples/rankFactors'`

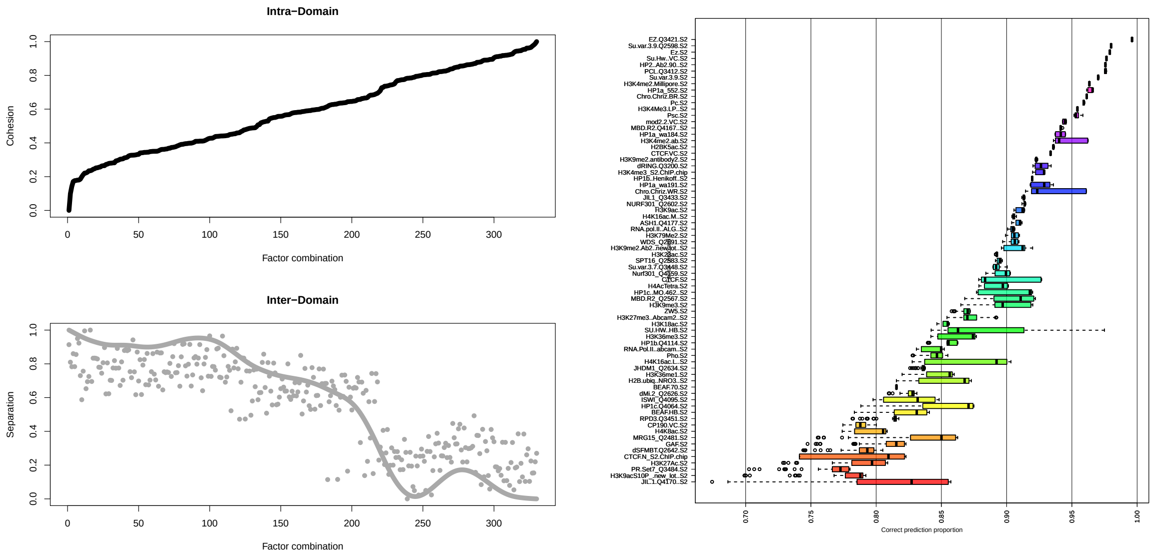


Figure 2: Left: From left to right, resulting domain integrity (Cohesion/Separation) based on computation of Intra and Inter-Domain distances for each combination of 4 factors within the HP1a repression domain. Results are sorted by Intra-Domain distance, which is an informative concept regarding cohesion of the observed domains in the map. Potential candidates are those among the leftmost factor combinations minimizing Intra-Domain distance but still providing meaningful biological content. Right: Factor ranking based on logistic regression of functional relationship with genetic elements when no domain information is available. Left axis, from top to bottom, indicates factor removed at each iteration, which is the one having the higher correct prediction rate based on information from those ones remaining below, until the number of remaining factors is lower than a given threshold. Boxplots in each line indicate the obtained prediction rates for all removed factors on top, based on the information from those on the bottom. Notice that as we remove more factors, general prediction rates get lower.

# 3 COMPARING CHROGPS-FACTOR MAPS

In chroGPS-factors maps epigenetic factors are represented over a low-dimensional space based on a similarity measure based on their observed co-occurences at whole genome level. However, this is only a global picture of epigenetic factor colocalization and thus does not reflect the possible mechanisms going on in certain kind of genomic regions (i.e. coding regions, promoters, transcription start sites, enhancers etc). Epigenomic factor colocalization is indeed a complex scenario, playing a critical role in gene regulation and silencing, transcriptional replication, genome structure and repair, etc. A very straight-forward exercise is to observe distribution of these epigenetic marks based on factor overlaps happening only over certain specified regions of interest, to be used at subsequent comparisons. Even though the methodology to generate chroGPS-factors maps is available in our previous work, first we will provide a brief reminder on how these maps are generated.

## 3.1 Generating factor maps

The main ingredient for generating chroGPS-factors maps is a collection of genomic intervals from our epigenomics data (putative binding sites, enriched regions, etc) in *GenomicRanges* format, in the shape of a *GRangesList* object, belonging to the experimental condition we want to study (i.e. cell line, patient, etc). Such information can be stored in independent plain tab-separated, BED or GFF files. Once our genomic intervals collection is loaded, we make use of the distGPS function to compute pairwise similarities / distances between them, as a way to assess whole genomic co-localization in our epigenomics background. At chroGPS we provide different metrics designed to work with genomic interval data to account for several scenarios where certain mathematical characteristics may be desirable, as well as the option to feed the workflow with user-defined metrics [Font-Burgada et al., 2014]. The resultant object contains a $n\ x\ n$ matrix of dissimilarities ranging between 0 and 1 that can be already used for visualization and analytical purposes (i.e. heatmaps, clustering).

The next step is to use the distance object as input to the core of our methodology. Multidimensional Scaling (MDS) technique used used to dissimilarity data in easy to visualize, easy to interpret, 2 or 3 dimensional graphical representations that account for the original relationships of similarity between the observed objects [Borg and Groenen, 2003]. We account for several MDS methods, and provide functions for parallel computation and goodness-of-fit optimization when several thousands of elements are present in our dataset.

The returned object can be represented in a 2 or 3 dimensional space where each element is located based on their similarity (i.e. co-occurrence) with all the other elements at whole genome level. This strategy already proved successful at reflecting the biological nature of epigenomic domains in *Drosophila melanogaster* S2 cells as illustrated in the left panel of Figure 3. MDS objects returned can be plotted directly using the package provided method. Additionally, the function *getPoints* can be used to retrieve values for each element present in the map and used for producing custom or enhanced graphical outputs or exporting interactive HTML5 graphics with RStudio if desired (See Supplementary section 5).

Supplementary R code and data available at:
`https://github.com/singlecoated/chroGPS2/tree/master/examples/chroGPS-factors'`

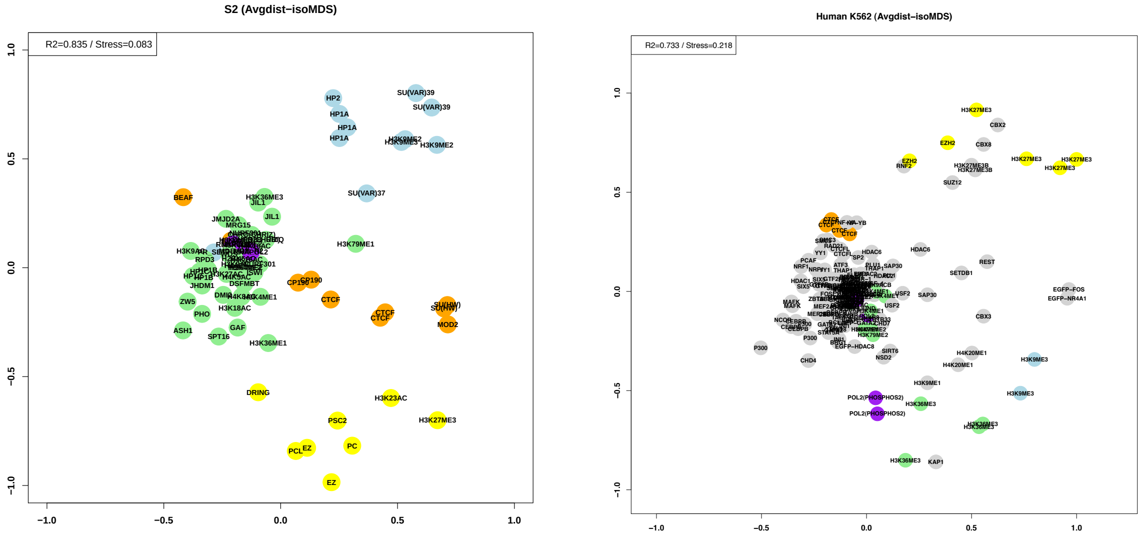


Figure 3: chroGPS-factors map of *Drosophila melanogaster* S2 cells (left) and human K562 leukemia cell line (right).

## 3.2 Generating region-specific factor maps (Promoters)

The basis for this case study is the same epigenetic factor collection used to generate the *Drosophila melanogaster* map above, plus additional information providing the genomic regions in which we would like to focus. In our case we will focus our study over gene promoters by obtaining UCSC dm3 genes and generating a new *GenomicRanges* collection with promoter regions (defined as 1kb upstream of the TSS).When these two objects are provided as input for the *distGPS* function, similarities between epigenetic factors are only taken into account when their overlap falls over the specified regions of interest. Thus, co-occurrences taking place at coding or intergenic regions will be ignored. If more control over region-based filtering is needed, interval-based operations from the IRanges and *GenomicRanges* packages offer advanced options to address it.

Straightforward visualization of the map generated in this way (4) already provides some insights on potential regulatory changes happening at promoter regions. Direct visual comparison with the whole genome *Drosophila melanogaster* S2 map [Font-Burgada et al., 2014] already highlights interesting changes in the conformation of some epigenomic domains, such as location of the Heterochromatin Protein complexes. However, it is desirable to perform further analitical work to offer the user quick and unbiased identification of significant changes when comparing region-based maps between two collection of regions or against a whole-genome scenario.

Supplementary R code and data available at:
`https://github.com/singlecoated/chroGPS2/tree/master/examples/diffPromoters'`

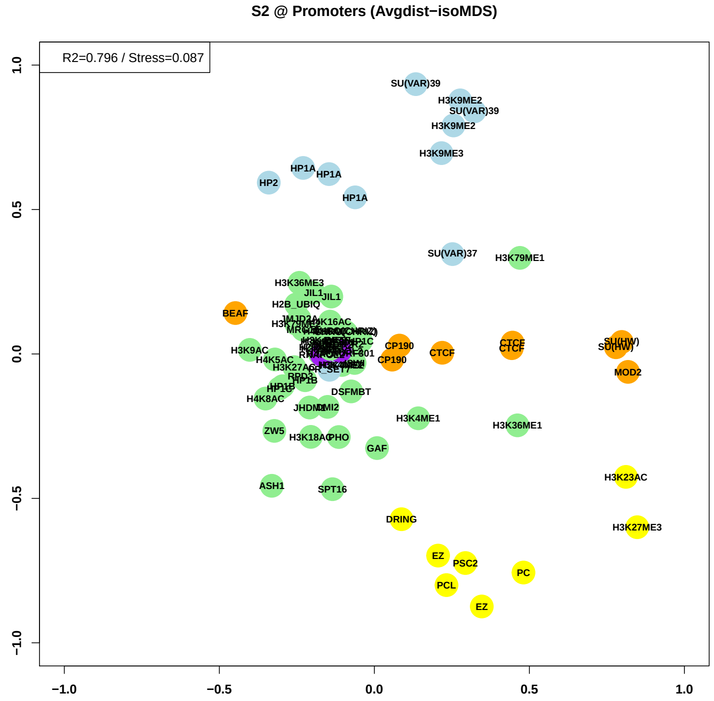


Figure 4: chroGPS-factors map of factor co-localization over *Drosophila melanogaster* S2 promoters.

### 3.3 Comparing factor maps between conditions (Origins of Replication)

Procrustes allows integration of epigenetic maps coming from different biological backgrounds, as well as adjustment of undesired biases due to technical effects [Borg and Groenen, 2003; Kendall, 1989]. Furthermore, it can be used to identify differences between epigenomic factor maps generated at different conditions (healthy/disease, control/treated), coming from distinct biological backgrounds, or being snapshots of different points in time, such as developmental stages or timings in origins of replication. We will focus on this elements to illustrate comparison of chroGPS-factors maps.
Origins of replication are particular genomic sequences at which replication of DNA starts in living eucharyote and prokaryote organisms, and of DNA-RNA in viruses. These sequences are recognized by specific proteins, that recognize, unwind and begin to copy the genomic sequence. In the following steps we illustrate how to use chroGPS to compare the epigenomic landscape at these different time points.

First, we start by loading Origins of Replication data from *Drosophila melanogaster* S2 cells available in modENCODE [Celniker et al., 2009], and contains location for Origins of Replication at Early, Early Mid, Late Mid and Late cell replication time points, stored in four different BED files. The other data to perform our study will be our already familiar collection of epigenomic elements (genomic intervals) in all conditions we would like to compare. In our case we will use *Drosophila melanogaster* S2 modENCODE binding sites and will filter them according to the different collections of Origins of Replication regions we just loaded. This can be easily performed using basic IRanges overlap operations.

