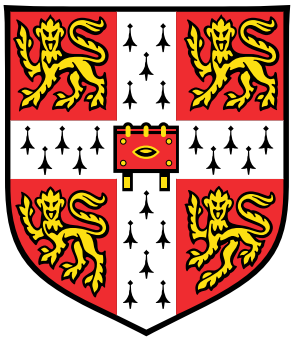
Applications for ChIP-sequencing data reusability



Thomas Frédéric Turlough Rensch

EMBL - European Bioinformatics Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Darwin College

February 2016

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution. It does not exceed the prescribed word limit for the Degree Committee for the Faculty of Biology.

Thomas Frédéric Turlough Rensch
February 2016

To Yasmina

Acknowledgements

It is a pleasure to thank those who made this thesis possible, beginning with my supervisor, Paul Flicek. His support and guidance were crucial in my scientific accomplishments, including but not only, the writing of my thesis. His trust and the freedom he gave me to explore the world of scientific research certainly challenged me, but from this I have learnt and developed immensely. All of my research was performed in the Flicek research group, which provided an invaluable environment in which to perform my work. For creating this environment and the incredible support through discussions, group meetings, presentations and social gatherings I would like to thank all of the past and present members of the Flicek group I encountered: Petra Schwalie, André Faure, Albert Vilella, David Thybert, Graham Ritchie, Camille Berthelot, Emily Wong, David Martin-Galvez, Maša Roller, and Dhoyazan Azazi. I would also like to thank our experimental collaborators from the Odom group, in particular, Duncan Odom for always making science exciting and his precious help in polishing my manuscript, and Diego Villar Lozano, who performed many experiments to support my research project and contributed significantly to my scientific paper. I am also very grateful to Nick Goldman, Wolfgang Huber, and Bertie Gottgens, all of whom constituted my thesis advisory committee and provided valuable feedback and suggestions throughout my PhD.

I am most grateful to the entire EBI "predoc" community, who all together create an incredible working and social environment that continuously provides scientific advice and support in many other aspects of the PhD life. In particular, I would like to thank the predocs of my year with whom I had the most exchanges: Nils Kölling, Konrad Rudolph, Kevin Gori, Ewan Johnstone, and Michael Menden. In this context, I would also like to express my gratitude to the other predoc representatives, Sander Timmer, Nils Kölling, and Maria Xenophontos for our numerous discussions and negotiations on a wide variety of matters related to the EBI PhD program, constantly dedicated to improving the student experience.

My heartfelt thanks also to my friends of the Cambridge Swiss Society and my fellow drivers of the Cambridge University Automobile Club. As we shared wonderful times, from enjoying

a cheese fondue to going around a track at 80mph, I am indebted to them for the fun we have shared, their friendship which I will always treasure, and for helping me maintain a healthy balance in life. Special thanks to my fellow Swiss Society Committee members: Nicolas Huguenin, Christine Hänni, Elisa Hemmig, and Bryan Ormond for their great support.

Absolutely none of this would have been remotely possible without my parents: Jean Jacques and Blánaid Rensch, who, from my earliest childhood, have always encouraged me to aim to achieve the best I can and have constantly helped, guided and supported me through all of my life decisions. My brother, Colla Rensch, has been equally important in my success, providing me with plenty of motivation as we shared our respective academic challenges, as well as his eternal fraternal support. Heartfelt thanks too to my friends, Michaël Portner and Yann Borie, back in Switzerland, who regularly share a bit of Swiss life with me.

Last and certainly not least, I am eternally grateful to my beloved partner, Yasmina Monteiro, who joined me for this incredible adventure in the United Kingdom. I cannot put into words how much having her by my side throughout my PhD has been vital to my success in this endeavour.

Abstract

Since the development of Next Generation Sequencing techniques, the field of bioinformatics has been producing significant amounts of data. Most of this data is free to access by anyone and stored in enormous databases such as the European Nucleotide Archive. Although this data has predominantly been generated in the context of solving a particular biological problem, it can be reused to explore different questions. My thesis describes two projects in which I design and execute experiments that make use of previously generated chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) datasets.

The first study explores mitochondrial heteroplasmy—the presence of more than one mitochondrial DNA (mtDNA) variant in a cell or individual. This occurs mostly because of the high mutation rate of the mtDNA and limited repair mechanisms in the mitochondrion. Motivated by mitochondrial diseases, the phenomenon has mostly been studied in medical contexts with human samples. To place these results in an evolutionary context and to explore general principles of heteroplasmy, I performed a large cross-species evaluation of the phenomenon. I developed a novel approach to detect mitochondrial heteroplasmy in ChIP-seq datasets, which include concomitant mtDNA sequenced in the experiment. In addition to validating that ChIP-seq data is an appropriate source to identify heteroplasmies, the results suggest that heteroplasmic positions across vertebrates have similar characteristics to human heteroplasmies. Largely consistent with previous studies, the results provide valuable insights into mitochondrial heteroplasmy.

The second project investigates the deconvolution potential of complex tissue ChIP-seq data. ChIP-seq studies have been predominantly performed using homogeneous cell lines because of the considerable amount of cells (of the same cell type) required to obtain reliable and biologically interpretable results. However, most biological tissues are complex and composed of many cell types. I explored a computational approach to model the behaviour of complex tissue ChIP-seq aiming to estimate the cell-type specific DNA-protein binding profiles directly from complex tissue ChIP-seq experiments. I performed in-silico experiments with publicly available datasets in order to simulate complex tissue data and

applied several computational methods attempting to deconvolve this data. Although the results are inconclusive, they provide insights into potential deconvolution approaches that could be applicable in the future.

As a whole, my thesis describes novel uses of existing datasets that have enabled the gain of deeper insight into biological systems and the assays used to measure them.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Biological research has evolved rapidly in the past few decades. Although the discovery of cells as units of complex organisms [1] and the theory of evolution are a couple of centuries old [2], until recently most research work has been qualitative rather than quantitative. Modern biology however has now branched out into several more quantitative subfields such as biophysics [3], bioengineering [4], and bioinformatics [5].

1.1 Bioinformatics

Bioinformatics is a field of interdisciplinary science in which computational approaches are devised and applied to solve biological questions. Although the field emerged from biology, when biologists started using computational tools and models to make sense of the novel complex data that was being produced, scientists from many different backgrounds (such as Computer Science or Physics) then joined the field and contributed to these new approaches in order to tackle the challenging biological questions [5]. The heterogeneity of specialists working together is what makes this field of research so special. I believe it is responsible for the significant progress that has been accomplished over a relatively short period of time and that the variety in the approaches to solve the biological questions, as well as the combination of methods, greatly increases the success rate in making scientific discoveries.

It is unclear exactly when Bioinformatics as a field of scientific research was born, but a key "trigger event" was when the complete amino acid sequence of the Insulin peptide was sequenced by Frederick Sanger in 1951 [6]. Several other peptide sequences were discovered shortly after but comparing and analysing them by hand was complicated and time consuming.

A few years later, scientists began to use computers to perform sequence analysis, for the first time. Margaret Oakley Dayhoff was a pioneer in the field of bioinformatics as she wrote programs to determine the complete amino-acid sequence of proteins [6]. Thirty years later, in 1981, the first significant part of the human genome, the human mitochondrion genome, was sequenced [7]. This circular piece of DNA, sometimes considered as a chromosome, was the easiest target to sequence due to the high number of copies and its short length (see below and Chapter 2 for more details). It took an additional 30 years for the full set of human chromosome sequences to be published in the year 2001 [8]. Today (2016), in addition to a refined and more detailed human genome version, high quality genomes of many other species are publicly available [9] and used by many scientists to further our knowledge in areas such as medical and evolution research [10–12].

Understanding how we humans have evolved, from floating unicellular organisms into gigantic multicellular individuals constituted of trillions of cells, is no easy task. Since the sequencing of our genome, we have built and continued to build the foundation of our knowledge and are expanding our understanding of the evolutionary process further than ever before. Although, whether life started with DNA or RNA is currently under debate [13], there is no longer any question regarding the importance of the role that DNA plays in evolution [14]. Moreover, since DNA is at the core of our functioning, it is likely that many explanations regarding our evolution are to be found in our genomes. In addition to gaining deeper insight into of how we became what we are, improving our knowledge of the inner workings of our DNA could lead to many medical breakthroughs. In recent years, we have discovered that many diseases are due to problems in our genomes [15], some of which are linked to minor genomic defects [16]. Although we have not yet discovered how to cure these diseases, we are currently trying to further our knowledge of the underlying processes that lead to the disease, which could potentially reveal key preventive solutions. So understanding how our complex organism has come to be, how it works, and how it will likely evolve in the future, in addition to satisfying our scientific curiosity, will also likely improve the quality of life for many people in this world.

Bioinformatics is a broad field of research with unclear borders. As mentioned above, it consists in exploring computational methods to make sense of modern and complex biological data, and it encompasses research such as image analysis methods for high throughput microscopy [17] or physical modelling of three-dimensional protein folding [18]. Although the field has expanded over the years, one of the core topics in bioinformatics remains sequencing data analysis [19], which is the focus of my research. There are several approaches to sequence analysis, several types of sequences that may be analysed (DNA,

RNA and protein) as well as numerous contexts of research. In the following sections, I will begin by summarising the current theory of cell biology in the context of genome to cell function, then I will briefly describe a few research areas before providing further details into protein-DNA binding interactions, which are essential to understanding the context of my thesis.

1.2 Evolutionary genomics

As previously stated, the main aim of bioinformatics is to further our knowledge of biological function. Each and every biological being that can be observed today is the result of evolution, which took place over millions of years of reproduction or replication [20]. For this reason, it is crucial to understand where "we" come from and how we have become what we are now. It is unclear who discovered the concept of evolution and hereditary traits, but famously it was Charles Darwin who made the first significant step. In 1859, he published his thoughts and arguments regarding the similarity in different bird species that he observed during his voyage in the Galapagos Islands [2] and introduced the concept of natural selection as a driver for evolution. Then, although evolution was widely accepted in the scientific community, for a number of years, it was believed that proteins carried the hereditary material. Friedrich Miescher, a Swiss physician and biologist not only discovered DNA (which he called "nuclein") but that variations in the DNA caused changes in bacteria types, thereby showing that DNA was a carrier of hereditary information [21].

Evolution has occurred over millions of years, while the field of genomics is less than 100 years old. This means that studying evolution using genomics is comparable to studying an entire film by only looking at the last frame (although genomes do carry traces of evolution). Luckily though, individuals from different species groups (e.g. mammals) are very similar to each other organically and even more so from a cellular point of view. To understand evolution, in the same way as Darwin compared birds, it is now common practice to perform large-scale genomic studies covering many species [22]. This enables us to comprehend which parts of the genome are identical in which species, and when and how speciation events might have occurred. From these genomic differences between species scientists have been able to establish the most likely evolutionary tree for all species [9] (see figure 1.1). Researchers have been able to establish and date most of the speciation points [23]. Although there exists a debate regarding the finer details of the exact genomic evolutionary tree [24], it was found to be largely in accordance with most of the early phenotype-based analysis.

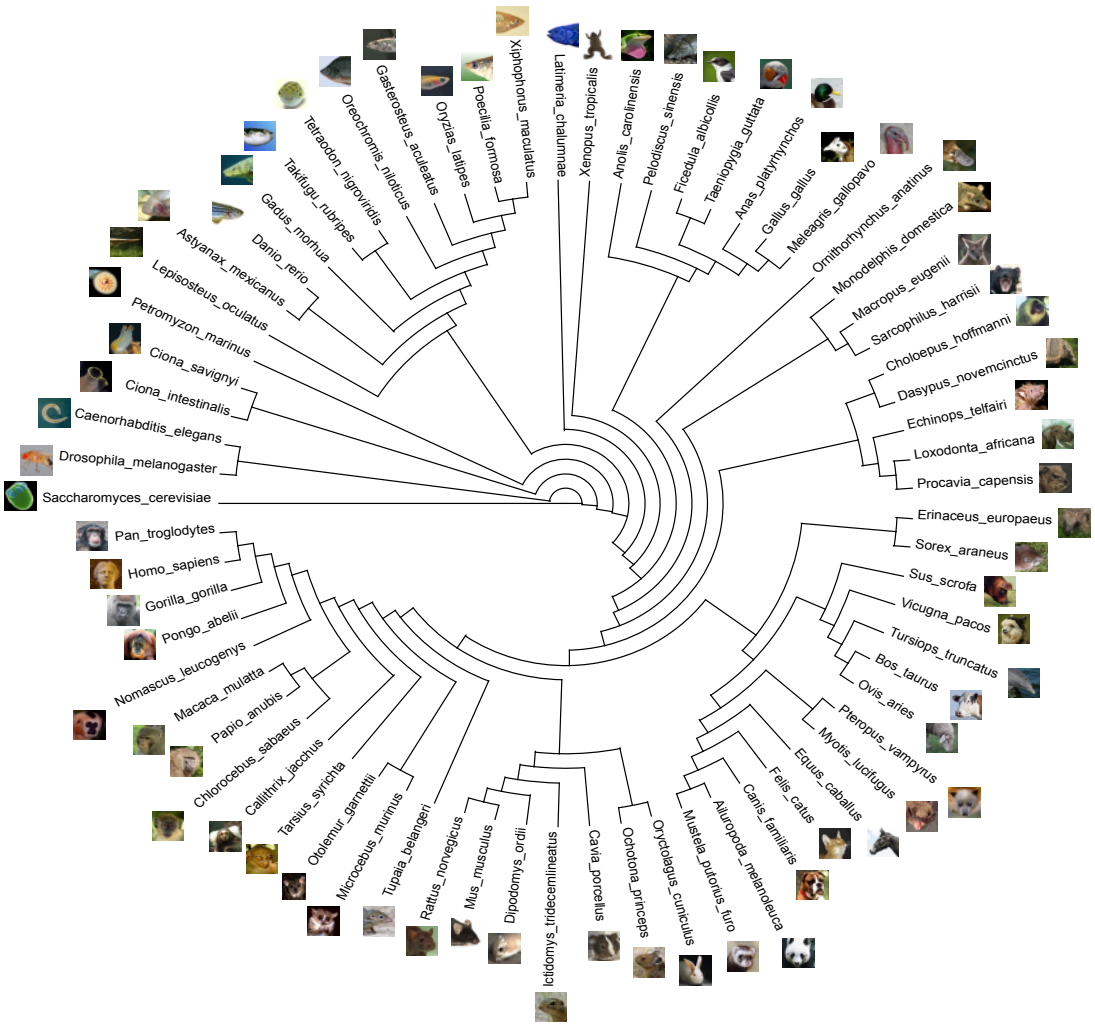


Fig. 1.1 The Ensembl Species tree. This evolutionary tree is produced and maintained by the Ensembl Compara team that specialises in comparative genomics. It is built using multiple alignments (see section 1.9.3) of the species' genomes. It shows the evolution patterns of all the Ensembl species, from human (*Homo sapiens*) to yeast (*Saccharomyces cerevisiae*). The graphical image of the tree was generated using Dendroscope [25]. The *Homo sapiens* species is located just below "nine o'clock". Obtained from Ensembl (v. 83) [9].

The latter suggesting, for example, that it was likely that we shared closer common ancestors with primates than with birds.

Studying evolution is key to understanding how we humans function. By investigating how we differ from other species, we can discover genotype to phenotype links, which can then lead to knowledge about how organs have evolved according to the environment. Evolution has come a long way and although we are significantly different from the first life on earth, which were tiny unicellular organisms, much of the cellular structure and organelles have remained similar [26]. Our understanding of how cells work and communicate has progressed incredibly in the last decade but many unknowns still exist. The main drive to gain such deep insight is to solve issues that relate to our health, such as how cells degenerate in cancer [27] or the reason why we are not equipped to deal with Alzheimer's or Parkinson's diseases [28, 29]. By pursuing fundamental research we are participating in the global push towards knowledge and this, although indirectly, has and will contribute to medical progress.

1.3 The central dogma of molecular biology

Cell biologists have understood the inner workings of cells for numerous years. The various organelles have been viewed under microscopes and without knowing the details of the mechanisms, it was known that the genome (DNA) is located in the nucleus, that it is transcribed into RNA, that the RNA floats out into the cellular matrix before being translated into proteins. This process, called the central dogma of molecular biology, still holds true today (see figure 1.2). It is at the core of almost all living organisms and thus key to understand of how these function. RNA viruses store their genetic material in RNA and thus do not exactly follow the central dogma however, it is debated whether they are living organisms [30]. Although first described in 1970 [31], we have since discovered many details of how each step works and are still investigating these mechanisms with the help of modern technology. The information stored in the DNA is not only used but replicated many times in the life of a cell, in addition to being propagated to the cell's offspring. Each of the underlying processes are complex and involve several different molecules and organelles. Each part of the central dogma is described in detail below.

