Software development and analysis of High Throughput Sequencing data for genomic enhancer prediction

Juan González-Vallinas Rostes

Director: Dr. Eduardo Eyras

Department: Experimental and Health Sciences

TESI DOCTORAL UPF / 2013

Barcelona

A mi abuelo.

Agradecimientos

A mis padres Ramon y Pilar, mi novia Kunmi, sin la cual esto no habría sido posible (y a toda su familia) mis hermanos Rafa y Gema, mi abuela Pilar, mi tios Pedro y Alicia, mis primos Juan y Marta, mis otros tios Pedro y Maria, mis primas Maria, Paloma, Elena y Margarita (y el resto de la familia de Valladolid) por su amor en los buenos momentos y por su apoyo en momentos duros. Especialmente a mi abuelo, que tanto hizo por mi y que no pudo llegar a ver esto.

A la gente de la cuadrilla en Donosti, con los que crecí y viví mis primeras experiencias en este mundo. A pesar de que os veo muy poco ultimamente, sois amigos para toda la vida.

A la cuadrilla de Bilbao, que hizo mi vida universitaria una experiencia fantástica, y me enseñaron a comportarme un poco mejor en el mundo real fuera de la manada de desajustados sociales que normalmente le rodean a uno en una facultad de Ingeniería Informática.

A los amigos de Dublin y Mexico, porque me enseñasteis lo que era el mundo globalizado.

A los amigos del master de Bioinformática, porque meter biólogos, químicos e informáticos en un mismo aula es siempre divertido.

A los viejos amigos Dani, Joni, Manso, Iturri, Focus y Lluvia, y todos los demás Bilbainos que estuvieron una temporada afincados en Barcelona. Desde que os fuisteis mi estancia en Barcelona no ha sido lo mismo.

Al grupito de las comidas en la terraza, que ha sido una de mis pocas fuentes de relax en los últimos meses. Os voy a echar de menos. Gael, Isaac, Babita, Ferran, Jorge, Leone y Endre.

A Nico, Amadis, Eneritz, Colin, Michi y especialmente Steve, por darme la perspectiva que muchas veces me falta.

A Nuria, Mariona y Arnau por dejarme entrar en el laboratorio y enseñarme el otro lado de los experimentos.

A los amigos de Perspicalia, Alfredo, Joni, Diego y Emilio, que ha sido el único en leerse parte de esta tesis y darme feedback (además de Eduardo).

A Christian, Maik, Armand, Juan Ramón (Meneu) y todos los programadores que me ayudaron a mantener la cordura y no sentirme como el

único loco gritando en el desierto del bizarro mundillo de la programación científica. También a Imma, por ser la única matemática en la sala (menos mal).

A Alba, por su inestimable ayuda con la burocracia de la universidad, organización de eventos y comidas en la terraza durante la primera etapa de mi estancia en el PRBB. A toda la gente de su grupo con quien tantos buenos ratos he pasado.

A Grooveshark, Wikimedia Foundation e Internet en general.

A Yoseph y Jordi por acogerme en Filadelfia y ayudarme tanto a nivel personal y profesional. Al grupo de expats de filadelfia, espero veros otra vez algún día.

Especialmente a todos los que no están en estos agradecimientos y deberian estar. En serio, disculpad. Soy despistado, pero no olvido.

Finalmente

Gracias a Eduardo Eyras, por aguantarme todo este tiempo y darme la oportunidad de hacer ciencia.

Preface

While the results of this dissertation were produced during the PhD Program of Biomedicine of Universitat Pompeu Fabra and could be broadly classified as basic biology research, one of the main results has been the development of software for the analysis of High Throughput Sequencing (HTS) data. Because of this, before introducing the work of the thesis I will like to define where it stands in the scientific spectrum, and make some personal observations.

Bioinformatics and *Computational Biology* have evolved from being simply popular terms to fully grown disciplines that name whole University departments. However, up to date, there seems to be a confusion on what exactly *Bioinformatics* is, and the scientific community seems to struggle with the difference between *Computational Biology* and *Bioinformatics*. The greater the confusion, the more popular these terms have become, to the point when the United States National Institute of Health (NIH) stepped in, defining in 2000 "Bioinformatics" as:

> *Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.*

This broad definition officially made *Bioinformatics* an umbrella word to define any department in the life sciences using or producing software as a tool for research. This will include researchers traditionally in the fields of Evolutionary Biology, Genetics, Genomics, Proteomics (and any other -omics field), Systems Biology, Ecology, Molecular Biology...

In my personal opinion, the reason why this confusion persists is because *bioinformatics* may be a bit too broad to describe a field of research. I suspect that the tendency to define *bioinformatics* as a field comes from the considerable difference in expertise required for the usage of these particular tools: experimental laboratory training, or *wet lab*, compared with software usage and development training, or *dry lab*.

While it is obvious that the tools and the expertise required to operate a *wet lab* or a *dry lab* are significantly different, there is no real reason to differentiate researchers that use computers and software from the rest, the same way that it does not make sense to separate researchers that use different animal models to study the same problem. Not only the use of computational and laboratory techniques are not mutually exclusive, they are also

increasingly complementary. For example, in a standard High Throughput Sequencing (HTS) experiment, both experimental and computational tools *and researchers* are needed to successfully complete the standard, basic pipeline that all HTS projects share.

In consequence, I would like to define this thesis as part of the field of genomics as defined by the World Health Organization *the study of genes and their functions, and related techniques*, in the sub-field of Regulation of gene expression. The technologies used and developed during this thesis could still be broadly classified as *bioinformatics techniques*. In addition, some of the techniques fall into the category of *Algorithms* and *Software Engineering*.

Finally, I will like to end this section with a personal opinion. I fear that the artificial segregation of researchers studying the same problems with different tools in *bioinformatics* departments may create an artificial scientific division. I insist that I have no other proof of this than my personal experience in the field, but I worry that *dry lab* research groups are becoming increasingly detached from *wet lab* groups, creating their own body of knowledge through their own ecosystem of journals, and increasingly not listening and reading what the "others" are doing. The ENCODE project (one of the biggest genomics releases in years) and some of its media hyped conclusions have produced some very bitter responses (See for Example Dan Graur's *On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE*), this could very well be a symptom of the tendency I am describing. I believe that this situation is ultimately harmful for the global scientific productivity.

Abstract

High Throughput Sequencing technologies (HTS) are becoming the standard in genomic regulation analysis. During my thesis I developed software for the analysis of HTS data. Through collaborations with other research groups, I specialized in the analysis of ChIP-Seq short mapped reads. For instance, I collaborated in the analysis of the effect of Hog1 stress induced response in Yeast and helped in the design of a multiple promoter-alignment method using ChIP-Seq data, among other collaborations.

Making use of expertise and the software developed during this time, I analyzed ENCODE datasets in order to detect active genomic enhancers. Genomic enhancers are regions in the genome known to regulate transcription levels of close by or distant genes. Mechanism of activation and silencing of enhancers is still poorly understood. Epigenomic elements, like histone modifications and transcription factors play a critical role in enhancer activity. Modeling epigenomic signals, I predicted active and silenced enhancers in two cell lines and studied their effect in splicing and transcription initiation.

Resumen

Las tecnologías *High Throughput Sequencing* (HTS) se están convirtiendo en el método standard de análisis de la regulación genómica. Durante mi tesis, he desarrollado software para el análisis de datos HTS. Mediante la colaboración con otros grupos de investigación, me he especializado en el análisis de datos de ChIP-Seq. Por ejemplo, colaborado en el análisis del efecto de Hog1 en células de levadura afectadas por stress, colaboré en el diseño de un método para el alineamiento múltiple de promotores usando datos de ChIP-Seq, entre otras colaboraciones.

Usando el conocimiento y el software desarrollados durante este tiempo, analicé datos producidos por el proyecto ENCODE para detectar enhancers genómicos activos. Los enhancers son áreas del genoma conocidas por regular la transcripción de genes cercanos y lejanos. Los mecanismos de activación y silenciamiento de enhancers son aún poco entendidos. Elementos epigenómicos, como las modificaciones de histonas y los factores de transcripción juegan un papel crucial en la actividad de enhancers. Construyendo un modelo con estas señales epigenómicas, predije enhancers activos y silenciados en dos lineas celulares y estudié su efecto sobre splicing y sobre la iniciacion de la transcripción.

Contents

Acknowledgements v
Preface ix
Abstract xi

1 Introduction **1**
1.1 The state of the art in Software Engineering for the life sciences 1
1.1.1 The Information Revolution 1
1.1.2 The case of Software Engineering (SE) 2
1.1.3 Code 3
1.1.4 Credit (and open source) 7
1.1.5 Things are not that bad 8
1.2 High Throughput Sequencing (HTS) for genomic regulation 9
1.2.1 DNA Sequencing 9
1.2.2 High Throughput Sequencing 10
1.2.3 HTS for Genomic Regulation 13
1.2.4 Mnase-Seq 14
1.2.5 ChIP-Seq 16
1.2.6 RNA-Seq 21
1.2.7 CLIP-Seq 23
1.3 ENCODE 24

Objectives **26**

2 Pyicos/Pyicoteo **28**

3 Pyicoteolib **80**
3.1 pyicoteolib.core 80
3.2 pyicoteolib.utils 83
3.3 Other interesting remarks 84

4 Regulation of alternative transcription and splicing by intragenic enhancers 86

Discussion 159

Conclusions 162

A Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling 164

B Use of ChIP-Seq data for the design of a multiple promoter-alignment method 166

C Mapping of HITS-CLIP data to the mouse genome 168

Bibliography 171

Chapter 1

Introduction

1.1 The state of the art in Software Engineering for the life sciences

1.1.1 The Information Revolution

Scientific research, as many other professions, has been going through a revolution because of information technologies, in particular data analysis through computational means and online communication.

The Internet now allows for scientific publications to achieve global reach at a fraction of the cost of the printing press. Without factoring in marketing costs, publishing text and static images for a global audience only requires an Internet connection and a publicly accessible server with little storage (less than one Gigabyte, and a few Gigabytes of disk in the worst case scenario). While the costs of maintaining a reliable and accessible server cannot be ignored, this cost is trivial compared with paper printing for a global audience.

The TCP/IP protocol and the modern computer both were designed and developed at first at universities and research institutions. According to Leiner et al. the scientific community was also an early adopter of the new communication technologies[36]

Without the scientific community the development and adoption of these new technologies would have been impossible, but paradoxically, the com-

munity is still heavily dependent on a publication and credit system that uses editorial and distribution restrictions still attributable to the printing press. Big organizational changes are slow. The bigger and better established the organization is, the more challenging it is. This is the current case of Academic Publication.

The resilience of Academic Publication to change is not a new issue. In 1969, during the beginnings of ARPANET (one of the predecessors of The Internet) this was already perceived as an issue by the community. As Leiner writes[36]:

> *The beginnings of the ARPANET and the Internet in the university research community promoted the academic tradition of open publication of ideas and results. However, the normal cycle of traditional Academic Publication was too formal and too slow for the dynamic exchange of ideas essential to creating networks*

The cultural successor of this "open publication tradition" is the so called *Open Science* movement, which is a loose term to define multiple initiatives that aim to improve scientific publication and standards to allow easier access of the public to research and improve the productivity and professional standards of the scientific community.

1.1.2 The case of Software Engineering (SE)

The issue of Academic Publication becomes increasingly detached from the scientific realities and needs when the main result of a research project implies publishing software. This issue has recently been addressed by The Science Code Manifesto (http://sciencecodemanifesto.org/). This manifesto is a document endorsed by close to 1000 researchers worldwide (as of early 2013). According to the original author, it is inspired in both Open Source and the Open Science movements. While I fully endorse the 5 main issues raised by the manifesto (Code, Copyright, Citation, Credit and Curation), for the scope of this dissertation I will focus mainly in Code and Credit, and will make specific observations regarding the particular issue of Software Engineering.

1.1.3 Code

Recent studies have shown that scientists spend typically a significant amount of their time developing software (around 30% or more). However, 90% or more of them are primarily self-taught[15], therefore lack exposure to software engineering common practices.

Software Engineering (SE) is the body of knowledge concerning the design, implementation and maintenance of software, as well as testing and quality standards[27]. This body of knowledge has been developed by Computer Science departments and the private software industry and in most cases (with some notable exceptions) it continues to be ignored by the majority of research groups producing software in the genomics field. The quality of the software published is dismal in many cases (if accessible at all)[65]. Basic testing, quality assurance and usability practices common in the software development industry (to the point that companies, no matter how small, that do not follow them are not considered professional) are systematically ignored. The simple statistics shown in Figure 1.1, published in Nature News[45], point out the systematic lack of basic SE training of scientists.

There is a variety of different and important issues regarding code standards, I will try to briefly cover the main ones.

Object Oriented Programming

Object-oriented programming (OOP) is a programming paradigm that allows for greater abstraction and modularity of code. The main concepts are *classes*, which could be seen as "templates" or "blueprints" of their instances (also known as *objects*) that are loaded into memory and interact with other objects. A class can inherit functionality from another class or from multiple classes, allowing for abstraction. Objects are supposed to abstract the user from what is inside them through their *methods*, allowing for modularity. OOP, while not an indispensable tool for many researchers, its a useful framework for abstraction when programming complex systems. I see OOP as a useful tool for some programming problems, but not like the only way of doing so. The software I developed during this thesis work follows OOP in some of its core functionality.

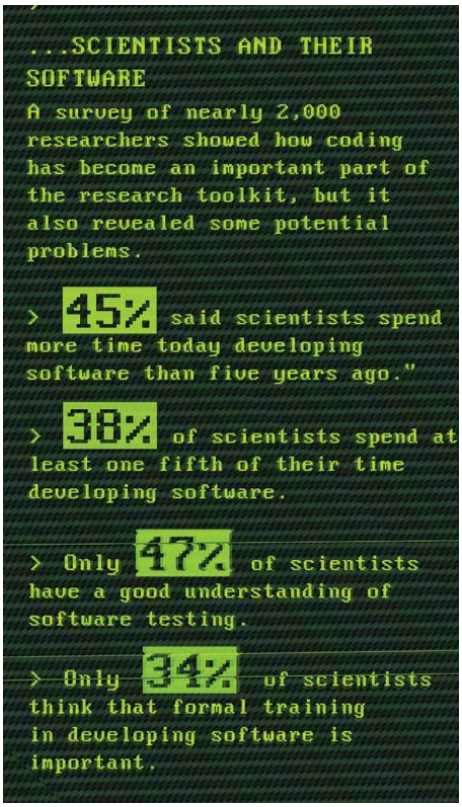


Figure 1.1: Software development and scientists

Version Control Systems (VCS)

Version control systems (VCS) are programs that track and store changes in a file or set of files. Software development using automated VCS can be traced back to the seventies with SCSS in Unix systems[5]. Modern VCS can be both centralized (with a central repository system) and distributed, where every user of the system has a local repository and changes by participants are periodically updated (merged). Common features of modern VCS are:

1. User access management
2. Source changes documentation
3. Delta compression (Storing only *changes* of files)
4. Networking capabilities
5. Efficient and controlled merging and branching

6. Atomic operations (either a change in the repository succeeds completely or gets reverted)

Currently, releasing software under version control is not a requirement for most scientific journals accepting software packages as the main product. On the other hand, it is globally accepted by the software development community (both professional and amateur, close-sourced and open sourced) as an essential tool.

While 1) might not be relevant to small science projects, the rest of the features of these systems are relevant to every software project, including one-person projects. It has been shown that software development productivity increases up to 40% when VCS is used (without involving teamwork)[6].

The reason why small science software projects are so common is also discussed in the Credit (and open source) section.

I want to further emphasize the benefits of VCS, even for one developer projects. *Documenting* what changes had been made to the code, when and why, can be crucial to spot a software error introduced months ago. *Delta compression* allows for significant storage space saving (and help with file order). *Network capabilities* makes easier accessing remote servers, serving both as backup and syncronization between computers (like a workstation and a laptop, for example). *Merging* and *branching* are useful for the single developer for exploration of alternative branches of development, for example like the usage of a library whose performance and reliability cannot be fully determined until the program has been written and tested.

Not only software development benefits from version control. As anecdotal evidence, this thesis has been written using a VCS called git[23]. Furthermore, sharing not only the final computer code to generate a result, but also the *evolution* of such code leads to increased scientific transparency and reproducibility.[53]

End User Software documentation

> *If it's not documented it doesn't exist*
>
> —Mike Pope, programming documentation writer (Microsoft, Amazon)

"Good software with poor documentation" is an oxymoron, since even seemingly simple software (according to the developer) is rarely self explanatory. Good documentation saves time to both users and developers, since is needed not only to help users get their desired results, but also helps developers not spend their time answering frustrated users. If a user fails to understand how to use a piece of software, this software is useless to the user. If this is the case for most users, the software is by definition mostly *unusable*. The reason for a piece of software to be unusable can very well be more technical (errors/bugs, too many resources needed, usability...), but it doesn't really matter if the end result (potential user lost) is the same.

Documentation starts in the code itself. *Literate programming* is an approach to programming proposed by Donald Knuth[32] that encourages writing of code as understandable as possible together with technical documentation. The next layer of documentation is comment documentation. Complex and non self-evident lines of code should be documented with comments in order to make it easier to come back and allow other developers to access them. After this, documenting the functions, classes and modules in a comprehensive way is necessary if the code is building a software library, written to be reused in other programs. Finally, the end user documentation is the documentation that explains to *users* how to use the software, documenting the program, making clear what options there are, what input it requires and what output is expected from it.

Specially in scientific environments, a user of a piece of software can range from someone that clicks one single *Run* button, to a programmer that is using a library of sub-routines for his own scripts. This is why documentation needs to be tailored for the type of end user it is targeting.

The software I have developed during my thesis (Chapters 2 and 3) has both an extensive user documentation and a library documentation. I also tried to follow Dr. Knuth principles of literate programming and comment the code, while I have to admit that I am not as disciplined in this topic as I would like to be.

1.1.4 Credit (and open source)

> *If programmers deserve to be rewarded for creating innovative programs, by the same token they deserve to be punished if they restrict the use of these programs.*
>
> —Richard Stallman, activist and founder of GNU open Source foundation

Why is scientific software quality sub-par? Some possible reasons have already been mentioned: Cultural inertia, lack of standards and training... However, in my opinion the problem will persist until there is a clear and quantifiable **credit** system for producing quality software; and more importantly, give incentives and recognition to researchers **collaborating in** (not starting) open source collaborative science projects.

The open source model for software development, where a group of programmers join in the same software project and develop it together by sharing the code, increases collaboration and is the perfect fit for the publicly funded scientific software production enterprise. And while it is true that most of the researchers publishing free software publish the code too, the collaborative side of the open source movement is missing in most cases, since most projects are undertaken by small research groups (with one single programmer normally)

As an example, let's consider recent developments in the task of mapping short genomic sequences (36-100 nucleotides long) coming from HTS experiments to a reference genome. The software tools that perform this task are commonly called *mappers*. There are currently at least 75 mappers published in peer-reviewed journals[20]. In the case of mappers for the specific task of spliced junctions, the count is around 31[2]. If some of the smaller groups had some kind of incentive to participate in the improvement of already started, open-source mappers projects, these tools would be of greater quality and make the task of selecting which one to use an easier endeavour. The problem is that small research groups need papers for survival, and PhD students and Postdoctoral fellows need first author papers. As a result, collaborating in open source projects is disincentivized by the publishing system itself. This is a big challenge that if overcome, will dramatically improve the quality standards of scientific software.

1.1.5 Things are not that bad

I am sometimes told by my peers that I can be a bit of a pessimist, this is why I will try to make an effort to end this section on a more positive note. It seems to me that slowly but surely, things are changing for the better. For example, current Marie Curie grants by the European Union encourage the publication of code in a repository written for the publication of a paper[19]. All digital journals like PLoS and BMC are rising in popularity and there are new technical journals specialized in this topic, like the "Source Code for Biology and Medicine" journal[58]. Also, open-source collaborative platforms like Bioconductor[22], and Galaxy[24], while they do not explicitly encourage collaboration between programmers and small research groups, are in the right direction of collaborative software development.

1.2 High Throughput Sequencing (HTS) for genomic regulation

1.2.1 DNA Sequencing

> *Scientific research is one of the most exciting and rewarding of occupations. It is like a voyage of discovery into unknown lands, seeking not for new territory but for new knowledge. It should appeal to those with a good sense of adventure*
>
> —Frederick Sanger

Deoxyribonucleic Acid polymers (DNA) were established as the genetic material by Avery et al. in 1944[7]. In 1949, Sanger determined that the **sequential** order of the nucleotides a particular chain of DNA nucleotides in the insulin molecule was critical for its functioning[56]. After these two discoveries, even before knowing about genomic transcription or the DNA double-helix structure, a technique that will allow to determine a sequence of DNA (or in other words, being able to perform **DNA sequencing**) became one exciting problem waiting to be solved.

In 1975, Sanger and Coulston developed the first DNA sequencing technique, which they called *plus and minus*, using a reaction involving *Escherichia Coli* DNA polymerase I[57]. In 1977 Maxam and Gilbert developed another technique (called Maxam-Gilbert) for the sequencing of DNA based on chemical degradation of nucleotides instead of enzymatic activity[44]. The same year, Sanger et al. improved their enzymatic *plus-minus* technique by adding dideoxynucleotide triphosphates as DNA chain terminators. While Maxam-Gilbert chemical sequencing was also used, Sangers *dideoxy* method became the most widely adopted method in the years to come[26].

Automation of DNA sequencing

Originally, the *dideoxy* method was a manual laboratory process that allowed the sequencing of a single chain of approximately 100 DNA nucleotides. Improvements in the technique, like the use of gel electrophoresis allowed for longer sequences (around 400 nucleotides) and greatly increased productivity. This increase of productivity led to data management

problems. As Clyde A. Hutchison III describes[26] working in the seventies on the bacteriophage psi genome:

> *Beginning with psiX, the management and analysis of sequencing data became a major undertaking. The original psiX data was in the notebooks of nine different workers each concerned with particular portions of the molecule. Michael Smith, on sabbatical in the Sanger group, had a brother-in-law named Duncan McCallum who was a business computer programmer in Cambridge. He wrote the first programs to help with the compilation and analysis of DNA sequence data (in COBOL).*

Sequencing automation was first achieved in 1986 by a Caltech group together with Applied Biosystems[61]. After this, followed the expressed sequence tag (EST)[1] approach to gene discovery, that reinforced the tendency for logarithmic growth in sequencing (Figure 1.2)

While automation of sequencing can be seen now as an indispensable step that modern genomics heavily relies on, it comes at a price. Large scale sequencing meant the need of more sophisticated computation and statistical analysis.

The Human Genome Project[28] has been a critical hallmark for the history of sequencing. It was probably one of the most publicised scientific achievements in the last decade, a competition between the public[34] and the private[63] sectors of science, ending up in a disputed tie.

The project, in combination with competition, became again a driving force of innovation in DNA Sequencing. From 1975 to 2001, the sequencing capacity increased from a few hundred nucleotides up to the more that 4 gigabases that the human genome allows. The public awareness and force of innovation in DNA Sequencing, combined with the experience obtained by many scientists and the subsequent interest and funding in the field were critical for the development of High Throughput Sequencing technologies.

1.2.2 High Throughput Sequencing

> *The first rule of any technology used in a business is that automation applied to an efficient operation will magnify the efficiency. The second is that automation applied to an inefficient operation will magnify the inefficiency.*

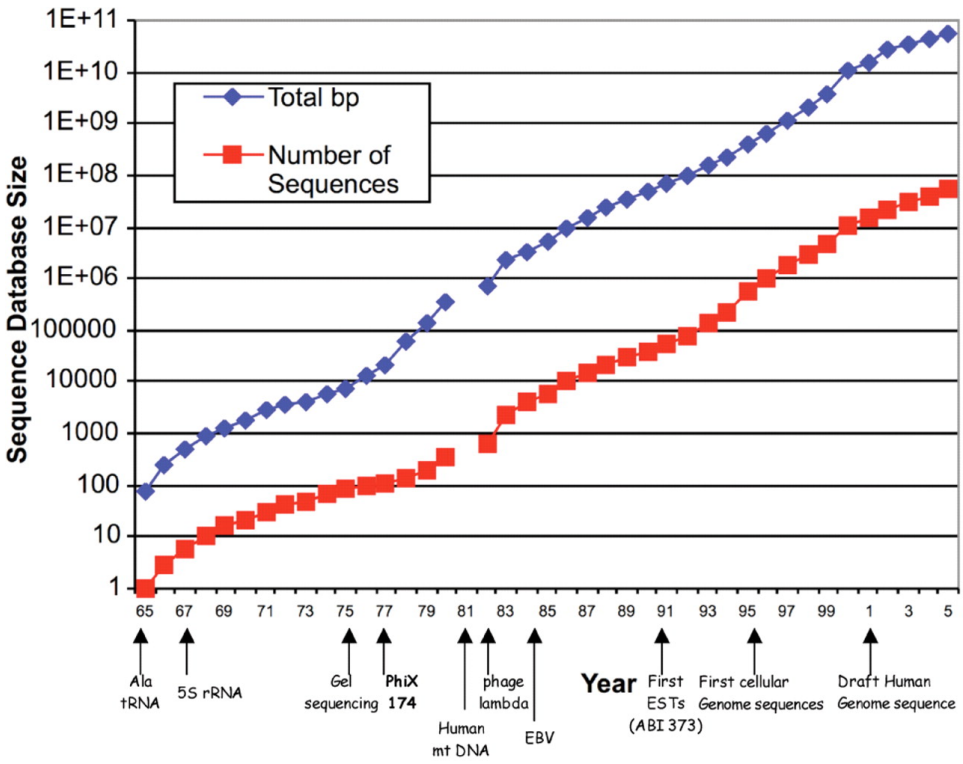


Figure 1.2: Technology and bigger genomes have been the main driving forces of the logarithmic increase of DNA sequencing output/throughput (Adapted from Hutchison 2007)[26]

—Bill Gates, Microsoft founder

The funding and interest on the Human Genome Project accelerated the development of High Throughput Sequencing technologies (HTS), also known as Next Generation Sequencing (NGS) and Deep Sequencing. HTS is a set of technologies that differs from the classic Sanger methods because of its high throughput and sequencing cost reduction[40].

Many HTS platforms based on different approaches have been developed in parallel. For example, *454 Sequencing* is based on *pyrosequencing* technology[55], the *SOLiD* platform (developed by Applied Biosystem) is based on *Sequencing by Oligonucleotide Ligation and Detection* and the *Solexa/Illumina* platforms (developed by Solexa, now owned by Illumina) are based on sequencing by synthesis (SBS) technology[11] (Figure 1.3)

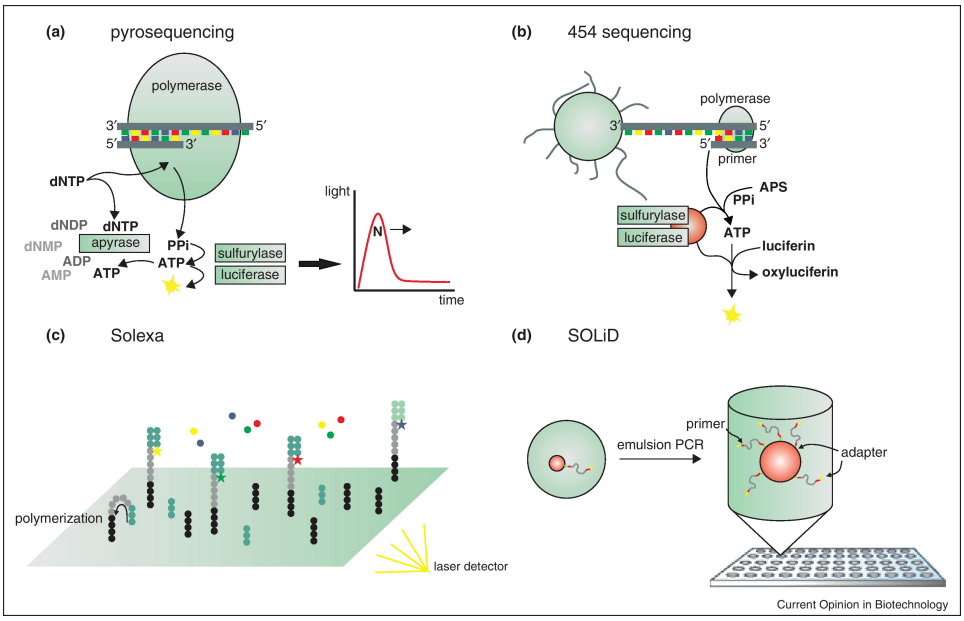


Figure 1.3: Different technologies used by High Throughput sequencing platforms (Figure from Mutz et al. 2013)[47]

However, as it happened with sequencing automation, increasing productivity has some side effects, since the cost reduction comes as a trade-off with read length and quality. For instance, in the case of *Solexa/Illumina* platform, the ones I have used in the projects and collaborations described in this thesis, the average read length started to be 36 nucleotides and only recently reads of aproximately 100 nucleotides can be obtained. HTS reads come with new technical problems that need new computational tools.

Because of its low cost and wide adoption, this new sequencing technology has surpassed its original purpose of DNA sequencing and assembly and is now being used for a greater variety of tasks, like novel diagnostic methods and therapy selection for cancer[13]. However, the focus of my work has been on the analysis of HTS data for genomic regulation, therefore the use of these set of technologies in this particular case is for the detection of RNA transcription and epigenetic elements. In particular, I developed software and analyzed HTS data with the following goals:

- Protein-DNA interaction detection
- Protein-RNA interaction detection
- Differential RNA transcription quantification

From now on, I will focus on the particular use of HTS for genomic regulation. Also, since all my work has been done using Solexa/Illumina data, all that I write from now on should be considered with this in mind. I would like to note though that most of the software and analysis techniques I developed will probably be useful with other similar sequencing platforms.

In the following section I will also introduce the relevant HTS protocols that I used during my work and the biological mechanisms that I studied with them.

1.2.3 HTS for Genomic Regulation

Genome wide analysis of gene expression and transcription factor binding has been dominated since the nineties until very recently by microarray hybridization technologies[59]. Microarray technologies are better established and more mature, and they are by no means obsolete. However, HTS technologies have some advantages over microarrays. In addition to a reduced cost there is no problem with cross-hybridization, there is no need to construct a plasmid library, it has the potential to study genomes at single nucleotide resolution and has possibly better sensitivity. All these provide enough incentive for many research groups and companies to change the microarrays for HTS based techniques[29].

High Throughput Sequencing protocols, even if we limit ourselves to the experiments used for the analysis of genomic regulation, can be very different to each other and used for different tasks. However, at the *in silico* part of the experiment, there are many common shared procedures among them.

Sequencing machines output a list of DNA sequences and quality scores. The usual process after this is mapping to a reference genome or transcriptome. The result of mapping is a list of coordinates corresponding to the putative position in the reference of the sequences. Sequence alignment software like BLAST[49] is not appropriate for the mapping of massive amounts of short sequence reads, accordingly, new mapping tools have been developed in the last few years[20].