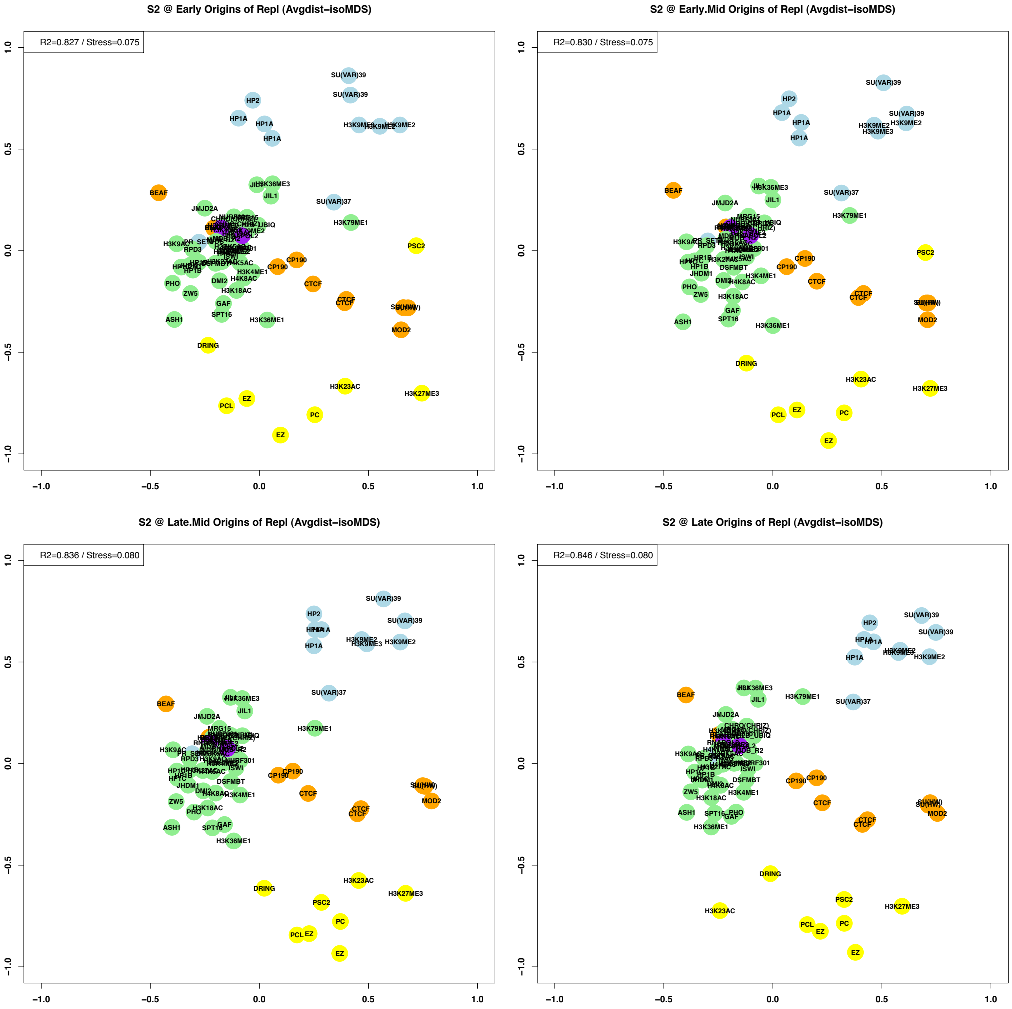


Figure 5: chroGPS-factors map showing the epigenomic landscapes *Drosophila melanogaster* S2 cell lines at Early (top left), Early-Mid (top right), Mid-Late (bottom left) and Late (bottom right). Upon visual inspection we can already observe changes in location and distribution of some epigenetic factors and domains.

The next step is to generate regular chroGPS-factors maps of each joint dataset as described in section 3.1. Map comparison is performed using Procrustes to measure and rank changes between two given maps in an unbiased manner, which is done automatically with the *diffGPS.factors* function. This function performs the dual task of finding an optimal adjustment in shift, scale and rotation so that both maps are matched as much as possible while maintaining relative distances between the respective factors of each one and providing the user with the internal sum of error metrics of the Procrustes algorithm. As a result, we obtain both a graphical 2D chroGPS-factors map where distances between replicates of the same element in both backgrounds are highlighted, and a ranked list of Procrustes errors for all common factors involved in the map. Statistical significance of the observed changes can be assessed via Mantel-Hazel and overlap permutation tests. Figure 6 illustrates results for Procrustes differential analysis of the transition between Early-Mid and Late time points.

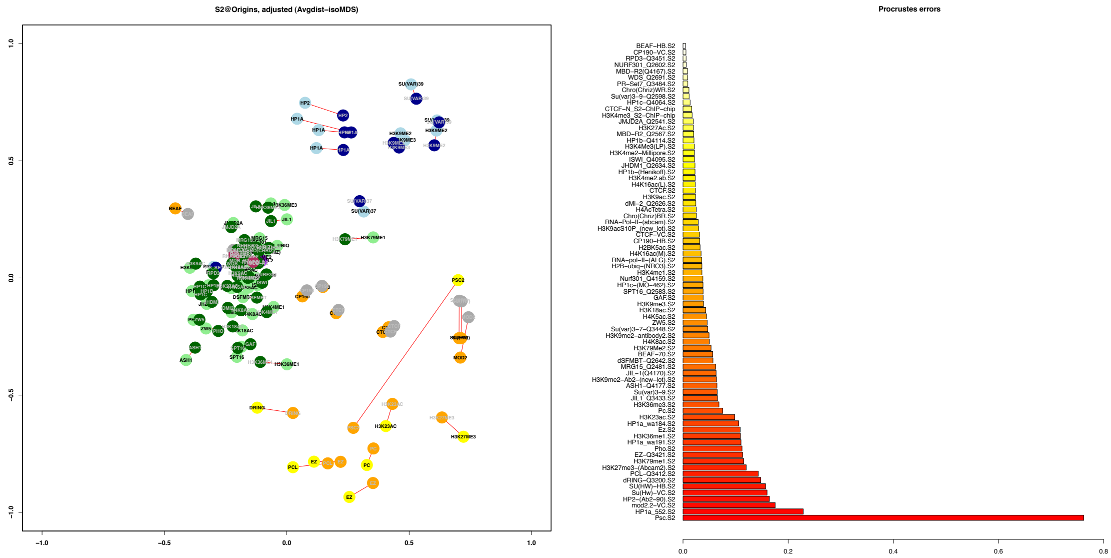


Figure 6: Differential chroGPS-factors map of *Drosophila melanogaster* S2 cells at Early vs Late Origins of Replication (left), and Procrustes normalized errors (right). In this differential map, PSC2, a Polycomb related protein involved in non-heterochromatic gene silencing, is strongly shifted from the boundary-insulator region (Early) towards the Polycomb repression one (Late).

Supplementary R code and data available at:
`https://github.com/singlecoated/chroGPS2/tree/master/examples/diffOrigins'`

## 3.4 Comparing factor maps between species

### 3.4.1 Introduction

In recent years, the ENCODE Project [Dunham et al., 2012] has widely expanded the availability of epigenomics data systematically for a huge collection of human cell lines and tissues, at different level of completion (Tiers). In this regard, K562, H1-hESC and GM12878 represent the three Tier1 'complete' epigenomes. K562 cells were the first human immortalised myelogenous leukemia line to be established. K562 cells are of the erythroleukemia type, and the line is derived from a 53-year-old female CML patient in blast crisis. Embryonic stem cells (ES) are pluripotent stem cells derived from the inner cell mass of a blastocyst, an early-stage preimplantation embryo. GM12878 is a B-lymphocyte, lymphoblastoid cell line immortalized via Epstein-Barr Virus, from the International HapMap Project - CEPH/Utah genotype. This, together with available data in mice and specially in fruit fly coming from the modEncode project, offers an interesting scenario to explore functional hypothesis regarding epigenomic differences between two distant species.

The proof of principle for this exercise was already introduced in our original chroGPS publication [Font-Burgada et al., 2014]. Basically, instead of using Procrustes adjustment to remove biases due to different cell or technological backgrounds with a global MDS map generated from a jointly computed distance matrix, we will adjust two complementary distance matrices for each background separately, as joint overlaps between factors from different genomic backgrounds cannot be computed. Since Procrustes will adjust different configurations to match each other, having a separate distance matrix for each organism backgrond is enough to perform a valid adjustment, as long as enough common landmark elements are present in each one.

### 3.4.2 Integrating and comparing factor maps from different organisms

In order to start, we will need to generate separated chroGPS-factors maps for each biological background as defined in our case by the already known *Drosophila melanogaster* S2 modENCODE collection and the Human K562 collection available from ENCODE. Bear in mind that in this case study we will be integrating together data from two different biological but also technical backgrounds, as fly data is composed of ChIP-chip determined binging sites whereas the human one comes from ChIP-Seq techniques, providing much better resolution at identification of potential binding sites. As we will see, conservation of relative distances between epigenomic factors allows successful integration of both datasets accounting for systematic biases in an unsupervised manner.

The first step once our datasets are loaded in the respective *GenomicRanges* lists is to generate separated chroGPS-factors maps using the steps already described. Again, we will use the *diffGPS.factors* function to obtain both a graphical map with highlighted distances between common factors in both backgrounds as well as a ranked list of Procrustes errors. As described in [Font-Burgada et al., 2014], efficient matching of both maps with Procrustes depends on the existance of enough common anchor elements. It is therefore advisable to use the *domainDist* function already described previously to assess replicate coherence and identify potential outliers.

As opposed to previously presented case studies, the genomic background of the two compared maps here is different. Therefore, combining the collections of epigenomic factors for joint computation of similarities between epigenomic elements is not possible, and we have to rely on integration of both independently generated maps using Procrustes adjustment. This is specially important in case of integrating backgrounds with big differences in the number of mapped factors or very few common elements between them, as even though a Procrustes adjusted joint map can always be computed, conservation of the relative configuration on both maps is important for obtaining meaningful biological interpretation of the results.

Direct comparison of backgrounds in this case study already pointed to a potentially interesting difference in the biological role of CTCF boundary elements between fly and human. Figure 7 illustrates the main result of this case study, presenting a joint Procrustes adjusted map for human K562 and Drosophila S2, which present some remarkable similarities, with apparently some domains well conserved between both species. However, it is also

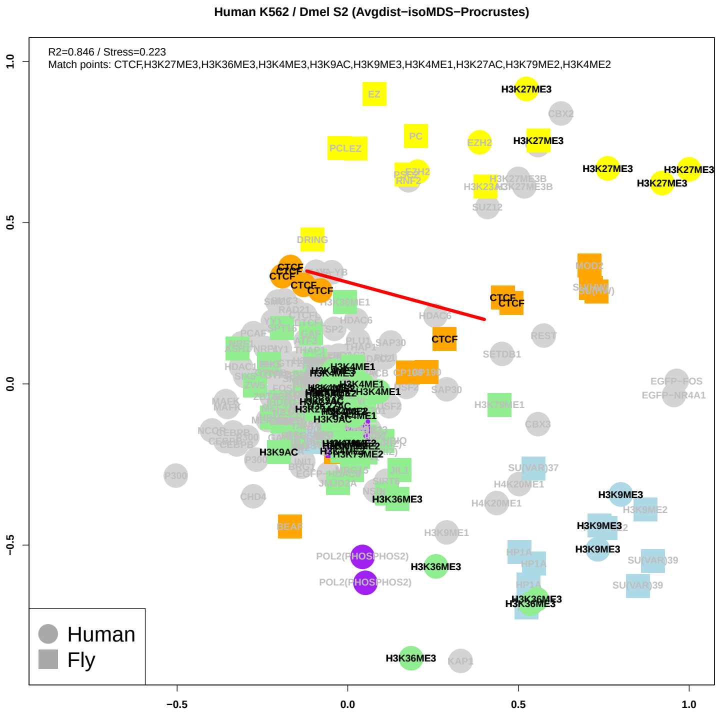


Figure 7: Differential chroGPS-factors map showing the epigenomic landscape differences in *Drosophila melanogaster* S2 cells and human K562 cell line. Notice the different location of CTCF elements in human (circles) and fly (squares), indicated by the red line.

clear that they present some strong differences, as is the apparent lack of conservation in the HP1 heterochromatin-related domain in human, with CBX heterochromatin related proteins spreading around several domains. Another more subtle difference at first sight is related to location of CTCF elements. CTCF is a well known insulator strongly associated with chromatin organization and chromatin loop formation, well located in the very center of the boundary-insulator domain in Drosophila, whereas apparently it locates close to the human BEAF element in the human map, surrounding the active transcription domain.

This finding prompted us to verify if this could be related to transcriptional orientation around CTCF sites. Briefly, we categorized all Drosophila dm3 intergenic regions based on orientation of adjacent transcription start and end sites, giving place to the three mutually exclusive categories of Head-to-Head (HH), Tail-to-Tail (TT) and Others (NN). Head-to-Head sites were those with adjacent TSS in opposite directions, while sites classified as Tail-to-Tail were those where transcription ended in the same adjacent region. Finally, NN indicate regions where a TXE is shortly followed by a TSS going in the same direction. Notoriously, as seen in Figure 8 such a basic classification of CTCF sites already unveiled a distribution for HH CTCF similar the one observed in human K562, while TT elements tend to locate in a region between external insulator elements and heterochromatin related proteins. NN CTCF elements shown almost the same behaviour as the original CTCFs. These findings go in par with recent works related to the complex nature of CTCF elements and their potential dual role in gene regulation and genome structure.