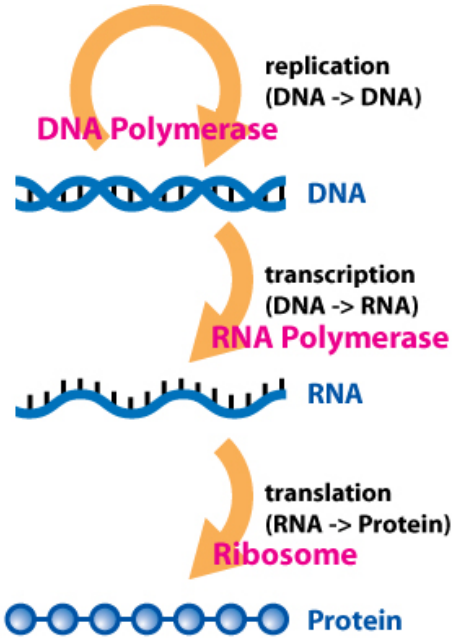


Fig. 1.2 The central dogma of molecular biology. The main transitions of the dogma are depicted above. During replication, the DNA polymerase copies the DNA. The DNA is transcribed into RNA by the RNA polymerase. Finally proteins are translated from RNA by a more complex process involving tRNAs and Ribosomes. Other transitions such as RNA to DNA and RNA to RNA also exist but are not as common and have thus been omitted from this schematic. Obtained from [33].

1.3.1 Deoxyribonucleic acid — DNA

Deoxyribonucleic acid (DNA) is the material used to store all of the information required by a cell to function and replicate. From a bio-chemical point of view, DNA is a covalently linked chain of molecules called nucleotides. Each nucleotide is composed of a nitrogen molecule with a base attached, a phosphate group, and a monosaccharide sugar. The backbone of the chain is constituted of sugar and phosphate groups covalently linked to each other, which creates a very stable macromolecule. There are four different DNA base types: A: adenine, T: thymine, C: cytosine, and G: guanine. Additionally, most DNA strands exist in a double strand formation, bases from each strand creating hydrogen bonds with bases from the opposite one. Each base only forms such bonds with one other base, meaning that the sequence of the second strand is dependent on the sequence of the first strand and vice-versa. Although some DNA strands are circular, those that are not have differing endings. On one side it is the 3' carbon atom from the ribose that is free, while on the other side it is the 5'. The three-dimensional structure of the double-stranded molecule, which was discovered in 1953 by Watson and Crick, forms a double stranded helix [32]. This helicoid structure provides additional molecular stability.

Most of a cell's DNA is located in its nucleus and the genetic information is encoded in the sequence of the nitrogen bases. This information exists in different forms. Some sequences contain blocks of information that will be used at the next level, generally outside the nucleus. These blocks will be transcribed into RNA (see section 1.3.2). Other parts of sequence are used as start and stop elements placed before and after previously mentioned information blocks. Some sequences serve as docking areas for particular cellular machinery such as the RNA polymerase (see below). Since the DNA contains all of the vital information required by a cell, it needs to be copied and passed on to the next cell when cell division occurs. This means that the information is copied and read many times and errors may accumulate. This is why the double stranded aspect of DNA (as described above) is extremely important; the second sequence contains the same information as the first and the cell makes use of this property to repair DNA, making sure both strands are consistent. The double strand is also used in the replication process that occurs before each cell division. Both strands are unwound and the DNA polymerase binds to each strand, using one as a template to assemble the respective opposite strand. This process yields two entire copies of the initial DNA, each copy consisting of one "old" strand and one newly created strand.

Since the genetic information is encoded in the sequence of the four bases, it is often represented as a chain of the four corresponding letters e.g. "ATCGGGCTA". Although DNA is double stranded, since the two strands contain essentially the same information, most of

the time DNA is represented as a single string of the four letters. To avoid confusion, DNA is always written from the 5' to the 3' end.

Genomics

The first mention of the term "Genomics" occurred in 1986 at a conference on the feasibility of mapping the entire human genome. During a discussion around starting a new scientific journal, Thomas Roderick, a geneticist, suggested calling it "Genomics" (which they did) [34]. The scientific field of genomics developed around the focus to study the structure and function of genomes. The field's key trigger event was the arrival of Next-Generation Sequencing methods (see section 1.8.1), which enabled genome-wide experiments and analysis to be performed [35]. Genomics includes many sub-categories of research such as the determination of new genomes [36] and the improvement of existing sequences [37], genetic mapping (studying gene clusters) [38] and genome-wide studies of various events (e.g. protein-binding or SNPs) [39, 40]. Another large field of study, evolutionary genomics, focuses on genomics in the context of evolution, comparing genomes between species and investigating potential ancestor genomes. Although some overlap exists, the study of single genes and their potential effects is generally categorised in the field of genetics. Genomics focuses on the study of DNA sequences, which as described above, are at the core of cellular function and contain most of the hereditary information. However, to obtain higher quality results or to broaden the scope of the research, the analysis of DNA is often combined with analysis of RNA or Protein sequences. Scientific fields focusing on the study of RNA and Protein sequences have also been largely developed and are described in more detail below (section 1.3.2 and section 1.3.3).

1.3.2 Ribonucleic acid — RNA

To access and use the information stored in DNA, the previously mentioned blocks of information are copied onto a different biological support called RNA. Ribonucleic acid (RNA), is a molecule that is very similar to DNA, but also differing in a number of key aspects. RNA does not contain thymine bases. Instead, it contains uracil ("U") bases, which pair with adenine, thus replacing thymine. Additionally, the sugar group included in its backbone is a ribose instead of a deoxyribose. This makes the whole molecule less chemically stable and it degrades over time due to hydrolysis. RNA exists as a single strand molecule and forms bonds with itself by folding and forming helicoidal structures with itself. It is

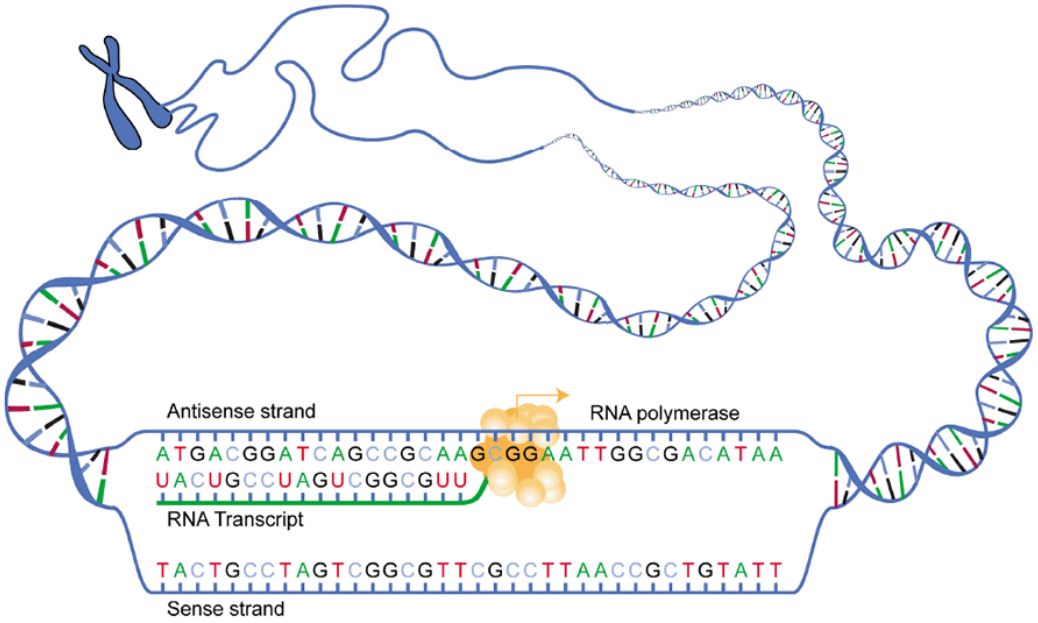


Fig. 1.3 DNA transcription into RNA. This figure depicts the RNA polymerase creating an RNA sequence (called a transcript). The DNA of the chromosome (visible in the top left) forms a double helix that has been unwound and opened by the RNA polymerase, which is using one of the DNA strands as a template to assemble the new RNA transcript. Once the transcript is terminated, it will float away and the DNA will "close" and reform a double helix with the other strand. Obtained from [41].

by forming such complex structures that it can act as structural building blocks as well as catalysers for certain reactions.

The RNA polymerase works in a similar manner to the DNA polymerase. It unwinds the DNA and uses a single strand as a template. The "second" strand however is assembled in RNA, and the DNA rewinds as the copy completes. The process of "reading" DNA and "copying" the sequence into RNA is called transcription and is depicted in figure 1.3. The RNA is generated in the nucleus, but most of it will float out of the nucleus through the nuclear pores. It is also important to note that RNA molecules are often subject to post-transcriptional modifications. These can add sections to the transcript, such as polyA tails (to avoid degradation), or remove sections such as intronic regions (transcribed regions that are not part of the final RNA product).

The chemical stability of the RNA sequences is often used to the cells' advantage, since most RNAs are used as templates to build proteins. Once the protein has been built, the RNA degrades and its blocks are reused to produce new RNAs. The process of translating DNA to RNA is mostly unidirectional. However, there are exceptions such as the infamous human immunodeficiency virus (HIV), which uses a Reverse Transcriptase to transcribe its RNA genome into DNA, that is then integrated into the host's genome [42].

Cells use RNA molecules for multiple purposes. Some RNA sequences act as information vectors to be further used to create proteins, these RNAs are known as messenger RNAs (mRNAs) (see below). Others, which are not used to encode protein sequences, are known as non-coding RNAs (ncRNAs). NcRNAs are used as structural blocks to build organelles such as Ribosomes (used to build proteins, see below). In addition to being structural, ncRNAs may also act as catalysers, enabling chemical reactions necessary to cell function. Finally, very small ncRNA molecules known as micro-RNAs are thought to be involved in the regulation of different processes in the cell.

Transcriptomics

The collection of RNA materials that a cell produces is called the transcriptome. From organelle building blocks, to catalytic sites and protein generation, the transcriptome is at the core of cellular function. The field of transcriptomics is the study of RNA sequences and they provide a sense of which blocks of genomic information are being used by the cell. However, several aspects of the process from DNA to RNA are complicated, for example pieces of RNA are generally processed to remove unnecessary parts (introns). Modern techniques are

able to identify different types of RNA and are also capable of quantifying the RNA content, hence providing insights into which genes are transcribed more than others [43].

1.3.3 Proteins

The final step in the central dogma process is the use of the mRNAs to produce proteins. This process is known as translation and is shown in figure 1.4. Messenger RNAs (mRNAs) are translated into proteins. Proteins, similarly to RNAs may be used for several purposes such as catalysers and structural blocks. However, they are no longer chains of nucleotides but chains of amino acids. There exist 20 different types of amino acids, therefore enabling a greater diversity for proteins. Obviously, the process of reading a four-element based RNA to assemble a 20-element based protein is a more complex process than the DNA or RNA "copying". In fact, three RNA bases correspond to one amino acid, and since there are more than 20 three-base combinations, some combinations code for the same amino acid. The cellular machine responsible for assembling proteins from RNA information is called a Ribosome and is constituted of a combination of ncRNAs and proteins. They read mRNAs and create chains of amino acids according to the mRNA sequence. A particular type of ncRNAs, tRNAs are used to transport amino-acids to the ribosome, which catalyses them together one by one. Proteins fold as they are being created, forming different types of bonds with themselves. Proteins will then also often bond with other proteins forming complex structures that constitute most of the cellular-machinery. They act as catalysers, they form structures to build complexes (similarly to RNA structures), and are involved in many cellular reactions. Proteins are constituted by more building blocks than RNAs and consequently enable more diverse structures. They may also bind to a larger range of molecules with various biochemical bonds (e.g. hydrogen bonds).

Proteomics

The study of proteins, their structure and function is known as Proteomics. The proteome is the most variable set when comparing to the genome or the transcriptome, and constantly changes as the cells react to different situations and environments. Understanding how proteins fold and come together to build complexes that are then able to perform complex tasks is very challenging. Modern techniques such as Xray crystallography and mass-spectrometry, are used to perform such experiments. Similarly to RNAs proteins are also modified after having been created and this phenomenon is called post-translational modification.

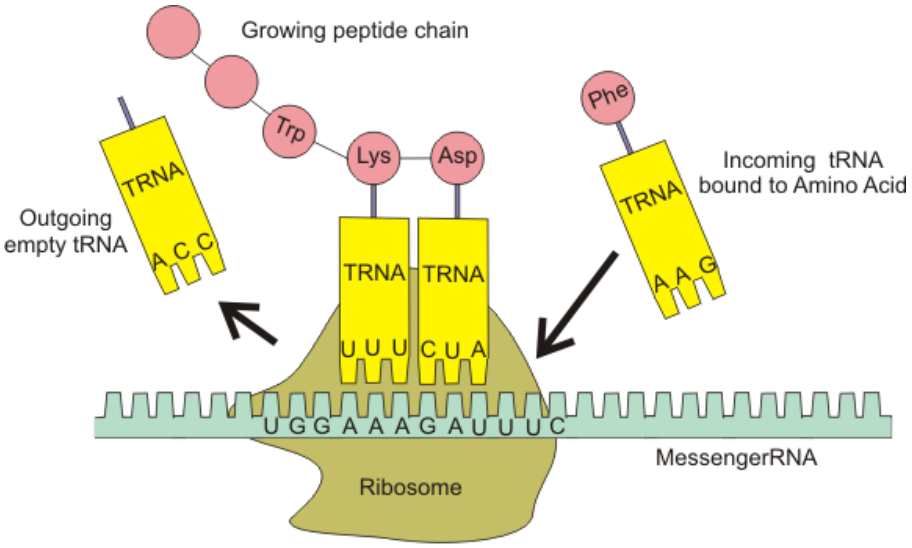


Fig. 1.4 mRNA translation into protein. This figure shows the Ribosome (a complex molecular machine, see section 1.3.3) reading an mRNA transcript and recruiting the corresponding tRNAs. Each type of tRNA carries a particular amino acid. The peptide chain will fold and float away once terminated. Modified from [44].

1.3.4 Phenotype

Finally, the combination of the various cellular components generally vary depending on the function of the cell. This leads to cells having different contents and shapes. Cells are said to differentiate as they become more specialised, thereby optimising their content for their specific tasks. For example, human erythrocytes (red blood cells) do not have a nucleus and are optimised to transport oxygen, while neurones have tree-like shapes with many dendrites to increase communication channels (see figure 1.5). The phenotype is the set of observable characteristics of a cell, tissue, or even organism and is often considered the final aspect in terms of biological mechanisms; all of the underlying processes "create" the organism phenotype.

Medical genomics

Scientific discoveries in all three fields (Genomics, Transcriptomics, and Proteomics) have a high impact but it is when they can be linked to each other that true progress is made in understanding the complex procedure that is cellular function. Many diseases have been associated to genome abnormalities. Major genomic impacts, such as an additional chromosome (e.g. Down syndrome) [45], as well as minor changes such as a small number of different base pairs in the genome DNA [46], can lead to diseases. Moreover, cancer is a form of cellular degeneration, wherein cells reproduce anarchically, which is initiated by small changes in normal DNA. Currently it is clear that we are observing DNA situations which are likely to induce disease and in most cases we cannot do much to prevent said disease. Areas of research in medical genomics are attempting to edit DNA to lower the prevalence of disease in such cases [47].

1.4 The genome and its elements

Although, occasional references to methods from transcriptomics and proteomics are made in this thesis, the main research focus here is on genomics. Hence, all of the data generated, processed and analysed, in the context of this thesis, consists of DNA sequencing data. The following subsections provide a more detailed view of the organisation of the DNA within cells.