With the mapped reads available, depending on the objective of the study, the procedure is different in many cases. However, many of the pipelines and normalization methods are common between technologies (See for example the common steps between ChIP-Seq, FAIRE-Seq and Dnase-Seq in

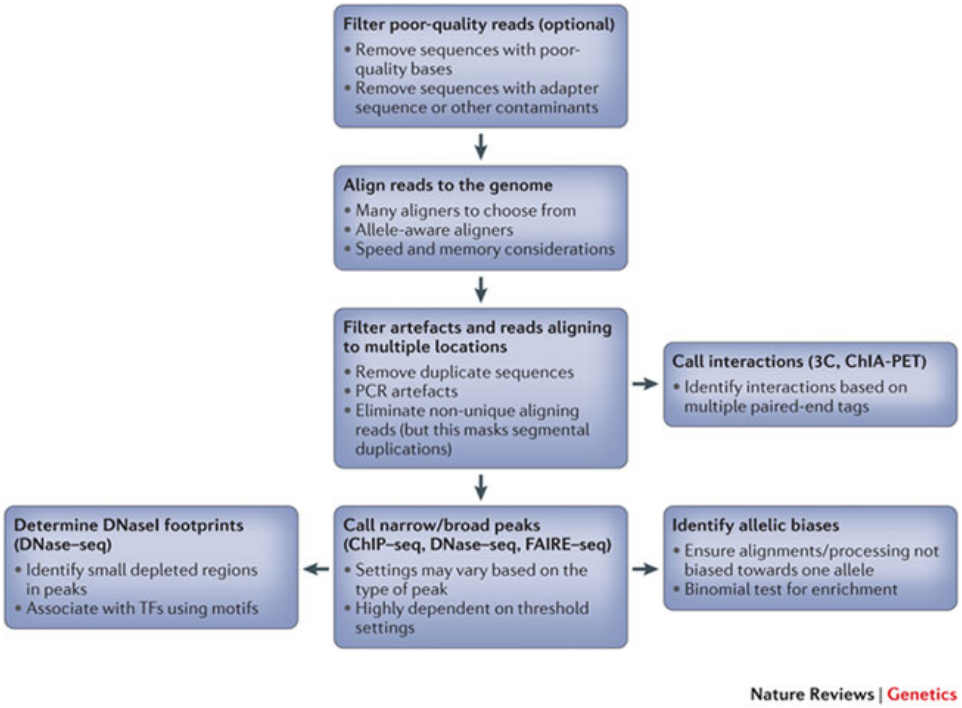


Figure 1.4: Variations of the ChIP-Seq protocol and their different uses (Figure from Furey et. al 2012)[21]

Figure 1.4). This is why I developed software as modular and extensible as possible (Chapter 2 and 3)

Reusability of methods does not only happen between HTS analysis protocols. Many of the normalization methods for microarrays have nowadays been adapted to HTS data analysis. For instance Samr[62] and SamSeq,[37] developed by Tibshirani's lab, Conditional Quantile Normalization[25] by Irizarry's lab to name a few.

1.2.4 Mnase-Seq

Micrococcal nuclease (Mnase) is an enzyme that cleaves and digests DNA until it reaches an obstacle, normally a nucleosome[67]. Digestion of chromatin using Mnase has been used to study chromatin since at least 1974[50]. Mnase-chip was then developed in combination with microarray technology, and more recently, Mnase-seq[69] in combination with HTS technology. Mnase-seq has been used to determine nucleosome positioning genome-

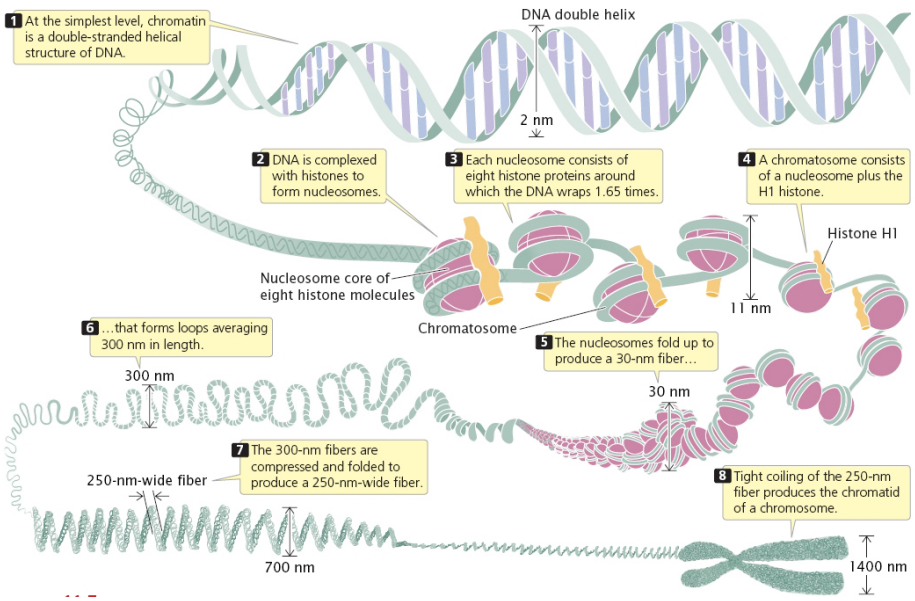


Figure 1.5: Chromatin and its different levels of organization. (Figure from Pierce 2005)[52]

wide.[67] We analyzed Mnase-seq data using our software for determining relative occupancy of nucleosomes in Yeast (see Appendix A).

Mnase-Seq uses: Nucleosome occupancy

The DNA-protein complex in eukaryotic cells that gives structure to the genome and allows for the different levels of structure complexity of the chromosomes is called Chromatin (Figure 1.5). The building units of chromatin are the nucleosomes: 145-147 bases of DNA wrap around each nucleosome protein complex in a left-handed superhelical manner.[41].

The protein core is made of histones. The abundant types of histones are H1, H2A, H2B, H3, and H4. H1 histone is not part of the “core” nucleosome, instead it is part of the linker and participates in the stability of the chromatin structure, while H2A, H2B, H3, and H4 form the core nucleosome octamer. In addition to the common histones, there have been many variants reported, some of which are though to have important regulatory functions, like H2A.Z, a variant of H2A that has been reported to be highly enriched in regulatory elements[10].

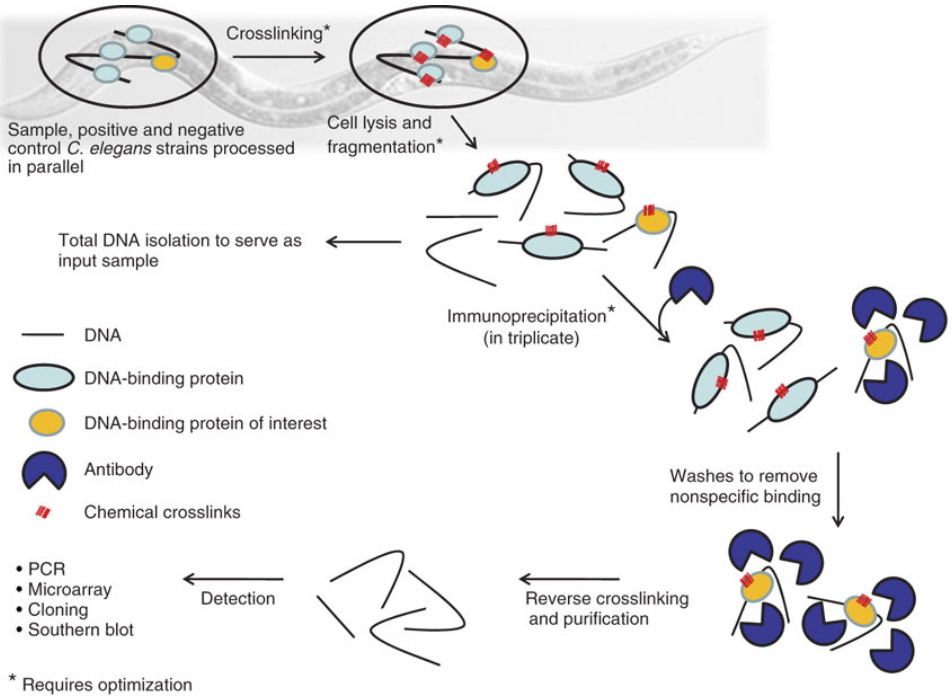


Figure 1.6: Chromatin Immunoprecipitation technique[46]

1.2.5 ChIP-Seq

Chromatin Immunoprecipitation (ChIP)[31] followed by High Throughput Sequencing (ChIP-Seq) allows for the detection of protein-DNA interactions.[29]. The experimental part of the protocol is described in Figure 1.6. It consists in first cross-linking the chromatin, then sonicating the DNA in order to slice it into smaller pieces to then immunoprecipitate the proteins of interest attached to the DNA through a specific antibody. The resulting sequences are then purified and prepared for sequencing using PCR amplification. In order to immunoprecipitate a particular protein using ChIP, a specific antibody that exclusively targets the protein is needed.

Some of the pioneer ChIP-Seq analysis[54;14] assumed a uniform read distribution when determining if read enriched sites where significant or not. There are now several documented biases of the technology, including *mappability*, *GC content bias*, *chromatin accessibility* (open chromatin is easier to shear[30]) and *copy number variation*[38], which account for non-uniform read distributions.

The mappability problem is a side effect of the short reads, because small DNA sequences with low complexity are more likely to be repeated in the genome, and in consequence regions of the genome with low complexity will get less or even no reads. Copy number variation is another problem happening because of short reads, since repeated regions of the genome larger than the read length cannot be uniquely mapped. The GC content bias is complex to correct, because it is sample-specific. In one sample, regions of the genome with many reads may have a higher (or maybe lower) content of GC.[51]. Normalization methods aim to correct these biases, but they will never be perfect, since in most cases it is impossible to delete noise without also deleting some signal. There are other two strategies that can be used in order to improve the quality of ChIP-Seq data:

Control experiments

By either immunoprecipitating DNA with a non specific antibody like immunoglobulin G (IgG) antibody (*Control*) or using raw Input DNA without immunoprecipitation (*Input*). As with normalization methods, both Control and Input have limitations: Control experiments output fewer sequences, so in the amplification phase some sequences can be extremely amplified, leading to very large *blocks* of reads. Input experiments, while they output more DNA and do not get as distorted by PCR amplification, they are biased towards nucleosomes.

Multiple replicas

Producing many replicas is probably still time consuming and expensive for most experimental groups. However, it might be the best solution for correcting some of the biases (see also *Discussion*).

Despite of these limitations, ChIP-Seq technology is an increasingly popular method for DNA-protein interaction detection. Since ChIP-Seq allows for the exploration of the whole genome without a library construction, many whole genome transcription factor mappings using ChIP-Seq have been carried out in the recent years. This can be confirmed looking at the data submission rates. Many of the independent genomic studies datasets are submitted to the Gene Expression Omnibus (GEO) database[9], maintained by NCBI. ChIP-seq submission rates have increased dramatically since 2008,

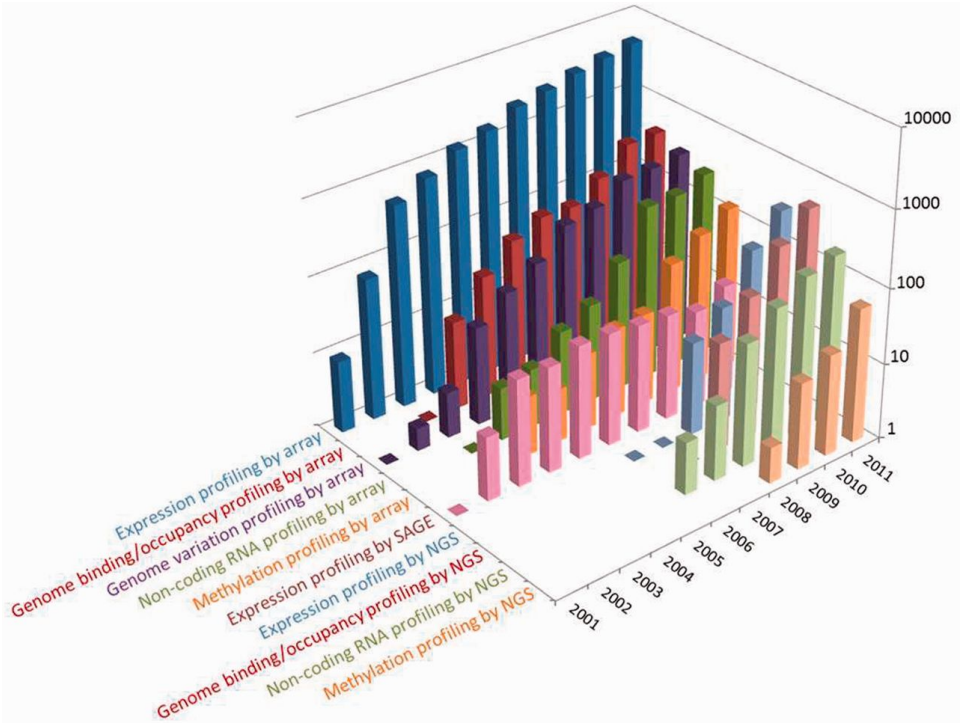


Figure 1.7: Distribution of the number and types of different experimental studies released by GEO each year since inception (Figure from Barrett et al. 2013)[9]

to the point that they are growing at a higher rate than the competing microarray based technology ChIP-chip (as illustrated in Figure 1.7, under the section *Genome binding / occupancy profiling by NGS*)

This shows that ChIP-Seq is a useful, rapidly maturing technology. ChIP-Seq is the HTS data type I have worked with the most. The software I developed at first focused mainly of ChIP-Seq (Chapter 2), but then expanded to other technologies. I have analyzed ChIP-Seq data for most of my scientific collaborations (Appendix A, B) and it is the main type of data I use for the detection of genomic enhancers (Chapter 4)

ChIP-Seq uses: Transcription factor detection

Transcription factors (TFs) are proteins that are of critical importance for genomic transcription regulation. TFs bind on promoter and enhancer regions (transcription factor binding sites, or TFBS) and can either activate

or repress transcription[35]. TFs with relatively short binding sites (<100 bp) can be detected using a process called "peak calling", where a set of statistical and heuristic methods are combined in order to find "peaks" of signal, normally with a Gaussian-like shape (See Chapter 2). Part of the software I developed included a *peak caller* algorithm (See Chapter 2), that I used to evaluate the detection of a multiple promoter-alignment method (See Appendix B).

Another approach to detection of transcription factors is differential enrichment between two conditions. We did this in a collaboration in Yeast cells before and after stress for Hog1 transcription factor and RNAPII, where we observed stress-mediated down-regulation of transcription (See Appendix A), and for the prediction of active enhancers (Chapter 4)

ChIP-Seq uses: Histone Modification detection

Histones are the building blocks of nucleosomes. The first post-translational histone modification was demonstrated by Vincent Allfrey in 1964[3]. In 1997, the nucleosome structure was shown at a 2.8 Angstrom resolution by Luger et. al using X-ray crystallography[41]. This image (Figure 1.8) showed that the N-terminal tails of the histones H2A, H2B, H3 and H4 portrude from the nucleosome, allowing for interaction with other nucleosomes and DNA itself. Since then, the relevance of histone modifications in genomic regulation has been shown to be critical for the manipulation and expression of DNA. There is a long list of reported modifications, the most prominently studied being histone acetylation, methylation and phosphorylation[8].

The two main mechanisms through which histone modifications can affect gene expression are direct structural perturbation and regulation of the binding of transcription factors. For instance, acetylation and phosphorilation are known to reduce the charge of histones, effectively opening chromatin, which facilitates the binding of TFs[8]. A second example involving regulation by binding of chromatin factors can be found in the regulation by enhancers. Tissue specific TFs and acetyltransferase P300 bind to regions of the genome marked with H3K4me1 and H3K4me2, regulating enhancer activity.[68] (See also Chapter 4).

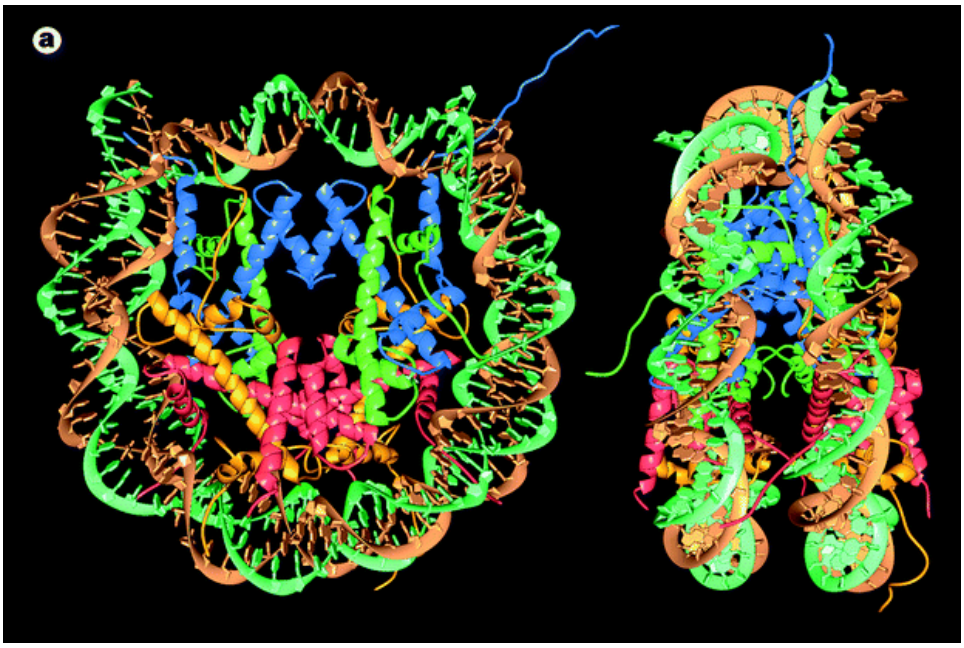


Figure 1.8: Original image published by Luger et. al, showing the structure of DNA wrapped around of histones. The N-terminal tails can be seen as small strings coming out of the structure[41]

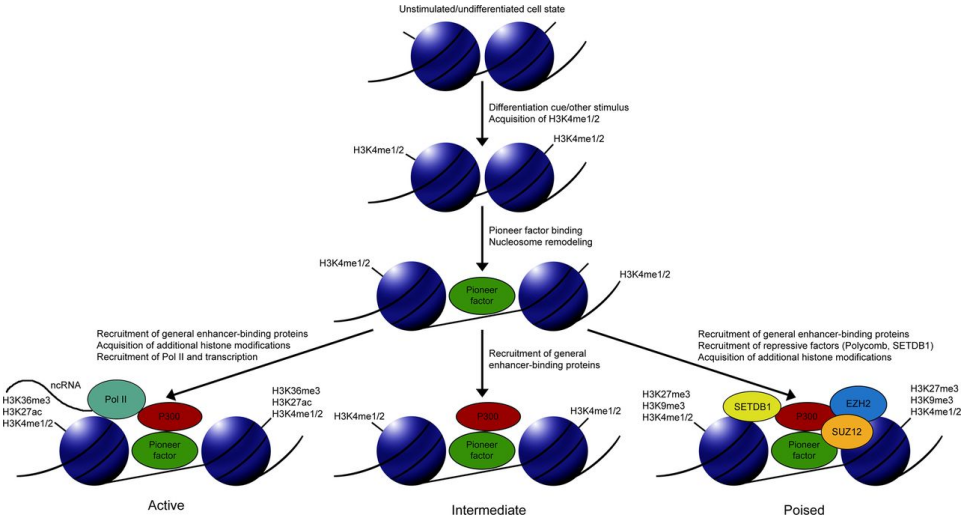


Figure 1.9: A pioneer factor promotes nucleosome remodeling by facilitating nucleosome sliding. This chromatin remodeling facilitates access of general enhancer-binding proteins like P300 and additional factors involved in the establishment of different enhancer states. (Adapted from Zentner and Scacheri)[68]

1.2.6 RNA-Seq

As it was the case of Protein-DNA interaction detection, since the mid-1990s microarray based sequencing technologies dominated the task of RNA quantification. Recently, RNA-Seq was developed as an attractive alternative. In its non-extended form, RNA-Seq uses Illumina/Solexa to sequence millions of fragments of cDNA coming from reverse transcription of RNA[48] (Figure 1.10)

As it happened with ChIP-Seq, at first RNA-Seq technology was expected to have very little to no biases[64]. However, mappability, GC content bias and other biases mentioned in the *ChIP-Seq* section also apply on RNA-Seq. In addition to these technical biases, there is another technical bias on the along mRNA and an added biological complexity coming from multiple transcripts overlapping in the same region by the same or different genes, and different kinds of sense and antisense RNA. Despite these challenges, RNA-Seq is widely used in the exploration of RNA transcription.

RNA-Seq uses: Alternative Splicing prediction

Alternative splicing is a process through which a single gene can produce multiple different mRNAs by inclusion or exclusion of exons during the splicing process of the pre-mRNA. It is an important mechanism believed to be key to higher eukaryotes transcript diversity and is regulated by splicing factors, interacting with the pre-mRNA[33].

There is increasing evidence that splicing and transcription are not, as initially believed, independent processes. They seem to happen predominantly together, allowing for interaction between the transcriptional and splicing machineries[16]. Furthermore, there is evidence that there is an influence of nucleosome positioning and histone modifications through and effect of RNAPII elongation and partly through a direct interaction of splicing factors with chromatin, and in alternative splicing[4].

Using RNA-Seq data, we measured the levels of differential alternative splicing between two cell lines in order to compare them with our active enhancer predictions. (See Chapter 4).

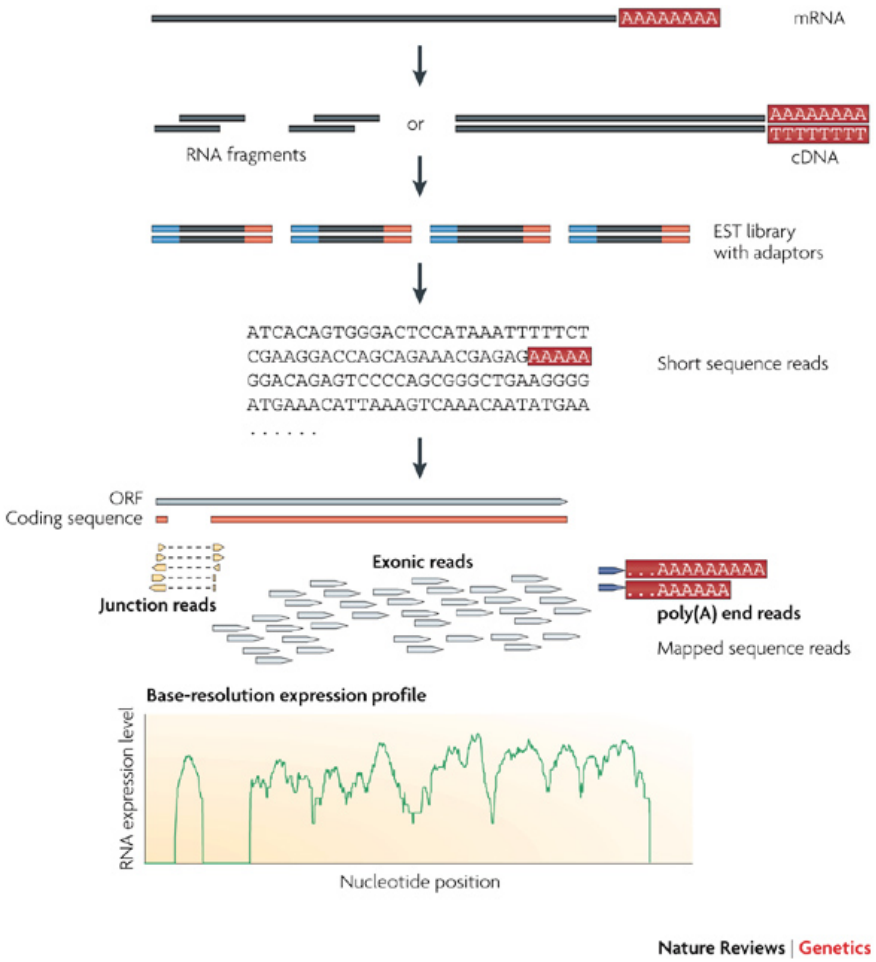


Figure 1.10: RNA sequences converted into a library of cDNA fragments. Sequencing adaptors are added to each cDNA fragment sequenced using HTS sequencing technology. (Adapted from Wang et al. 2009)[64]

RNA-Seq uses: RNA quantification

Comparing the relative number of normalized counts between two conditions in a particular genomic region, one can determine up and down regulation of RNA from a RNA-seq experiment. Using our method for relative enrichment (Chapter 2), I measured the regulation of RNA expression in intergenic regions of the genome in order to determine whether our putative active genomic enhancers produce enhancer RNAs (eRNAs). (Chapter 4)

1.2.7 CLIP-Seq

Cross-linking and immunoprecipitation a protein bound to RNA (CLIP-Seq)[39] has been used for the analysis of Protein-RNA interactions. Due to the level of noise produced by this experiment and the lack of a reliable control (Inmunoprecipitating with a non specific antibody outputs almost no reads[66]), I implemented a bootstrapping method that models a technical background on genes called *modified FDR*[66] for the analysis of CLIP-Seq data. (Chapter 2) and used it in order to detect the interaction between RNA and the RNA binding protein G3BP.

1.3 ENCODE

From the release of the ENCODE pilot project in 2007[12] to their final, media amplified release in 2012[17], ENCODE has become the most ambitious repository of exploratory genomic datasets to date. The ENCODE project is a global collaborative effort started in 2003 by the National Human Genome Research Institute (NHGRI), the same institution that was in charge of launching the Human Genome Project 1989[12], and is therefore considered the direct heir of the human genome project, together with the 1000 genomes project datasets.

In its last release, 1,640 datasets where provided involving 147 different human cell types. All datasets follow the same quality standards, both at the experimental and computational aspect of the experiments. The release has been done for both human (referenced to human assemblies hg18 and hg19) and mice (assembly mm9). The amount of data is vast, diverse and therefore complex to represent in a single table (Figure 1.11) To solve this problem, the ENCODE consortium has produced interactive web tools like the RNA dashboard (http://genome.crg.es/encode_RNA_dashboard)[42] and the ChIP-Seq experiment matrix (http://encodeproject.org/ENCODE/dataMatrix/)[18] to explore, visualize and download the data released.

The ENCODE project is still ongoing and is now it its third phase, aiming to produce data for all genomic and epigenomic datasets as well as for all protein-RNA interactions for all cell lines listed, dividing the cell lines by priority given in three Tiers: Tier 1, Tier 2 and Tier 3. I analyzed data for multiple histone modification and transcription factor datasets for two of the main cell lines in the ENCODE project (Tier 1), the leukemia line *K562* and the lymphoblastoid cell line *GM12878* (Chapter 4).

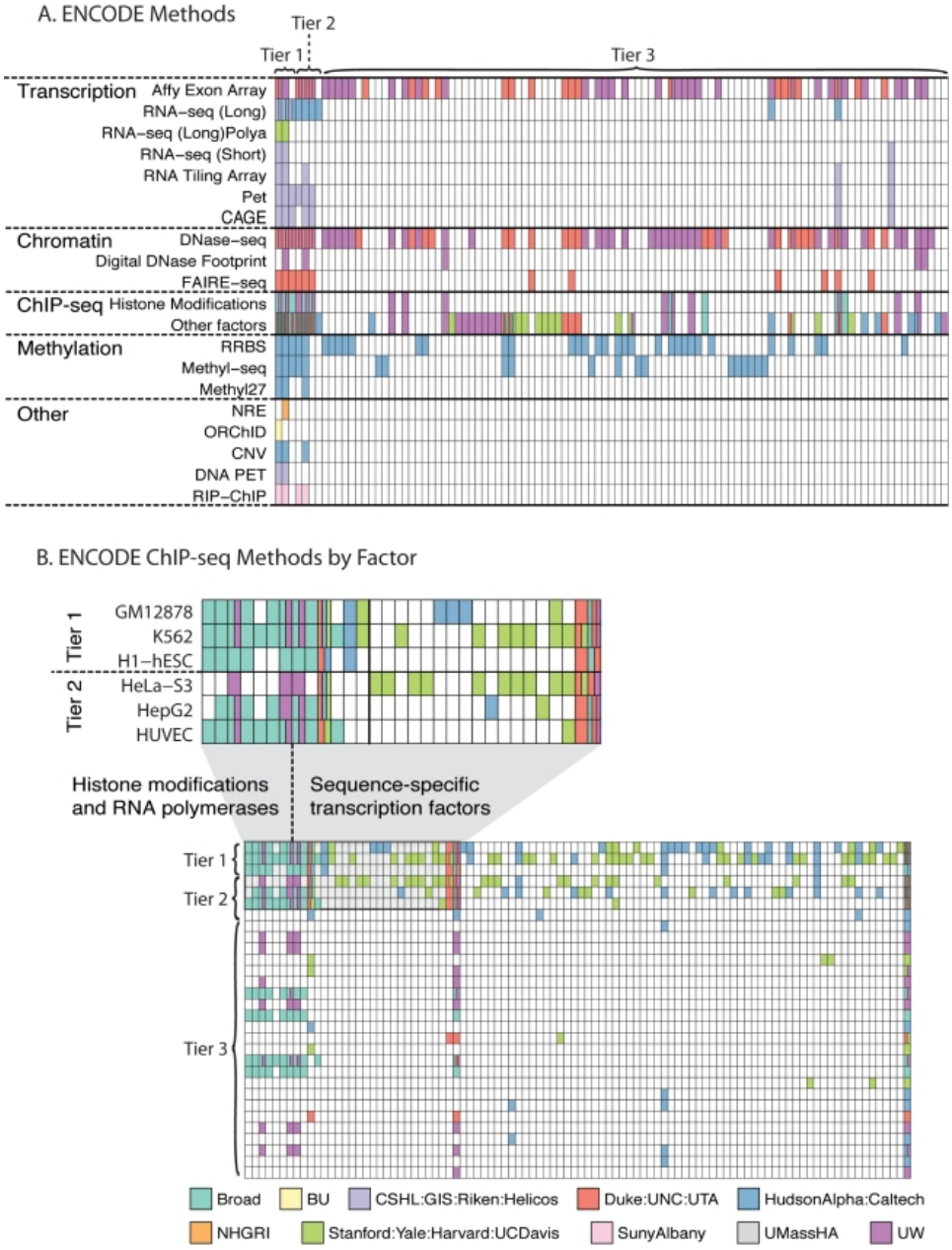


Figure 1.11: A data matrix representing all ENCODE data types. Each row is an experimental technique and each column is a cell line. Colors represent the produced datasets as of 2010. B) Data generated by ChIP-Seq are split into a second matrix where each row represents a cell type and each column represents the factor or histone modification to which the antibody is raised. (Adapted from *A user's guide to the encyclopedia of DNA elements*, ENCODE Consortium, 2010)

Objectives

This thesis work had 2 stages: The first stage was the development of software and analysis pipelines for HTS datasets. The second stage was the usage of this software (and the HTS analysis expertise obtained) for the detection of active genomic enhancers.

Development of HTS software for the analysis of HTS datasets

- The development of genomic coordinates manipulation software.
- The development of a ChIP-Seq peak calling algorithm.
- The development of software for differential enrichment to compare relative changes of HTS signals between conditions.
- The design of pipelines for the automation of HTS data analysis.
- Demonstration of the usefulness of this software in scientific collaborations.

Detection of intragenic active genomic enhancers and study their effect in RNA processing

- Determination of relevant histone modifications and transcription factor in active enhancers using human ENCODE data.
- Using relative enrichment of the selected features to build a model to predict activated and silenced enhancers (relative between two cell lines) in the intergenic space of the genome.
- Verify that our predictions have properties of known active enhancers.
- Use the model for the intergenic enhancers, to predict in the active intragenic enhancers.

- Explore the effect of intragenic active enhancers in transcription and splicing.

Chapter 2

Pyicos/Pyicoteo

This chapter is about Pyicos (now named Pyicoteo suite), a collaborative project in the Regulatory Genomics group. I did the initial design of the software arquitecture, the implementation and the performance benchmarking, while Sonja Althammer did testing, analysis benchmarking, and was the first user of the software. Together with Eduardo Eyras, we both designed the software depending on the analysis needs. Pyicos is a flexible tool for the analysis of HTS mapped reads and can be used for basic manipulation of HTS files, manipulation of genomic coordinates, peak calling and enrichment analysis between two conditions. Cecilia Ballare with Miguel Beato produced some of the experimental data that we used to test our software.

We published 2 articles:

González-Vallinas J*, Althammer S*, Eyras E. Pyicos: A Flexible Tool Library for Analyzing Protein-Nucleotide Interactions with Mapped Reads from Deep Sequencing. Bioinformatics for Personalized Medicine. 2012; 6620: 83-88.

http://link.springer.com/chapter/10.1007/978-3-642-28062-7_9

Althammer S*, González-Vallinas J*, Ballaré C, Beato M, Eyras E. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. Bioinformatics. 2011 Dec 15;27(24):3333–40.