Supplementary R code and data available at:
`https://github.com/singlecoated/chroGPS2/tree/master/examples/diffSpecies'`

### 3.5 Integration of Hi-C Topological Associated Domains (TADs) data

Topologically associating domains (TADs) are a way of understanding the organization of mammalian genomes as being split up into "chromosome neighborhoods" within which most enhancer-promoter contact occurs. They can range in size from thousands to millions of DNA bases. TADs are separated from each other by boundary regions enriched for transfer RNA genes and for binding of the transcription factor CTCF. In [Gomez-Marin et al., 2015], authors provide a whole genome distribution of TADs over polytenic chromosomes in *Drosophila*. Integration of these results into chroGPS maps are addressed in two ways. First, TAD regions are directly incorporated into chroGPS factors map as an additional factor. Location of TADs is not surprisingly strongly associated with CTCF elements (Figure 8), comprising almost the exact center of the insultator-boundary domain.

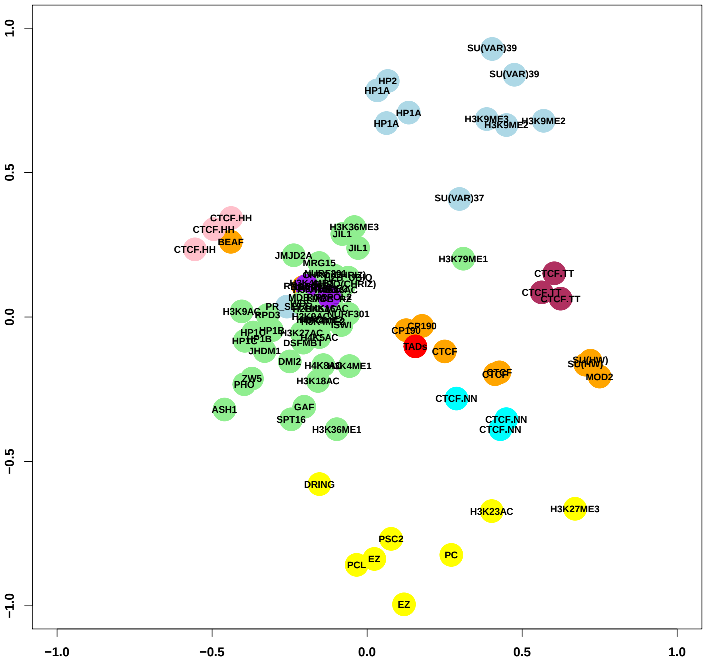


Figure 8: chroGPS-factors map *Drosophila melanogaster* S2 cells including TAD information and CTCF sites categorized categorized according to adjacent TSS/TXE configuration.

# 4 COMPARING CHROGPS-GENES MAPS

In chroGPS-genes maps, all genes in the genome are assigned a unique epigenetic 'profile', based on the epigenetic marks present at gene level in a certain biological scenario. Gene maps can also be used to highlight differences happening between different biological backgrounds such as different cells or tissues, diseases, and developmental or differentiation stages. In the following example we will explore several of these scenarios, using chroGPS to study differences between different epigenomic landscapes, as well as introduce the necessary methodology for performing this task.

## 4.1 Generation and functional analysis of chroGPS-genes maps.

The procedure to generate and analyze chroGPS-genes maps is detailed in our previous publication [Font-Burgada et al., 2014] as well as in the chroGPS package vignette, however here we offer a summarized how-to highlighting the general steps.

The initial set-up for generation of chroGPS-genes maps is a collection of epigenetic factors in the shape of a *GenomicRanges* list, and a set of genetic elements (i.e. genes). chroGPS-genes are aimed at visualization and analysis of genetic elements based on their epigenetic state, that is, the epigenetic marks they present. In this example we will focus on genes (as defined by the longest possible transcript for each gene in the UCSC dm3 genome. The base object for generating chroGPS-genes maps is a matrix of $G$ rows (usually genes) and $F$ columns (epigenetic factors), and where each cell $Gi,Fj$ of the matrix will take a value of 1 if a certain epigenomic mark j is reported for gene i, and 0 otherwise.

The resultant binary matrix is given as the main argument to the *distGPS* function, that will compute pairwise similarities/distances between rows of the matrix (in our case, genes), based on their epigenetic profiles. Thus, genes sharing a high number of common factors will present a high similarity value (or small distance), whereas ones sharing very few ones will be the opposite. As with chroGPS-factors, we offer different similarity metrics and ways to weight and penalize presence and absences of epigenetic factors when computing distances, as well as different ways to deal with technical replicates of the same epigenetic factor.

The following step is to use the resulting distGPS object to generate a low-dimensional representation using MDS, as we did in chroGPS-factors. In order to speed up computation of the MDS solution with gene maps that comprehend thousands of genes, we offer parallel computation using a random split-and-combine approach that also offers optimization of the resulting map in order to maximize goodness-of-fit and removes potential undesired effects from the randomization applied in the parallel computation.

After generation of the map, the next natural step in our approach is to perform a hierarchical clustering analysis using the clusGPS function, which by default performs a hierarchical clustering of the original distance matrix using average linkage. Methods for unsupervised determination of underlying number of clusters in our scenario and Bayesian non-parametric density estimation for assigning posterior probabilities of cluster identity for each element in the map are provided [Jara et al., 2011]. Final map visualization can also incorporate additional useful information such as expression values by means of palette color scales in the provided plot function, and point size to reflect cluster classification uncertainty. Functional analysis of the identified clusters can also be performed by means of assessing their enrichment / depletion on epigenetic marks when compared to the whole map using the *profileClusters* function. Figure 9 shows final result of this approach as analyzed in [Font-Burgada et al., 2014].

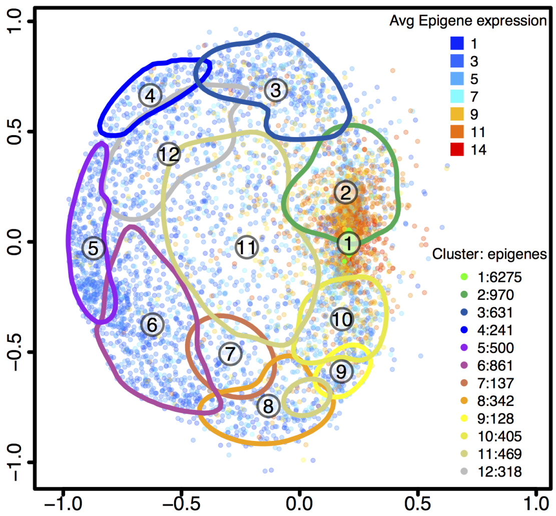


Figure 9: chroGPS-genes map of S2 cells. The map is analyzed using hierarchical clustering with average linkage. Clusters corresponding to 50 percent between-cluster distance are shown after unsupervised merging. The epigenetic state of each cluster is determined based on the log2 enrichment/depletion ratio of each factor. 'Epigenes' are colored according to their average log2 RMA expression levels in S2 cells [Kessler et al., 2015], and they distribute clearly according to Active and Repression domains. Legends on the right side indicate number of genes in each cluster and average gene expression level of each gene point.

## 4.2 Introduction to chroGPS-genes differential maps

Generation of differential chroGPS-genes maps to compare two epigenomic data sets is performed in a very similar way to the ones presented above. First, common factors between both backgrounds to compare are selected, usually after performing some operation to unify factor replicates when present. Then, for genes with at least one epigenetic mark in each background we compute pairwise distances between their epigenetic profiles using our metric of choice and a 2 dimensional map is generated via MDS. This allows to keep trace both of the epigenetic and background identity for each gene in the analysis. Since genes can be identified easily on the map by means of its epigenetic profile, it is straightforward to know which genes from different conditions present the same or very similar profiles (thus they are located in the same exact point or very close in the map), and which ones present significant differences (and therefore differ strongly in their location).

Downstream functional analysis of this differential map is essentially done in the same way as in a regular chroGPS-genes one. Briefly, hierarchical clustering is performed over the mathematical distances computed between each pair of unique epigenetic profiles. The identified clusters in this first step are then subject to an unsupervised merging process in order to further refine cluster definition and reduce granularity inherent to this method. Finally, after a final clustering configuration is obtained, each gene is assigned a posterior probability of correct classification by means of Bayesian density estimation procedures. In this way, we present not just a (simple and elegant) method to obtain an average robustness / reproducibility score for each one of the clusters and also for the global clustering solution, but also to obtain a posterior probability value indicating how sure we are that a given gene is 'well' located within its assigned cluster. This strategy proves very useful also not just to refine selection of those genes changing clearly from any two given clusters (i.e. genes suffering a strong change in epigenomic identity), but can also be used with differential maps to assign a probability score in order to rank gene changes involved in cluster changes of interest, providing us with an intuitive method to select potential candidate genes for further analysis in a scenario where two different backgrounds want to be compared.

Supplementary R code and data available at:
`https://github.com/singlecoated/chroGPS2/tree/master/examples/chroGPS-genes'`

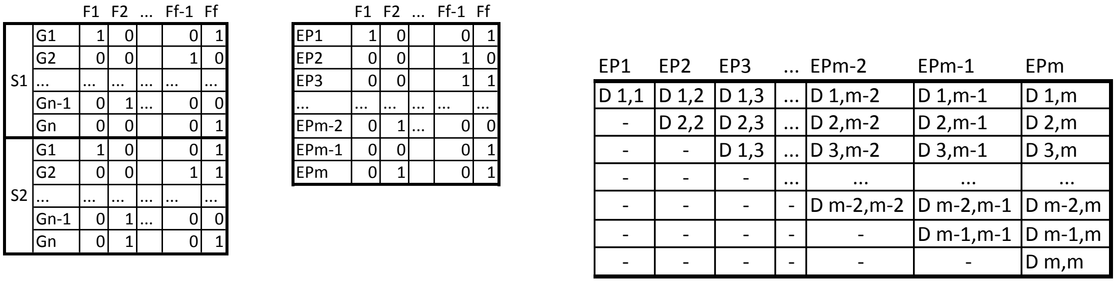
### 4.3 Comparing *Drosophila melanogaster* S2 and BG3 cell lines

We illustrate this with a case study comparing two well characterized *Drosophila melanogaster* cell lines, S2 and BG3. Schneider 2 cells, usually abbreviated asS2 cells, are one of the most commonly used Drosophila melanogaster cell lines. S2 cells were derived from a primary culture of late stage (20–24 hours old)*Drosophila melanogaster* embryos, likely from a macrophage-like lineage. BG3, however, is a cell line derived from central nervous system of third instar larvae.

To demonstrate how this strategy allows to identify changes related to biological differences between these two cell lines, following generation and functional analysis of the differential map involving common epigenetic factors present in S2 and BG3 we iterate over gene groups involved in cluster transitions between both backgrounds, and proceed to perform a Gene Ontology enrichment analysis for all genes involved in the observed cluster transitions.

The usual scenario in which to perform a differential chroGPS-genes map is that one involving studying epigenetic profiles over genes in two different conditions, biological or technical backgrounds from the same species, even though one could use the same methodology to address for instance changes in regulatory mechanisms in promoter vs coding regions or origins of replication, by simply recomputing binding site assignments under different genomic locations, etc. The basic elements to generate this map are the two collection of epigenetic elements which we want to compare at our desired level to use for generation of the necessary binary matrices described previously, or in its defect, two already computed matrices of 0s and 1s representing epigenetic profiles for each genetic element in each condition. Only common epigenetic factors mapped in both conditions will be taken into account (and therefore, only common genetic elements with at least 1 mapped factor in each condition are used). The *diffGPS.matrix* function takes raw *genes* x *factors* matrices for both backgrounds and returns a unique matrix containing unique epigenetic profiles for them (see 10).