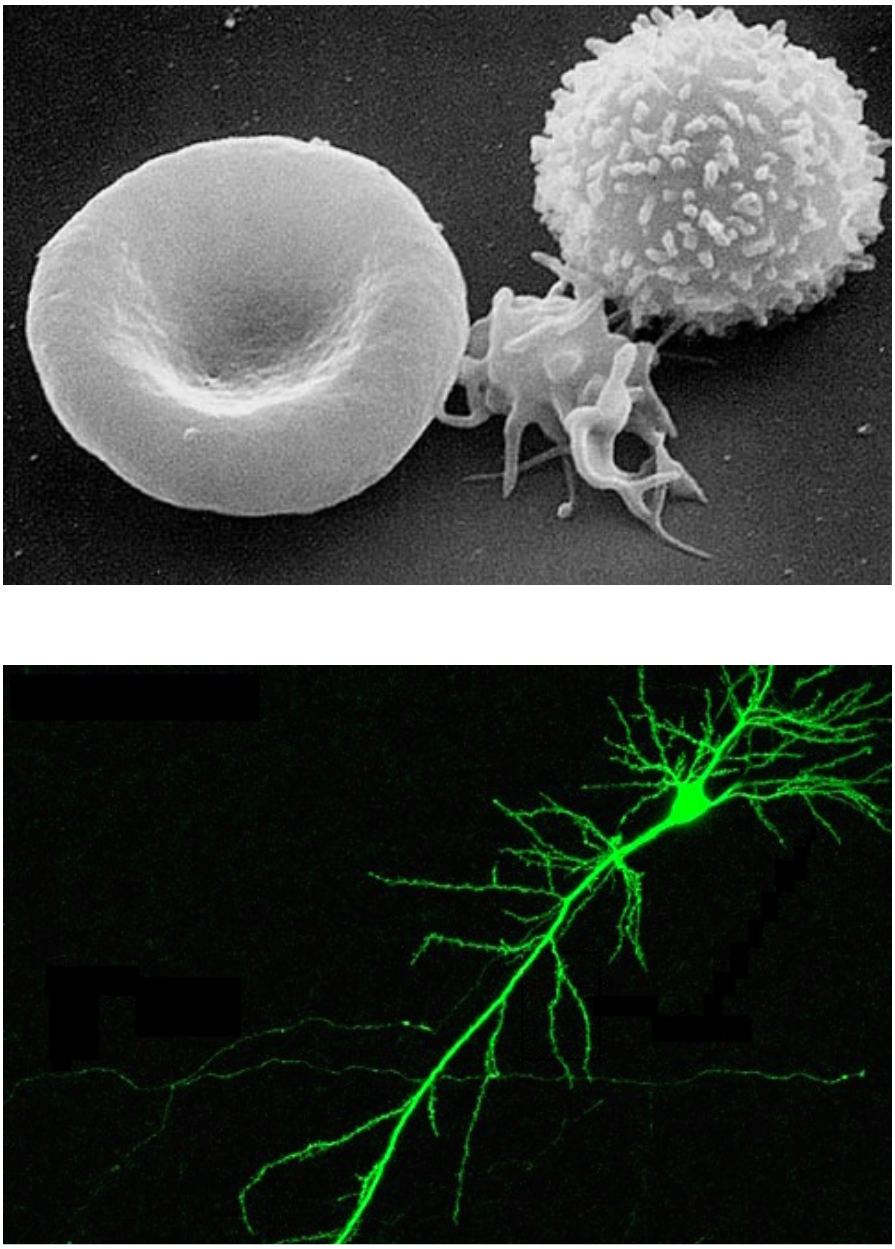


Fig. 1.5 Different cell phenotypes. The top image shows the three main blood cell types, respectively (from left to right): Erythrocytes (red blood cells), Thrombocytes (platelets), and Leukocytes (white blood cells). The bottom image shows a neuron (brain cell). These pictures illustrate the variation in cell phenotypes (shapes). Both images were produced using high resolution microscopy techniques (Electron microscopy and confocal microscopy). Modified from [51, 52]

1.4.1 Genome

The genome refers to the entire genetic material contained within a cell or an organism. This almost always consists of DNA, except, for example, in RNA viruses such as HIV (see section 1.3.2). The length or size of a genome is defined by the total number of non-redundant bases, thus counting the bases of one of the strands only (DNA almost always exists in a double stranded form, see section 1.3.1). The size of different genomes varies significantly between species; from a hundred kilo (10^3) bases in some bacteria [48] to a hundred giga (10^9) bases in particular fish [49], the human genome contains just over 3×10^{10} base pairs [8]. Studies have found some correlation between the size of genomes and the complexity of their respective organisms, however it does not hold as the genomes become very large [50]. In addition to the size of genomes, there are a number of different characteristics of the DNA sequences that have been (and are still being) studied. Below is a list of some of these characteristics.

- **Repetitive content:** Some regions of the genome may contain segments of highly repetitive nature. Relatively short sequences (generally less than 100 bases) are repeated many times. Several different mechanisms have been shown to create these repetitive sections, such as problems when the DNA polymerase "slips" and copies a section multiple times. Small bacterial genomes often have no repetitive segments at all, while the very large genomes tend to have a large portion of such segments. Although for a long time, such regions were deemed non-functional, several recent studies have shown the opposite [53, 54].
- **GC-content:** The GC-content is the fraction of bases within a region that are either cytosine (“C”) or guanine (“G”). The number of hydrogen bonds between opposite bases is higher for guanine-cytosine than for adenine-thymine, thus DNA with a higher GC-content is more stable biochemically. The GC-content varies in different species and a number of studies have investigated the GC-content of a species’ genome in relation to the temperature of their usual environment [55, 56]. It has also been shown that regions with a high number of protein-coding genes, have a higher GC-content [57].
- **Ts/Tv rate:** The rate of transitions to transversions (see section 1.4.3) is another characteristic that has been studied in different genomes. Theoretically, due to the chemical properties of the bonds, the number of transitions should be higher than the number of transversions. This is generally true, however the exact ratio varies across species as well as across different genomic contexts [58].

1.4.2 Genes and gene regulation

A genome is far more than a long chain of nucleotides since information is encoded within the sequence of bases (see section 1.3). Many scientists are still working hard to unveil the details regarding the storage, usage and interpretation of this complex information. To be useful, such quantities of data need to be carefully organised. The main information blocks encoded in the genome are genes, of which exist several different types. Protein-coding genes are DNA sequences that will be transcribed into mRNA transcripts before being translated into proteins (see section 1.3). Non-coding RNA genes are DNA sequences that will only be transcribed into RNA but will not be further translated. There are many different types of non-coding RNA genes, some are long and will form structures, while others may be extremely short and interfere with cellular processes, for example by binding to mRNA transcripts (which can lead to the degradation of the mRNA). To provide a reference of the number of distinct genes, there are just over 20,000 protein-coding genes in the human genome (20,313 listed in Ensembl v.83 [9]). However, it is still unclear how many non-coding RNA genes exist, especially since there are many different microRNA genes [59]. Overall, it appears that about 75% of the genome is transcribed from DNA to RNA, but only about 1% of the genome will be further translated into proteins [60]. Nonetheless, non-transcribed DNA sequences are still found to play important roles in terms of cellular function, for example acting as binding locations for regulatory proteins.

Gene regulation

In a living organism, almost every cell capable of division contains the same DNA material packed into identical chromosomes. However, these cells are far from being the same in any other way. Since they have different tasks to perform in the organism, they do not require the same proteins to function optimally. In order to produce more, fewer or no proteins at all, genes are regulated in various ways by many a range of elements at each level. At the DNA level, the first elements of gene regulation are the histone proteins around which the DNA is folded; they need to be ejected to give access to the RNA polymerase so it can bind and initiate transcription.

More complex processes involving many different DNA binding proteins also have an influence on gene regulation. Transcription regulators are proteins, which generally bind to the promoter location of genes to facilitate or block the RNA polymerase from docking. Usually a combination of different proteins bind around the promoter, interacting with each

other. Transcription factors are necessary to enable the "docking" of the RNA polymerase, while other regulators will accelerate, slow or completely block the polymerase.

The addition of a methyl group to the DNA strand (DNA methylation) is another process responsible for gene regulation. This transformation in itself renders the binding of the RNA polymerase more complicated, but additionally specific proteins that completely block the process often bind these methylated sequences. Moreover, the proteins which are responsible for organising and packing the DNA called histones (see section 1.4.4), also play a key role in gene regulation. Their structure can be chemically modified and these "histone modifications" have been shown to impact chromatin structure and transcription regulation [61]. Gene regulation at the DNA level is known as transcription regulation. Processes of regulation at the RNA and protein levels are other key processes for cellular function, however they are not relevant to this thesis and will thus not be covered here.

In this thesis, I make use of ChIP-seq data from a variety of transcription factors and histone modifications. In Chapter 2, most of the data comes from CTCF, CEBPA, HNF4A (as well as histone marks H3K4me3 and H3K4ac), while in Chapter 3, I use mostly CTCF, CEBPB and P300. A brief description of these binding proteins follows:

- **CTCF** is a highly conserved and ubiquitously expressed transcription factor. It has multiple functions, acting in the regulation of transcription (both as an activator and a repressor), but also as an insulator (blocking the activity of enhancers). It also plays a key architectural role in the three-dimensional organisation of the genome, by binding strands of DNA together to form chromatin loops (see section 1.4.4) [62].
- **CEBPA** and **CEBPB** are two transcription factors that play key roles in liver function, however they are also expressed in other tissues such as lung, pancreas, or skeletal muscle. They bind the same core DNA sequence but have divergent functions (CEBPA is highly expressed in healthy liver tissue, while CEBPB is up-regulated during liver regeneration (post-injury for example). Both regulate many metabolic genes and bind DNA as a homodimer (two identical proteins forming a larger protein complex before binding DNA) or as a heterodimer with each other (one CEBPA and one CEBPB protein) [63].
- **HNF4A** is highly conserved and regulates many hepatocyte-specific genes. It plays a key role in the differentiation of blood cells. It is mainly expressed in liver and kidney but also in other tissues such as the small intestine and colon. HNF4A binds DNA as a homodimer. [64]

- **EP300** is a transcriptional co-activator as it binds transcription factors to activate transcription. It is expressed in multiple tissues and regulates the activity of many genes, playing a key role in regulating cell growth and division, as well as cell differentiation. Mutations affecting this protein have been linked to several cancers, since without EP300, cells cannot regulate growth and division, which leads to the formation of tumours. [65]

1.4.3 Genetic mutations

Although double stranded DNA is a chemically stable molecule, it is constantly under stress, being unwound and rewound for replication and/or transcription processes as well as being bound and unbound by many different types of proteins (see section 1.4.2 and section 1.4.4). Occasionally, processes do not perform as intended and minor problems arise. For example, the replication process may skip a number of bases, or "stutter" and copy a region multiple times. Sometimes the double strand breaks and needs to be reattached. Many cellular mechanisms exist to deal with these issues and most often they do so very well. However, some mistakes are not repaired and get passed on to other cells, and to the progeny. These "errors" are called mutations. Mutations can consist in the addition or disappearance of bases to the DNA sequence; these are respectively called insertions and deletions. Some mutations will have no incidence on the cell or organism, while others will have different levels of effects and some will not even be viable. It is most likely that these random genomic mutations have driven the evolutionary process. Logically, most of the mutations that living individuals accumulate, have no or only minor effects, and these are generally also the smallest. Mutations can be as small as the change in a single nucleotide, these are called point mutations and they have specific appellations depending on which type of base is replaced by which other; a transition occurs in cases where adenine and guanine or cytosine and thymine are exchanged, while any other type of change consists in a transversion. In terms of chemical theory, transitions are more likely to happen than transversions and this is what is generally observed [58].

Since genomes from a population (within a species or sub-group for example) are extremely similar (see section 1.4.5), the study of genomic mutations within the individuals is key in both evolutionary and medical contexts. From a population perspective, variants of single nucleotide mutations present in multiple individuals are known as single nucleotide polymorphisms (SNPs), while larger mutated segments (from 1kb to 3mb) differing between individuals are known as genomic structural variation [66]. In the human genome, ~5% of

bases are annotated as regions of structural variation, while ~10 million bases are as SNPs [67, 68]. Larger genomic variation, called chromosome abnormality, generally has severe phenotypic impact and will most often be lethal. Trisomy 21, known as down syndrome is an example of a genome abnormality (presence of three instead of two copies of chromosome 21) [45]. In the second chapter of this thesis, I investigate a particular phenomenon which arises from point mutations on the mtDNA genome and that is analogous to studying SNPs, because individuals (and cells) contain a "population" of mitochondria (see Chapter 2).

1.4.4 Genomic structure

As described above (see section 1.4.1), most genomes consist in extremely large sequences of DNA. All of this DNA needs to be organised and structured physically within the cell in order for its content to be readily accessible but also to minimise its volume. Specific proteins, called histones, are key elements in this process. There are four different types of histones and two of each bind together to form an octomer, which along with the DNA constitute the nucleosome complex (see figure 1.6). In the nucleus, the double stranded DNA is bound to many histone octomers, each forming a nucleosome complex. An analogy often used to described this process is that the DNA is like string wrapped around many beads. With the support of an additional histone protein, the nucleosomes are then compacted together forming a larger structure, similarly to threads within a rope. This complex comprising the DNA and proteins together is called Chromatin. Finally, additional proteins come into play to create the highest genomic structures, known as chromosomes. These complex DNA structuring systems are only present in Eukaryotes. Prokaryotes, which typically have smaller genomes, have different proteins, called nucleoid proteins, to organise the genome. Chromatin plays three main roles for the genome:

1. **Compaction:** Due to the large quantity of DNA present in most cells, if these molecules were left to float randomly, they would occupy a space significantly larger than the volume available for DNA in cells [69]. For this reason, DNA compaction is simply essential to cellular function.

2. **Protection and repair:** Since DNA is at the core of cellular function, it is vital to protect it adequately. Being compacted and bound by many proteins prevents significant DNA damage from taking place, in addition to preventing DNA degradation. Histones also play a key role in DNA repair by attracting repair proteins to damaged sites.

3. **Regulation:** The chromatin is constantly being remodelled, for example to provide access to different regions, to transcription factors and polymerases. This process is done by ejecting the histones to free the DNA. Such regions of accessible DNA are known as open chromatin.

Generally, genomes are packed into several chromosomes, the overall number of which varies between species. Males of a particular ant species only have one chromosome [70], while some crab species have 254 [71]. In most species, multiple copies of the genome are present in each cell, meaning there are a number of identical chromosomes. Species, for which two copies are present, are known as diploid and in such cases, chromosomes are known as pairs of chromosomes. For example humans are diploid and have 23 pairs of chromosomes. Triploids have three copies of their genome and tetraploids have four, while species with more than three or four copies are often referred to as polyploids. For a number of species, chromosomes also play a key role in sex determination. In fact, sex is determined genetically in all mammalian species as well as a number of vertebrates. For many animals, this translates into the presence, absence or different number of particular chromosomes known as sex chromosomes.

1.4.5 Sequenced genomes

As previously described in section 1.4.3, individuals within a species have extremely similar genomes. For the purpose of inter-species genomic comparisons, scientists have defined reference genomes per species. Sequencing genomes has been a scientific field in itself, which started when the first small bacterial genomes were sequenced and culminated with the sequencing of the human genome in 2001 (see section 1.1). Moreover, as the cost of sequencing is decreasing [72], increasingly more genomes are being sequenced. Ensembl (v.83), a large set of genomic analysis tools which includes large genomic databases, currently stores the annotated genomes of 87 vertebrate species [9]. Other specific databases from the same project, such as EnsemblPlants or EnsemblBacteria, list considerably more non-vertebrate species' genomes [9].

The recent growth in the number of sequenced genomes is significant and this benefits the scientific community greatly. First and foremost, obtaining the entire genomic sequence of a species enables novel research to be produced, furthering our understanding of genome function, potentially even unveiling phenotypic associations. A second, but equally important point is that a novel genome constitutes a reference, which can be used to rapidly align NGS-produced short sequencing reads of other individuals of the same species, thus broadening

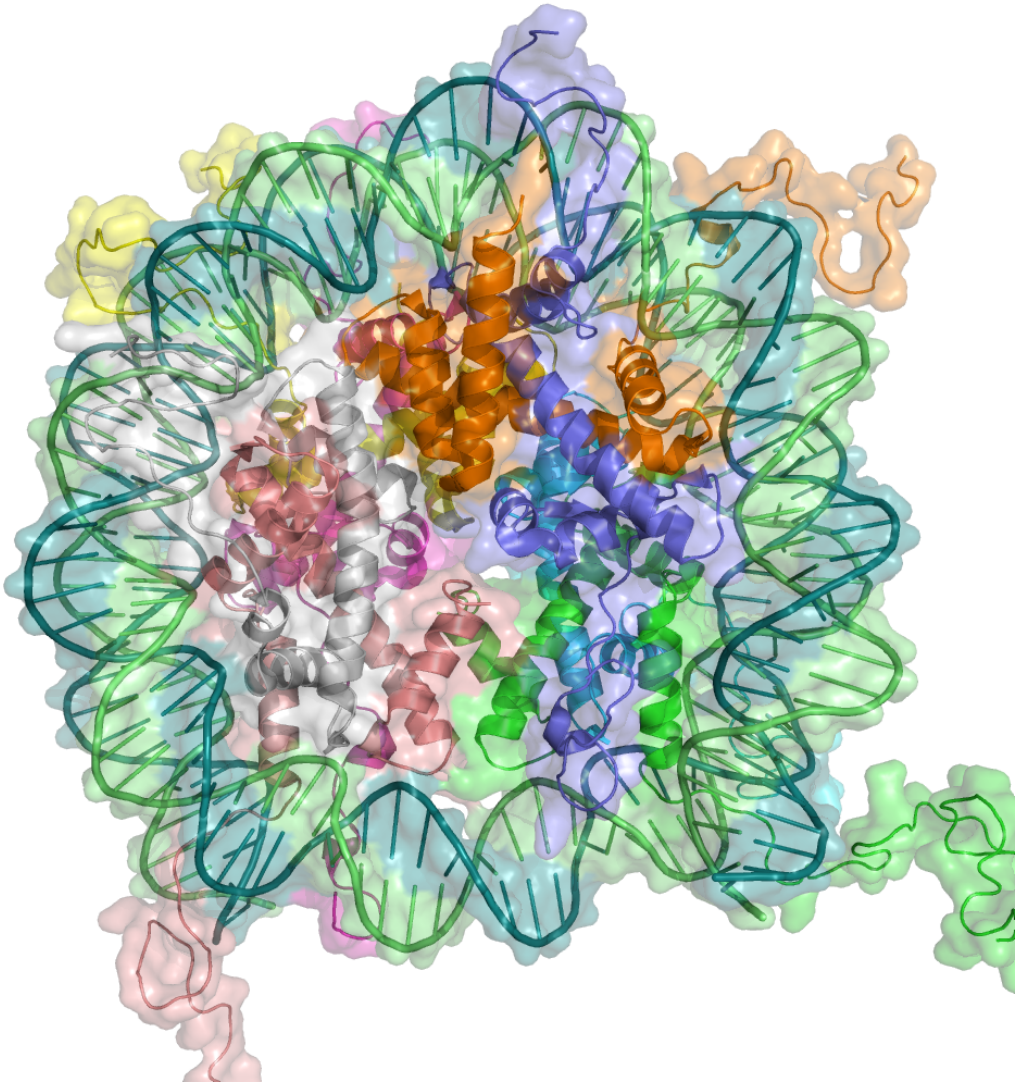


Fig. 1.6 Nucleosome schema. This is a representation of a nucleosome, the DNA double helix is wrapped around the four main histone proteins that are visible at the centre. The backbone of the DNA and the protein secondary structures are colour in slightly darker colour, while the overall three-dimensional shape is shown in light colour in the background. Obtained from [74].

the data for a particular species (see section 1.8.1). Moreover, it also allows the use of NGS to explore closely related species, using the reference genome to map the reads. This approach has been carried out many times successfully and works well when closely related genomes do not have significant changes. In one of my projects (see Chapter 2), I made use of many reference genomes to map reads from individuals to their respective species, but I also mapped reads from individuals, for which a reference genome does not yet exist, to closely related species. Furthermore, it is possible to create multiple alignments (see section 1.9.3) by aligning genomes from different species to each other. I created a multiple alignment of different vertebrates in the context of my mitochondrial heteroplasmy project (see figure 2.3). This can, for example, reveal sequences that are conserved across species, such as ortholog genes (genes that share the same ancestral gene and usually retain their function as well). Furthermore, scientists have also been able to take advantage of the combination of many sequenced genomes to create gene prediction models, which are able to locate genes in species without annotated genomes [73].