Available from:

http://bioinformatics.oxfordjournals.org/content/27/24/3333.long

The documentation can be found at:

http://regulatorygenomics.upf.edu/pyicoteo

Chapter 3

Pyicoteolib

Pyicoteolib is a library for the analysis of HTS mapped reads. It contains the building blocks of the different tools of what I have named the Pyicoteo suite (a refactoring of the old Pyicos project). In addition of being part of the library for the implementation of the Pyicoteo suite tools, Pyicoteolib is designed to be used in custom python scripts for the analysis of HTS data. In this chapter I will describe the some of the more technical details and interesting modules and classes of the implementation of the library. Pyicoteolib is implemented in the Python programming language and compatible with CPython ¿2.6 and PyPy 2. In the Python convention, logical groups of functions and/or classes are grouped into *modules*

3.1 pyicoteolib.core

Pyicoteolib.core is a module that contains the most basic classes in the library, the holder of reads and regions of reads, the readers of the different formats and the exceptions. The two main and more interesting classes are ReadCluster and ReadRegion. A ReadCluster instance is an object that can contain one read or a group of **overlapping** reads. It can can be initialized to read in multiple formats, as read-like formats, like BED, Sam, BAM, eland and custom text formats by specifying the name, start, end and strand column positions of a plain delimited file and histogram like formats (our own bedpk format and different versions of the UCSC wiggle format).

The object has a format agnostic internal representation of the data, and therefore it is easily extensible to new formats. Instances of the ReadCluster object can be added, compared using common comparison operators ($<$ $>$ $==$ and !=) and subtracted to other ReadCluster objects. ReadCluster also implements common useful python idioms, for example, len() to determine the length of a ReadCluster and str() to obtain a string description, and are iterable with standard Python syntax (for example, in a for loop for length, height in read_cluster will iterate through the different levels of the ReadCluster, outputting the length and the number of). Both the subtraction and addition of ReadCluster objects are optimized to be memory efficient and as fast as possible (See the Supplementary material of the Pyicos Bioinformatics paper for details on the subtract algorithm).

Other useful functions of the ReadCluster object include:

- extend() and push(), to extend and displace the reads in the strand direction respectively
- trim() to trim the borders of the histogram given a threshold
- split() to divide into different sub-clusters of reads based on a ratio parameter between local maxima an minima
- get_max_height() and get_max_height_pos() in order to get the maximum number of reads overlapping in the same position and the genomic coordinate of the height respectively
- overlap() returns a ratio of overlap between two ReadCluster objects overlap.

The ReadCluster object is optimized following the design principles of compression and "lazy loading". When a ReadCluster object is given a list of reads with the same genomic coordinates, it stores them in a compressed list of coordinates. As an example, lets assume we have these three reads (columns correspond to "chromosome", "start" and "end" coordinates respectively)

chr1 1 100
chr1 1 100
chr1 1 100
chr1 50 150

After reading into the ReadCluster, the representation in the object will look like this.

chromosome = ''chr1''

readList begin

occurrences: 3, start: 1, end: 100
occurrences: 1, start: 50, end: 150

end readList

This compression, while simple, is critical for HTS analysis, since duplicated short reads can sometimes reach the millions. Storing millions of reads uncompressed in the same object will mean having performance and memory problems.

The second stage of the optimization of the ReadCluster class is the "lazy loading" of the clustering of the reads. Many useful functions in the ReadCluster object require of a clustering calculation (for example, split()), but some other do not. The reason why this is not calculated at the initialization of the object (lazy loading) is that in cases with many reads, it is computationally intensive to do so, and in some cases it is not really needed. Continuing with the example above, after clustering, the representation of the data in memory will be as follows.

readList = Empty

clusterLevels begin

start: 1, end: 50, number of reads: 1
start: 51, end: 100, number of reads: 4
start: 101, end: 150, number of reads: 3

end clusterLevels

The second basic class is the ReadRegion. An instance of this class can contain multiple ReadCluster classes, and has a higher level functionality, like getting normalized counts of reads in a region, shuffling them for randomization purposes or generating density visualizations of the reads. As for the ReadCluster, it also has a lazy loading system that stops the ReadRegion object from creating all ReadCluster objects on initialization, unless needed.

Some useful functions of this class are:

- shuffle_tags(), that mixes the read positions in the region randomly

- percentage_covered(), that returns the percentage of the region covered by reads
- swap(), a function that given two regions swaps the reads randomly between them taking into consideration a ratio parameter for the probability of landing in one particular region
- normalized_counts(), that returns a normalized count of the reads based on the chosen normalizations (region length and/or total number of reads in the dataset).

3.2 pyicoteolib.utils

In the utils module there are some utility classes and functions used for the parsing, filtering and sorting of big text files. Because we wanted Pyicoteolib to have as little external dependencies as possible, instead of using an statistical python package like scipy (http://www.scipy.org), the module some statistical calculation functions were included (a Pearson correlation and the calculation of Poisson probability). The most interesting classes that I have used in my analysis scripts outside of the Pyicoteo suite and that could be of interest for the community are the file iteration classes, namely SortedFileReader (Figure), SortedFileCountReader (Figure) and DualSortedReader and the sorting class BigSort. All these classes have in common the principle of working with big files without loading them fully into memory.

BigSort is a class designed to sort text files without loading them fully into memory. This is particularly useful for HTS datasets, since they normally occupy several Gigabytes in disk. It accepts as input a sorting pattern (in functional programming mode) matching the selected HTS format. It uses temporary files, which size can be modified through configuration parameters, sorts them and finally merges them into a single, sorted file.

Taking advantage of the fact that after passing a big file through the BigSort algorithm we can be certain that the file is sorted, I implemented the *sorted files* iteration classes SortedFileReader, SortedFileCountReader and DualSortedReader. SortedFileReader holds a cursor position of the file in disk and a file path. Given a ReadRegion, it iterates through the file starting on the cursor position, and will return the overlapping reads in form of ReadCluster objects found in the position that overlap

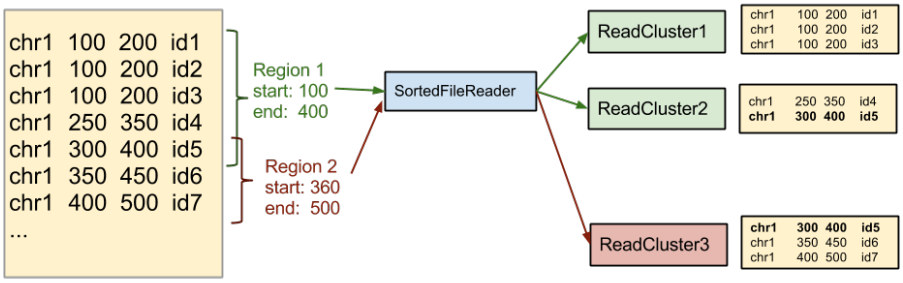


Figure 3.1: Representation of the behaviour of a SortedFileReader instance. The yellow rectangles represent sorted files with genomic coordinates (chromosome, start, end and identifier). The SortedFileReader is fed a list of regions sequentially and yields one ReadCluster object per group of overlapping reads. The read line in bold corresponding to identifier *id5* is shared between two overlapping regions.

with the start and end positions of the region specified (including). The cursor will be left behind the position of the start of the last ReadRegion fed to the SortedFileReader. The behavior of the cursor allows for the iteration through the file using overlapping regions without loading the file into memory and without iterating the file more than once. The SortedFileCountReader class is similar to SortedFileReader, but instead of returning ReadCluster objects, it returns the total number of reads encountered in the region. This again is a performance optimization, since not generating ReadCluster objects makes for a much faster iteration through the file. The DualSortedReader can be fed two sorted files, and will return them in order without loading them entirely into memory.

The functionality of the file iteration classes of the utils module is critical for being able to read big files with overlapping regions without loading them entirely into memory while at the same time not compromising processing time significantly because the file is read only once.

3.3 Other interesting remarks

Because our interest was not to depend on external libraries, I programmed native python implementation for the reading of the binary BAM format. The Pyicoteo suite and Pyicoteolib are not compatible yet with Python 3,

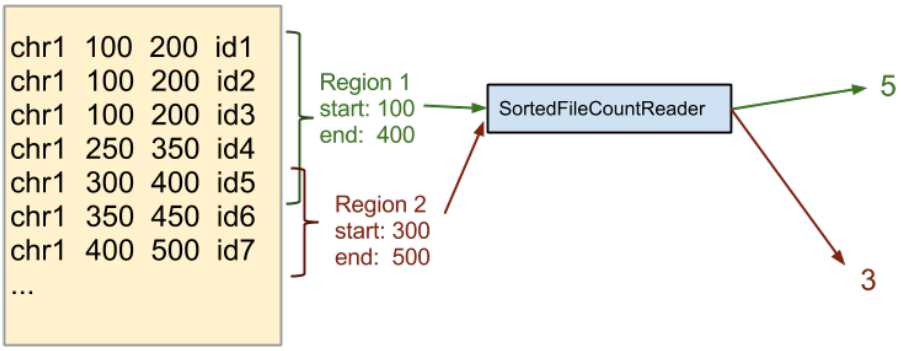


Figure 3.2: The SortedFileCountReader class has a similar behaviour than the SortedFileReader class, with the difference that instead of yielding ReadCluster objects, it yields integers with the number of overlapping reads.

but has been made compatible to run with PyPy. PyPy is a fast, compliant alternative implementation of the Python language (2.7.3) and thanks to its Just-in-Time compiler, Python programs often run faster on PyPy (http://pypy.org/)

Chapter 4

Regulation of alternative transcription and splicing by intragenic enhancers

(At the time of deposit, the manuscript was under preparation)

Evidence for an effect of active intragenic enhancers in the regulation of alternative transcription initiation and splicing

Juan González-Vallinas[1], Eduardo Eyras[1,2]

[1]Universitat Pompeu Fabra, Dr Aiguader 88, E08003 Barcelona, Spain
[2]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, E08010 Barcelona, Spain

Abstract

Background

Alternative splicing (AS) is a key mechanism to generate functional diversity in most eukaryotic cells. The regulation of alternative splicing has been generally thought of being primarily controlled by the activity of splicing factors and by the elongation rate of the RNA polymerase II. Recent evidence has highlighted a complex network of AS regulation that involves interactions between RNA, chromatin and protein factors. Transcriptional enhancers, which are characterized by specific epigenetic patterns, can regulate gene transcription from afar or inside genes. Intragenic enhancer activity modifies locally the chromatin, thereby potentially affecting the processing of the nascent RNA.

Results

Using high-throughput data from ENCODE we have compared the chromatin patterns of the tumoral K562 and normal GM12878 cell-lines to build a predictive model of active enhancers. We use a novel approach based on the relative signal changes between these 2 cell-lines. We therefore exclusively predict enhancers that are active in one cell type but silent in the other. Using this approach, we identified 10,365 intergenic enhancers that are active K562 but silent in GM12878, and 9,777 intergenic enhancers that are silent in K562 but are active in GM12878. We validate our predictions by showing that they have general properties of known enhancers. In particular, active enhancers produce long (>200nt) nuclear RNAs and correlate with DNaseI and RNAPII signals. Using this model, we also predict 11,055 (11,917) intragenic active enhancers in K562 (GM12878) that are silent in the other cell line. We relate the activation and silencing of intragenic enhancers with the expression and splicing changes of

the host genes. We found that intragenic enhancers activate alternative transcription initiation sites. Moreover, active enhancers nearby alternative exons are associated to exon inclusion, whereas enhancers that are silent relative to the other cell line are mainly associated to skipping events.

Conclusions

The activation or silencing of intragenic transcriptional enhancers can modulate the expression of the host genes as well as the splicing regulation of nearby exons, likely by modifying the local state of the chromatin.

Introduction

The activity of the RNA Polymerase II (RNAPII) is highly connected with RNA processing (Neugebauer 2002). Indeed, there is evidence that the majority of pre-mRNA splicing takes place as transcripts are synthesized by RNAPII (see e.g. Carmo-Fonseca and Carvalho 2007). An important consequence of this is that RNAPII activity can affect how splicing occurs (Kornblihtt 2007, Ip et al 2011). Numerous studies have identified exons with splicing variability due to the modulation of the activity of RNAPII (Kornblihtt 2007). This can occur through interactions between the splicing and transcription machineries (de la Mata and Kornblihtt 2006; Loomis et al. 2009) or through mechanisms that affect RNAPII elongation (de la Mata et al. 2003, Ip et al. 2011, Shukla et al. 2011). The chromatin state, characterized by specific histone post-translational modifications, can affect RNAPII elongation and, in turn, splicing (Batsche et al. 2006, Schor et al. 2009, Allo et al. 2009, Saint-Andre et al. 2011, Shukla et al. 2011). Chromatin and splicing can also be related through the recruitment of splicing factors by chromatin adaptor proteins (Sims et al. 2007, Loomis et al. 2009, Gunderson et al. 2009, Luco et al. 2010) or by the modulation of the chromatin state through splicing activity (Lin et al. 2008, Zhou et al. 2011, de Almeida et al. 2011, Kim et al. 2011).

It has been estimated that more than 90% of human genes produce alternatively spliced transcripts (Pan et al. 2008, Wang et al. 2008). However, it is still unknown the extent to which chromatin regulation impacts alternative splicing in the cell. Besides the mechanisms described so far, other processes involving chromatin changes may influence RNAPII activity and splicing, like the spatial organization of the genome (Dixon et al. 2012) or ageing (Han and Brunet 2012). Transcriptional enhancers, which generally regulate transcription of genes from afar, are also characterized by specific chromatin signatures, which may differ depending of whether the enhancer is active or not (Heitzman et al 2007, Heintzman et al. 2009, Creyghton et al. 2010, Rada-Iglesias et al. 2011, Bonn et al. 2012). Interestingly, it was shown before that a transcriptional enhancer activity could modulate RNAPII elongation and thereby affect splicing in a reporter gene (Kadener et al. 2002). Putting these various results together, we hypothesized that the activity of transcriptional enhancers, in particular, when they are at intragenic regions, could affect the alternative splicing of genes. In this article, we study how the activity of transcriptional enhancers can modulate transcription initiation and

alternative splicing by means of the associated changes in the chromatin state.

Transcriptional enhancers have been generally characterized by studying the genome-wide binding of the acetyltransferase P300, an ubiquitous enhancer co-activator (Heintzman et al. 2012, Visel et al. 2009; Blow et al. 2010). However, not all enhancers show P300 activity (see Maston et al. 2012 and references therein). Enhancers have also been characterized by their chromatin state (Boyle et al. 2008, Lupien et al. 2008, Heintzman et al 2007, Heintzman et al. 2009). Although the mono-methylation of histone 3 lysine 4 (H3K4me1) has been identified to be an important signature for enhancers (Heintzman et al. 2007), this mark might not be sufficient for enhancer activation (Creyghton et al. 2010, Pekowska et al. 2011). In fact, recent evidence shows that other marks like H3K27ac (Heintzman et al. 2009, Creyghton et al. 2010, Rada-Iglesias et al. 2011, Bonn et al. 2012) and H3K4me3 (Pekowska et al. 2011, Bonn et al. 2012) may be necessary for enhancer activity. Additionally, the recruitment of RNAPII and the concomitant production of enhancer-associated RNAs (eRNAs) have also been associated to active enhancers (de Santa et al. 2010, Kim et al. 2010, Creyghton et al. 2010, Rada-Iglesias et al. 2011, Bonn et al. 2012). Thus, the activation or silencing of enhancers is associated to specific chromatin signatures.

Although enhancers are typically defined to regulate gene transcription at a distance, about 50% of enhancers predicted by high-throughput methods lie within protein-coding genes (Heintzman et al. 2007) and some overlap exons (Ritter et al. 2012, Birnbaum et al. 2012). Intragenic enhancers can regulate the host gene (Ritter et al. 2012), a different nearby gene (Birnbaum et al. 2012) or can act as alternative promoters (Kowalczyk et al. 2012). These results raise the question of whether intragenic enhancers, upon activation or silencing, by means of the associated local changes of the chromatin state, may affect the alternative splicing of nearby exons, possibly through the modulation of RNAPII elongation.

In this work we investigate whether activation or silencing of enhancers inside genes affect the alternative splicing of nearby exons. Measuring the relative differences in histone marks and activity of various complexes between two cell lines, we build a computational predictive model for active and silent enhancers, using high-throughput epigenomic data from ENCODE (Dunham et al. 2012).

Our predictions show enhancer-like properties according to positional distribution, correlation with gene expression and production of eRNAs. By applying our method to intragenic regions, we predict 10,365 active and 9,777 silent intragenic enhancers.

Results

We started our analysis by building a computational model to detect enhancers that are either active or silent relative to two samples. The model is based on the relative differences in chromatin marks between the two cell lines using a sliding window genome-wide. Considering the length distribution of the experimentally validated VISTA enhancers (Visel et al. 2007) (Supplementary figure 1A), we used 1500bp length windows along the entire human genome, overlapping 500bp. Additionally, we considered only those windows that are located at least 500bp of any annotated gene. This resulted in a total of 3,086,047 overlapping windows. For all these windows, we considered the relative enrichment of a number of histone marks and protein factors, as well as Histone variant H2A.Z and RNAPII (Supplementary Table 1), comparing K562 with GM12878. We assigned to each region a list of attributes, consisting of the 17 enrichment z-scores for the relative enrichment of the ChIP-Seq signals (Methods). Additionally, we considered the z-scores obtained for a Control ChIP-Seq experiment and a DNA Input sample from the same cell lines.

We hypothesized that, since active enhancers are associated with specific histone signals that distinguish them from poised enhancers and from arbitrary genomic regions, the calculated features would group together regions with similar activity. Accordingly, the clustering of regions according to chromatin features would produce different types of regions, among which we should find active enhancers. In order to determine which features are relevant for such classification, we performed feature selection with the Boruta algorithm (Kursa et al. 2010) (Methods). This method finds relevant features by measuring the relevance of attributes with respect to a reference attribute. That is, we considered one signal as a correlation feature against which all other signals are compared. This correlation feature works as proxy for the enhancer activity.

We considered two of the main epigenomic marks related to active enhancers: H3K27ac (Creyghton et al. 2007) (Figure 1B) and H3K4me3 (Pekowska et al. 2011) (Supplementary figure 3A). We found the signals H3K4me1 and H3K4me2, which have been observed to be always present in enhancers, active or not (Pekowska et al. 2011). We also consistently found H2A.Z, which is a histone variant associated to open chromatin and has been observed in promoters and enhancers in association with H3K4 methylation (Ku et al. 2012). We also found P300, which is ubiquitously present in enhancers (Visel et al. 2009). We argue that these signals are consistently associated in the activation of enhancers, and possibly sufficient and general enough to characterize active enhancers in any given cell type.

Interestingly, when P300 or H3K4me1 are used as a correlation feature, the signals H3K27ac and H3K4me3 are not the most significant features (Supplementary Figures 3B and 3C). Additionally, P300 seems to associate with the largest subset of features . This is consistent with experimental evidence showing that P300 associates to enhancers ubiquitously (Wang et al. 2005, Heintzman et al. 2009, Visel et al. 2009) and that enhancers with H3K4me1 and/or P300 occupancy do not always imply activation (Creyghton et al. 2010, Pekowska et al. 2011). In fact, H3K4me1 precedes enhancer-binding factors and P300 may be present in poised and intermediate enhancer states (Zentner et al. 2012). When correlated against the other 2, these 3 features consistently appeared above all technical and biological controls in importance when not used as a correlation class, along with H3k4me1 and H3k4me2 and the histone variant H2A.Z. Although RNAPII and H3K36me3 have been detected before on enhancers (de Santa et al. 2010, Kim et al. 2010), we did not find them as strong predictors of enhancer activity. Interestingly, we did find a strong correlation of the transcription factor PU.1 with H3K27ac but not with H3K4me3. Moreover, although enhancers are modulated by transcription factors in a tissue and developmental state specific manner (Ong et al. 2011), we wanted to find a general description of active enhancers. Accordingly, we did not include PU.1 in the model, and decided to select only the features that scored consistently above the technical and biological controls for both H3K27ac and H3K4me3 feature selection runs, including both of them. Therefore, the features we selected for the statistical model for the prediction of active enhancers are: P300, H3K27ac, H3K9ac, H3k4me1-2-3 and H2A.Z.

In order to validate our feature selection process, we used as correlation class the signal

Control signal (Supplementary Figure 3D) and H4K20me1 (Supplementary Figure 3E). The Control ChIP-Seq experiment with no specific antibody did not correlate significantly with any of the other features. On the other hand, H4K20me1, which has been associated to transcription repression and heterochromatin (Balakrishnan et al. 2010) (Beck et al. 2012) but not to enhancer activity, shows some correlation with EZH2, but no correlation with any other signal.

Intergenic enhancers were predicted using overlapping 1500bp windows that are further away than 500bp from any gene locus (Supplementary Figure 2A) (Methods). We clustered using mclust (Fraley et al. 2007). The model analysis shows that there are mainly three clusters and that considering more clusters does not improve significantly the quality of the clustering (Methods) (Supplementary Figure 4A)**.** We thus considered three enhancer classes: One of these classes, which we call active, is characterized for being enriched in H3K4me3 and H3K27ac (Figure 1B), which is considered to represent enhancers that appear to be activated in K562 cells but were silent in GM12878. A second class, which we call silent, is characterized by a depletion of the same previous marks, which we can interpret as enhancers that are silent in K562 but were active in GM12878. Finally, there is a third cluster of regions with small or no changes in most of the signals, indicating that these regions do not have differential activity between the two cell lines. This class is referred to as no-change. These three groups (active, silent, no-change) define thus the three predictable enhancer classes. With the classification model selected by mclust, we classify genome-wide to predict active and silent intergenic enhancers. This resulted in 66,079 windows active in K562 (silent in GM12878) and 64,436 windows silent in K562 (active in GM12878).

In order to evaluate the accuracy of our predictions, we compared our predicted active and silent windows with the enhancer windows predicted for K562 by ChromHMM (Ernst 2011). The majority of our enhancers predicted as active in K562 or GM12878 overlap with ChromHMM windows labeled as weak or strong enhancers (Supplementary Figure 5A and 5B). Interestingly, a number of our predicted active windows overlap with regions labeled as transcriptionally active. On the other hand, when we compared our active windows with ChromHMM labels in the other cell line, as expected, the majority corresponds to ChromHMM

silent windows (Supplementary Figures 5C and 5D).

We also observed that increasing the threshold on the posterior probability of our predictions increases the agreement with ChromHMM enhancers and decreases the coincidence with other categories (Supplementary Figures 5E and 5F). Interestingly, there is also a slight increase with regions labeled as promoter by ChromHMM. In contrast, when comparing the active enhancers in one cell line with the ChromHMM labels from the other cell line, we find no correlation with the posterior probability (Supplementary Figures 5G and 5H). Based on these comparisons, we decided to keep windows predicted with a posterior probability of > 0.95, which resulted in 36,301 active enhancers and 37,859 silent filtered windows in intergenic regions.

Overlapping windows were clustered into 16,646 active and 16,328 silenced enhancers, respectively, which distribute evenly along the genome (Supplementary Figure 6A). These clusters had mean length of 3,053bp and the majority of them (87.65%) were shorter than 5kb (Supplementary Figure 6B). There were also 273 (1.38%) predictions longer than 10kb, which may correspond to large-scale chromatin domains (Dixon et al. 2012, Han and Brunet 2012) or to clusters of enhancers (or super-enhancers) (Whyte et al. 2013). We filtered out those predictions longer than 5kb, resulting in 10,365 active enhancers and 9,777 silenced enhancers, with mean lengths of 2,704.6bp and 2,588bp, (median lengths of 2,500bp and 2,000bp), respectively . These average lengths are in agreement with previous analyses of enhancers from ChIP-Seq data of histone marks and protein factors (Pekowska et al. 2011, Bonn et al. 2012, Birnbaum 2012).

We next studied the signals that did not show strong correlation with enhancer activity. Interestingly, PU.1 and RNAPII correlate with the predictions, and 25.3% and 20.1% of the active enhancers in K562 show a significant relative difference (left-tailed p-value < 0.01) in PU.1 (Figure 2A) and RNAPII (Supplementary Figure 7A), respectively. In contrast, H3K27me3 shows a weak inverse correlation with enhancer activity and 6.5% of the silent enhancers in K562 show a significant depletion (right-tailed p-value < 0.01) of H3K27me3 (Supplementary Figure 7B). We also studied CTCF and H3K36me3, which, even though they have been detected before on enhancers (de Santa et al. 2010, Kim et al. 2010), we did not

find them as strong predictors of enhancer activity and were not included in our predictive model. For both of them we find a weak correlation with enhancer activity and only 7.4% and 4.6% of active enhancers in K562 show a significant enrichment in CTCF and H3K36me3, respectively (Supplementary Figures 7C and 7D).

To validate our predictions, we used DNaseI signal, which has been used before for the genome-wide detection of enhancers (Boyle et al. 2008, Lupien et al. 2008, Heintzman et al. 2007). We therefore used the DNaseI data from the ENCODE project for the same cell lines to calculate the relative enrichment in predicted enhancer regions. We found a correlation of the relative enrichment of open chromatin with active enhancers (Figure 2B). Moreover, 53.6% and 46.7% of the active and silenced enhancers show a significant enrichment (left-tailed p-value < 0.01) and depletion (right-tailed p-value < 0.01), respectively, in DNaseI signal (Figure 2B).

RNAPII bound enhancers can produce RNAs (Kim et al. 2010, De Santa et al. 2010). These enhancer-associated RNAs (eRNAs) have been described as being of approximately 2kb and non-polyadenylated (polyA-), but polyA+ RNAs have been detected too (Kim et al. 2010; Creyghton et al. 2010, Rada-Iglesias et al. 2011, Kowalczyk et al. 2011). We therefore investigated whether our predicted active enhancers produce RNAs as well. We analyzed the relative enrichment of RNA-Seq datasets from ENCODE (Dunham et al. 2012) for polyA+ and polyA- RNAs, and separately for short (<200nt) and long (>200) and according to their origin, nuclear, cytosolic, and total cell. We found that enhancer activity correlates with the production of polyA+ (Figure 2C) and polyA- (Supplementary Figure 8A) long (>200bp) nuclear RNAs, compared to silent enhancers and to the set of unchanged regions. This increase is much larger than for the other fractions and types of RNAs, for which we did not find such a significant difference (Supplementary Figure 8B). Surprisingly, there seems to be also an increase in cytosolic polyA+ (Supplementary Figure 8C) and this difference almost disappears for cytosolic polyA- RNAs (Supplementary Figure 8D) and short RNAs (<200bp) (Supplementary Figure 8E and 8F). Interestingly, not all enhancers predicted as active seem to generate eRNAs. About 26.4% and 32.1% of the predicted active enhancers in K562 have a significant (left-tailed p-value < 0.01) increase of nuclear polyA+ and polyA-, respectively. In

comparison, we only found a 1.25% of active enhancers with significant (left-tailed p-value < 0.01) increase for short nuclear RNAs. For cytosolic polyA+, 18.7% of the predicted active enhancers in K562 have a significant (left-tailed p-value < 0.01) increase, but in contrast, only 5% of short total RNAs and 9.2% of polyA- cytosolic RNAs show a significant enrichment.

Although enhancers can regulate genes from afar, and one enhancer can regulate multiple genes (Li et al. 2012) they are enriched upstream of genes (Visel et al. 2009). We therefore connected enhancers to genes by choosing for each enhancer the closest gene in either direction. With this approximation, active intergenic enhancers show enrichment at distances close to TSSs compared to random regions (Methods) and to silent enhancers (Figure 2D). Using the same enhancer-TSS pairs, we then calculated the relative change in gene expression measured from RNA-Seq data from ENCODE (Methods). We observe that genes with activated enhancers at a distance between 2kb and 10kb show up-regulation, whereas genes with silenced enhancers in the same distance range show down-regulation (Figure 2E). Moreover, this association is conserved when the distance range of the enhancers is extended to be between 10kb and 100kb from the closest gene (Supplementary Figure 9A). Further support for transcription activity in association to our predicted enhancers was found measuring RNAPII around the TSS: relative density of RNAPII around the TSS in genes close to predicted enhancers correlate with enhancer activity (Supplementary Figure 9B).

Additionally, we searched for some evidence of direct physical interactions for the enhancer-TSS pairs calculated above by using ChIA-PET data for RNAPII (Li et al. 2012). Although only a small fraction of activated enhancers have ChIA-PET links to TSS regions (1.6%), there is an enrichment over silenced enhancers and randomized regions (Supplementary Figure 9C), indicating that predicted active enhancers tend to have more ChIA-PET links than expected by chance, and more than when enhancers are silenced.

Since we are comparing a leukemia cell line (K562) with a normal blood derived cell line, we further investigated whether enhancers active in K562 and not in GM12878 have any association to genes that have been involved in cancer. Using the cancer gene census (Futreal et al. 2004), we found that enhancers predicted to be activated in K562 are enriched

for genes linked to cancer, compared to random regions and to silent enhancers (Figure 2F). In summary, these analyses indicate that our predicted enhancers show properties of active enhancers. We therefore set out to predict intragenic enhancers using the same computational model.

Intragenic enhancers affect the pre-mRNA regulation of the host gene

Active enhancers regulating the expression of nearby genes have been observed in exons (Ritter et al. 2012, Birnbaum et al. 2012) and about 50% of enhancers predicted by high-throughput methods lie within protein-coding genes (Heintzman et al. 2007). Additionally, by comparing the overlap of VISTA elements with the annotation in GENCODE (version 7) (Harrow et al. 2012), we observe that there is no bias for intragenic or intergenic regions (Supplementary Figures 1B and 1C). All these evidences indicate that intragenic enhancers represent an important regulatory component of the genome. Accordingly, we decided to apply our predictive model to localize putative intragenic enhancers that are activated in K562 relative to GM12878, and vice versa.

In order to predict intragenic active enhancers, we considered 1.5kb sliding windows inside genes, starting 500kb downstream of the first TSS and eliminating all windows that overlap with a 1kb region around every annotated alternative TSS (Supplementary Figure 2B). This resulted in an initial set of 2,206,307 possible 1.5kb windows. For these predictions we used the same features as for the intergenic enhancers. Additionally, we used the exact same seed used in the intragenic prediction for the training, a set of 15,000 intergenic regions for the clustering of intragenic regions (Methods), composed of 552 active enhancers, 616 silenced enhancers and 13,832 regions with no significant difference between the cell lines. Using the clustering approach as with the intergenic predictions, we predicted 73,080 active and 92,225 silenced regions. As we did previously with intergenic enhancers, we compared our predicted intragenic predictions with ChromHMM predictions with similar results (Supplementary Figures 11A to 11H). Accordingly, we also kept only windows predicted with posterior probability > 0.95, resulting in 42,297 active and 55,624 silent intragenic enhancer regions. After clustering overlapping regions, we obtained 17,791 active intragenic enhancers in K562

relative to GM12878 and 21,108 active intragenic enhancers in GM12878 relative to K562, falling inside of a total of 5,162 genes (10.11% of all genes) and 5,933 (11.61%) genes, respectively. The mean length of these predictions is 3,665bp, with the majority (82.81%) being shorter than 5kb (Supplementary Figure 10). As before, we kept those shorter than 5kb, resulting in 11,055 activated and 11,917 silenced candidate intragenic enhancers.

Our predicted intragenic enhancers tend to occur in separate genes, with only 29.2% of the genes hosting enhancers of both types. The majority of intragenic activated (78.24%) or silenced (80.61%) enhancers fall in intronic regions, and 26.02% of activated and 22.07% of silenced enhancers overlap an exon. However, comparing the proportion of exonic and intronic regions covered by enhancers with the actual proportions in genic regions, we find no preference for exons or introns (Supplementary Methods). When we looked at the position of the intragenic enhancers, we observed a preference for active and silenced enhancers to fall in the first intron (Supplementary Figures 12A to 12F). However, this effect can be explained by the fact that first introns are on average longer in human (Supplementary Methods) (Bradnam and Korf 2008). We further checked whether genes hosting predicted enhancers tend to show significant differential expression between the two cell lines. Similarly as before for enhancers linked to genes, we find a correlation of the relative expression change of genes hosting active or silenced enhancers. Specifically, 23.8% of 5,162 genes with only active enhancers show a significant expression up-regulation, whereas 34.5% of the 5,933 genes with only silent enhancers show a significant expression down-regulation (Methods).