This combined matrix is given as input to the *distGPS* function, together with the re-



|  |  | F1 | F2 | ... | Ff-1 | Ff |
|---|---|---|---|---|---|---|
| S1 | G1 | 1 | 0 |  | 0 | 1 |
|  | G2 | 0 | 0 |  | 1 | 0 |
|  | ... | ... | ... | ... | ... | ... |
|  | Gn-1 | 0 | 1 | ... | 0 | 0 |
|  | Gn | 0 | 0 |  | 0 | 1 |
| S2 | G1 | 1 | 0 |  | 0 | 1 |
|  | G2 | 0 | 0 |  | 1 | 1 |
|  | ... | ... | ... | ... | ... | ... |
|  | Gn-1 | 0 | 1 | ... | 0 | 0 |
|  | Gn | 0 | 1 |  | 0 | 1 |

|  | F1 | F2 | ... | Ff-1 | Ff |
|---|---|---|---|---|---|
| EP1 | 1 | 0 |  | 0 | 1 |
| EP2 | 0 | 0 |  | 1 | 0 |
| EP3 | 0 | 0 |  | 1 | 1 |
| ... | ... | ... | ... | ... | ... |
| EPm-2 | 0 | 1 | ... | 0 | 0 |
| EPm-1 | 0 | 0 |  | 0 | 1 |
| EPm | 0 | 1 |  | 0 | 1 |

| EP1 | EP2 | EP3 | ... | EPm-2 | EPm-1 | EPm |
|---|---|---|---|---|---|---|
| D 1,1 | D 1,2 | D 1,3 | ... | D 1,m-2 | D 1,m-1 | D 1,m |
| - | D 2,2 | D 2,3 | ... | D 2,m-2 | D 2,m-1 | D 2,m |
| - | - | D 1,3 | ... | D 3,m-2 | D 3,m-1 | D 3,m |
| - | - | - | ... | ... | ... | ... |
| - | - | - | - | D m-2,m-2 | D m-2,m-1 | D m-2,m |
| - | - | - | - | - | D m-1,m-1 | D m-1,m |
| - | - | - | - | - | - | D m,m |

Figure 10: Left: stacked binary matrix with presence/absence information for common genes and epigenetic factors from backgrounds S1 and S2. Center: The unique epigenetic profiles for both backgrounds are used to compute a unique similarity/distance matrix based on pairwise similarity between the vectors for all epigenetic profiles (right).

spective labels identifying each background. The resulting object is used to generate a low-dimensional map of all genetic elements based on the similarity of their epigenetic profiles using MDS. All other methodologies used with regular chroGPS-genes maps (clustering, etc) are available for differential maps as well. To sum up, the map just represents a certain number of genetic elements on space based on similarity of their epigenetic profiles. Thus, two genes sharing exactly the same epigenetic profile will be located at exactly the same point, whereas changes in their profiles will translate into shifted positions. These changes and their potential functional effects are indentified and analyzed with the *diffGPS.genes* function.

This function takes as input the binary matrix with combined epigenetic profiles for both backgrounds and the clustering object produced by *clusGPS*, and retrieves cluster identities and posterior probabilities of classification. This information is provided for each genetic element analyzed and for both conditions of interest, and therefore it is straightforward not only to have a trace of all epigenetic cluster changes taking place for all analyzed genetic elements, but also to rank those changes based on confidence estimates as provided by the posterior probabilities of classification.

Once cluster transitions are obtained, a very straightforward approach for downstream analysis and is to perform Gene Ontology enrichment analysis via hypergeometric testss using the provided *enrichGPS* function. Alternatively, the differentially located gene lists can be analyzed with other functions such as the *getEnrichedGO* from the *ChIPpeakAnno* package, or using, DAVID, Gene Set Enrichment Analysis or others. In order to illustrate this, we show results of our provided function *enrichGPS* to investigate potentially interesting Biological Process terms enriched among genes involved in cluster transitions.

Between all cluster transitions analyzed (See supplemental Table 1), in Figure 11 we present the differential map highlighting results for genes involved in cluster transition 5 (S2) to 2 (BG3). Top Gene Ontology enriched results from the list of 70 genes involved in this cluster transition present a strong and statistically significant enrichment in processes such as axon guidance, neuron growth and other central nervous system ones.

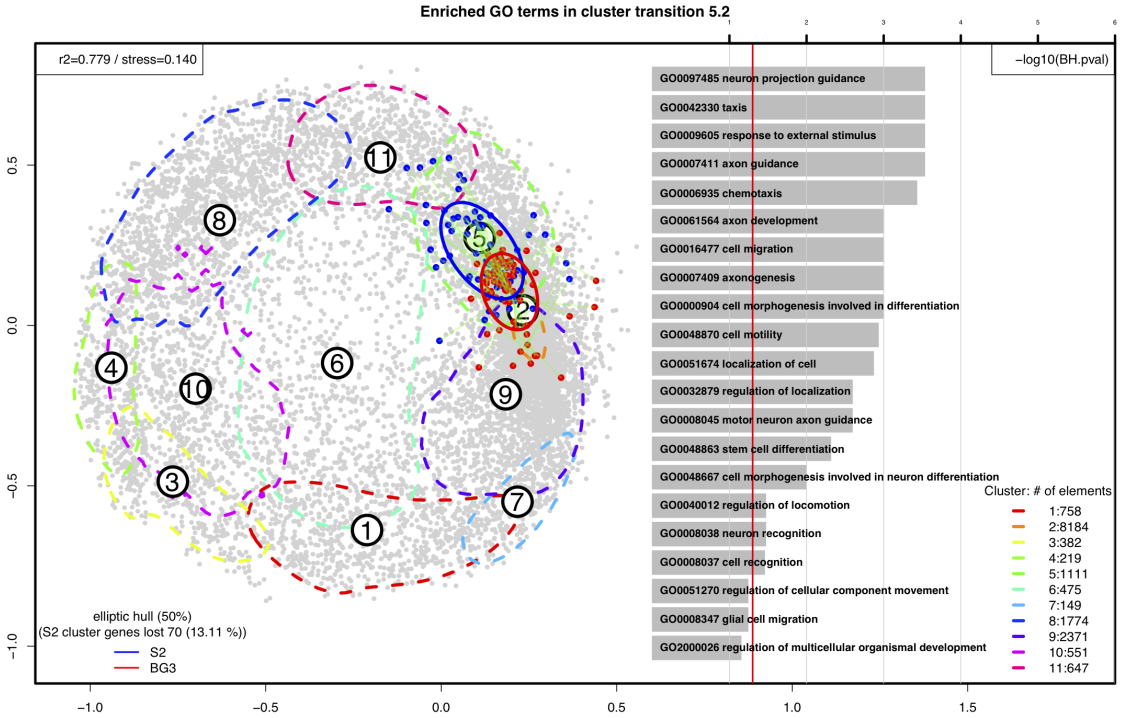


Figure 11: chroGPS-genes differential map of *Drosophila melanogaster* S2 and BG3 cells showing top 20 Biological Process enriched terms for genes involved in cluster transition 2/5.

Supplementary R code and data available at:
`https://github.com/singlecoated/chroGPS2/tree/master/examples/diffGenes'`

# 5 SOME NOTES ON VISUALIZATION

## 5.1 Exporting chroGPS maps to XGMML / Cytoscape

Maps generated with chroGPS can be exported to GraphML XGMML format using the provided funtion gps2xgmml. These maps can be imported with several R network analysis packages such as *igraph* and are also suitable to use with external network visualization and analysis solutions like Cytoscape. Additional information regarding distances between elements in the map and all other information can be easily exported into tabular format files and assigned to the imported network nodes using the standard procedures for each software.

## 5.2 Exporting chroGPS maps to HTML5 with plotLy, RStudio and Shiny

RStudio is an integrated development environment (IDE) for R available in open-source and commercial editions and which runs both in desktop and server mode under the most used operating systems. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. Shiny is an R package that makes it easy to build interactive web apps straight from R. You can host standalone apps on a webpage or embed them in R Markdown documents or build dashboards. You can also extend your Shiny apps with CSS themes, htmlwidgets, and JavaScript actions. Seamless RStudio and Shiny integration makes them a good solution to export graphical results into interactive web pages. Plotly, also known by its URL, Plot.ly, is an online data analytics and visualization tool that provides online graphing, analytics, and statistics tools as well as scientific graphing libraries for Python, R, MATLAB, and other languages.

Supplementary R code and data available at:
`https://github.com/singlecoated/chroGPS2/tree/master/examples/viz'`

## References

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol., 28(10):1045–1048.

Borg, I. and Groenen, P. (2003). Modern multidimensional scaling: Theory and applications. Journal of Educational Measurement, 40(3):277–280.

Celniker, S. E., Dillon, L. A., Gerstein, M. B., Gunsalus, K. C., et al. (2009). Unlocking the secrets of the genome. Nature, 459(7249):927–930.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414):57–74.

Ernst, J. and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol., 33(4):364–376.

Font-Burgada, J., Reina, O., Rossell, D., and Azorin, F. (2014). chroGPS, a global chromatin positioning system for the functional analysis and visualization of the epigenome. Nucleic Acids Res., 42(4):2126–2137.

Gel, B., Diez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., and Malinverni, R. (2016). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics, 32(2):289–291.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology, 5:R80.

Gomez-Marin, C., Tena, J. J., Acemel, R. D., Lopez-Mayorga, M., Naranjo, S., de la Calle-Mustienes, E., Maeso, I., Beccari, L., Aneas, I., Vielmas, E., Bovolenta, P., Nobrega, M. A., Carvajal, J., and Gomez-Skarmeta, J. L. (2015). Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc. Natl. Acad. Sci. U.S.A., 112(24):7542–7547.

Jara, A., Hanson, T. E., Quintana, F. A., Muller, P., and Rosner, G. L. (2011). DPpackage: Bayesian Non- and Semi-parametric Modelling in R. J Stat Softw, 40(5):1–30.

Kendall, D. G. (1989). A survey of the statistical theory of shape. Statistical Science, 4(2):87–99.

Kessler, R., Tisserand, J., Font-Burgada, J., Reina, O., Coch, L., Attolini, C. S., Garcia-Bassets, I., and Azorin, F. (2015). dDsk2 regulates H2Bub1 and RNA polymerase II pausing at dHP1c complex target genes. Nat Commun, 6:7049.

Zhu, L. J., Gazin, C., Lawson, N. D., Pages, H., Lin, S. M., Lapointe, D. S., Green, M. R., et al. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics, 11:237.

# Chapter 4

# Discussion and Conclusions.

It is necessary to provide efficient and intuitive tools to visualize epigenomics data [Marx, 2015]. Recent advances in the ability to generate, analyze and share vast amounts of genomic and epigenomics information in a wide range of scenarios gave place to unprecedented opportunities to explore and test biological hypothesis [Celniker et al., 2009; Dunham et al., 2012; Bernstein et al., 2010]. There is, however, a need to develop intuitive, user friendly and computationally feasible tools to visualize these data, and to encode the obtained conclusions from its functional and comparative analysis as visual information that can be easily interpreted, capuring and reflecting the underlying structure in a manner which is unbiased, to therefore 'tell the truth and nothing but the truth' [Bertin, 1967].

chroGPS is a computational framework to measure, visualize and compare epigenetic similarity between factors and/or genetic elements and to represent them in low dimensional maps using Multidimensional Scaling (MDS), Procrustes and clustering techniques [Borg and Groenen, 2003; Kendall, 1989; Tan et al., 2005]. As a proof of principle, we used data generated by the modENCODE project [Celniker et al., 2009] in the *Drosophila* S2 and BG3 and human K562 cell lines [Dunham et al., 2012] to present several case studies using two types of maps: chroGPS-factors and chroGPS-genes.

**Generating chroGPS-factors maps**

chroGPS-factors describes similarities between epigenetic factors and inform about their functional association (fig. 4.1). A similarity measure between factors is computed based on genomic distribution of their binding profiles using the average interval overlap (iOverlap) metric [Font-Burgada et al., 2014]. These similarities are then represented in 2 or 3 dimensional maps using non-metric MDS (isoMDS) that jointly estimates a non-parametric monotonic relationship between the original and the graphical Euclidean distances in the map. The resulting representation accurately describes known epigenetic states (fig. 4.1): active chromatin (in green), boundary / insulator function (in grey), HP1 heterochromatin related repression (in blue) and Polycomb (PC) dependent silencing (in yellow).

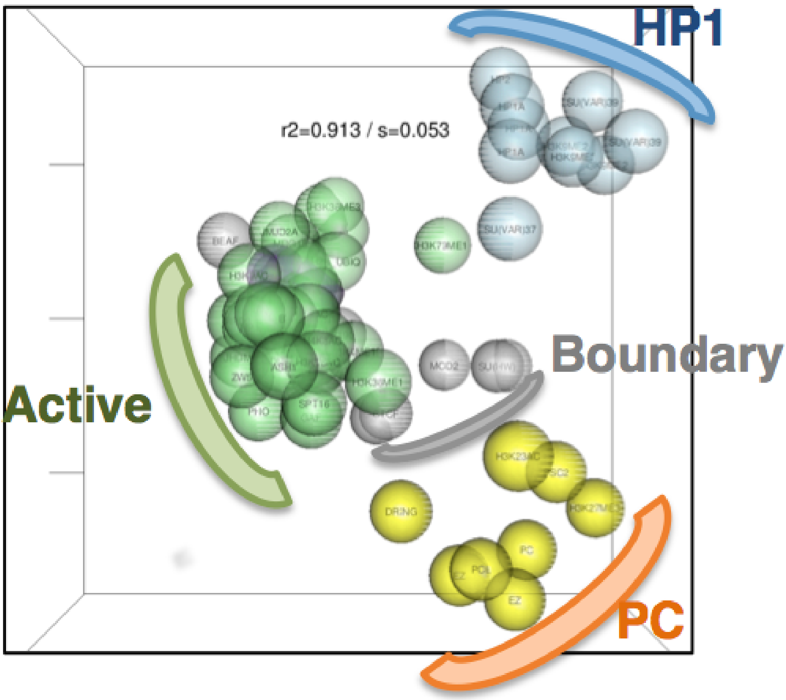


Figure 4.1: chroGPS-factors 3D map of *Drosophila* S2 cells. Similarity between 76 individual epigenetic factors, as determined from their genomic profiles using iOverlap, is represented in a 3D epigenetic map using isoMDS. Factors are colored according to their biological activity: regulation of transcription (green), boundary/insulator function (grey), HP1 (blue) and Polycomb (PC)-dependent silencing (yellow). RNApol II is indicated in purple. The squared Pearson correlation ($R^2$) between original and approximated distances and the classical *stress-1* function (section 1.5.1) are indicated.