Sequencing the entire genome of most species is a challenging task due to the length of the sequence. As previously mentioned, the human genome, for example, is about 3 billion nucleotides long. Several methods and technologies are available to perform this task. However, none of them are able to sequence more than a couple of hundred bases at one time (see section 1.8.1). This means that smaller sequenced regions must be overlapped and joined to recover the sequence of the entire genome; this process is called genome assembly (see section 1.9.2). The number of short read sequences that need to be merged together in order to form the genome is enormous, making this process very complicated. The first step of genome assembly is to build small clusters of overlapping reads, attempting to obtain the longest continuous sequence. These clusters are called contigs. When reads obtained from the same sample, corresponding to a small genomic region, fit (with overlapping bases) in two different contigs, it is possible to know that these contigs are located close to each other (due to the length of the small region). Such contigs are linked to each other, even if the exact sequence between those is still unknown, regions such as these are called gaps. The groups of linked contigs and gaps are called scaffolds. The aim of genome assembly, is to obtain scaffolds that correspond to each chromosome, with no (or very few) gaps within them remaining.

Achieving this is very challenging due to a number of factors such as highly repetitive regions that are impossible to assemble or sequencing errors, which induce errors in the assembly. In reality, genomes are often considered fully sequenced when more than ~95% of the nucleotides are sequenced. The term “quality” of a genome, is often used to describe the

current state of the assembly; low quality genomes will still be constituted of a considerable amount of scaffolds compared to high quality genomes. Since scientists have access to genomes of various qualities, it is an important factor to take into consideration while performing analysis (especially in terms of read mapping, see section 1.8.2). In the work later described in Chapter 2, I make use of high and low quality genomes and I observe a link between potentially poor results with the low quality genome samples (see section 2.4.2).

1.5 The Mitochondrion

The mitochondrion is the organelle (cellular sub-unit) that is responsible for most of the cells supply in energy (transported with molecules of adenosine triphosphate — ATP). They also play a role in other cellular functions, such as cell growth, cell division and cell death. Mitochondria are present in almost all eukaryotic cells and are relatively different to the other organelles. A schematic of the mitochondrion structure and an electron-microscopy picture are shown in figure 1.7. Mitochondria are wrapped in a double phospholipid membrane and contain their own genome. In fact, they are the result of an endosymbiosis absorption of an ancient type of prokaryote organism. This endosymbiotic event took place early in evolution and this explains why most species have them [75]. The endosymbiosis is the reason mitochondria have a double membrane, that was formed upon inclusion, as well as a circular, bacteria-like genome (see below: section 1.5.1). There are many copies of this genome within the mitochondrial matrix and this is necessary because the mitochondria divide via a "budding" mechanism. To rephrase, the membrane closes upon itself and forms two different mitochondria (which both need copies of the genome to continue their existence). The reverse of this process may also occur (when two mitochondria fuse to become one). The overall number of mitochondria in a cell can vary significantly and cells that require a lot of energy, such as muscle cells, generally contain a much higher number [76]. Finally, a key characteristic of the mitochondrion organelle, is that, in most animals, they are inherited maternally. This means that all of the mitochondria within a cell or an organism are "copies" of the maternal organelles. Paternal mitochondria are eliminated during the meiosis division (sexual reproduction) with several different mechanisms in different species. In mammals for example, only a small number of mitochondria enter the egg via the sperm and they are "marked" with ubiquitin (a small regulatory protein that attracts destructive cellular mechanisms) and destroyed [77]. The maternal inheritance of mitochondria means its genome is also maternally inherited and, in this context, there have been many genomic studies focusing on the mitochondrial genome (see below, section 1.5.1).

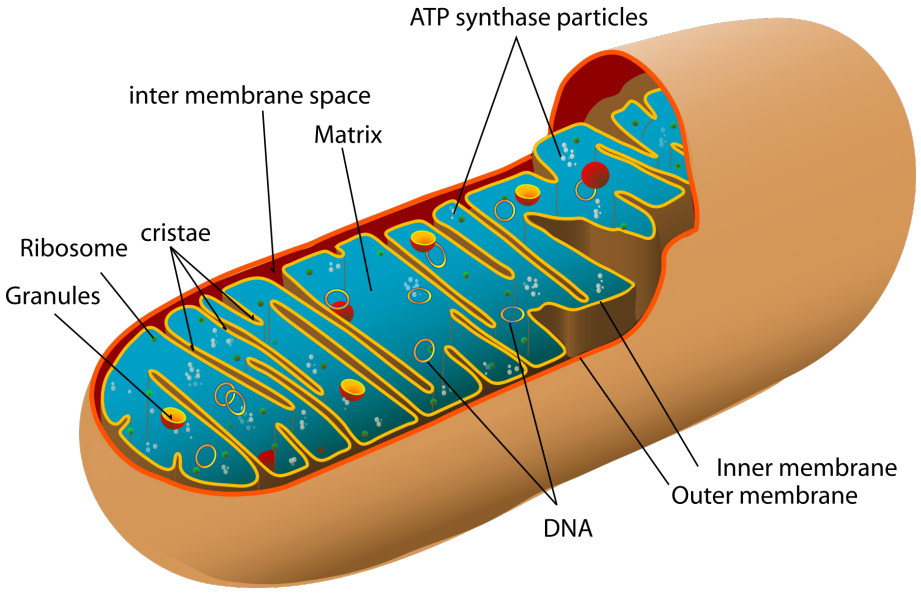


Fig. 1.7 The Mitochondrion. The picture at the top was generated with electron-microscopy and is a cross section of two mitochondria (from mammalian lung tissue). The structure of the organelle is clearly visible with the smooth round outer phospholipid bilayer membrane and the second folded inner membrane which has a much larger surface. The larger surface and the compartments it creates are key to the energy production function of the mitochondrion. Obtained from [80, 81].

1.5.1 The mitochondrial genome — mtDNA

The mitochondrial genome, or mtDNA, is sometimes referred to as a chromosome and is part of the hereditary DNA. However, it is quite different to the other chromosomes in many ways; in fact, it closely resembles a bacterial genome. It is a closed circular molecule of DNA and in mammals it is ~16kb (kilo bases) long. This genome is not bound and packed by histone proteins as the other chromosomes. Instead a different set of proteins (nucleoid proteins) organise and pack the DNA to keep it from floating around in the mitochondrial matrix. The mammalian mtDNA encodes for 13 protein-coding genes, as well as 24 non-coding RNA genes; enough to build the cellular machinery necessary to transcribe and translate those genes (see figure 1.8). However, of the full set of proteins the mitochondrion uses to function, the large majority (~70%) are not produced via the mtDNA, but are instead created in the cell's nucleus and imported through membrane pores from the cellular matrix. Another interesting aspect of mtDNA compared to nuclear DNA is the Ts/Tv ratio (see section 1.4.1), which is generally higher in the mtDNA [78]. In Chapter 2 (see section 2.4.6), I investigate the Ts/Tv ratio in the mtDNA of many species and make a similar observation. The GC-content of the mtDNA has also been studied in many species and is quite variable [79].

Due to the high copy-number of mtDNA present in each cell and to the short length relative to the other chromosomes, the mitochondrial genome was in fact the first human "chromosome" to be sequenced (see section 1.1). Since then, the mtDNA has been the subject of countless scientific studies in various contexts. The mtDNA is highly conserved and this is further discussed in Chapter 2 and visible on the multiple alignment in figure 2.3. This characteristic combined with the fact that it is solely inherited maternally has made the mtDNA an ideal target for evolutionary studies [82]. Since mitochondria are the main producers of ATP within the cells, mtDNA defects are known to create serious issues and a number of mtDNA variations have been linked to metabolic diseases. Additionally, several cancers and other neurodegenerative diseases have also been associated with particular mtDNA regions or nucleotides [83]. Recently (2015), the UK government was the first country to approve a three parent in-vitro fertilisation approach, which is required when the mother has defective mitochondria [84]. In this case, the mitochondria can be donated by an external person (usually female for technical reasons), thus the offspring will have the usual mix of DNA (maternal and paternal) in its nuclear chromosomes, while its mtDNA will belong to another (female) person. The mitochondrial genome has been the subject of research in other fields such as forensics or DNA barcoding [85, 86]. The high copy-number of mtDNA, increases the chance of recovering the genome from old or low quality samples, which is the reason

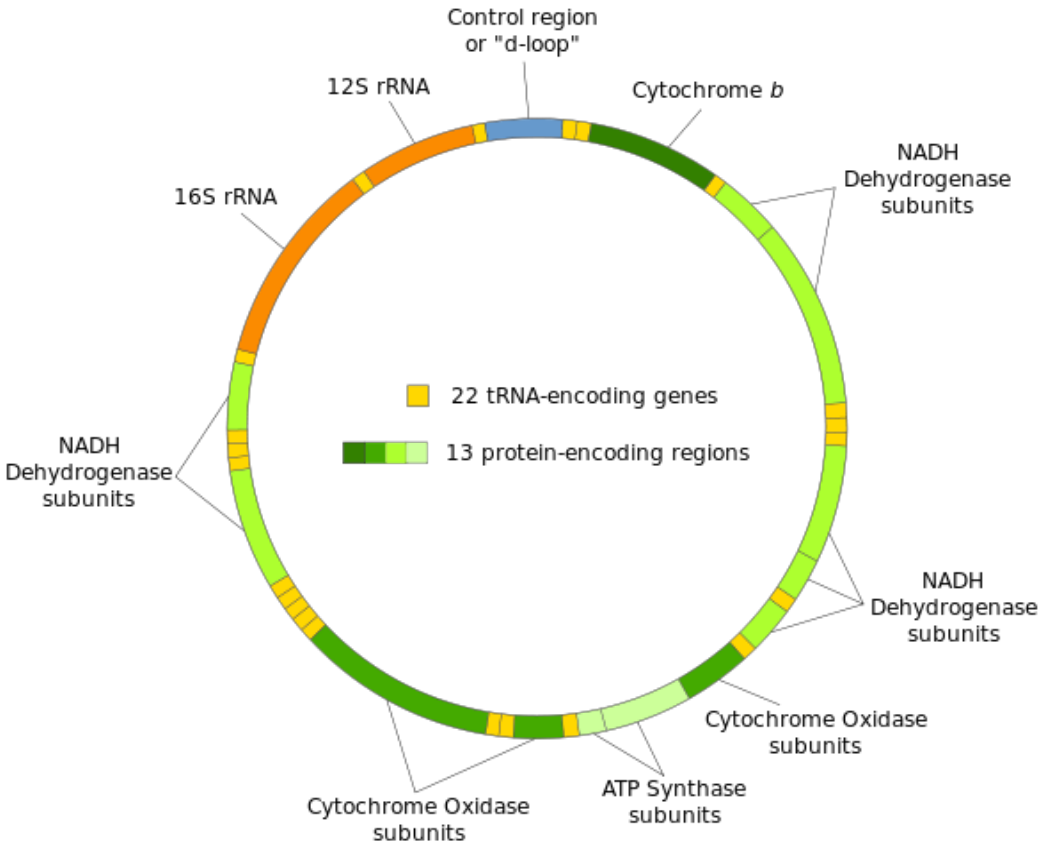


Fig. 1.8 The mitochondrial DNA — mtDNA. The mtDNA is a circular (bacterial like) genome. It is ~16kb long in mammals and contains a small number of protein-coding genes and non-coding RNA genes. The control region, which is entirely non-coding, contains the origin of replication and is the most variable mtDNA region (see Chapter 2). The mitochondrion makes use of these "self-produced" proteins but requires a much larger quantity to function (the rest is imported from the cellular matrix), see section 1.5. Obtained from [87].

why the field of forensic science has extensively investigated mtDNA. On the other hand, the concept of DNA barcoding, which is to use a standardised short genomic region to quickly identify and classify species, focuses on mtDNA sequences since it is present in almost all eukaryotes.

Although, mtDNA has been the focus of many studies, it is common in modern genome-wide assays to remove the mtDNA sequences from the analysis (either from the biological sample directly during DNA extraction and purification or computationally by removing reads that align to the mtDNA). The main reasons for doing this are the sequencing costs and the noise in the data generated by the mtDNA reads. The high copy-number of the mtDNA means that it is sequenced significantly more than the other chromosomes and since the cost of sequencing is generally expensive, removing the mitochondria from the sample before sequencing can reduce cost. Similarly, the high number of sequencing reads which are coming from the mtDNA can create noise in the resulting data. As part of this thesis, I investigate a particular type of mtDNA variation taking advantage of this high number of mtDNA sequenced reads (see Chapter 2).

1.6 Organic tissues

As described above (see section 1.3), DNA is at the core of cellular function producing the proteins that the organelles need to operate the various cellular mechanisms. We also know that there are different types of cells (see figure 1.5) that express different genes to perform different functions (see section 1.4.2). Tissues, which combine to form organs (the largest elements composing organisms), are composed of many cells. More importantly, organs are also composed of different cell types. For example, the brain is made up of two main cell types, neurones and neuroglia. However, these can be subcategorised as there are two different types of neurones (type 1 and type 2) and several types of glia. The cellular composition of tissues is extremely important to consider when performing biological experiments that investigate cell specific characteristics, such as gene expression or protein binding profiles (see section 1.9.1 and section 1.8.3).

Until recently, single cell assays for RNA-seq [88] and ChIP-seq [89] were not available (see section 1.8.3 and section 1.9.1) and to obtain good quality interpretable data, these experiments required a high quantity of material (respectively RNA or DNA) and thus used a large number of cells. Tissues with lower cellular complexity are thus better targets for such experiments and liver, which is composed mostly (~70%) of a single cell type (hepatocyte)

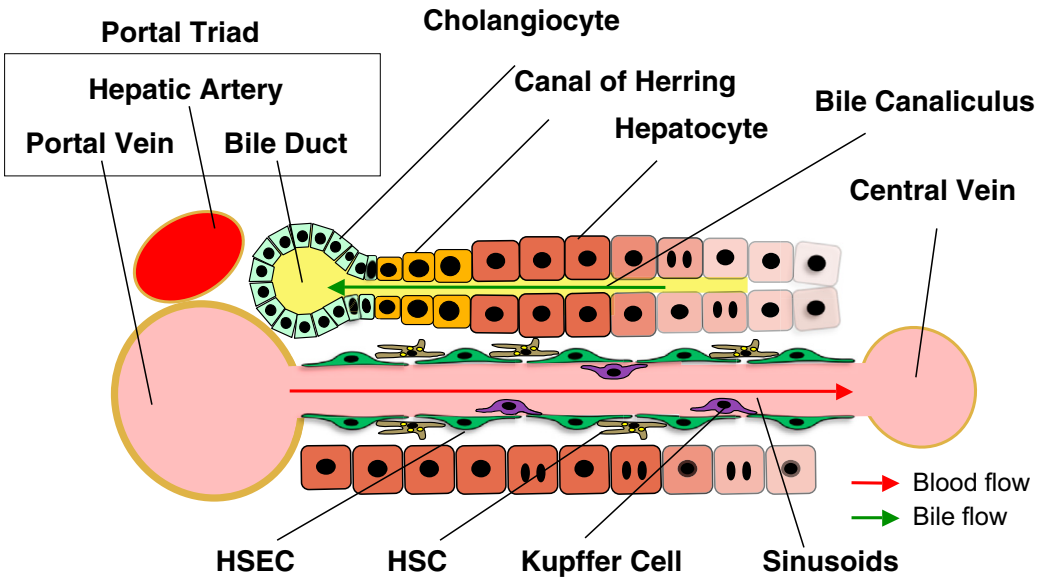


Fig. 1.9 Cell type constituents of liver. This figure depicts a small region of liver tissue, including the liver cells and the blood vessels. The liver tissue is relatively homogenous as it contains mostly (~70%) hepatocytes and is thus an ideal target for ChIP-seq and RNA-seq experiments. Obtained from [91].

is such a tissue (see figure 1.9) [90]. Sorting the cells before performing the experiment is sometimes possible, especially when cells are in suspension (blood cells for example), but this process is very challenging in other contexts. Cells have been cultured in laboratories for many years and cell lines in particular (populations of identical cells descending from the same cell) are ideal targets for biological experiments and have been used extensively. In the context of this thesis, I use data that has been generated from liver samples (Chapter 2) and cultured cell lines (Chapter 3).