Intragenic enhancers have been observed to regulate the expression of other genes (Birnbaum et al. 2012). On the other hand, the activation or silencing of an enhancer implies a change in the local chromatin state, which can then affect the transcription activity of the host gene (Ritter et al. 2012, Kowalczyk et al. 2012). We thus hypothesized that this chromatin change would then reflect in an effect at the level of the pre-mRNA regulation, possibly at the level of transcription initiation (Kowalczyk et al. 2012) or at the level of splicing (Luco et al. 2010). We thus first tested whether the activation or silencing of internal enhancers may produce the activation or repression of a downstream internal TSS. We considered all active and silenced enhancers that fall between the most upstream TSS (TSS1) and the first internal

annotated TSS (TSS2), such that the distance TSS1-TSS2 was longer than 20kb. This resulted in a total of 870 TSS1-TSS2 pairs, from which 113 (13%) had at least one active enhancer and 135 (15.52%) had at least one silenced enhancer between both TSSs. When an active enhancer is present between the two alternative TSSs, we observe that generally both TSS1 and TSS2 show an increase in RNAPII density in K562 relative to the other cell line (Figure 3A). This indicates that activation of an intragenic enhancer can affect both TSSs, not only the downstream one. Conversely, when a silent enhancer is present between both TSSs, the relative level of RNAPII tend to decrease at both TSSs relative to the other cell lines (Figure 3B), which corroborates the previous finding for GM12878 cells. Interestingly, this effect persists for other downstream alternative TSS events (Supplementary Figures 13A to 13C), indicating that intragenic enhancers can activate internal TSSs, but also affect transcription of the most upstream TSS to some extent.

As for intergenic enhancers, we used ChIA-PET for RNAPII in K562 to validate a possible direct interaction between our intragenic enhancers and the first TSS of each gene. As obtained before for intergenic enhancers, we observe a higher density of ChIA-PET links for active enhancers than for silent ones. However, in this case there is no clear distinction between the putative enhancers and the intragenic randomized positions (Supplementary Figure 14), probably due to the high RNAPII activity inside genes.

The change in chromatin state induced by the activation or silencing of an enhancer may change the processing of the pre-mRNA. For instance, it is known that intragenic chromatin states can affect RNAPII elongation, which in turn produce changes in alternative splicing (Schor et al. 2009, Allo et al 2009). We therefore hypothesized that intragenic enhancers that are active in a cell line relative to the other one may be associated with differences in the inclusion levels of nearby exons relative to the two same cell lines. Accordingly, we measured for all genes the differential inclusion of exons between K562 and GM12878 using cytosolic RNA-Seq polyA+ data from ENCODE (Methods). Using as cut-off p-value < 0.05 and $\log_2$-fold change > 8 (Supplementary Figure 15), we detected a total of 1,363 and 3,114 exons significantly included and skipped in K562 relative to GM12878. We considered the top (increased inclusion) and bottom (decreased inclusion) 1,363 exons with a significant

inclusion change and with no significant change in expression of the gene. From these two exon sets, 235 (17.1%) and 249 (19.1%) have an active or silent enhancer, respectively, in the region spanning the exon and 3kb upstream and downstream of the exon (Methods). For this subset of exons, we found an association between active enhancers and increase inclusion. From the 162 regulated exons having an active enhancer downstream, 127 exons had increased inclusion and 35 exons decreased inclusion in K562 relative to GM12878 (Figure 4A). Similarly, from the 92 exons with an active enhancer upstream, we found 73 exons with increased inclusion and 19 exons with decreased inclusion in K562 relative to GM12878 (Figure 4B). From these cases, 64% and 48.1% from the first and second group, respectively, show a significant change in gene expression. This pattern is replicated by the enhancers silent in K562 but active in GM12878: from the 129 regulated exons having a silent enhancer downstream, we found 19 with increased inclusion and 110 exons we decreased inclusion in K562 (Figure 4C). Similarly, from the 101 exons having a silent enhancer upstream, we found 16 exons with increased inclusion and 85 exons with decreased inclusion (Figure 4D); and these two groups, 55.0% and 64.4%, respectively, show a significant change in gene expression. As a comparison, 47 and 55 control exons that did not change splicing (p-value < 0.05, absolute $\log_2$-fold < 0.1) had a silenced enhancer upstream and downstream, respectively (Figure 4E).

As an example, we show the case of a regulated exon in the gene LTBP (Figure 4E), where two active enhancers in K562 correlate with three differentially included exons that are highly included in K562 relative to GM12878 ($\log_2$-fold 13.5, 14.17 and 13.90 downstream)

In conclusion, we have found a possible association between the activity of intragenic enhancers and the regulation of the pre-mRNA. In particular, the activation of intragenic enhancers can affect the activity of internal transcription start sites, and the inclusion of nearby exons.

Methods

Datasets. Annotated human enhancers with a homologous enhancer in mouse that had been experimentally validated were downloaded from VISTA (Visel et al. 2007). The gene set was obtained from the 7th release of GENCODE, which is based on the assembly GRCh37 (hg19) and was included in the Ensembl release 62. ChIP-Seq and RNA-Seq datasets were downloaded from ENCODE (Dunham et al. 2012) for K562 and GM12878 cells. The datasets used were: ChIP-Seq for CTCF, EZH2, P300, Pol2, PU.1, STAT1, H3K9ac, H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K36me3, H3K79me2, H4K20me1 and H2A.Z; one Control ChIP-Seq experiment and one input experiment; and .RNA-Seq for short (<200nt) and long (>200nt), polyA+ and polyA- RNAs from whole cell, nucleus and cytosol. All datasets were downloaded in the form of mapped reads to the reference hg19 genome in BAM format.

Relative enrichment calculation. Our model of active enhancers is based on intergenic regions away from gene loci. We used sliding windows of 1,500nt, with a slide shift of 500nt. In order to avoid mixing enhancer signal with genic and promoter signals, we discarded windows that were closer than 500nt to an annotated TSS. Similar approaches were applied to intergenic (Supplementary Figure 2A) and intragenic (Supplementary Figure 2B). Although there are more intergenic windows (~3x10^6 vs ~2.2x10^6) in both cases the amount of windows with signal was similar, ~1.5 million windows, which were then kept for further processing. We used the relative enrichment of epigenetic signals between 2 cell lines in order to predict active enhancers in K562 (relative increase of activation marks in K562 with respect to GM12878) and silent enhancers in K562 (relative decrease of activation marks in GM12878 with respect to K562). Full-quantile normalization for counts and GC content was applied using EDASeq (Risso et al. 2011). GC content in each region was calculated as the proportion of G+C in the 1500nt window. After normalization, the z-score of the relative enrichment of the ChIP-Seq signals between K562 and GM12878 was calculated with Pyicoenrich command (Althammer et al. 2011). This method defines a vector of z-scores per region, each z-score indicating the significance of the relative enrichment for each ChIP-Seq experiment, which we refer to as attributes. A positive z-score in for a region indicates an increased in ChIP-Seq signal in K562 relative to GM12878 in that region, whereas a negative

z-score indicates a decreased signal in K562 relative to GM12878. Z-scores close to zero indicate no significant differences between the cell lines. For all datasets, except for the ChIP-Seq with non-specific antibody and for the RNA-Seq datasets, we used replicates. The relative enrichments were calculated with respect to the distribution described by the comparison between replicates. When replicates were not available, these were simulated by mixing and dividing the two conditions by random sampling.

Feature Selection. Feature selection was performed using Boruta (Kursa et al. 2010), which is a feature selection wrapper algorithm that uses random forests (Liaw et al. 2001) to find relevant features by measuring the relevance of attributes with respect to a reference attribute (or correlation class) in comparison with a random model extracted from the original dataset. Boruta requires a correlation class to perform the feature selection, i.e. it uses one single feature to evaluate the rest against it, using the Random Forest search. We performed this analysis for each of the individual marks: H3K27ac, H3k4me3, etc. In each case, the correlation was performed 10 times using normalized counts on a subset of 5000 intergenic windows, sampled randomly for each one of the 10 iterations. As control signals, in addition to the random technical controls added by Boruta, the ChIP-Seq signal from a non-specific antibody (Dunham et al. 2012) was added in the analysis.

Window clustering. Intergenic windows were clustered according to their z-scores in the selected features using mclust (Fraley et al. 2007). Mclust is based on finite normal mixture modeling and uses the Bayesian Information Criterion (BIC) (Schwarz 1978) for model optimization. As seed for mclust clustering, 15000 randomly selected windows were given as input. Various seed selections of this size did not change the results significantly. Calculating the BIC score for different models and different number of clusters, this plateaus at around 3 clusters for most models (Supplementary Figure 4A). This indicates that there are mostly three mean classes, two that correspond to active and silenced enhancers, and a number of intermediate states are not clearly delimited and conform a gradient of multiple states, which are combined into a single class, also mixed with regions of no chromatin activity. This observation is reinforced by the uncertainty plot, which shows that more “certain” regions are on the extreme values of the correlation of the features (Supplementary Figure 4B). The final model used for clustering was the centroid type VEV, which creates clusters that correlate with Variable volume, Equal shape, and Variable orientation. The algorithm that Mclust uses

for predicting this model is Expectation-Maximization.

Intragenic enhancers. Intragenic enhancers were calculated similarly to intergenic enhancers, but using a set of 15,000 intergenic windows as seeds. These seed windows correspond to 552 active, 616 silenced and 13,832 no-change regions. The clustering was performed using the same model as for the intergenic windows.

Linking enhancers to genes. In order to connect the predicted (active and silent) enhancers with genes, enhancers were linked to the closest TSS on either direction. Random enhancer positions were generated by placing each enhancer in a random location in the same chromosome, avoiding gaps, genic regions, other random locations, and not closer than 500nt from any gene. Random intragenic enhancers were generated similarly, by placing the intragenic enhancers in a random location inside the same gene, avoiding regions of 1kb around any alternative TSS and avoiding other random enhancers. ChIA-PET data for RNAPII in K562 cells from the ENCODE (Dunham et al. 2012) was also used to link predicted enhancers with nearby genes. An enhancer was considered connected to a gene if there were at least 3 ChIA-PET pairs connecting both the predicted enhancer and the region of 1kb around the TSS of the gene. For this calculation, only enhancers that were between 2kb and 100kb from a TSS were considered.

Cancer related genes. For those genes connected to active enhancers, the proportion of them associated to cancer was calculated using the Cancer Gene Census from the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/genetics/CGP/Census/) (Futreal et al 2004).

Alternative transcription and splicing. For every gene in the GENCODE (version 7) annotation, the most upstream TSS (TSS1) and all the alternative TSS positions (TSS2, …) were considered. Each pair TSS1-TSS2, TSS2-TSS3, TSS3-TSS4, …, was considered as an alternative transcription event. RNAPII relative enrichment levels were measured around each TSS using the same normalization and calculation method as before. To control possible association with upstream enhancers, we discarded all alternative TSS events that have a predicted intergenic enhancer (active or silent) 100kb upstream of the gene. Differential inclusion levels were calculate for Gencode exons using DEXSeq (Anders et al. 2012) with

polyA+ cytosolic RNA-Seq.

Discussion

Using a method that measures the relative enrichment and depletion of multiple signals coming from ChIP-Seq experiment, and using a semi-supervised approach, we predict enhancers that are active relative to a different condition or cell type. We also predict the opposite class, which corresponds to enhancers that appear to have been silenced, but are active in the other condition or cell type. We have obtained 21,420 enhancers that are active in K562 but not in GM12878 cells, which we call active, and 21,694 enhancers that are active GM12878 but not in K562, which we call silent enhancers. Our model provides a novel approach, because while comparable with other methods in the objective of using chromatin states to predict regulatory elements, we focus on only exclusively predicting active enhancers and silent enhancers in one condition respect of the other. We do so by first predicting exclusively in intergenic regions (in order to avoid mixed signals coming from transcriptional sites). Also, our method is able to predict relative activation and silencing between two conditions, unlike other methods. However, our method is blind to enhancers active or silent in both conditions.

The estimation of active enhancers in a given cell type is very much dependent on the technique used to detect them (Zentner et al. 2012). Although activation of enhancers is generally associated to a number of histone modifications, which enhancers are active is cell type specific and only a small fraction of the many candidate enhancers previously identified using a variety of techniques (ENCODE Consortium 2012) may be active in a given cell. For instance, Heintzman et al. found 24,566 putative enhancers in K562 cells with about 20% of them overlapping putative enhancers detected in HeLa cells (Heintzman et al. 2009). We predict 21,420 putative enhancers that are active in K562 and that are non-active or poised in GM12878, and 21,694 putative enhancers that are active in GM12878 but they have been repressed in K562. In contrast, ChromHMM (Ernst et al. 2010) predicts more than 60,000 non-abutting genomic regions to be strong enhancers and about three times as much for weak enhancers. There are two main reasons for this discrepancies: the resolution of the

genome segmentation is very different, and we only predict enhancers that are active in one condition, but not in the other.

Our enhancers are H3K27ac dependent and are defined almost entirely by chromatin signals. The features found for our classes confirm that active enhancers are characterized not only by the presence of H3K4me1, but also by the presence of H3K27ac, H3K4me3 and RNAPII (de Santa et al. 2010, Kim et al. 2010, Rada-Iglesias et al. 2011, Bonn et al. 2012). We also observed that active enhancers show an enrichment of the histone variant H2A.Z, which has been identified to demarcate regulatory regions (Jin et al. 2009). Interestingly, we also found a strong correlation of the transcription factor PU.1, and a significant occupancy of PU.1 in 25.3 % of active enhancers. PU.1 has been shown before to be an essential co-factor for enhancer activity (Ghisletti et al. 2010, Heintzman et al. 2009, Robertson et al. 2008), is known to bind to H3K4me1 sites in macrophages and B cells in a cell-specific manner (Heinz et al. 2010, Jin et al. 2011) and has sequence similarity lymphoid-specific enhancer (Uniprot 2013) (See also supplementary methods). Here we find it strongly associated with potentially active enhancers in a leukemia cell line.

We found that CTCF and EZH2 and the histone marks H3K36me3 and H4K20me1 do not seem to play any general role in enhancer activation. H3K27me3 is the only feature that shows a pattern of depletion in active enhancers and of enrichment in silent enhancers, but mainly in long enhancer-like regions, which may be related to other regulatory mechanisms.

We additionally found that predicted enhancer activity correlates with production of enhancer-associated RNAs (eRNAs). Moreover, these are mostly long nuclear RNAs and can be polyA+ as well as polyA-, although we found cytosolic RNAs too. Additionally, we observed that not all active enhancers produce eRNAs. On the other hand, although RNAPII and H3K36me3 have been detected on enhancers as well (de Santa et al. 2010, Kim et al. 2010), in relation to eRNA production, we did not find them as strong predictors of enhancer activity, and only 14.9% of active enhancers have a significant increase in the occupancy of RNAPII.

The model built on intergenic enhancers was applied to predict intragenic enhancers. Since our model allows us to distinguish between activation or silencing, we can obtain associations to relative differences between the same two conditions in expression and splicing. We have shown that intragenic transcriptional enhancers, upon activation or silencing, may affect the activity of alternative transcription start sites. Surprisingly, we observe that intragenic enhancers can affect the most upstream TSS too. This generalizes previous findings indicating that intragenic enhancers can act as internal alternative promoters (Kowalczyk et al. 2012).

We also found that to a small extent, intragenic enhancers, upon activation or silencing, also associate to the regulation of the inclusion of nearby exons. These changes may be mediated by the changes in the RNAPII elongation produced by the chromatin change or by the recruitment of splicing factors, mediated by the chromatin signals. We observed that active enhancers, which show an increase of H3K4me3, H3K27ac and H3K9ac, among others, but not of H3K36me3, are associated to an increase in exon inclusion. Moreover, this seems to occur for enhancers upstream or downstream of the regulated exon. Interestingly, these signals have not been directly associated before to splicing. They are open chromatin marks, so one would assume that they should not represent any roadblock for RNAPII. Accordingly, it would seem that the expected outcome would be the opposite, i.e. reduced inclusion; possibly mediated by a fast RNAPII. Interestingly, splicing changes due to RNAPII elongation changes can be in both directions (Ip et al. 2010); and we find both effects on splicing too, although one is overrepresented with respect to the other.

Our results also indicate that the effect that intragenic enhancers may have in the differences in splicing observed between two cell types may be very limited. Additionally, a considerable proportion (51.1% for genes containing differentially included genes in K562 and 52.9% for differentially excluded genes) of splicing changes occur in genes that change expression. This indicates that the main effect of the activation of enhancers may be related to the activation of alternative transcription in the gene and alternative splicing is byproduct of that.

Our work illustrates the effectiveness of our method in relating relative activity of enhancers to RNA processing mechanisms, like transcription activation and splicing. This presents several advantages over other methods that find direct associations of the level of histone marks with the level of gene expression or exon inclusion. In our case, regardless of the level of inclusion in one condition, we can relate the relative change in inclusion with the activity of one or more factors, which in turn provides directionality to the association. In this sense, we find that the chromatin signals generally linked to the activation of enhancers can activate alternative internal transcription start sites. Additionally, we find that this activity often co-occur with the increase of inclusion in nearby exons. Altogether, our analysis suggests that transcription and splicing may be often coupled through the activation or silencing of intragenic enhancers.

Figure captions

Figure 1. A predictive model of active enhancers A) Feature selection using H3K27ac as a correlation class. The bars represent the average importance score per feature after averaging over 10 random samples of 5000 intergenic windows extracting from all intergenic windows with signal in at least one cell type. Red labels and bars indicate the minimum (randMin), mean (randMean) and maximum (randMax) of the simulated replicates, as well as the ChIP-Seq with a non-specific antibody (Control). The red dashed line separates the relevant features (in blue) from the non-relevant features (in grey). Feature selection and graphical representation was performed with Boruta (Methods). **B)** Scatter plot of the intergenic windows according to relative enrichment z-scores for every pair of selected feature (x and y axes). Each dot represents a window and windows are separated according to the three clusters (Methods): active enhancers are represented in green, unchanged enhancers in blue and silent enhancers in red. The black centroids show the center and standard deviation of the correlation between different features. Clustering and graphical representation were performed in mclust (Methods).

Figure 2. Properties of predicted intergenic enhancers. A) Relative enrichment of PU.1 at active and silent enhancer, as well as for regions of no-change in chromatin. The violin plot describes the distributions for the z-score of the relative enrichment along the y-axis. Positive z-score values mean enrichment in K562, while negative z-scores mean enrichment in GM12878. **B)** Distributions of the relative enrichment for polyadenylated long (>200nt) nuclear RNA on active, silent and regions with no change in chromatin signals. **C)** Distributions of the relative enrichment of polyadenylated long (>200nt) nuclear RNA on active, silent and regions with no change in chromatin signals. Distributions are represented as before. **D)** Percentage of enhancers at a given distance from the TSS, for active (blue), silent (green), as well as for the corresponding randomized sets (red and cyan) (Methods). **E)** Relative expression change in genes associated to enhancers by proximity to the TSS.. The violin plot describes the distributions of z-score of the relative enrichment along the y-axis for active and silent enhancers, as well as for no-change regions. Genes where linked to nearest putative enhancers within a distance range between 2kb and 10kb. **F)** Cumulative distribution of enhancer nearby genes related to cancer in terms of the distance between the TSS and the closes enhancer. The comparison is made between active and silent predicted enhancers, and the corresponding randomizations.

Figure 3. Intragenic enhancers and alternative Transcription initiation. A) Relative enrichment z-scores of RNAPII on the most upstream TSS (TSS1) when there is active (left red violin plot) and silent (right blue violin plot) when the enhancer sits between both TSSs and at a minimum distance of 1000 nt from either one. The yellow violin plot in the middle yellow represents the z-score distribution of all TSS1-TSS2 pairs, with or without predicted enhancers, with a dashed green line representing the median value of this distribution. **B)** Relative enrichment z-scores of RNAPII on the second alternative TSS (TSS2) in TSS1-TSS2 events. The conventions are as in A). **C)** A TSS1-TSS2 event in gene MAGED1 with a silent enhancer in K562 close to TSS2 (green box). RNA-Seq data from the same cell lines show lower RNA-Seq density in K562.

Figure 4. Effect of intragenic enhancers on alternative splicing. This figure describes the found combinations of enhancer type and alternative splicing regulation. We shod the significant inclusion and skipping events in K562 associated to **A)** an active enhancer

downstream **B)** an active enhancer upstream, **C)** a silent enhancer downstream; and **D)** a silent enhancer upstream. **E)** Example of an inclusion event in the gene LTBP4 with an active enhancer upstream/downstream in K562.

Acknowledgements

The authors would like to thank E. Furlong and Y. Barash for useful discussions. This work was supported by grants from Plan Nacional I+D (BIO2008-01091 and BIO2011-23920) and Consolider programme (CSD2009-00080) from the Spanish Ministry of Science, and by a grant from the Sandra Ibarra Foundation for Cancer (FSI 2011-035). JGV was supported by an FPI grant from the Spanish Ministry of Science (BES-2009-018064).

References

Alexander RD, Innocente SA, Barrass JD, Beggs JD. Splicing-dependent RNA polymerase pausing in yeast. Mol Cell. 2010 Nov 24;40(4):582-93.

Allo M, Buggiano V, Fededa JP, Petrillo E, Schor I, de la Mata M, Agirre E, Plass M, Eyras E, Elela SA, Klinck R, Chabot B, Kornblihtt AR. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. Nat Struct Mol Biol. 2009 Jul;16(7):717-24.

Alló M, Schor IE, Muñoz MJ, de la Mata M, Agirre E, Valcárcel J, Eyras E, Kornblihtt AR. Chromatin and alternative splicing. Cold Spring Harb Symp Quant Biol. 2010;75:103-11.

Althammer S, González-Vallinas J, Ballaré C, Beato M, Eyras E. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. Bioinformatics. 2011 Dec 15;27(24):3333–40.

Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012 Oct;22(10):2008-17. doi: 10.1101/gr.133744.111.

Balakrishnan L, Milavetz B. Decoding the histone H4 lysine 20 methylation mark. Crit. Rev. Biochem. Mol. Biol. 2010 Oct;45(5):440–52.

Batsche, E., Yaniv, M. & Muchardt, C. The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat Struct Mol Biol* **13**, 22-29 (2006).

Bégay V, Smink J, Leutz A. Essential requirement of CCAAT/enhancer binding proteins in embryogenesis. Mol. Cell. Biol. 2004 Nov;24(22):9744–51.

Beck DB, Oda H, Shen SS, Reinberg D. PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. Genes Dev. 2012 Feb 15;26(4):325–37.

Birnbaum, R. Y., Clowney, E. J., Agamy, O., Kim, M. J., Zhao, J., Yamanaka, T., Pappalardo, Z., et al. (2012).

Coding exons function as tissue-specific enhancers of nearby genes. *Genome research*, 1059-1068.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. a, Holt, A., Plajzer-Frick, I., et al. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nature genetics*, *42*(9), 806-10.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., et al. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, *132*(2), 311-22.

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., et al. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature genetics*, *44*(2), 148-56.

Bradnam KR, Korf I. Longer First Introns Are a General Property of Eukaryotic Gene Structure. PLoS ONE. 2008 Aug 29;3(8):e3093.

Carmo-Fonseca M, Carvalho C. Nuclear organization and splicing control. Adv Exp Med Biol. 2007;623:1-13.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *PNAS* 21931-6. doi:10.1073/pnas.1016071107

de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR. A slow RNA polymerase II affects alternative splicing in vivo. Mol Cell. 2003 Aug;12(2):525-32.

de la Mata M, Muñoz MJ, Alló M, Fededa JP, Schor IE, Kornblihtt AR. RNA Polymerase II Elongation at the Crossroads of Transcription and Alternative Splicing. Genet Res Int. 2011;2011:309865.

de Almeida SF, Grosso AR, Koch F, Fenouil R, Carvalho S, Andrade J, Levezinho H, Gut M, Eick D, Gut I, Andrau JC, Ferrier P, Carmo-Fonseca M. Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. Nat Struct Mol Biol. 2011 Jul 26;18(9):977-83.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., et al. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology*, *8*(5), e1000384. doi:10.1371/journal.pbio.1000384

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012 Apr 11;485(7398):376-80.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74.

Eckner R. p300 and CBP as transcriptional regulators and targets of oncogenic events. Biol Chem. 1996 Nov;377(11):685-8. Review.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011 May 5;473(7345):43–9.

Fan JY, Gordon F, Luger K, Hansen JC, Tremethick DJ. The essential histone variant H2A.Z regulates the equilibrium between different chromatin conformational states. Nat Struct Biol. 2002 Mar;9(3):172-6.

Fraley C, Raftery A. MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. 2007.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nature Reviews Cancer. 2004 Mar 1;4(3):177–83.

Ghisletti S, Barozzi I, Mietton F, Polletti S, De Santa F, Venturini E, Gregory L, Lonie L, Chew A, Wei CL, Ragoussis J, Natoli G. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. Immunity. 2010 Mar 26;32(3):317-28.

Gomes NP, Espinosa JM. Gene-specific repression of the p53 target gene PUMA via intragenic CTCF-Cohesin binding. Genes Dev. 2010 May 15;24(10):1022-34.

Gunderson FQ, Johnson TL. Acetylation by the transcriptional coactivator Gcn5 plays a novel role in co-transcriptional spliceosome assembly. PLoS Genet. 2009 Oct;5(10):e1000682.

Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012 Feb 15;482(7385):339-46. doi: 10.1038/nature10887.

Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012 Apr;13(2):204–16.

Han S, Brunet A. Histone methylation makes its mark on longevity. Trends Cell Biol. 2012 Jan;22(1):42-9.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. 2012 Sep;22(9):1760–74.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*(7243), 108-12.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, *39*(3), 311-8.

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

Ip JY, Schmidt D, Pan Q, Ramani AK, Fraser AG, Odom DT, Blencowe BJ. Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. Genome Res. 2011 Mar;21(3):390-401.

Jia L, Landan G, Pomerantz M, Jaschek R, Herman P, Reich D, et al. Functional Enhancers at the Gene-Poor 8q24 Cancer-Linked Locus. PLoS Genet. 2009;5(8):e1000597.

Jin F, Li Y, Ren B, Natarajan R. Enhancers: multi-dimensional signal integrators. Transcription. 2011 Sep Oct;2(5):226-30.

Kadener S, Fededa JP, Rosbash M, Kornblihtt AR. Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. Proc Natl Acad Sci U S A. 2002 Jun 11;99(12):8185-90.

Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods. 2010;7(12):1009–15.

Kowalczyk, M. S., Hughes, J. R., Garrick, D., Lynch, M. D., Sharpe, J. a, Sloane-Stanley, J. a, McGowan, S. J., et al. (2012). Intragenic enhancers act as alternative promoters. *Molecular cell*, *45*(4), 447-58.

Kornblihtt AR. CTCF: from insulators to alternative splicing regulation. Cell Res. 2012 22(3):450-2.

Ku M, Jaffe JD, Koche RP, Rheinbay E, Endoh M, Koseki H, Carr SA, Bernstein BE. H2A.Z landscapes and dual

modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. Genome Biol. 2012 Oct 3;13(10):R85.

Jin C, Zang C, Wei G, Cui K, Peng W, Zhao K, Felsenfeld G. H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. Nat Genet. 2009 Aug;41(8):941-5.

Jin, F., Li, Y., Ren, B., & Natarajan, R. (2011). Enhancers: multi-dimensional signal integrators. *Transcription*, *2*(5), 226-30.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. a, et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182-7. Nature Publishing Group. doi:10.1038/nature09033

Kim S, Kim H, Fong N, Erickson B, Bentley DL. Pre-mRNA splicing is a determinant of histone H3K36 methylation. Proc Natl Acad Sci U S A. 2011 Aug 16;108(33):13564-9.

Kornblihtt AR. Coupling transcription and alternative splicing. Adv Exp Med Biol. 2007;623:175-89.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010 Oct;11(10):733–9.

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. Cell. 2012 Jan 20;148(1–2):84–98.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map Format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

Lin S, Coutinho-Mansfield G, Wang D, Pandit S, Fu XD. The splicing factor SC35 has an active role in transcriptional elongation. Nat Struct Mol Biol. 2008 Aug;15(8):819-26.

Loomis RJ, Naoe Y, Parker JB, Savic V, Bozovsky MR, Macfarlan T, Manley JL, Chakravarti D. Chromatin binding of SRp20 and ASF/SF2 and dissociation from mitotic chromosomes is modulated by histone H3 serine 10 phosphorylation. Mol Cell. 2009 Feb 27;33(4):450-61.

Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. Epigenetics in alternative pre-mRNA splicing. Cell. 2011 Jan 7;144(1):16-26.

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. Regulation of alternative splicing by histone modifications. Science. 2010 Feb 19;327(5968):996-1000.

Lupien, M., Eeckhoute, J., Meyer, C. a, Wang, Q., Zhang, Y., Li, W., Carroll, J. S., et al. (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, *132*(6), 958-70.

Maston GA, Landt SG, Snyder M, Green MR. Characterization of enhancer function from genome-wide analyses. Annu Rev Genomics Hum Genet. 2012 Sep 22;13:29-57.

Mitchell T. *Machine Learning*, TheMc-Graw-Hill Companies, Inc., 1997.

Neugebauer KM. On the importance of being co-transcriptional. J Cell Sci. 2002 Oct 15;115(Pt 20):3865-71.

Ong C-T, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. 2011 Apr;12(4):283–93.

Padgett RA. New connections between splicing and human disease. Trends Genet. 2012, 28(4):147-54.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008 Dec;40(12):1413-5.

Pekowska A, Benoukraf T, Zacarias-Cabeza J, Belhocine M, Koch F, Holota H, et al. H3K4 tri-methylation provides an epigenetic signature of active enhancers. EMBO J. 2011 Oct 19;30(20):4198–210.

Pennacchio, L. a, & Visel, A. (2010). Limits of sequence and functional conservation. *Nature genetics*, *42*(7), 557-8.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. a, Flynn, R. a, & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, *470*(7333), 279-83. Nature Publishing Group. doi:10.1038/nature09692

Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics. 2011 Dec 17;12(1):480.

Ritter, D. I., Dong, Z., Guo, S., & Chuang, J. H. (2012). Transcriptional Enhancers in Protein-Coding Exons of Vertebrate Developmental Genes. (M. Schubert, Ed.)*PLoS ONE*, *7*(5), e35202.

Robertson AG, Bilenky M, Tam A, Zhao Y, Zeng T, Thiessen N, Cezard T, Fejes AP, Wederell ED, Cullum R, Euskirchen G, Krzywinski M, Birol I, Snyder M, Hoodless PA, Hirst M, Marra MA, Jones SJ. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. Genome Res. 2008 Dec;18(12):1906-17.

Saint-André V, Batsché E, Rachez C, Muchardt C. Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. Nat Struct Mol Biol. 2011 Mar;18(3):337-44.

Schor, I. E., Rascovan, N., Pelisch, F., Allo, M. & Kornblihtt, A. R. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. *Proc Natl Acad Sci U S A* **106**, 4325-4330 (2009).

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature. 2011 Nov 3;479(7371):74-9.

Sims RJ 3rd, Millhouse S, Chen CF, Lewis BA, Erdjument-Bromage H, Tempst P, Manley JL, Reinberg D. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. Mol Cell. 2007 Nov 30;28(4):665-76.

Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978 Mar;6(2):461–4.

Visel A, Rubin EM, Pennacchio LA. Genomic Views of Distant-Acting Enhancers. Nature. 2009 Sep 10;461(7261):199–205.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. 2007 Jan;35(Database issue):D88–92.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009 Feb 12;457(7231):854–8.

Wang, Q., Carroll, J. S. & Brown, M. Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. Mol. Cell 19, 631–642 (2005).

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008 Nov 27;456(7221):470-6.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013 Apr 11;153(2):307-19.