After generation of the *Drosophila melanogaster* S2 map, we used Procrustes to integrate in-house generated ChIP-seq data, in order to assess functional relationships between the novel and previously undescribed insulator / boundary factors Ibf1 and Ibf2 (fig. 4.2) [Cuartero et al., 2014; Cuartero i Betriu, 2014], and several proteins involved in regulation of H2Bub1 and involved in RNA-Polymerase-II pausing at HP1c complex target genes (figs. 4.2 and 4.3) [Kessler et al., 2015; Kessler, 2014]. Successful integration of these data contributed to obtain a visual representation of their global scenario within *Drosophila* S2 cells.

### Generating chroGPS-genes maps

chroGPS-genes integrates epigenetic marks at gene level and describes the epigenetic context of gene expression and function (fig. 4.4). In this case, epigenetic similarity between genes is defined based on the epigenetic marks that they have in common and is determined using the Tanimoto metric [Font-Burgada et al., 2014]. chroGPS-genes maps may contain tens of thousands of points, which constitutes a high-dimensional problem posing important computational challenges for MDS representation. To overcome these limitations, we developed a novel two-step procedure, *BoostMDS*, that finds an initial solution by splitting the distance matrix into smaller partially

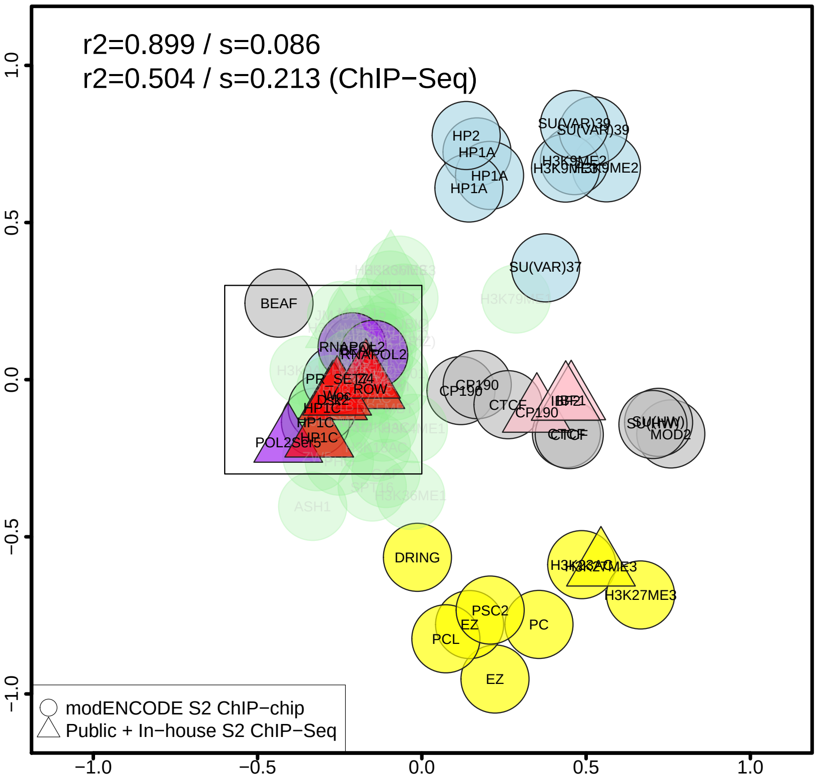


Figure 4.2: 2-dimensional version of the map in (fig. 4.1). After determination of whole-genome binding sites by performing ChIP-seq bioinformatics analysis, several factors of interested were integrated in the map to assess their potential functional relationships from a global perspective, using Procrustes to identify and adjust for systematic technical biases between ChIP-chip and ChIP-seq data. In pink: the novel insulator / boundary related factors Ibf1 and Ibf2 [Cuartero et al., 2014; Cuartero i Betriu, 2014], which localized precisely with modEncode related proteins CP190 and CTCF, showing a clear and strong co-localization with CP190 and suggesting functional relationship with these elements from this chromatin domain. In red: HP1c, WOC, ROW, Dsk2 and Z4 [Kessler et al., 2015; Kessler, 2014], well located within the active region surrounding HP1c and RNA-Polymerase-II.

overlapping sub-matrices, which are then computationally tractable. Since each pair of adjacent distance matrices share a certain number of common anchor points, Individual solutions can be adjusted sequentially using Procrustes to conform a global representation (fig. 4.4). Also, as genomic interval data objects are usually sorted by chromosomal coordinates, adjacent genes in the original data matrix tend to present very similar epigenetic profiles, which would translate in closely spaced MDS points. In order to ensure that most of the selected anchor overlapping points are distributed equally across the whole map for good Procrustes performance, the original rows in the data are shuffled prior to distance computation. Furthermore, to prevent this random shuffling having any undesired effects on the final result and to ensure result reproducibility, the obtained solution is then refined by formally maximizing the $R^2$ coefficient via a gradient search algorithm [Strickert et al., 2005].

In chroGPS-genes maps each point represents a group of genes that share an identical set of epige-

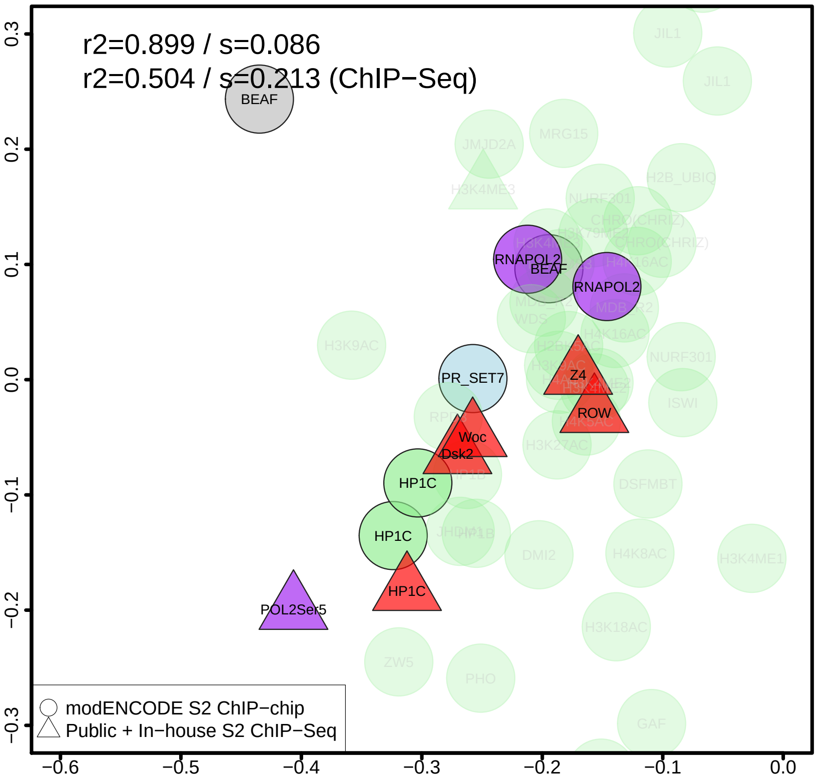


Figure 4.3: Zoom-in view of the active transcription region in (fig. 4.2) to show detailed location of HP1c, WOC, ROW, Dsk2 and Z4 ChIP-seq in-house generated factors.

netic factors or, in other words, an elusive epigene [Bird, 2007]. Functional analysis is performed using hierarchical clustering with average linkage to identify groups of genes with similar epigenetic profiles, an approach which returned a very large number of clusters with a high degree of spatial overlap [Font-Burgada et al., 2014], complicating interpretation. To solve this, we performed an unsupervised cluster merging process, by joining adjacent clusters based on Bayesian density estimation [Jara et al., 2011] until their degree of overlap dropped swiftly, greatly improving the Correct Classification Rate (CCR) [Font-Burgada et al., 2014]. By using this approach, 12 clusters can be identified that are specifically enriched/depleted in particular epigenetic factors and describe distinct epigenetic states (Figure 2): actively transcribed genes (clusters 1 and 2), HP1-silenced (clusters 3 and 4) and PC-silenced genes (clusters 5, 6, 7, 8, and 10, fig. 4.4).

### Comparing chroGPS-factors and chroGPS-genes maps

The use of chroGPS is not limited to the examples described above since it is a general tool applicable to a wide range of situations. In particular, chroGPS maps can be generated for a particular developmental process or disease, as well as based on certain genomic regions of interest such as gene promoters [Reina et al., 2017] in order to study functional relationships between epigenetic factors under specific circumstances. These maps can identify genetic / epigenetic transitions / alterations of the analyzed elements between different backgrounds or conditions. So, for instance,

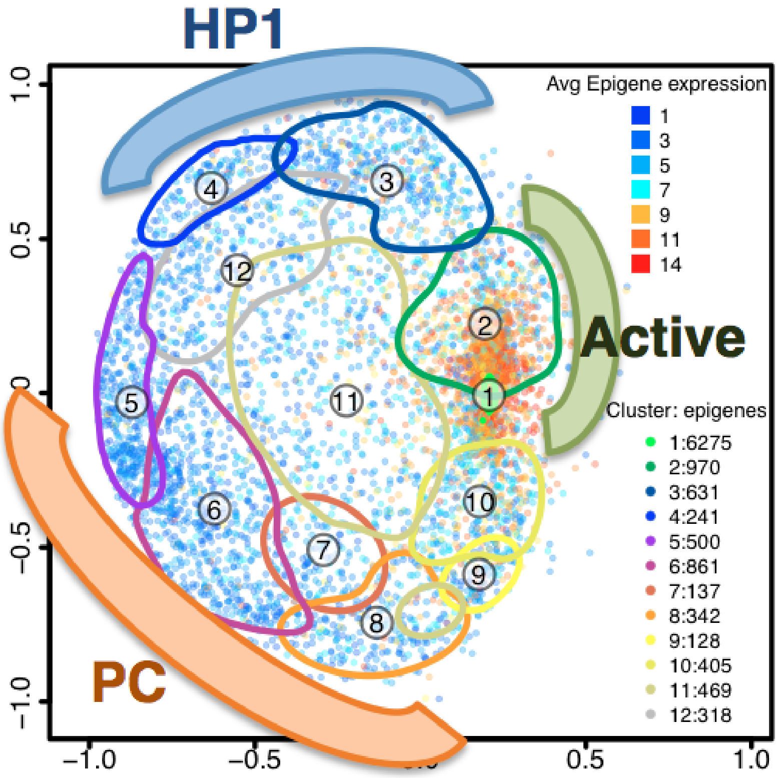


Figure 4.4: chroGPS-genes map of S2 cells. The map is analyzed using hierarchical clustering with average linkage. Clusters corresponding to 50 percent between-cluster distance are shown after unsupervised merging. The epigenetic state of each cluster is determined based on the log2 enrichment/depletion ratio of each factor. 'Epigenes' are colored according to their average log2 RMA expression levels in S2 cells [Kessler et al., 2015], and they distribute clearly according to Active and Repression domains. Legends on the right side indicate number of genes in each cluster and average gene expression level of each gene point.

comparing chroGPS-genes maps from normal and affected cells in a given disease could identify which genes are changing epigenetic status and in what direction(s).

Comparing maps when they contain a reduced number of items, such as in chroGPS-factors maps, is relatively straight-forward. Maps can be superimposed by using the factors shared between different maps as anchor points using Procrustes adjustment [Kendall, 1989]. Procrustes superimposes two sets of points by altering their center, scale and orientation, while preserving relative distances within each set, and relevant differences between sets can be obtained via Procrustes sum of squared errors. This approach has been sucessfully applied to integrate and compare the epigenetic landscape of *Drosophila* S2 cells at different time points during replication [Reina et al., 2017](fig. 4.5), but can be extended to compare different cell lines or tissues. In this way, factors that significantly change their relative position in the maps could be readily identify and the observed alterations assesed from a functional perspective.