1.7 Data generation and reusability

Since the arrival of high performance machines to sequence DNA, scientists worldwide are sequencing continuously. The increase of the total amount of data generated by the field is often said to grow faster than Moore's law, a prediction made by a prominent computer scientist, predicting that the amount of transistors per chip (fixed size) would double every two years, which is still true today (2016). In this context, the field of genomics is often compared to other huge data generating fields or companies and is foreseen to become one of the largest within 10 years [92]. This creates serious challenges for scientists, as they have to analyse more and more data, which also needs to be stored efficiently to enable rapid access.

As the global genomic data amount increases, it is becoming more and more interesting to find ways to reuse this data. Data can be reused in many ways but the most common is to use previously generated data as benchmarks for current studies. Another common method is to combine previously generated data with newly generated data to maximise the examined dataset. In this thesis, I explore two different approaches to reusing publicly available data. First, using previously generated data to answer an entirely different biological question (see Chapter 2), and second to explore a potential computational model to further our understanding and extend a particular experimental method (see Chapter 3).

1.8 Chromatin Immunoprecipitation followed by sequencing (ChIP-seq)

Chromatin Immunoprecipitation followed by sequencing is a modern method used to detect genomic locations bound by proteins such as transcription factors or histones (see section 1.4.2). This method relies on a number of other modern genomic assays; mainly

on being able to sequence DNA (see section 1.8.1) and to map sequenced DNA reads to a reference genome (see section 1.8.2). Therefore, before going into further details about the ChIP-seq method, I first describe the process of sequencing DNA and mapping reads. Additionally, after explaining the ChIP-seq method, I also give an overview of several other genomic analysis methods that are mentioned in this thesis. Finally, since most technologies involved in genomics research are continuously improving, after describing the current state of the art methods, I provide an overview of a number of potential new methods that are likely to be developed in the future.

1.8.1 DNA sequencing

The first DNA sequencing method was developed by Ray Wu in 1970 [93] and then optimised by Frederick Sanger in 1975 [94]. This technique, known as Sanger sequencing, takes advantage of the DNA replication system, which is used when cells divide (see section 1.3.1). This method requires four different samples each containing: a single strand of the DNA to be sequenced, a primer (a short sequence of DNA, usually a dozen bases long, which matches the start of the target DNA strand), DNA polymerase and nucleotides with each bases (ATCG), all are necessary for DNA replication. In addition to this, modified DNA nucleotides that do not contain the chemical properties (3'-OH group) to be bound by additional nucleotides (terminators) are added to each mix, but only nucleotides with a particular type of base per mix. Each mixture will create replicates of varying length that always terminate with a specific base. Comparing the lengths of all the different created fragments, using gel electrophoresis for example, enables the determination of the exact sequence of the initial strand of DNA. This technique yields high quality sequencing results and can be used to sequence DNA up to ~800 base pairs long. Although this method is relatively old, costly (per basepair sequenced), and time consuming, it is still widely used today in contexts that require long sequencing reads (see section 1.9.2), for example to validate results obtained from modern technologies that are often more error prone (see section 2.4.9). Sanger sequencing is also used in many projects that do not require large quantities of DNA sequencing. Most DNA sequencing methods rely on the same underlying principle and make use of DNA polymerase. For this reason, DNA sequencing generally occurs from the 5' to the 3' end DNA fragments.

Next generation sequencing — NGS

Next generation sequencing or NGS, refers to modern high throughput sequencing technologies that enable the sequencing of large amounts of DNA in a short amount of time. Several methods have been developed and are described below. They each have different properties that induce advantages and disadvantages. However, in recent years, one particular technology, referred to as Illumina sequencing (after the company that bought the technology), which was developed in Cambridge in the mid-1990s by Shankar Balasubramanian and David Klenerman, has completely dominated the field [95]. Illumina sequencing relies on the same underlying principle of Sanger sequencing (labelled base termination). However, modern technology now enables different fluorescent dyes to be attached to the different bases and therefore different mixtures are no longer necessary. Another novel element is that the terminators are "temporary", meaning they can be removed and the process of DNA synthesis can continue. A camera is used to identify which label has been attached before the removal of the terminators takes place. To facilitate the automation and increase the throughput of the process, the DNA fragments to be sequenced are attached to a surface (called a flow cell), in an ordered manner. The flow cell is similar to a grid in which each square can contain a specific DNA fragment to be sequenced. Before sequencing starts, a cloning process takes place to increase the number of copies of the DNA fragments per cells (called clusters). Once the sequencing starts, and after enough time for nucleotides to bind, a camera takes a picture, before the terminators are removed and the sequencing continues. Each picture taken by the camera will show, for each cluster, a particular colour, corresponding to the base attached at that time.

Although some technologies are better in particular aspects than Illumina sequencing, this method is the best compromise (quality, throughput, cost, etc.) and this is the main reason it is so widespread today. Nonetheless, the main drawback that Illumina (and most NGS technologies) have is the maximum length of the fragments of DNA that can be sequenced (currently in the low hundreds base pairs maximum). This means that longer sequences of DNA are broken down into small fragments to be sequenced and the resulting sequenced short fragments, called sequencing reads, are then required to be assembled (see section 1.9.2) or aligned to a reference genome (see section 1.8.2).

Other NGS methods

Other NGS approaches based on different technologies have been and are still being developed. Although I do not use any data generated by such technologies, these methods

are being developed and used in particular contexts and each have different advantages and disadvantages. A brief overview of the main alternative next-generation sequencing technologies is provided below.

- **Pyrosequencing:** this technology was developed in the late-1990s and is also based on the underlying concept of Sanger sequencing [96]. Instead of fluorescent labels, enzymes (sulfurylase and luciferase) are used to emit light. A cloning step is also performed prior to sequencing to ensure enough light is emitted (similarly to Illumina sequencing). This method sits between Sanger sequencing and Illumina sequencing both in terms of read length and cost. The main disadvantage is the difficulty in sequencing homopolymers (sequences of multiple identical bases e.g. AAAAA). In recent years, as Illumina has increased its maximum read length, this technology has lost considerable interest.

- **SOLiD sequencing:** Sequencing by Oligonucleotide Ligation and Detection (SOLiD) uses a different approach. Instead of using DNA polymerase to synthesise a new strand, it uses DNA ligase (an enzyme used to repair DNA) to attach different labeled probes [97]. Probes are attached depending on their bases and fluorescence enables the determination of the attached probe. This technique has a very low cost per base however it is more time consuming and has problems with palindromic sequences (sequences that are identical on the reverse strand, e.g. "GATC" will match "CTAG" but if both are read from 5' to 3', they both read "GATC").

- **Ion Torrent:** Ion semiconductor sequencing (also known as Ion Torrent), is a sequencing by synthesis method that does not require any modified nucleotides (neither labeled bases, nor terminators) and no optical devices either [98]. This method relies on detecting the release of hydrogen ions, which takes place when a nucleotide is added to the DNA sequence by the DNA polymerase. Cycles during which nucleotides with the same base are added to the mixture take place until a hydrogen ion is detected. This enables a step by step synthesis, for which each base added is known (by which cycle the method was in when the hydrogen was detected). The main advantage of the Ion Torrent technology is the low cost of the equipment. Even though the cost per base is quite high compared to other methods, it enables research labs with limited funds to explore short sequences like bacteria for example. Similarly to pyrosequencing, this approach has issues with detecting homopolymers.

- **SMRT:** Single Molecule Real Time (SMRT) sequencing, sometimes referred to as PacBio sequencing, was commercialised by Pacific Biosciences in 2011 [99]. This method also relies on sequencing by synthesis with fluorescent labelling, but takes

advantage of nanotechnology. Zero-mode waveguides (ZMWs), nanostructures that allow the observation of small amounts of light are used to observe the fluorescence of a single labelled base. This means that no cloning is necessary in this process. This technology produces very long reads (thousand of base pairs) at a rapid pace. SMRT does also have disadvantages. For instance, the overall throughput is limited in comparison with other high throughput technologies, such as Illumina sequencing and the cost per base of this method is high (the reagents necessary are very expensive).

- **Nanopore sequencing:** This technique also relies on nanotechnology and in particular nanopores (small holes) that are just big enough to let single nucleotides through [100]. When an electric current is run through the nanopore structure, the current fluctuates depending on the shape of the pore. Therefore, as a DNA molecule passes through the nanopore (one base at a time), the current will vary according to the base present in the pore. This makes it possible to determine the sequence of the DNA strand [101]. Nanopore sequencing is the most recent sequencing technology to be commercially available (since 2015). This technique does not require labelling and, similarly to SMRT, produces very long reads (thousands of base pairs) and does so quickly. It is important to note that this approach is the only one that determines the sequence of the "exact" target molecule of the DNA, all other methods detect the matching reverse strand. Finally, although this technology is extremely promising, it is still in its infancy and currently the detection error rates are much higher than other methods.

Except for a number of validation experiments performed using Sanger sequencing, all of the data processed and analysed in the context of this thesis has been generated with Illumina sequencing. However, it is important to note that even though Illumina sequencing technology has dominated the world of large scale DNA sequencing in recent years, novel approaches such as SMRT and Nanopore sequencing could have a significant impact on bioinformatics research in the future. In a similar way that the arrival of NGS revolutionised the approach to genomics (see section 1.1) these technologies could completely change our investigative approaches. Future directions of sequencing technology and their potential impacts are further discussed in the conclusion of this thesis (see section 4.2).

1.8.2 Short read mapping

Modern sequencing techniques yield short sequences that range from a few dozen base pairs up to a few thousand more recently (see section 1.8.1). Performing sequence assembly to obtain the entire sequence of the initial piece of DNA from just the short sequencing reads

is often impossible and requires the use of different methods, which yield longer reads (see section 1.4.5 and section 1.9.2). However, if the sequence of the piece of DNA we are sequencing is already known and only a small number of bases might differ from the known sequence, then the task becomes simpler. This is the case when we sequence the genome of an individual for which the reference genome is available (see section 1.4.5). The aim of the process is now to find exactly where in the sequence the short read belongs. Similarly to sequence assembly, advanced mathematical concepts such as the Burrows-Wheeler transform (see below) have also been applied and algorithms developed to perform this in the most efficient way [102, 103]. Repetitive regions are also a challenge for mapping reads and unfortunately scientists are often required to ignore such regions since it is impossible to determine their exact length (which may vary between individuals) and read coverage.

Burrows-Wheeler transform

The Burrows-Wheeler transform, or BWT, is a process in which the characters of a piece of text are rearranged in a way that if the initial chain of characters contained several repeating subsequences (e.g. words such as "the", "and", "to", or "of", which appear many times in english text), the output text will contain multiple repeats of a single character. The transform is obtained by sorting the rotations of the input text and taking the last character of each rotation to form the output. Usually, a lexicographical order, which defines the order of all characters is used to sort the rotations to ensure that special characters such as ',' and '@' are part of the order. To ensure the reversibility of the transform, a unique end-of-file (EOF) marker needs to be used (or added by the algorithm). Below, table 1.1 shows how the string "^BANANA|" is transformed with BWT to "BNN^AA|A".

Initial string	All rotations	Sorted rotations	Resulting string
	^BANANA|	ANANA|^**B**	
	|^BANANA	ANA|^BA**N**	
	A|^BANAN	A|^BANA**N**	
	NA|^BANA	BANANA|^	
^BANANA|	ANA|^BAN	NANA|^B**A**	BNN^AA|A
	NANA|^BA	NA|^BAN**A**	
	ANANA|^B	^BANANA|	
	BANANA|^	|^BANAN**A**	

Table 1.1 Burrows-Wheeler transform example, modified from [104].

One of the main fields that has taken advantage of BWTs is "text compression"; as repeating characters are easy to compress (e.g. a trivial but non-optimal way of doing this is to use numbers: "aaaabbccc" becomes "4a2b3c"). However, in bioinformatics the usage of another key property of BWTs is used to perform efficient matching of strings in text. In fact, the sub-strings of a piece of text that have identical prefixes appear together after the text is transformed with BWT. This enables DNA alignment programs such as BWA (Burrows-Wheeler Aligner) [103] to efficiently look for positions in the genome to which the sequenced reads map. They make use of the array generated to build the BWT. The table 1.2 depicts this array for the string "googol$" ("$" here is used to as an EOF marker). By keeping the indices of each row in the BWT table it is possible to find the exact positions within the initial text at which the sub-strings are present. For example, the sub-string "go" is present in the range [1-2] in the sorted rotations and the sorted indices (3,0) correspond to the exact positions of the sub-string "go" within the initial text "googol$".

Initial indices	All rotations	Sorted rotations	Sorted indices
0	googol$	$googol	6
1	oogol$g	gol$goo	3
2	ogol$go	googol$	0
3	gol$goo	l$googo	5
4	ol$goog	ogol$go	2
5	l$googo	ol$goog	4
6	$googol	oogol$g	1

Table 1.2 Mapping sub-string with Burrows-Wheeler transform example, modified from [103].

Combining this BWT read mapping technique with graph theory algorithms (used to search the array) and memory reduction techniques (to store the array efficiently) is the approach most popular read mapping programs used today, such as BWA [103], use to map sequencing reads in an efficient manner.

1.8.3 Chromatin Immunoprecipitation followed by sequencing (ChIP-seq)

Many different types of proteins bind the DNA to perform various tasks (see section 1.4.2 and section 1.4.4). To study particular protein-DNA interactions, a method was developed to

locate the genomic regions to which specific proteins bind [105]. This technique is based on the concept of immunoprecipitation, a process used to purify proteins. Antibodies (that bind to specific proteins) are used to attach the proteins to bead structures for precipitation. Since proteins within a cell are continuously binding and unbinding DNA, the first step of chromatin immunoprecipitation is to cross-link (fix) all of the bound proteins to the DNA. This is usually done with formaldehyde and with a sample of a large number of living cells. To enable precipitation, the second step of the process is to break all of the DNA links between the proteins. Since proteins protect DNA, methods that randomly cut DNA (sonication for example) can be used, resulting in a mixture of proteins (still bound to DNA) and small DNA fragments. Immunoprecipitation is then performed to precipitate only the proteins of interest (including the DNA to which they are bound). After the precipitation step, unfixing the proteins frees all of the previously bound DNA material that can be analysed. The first version of chromatin immunoprecipitation used a microarray based technique to identify this DNA by hybridising it to a set of probes (ChIP on chip). This technique enabled the identification of the binding motif, which could then be searched for in the reference genome in order to study genome-wide binding patterns. The modern version of this assay (first published in 2007) makes use of sequencing (using NGS technology) instead of a microarray chip, and is thus called chromatin immunoprecipitation followed by sequencing (ChIP-seq) [106]. The short reads, produced from the sequencing of the protein-bound DNA fragments, are then mapped to the species genome. This creates piles of short reads (called ChIP-seq peaks) that map around the protein bound genomic regions. From this data, it is possible to analyse protein specific binding profiles in a genome wide manner. The whole process is illustrated in figure 1.10.

In reality, the antibody precipitation step is not 100% efficient and the resulting DNA fragments that are sequenced cover much larger regions than the protein bound regions. Nonetheless, protein bound regions are enriched in mapped reads and peaks are still detectable. Software using various mathematical models have been developed to detect these read enriched areas [107, 108]. One of the major biases for ChIP-seq are regions of open chromatin, which are also enriched with sequencing reads. For this reason, it is common practice to perform a “naked” ChIP-experiment (no cross linking or precipitation) or a “mock ChIP” experiment (use of a “mock” antibody that does not attach to any proteins) in parallel (same sample) to the normal ChIP-seq experiment [109]. This type of control experiments enable the comparison between both resulting peak tracks and thus the removal of any peak that is likely not the result of a protein bound region. Most ChIP-seq analysis programs (often called peak callers) are designed to make use of such control experiments.