Wiench, M., John, S., Baek, S., Johnson, T. a, Sung, M.-H., Escobar, T., Simmons, C. a, et al. (2011). DNA methylation status predicts cell type-specific enhancer activity. *The EMBO journal*, *30*(15), 3028-39.

Zentner GE, Scacheri PC. The chromatin fingerprint of gene enhancer elements. J Biol Chem. 2012 Sep 7;287(37):30888-96.

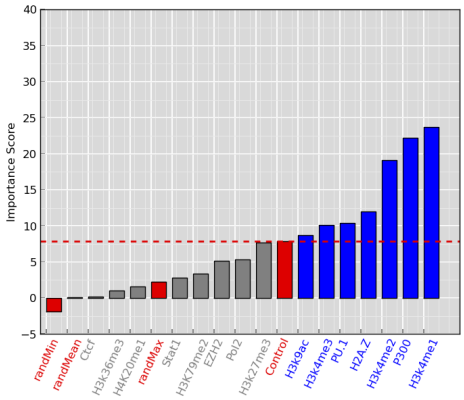
Zhou HL, Hinman MN, Barron VA, Geng C, Zhou G, Luo G, Siegel RE, Lou H. Hu proteins regulate alternative splicing by inducing localized histone hyperacetylation in an RNA-dependent manner. Proc Natl Acad Sci U S A. 2011 Sep 6;108(36):E627-35.

Uniprot PU.1: http://www.uniprot.org/uniprot/P17947

Wen C-J, Xue B, Qin W-X, Yu M, Zhang M-Y, Zhao D-H, et al. hNRAGE, a human neurotrophin receptor interacting MAGE homologue, regulates p53 transcriptional activity and inhibits cell proliferation. FEBS Lett. 2004 Apr 23;564(1-2):171–6.

Figure 1

A


Importance Score
40
35
30
25
20
15
10
5
0
−5
randMin
randMean
Ctcf
H3k36me3
H4K20me1
randMax
Stat1
H3K79me2
EZH2
Pol2
H3k27me3
Control
H3k9ac
H3k4me3
PU.1
H2A.Z
H3k4me2
P300
H3k4me1

B

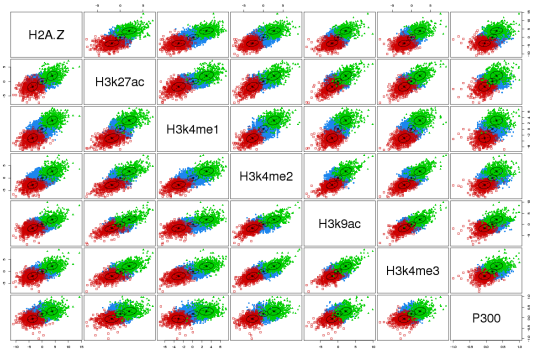

H2A.Z
H3k27ac
H3k4me1
H3k4me2
H3k9ac
H3k4me3
P300

Figure 2

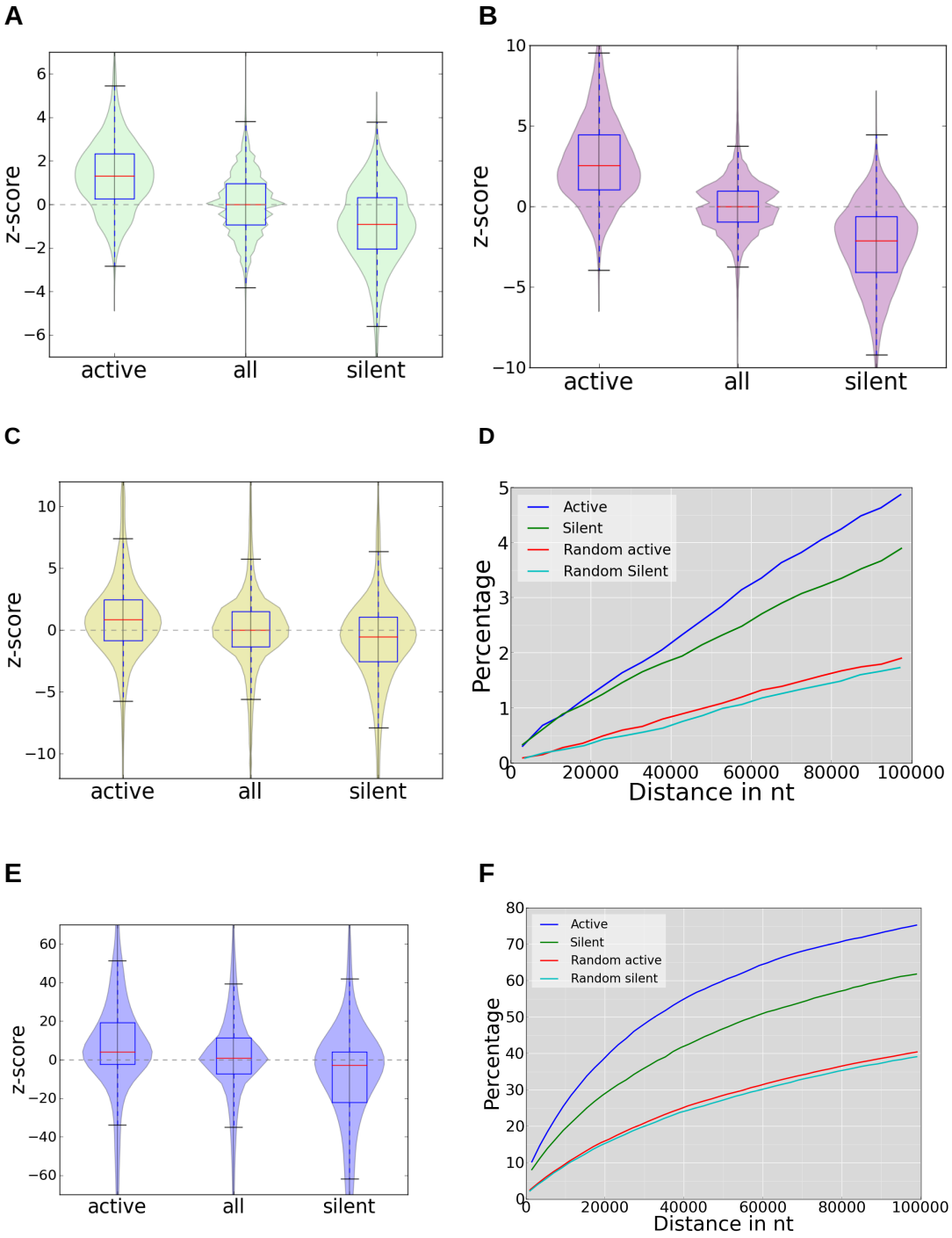
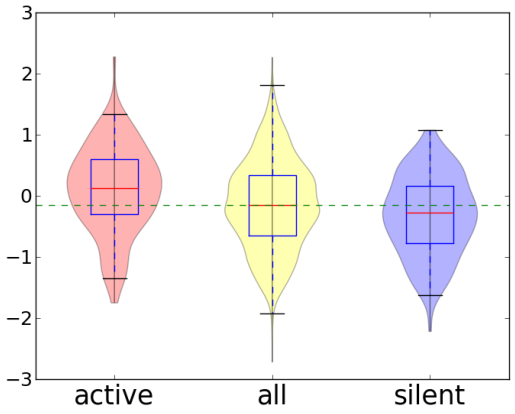
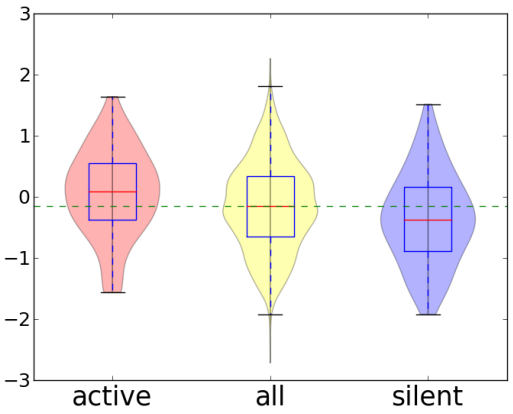

A
z-score
active
all
silent
B
z-score
active
all
silent
C
z-score
active
all
silent
D
Active
Silent
Random active
Random Silent
Percentage
Distance in nt
E
z-score
active
all
silent
F
Active
Silent
Random active
Random silent
Percentage
Distance in nt

Figure 3

A


3
2
1
0
−1
−2
−3
active
all
silent

B


3
2
1
0
−1
−2
−3
active
all
silent

C

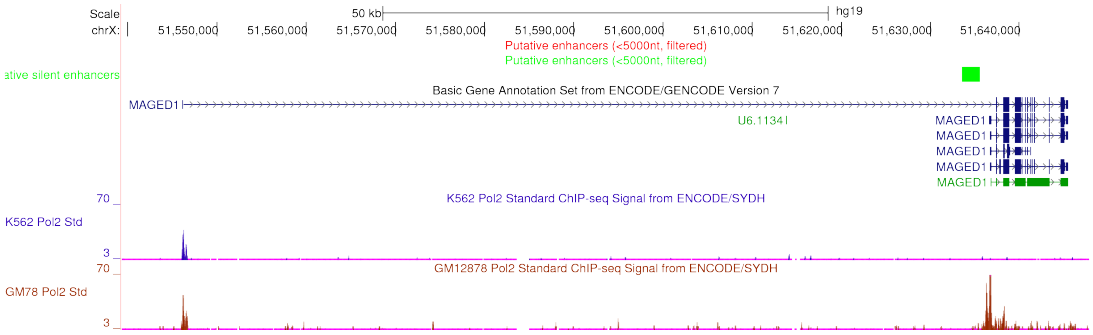
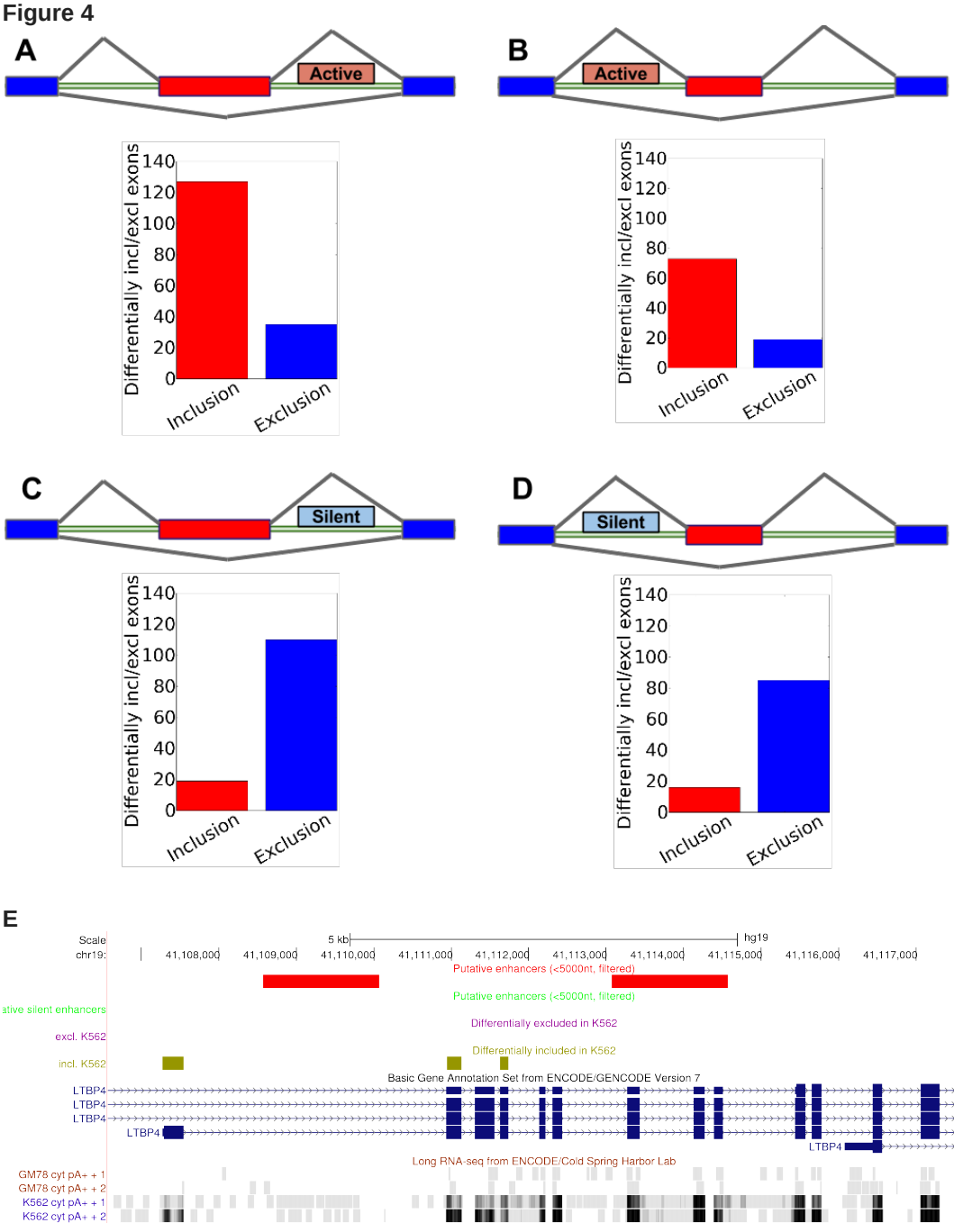
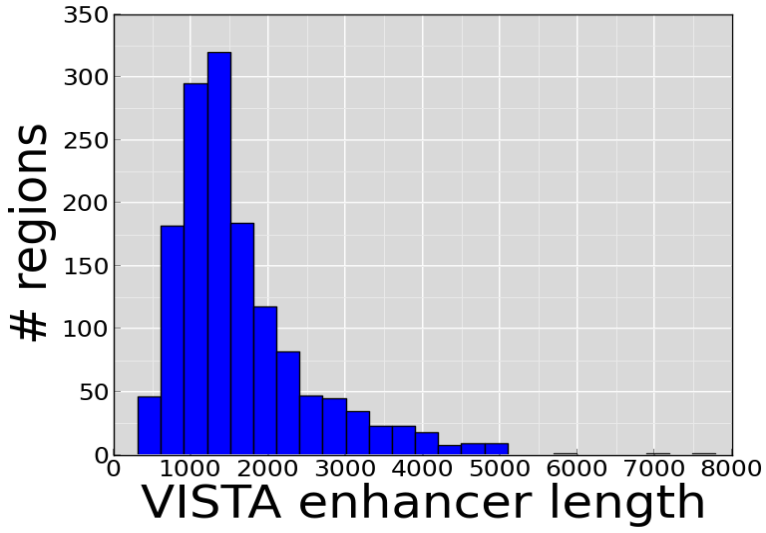

Scale
chrX:
50 kb
hg19
51,550,000
51,560,000
51,570,000
51,580,000
51,590,000
51,600,000
51,610,000
51,620,000
51,630,000
51,640,000
Putative enhancers (<5000nt, filtered)
Putative enhancers (<5000nt, filtered)
ative silent enhancers
Basic Gene Annotation Set from ENCODE/GENCODE Version 7
MAGED1
U6.11341
MAGED1
MAGED1
MAGED1
MAGED1
MAGED1
70
K562 Pol2 Standard ChIP-seq Signal from ENCODE/SYDH
K562 Pol2 Std
3
70
GM12878 Pol2 Standard ChIP-seq Signal from ENCODE/SYDH
GM78 Pol2 Std
3

Figure 4

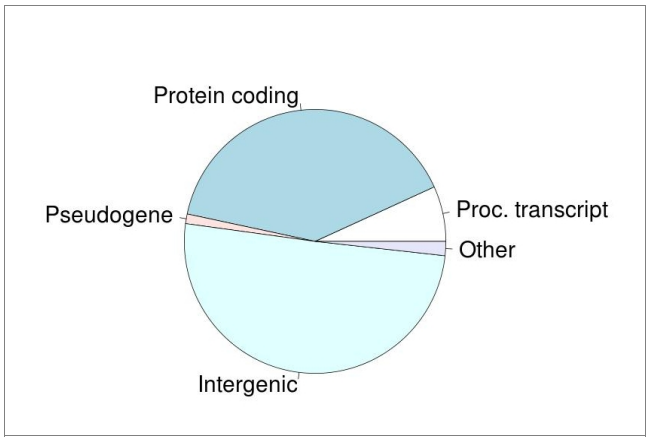

A
Active
Differentially incl/excl exons
140
120
100
80
60
40
20
0
Inclusion
Exclusion
B
Active
Differentially incl/excl exons
Inclusion
Exclusion
C
Silent
Differentially incl/excl exons
Inclusion
Exclusion
D
Silent
Differentially incl/excl exons
Inclusion
Exclusion
E
Scale
chr19:
5 kb
hg19
41,108,000
41,109,000
41,110,000
41,111,000
41,112,000
41,113,000
41,114,000
41,115,000
41,116,000
41,117,000
Putative enhancers (<5000nt, filtered)
Putative enhancers (<5000nt, filtered)
ative silent enhancers
Differentially excluded in K562
excl. K562
Differentially included in K562
incl. K562
Basic Gene Annotation Set from ENCODE/GENCODE Version 7
LTBP4
LTBP4
LTBP4
LTBP4
LTBP4
Long RNA-seq from ENCODE/Cold Spring Harbor Lab
GM78 cyt pA+ + 1
GM78 cyt pA+ + 2
K562 cyt pA+ + 1
K562 cyt pA+ + 2

Supplementary Material

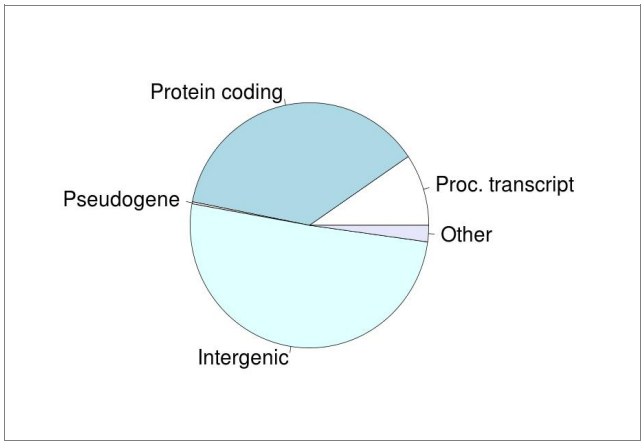
Properties of VISTA enhancers



Supplementary Figure 1A. Length distribution of human VISTA enhancers. The average length of the VISTA regions is 1637.9 nt, the median is 1383 and the standard deviation is 891 nucleotides. Out of the 1447 experimentally validated regions, only 6 (0.41%) are above 5000 nucleotides.

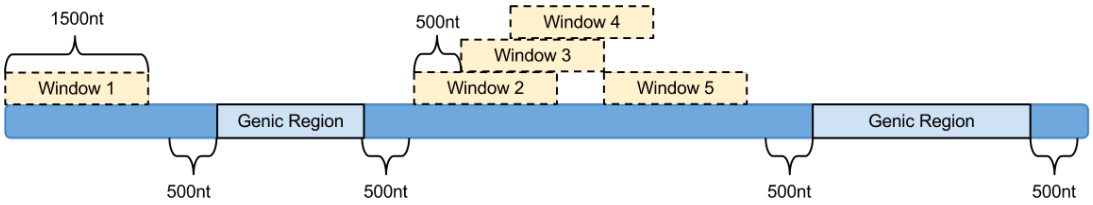


Supplementary Figure 1B. GENCODE coverage of the genome. Pie chart representing the percentage of bases covered by different annotated elements in the human genome (GENCODE V7): protein coding genes, Processed Transcripts, Pseudogenes, regions and Other, which summarizes relatively rare annotated elements. These elements are: Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes (active and silent), several types of small non-coding RNA and lincRNA. The rest of non annotated elements by GENCODE are classified as intergenic

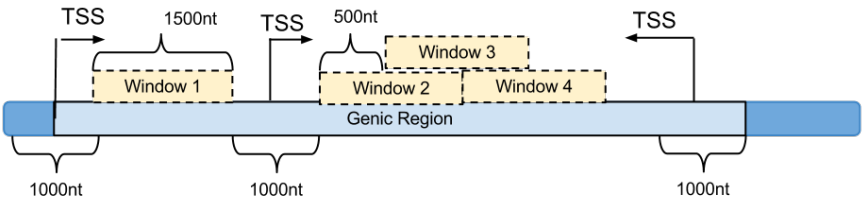


Supplementary Figure 1C. VISTA elements positioning. Percentage of the VISTA regions that fall in one of the categories described above.

Exploratory Windows

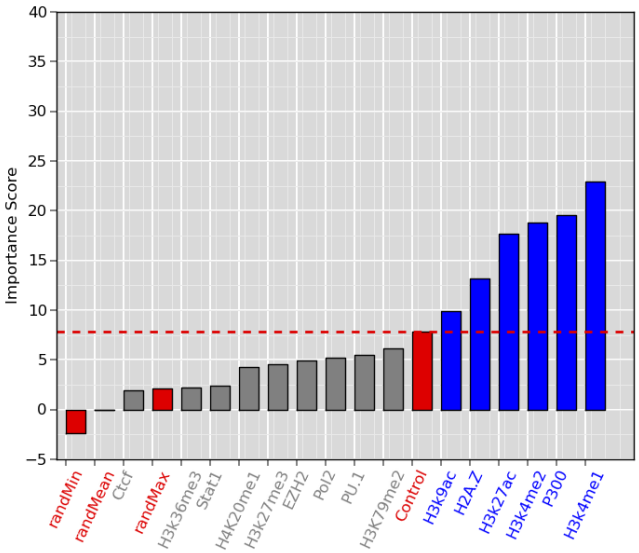


Supplementary Figure 2A. Intergenic windows representation. The exploratory intergenic windows have a size of 1500nt are overlapping 500nt from each other. In order to avoid mixing with promoter signal, they are at least 500nt away from Genic regions.

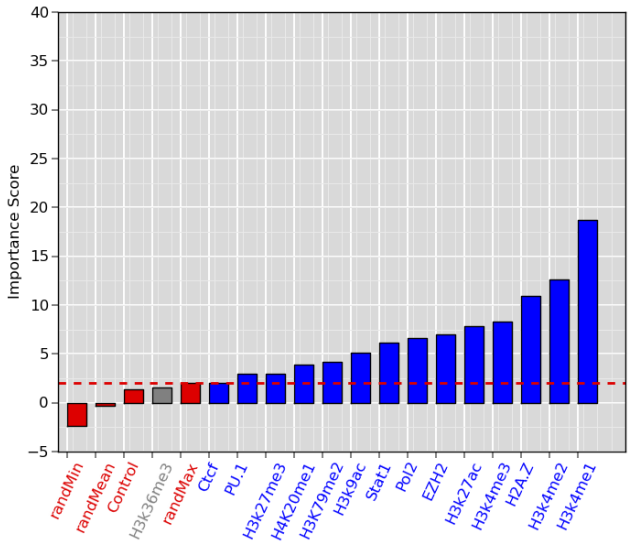


Supplementary Figure 2B. Intragenic windows representation. Same size as the intergenic windows. As in with the intragenic windows, other and are at least 500nt away from Genic regions.

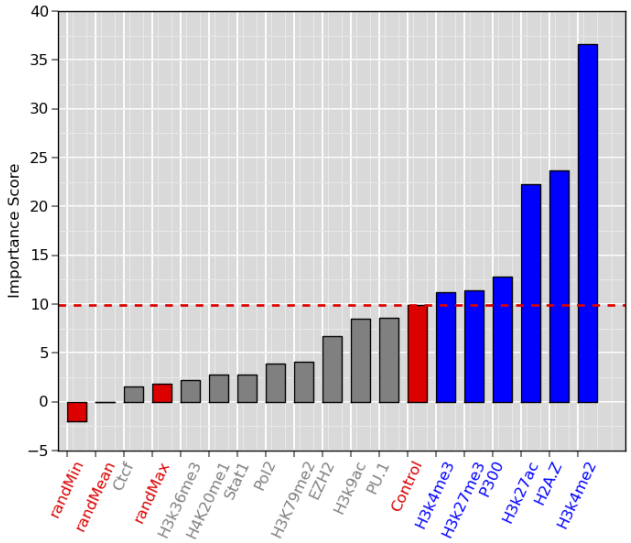
Feature Selection with Boruta



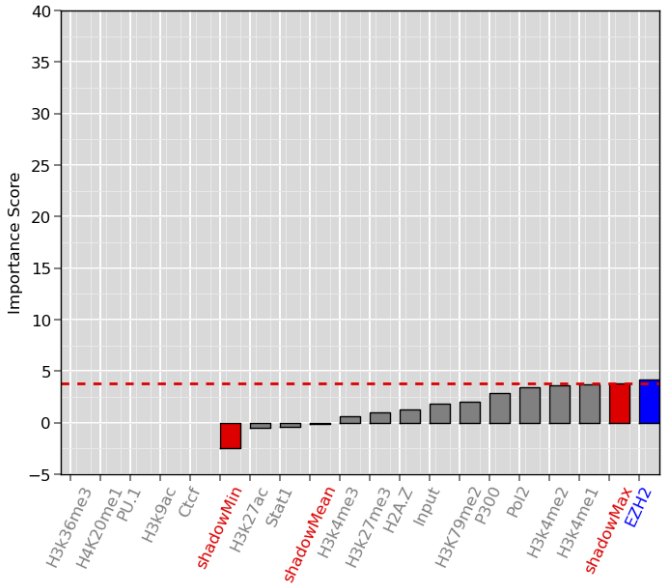
Supplementary Figure 3A - Feature selection using H3K27ac as a correlation class. The bars represent the average importance score per feature after averaging over 10 random samples of 5000 intergenic windows extracting from all intergenic windows with signal in at least one cell type. Red labels and bars indicate the minimum (randMin), mean (randMean) and maximum (randMax) of the simulated replicates, as well as the ChIP-Seq with a non-specific antibody (Control). The red dashed line separates the relevant features (in blue) from the non-relevant features (in grey).



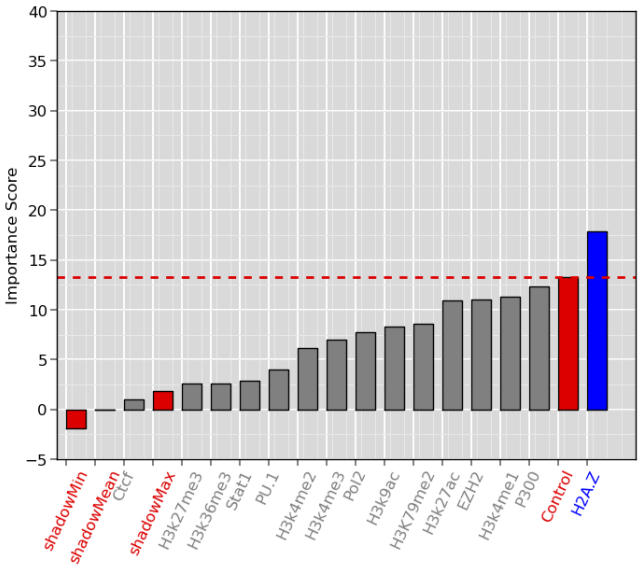
Supplementary Figure 3B Feature selection average scores using P300 as the correlation class



Supplementary Figure 3C - Feature selection average scores using H3K4me1 as the correlation class



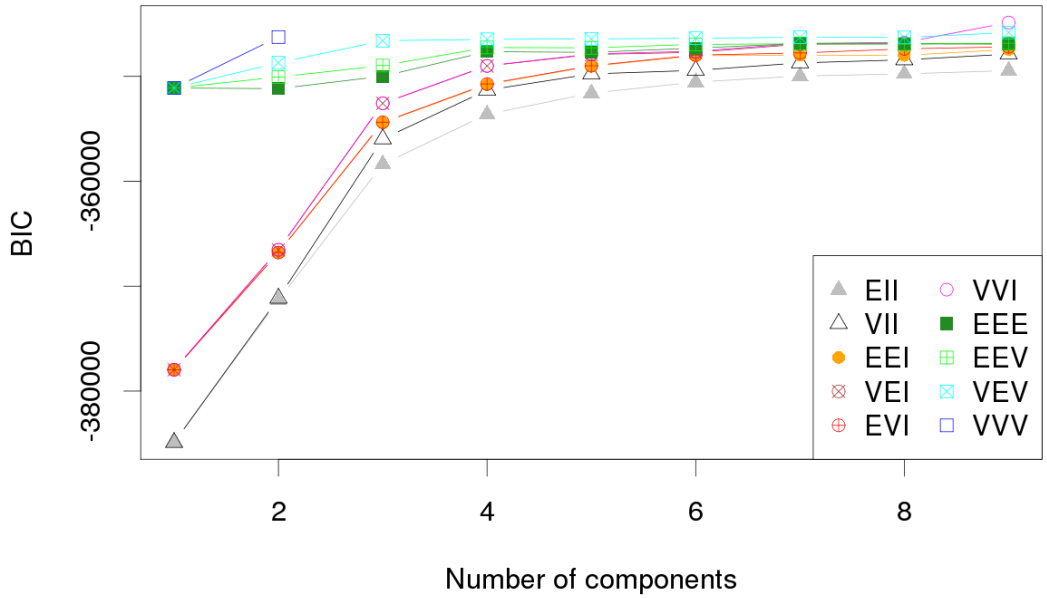
Supplementary Figure 3D - Feature selection average scores using Control as the correlation class



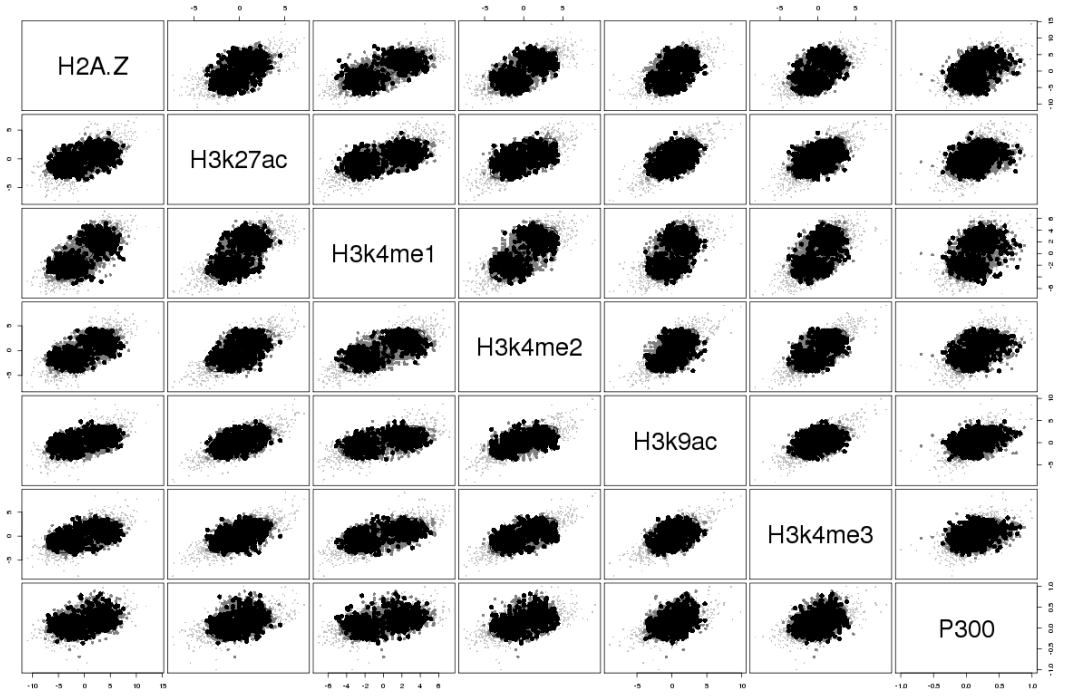
Supplementary Figure 3E - Feature selection average scores using H4k20me1 as the

correlation class.