In order to compare chroGPS-genes maps, a similar approach is used, by generating maps considering all possible epigenetic profiles for genes from both conditions. Then, functional analysis

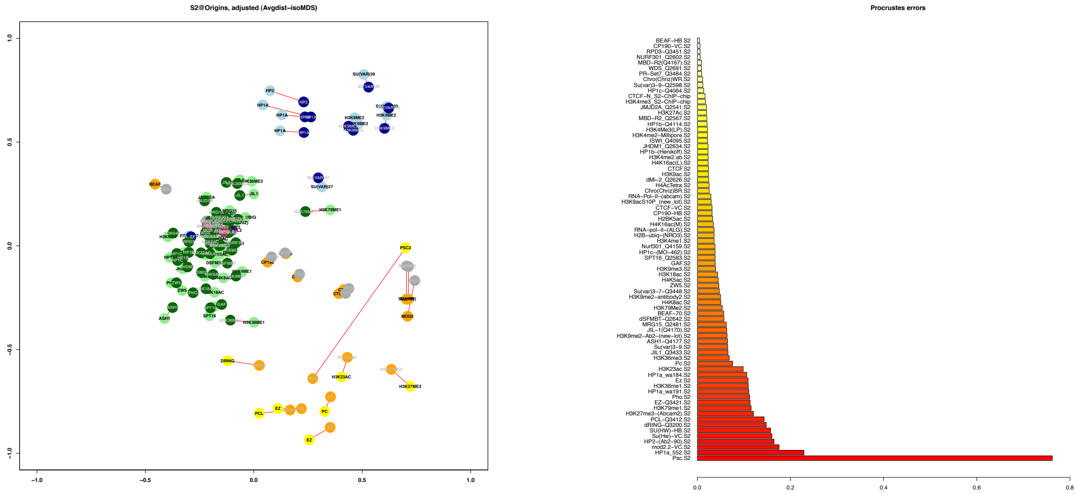


Figure 4.5: Differential chroGPS-factors map of *Drosophila melanogaster* S2 cells at Early vs Late Origins of Replication (left), and Procrustes normalized errors (right). In this differential map, PSC2, a Polycomb related protein involved in non-heterochromatic gene silencing, is strongly shifted from the boundary-insulator region (Early) towards the Polycomb repression one (Late).

is performed using hierarchical clustering with average linkage, unsupervised cluster merging and Bayesian density estimation [Jara et al., 2011] to identify and characterize groups of genes presenting significant changes between both conditions [Reina et al., 2017]. Akin with results obtained from other high-throughput technologies, candidate gene lists can be used for further downstream functional analysis, such as Gene Ontology [Ashburner et al., 2000] enrichment [Ashburner et al., 2000] or Gene Set Enrichment Analysis [Subramanian et al., 2005], as well as compared with other genomics data such as changes in gene expression or regulation to further explore interesting hypothesis. We illustrate the use of this strategy to identify and characterize genes potentially involved in epigenomic changes between third instar larvae *Drosophila* S2 cells and BG3 neuronal tissue (4.6).

### Assessing the core set of factors to generate biologically meaningful epigenomic maps

Obtaining data on a large number of epigenetic factors for multiple conditions may not always be affordable both technically and economically. In this regard, it must be noticed that epigenetic states are generally determined by the concurrence of several epigenetic factors that, from this point of view, provide redundant information [Ernst and Kellis, 2015]. Therefore, it seems feasible to select a reduced set of factors to generate maps without losing their resolution and biological information. Data availability which allowed generation of chroGPS maps for model organisms also allows addressing this question in an unbiased manner. In detail, when chromatin domain information is available for an organism of interest, we assess cohesion and separation (fig. 1.15) of these domains in the map to select those candidate groups of factors which provide more information to the obtained representation. Additionally, or when chromatin domain information is not available or unknown, we use logistic and linear regression to estimate the predictive power of groups of marks, helping in the selection of potential candidates for de-novo epigenome mapping, so that they lead to a meaningful and useful map from a biological point of view [Reina et al., 2017](fig. 4.7).

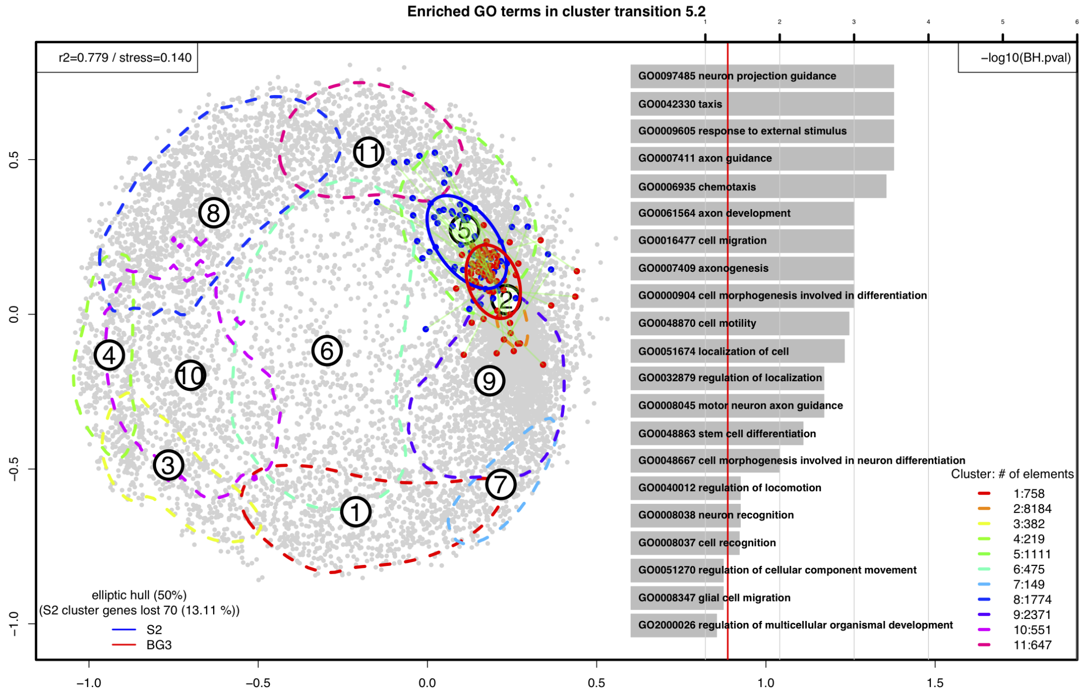


Figure 4.6: Differential chroGPS-genes map of *Drosophila melanogaster* S2 and BG3 cell lines. Red and blue ellipses illustrate genes changing from cluster 5 in S2 (blue ellipse, moderate HP1 repression) to cluster 2 in BG3 (red ellipse, transcriptionally active). Dashed lines indicate probability contours containing 50 percent of genes for each cluster. On the right, -log10 of Benjamini-Hochberg hypergeometric test pvalues for the top 20 enriched Gene Ontology terms for genes involved in the analyzed cluster transition are shown. Top left legend indicates $R^2$ coefficient and *stress* - 1 (section 1.5.1) values for the shown MDS solution. Bottom left legend indicates proportion of genes lost in the S2 (blue ellipse) cluster. Selected genes are specially enriched in nervous system and cell differentiation categories.

### Merging epigenetic data from different species

Available data for the generation of epigenetic maps in disease relevant genomes (i.e., humans, mice) is not abundant. However, considering the high evolutionary conservation of the main epigenetic factors and their functional interactions [Woo and Li, 2012], it seems reasonable to explore the possibility that data obtained in different species could be merged to generate high resolution species-independent maps.

Procrustes allows successful integration of epigenetic information from different conditions, technological or biological backgrounds into biologically meaningful chroGPS-factors maps, thanks to the strong conservation of certain epigenetic factor relationships [Font-Burgada et al., 2014]. As a side result, factors in which this degree of conservation is not present can be easily identified [Reina et al., 2017]. Interestingly, since such conservation is also observed at an inter-species level [Woo and Li, 2012], we used the same methodology to efficiently integrate inter-species information. As a result, these techniques allowed to highlight potential significant similarities and differences between epigenetic programs across species, pointing towards potentially interesting functional dif-

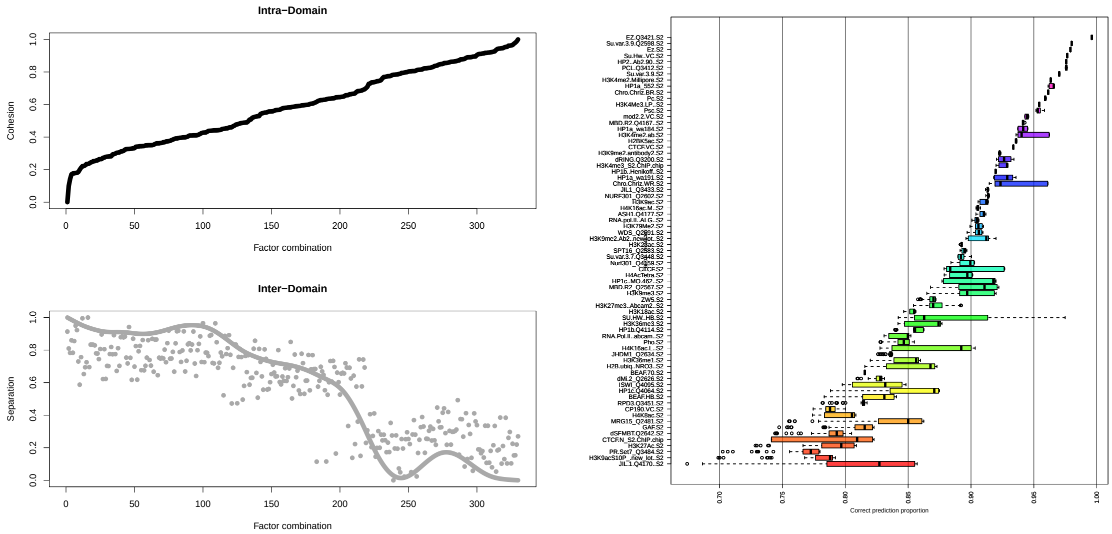


Figure 4.7: Left: From left to right, resulting domain integrity (Cohesion/Separation, fig. 1.15) based on computation of Intra and Inter-Domain distances for each combination of 4 factors within the HP1a repression domain. Results are sorted by Intra-Domain distance, which is an informative concept regarding cohesion of the observed domains in the map. Potential candidates are those among the leftmost factor combinations minimizing Intra-Domain distance but still providing meaningful biological content. Right: Factor ranking based on logistic regression of functional relationship with genetic elements when no domain information is available. Left axis, from top to bottom, indicates factor removed at each iteration, which is the one having the higher correct prediction rate based on information from those ones remaining below, until the number of remaining factors is lower than a given threshold. Boxplots in each line indicate the obtained prediction rates for all removed factors on top, based on the information from those on the bottom. Notice that as we remove more factors, general prediction rates get overall lower.

ferences between certain epigenetic factors (fig. 4.8).

This finding prompted us to verify if this could be related to transcriptional orientation around CTCF sites. Briefly, we categorized all Drosophila dm3 intergenic regions based on orientation of adjacent transcription start and end sites, giving place to the three mutually exclusive categories of Head-to-Head (HH), Tail-to-Tail (TT) and Others (NN). Head-to-Head sites were those with adjacent TSS in opposite directions, while sites classified as Tail-to-Tail were those where transcription ended in the same adjacent region. Finally, NN indicate regions where a TXE is shortly followed by a TSS going in the same direction. Notoriously, as seen in fig. 4.9, such a basic classification of CTCF sites already unveiled a distribution for HH CTCF similar the one observed in human K562, while TT elements tend to locate in a region between external insulator elements and heterochromatin related proteins. NN CTCF elements shown almost the same behaviour as the original CTCFs.