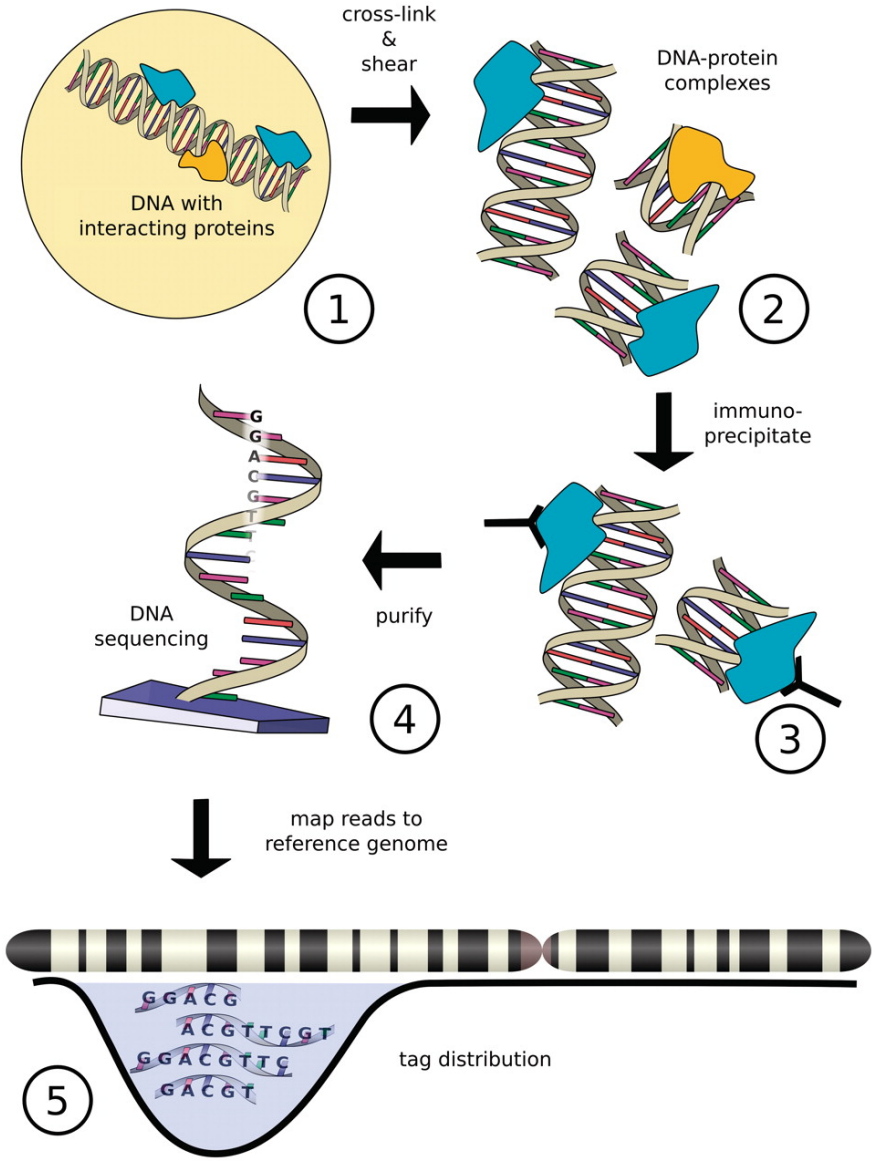


Fig. 1.10 The ChIP-seq experimental method. The first step of a ChIP-seq experiment consists in cross-linking the DNA-binding proteins to the DNA before shearing the DNA, usually using sonication. Antibodies are then used to "pull-down" the proteins of interest along with the DNA to which they are bound (immunoprecipitation). Following this, the DNA is purified and sequenced. Finally, the short sequencing reads are mapped to the reference genome and regions that were bound by the protein of interest are enriched and form read peaks. Obtained from [110].

Since the ChIP-seq DNA fragments are generally sequenced from the 5' to the 3' end (see section 1.8.1), the pile of reads will in fact be slightly shifted towards the 5' end (on each strand), resulting in two shifted peaks on each strand that surround the "true" protein binding site. Since the size of this shift is dependant on the average sequenced fragment size, modern peak-calling software estimate the value of this shift and then move the peaks accordingly to form a single pile of reads over the correct binding site (see figure 1.11). All of the ChIP-seq peak calling executed in the context of this thesis was done with the Model-based Analysis for ChIP-seq (MACS) software [108].

The number of cells required to obtain good results with ChIP-seq varies depending on several parameters such as the antibody quality. Recent studies have shown that in certain circumstances, it is possible to run the experiment with a limited number of cells [111]. However, most ChIP-seq experiments use a large number of cells to ensure the resulting read peaks are well defined. Last year (2015), for the first time, a ChIP-seq experiment was achieved on individual cells (single-cell ChIP-seq) [89]. Although this achievement demonstrates promising technological progress, there are many challenges and limitations in performing single-cell ChIP-seq. I discuss a number of these in Chapter 3 and in the conclusion of my thesis.

Model-based Analysis of ChIP-Seq — MACS

Model-based Analysis of ChIP-seq (MACS) is a software designed to identify protein bound regions on the genome using ChIP-seq data [108]. It performs two key steps to precisely determine each bound location. The first consists in evaluating the peak shift distance between peaks from both DNA strands before shifting the reads accordingly to form unique peaks covering the protein binding sites (see section 1.8.3). In the second step of analysis, MACS uses a Poisson distribution to model read counts along the genome and calculate a p-value per potential peak that the user can then use to filter out unlikely peaks.

To identify the shift size, MACS first makes use of a sliding window approach to identify regions with significant read enrichment versus the control data (see section 1.8.3). It then randomly selects 1,000 of these highly enriched regions and makes the assumption that these correspond to protein bound sites. Each of these regions, is composed of forward and reverse strand reads, each set forming different peaks. Aligning these 1,000 "peak pairs" by the centre point between the forward and reverse strand peaks, generates two (shifted) distributions of reads. MACS then uses the distance between the summits of these distributions as the

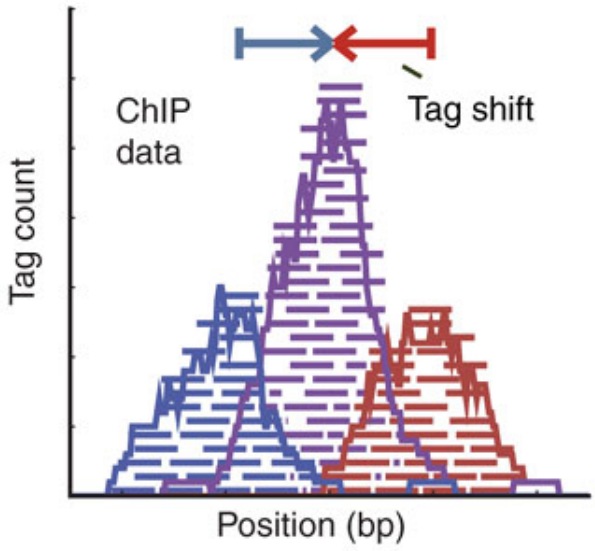


Fig. 1.11 The strand shift of ChIP-seq reads. The pile of aligned reads obtained from ChIP-seq are shifted towards the 5' end, which means that protein bound regions are covered by two read peaks (one for each strand). The blue and read peaks illustrated show each of these "strand specific" peaks. Modern peak-calling algorithms estimate and correct for this shit, generating merged that overlap the "true" bound region. The purple peak in this figure shows this merged peak. Modified from [112].

peak-shift distance d, as illustrated in figure 1.12. Finally, it shifts all the aligned reads towards the 3' end by half the estimated shift distance, $d/2$.

After having performed the shifting of the reads, MACS uses a sliding window of length $2d$ to go over the entire genome and identify regions with significant read enrichment. MACS uses a Poisson distribution to model read counts (a discrete probability distribution that is often used to model the number of times something occurs within an interval of time or space). The Poisson distribution has only one parameter (λ) that captures both the mean and the variance of the distribution, and MACS uses 10^5 as a default value for λ. For each window, MACS calculates the p-value of the window (see section 1.10.1) to determine if a region has significant read enrichment. The set of detected peaks corresponds to these significantly enriched regions (adjacent regions with significant read enrichment are merged).

Furthermore, to avoid local biases, MACS does not use a single distribution to model read coverage over the entire genome, but dynamically estimate the Poisson distribution for the neighbouring region of the current window (regions of 1,000, 5,000, and 10,000 base pairs centred at the current window) and then selects the Poisson parameter for the window as: $\lambda_{local} = max(\lambda_{default}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k})$. This ensures local genome biases do not influence the peak calling procedure, for example "small" peaks in low coverage regions can still be detected.

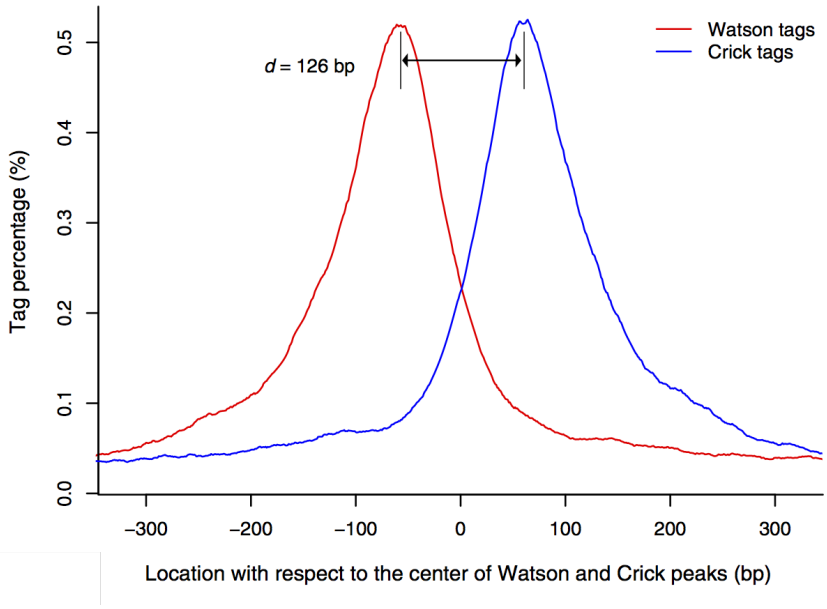


Fig. 1.12 MACS estimates the peak shift distance. To estimate the peak shift distance, MACS randomly selects 1,000 highly enriched regions and aligns their respective forward read (Watson) and reverse (Crick) read peaks by the centre. This figure shows such an alignment, and the shift distance in this case is: $d = 126bp$. Modified from [108].

1.9 Other bioinformatics methods

This section provides an overview of additional bioinformatics methods that are relevant to the work described in this thesis.

1.9.1 RNA-sequencing

RNA-sequencing (RNA-seq) is an experimental method that takes advantage of NGS technology to perform gene expression analysis that was first published in 2008 [113]. It aims at sequencing the RNA content of cells. Since almost all DNA sequencing technologies rely on "sequencing by synthesis" (see section 1.8.1), it is not possible to directly sequence RNA transcripts with the same protocols. RNA-seq relies on an enzyme called a reverse transcriptase, mainly found in viruses, that reverse-transcribes RNA into DNA. Using this enzyme, it is possible to create DNA fragments that are complementary to the RNA transcripts; these are called cDNAs. The cDNA fragments can then be sequenced by "standard" NGS sequencing methods. Similarly to ChIP-seq, the produced sequencing reads can be mapped to a species genome and gene sequences will show enrichments of reads. Different RNAs have different characteristics that can be used to filter them from the initial sample to perform a more specific assay. For example, mRNAs almost always contain a polyA tail (see section 1.3.2) that can be used to assay protein-coding transcripts only [114]. Additionally, as ~90% of all RNA is rRNA, it is often filtered out (to increase the detection power of the other RNA types) [115].

Since many RNA transcripts are post-transcriptionally modified (see section 1.3.2), mapping the reads to the reference genome can prove challenging. Another solution, is to map the reads to a reference transcriptome [116]. Similarly to reference genomes (see section 1.4.5), reference transcriptomes are also produced and maintained. Assembling the cDNA reads "de novo" is the third possible approach to retrieve the RNA sequences [117]. Recent advances in RNA-seq analysis, have developed methods to quantify gene expression and to identify genes that are differentially expressed in different samples [118]. Another recent novelty has been the achievement of performing the RNA-seq protocol on individual cells (single-cell RNA-seq), and although this technique is still being developed the results are promising [88].

1.9.2 Sequence assembly

No technology is currently capable of sequencing large sequences of DNA, such as chromosomes or even entire genomes (see section 1.8.1). This means that in order to obtain the sequence of a long region, it is necessary to break it down and sequence the smaller fragments. This is usually repeated several times to increase the sequencing coverage (the number of times a sequence is sequenced). The process of putting the short sequencing reads back together is called sequence (or genome) assembly [119]. The basic principle of this is to look for overlapping regions between the reads and to merge these reads into continuous segments (contigs). Sequence assembly can be performed with the help of existing similar sequences (genomes of closely related species for example). Alternatively, it can be done without any external support, just using the short sequencing reads, and this approach is called de novo assembly. Several sequencing methods have been used to sequence and assemble genomes, but usually technologies that produces longer reads are preferred (especially for de novo assembly). The human genome was almost entirely built from Sanger sequencing. Genome assembly is as a field of bioinformatics that has benefited significantly from advanced mathematical models, such as De Bruijn graphs (see below). However, despite the progress in the theoretical models used and the computational tools on which the algorithms run, assembling long sequences of DNA is still a complex task which can take a lot of CPU time. Luckily, this process can often be avoided by obtaining the reference genome from a species of interest, to which short sequencing reads can be mapped to obtain the entire sequence of the long piece of DNA (see section 1.4.5 and section 1.8.2).

De Bruijn graphs

Graphs are mathematical structures consisting in vertices that are connected to each other with edges (see figure 1.13). They are used to represent pairwise relations (edges) between objects (vertices). When the relations are reciprocal, a simple line is used to draw edges, while when they are unidirectional, an arrow is used; such graphs are respectively called undirected and directed graphs. De Bruin graphs are a particular type of directed graphs. They are defined by a set of symbols $S = (s_1, ..., s_m)$ and a dimension n. The mn vertices of a n-dimensional DeBruijn graph correspond to each possible arrangement of length n of the m symbols in S. In fact, a vertex is connected to all of the other vertices that have the same sequence shifted to the left once and any symbol of S at the right-most position. The figure 1.13 shows the DeBruijn graph for the sequence $S = 0, 1$ and we can observe for example that the node "01" is connected to the nodes "10" and "11".

As explained above (see section 1.8.1), it is common in bioinformatics to have a high coverage sequencing data made up of short sequencing reads and DeBruijn graphs are ideal to represent such redundant datasets (overlapping reads). For example, Velvet [120] is a software which makes use of De Bruin graphs to perform de novo assembly of short sequencing reads (see section 1.9.2). It converts the sequenced reads into shorter segments of equal length called k-mers before building a De Bruijn graph with them (all sequences of length 4 present in the sequenced reads would be in the set of 4-mers, this is shown on figure 1.14). Identical k-mers are all represented by a single node, meaning the memory required to store the graph is used efficiently. Furthermore, nodes that are separated only by a single edge can be merged, which simplifies the data structure even more. Finally, sequencing errors or biological variants can be identified as loops in the graph. For the purpose of assembly, loops are generally removed following specific criteria such as keeping the path with the highest coverage. Similarly, paths that diverge but do not reconnect to the main sequence path can be eliminated. The figure 1.14 shows the process of creating k-mers (4-mers in this particular instance) and then building a DeBruijn graph from these k-mers, and figure 1.15 illustrates the De Bruin graph after the simplification step (merging nodes that are separated by a single edge).

1.9.3 Multiple alignments

From an evolutionary perspective, it is very interesting to compare different sequences that share a common ancestor, exploring how these have diverged over time. One method of comparison is to directly use the sequences, just in terms of the bases, and explore which regions look the same and which others have changed significantly. Using the similar (or even identical) sub-regions, it is possible to align sequences. The process of aligning two sequences is called pairwise alignment. The computational requirement to execute this task is generally low and for extremely small sequences pairwise alignment has even been done by hand. Aligning multiple sequences (multiple alignments) is done in the same manner, except with the aim to compare multiple sequences to each other. This becomes an increasingly challenging computational task that complex probabilistic models have been applied to solve [121, 122]. In Chapter 2, a multiple alignment of mtDNA genomes is used to show the location of genomic events in multiple species (see figure 2.3).

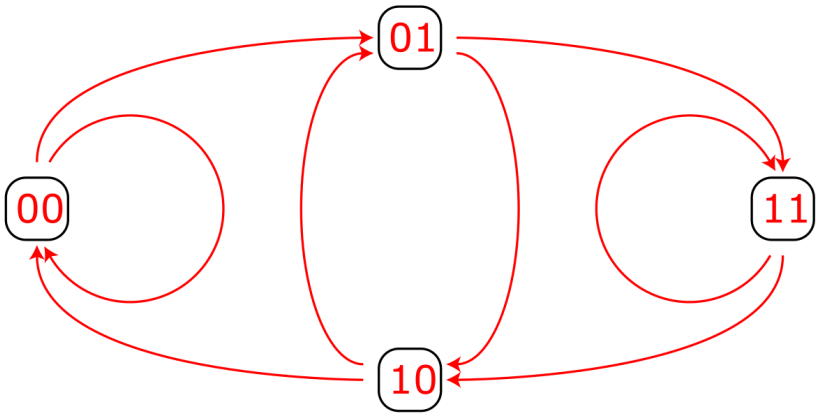


Fig. 1.13 Example De Bruin graph. This is an example De Bruin graph for the sequence $S = \{0, 1\}$. In De Bruin graphs, vertices are (directly) connected to all of the other vertices that have the same sequence shifted to the left once and any symbol of S at the right-most position. For example, here the vertice "01" is connected to the vertices "10" and "11". Modified from [123].