Intergenic clustering

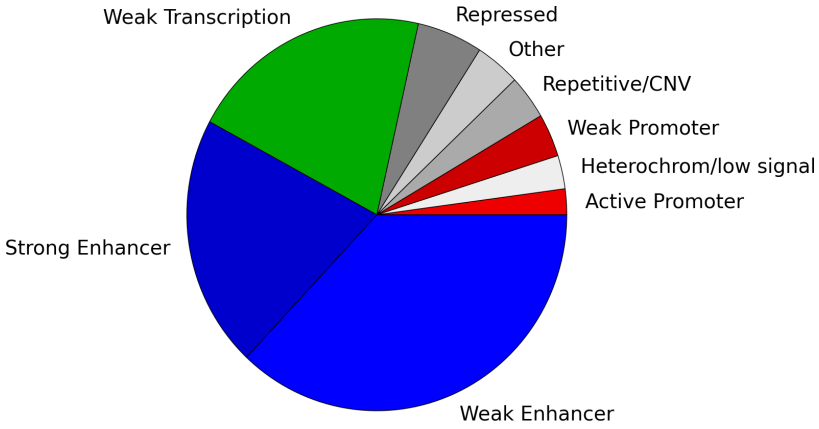


Supplementary Figure 4A. BIC information criterion calculation by Mclust. The X axis represents the number of clusters, the Y axis the BIC score. Every line correspond to a different type of model. The model that scores higher and plateaus faster is VEV (Variable Volume, Equal Shape, Variable Orientation).

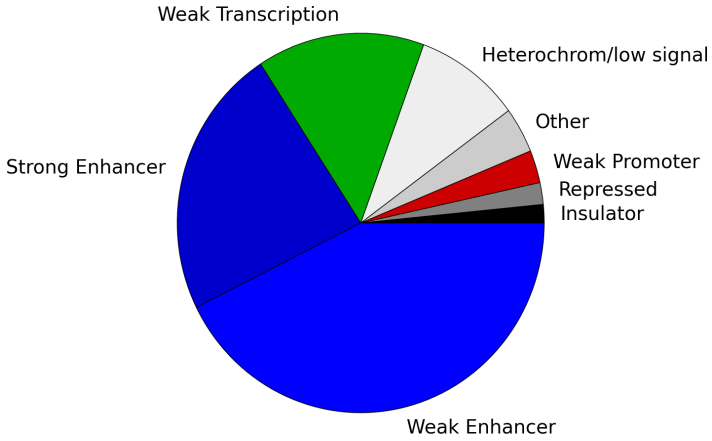


Supplementary Figure 4B. Mclust visual representation of the uncertainty. Dark dots represent more uncertainty (less probable to be good predictions), lighter dots are more certain (more probable to be good predictions). As expected, bigger differences between K562 and GM12878 levels correlate with less uncertainty.

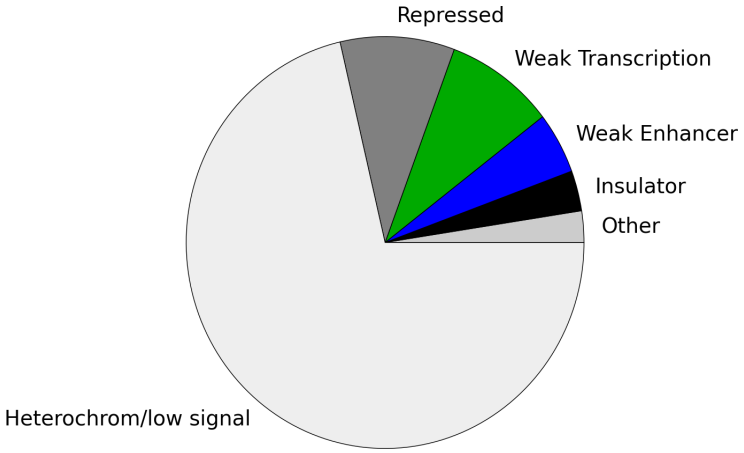
Comparing intergenic predictions with ChromHMM predictions



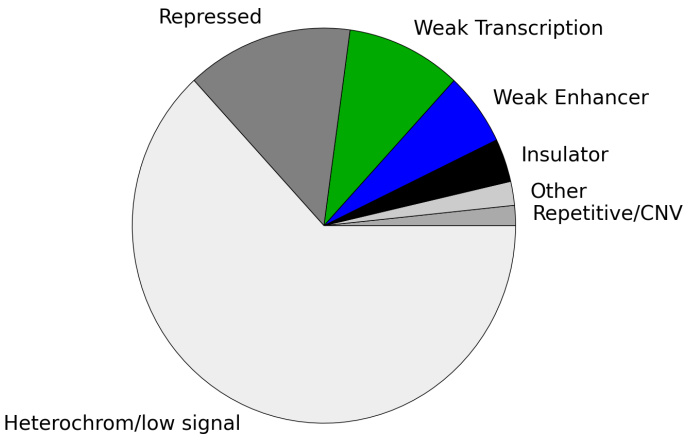
Supplementary Figure 5A. Pie chart showing the overlap of our active intergenic enhancer windows with different ChromHMM (Ernst et al. 2011) classes for K562. In blue, predictions that correlate with enhancer predictions, in green predictions that correlate with transcription predictions, in red correlation with promoter regions and in grey Repressed, Repetitive and others.



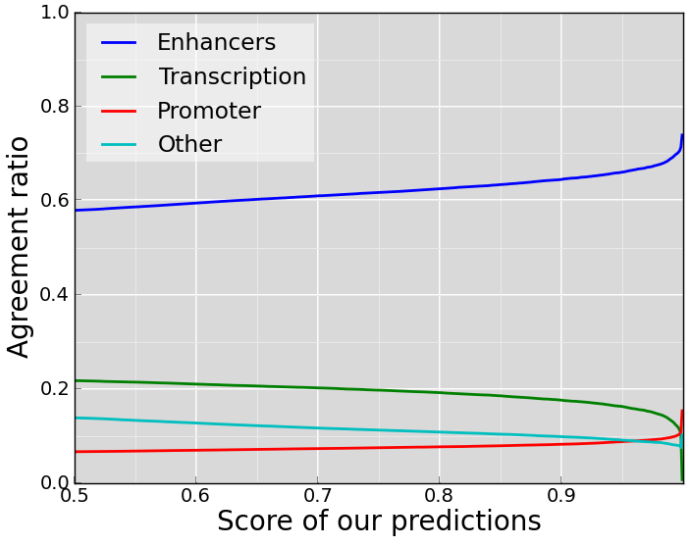
Supplementary Figure 5B. Pie chart showing the overlap of our silent intergenic enhancers with ChromHMM classes for GM12878.



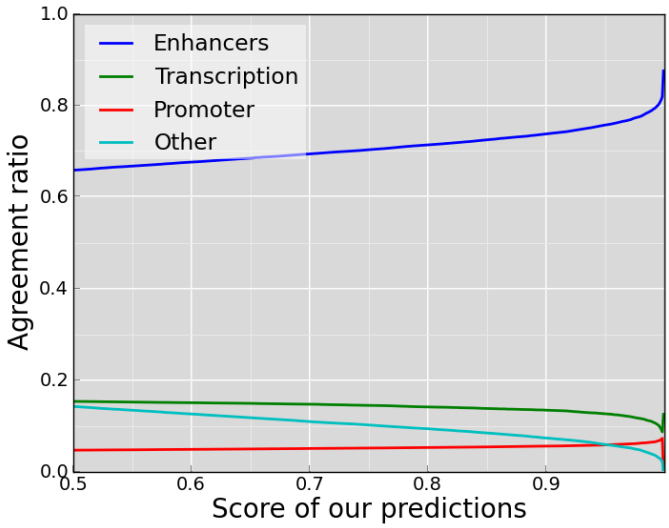
Supplementary Figure 5C. Pie chart showing the overlap of our active intergenic enhancers with ChromHMM classes for GM12878. In blue, predictions that correlate with enhancer predictions, in green predictions that correlate with transcription predictions, in red correlation with promoter regions and in grey Repressed, Repetitive and others.



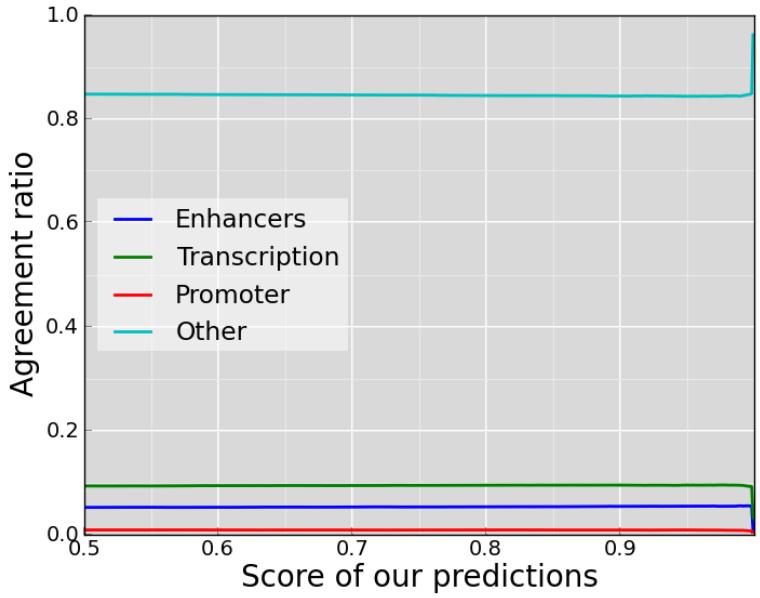
Supplementary Figure 5D. Pie chart showing the overlap of our silent intergenic enhancers with ChromHMM classes for K562. In blue, predictions that correlate with enhancer predictions, in green predictions that correlate with transcription predictions, in red correlation with promoter regions and in grey Repressed, Repetitive and others.



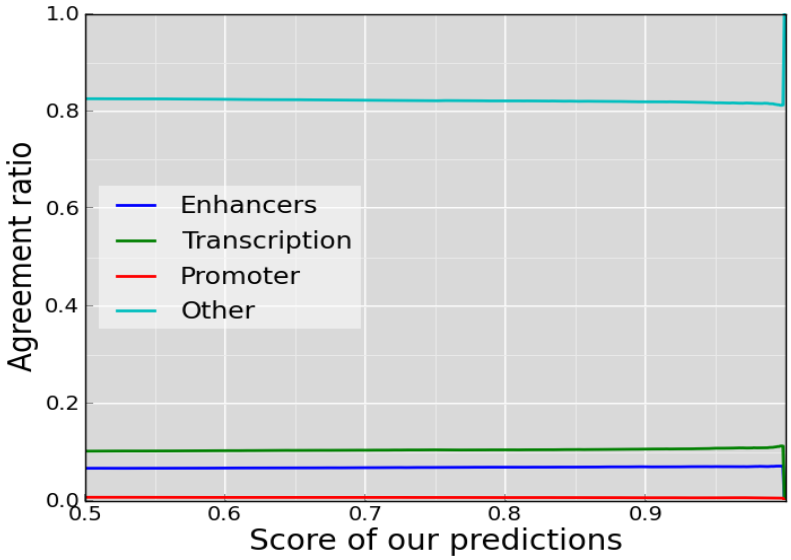
Supplementary Figure 5E - Proportion of our enhancers that are labeled by ChromHMM in K562 (y axis) as a function of the prediction score in K562, given in terms of the posterior probability for the class (x axis) The blue is the proportion of our active enhancer predictions that fall in windows labeled as weak or strong enhancers by ChromHMM in K562, green for all transcription related classes (weak, transition and elongation) and red for promoter (weak, poised and active). In Cyan we include all other classes predicted by ChromHMM (polycomb repressed, insulator, heterochromatin, repetitive, low signal)



Supplementary Figure 5F. Proportion of our enhancers that are labeled by ChromHMM (Ernst et al. 2011) in GM12878 (y axis) as a function of the prediction score in GM12879, given in terms of the posterior probability for the class (x axis) The blue is the proportion of our active enhancer predictions that fall in windows labeled as weak or strong enhancers by ChromHMM in K562, green for all transcription related classes (weak, transition and elongation) and red for promoter (weak, poised and active). In Cyan we include all other classes predicted by ChromHMM (polycomb repressed, insulator, heterochromatin, repetitive, low signal)

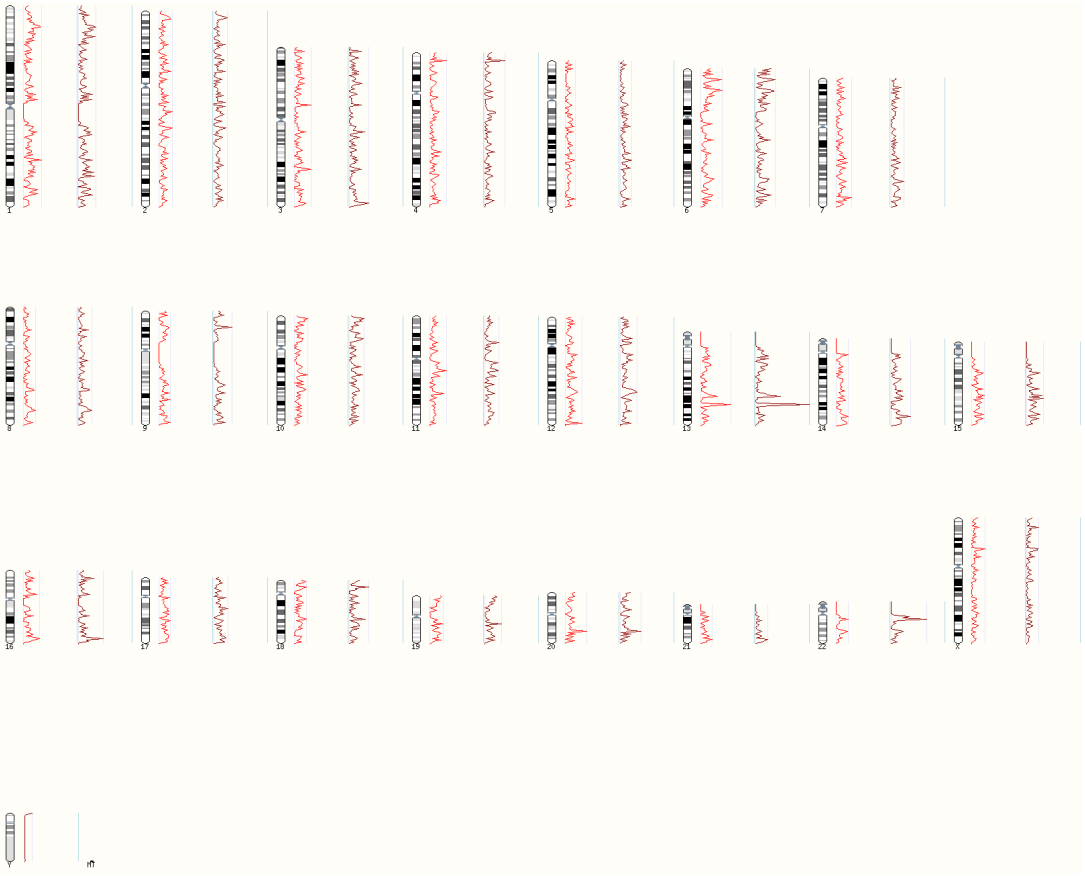


Supplementary Figure 5G. Proportion of our active enhancers in K562 that are labeled by ChromHMM (Ernst et al. 2011) in GM12878 (y axis) as a function of the prediction score in K562, given in terms of the posterior probability for the class (x axis) The blue is the proportion of in K562 our active enhancer predictions that fall in windows labeled as weak or strong enhancers by ChromHMM in GM12878, green for all transcription related classes (weak, transition and elongation) and red for promoter (weak, poised and active). In Cyan we include all other classes predicted by ChromHMM (polycomb repressed, insulator, heterochromatin, repetitive, low signal)

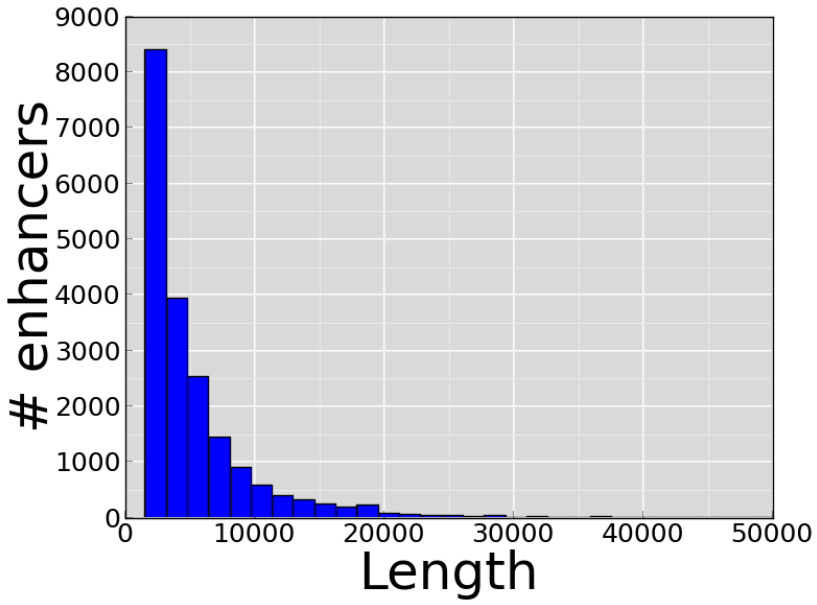


Supplementary Figure 5H. Proportion of our silenced enhancers in K562 that are labeled by ChromHMM (Ernst et al. 2011) in K562 (y axis) as a function of the prediction score in K562, given in terms of the posterior probability for the class (x axis) The blue is the proportion of in K562 our active enhancer predictions that fall in windows labeled as weak or strong enhancers by ChromHMM in K562, green for all transcription related classes (weak, transition and elongation) and red for promoter (weak, poised and active). In Cyan we include all other classes predicted by ChromHMM (polycomb repressed, insulator, heterochromatin, repetitive, low signal).

Properties of the intergenic enhancers

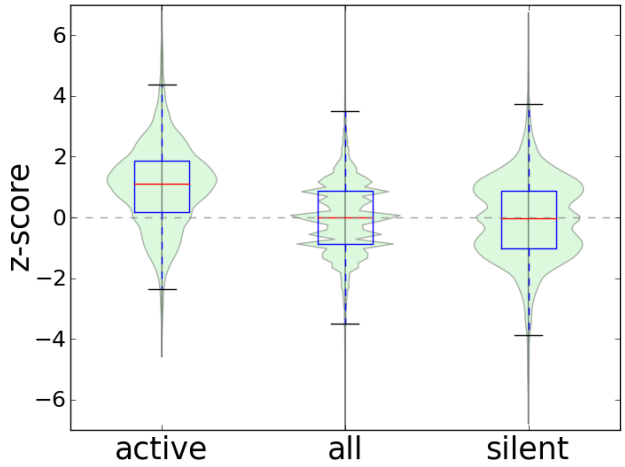


Supplementary Figure 6A. Genome wide distribution of intergenic enhancers. Profile of the density of predicted intergenic enhancers (activated and silenced) along the human karyotype. In dark red, predicted enhancers are represented for all lengths.. In lighter red, enhancers of lengths <5000bp are represented..

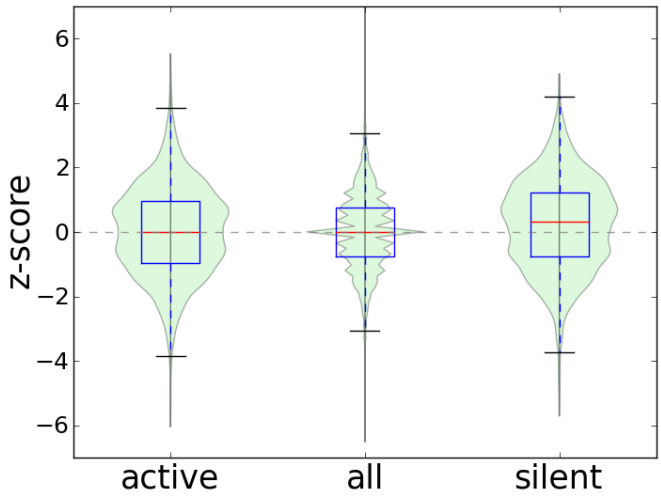


Supplementary Figure 6B - Length distribution of our predicted intergenic enhancers

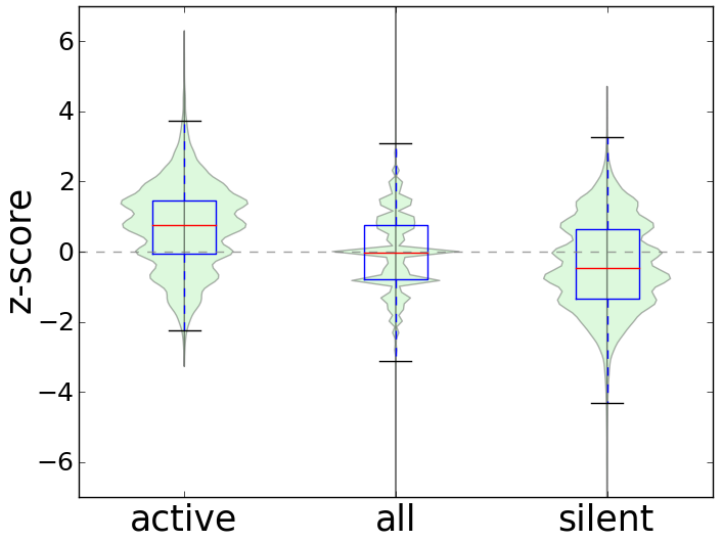
Correlation with other features



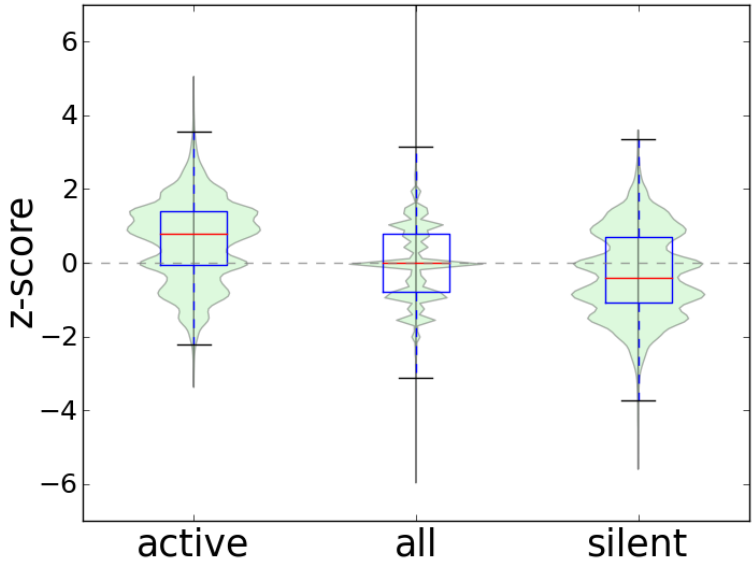
Supplementary Figure 7A. Relative enrichment of RNAPII in intergenic enhancers. Zscore distributions for our putative active and silent enhancers, as well as for all regions.



Supplementary Figure 7B. Relative enrichment of H3K27me3 in intergenic enhancers. Zscore distributions for our putative active and silent enhancers, as well as for all windows.

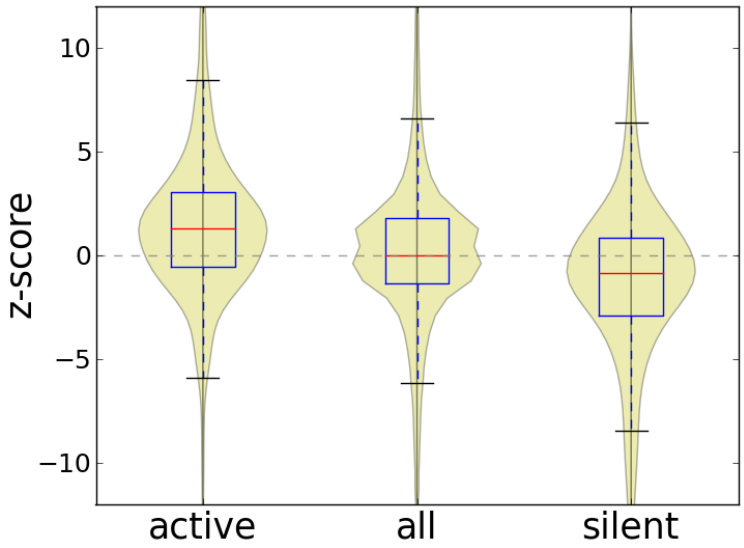


Supplementary Figure 7C. Relative enrichment of CTCF in intergenic enhancers. Zscore distributions for our putative active and silent enhancers, as well as for all enhancers.

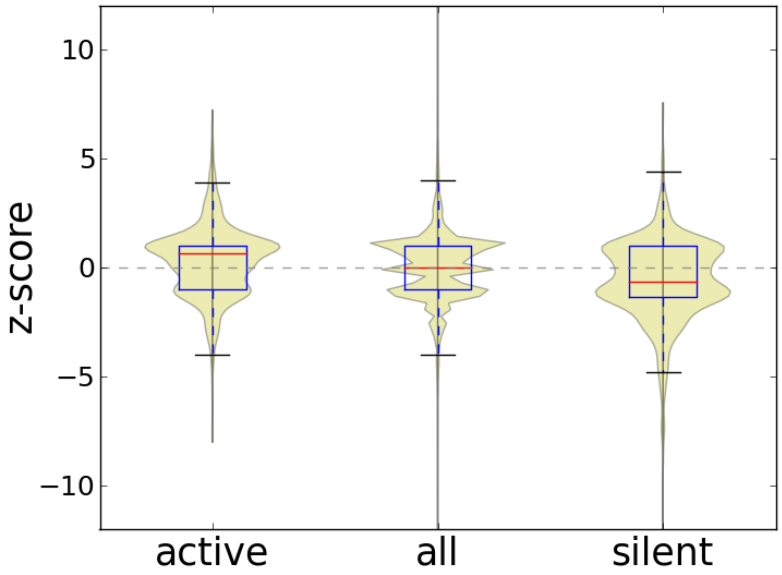


Supplementary Figure 7D. Relative enrichment of H3K36me3 in intergenic enhancers. Zscore distributions for our putative active and silent enhancers, as well as for all enhancers.

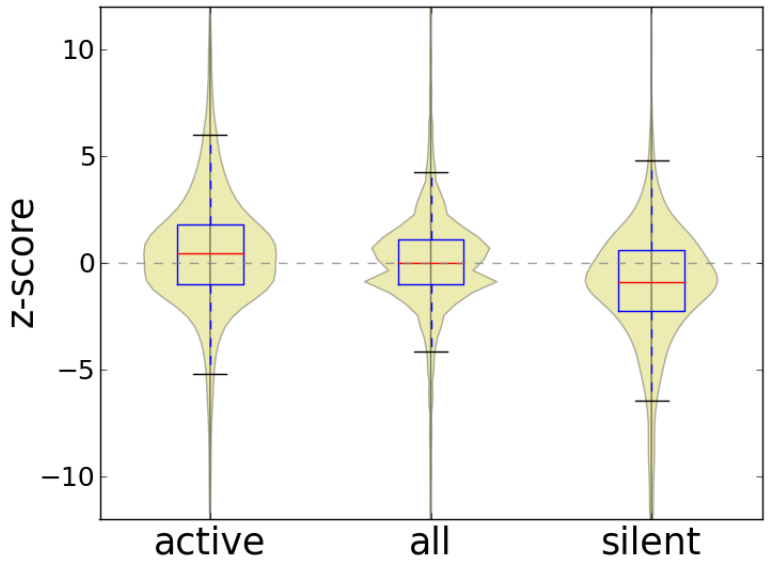
eRNA



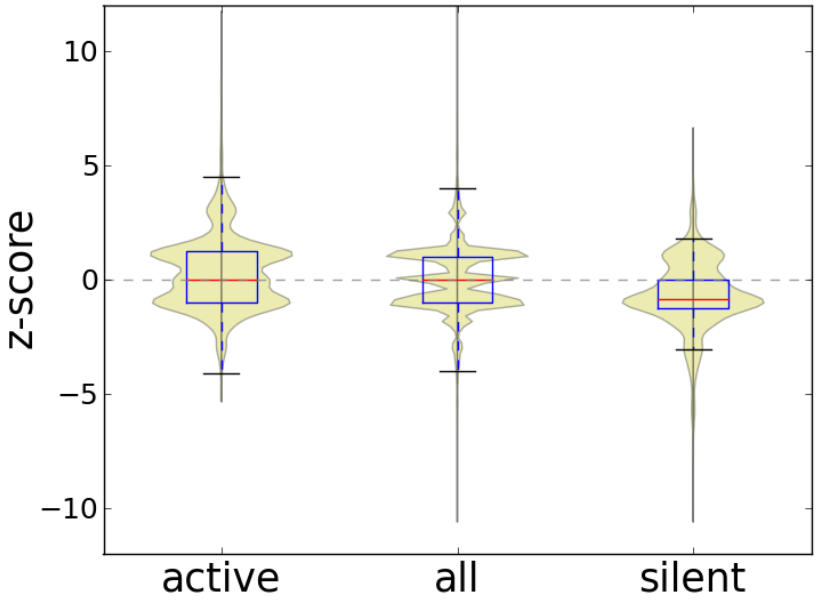
Supplementary Figure 8A. Nuclear polyA- eRNAs. **Violin plots for** Nuclear long (>200nt) poly A- RNA-Seq. Relative enrichment of nuclear poly A- reads in K562 relative GM12878 in active and silent enhancers, as well as the distribution of all z-scores



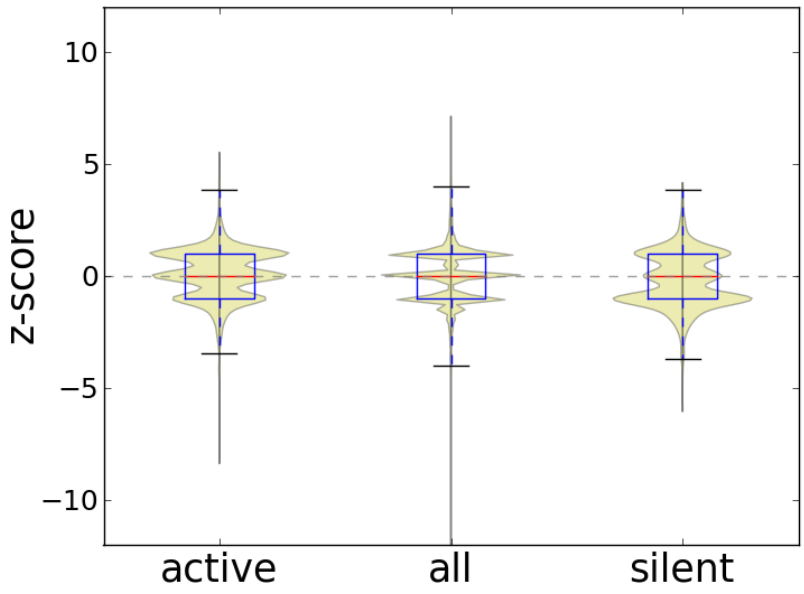
Supplementary Figure 8B. Relative K562 RNA levels of active enhancers and silent enhancers for **all RNA** (RNA total) types.



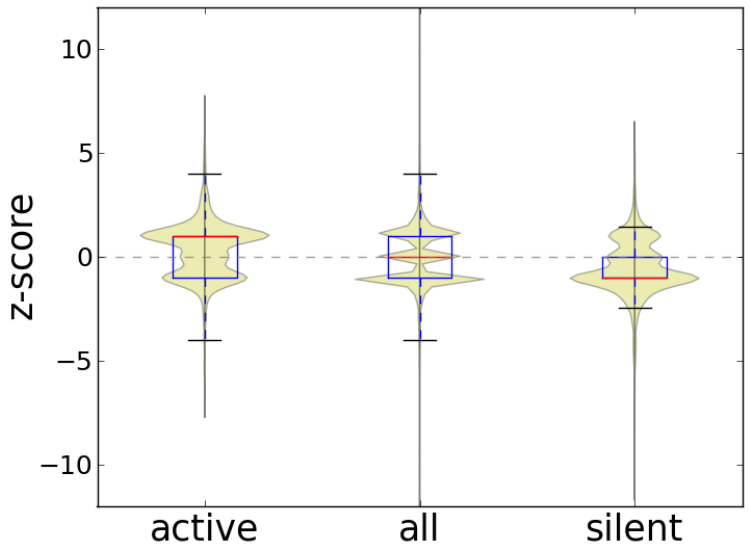
Supplementary Figure 8C. Cytosolic polyA+ eRNAs. Relative K562 RNA levels of active and silent enhancers for cytosolic polyA+ levels.