### Considerations

Even though computation of distances for generation of chroGPS-factors maps using parallel computing is not a major issue, it may be relatively slower for very large datasets with hundreds of

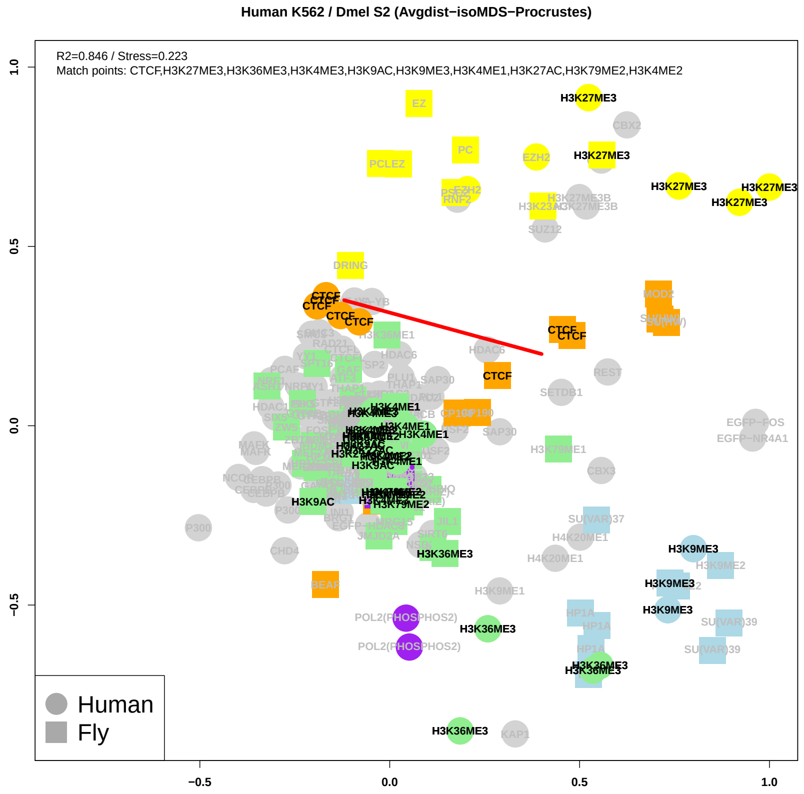


Figure 4.8: Joint Procrustes adjusted map for human K562 and *Drosophila* S2, which present some remarkable similarities, with apparently some domains well conserved between both species. However, it is also clear that they present some strong differences, as is the apparent lack of a clear HP1 heterochromatin-related domain in human, with CBX heterochromatin related proteins spreading between fruit fly HP1 and PC-related repression [Kaustov et al., 2011]. Another more subtle difference at first sight is related to location of CTCF elements. CTCF is a well known insulator strongly associated with chromatin organization and chromatin loop formation [Van Bortle and Corces, 2013], well located in the very center of the boundary-insulator domain in *Drosophila*, whereas apparently it locates close to active transcription group in the human map.

factors. However, when performing integration of additional data into well-known and studied factors maps to assess their functional relationships, it is enough to compute only distances for the new elements. In this way, we envision offering precomputed distance matrices for several chroGPS-factors maps in several organisms and cell lines, so that it is enough to integrate the new factors of interest to generate and analyze their functional relationships within the resulting maps.

Large datasets may pose a problem also in terms of used computer memory, specially when using R in UNIX/Linux environments, in which R memory management can seriously compromise the operating system. Several R packages, such as *bigmemory* can take care of working with large datasets using special indexed virtual memory management, and are recommended for hardware configurations with moderates amount of RAM.

chroGPS maps provide useful insight on epigenetic marks both at factor and gene level. However, there are also drawbacks that hamper interpretation. In this regard probably the most noticeable

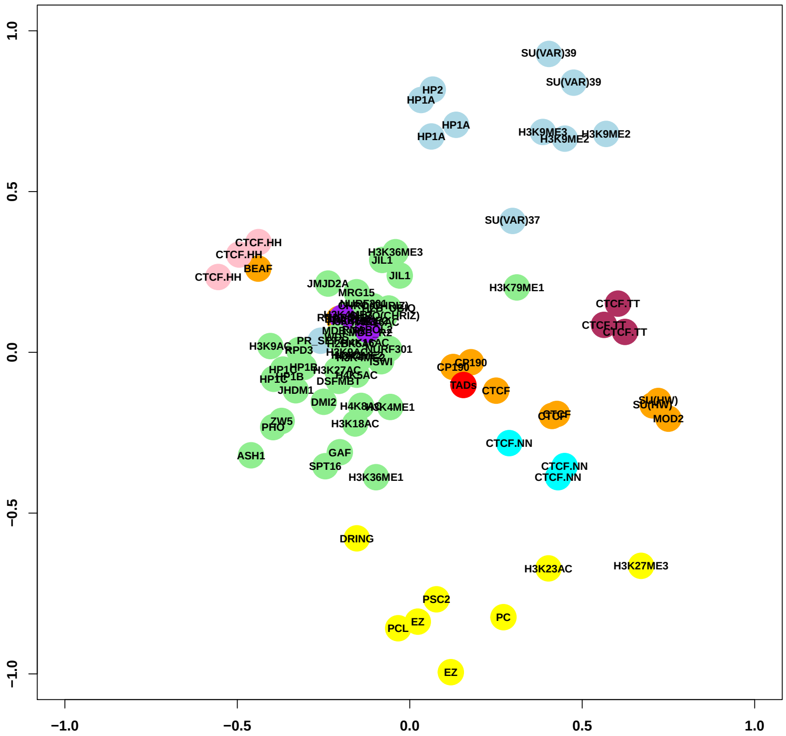


Figure 4.9: The adjacent/divergent TSS classification and its effect on *Drosophila melanogaster* CTCF location in the S2 2-dimensional chroGPS-factors map.

is the very high number of factors and genes that are present in the active transcription region in both maps. Further work is needed regarding the establishment of the 'core' factors defining relevant chromatin domains in order to eliminate potentially redundant information from elements carrying almost the same epigenetic profile information, which reduces interpretability of the map in some regions.

Additionally, even though ChIP-seq techniques account for a much higher resolution than previously used ChIP-chip, they are not free from issues either, ranging from antibody specificity problems to GC-content bias [Teng and Irizarry, 2017], something that highlights the importance of re-visiting existing data in order to perform adequate quality control analysis and, if needed, to reprocess and re-analyze it in order to obtain 'clean' curated versions of current public datasets.

The finding related to differential distribution of CTCF elements based on adjacent TSS/TXE orientation also points towards a need to further study the possible existence of binding site clusters within the same mapped factor that could relate to differentiated roles associated to distinct chromatin domains. This could be approached in two ways: 1) by assessing if a given factor presents binding site clusters with different intrinsic characteristics such as genomic width and/or intensity [Starmer and Magnuson, 2016]; 2) by assessing the possible relationship between binding site location with other genetic elements (TSS/TXE, promoters, enhancers or tandem repeated sequences). Both approaches can be supported by a computational methodology using chroGPS maps to estimate loss and gain in domain / cluster cohesion and separation, in order to converge on an optimal re-classification result.

Regarding generation of inter-species maps, while evolutionary conservation has been observed for several groups of epigenetic marks [Woo and Li, 2012], it cannot be taken for granted, as map domain identity may also be lost between species [Kaustov et al., 2011]. Therefore, accurate selection of epigenetic marks to act as Procrustes landmark points in this kind of integration is crucial, and have to be chosen based not only on an evenly spatial distribution of these points across the map, but also taking into account the existing biological knowledge available. Selection of landmark points for Procrustes analysis between maps from the same species is not a trivial task either, and even though it can be approached from a biological perspective based on domain conformation of the maps, in general some numerical estimation of landmark point 'goodness' is also desirable in order to help in the selection of 'bridge' datasets to accommodate different backgrounds in a unique map.

Finally, even though in recent years there was a massive increase in generation and publication of datasets with comprehensive genetic and epigenetic information, some of the consortiums who produced this information are no longer generating new data [Boley et al., 2014], something which poses a limitation in the generation of functional hypothesis to address using this technique, for instance to explore scenarios regarding developmental stages and their functional relationships with gene expression [Li et al., 2014]. It is therefore convenient to further refine and assess results from maps generated using both experimental and imputed data [Ernst and Kellis, 2015] in order to face this challenge.

### Closing remarks

chroGPS is available as a package for Bioconductor [Gentleman et al., 2004], a comprehensive collection of open-source libraries for computational analysis of 'omics' data using the R statistical language [R Development Core Team, 2008], and chroGPS2 will be available in the same repository. Both packages offer a computational framework based on Multidimensional Scaling and are designed to explore combinations of multiple data types, accounting for systematic biases and that can focus both on genetic elements and epigenetic factors. Our main contribution is enabling the integration and comparison of massive heterogeneous epigenetics data in a visually appealing and context-rich manner, by using techniques favoring interpretation and clear visual encoding of variable information and that relate well to human ability for visual perception of similarities and differences between elements, so that we can actually look at the underlying characteristics of the data [Cooley and Tukey, 1965; Bertin, 1967; Kendall, 1989; Borg and Groenen, 2003; Tan et al., 2005; Borg et al., 2012].

Aside from favoring representation of epigenetics information in a way that is easy to interpret, we also assessed the adequacy of multiple distance metrics from a statistical point of view, and provided algorithms to represent a large number of objects at high resolution using a computational effort manageable by a modern desktop or workstation computer by designing and implementing an algorithm to perform parallel computation of large dissimilarity matrices using Multidimensional Scaling, as well as offering further strategies to annotate the maps in order to enhance their interpretability [Font-Burgada et al., 2014; Reina et al., 2017].

chroGPS maps proved useful in a variety of situations, such as understanding functional interplays between existing and novel epigenetic factors in *Drosophila*, to assess and explore conservation and differences across S2 and BG3 cells, deriving testable hypotheses for novel factors, studying chromatin states at genes and the epigenetic regulation of complex pathways [Font-Burgada et al., 2014], to compare differences between biological backgrounds from an epigenetic point of view at factor and whole-genome level, to investigate epigenetic differences between species, and to assess the set of factors conforming chromatin domain identity [Reina et al., 2017].

Additionally, we used chroGPS to generate maps that consider only overlaps at specific locations (e.g. promoters or origins of replication) [Reina et al., 2017] to inform about epigenetic states and compare functional relationships occurring at the investigated elements, providing further insight at the potential regulatory mechanisms involved in the complex molecular, spatial and temporal choreography that exists within the cell nucleus of eukaryotic organisms, as a set of goggles to have a look under the surface of the vast ocean of epigenetics inexplicability.

# Bibliography

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular Biology of the Cell, Fourth Edition. Garland Science, 4 edition.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25(1):25–29.

Azzalini, A. and Menardi, G. (2014). Clustering via nonparametric density estimation: The r package pdfcluster. Journal of Statistical Software, Articles, 57(11):1–26.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., Zhang, J., et al. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput. Biol., 9(11):e1003326.

Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res., 21(3):381–395.

Bayona-Feliu, A., Casas-Lamesa, A., Reina, O., Bernues, J., and Azorin, F. (2017). Linker histone H1 prevents R-loop accumulation and genome instability in heterochromatin. Nat Commun, 8(1):283.

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. Methods, 58(3):268–276.

Berger, S. L. (2007). The complex language of chromatin regulation during transcription. Nature, 447(7143):407–412.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol., 28(10):1045–1048.

Bertin, J. (1967). Smiologie graphique : les diagrammes, les rseaux, les cartes. Mouton, Paris.

Bird, A. (2007). Perceptions of epigenetics. Nature, 447(7143):396–398.

Bock, C. and Lengauer, T. (2008). Computational epigenetics. Bioinformatics, 24(1):1–10.

Boley, N., Wan, K. H., Bickel, P. J., and Celniker, S. E. (2014). Navigating and mining modENCODE data. Methods, 68(1):38–47.

Borg, I. and Groenen, P. (2003). Modern multidimensional scaling: Theory and applications. Journal of Educational Measurement, 40(3):277–280.

Borg, I., Groenen, P. J., and Mair, P. (2012). Applied Multidimensional Scaling. Springer Publishing Company, Incorporated.

Brown, J., Pirrung, M., and McCue, L. A. (2017). FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics.

Canzar, S. and Salzberg, S. L. (2017). Short Read Mapping: An Algorithmic Tour. Proc IEEE Inst Electr Electron Eng, 105(3):436–458.

Cao, X. and Zhong, S. (2013). Enabling interspecies epigenomic comparison with CEpBrowser. Bioinformatics, 29(9):1223–1225.

Carroll, T. S., Liang, Z., Salama, R., Stark, R., and de Santiago, I. (2014). Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Front Genet, 5:75.

Celniker, S. E., Dillon, L. A., Gerstein, M. B., Gunsalus, K. C., et al. (2009). Unlocking the secrets of the genome. Nature, 459(7249):927–930.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation, 19(90):297–301.

Cox, T. and Cox, M. (1994). Multidimensional scaling. Chapman and Hall, London.

Cox, T. F. and Cox, M. (2000). Multidimensional Scaling, Second Edition. Chapman and Hall/CRC, 2 edition.

Cremer, T. and Cremer, M. (2010). Chromosome territories. Cold Spring Harb Perspect Biol, 2(3):a003889.

Cuartero, S., Fresan, U., Reina, O., Planet, E., and Espinas, M. L. (2014). Ibf1 and Ibf2 are novel CP190-interacting proteins required for insulator function. EMBO J., 33(6):637–647.

Cuartero i Betriu, S. (2014). Identificacio i caracteritzacio de nous factors associats als elements insulator de Drosophila melanogaster. PhD thesis, Universitat de Barcelona.

D., A., L., C. M., T., J., and D., R. (2015). Epigenetics, 2nd Edn. . Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press., 2 edition.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet., 14(6):390–403.

Diaz, A., Nellore, A., and Song, J. S. (2012). CHANCE: comprehensive software for quality control and validation of ChIP-seq data. Genome Biol., 13(10):R98.

Dixon, J. R., Gorkin, D. U., and Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. Mol. Cell, 62(5):668–680.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414):57–74.