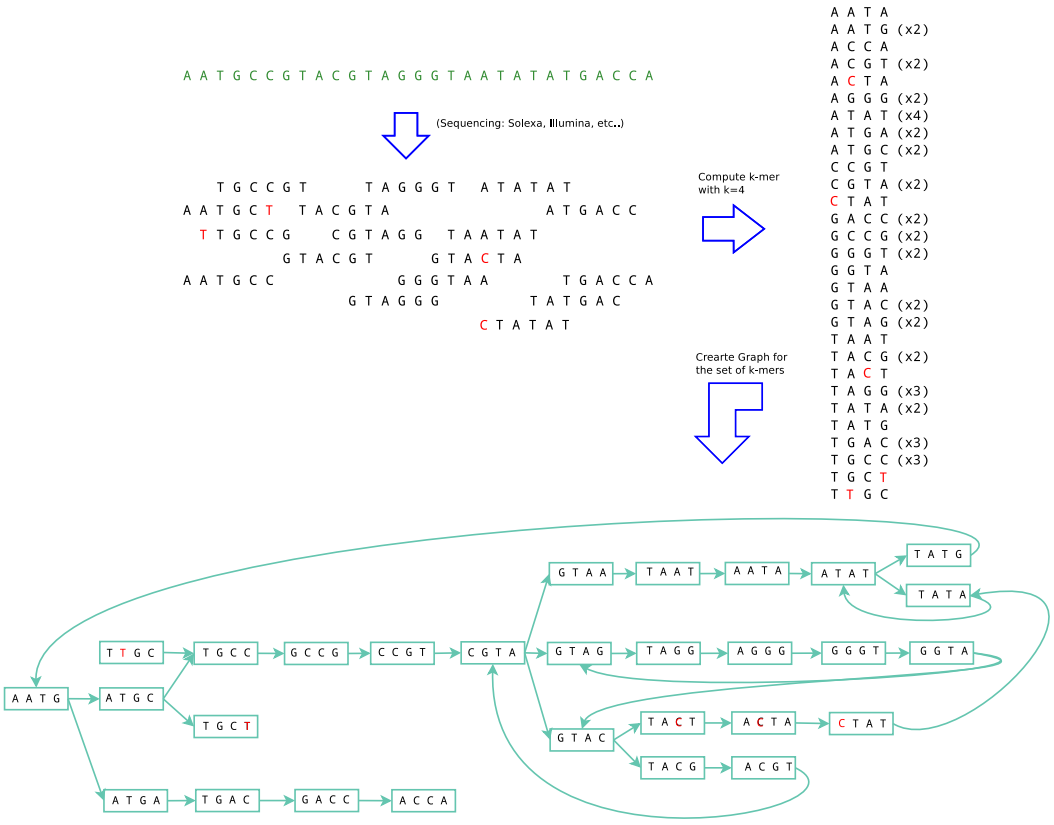


Fig. 1.14 Velvet De Bruijn gaph building process. This example shows an initial sequence and the respective short reads generated by the next generation sequencing process. The set of corresponding k-mers (4-mers in this case) are listed and the initial (non-simplified) De Bruijn graph is illustrated. Obtained from [124].

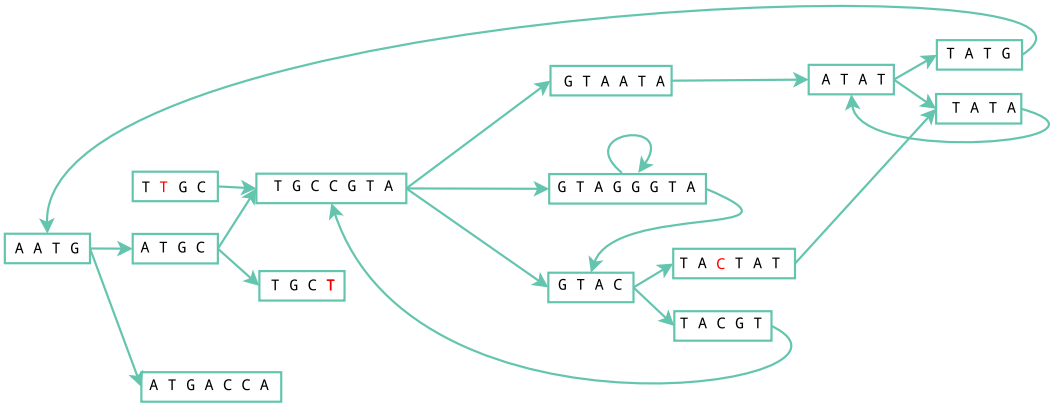


Fig. 1.15 Example De Bruin graph simplified. The De Bruijn graph illustrated here corresponds to the graph from figure 1.14 after simplification. Nodes connected to each other by a single vertices have been merged together. Obtained from [125].

1.10 Technical background

1.10.1 Statistical methods

A common theme in biological research is to compare different experimental outputs, to observe the differences between healthy and diseased samples for example. Such comparisons are generally either done in order to demonstrate that the results are different, or on the contrary, that they are similar to other results. In order to ensure that these types of comparisons are made in a consistent manner, theoretical methods, called statistical tests, are generally applied. Although many different tests have been devised for different situations, such as the number of data dimensions, they are all based on the same underlying mechanism. A null hypothesis is formulated, and will most often be that the two datasets are not different or that there is no relationship between them. The test will then calculate the probability of observing our sampled results (or more "extreme" results) if the null hypothesis is in fact true. This probability is often called a "p-value" and also serves as a measure of significance of the difference between the datasets; the smaller the p-value is, the more different the data are. Furthermore, thresholds are used by scientists to decide if the null hypothesis is likely to be correct or not, the most common threshold is 5%, meaning the null hypothesis will be rejected only if the p-value is smaller than 5%.

The main statistical test used in the work described in this thesis is Fisher's exact test, which is a test suited to compare two datasets of categorical data having only two levels, for example the presence or absence of a characteristic. The null hypothesis to be accepted or rejected is that the two datasets are independent. In other terms, Fisher's exact test can be applied to a 2 x 2 contingency table (e.g. table 1.3) and the p-value is calculated according to the following equation ($n = a+b+c+d$):

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

For example, if we have two different samples of cells and we are investigating the expression of a particular protein, we can build a 2 x 2 contingency table (see table 1.3). In the first sample, there are a cells that express the protein and c cells that do not, similarly in sample 2, there are respectively b cells that express the protein and d cells that do not. The p-value can then be calculated according to the equation above and the null hypothesis (that the expression of the protein is independent from the cell sample) can be accepted or rejected depending on the chosen threshold.

	Sample 1	Sample 2	Totals
# cells protein is expressed	a	b	a + b
# cells protein is not expressed	c	d	c + d
Totals	a + c	b + d	n = a + b + c + d

Table 1.3 Contingency table example.

Another test used in this thesis is the Pearson's chi-squared test (or χ^2test), which is also used on categorical datasets, but aims to identify if a sampled dataset is different from a theoretical distribution (this test can also be used to assess if two categorical variables are independent, e.g. within a population are eye colour and hair colour independent). In a χ^2test, the test statistic follows a χ^2 distribution when the null hypothesis is true. A χ^2 distribution with k degrees of freedom is the distribution of a sum of the squares of k independent normal random variables.

For example, if the category proportions of a dataset are known but these proportions differ in a specific sub-sample, the chi-squared test can be used to assess whether this variation is due to chance or not (if the sub-sample proportions are or aren't significantly different to the dataset proportions). The categories must be mutually exclusive and their total probability must equal 1 (meaning that all categories must be included). The null hypothesis of the chi-squared test is that there is no difference between the distributions (sample and theoretical). The following equation is used to calculate the test statistic χ^2 (n = number of cells in contingency table, O_i =number of observations (in the sample) of category i, E_i = expected number of observations of category i (according to the theoretical distribution), N = total number of observations in the sample, Np_i = fraction of elements of category i in the theoretical distribution):

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} p_i \left(\frac{O_i/N - p_i}{p_i} \right)^2$$

The number of degrees of freedom k is calculated as the number of cells in the contingency table (n) minus the number of rows minus one times the number of columns minus one, $k = (r-1)(c-1)$. Finally, we can calculate the corresponding p-value using the χ^2 distribution with k degrees of freedom (see figure 1.16).

Finally, another important parameter to assess between two datasets is whether those data elements are correlated in any way, in other words are they dependent on each other. Pearson's correlation coefficient is a measure of linear dependency between two random variables

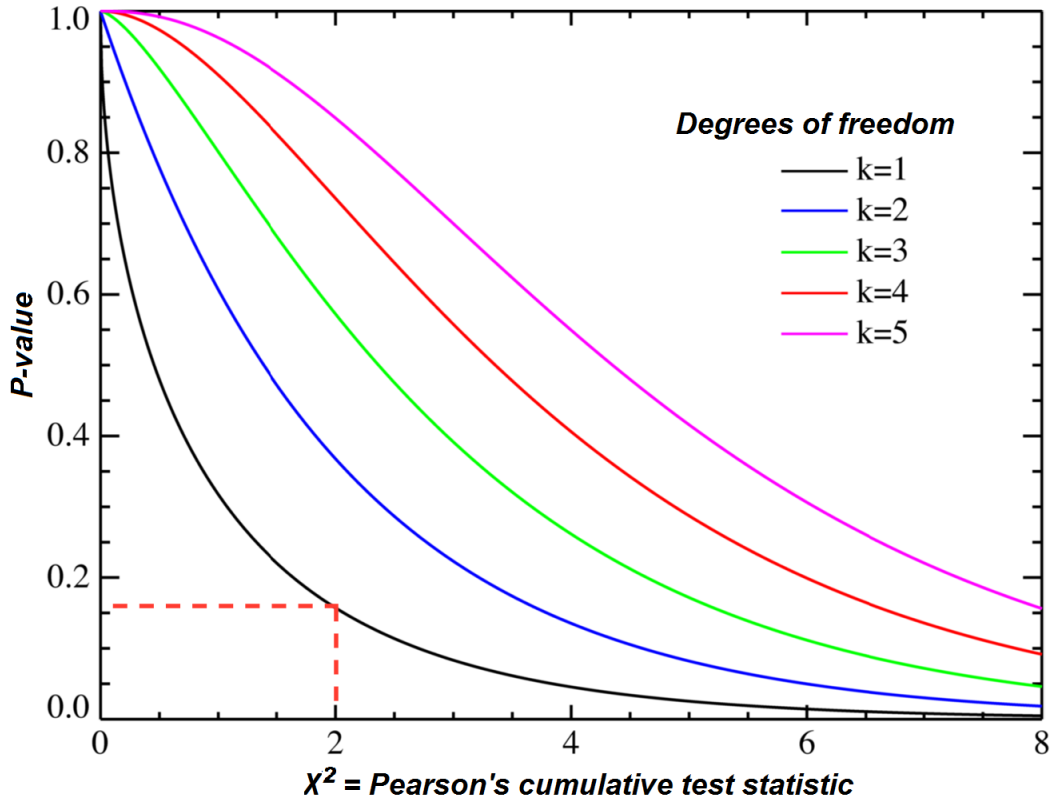


Fig. 1.16 This example illustrates the relation between the p-value and multiple χ^2 distributions. In this case, if we assume that $\chi^2 = 2$ and we have one degree of freedom ($k = 1$), then we can calculate the p-value from the respective χ^2 distribution, $p = 0.16$. Modified from [127].

(datasets). This coefficient (noted r for samples and ρ for populations) takes values between -1 and 1, with 1 corresponding to a perfect positive correlation, -1 corresponding to a perfect negative correlation, while it takes the value 0 when the data is uncorrelated (linearly). The coefficient is calculated by dividing the covariance of the two datasets by the product of their standard deviations. The covariance is a measure of joint variability of two random variables, it is positive when large values of one variable correspond to the large values of the second (and similarly the smaller values of one match the smaller ones of the other), and it is negative in the inverse situation (larger values of the first variable correspond to the smaller values of the second and vice versa). It is a measure of how much two variables are linearly associated. Mathematically it is the mean value of the product of the deviations of two variables with their respective means $cov(X,Y) = E[(X - E[X])(Y - E[Y])]$. Pearson's correlation coefficient is a normalisation of the covariance of two variables, it is their covariance divided by the product of their standard deviations. The standard deviation of a variable is a measure of its variability and is calculated as the square root of the variance, which is a variable's expected squared deviation from its mean ($\sigma = \sqrt{Var(X)} = \sqrt{E[(X - E[X])^2]}$). The Pearson's correlation coefficient is calculated according to the equation below (cov : covariance, E : expected value, σ : standard deviation).

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X - E[X])\,(Y - E[Y])\right]}{\sigma_X \sigma_Y}$$

1.10.2 Machine learning algorithms

Machine learning is a modern field of statistical/computational modelling, in which algorithms are designed to learn from the input data and make according predictions [126]. These programs are implemented to optimise their predictive methods as they process the data, thus "learning" from it and continuously evolving and improving. There exist two main types of machine learning models. The first are designed to be "trained" on a specific dataset, before being "applied" to the real data. The training dataset consists in data for which the correct predictions are known and this dataset should be representative of the real data. This approach is called "supervised learning" and in other words means that the predictive model is fitted to a training dataset and then used to make predictions on a different, generally much larger, dataset. Supervised learning models do not evolve after the training phase. On the contrary, the second models are designed to find patterns in datasets for which predictions have not been made and without the use of a training dataset. This approach, called "unsupervised learning" is generally used to find data clusters in large high-dimensional datasets.

Two supervised learning methods were used in the work described in this thesis (see Chapter 3):

- **Logistic Regression Model:** logistic regression is a probabilistic model used to categorise data in which multiple independent variables determine the category. In its simplest form, the data is categorised in a binary manner (two categories). It is analogous to linear regression (fitting a linear function of the form $f(x) = ax + b$ to a dataset minimising the sum of the perpendicular distances to each point), but instead a standard logistic function is fitted to the dataset of binary variables. The standard logistic function is a sigmoid curve ("S" shape) with values between 0 and 1 (thus interpretable as probabilities). The logistic regression equations is the following (e: natural logarithm base, β_0 and β_1: regression coefficients):

$$f(x) = \frac{e^{(\beta_0+\beta_1 x)}}{1 + e^{-(\beta_0+\beta_1 x)}}$$

  For example, figure 1.17 illustrates the logistic regression of a set of data points corresponding to the number of hours students have studied and whether they have passed or not. Similarly to other regressions, once the function is fitted (for example using a training set), it can be used to make predictions. From the function in figure 1.17, we can predict that the probability of passing the exam having studied 2 hours is 0.25. Multinomial logistic regression, a generalisation of binary logistic regression can be used with datasets that have more than two categories.

- **Support Vector Machine (SVM):** this categorisation model makes use of a geometric construction to categorise data in a binary manner. The data points of a dataset form an n-dimensional space, each point represented by a vector. For example, data points with two variables form a 2-dimensional space, and can be represented as points on a standard two axis plot. For linearly separable binary categorical data (data for which a hyperplane can be placed such that all data points from one category are on the same side of the hyperplane and all of the other points are on the other side), SVMs construct hyperplanes that optimally separates the data into 2 categories, maximising the distance between them. The distance between the two hyperplanes separating the categories (see figure 1.18), is called the margin and the maximum-margin hyperplane is the one that is in the middle of the two. The points, or vectors, that lie on these margins are called the support vectors. Predicting the category of a new datapoint can then be done by assessing on which side of the hyperplane does it lie. Formally, for a data set $(\vec{x_1}, y_1), ..., (\vec{x_n}, y_n)$ with $\vec{x_i}$ the vector of variables and

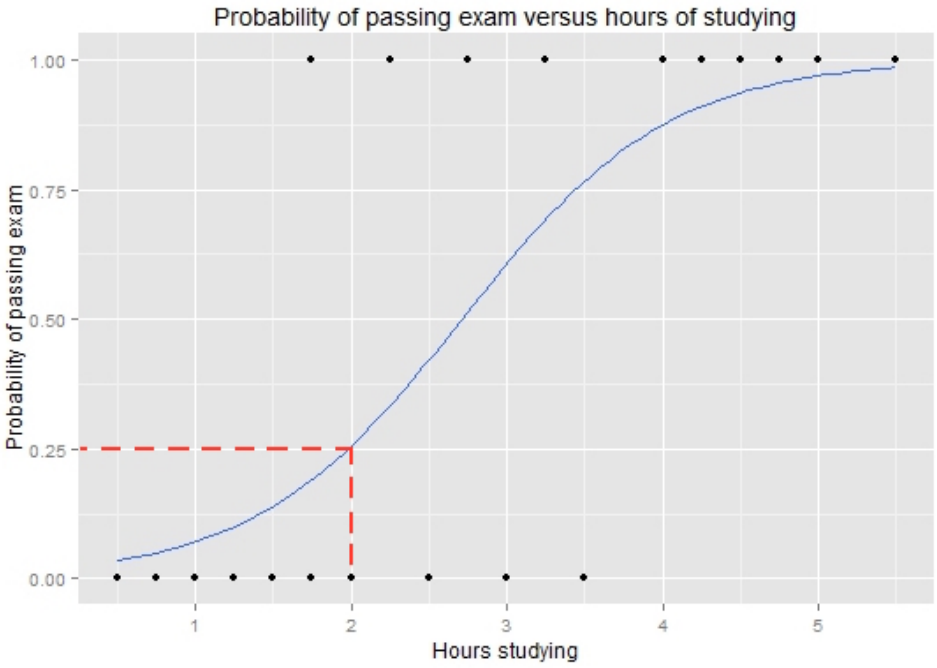


Fig. 1.17 This example illustrates the function obtained by logistic regression on a set of data points corresponding to the number of hours students have studied and whether they have passed or not. Similarly to other regressions, once the function is fitted (for example using a training set), it can be used to make predictions. For example, from this function we can predict that the probability of passing the exam having studied 2 hours is 0.25. Modified from [128].

y_i the category represented as 1 or -1, the equations of the margin hyperplanes are $\vec{w} \cdot \vec{x} - b = 1$ and $\vec{w} \cdot \vec{x} - b = -1$ with $\vec{w}$ normal to the hyperplane and $\frac{b}{\|\vec{w}\|}$ equal to the distance between the hyperplane and the origin (see figure 1.18). Since we assume linearly separable data, all of the points of each category should be above, respectively below their margins. Formally this means that $\vec{x_i} \cdot \vec{w} - b \geq 1$, if $y_i = 1$ and $\vec{x_i} \cdot \vec{w} - b \leq -1$, if $y_i = -1$, or (combination of both equations) $y_i(\vec{x_i} \cdot \vec{w} - b) \geq 1 \forall i$. The margin (distance between the margin hyperplanes) is equal to $\frac{2}{\|\vec{w}\|}$ (see figure 1.18) and thus maximising the margin means minimising $\|\vec{w}\|$. So the optimisation problem to find the maximum-margin hyperplane (to be used to separate the data) is: $min\|\vec{w}\|$ with $y_i(\vec{x_i} \cdot \vec{w} - b) \geq 1 \forall i$. Furthermore, predicting the category of an additional point $(\vec{x_{new}}, y_{new})$ with the previously determined optimal hyperplane defined by the two parameters $\vec{w}$ and b, can be done simply as $y_{new} = sgn(\vec{w} \cdot \vec{x_{new}} - b)$, with $sgn(x) = -1$ if $x < 0$, 0 if $x = 0$, and 1 if $x > 0$.