Supplementary Figure 8D. Cytosolic polyA- eRNAs. Relative K562 RNA levels of active and silent enhancers for cytosolic polyA- levels.

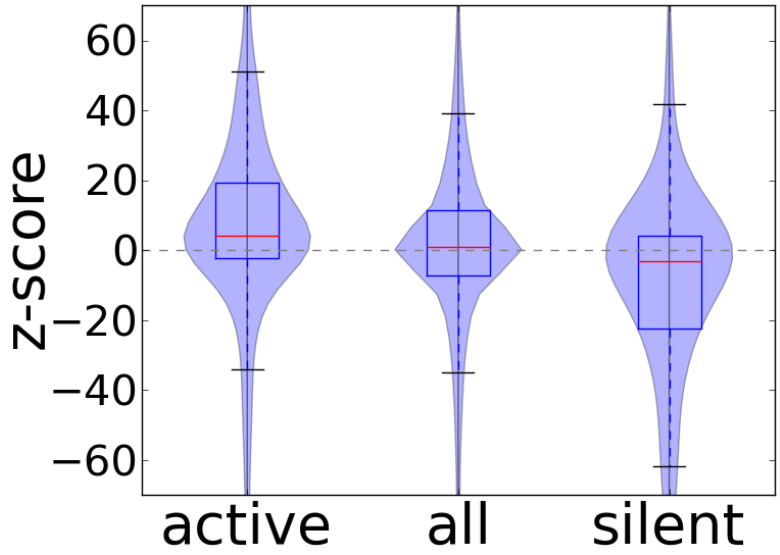


Supplementary Figure 8E. Short nuclear RNA. Relative K562 RNA levels of active and silent enhancers for short nuclear (<200nt) RNA.

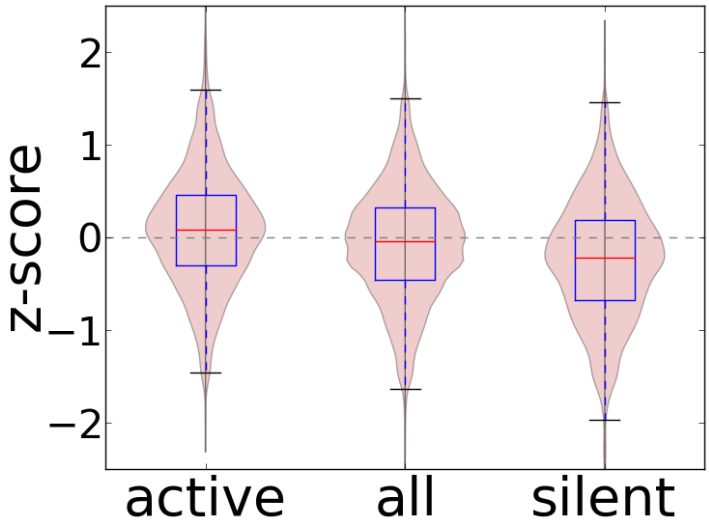


Supplementary Figure 8F. Short Total RNA. Relative K562 RNA levels of active and silent enhancers for short nuclear (<200nt) RNA.

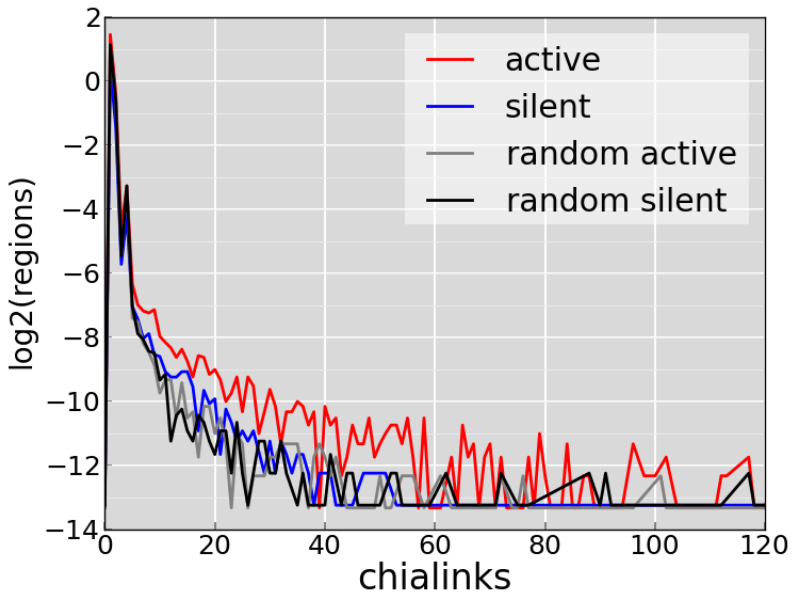
Association of intergenic enhancers and TSS activity



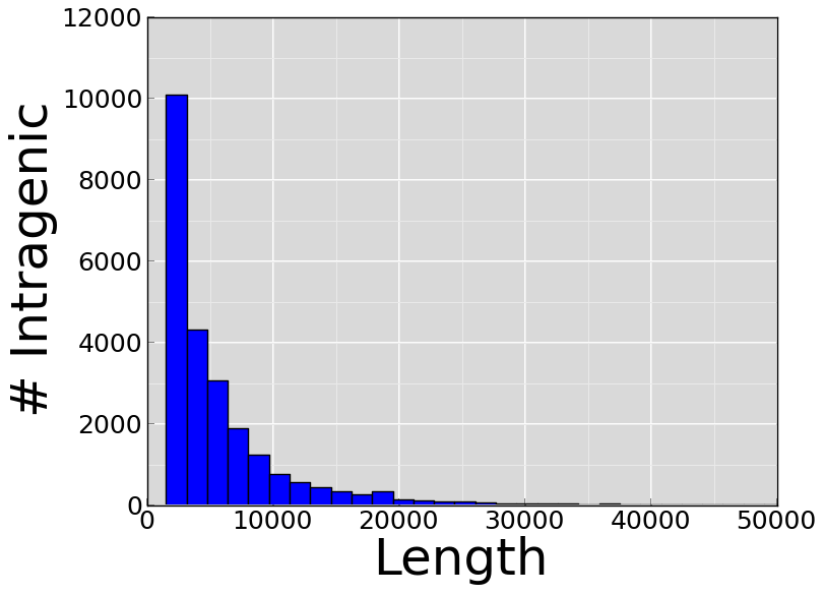
Supplementary Figure 9A. Relative expression change for genes associated to activated and silenced enhancers. The Y axis measures the relative expression change in terms of a Z-score for genes associated to activated (left violin plot) and silent enhancers (right violin plot), and for all genes (middle violin plots).. Genes were associated to the nearest predicted enhancer within a range of 10kb to 100kb from the TSS on either direction..



Supplementary Figure 9B. RNAPII activity associated to active and silent enhancers. The Y axis measures the relative change in RNAPII density in a 1kb window around the TSS in terms of a Z-score for genes associated to activated (left violin plot) and silent enhancers (right violin plot), and for all genes (middle violin plots). Genes were associated to the nearest predicted enhancer within a range of 10kb to 100kb from the TSS on either direction.

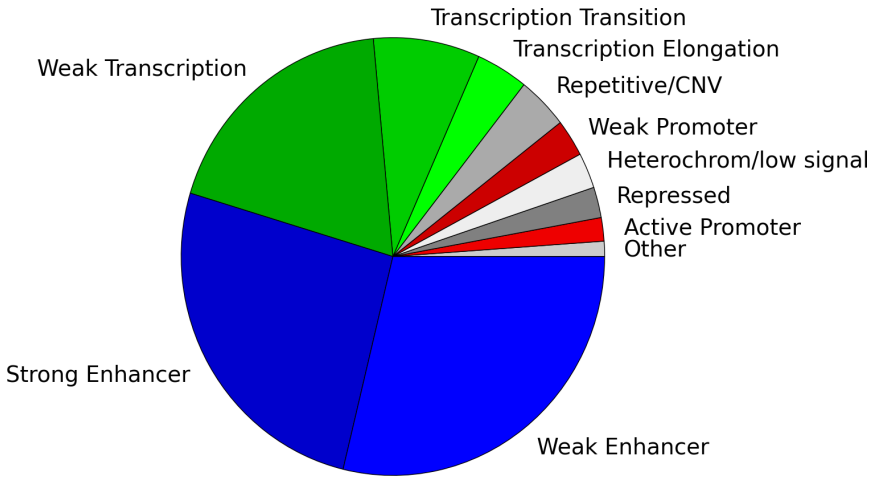


Supplementary Figure 9C. Percentage of intergenic predicted enhancers linked by ChIA-PET to a nearby TSS**.** In the Y-axis we plot in log2-scale the fraction of regions with ChIA-PET links to a nearby TSS, for activated, silenced, random activated and random silenced enhancers. TSS – enhancer pairs are considered when all elements are located at least 3 kilobases away and as far as 100 kilobases.

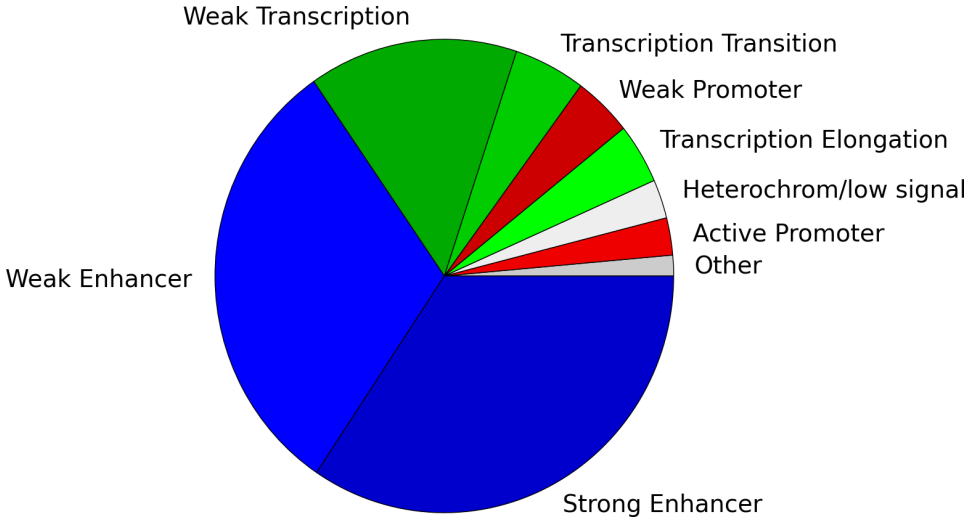


Supplementary Figure 10 - Length distribution of intragenic enhancers

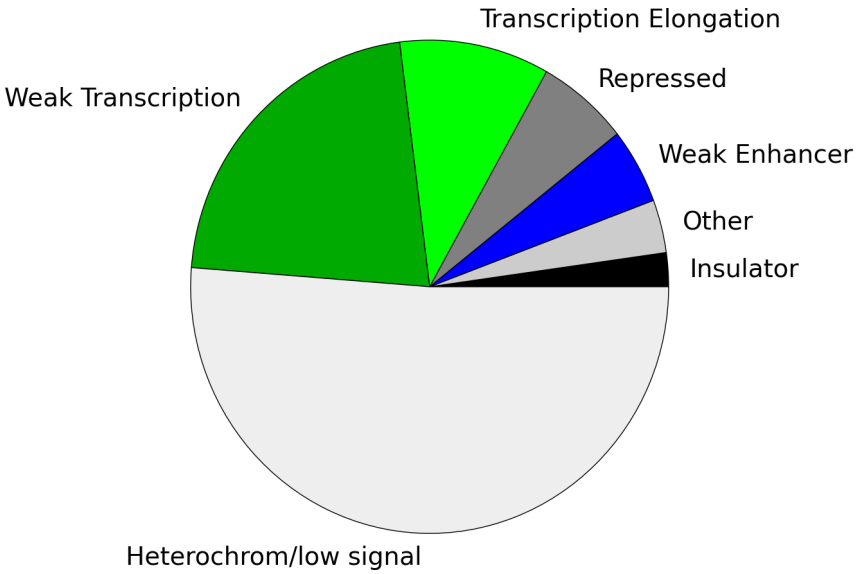
Comparison of our intragenic predictions with ChromHMM predictions



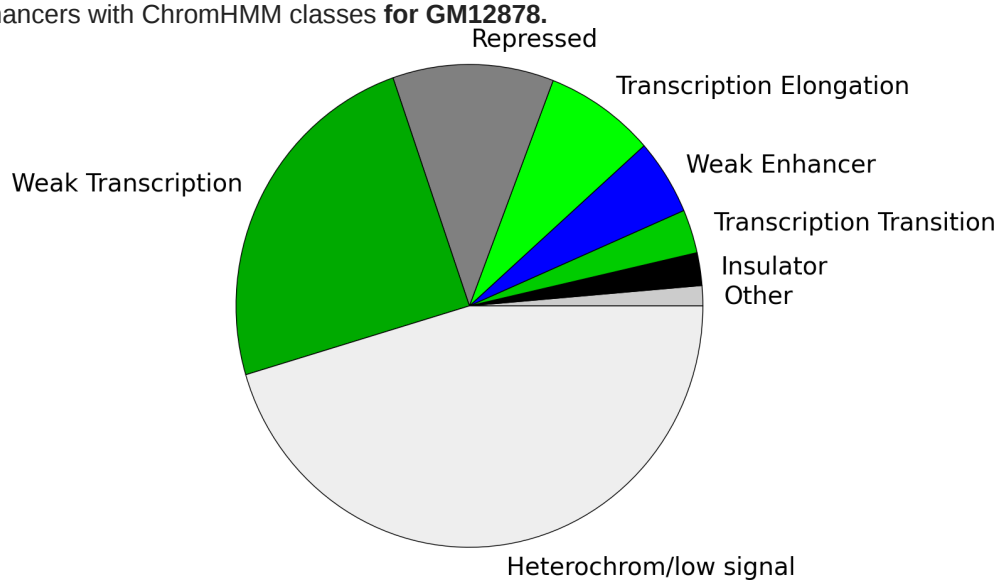
Supplementary Figure 11A. Pie chart showing the overlap of our **active intragenic** enhancers with ChromHMM classes **for K562.**



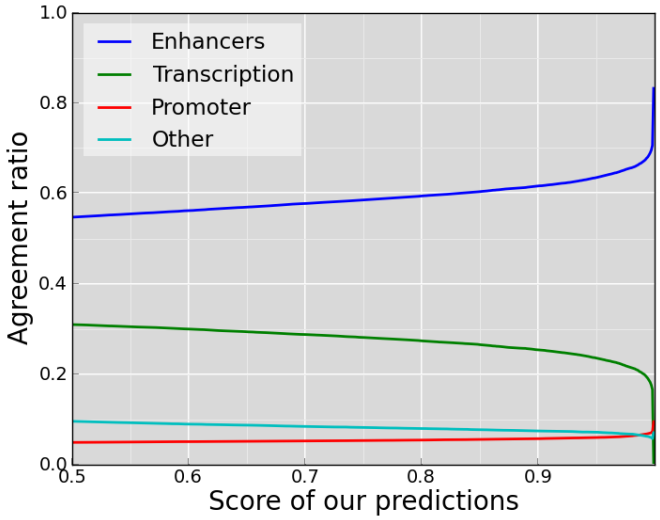
Supplementary Figure 11B. Pie chart showing the overlap of our **silent intragenic** enhancers with ChromHMM classes **for GM12878.**



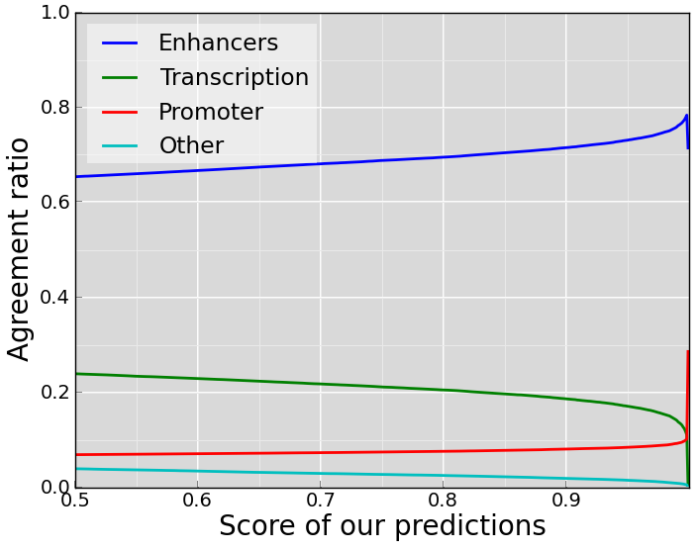
Supplementary Figure 11B. Pie chart showing the overlap of our **active intragenic** enhancers with ChromHMM classes **for GM12878.**



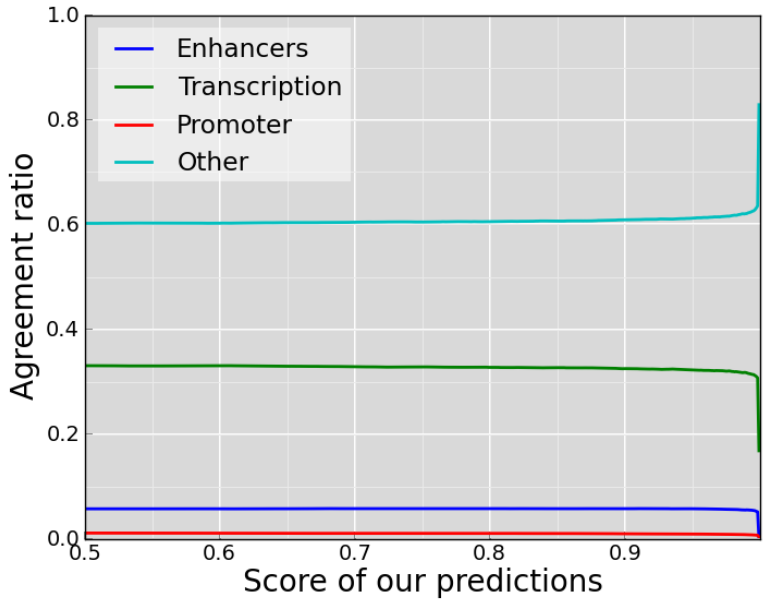
Supplementary Figure 11C. Pie chart showing the overlap of our **silent intragenic** enhancers with ChromHMM classes **for K562.**



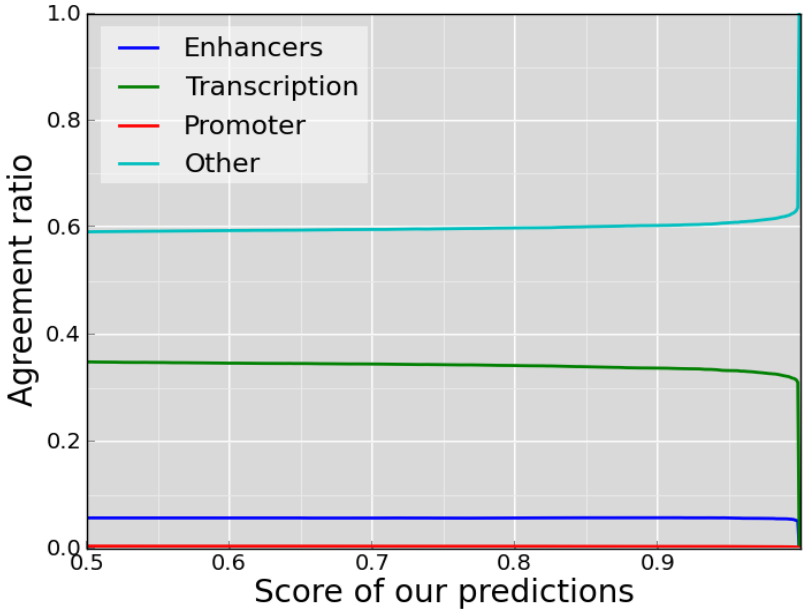
Supplementary Figure 11E – Agreement ratio evolution between of our intragenic **active** predictions and ChromHMM predictions **in K562**. For legend explanation see Supplementary Figure 7.



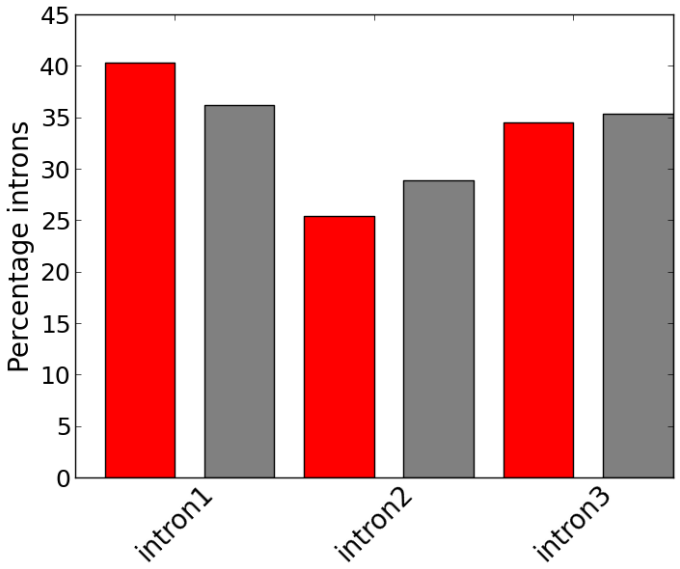
Supplementary Figure 11F – Agreement ratio evolution of our intragenic **silent** predictions with ChromHMM predictions i**n GM12878.**



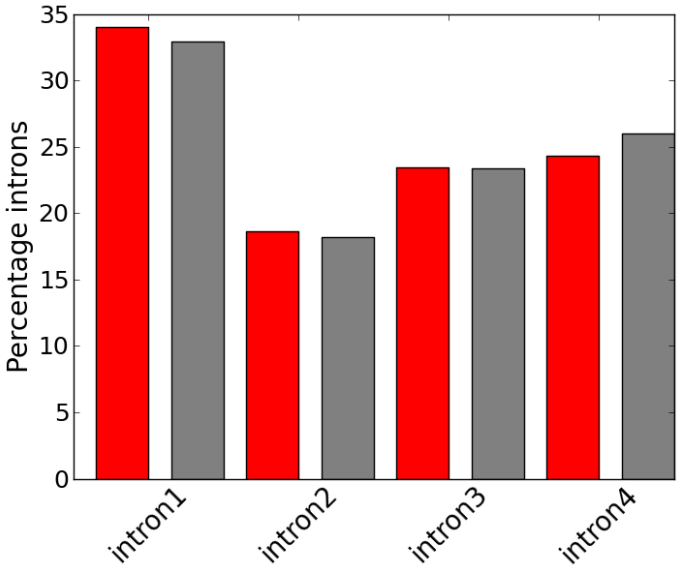
Supplementary Figure 11G – Agreement ratio evolution of our **intragenic active** predictions with ChromHMM predictions **in GM12878.**



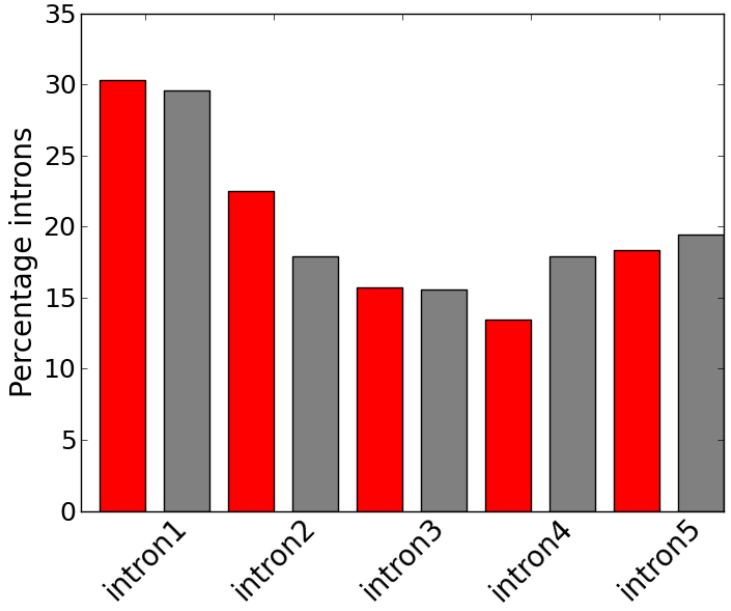
Supplementary Figure 11H – Agreement ratio evolution of our **intragenic silent** predictions with ChromHMM predictions **in K562.**



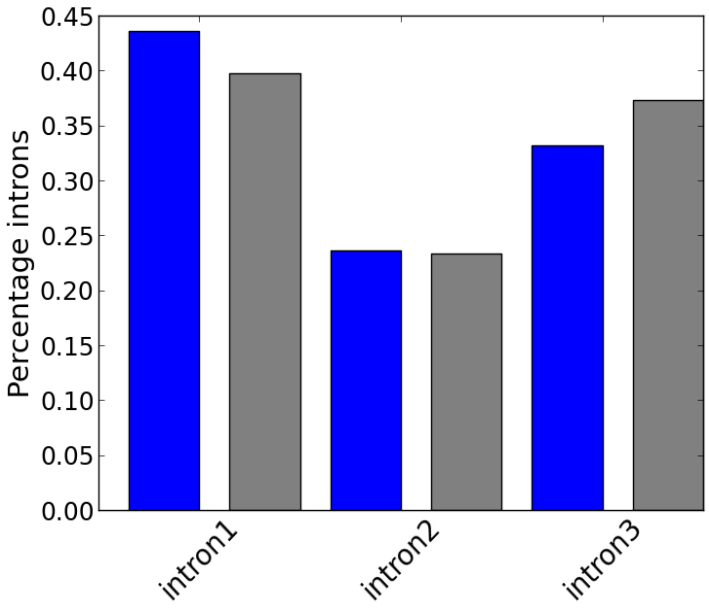
Supplementary Figure 12A – Positioning of intronic **active** putative enhancers. This barplot shows the percentage of introns in the first, second and third exon for active enhancers and randomized “active” positions in genes with exclusively **3 introns.** Red bars are for intragenic active enhancers, grey bars randomized intragenic “active” positions.



Supplementary Figure 12B – Positioning of intronic **active** putative enhancers. Positioning of elements in genes with exclusively **4 introns.** Red bars are for intragenic active enhancers, grey bars randomized intragenic “active” positions.

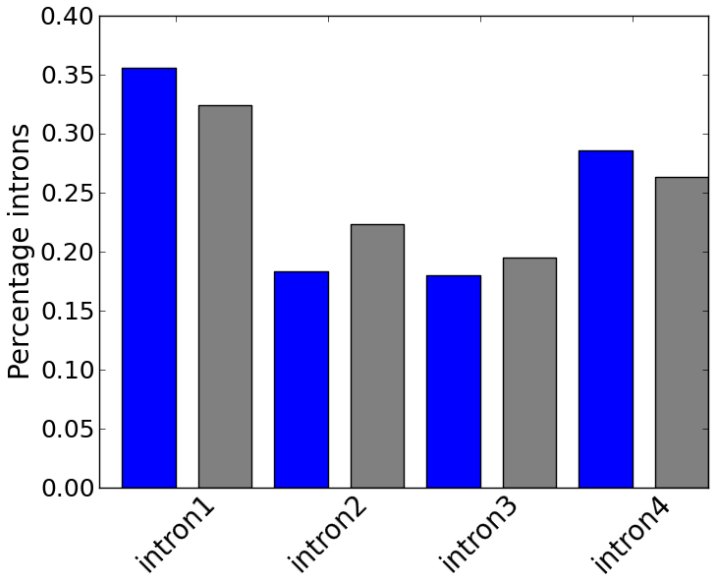


Supplementary Figure 12C – Positioning of intronic **active** putative enhancers. Positioning of elements in genes with exclusively **5 introns.** Red bars are for intragenic active enhancers, grey bars randomized intragenic “active” positions.

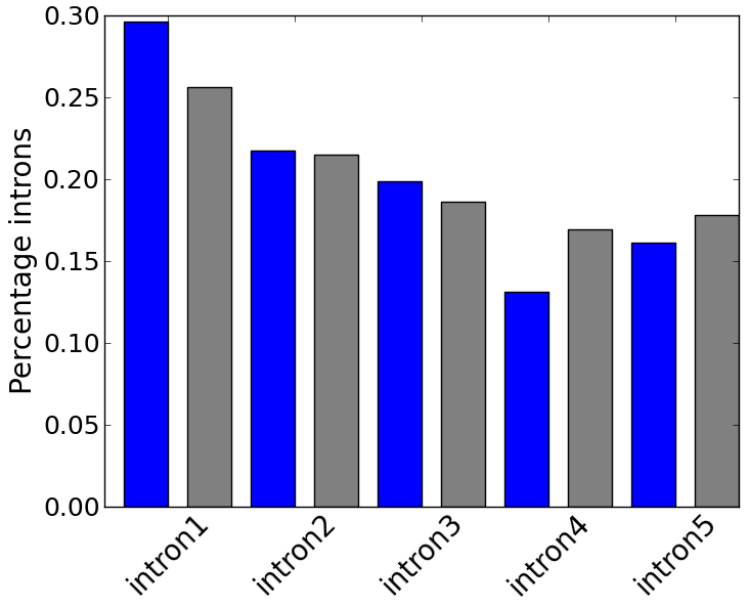


Supplementary Figure 12D – Positioning of intronic **silent** putative enhancers. Positioning of

elements in genes with exclusively 3 introns. Red bars are for intragenic active enhancers, grey bars randomized intragenic "silent" positions.

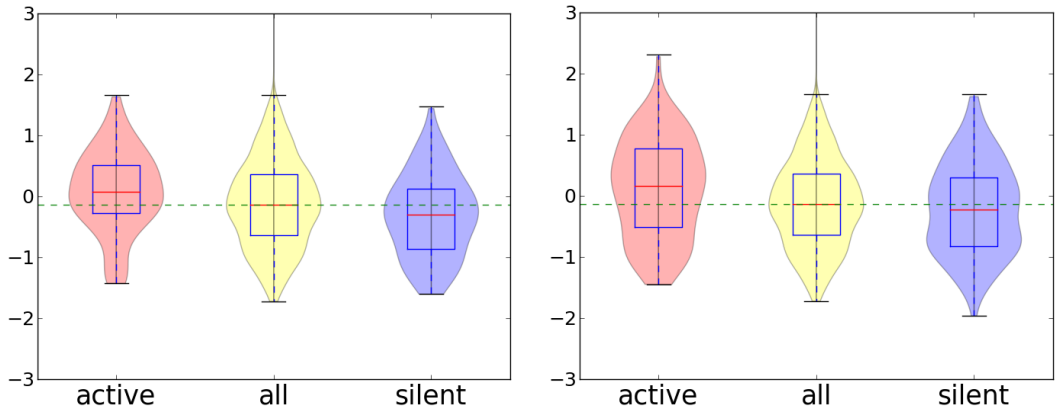


Supplementary Figure 12E – Positioning of intronic **silent** putative enhancers. Positioning of elements in genes with exclusively 4 introns. Red bars are for intragenic active enhancers, grey bars randomized intragenic "silent" positions.