Duta, N., Sonka, M., and Jain, A. K. (1999). Learning shape models from examples using automatic shape clustering and procrustes analysis.

Ebert, A., Lein, S., Schotta, G., and Reuter, G. (2006). Histone modification and the control of heterochromatic gene silencing in Drosophila. Chromosome Res., 14(4):377–392.

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.

Ernst, J. and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol., 33(4):364–376.

Feng, J., Liu, T., and Zhang, Y. (2011). Using MACS to identify peaks from ChIP-Seq data. Curr Protoc Bioinformatics, Chapter 2:Unit 2.14.

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., et al. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell, 143(2):212–224.

Font-Burgada, J., Reina, O., Rossell, D., and Azorin, F. (2014). chroGPS, a global chromatin positioning system for the functional analysis and visualization of the epigenome. Nucleic Acids Res., 42(4):2126–2137.

Fraley, C. and Raftery, A. E. (2006). Mclust version 3 for r: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics.

Friendly, M. (2008). The golden age of statistical graphics. Statist. Sci., 23(4):502–535.

Gaszner, M. and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nat. Rev. Genet., 7(9):703–713.

Gel, B., Diez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., and Malinverni, R. (2016). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics, 32(2):289–291.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology, 5:R80.

Greil, F., de Wit, E., Bussemaker, H. J., and van Steensel, B. (2007). HP1 controls genomic targeting of four novel heterochromatin proteins in Drosophila. EMBO J., 26(3):741–751.

Hardy, P. A., Zacharias, H., and Flemming, W. (2009). Walther Flemming on histology in medicine 1878: a newly discovered letter to his father. Ann. Anat., 191(2):171–185.

He, Y. and Wang, T. (2017). EpiCompare: an online tool to define and explore genomic regions with tissue or cell type-specific epigenomic features. Bioinformatics, 33(20):3268–3275.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., Glass, C. K., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell, 38(4):576–589.

Herr, J. (2014). Data visualization principles. In The 2nd EMBO Conference on Visualizing Biological Data (VIZBI 2014) - March 2014. EMBL-Heidelberg. Germany.

Jara, A., Hanson, T. E., Quintana, F. A., Muller, P., and Rosner, G. L. (2011). DPpackage: Bayesian Non- and Semi-parametric Modelling in R. J Stat Softw, 40(5):1–30.

Jones, B. (2012). Gene regulation: Transcription factor clutch control. Nat. Rev. Genet., 13(6):380.

Juan, D., Perner, J., Carrillo de Santa Pau, E., Marsili, S., Ochoa, D., Chung, H. R., Vingron, M., Rico, D., and Valencia, A. (2016). Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs. Cell Rep, 14(5):1246–1257.

Karnik, R. and Meissner, A. (2013). Browsing (Epi)genomes: a guide to data resources and epigenome browsers for stem cell researchers. Cell Stem Cell, 13(1):14–21.

Kaustov, L., Ouyang, H., Amaya, M., Lemak, A., Nady, N., Duan, S., Wasney, G. A., Li, Z., Vedadi, M., Schapira, M., Min, J., and Arrowsmith, C. H. (2011). Recognition and specificity determinants of the human cbx chromodomains. J. Biol. Chem., 286(1):521–529.

Kendall, D. G. (1989). A survey of the statistical theory of shape. Statistical Science, 4(2):87–99.

Kessler, R. (2014). Molecular and functional characterization of the HP1c complex in Drosophila melanogaster. PhD thesis, Universitat Pompeu Fabra.

Kessler, R., Tisserand, J., Font-Burgada, J., Reina, O., Coch, L., Attolini, C. S., Garcia-Bassets, I., and Azorin, F. (2015). dDsk2 regulates H2Bub1 and RNA polymerase II pausing at dHP1c complex target genes. Nat Commun, 6:7049.

Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M. Y., Luquette, L. J., Xi, R., Jung, Y. L., Park, R. W., Bishop, E. P., Canfield, T. K., Sandstrom, R., Thurman, R. E., MacAlpine, D. M., Stamatoyannopoulos, J. A., Kellis, M., Elgin, S. C., Kuroda, M. I., Pirrotta, V., Karpen, G. H., and Park, P. J. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature, 471(7339):480–485.

Kidder, B. L., Hu, G., and Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. Nat. Immunol., 12(10):918–922.

Klingenberg, C. P. (2015). Analyzing fluctuating asymmetry with geometric morphometrics: Concepts, methods, and applications. Symmetry, 7(2):843–934.

Kruskal, J. (1964a). Multidimensional scaling by optimizing a goodness of fit to a nonmetric hypothesis. Psychometrika, 29:1–27.

Kruskal, J. (1964b). Nonmetric multidimensional scaling: a numerical method. Psychometrika, 29:115–129.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res., 22(9):1813–1831.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics, Chapter 11:Unit 11.7.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9(4):357–359.

Li, G. and Reinberg, D. (2011). Chromatin higher-order structures and gene regulation. Curr. Opin. Genet. Dev., 21(2):175–186.

Li, J. J., Huang, H., Bickel, P. J., and Brenner, S. E. (2014). Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. Genome Res., 24(7):1086–1101.

Li, W. V., Razaee, Z. S., and Li, J. J. (2016). Epigenome overlap measure (EPOM) for comparing tissue/cell types based on chromatin states. BMC Genomics, 17 Suppl 1:10.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. DMBnet Journal, 17(1):10–12.

Martinez-Balbas, M. A., Tsukiyama, T., Gdula, D., and Wu, C. (1998). Drosophila NURF-55, a WD repeat protein involved in histone metabolism. Proc. Natl. Acad. Sci. U.S.A., 95(1):132–137.

Marx, V. (2015). Visualizing epigenomic data. Nat. Methods, 12(6):499–502.

Matharu, N. and Ahituv, N. (2015). Minor Loops in Major Folds: Enhancer-Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. PLoS Genet., 11(12):e1005640.

Maze, I., Noh, K. M., Soshnev, A. A., and Allis, C. D. (2014). Every amino acid matters: essential contributions of histone variants to mammalian development and disease. Nat. Rev. Genet., 15(4):259–271.

Mikula, M., Skrzypczak, M., Goryca, K., Paczkowska, K., Ledwon, J. K., Statkiewicz, M., Kulecka, M., Grzelak, M., Dabrowska, M., Kuklinska, U., Karczmarski, J., Rumienczyk, I., Jastrzebski, K., Miaczynska, M., Ginalski, K., Bomsztyk, K., and Ostrowski, J. (2016). Genome-wide co-localization of active EGFR and downstream ERK pathway kinases mirrors mitogen-inducible RNA polymerase 2 genomic occupancy. Nucleic Acids Res., 44(21):10150–10164.

Nakato, R. and Shirahige, K. (2017). Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief. Bioinformatics, 18(2):279–290.

Osipyan, H., Krulis, M., and Marchand-Maillet, S. (2015). A survey of cuda-based multidimensional scaling on gpu architecture. In Schulz, C. and Liew, D., editors, ICCSW, volume 49 of OASICS, pages 37–45. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik.

Patel, R. K. and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS ONE, 7(2):e30619.

Pawliczek, P. and Dzwinel, W. (2009). Parallel implementation of multidimensional scaling algorithm based on particle dynamics. In Wyrzykowski, R., Dongarra, J., Karczewski, K., and Wasniewski, J., editors, PPAM (1), volume 6067 of Lecture Notes in Computer Science, pages 312–321. Springer.

Planet, E., Attolini, C. S., Reina, O., Flores, O., and Rossell, D. (2012). htSeqTools: high-throughput sequencing quality control, processing and visualization in R. Bioinformatics, 28(4):589–590.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reina, O., Azorin, F., and Stephan-Otto, C. (2017). chroGPS2, differential analysis of epigenome maps in R. Unpublished manuscript.

Riddle, N. C., Minoda, A., Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Tolstorukov, M. Y., Gorchakov, A. A., Jaffe, J. D., Kennedy, C., Linder-Basso, D., Peach, S. E., Shanower, G., Zheng, H., Kuroda, M. I., Pirrotta, V., Park, P. J., Elgin, S. C., and Karpen, G. H. (2011). Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res., 21(2):147–163.

Robinson, M. D. and Pelizzola, M. (2015). Computational epigenomics: challenges and opportunities. Front Genet, 6:88.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., Ali, S., Chin, S. F., Palmieri, C., Caldas, C., Carroll, J. S., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature, 481(7381):389–393.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20(Supplement C):53 – 65.

Saksouk, N., Simboeck, E., and Dejardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. Epigenetics Chromatin, 8:3.

Salozhin, S. V., Prokhorchuk, E. B., and Georgiev, G. P. (2005). Methylation of DNA–one of the major epigenetic markers. Biochemistry Mosc., 70(5):525–532.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., and Cavalli, G. (2007). Genome regulation by polycomb and trithorax proteins. Cell, 128(4):735–745.

Sibson, R. (1972). Order invariant methods for data analysis. Journal of the Royal Statistical Society B, 34:311–349.

Spitz, F. and Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet., 13(9):613–626.

Starmer, J. and Magnuson, T. (2016). Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains. BMC Bioinformatics, 17:144.

Strickert, M., Teichmann, S., Sreenivasulu, N., and Seiffert, U. (2005). High-Throughput Multi-dimensional Scaling (HiT-MDS) for cDNA-Array Expression Data, volume 3696.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A., 102(43):15545–15550.

Suthar, N., jeet Rajput, I., and kumar Gupta, V. (2013). A technical survey on dbscan clustering algorithm.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics, 31(12):2032–2034.

Teng, M. and Irizarry, R. A. (2017). Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data. Genome Res.

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinformatics, 14(2):178–192.

Tie, F., Prasad-Sinha, J., Birve, A., Rasmuson-Lestander, A., and Harte, P. J. (2003). A 1-megadalton ESC/E(Z) complex from Drosophila that contains polycomblike and RPD3. Mol. Cell. Biol., 23(9):3352–3362.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. Psychometrika, 17:401–419.

Tsochatzidou, M., Malliarou, M., Papanikolaou, N., Roca, J., and Nikolaou, C. (2017). Genome urbanization: clusters of topologically co-regulated genes delineate functional compartments in the genome of Saccharomyces cerevisiae. Nucleic Acids Res., 45(10):5818–5828.

van Bemmel, J. G., Filion, G. J., Rosado, A., Talhout, W., de Haas, M., van Welsem, T., van Leeuwen, F., and van Steensel, B. (2013). A network model of the molecular organization of chromatin in Drosophila. Mol. Cell, 49(4):759–771.

Van Bortle, K. and Corces, V. G. (2013). The role of chromatin insulators in nuclear architecture and genome function. Curr. Opin. Genet. Dev., 23(2):212–218.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet., 10(4):252–263.

Wang, C., Zhan, X., Liang, L., Abecasis, G. R., and Lin, X. (2015). Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. Am. J. Hum. Genet., 96(6):926–937.

Weber, C. M. and Henikoff, S. (2014). Histone variants: dynamic punctuation in transcription. Genes Dev., 28(7):672–682.

Woo, Y. H. and Li, W. H. (2012). Evolutionary conservation of histone modifications in mammals. Mol. Biol. Evol., 29(7):1757–1767.

Xu, S., Grullon, S., Ge, K., and Peng, W. (2014). Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. Methods Mol. Biol., 1150:97–111.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., Flicek, P., et al. (2016). Ensembl 2016. Nucleic Acids Res., 44(D1):D710–716.

Yen, A. and Kellis, M. (2015). Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. Nat Commun, 6:7973.

Zhang, P., Du, J., Sun, B., Dong, X., Xu, G., Zhou, J., Huang, Q., Liu, Q., Hao, Q., and Ding, J. (2006). Structure of human MRG15 chromo domain and its binding to Lys36-methylated histone H3. Nucleic Acids Res., 34(22):6621–6628.

Zhang, Y., Su, J., Yu, D., Wu, Q., and Yan, H. (2013). EpiDiff: entropy-based quantitative identification of differential epigenetic modification regions from epigenomes. Conf Proc IEEE Eng Med Biol Soc, 2013:655–658.

Zhou, X., Li, D., Lowdon, R. F., Costello, J. F., and Wang, T. (2014). methylC Track: visual integration of single-base resolution DNA methylation data on the WashU EpiGenome Browser. Bioinformatics, 30(15):2206–2207.

Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E. A., Koebbe, B. C., Nielsen, C., Hirst, M., Farnham, P., Kuhn, R. M., Zhu, J., Smirnov, I., Kent, W. J., Haussler, D., Madden, P. A., Costello, J. F., and Wang, T. (2011). The Human Epigenome Browser at Washington University. Nat. Methods, 8(12):989–990.

Zhu, L. J., Gazin, C., Lawson, N. D., Pages, H., Lin, S. M., Lapointe, D. S., Green, M. R., et al. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics, 11:237.