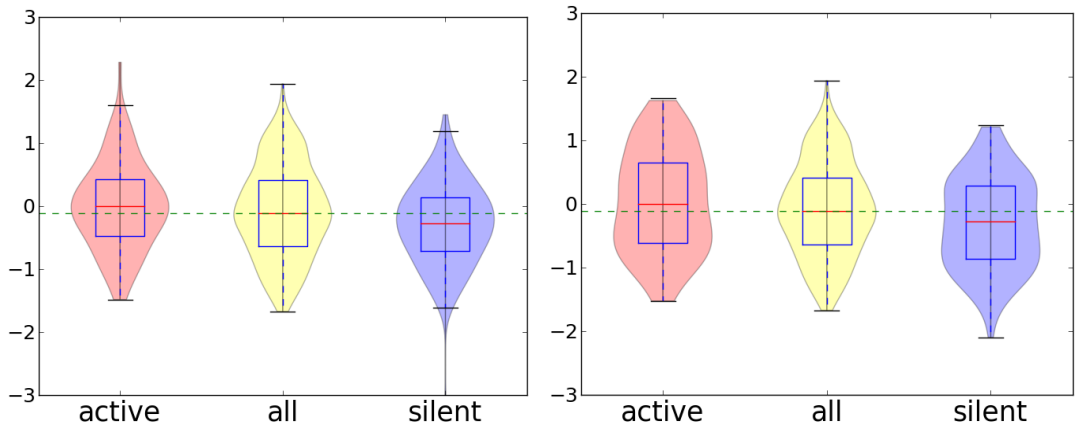


Supplementary Figure 12F – Positioning of intronic **silent** putative enhancers. Positioning of

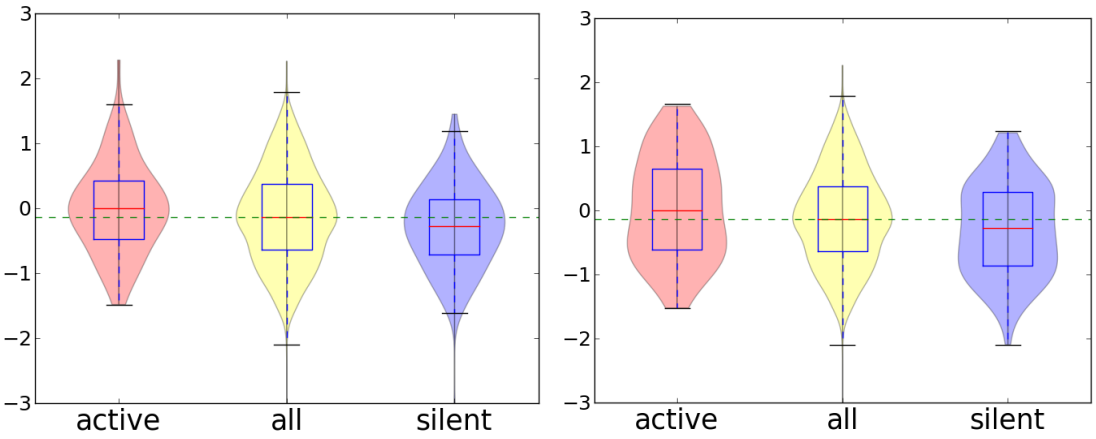
elements in genes with exclusively **5 introns.** Red bars are for intragenic active enhancers, grey bars randomized intragenic "silent" positions.



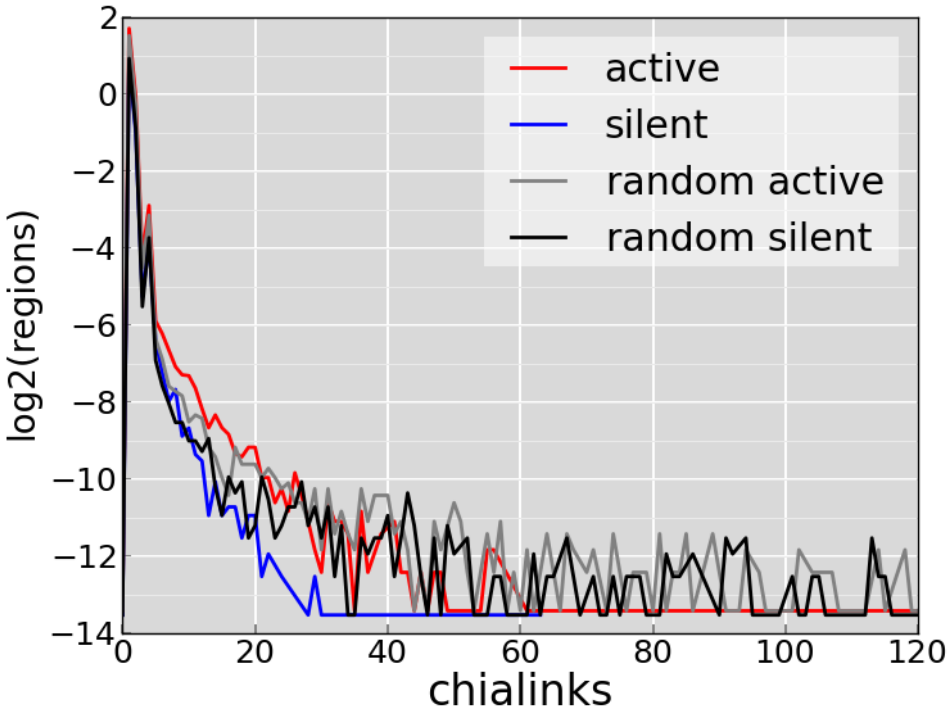
Supplementary Figure 13A – TSS levels of nearby enhancers for event 2. The left and rights figures corresponds to TSS1 and TSS2 respectively. The observed effect in Figure 3 is still present.



Supplementary Figure 13B – TSS levels of nearby enhancers for event 3.

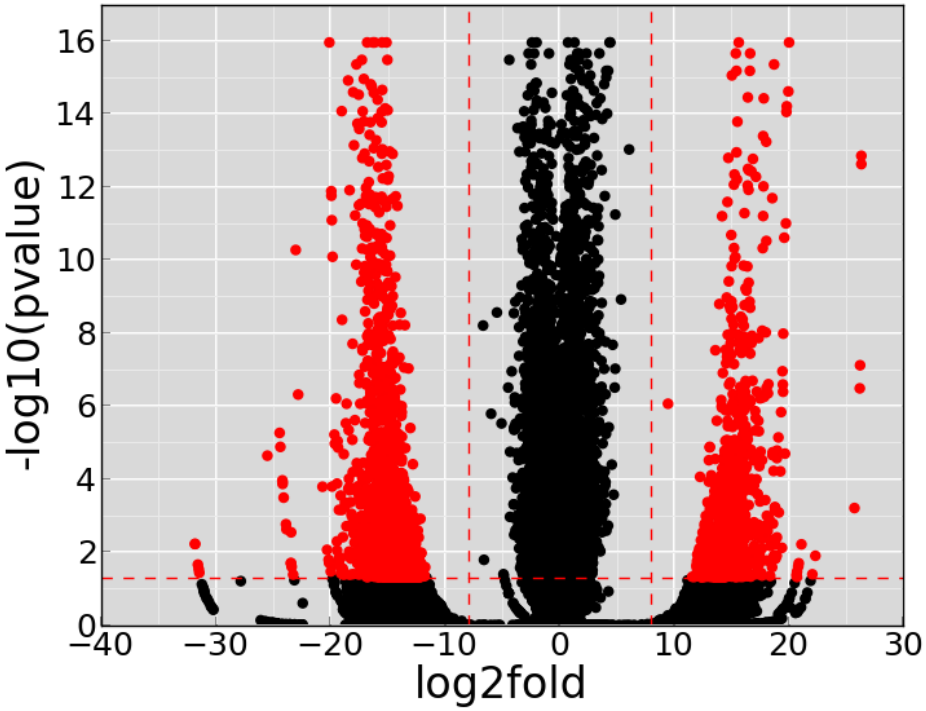


Supplementary Figure 13C – TSS levels of nearby enhancers for event 4.



Supplementary Figure 14 - ChIA-PET links for intragenic enhancers. In the Y-axis we plot in log2-scale the fraction of regions with ChIA-PET links to a nearby TSS, for activated, silenced, random activated and random silenced enhancers. TSS – enhancer pairs are considered

when all elements are located at least 3 kilobases away and as far as 100 kilobases.



Supplementary Figure 15 - Vulcano plots/fold enrichment distribution with cutoffs – We consider significant every exonic region with an absolute log2fold bigger than 8 and a p-value < 0.05 (-log10(p-value) = 1.3)

Supplementary Methods

Protocol	Dataset type	Experiments
ChIP-Seq	Histone Modifications	**H3K4me1**, **H3K4me2**, **H3K4me3**, **H3k9ac**, **H3K27ac**, H3K27me3, H3K36me3, H3K79me2, H4K20me1
	Transcription Factors	CTCF, EZH2, **P300**, Pol2, PU.1, Stat1
	Histone Variant	**H2A.Z**

SupplementaryTable 1. ChIP-Seq datasets from the ENCODE project we used in our analysis. We downloaded the BAM files with the mapped reads for both cell-lines K562 and GM12878. The datasets in bold are the ones used in the final model, after the feature selection study.

Feature selection using Boruta

As a check for the feature selection, we run the Boruta algorithm with the same parameters but using P300 as the correlation class. The results (Figure 1B) validate the enhancer activation model already proposed where P300 is an ubiquitous mark in enhancers in all states. Excluding the technical and biological controls and H3K20me1, the rest of the features seemed to be at least marginally relevant when correlated to P300. Running the selection algorithm with the H3K4me1 mark the biological control average Boruta score increased notably, suggesting that the mark is present in many repetitive regions along the genome, normally found to be “noisy” in ChIP-Seq experiments.

Randomized positions

For control of the experiments, we calculated 4 sets of randomized positions (intergenic/intragenic and active/silent putative predictions). These set where calculated keeping the same length, number of elements and restrictions of their counterparts. These restrictions are:

- In the case of intergenic elements, the randomized positions do not overlap any annotated

element, and their security distance from TSS (Supplementary Figure 2A)

- In the case of the intragenic elements, the randomized positions are inside intragenic elements, and do not overlap with TSS and their security distance (Supplementary Figure 2B)

Proportion of intragenic enhancers in introns and exons compared to proportion of exon and intron lengths

Counting the percentage of active and silenced enhancers nucleotides overlap with exons we found that 8.1% and 6.6% bases where exonic respectively. We then counted the number of exonic nucleotides covered by the GENCODE.V7 annotation also discarding around 1kb around TSS, to make the comparison fair with the putative enhancers and the percentage of exonic bases was 9.18%

First introns are longer on average

We calculated the average of every intron separating genes in the GENCODE.V7 annotation by number of introns, from 2 to 7 introns. The only case where the first intron doesn't seem to be longer in the 2 introns genes group.

```
Total genes with 2 intron(s): 5392
Intron 1 avg. length: 7263.61183234
Intron 2 avg. length: 7199.96735905

Total genes with 3 intron(s): 3089
Intron 1 avg. length: 7832.92651343
Intron 2 avg. length: 4888.83360311
Intron 3 avg. length: 6798.17578504

Total genes with 4 intron(s): 2296
Intron 1 avg. length: 7814.87108014
Intron 2 avg. length: 4712.29790941
Intron 3 avg. length: 4521.63719512
Intron 4 avg. length: 6246.63240418

Total genes with 5 intron(s): 1863
Intron 1 avg. length: 6881.82930757
Intron 2 avg. length: 5561.91250671
Intron 3 avg. length: 4552.50187869
Intron 4 avg. length: 4049.90982287
Intron 5 avg. length: 5628.71604938

Total genes with 6 intron(s): 1554
Intron 1 avg. length: 6973.46718147
Intron 2 avg. length: 5309.77284427
```

```
Intron 3 avg. length: 4375.55984556
Intron 4 avg. length: 4445.78635779
Intron 5 avg. length: 4006.58687259
Intron 6 avg. length: 4861.94144144

Total genes with 7 intron(s): 1290
Intron 1 avg. length: 6926.83333333
Intron 2 avg. length: 5174.78914729
Intron 3 avg. length: 4665.16744186
Intron 4 avg. length: 4541.41782946
Intron 5 avg. length: 3769.95968992
Intron 6 avg. length: 3817.72868217
Intron 7 avg. length: 4435.38294574
```

References

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011 May 5;473(7345):43–9.

mclust: an R package for normal mixture modeling [Internet]. [cited 2013 Apr 20]. Available from: http://www.stat.washington.edu/mclust/

Discussion

Biases in HTS datasets

For many HTS experiments technical biases remain as one of the biggest challenges to overcome. Ultimately, if the biases mentioned cannot be corrected at the source (the experimental side) and considering there is not always an appropriate control experiment, it is sometimes not possible to fully model the distribution of the reads. It has been shown that different RNA-Seq datasets follow approximately different counts distributions and therefore they can not be determined a priori[37].

A solution to this problem consists in using multiple biological replicas, to separate the signal from the biological variability. While obviously more expensive than using a control experiment, will probably become the standard in future HTS experiments as sequencing technology becomes more affordable and the technology matures.

Runtime optimization vs memory efficiency

A software analysis pipeline can be reduced from days to hours with the appropriate optimization techniques. It is a common misconception that choosing a low-level programming language like C is the straightforward solution for speed optimization. While it is true that low-level programming languages perform better on average than programming languages that have more functionality like garbage collection (to manage memory automatically) and dynamic typing (allowing for variables to hold more than one type), the time invested by a programmer in order to get the desired results may not be worth the effort in many cases. Ultimately, programming

languages are just tools, and the appropriate tool must be used for the appropriate problem. Furthermore, modern Just in Time (JiT) compilers that generate native machine code on execution time are dramatically optimizing runtime for popular high level programming languages like Java and Python, closing the performance gap.

During my thesis work, I tried re-coding critical parts of the Pyicoteo library in C and Cython (a variation of Python that translates directly to C code). There were some marginal improvements of performance, but the problems were not actually there. In many cases the critical performance bottlenecks were more mundane than that, maybe an inefficient loop or a program that checks too often into a database. By doing some software benchmarking you may find out that the problem is a vector that in some unusual circumstances gets too big. This is one of the main side effects of mapped reads saturation, and one of the most recurring problems that I found when working with HTS data. For example, it has been quite common to encounter HTS datasets that may have around 100 reads on average per genomic region, but an outlier region (normally artifactual, but not necessarily) may contain 2 million reads. This lead to huge delays, since the algorithm was designed to deal with a few thousand reads at a time. The solution did not involve re-coding the algorithm in a lower level language, but actually doing algorithm profiling to first identify the problem (this was not trivial) and then implementing a cache system for duplicated reads, which in a relatively short genomic region with high volume of reads happens quite often. This is why the algorithm implementation and optimization are so important.

Some of the computations in the Pyicoteo suite could be parallelized. One performance optimization that remains to be done is parallelizing some of these computations.

Memory efficiency vs runtime trade-off

We designed Pyicoteo with the principle of trying to be as memory efficient as possible. Memory efficiency allows for many more instances of Pyicoteo running concurrently and allows for analysis in smaller machines. However, there are two shortcomings of this implementation approach, which are disk usage and execution time. Pyicoteo makes heavy use of temporary files and can sometimes double the size of the datasets used. Slow access to disk can be a problem, specially if the software is running in a distributed system where the disks are saturated or further away from each other. In any case,

memory is normally smaller than disk, modern disks have terabytes of space and disk access technology is fast and reliable. memory is also getting bigger, but it will always be smaller than disk. The runtime trade-off is another inevitability of focusing on memory efficiency but at least in the problems I have focused on solving, it has not been a critical issue. As discussed above, the main bottlenecks I encountered where mostly algorithmic in nature.

Limitations of active enhancer detection using relative enrichment between two cell lines

Our method for prediction of active enhancers has some limitations. First, comparing two cell lines can only give limited results, since enhancers are known to be activated and silenced differently not only between tissues, but also between different stages of the cell cycle. While K562 (leukemia) and GM12878 (lymphoblastoid) are similar tissues, having time series data and more replicas will surely improve the outcome of studies like this one. Also, it has to be noted that HTS data is relatively new and that HTS technology still has problems that need to be overcome. Furthermore, while our method is interesting because we are able to see relative activation and silencing between two conditions and, unlike other methods, we are able to see relative activation, our method is blind to enhancers activated or silenced in both conditions. Finally, enhancer activation is still poorly understood, and a model like ours could help revealing more pieces of this puzzle.

Because we wanted to prove that histone modifications and transcription factors are sufficient for enhancer activation prediction, we did not take into consideration in our model other properties known to be associated to enhancer activation like DNA methylation, Dnase I sensitivity and sequence conservation. In fact, some of these properties are not necessarily specific of enhancers, as enhancers for example are not necessarily conserved and not all Dnase I sensitive sites are active enhancers.

Conclusions

The contributions of this thesis can be summarized as follows:

- I implemented a flexible tool for the analysis of HTS data (Pyicoteo), and used it for analyzing data in multiple scientific collaborations. The tool is open source and publicly available and is being used by the scientific community for the analysis of HTS data. I estimate it is being routinely used in the projects of our group and at least 3 or 4 more groups. It has been used in 4 published works, and there are at least another 6 in the publishing pipeline in different groups.
- The flexibility of Pyicoteo allows for it to be used in many different ways:
  - As a python library in custom scripts, as a genomics coordinates manipulation package
  - As a file manipulator and format converter
  - As a for *peak calling* for ChIP-Seq and CLIP-Seq
  - As a tool for measuring relative changes between two conditions with and without replica
- Pyicoteo is an ongoing project and it has been released as an Open Source project. The objective is to turn it into a useful tool for the community.
- Using a method that measures the relative enrichment and depletion of multiple signals coming from ChIP-Seq experiment and using a semi-supervised clustering approach I predicted enhancers that are active relative to a different condition or cell type.

- Active enhancers produce enhancer RNA (eRNA) of different types. Nuclear RNA seems to be both polyadenilated and not polyadenilated, but also there seems to be cytosolic polyadenilated RNA coming from enhancer sites. RNA produced also seems to be mostly
- Looking at the relative changes between leukemia and lymphoblastoid cell lines, we observe possible activation of cancer related genes and enhancers.
- The activation or silencing of intragenic transcriptional enhancers can modulate the expression of the host genes as well as the splicing regulation of nearby exons, likely by modifying the local state of the chromatin.

Appendix A

Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling

For this collaboration, I was part of the analysis of Hog1 and RNAPII ChIP-Seq datasets. We carried out relative enrichment analysis for the genome-wide localization study of RNAPII and Hog1. We showed that stress in Yeast cells reduces RNAPII levels in housekeeping genes and is enriched in stress-responsive genes. We also observed that Hog1 was required for this effect to happen, as we also found it in stress-responsive genes.

The article was published at:

Nadal-Ribelles M, Conde N, Flores O, González-Vallinas J, Eyras E, Orozco M, et al. Hog1 bypasses stress-mediated down-regulation of transcription by RNA polymerase II redistribution and chromatin remodeling. Genome Biol. 2012 Nov 18;13(11):R106.

Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3580498/

My contribution

ChIP reads were extended by 250 bp (a bit above fragment size) and normalized between sample and control using Pyicos. Enrichment of Hog1 and RNA Pol II was done by running the Pyicos enrichment protocol comparing untreated to treated samples. Hog1-dependent RNA Pol II recruitment was determined by comparing salt-treated samples with wild-type and hog1 strains. For all comparisons, enrichment was considered significant for a Z-score > 4 (p-value = 0.0001). MA plots were done using the Pyicos software, where M represents the log ratio of stressed versus non-stressed (y-axis) and A is the average of the log intensities (x-axis) of all the genes, for enrichment of Hog1 and RNA Pol II.

The percentage of nucleosome occupancy was determined by running the extension and normalization (Using TRPK) protocols using the first 200 bp downstream of the TSS, which encompasses the +1 nucleosome for all the gene clusters, using Pyicos.

Appendix B

Use of ChIP-Seq data for the design of a multiple promoter-alignment method

Pro-Coffee is a multiple aligner specifically designed for homologous promoter regions. I collaborated in the design of a novel validation procedure based on multi-species ChIP-seq dataset for different species of vertebrates by doing the ChIP-Seq mapping and peak detection in Human, Mouse, Dog and Chicken for transcription factors HNF4A and CEBPA.

The article was published at:

Erb I, González-Vallinas JR, Bussotti G, Blanco E, Eyras E, Notredame C. Use of ChIP-Seq data for the design of a multiple promoter-alignment method. Nucleic Acids Res. 2012 Apr;40(7):e52.

The online reference can be found at:

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3326335/#gkr1292-B4

My contribution

I downloaded the raw HTS reads coming from Illumina sequencing produced by Schmidt et al.[60] (http://www.ebi.ac.uk/arrayexpress/files/E-TABM-722/E-TABM-722.idf.txt). I constructed the assembly indexes for human (hg18), mouse (mm9), dog (canFam2) and chicken (galGal3) using

GEM index[43] (http://gemlibrary.sourceforge.net) and run the mapping of the data for both experiment and control sequence files using GEM allowing up to two mismatches, with default settings as quality filter.

I did the Peak Calling using Pyicos, in each case extending reads to the value calculated by the strand correlation algorithm, and using the operations of normalization, subtraction of control and filtering with a Poisson test based on the height of the peaks. We selected the p-value cut off based on benchmarking with other methods. Finally, I obtained the binding regions by centering regions of 100 nucleotides over the genomic coordinates of significant peaks.

Appendix C

Mapping of HITS-CLIP data to the mouse genome

I did the mapping and peak calling of CLIP-Seq data for the analysis of G3BP protein.

Publication:

Sophie Martin, Juan González-Vallinas, Nicolas Bellora, Manuel Irimia, Gracja Michlewski, Monika Raabe, Eduardo Eyras, Henning Urlaub, Ben Blencowe, Jamal Tazi, Javier Caceres. HITS-CLIP of G3BP Reveals a Ribonucleoprotein Complex that Preferentially Binds to Intron-retaining Transcripts in Mouse Brain and Influences their Expression Level in the Cerebellum (Submitted to PLOS Genetics)

My contribution

For each sample, we filtered out reads shorter than 21 nucleotides to avoid ambiguous mapping locations, as any random sequence shorter than 21nt is highly likely to be found in the genome. We did a 2 step mapping, first using as reference the mouse mm9 genome (NCBI Build 37 assembly, July 2007) using Bowtie v0.12.7 (Langmead et al. 2009) allowing up to 2 mismatches. Furthermore, to avoid missing any tags from mature transcripts, split reads that could correspond to exon junctions were mapped using the GEM split mapper (build 592) from the GEM library (http://gemlibrary.sourceforge.net) using the following filters: same strand,

same chromosome, and maximum of 100,000 bases in distance, based on a distance estimation between exons in mm9. Significant clusters of mapped reads where identified using the modified FDR method for CLIP-Seq (Yeo et al 2009) as implemented in Pyicos software version 0.9.9.2 (Chapter 2) using default parameters.

Bibliography

[1] Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., and Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science (New York, N.Y.)*, 252(5013):1651–1656. PMID: 2047873.

[2] Alamancos, G. P., Agirre, E., and Eyras, E. (2013). Methods to study splicing from high-throughput RNA sequencing data. *arXiv:1304.5952*.

[3] Allfrey, V. G., Faulkner, R., and Mirsky, A. E. (1964). Acetylation and methylation of histones and their possible role in the regulation of rna synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 51:786–794. PMID: 14172992.

[4] Alló, M., Schor, I. E., Muñoz, M. J., Mata, M. d. l., Agirre, E., Valcárcel, J., Eyras, E., and Kornblihtt, A. R. (2010). Chromatin and alternative splicing. *Cold Spring Harbor Symposia on Quantitative Biology*, 75:103–111. PMID: 21289049.

[5] Ambriola, V., Bendix, L., and Ciancarini, P. (1990). The evolution of configuration management and version control. *Software Engineering Journal*, 5(6):303 –310.

[6] Atkins, D., Ball, T., Graves, T., and Mockus, A. (2002). Using version control data to evaluate the impact of software tools: a case study of the version editor. *IEEE Transactions on Software Engineering*, 28(7):625 – 637.

[7] Avery, O. T., Macleod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid frac-

tion isolated from pneumococcus type III. *The Journal of experimental medicine*, 79(2):137–158. PMID: 19871359.

[8] Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395. PMID: 21321607.

[9] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(Database issue):D991–995. PMID: 23193258.

[10] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837. PMID: 17512414.

[11] Bennett, S. (2004). Solexa ltd. *Pharmacogenomics*, 5(4):433–438.

[12] Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress,

M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*,

447(7146):799–816. PMID: 17571346.

[13] Boyd, S. D. (2013). Diagnostic applications of high-throughput DNA sequencing. *Annual review of pathology*, 8:381–410. PMID: 23121054.

[14] Boyle, A. P., Guinney, J., Crawford, G. E., and Furey, T. S. (2008). F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics (Oxford, England)*, 24(21):2537–2538. PMID: 18784119.

[15] Climate Code Foundation (2013). Science code manifesto. http://sciencecodemanifesto.org/.

[16] Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gómez Acuña, L. I., Fiszbein, A., Godoy Herz, M. A., Nieto Moreno, N., Muñoz, M. J., Alló, M., Schor, I. E., and Kornblihtt, A. R. (2013). Transcriptional elongation and alternative splicing. *Biochimica et biophysica acta*, 1829(1):134–140. PMID: 22975042.

[17] Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Dunham, I., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kelllis, M., Khatun, J., Kheradpour, P., Kundaje, A., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Green, E. D., Good, P. J., Feingold, E. A., Bernstein, B. E., Birney, E., Crawford, G. E., Dekker, J., Elinitski, L., Farnham, P. J., Gerstein, M., Giddings, M. C., Gingeras, T. R., Green, E. D., Guigó, R., Hardison, R. C., Hubbard, T. J., Kellis, M., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Snyder, M., Starnatoyannopoulos, J. A., Tennebaum, S. A., Weng, Z., White, K. P., Wold, B., Khatun, J., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Giddings, M. C., Bernstein, B. E., Epstein,

C. B., Shoresh, N., Ernst, J., Kheradpour, P., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Ward, L. D., Altshuler, R. C., Eaton, M. L., Kellis, M., Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H. P., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Risk, B. A., Robyr, D., Ruan, X., Sammeth, M., Sandu, K. S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T. J., Reymond, A., Antonarakis, S. E., Hannon, G. J., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T. R., Rosenbloom, K. R., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kent, W. J., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Furey, T. S., Song, L., Grasfeder, L. L., Giresi, P. G., Lee, B.-K., Battenhouse, A., Sheffield, N. C., Simon, J. M., Showers, K. A., Safi, A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Kim, S. K., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Birney, E., Iyer, V. R., Lieb, J. D., Crawford, G. E., Li, G., Sandhu, K. S., Zheng, M., Wang, P., Luo, O. J., Shahab, A., Fullwood, M. J., Ruan, X., Ruan, Y., Myers, R. M., Pauli, F., Williams, B. A., Gertz, J., Marinov, G. K., Reddy, T. E., Vielmetter, J., Partridge, E. C., Trout, D., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., King, B., Muratet, M. A., Antoshechkin, I., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Gunter, C., Newberry, J. S., Levy, S. E., Absher, D. M., Mortazavi, A., Wong, W. H., Wold, B., Blow, M. J., Visel, A., Pennachio, L. A., Elnitski, L., Margulies, E. H., Parker, S. C. J., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Chrast, J.,

Davidson, C., Derrien, T., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Howald, C., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Kokocinski, F., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tanzer, A., Tapanari, E., Tress, M. L., van Baren, M. J., Walters, (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74. PMID: 22955616.

[18] ENCODE Consortium (2013). ENCODE ChIP-seq experiment matrix. http://encodeproject.org/ENCODE/dataMatrix/encodeChipMatrixHuman.html.

[19] European Union (2013). Marie curie actions - promoting excellence. http://ec.europa.eu/research/fp6/mariecurie-actions/action/excellence_en.html.

[20] Fonseca, N. A., Rung, J., Brazma, A., and Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177.

[21] Furey, T. S. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, 13(12):840–852.

[22] Gentleman, R. C., Carey, V. J., Bates, D. M., 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. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80. PMID: 15461798.

[23] Git Development Team (2013). Git. http://git-scm.com/.

[24] Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86. PMID: 20738864.

[25] Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics (Oxford, England)*, 13(2):204–216. PMID: 22285995.

[26] Hutchison, Clyde A, r. (2007). DNA sequencing: bench to bedside and beyond. *Nucleic acids research*, 35(18):6227–6237. PMID: 17855400.

[27] IEEE (2013). Software engineering body of knowledge (SWEBOK). http://www.computer.org/portal/web/swebok.

[28] International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945. PMID: 15496913.

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

[30] Kidder, B. L., Hu, G., and Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. *Nature immunology*, 12(10):918–922. PMID: 21934668.

[31] Kim, T. H. and Ren, B. (2006). Genome-wide analysis of protein-DNA interactions. *Annual Review of Genomics and Human Genetics*, 7(1):81–102. PMID: 16722805.

[32] Knuth, D. E. (1992). *Literate Programming*. Center for the Study of Language and Information Publica Tion.

[33] Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M. J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews Molecular Cell Biology*, 14(3):153–165.

[34] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L.,

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

409(6822):860–921. PMID: 11237011.

[35] Latchman, D. S. (1997). Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology*, 29(12):1305–1312.

[36] Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., and Wolf, S. (1999). A brief history of the internet. *arXiv:cs/9901011*.

[37] Li, J. and Tibshirani, R. (2011). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical methods in medical research*. PMID: 22127579.

[38] Liang, K. and Keleş, S. (2012). Normalization of ChIP-seq data with control. *BMC Bioinformatics*, 13(1):199. PMID: 22883957.

[39] Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469. PMID: 18978773 PMCID: PMC2597294.

[40] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012. PMID: 22829749 PMCID: PMC3398667.

[41] Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–260. PMID: 9305837.

[42] Maik Roder and Julien Lagarde (2013). ENCODE RNA dashboard. http://genome.crg.es/encode_RNA_dashboard/hg19/.

[43] Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, 9(12):1185–1188.

[44] Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564. PMID: 265521.

[45] Merali, Z. (2010). Computational science: ...Error. *Nature News*, 467(7317):775–777.

[46] Mukhopadhyay, A., Deplancke, B., Walhout, A. J. M., and Tissenbaum, H. A. (2008). Chromatin immunoprecipitation (ChIP) coupled to detection by quantitative real-time PCR to study transcription factor binding to DNA in caenorhabditis elegans. *Nature Protocols*, 3(4):698–709.

[47] Mutz, K.-O., Heilkenbrinker, A., Lönne, M., Walter, J.-G., and Stahl, F. (2013). Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology*, 24(1):22–30. PMID: 23020966.

[48] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, 320(5881):1344–1349. PMID: 18451266.

[49] NCBI (2013). BLAST (basic local alignment search tool). http://blast.ncbi.nlm.nih.gov/.

[50] Noll, M. (1974). Subunit structure of chromatin. *Nature*, 251(5472):249–251.

[51] Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772.

[52] Pierce (2005). *Genetics: A conceptual approach*. W. H. Freeman.

[53] Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine*, 8(1):7.

[54] Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651–657. PMID: 17558387.

[55] Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry*, 242(1):84–89. PMID: 8923969.

[56] SANGER, F. (1959). Chemistry of insulin; determination of the structure of insulin opens the way to greater understanding of life processes. *Science (New York, N.Y.)*, 129(3359):1340–1344. PMID: 13658959.

[57] Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–448. PMID: 1100841.

[58] scfbm (2013). Source code for biology and medicine. http://www.scfbm.org/.

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

[60] Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (New York, N.Y.)*, 328(5981):1036–1040. PMID: 20378774.

[61] Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. H., and Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679.

[62] Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121. PMID: 11309499.

[63] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K.,

Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–1351. PMID: 11181995.

[64] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

[65] Wren, J. D. (2004). 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics (Oxford, England)*, 20(5):668–672. PMID: 15033874.

[66] Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X.-D., and Gage, F. H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature structural & molecular biology*, 16(2):130–137. PMID: 19136955.

[67] Zentner, G. E. and Henikoff, S. (2012). Surveying the epigenomic landscape, one base at a time. *Genome biology*, 13(10):250. PMID: 23088423.

[68] Zentner, G. E. and Scacheri, P. C. (2012). The chromatin fingerprint of gene enhancer elements. *The Journal of biological chemistry*, 287(37):30888–30896. PMID: 22952241.

[69] Zhang, Z. and Pugh, B. F. (2011). High-resolution genome-wide mapping of the primary structure of chromatin. *Cell*, 144(2):175–186. PMID: 21241889.