If the data is not linearly separable, then a "soft margin" approach is necessary, introducing an additional parameter that identifies how many data points can be ignored to fit the hyperplane. Finally, to categorise more complex non linearly separable data, SVMs have been generalised to use non-linear geometric constructions such as polynomial hypersurfaces, but the underlying principles of the method stay the same.

The prediction performance of machine learning models is commonly demonstrated with a Receiver-Operating Characteristic (ROC) curve and expressed with the numerical value of the area under the ROC curve (AUC value). The ROC curve shows the true-positive prediction rate and the false-positive rate for a range of different model parameters. This means that a ROC curve following the main diagonal (AUC value of 0.5) represents “random classification” as we have the same number of correctly classified and wrongly classified data points. Thus, the better the predictive performance of the model is, the furthest away from the random line the ROC curve will be, and the highest the AUC value will be.

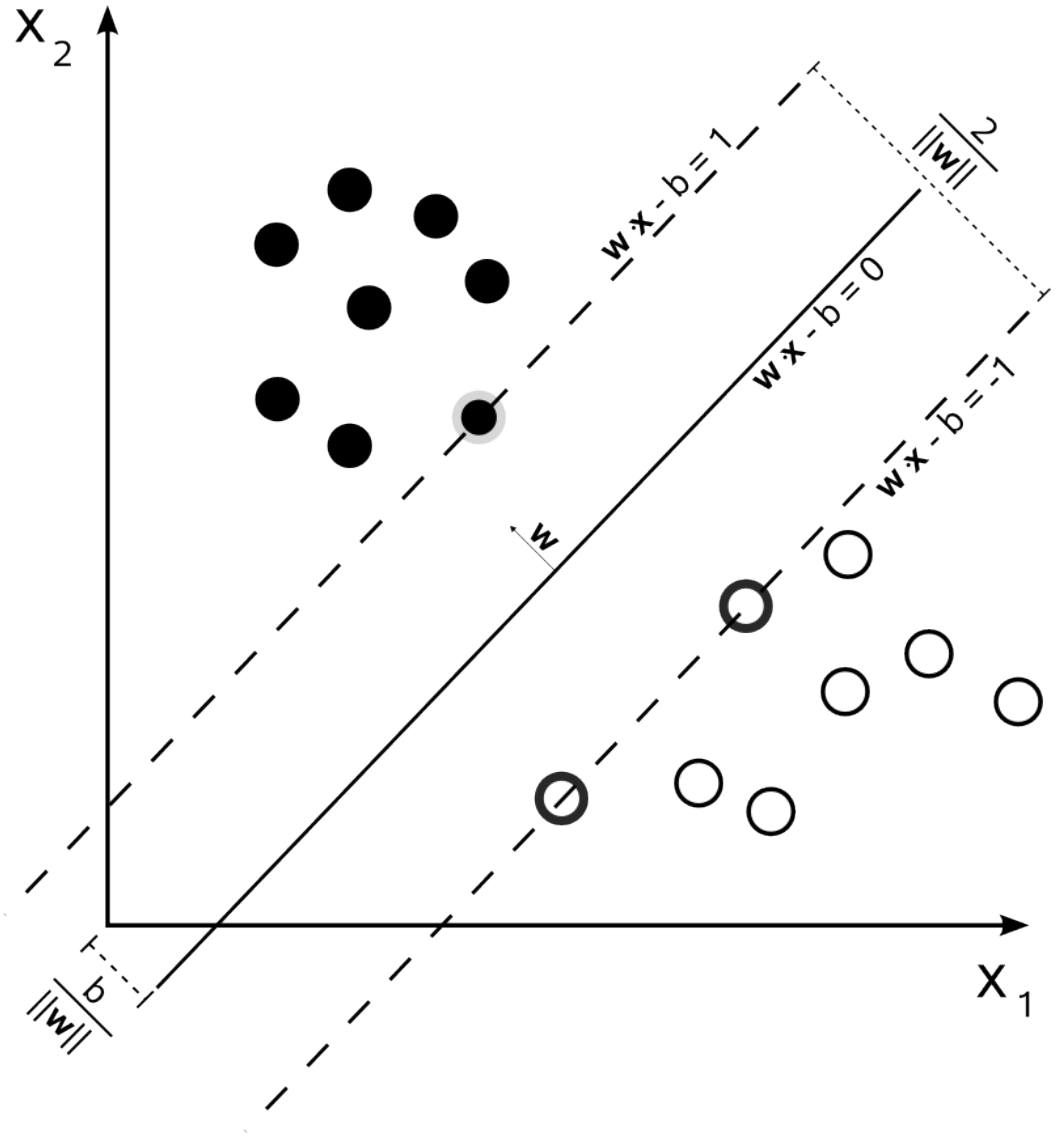


Fig. 1.18 This example illustrates the margin hyperplanes (dashed lines) and the maximum-margin hyperplane (solid line) for a binary categorical data set. The support vectors (data points that if removed would change their respective margin hyperplanes) are highlighted. The margin (distance between the margin hyperplanes) and the distance of the maximum-margin hyperplane and the origin are also shown. Obtained from [129].

1.11 Thesis Structure

In this introductory chapter of my thesis, I provide the necessary biological background to understand the context and motivation behind my research projects. I also describe the main experimental methods that are used to generate the data analysed throughout my work. Finally, an overview of the key mathematical tools that I apply to analyse data in my projects is also given.

In Chapter 2, I describe my first research project, in which I investigate mitochondrial heteroplasmy, a phenomenon consisting in point mutations on the mtDNA genome. Reusing ChIP-seq data previously generated and publicly available for multiple species, I first demonstrate that the mtDNA read coverage in ChIP-seq data is sufficient to detect heteroplasmies. Secondly, I develop and run a heteroplasmy detection algorithm based on a previous study, discovering 107 heteroplasmic positions in 14 species. Finally, I analyse the discovered positions and compare them with previous mtDNA research.

In my second research project, covered in Chapter 3, I look into the potential of ChIP-seq data deconvolution for complex tissues. Taking advantage of the ChIP-seq data generated in the ENCODE project, I first describe the creation of in-silico mixtures of data to emulate the behaviour of ChIP-seq in complex tissues. I then apply several mathematical models to attempt to deconvolve the peak signals obtained from the data. In a second approach, I make use of the same models but this time using a binning strategy instead of peaks to perform deconvolution.

Finally, in Chapter 4, I summarise my results and analyse my overall findings. I then provide an outlook into the potential future directions in which this research could evolve, in addition to discussing the potential advantages that technological advances could bring. Finally, I offer a few concluding remarks to end this thesis.

Chapter 2

Mitochondrial Heteroplasmy

2.1 Summary

Mitochondrial heteroplasmy, the presence of more than one mtDNA variant in a cell or individual is not as uncommon as previously thought. It is mostly due to the high mutation rate of the mtDNA and limited repair mechanisms present in the mitochondrion. Motivated by mitochondrial diseases, the phenomenon has mostly been studied in human samples and in medical contexts. To place these results in an evolutionary context and to explore general principles of heteroplasmy, we completed a large cross-species evaluation of heteroplasmy in mammals. We developed a novel approach to detect mitochondrial heteroplasmy in previously reported ChIP-sequencing datasets, which include concomitant mitochondrial DNA sequenced in the experiment.

In this project, we first demonstrated that the sequencing coverage of mtDNA in ChIP-sequencing experiments was sufficient for heteroplasmy detection. We then implemented a detection method by extension from a previous approach, which aimed to accurately detect heteroplasmies despite the error rate of NGS technology. Applying this method to 79 individuals from 16 species resulted in 107 heteroplasmic positions present in a total of 45 individuals. Further analysis revealed that the majority of detected heteroplasmies occurred in intergenic regions. In addition to validating that ChIP-sequencing data is an appropriate source to identify mitochondrial heteroplasmy, our results suggest that heteroplasmies across vertebrates tend to have similar characteristics as found for human heteroplasmies. Although largely consistent with previous studies, our results provide valuable insights into

mitochondrial heteroplasmy. This work is currently in the process of being published and all the experiments were performed by Diego Villar [130].

2.2 Background

Mitochondrial DNA (mtDNA) forms a circular molecule, which is located in the mitochondrial matrix [131]. In mammals, mtDNA is 16.5 kb long and contains 37 genes [132]. For the most part mtDNA either codes for proteins or for ribosomal RNAs and transfer RNAs, except for a 1kb stretch known as the control region, which contains one origin of replication and both origins of transcription [133, 134]. Several identical mtDNA copies (between 2 and 10 in human) are present in each individual mitochondrion, which means a single cell can contain hundreds to thousands of copies of mtDNA [132, 134]. The mtDNA was the first part of the human genome to be sequenced, and to this day is one of the most studied segments of DNA in human and in many other species [132, 134]. In addition to its high copy number, the mutation rate of mtDNA is significantly higher than that of nuclear DNA [133]. These properties make it common for an individual to have more than one mtDNA variant (more than one type of mitochondrial genome, each defined by a base pair sequence differing by at least one base): this phenomenon is known as heteroplasmy [135] and has been observed and studied in many species and contexts.

In human, hundreds of diseases are linked to point mutations in the mitochondrial genome [136]. Many of these mutations exist in a heteroplasmic state, and the extent of the disease symptoms vary according to the proportion of the deleterious allele [137]. Such diseases include many metabolic diseases, age-related neurodegenerative diseases such as Alzheimer's and Parkinson's, as well as several types of cancer [136, 138–141]. Research in fields such as population genetics and forensics has also focused on heteroplasmy as a way to investigate aspects of inheritance [142]. Studies of heteroplasmy in other taxa have been mostly performed for genetic barcoding or to investigate molecular evolution, but they have generally focused on small controlled datasets. Several studies have also performed cross-species comparisons in a limited number of closely related species, such as different types of bees [86, 143, 144].

Heteroplasmy was first reported in 1983 [145] and has been detected with a variety of methods including Sanger capillary sequencing [146] and pyrosequencing [147]. However, these sequencing methods are expensive and slow, which limited the number of studied samples. More recently, next-generation sequencing (NGS) has been used to study mito-

chondrial heteroplasmy with high-throughput data, and several computational approaches for heteroplasmy detection have been developed [148–151]. The main challenge in using NGS data to detect heteroplasmies is sequencing errors, which tend to be location-specific and thus can be confused with heteroplasmies. To avoid such biases, criteria for NGS-based heteroplasmy detection were developed using PhiX genome simulations and establishing different quality thresholds to identify heteroplasmic positions [148]. Since the heteroplasmy detection power increases with coverage, recent studies employing high coverage sequencing (>1000x) have adapted these criteria [20] (e.g. more lenient thresholds) as well as developed advanced probabilistic models to detect micro-heteroplasmies (i.e. positions with a minor allele ratio below 2%-5%) [150, 152]. In this chapter, we focus on detecting a higher level of heteroplasmy (>15%) using a modified version of the established criteria [148].

Previous heteroplasmy studies used targeted mtDNA sequencing. In most cases mtDNA was extracted from whole blood or buccal tissue, although recently a few studies have investigated a range of tissues [153, 154]. As the cost of sequencing continues to fall, the quantity of datasets being generated and stored is rapidly increasing. Among the benefits of public availability of sequencing experiments is their use to efficiently answer research questions not explored at the time of data generation. Here, we exploit a combination of previously generated and novel datasets resulting from Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) experiments to perform heteroplasmy detection across a range of vertebrate species. Although mtDNA is of the order of 0.1% of all DNA in a cell [155], the high copy number of the circular mitochondrial genome generally leads to it being sequenced many times in ChIP-seq experiments, resulting in a significant proportion of ChIP-seq reads covering the mtDNA.

We first show that ChIP-seq data provides suitable mtDNA coverage for heteroplasmy detection. We then apply a detection method based on the previously mentioned criteria to a collection of ChIP-seq datasets obtained from five previously published studies combined with novel ChIP-seq data (see 2.3.1) and comprising a total of 79 individuals from 16 species. Our findings provide several insights into mtDNA heteroplasmy over a large portion of the mammalian phylogeny.

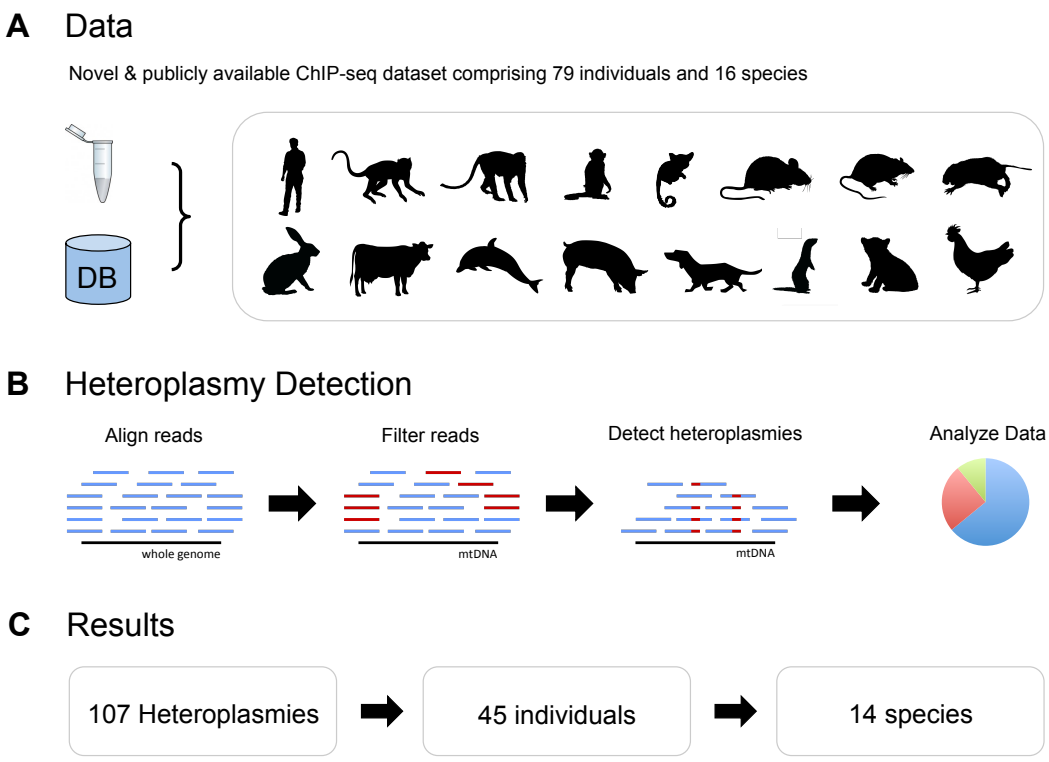


Fig. 2.1 Heteroplasmy detection workflow. The raw read files obtained from the ChIP-seq experiments were first aligned to the respective reference genomes. The aligned reads were then pre-processed, filtering out duplicated reads and extracting reads mapping with a high quality score to the mtDNA. The heteroplasmy detection algorithm was then applied to all the samples. Finally, we analysed the genomic properties of heteroplasmic positions across our vertebrate dataset.

2.3 Materials and methods

2.3.1 Data

We performed ChIP-seq experiments for CEBPA, H3K4me1, H3K4me3, H3K27ac and total histone H3 on a collection of samples from multiple species. Raw sequencing files and detailed experimental protocols have been uploaded to the ArrayExpress database [156] (see Supporting Data for accession numbers). Briefly, chromatin immunoprecipitation was carried out from 0.1-0.5g of liver tissue, using antibodies against H3K4me3 (millipore 05-1339), H3K27ac (abcam ab4729), H3K4me1 (abcam ab8895), total histone H3 (abcam ab1791) or CEBPA (Santa Cruz Biotechnology sc-9314). Histone mark ChIP experiments were performed with automated 96-well protocols in an Agilent Bravo liquid handling robot [157]. A manual version of the protocol was used for CEBPA experiments to allow for higher chromatin input.

The previously published data obtained from [40, 158–161] was obtained from the ArrayExpress database (see section 2.3.11 for accession numbers). We performed the preprocessing and aligning of the reads (see section 2.3.2) so only raw read files (in FASTQ format) were downloaded. A list of these datasets is provided in table 2.1 and experimental details regarding all of these experiments are described in detail in their respective publications, as well as in the protocols in ArrayExpress.

2.3.2 Pre-processing and read alignment

Raw sequencing reads (FASTQ files) were aligned to the whole genome of their respective species obtained from Ensembl (v. 81) [9]. The human samples were aligned to the human reference genome used in the 1000 Genomes Project (GRch37) [162]. Some species were aligned to closely related species' genomes. Specific assemblies and files used for alignment are listed in Appendix A (see figure A.2). Finally, the samples of species, for which a full genome was not available (*O. garnettii, H. glaber, and M. furo*), were aligned to the reference mitochondrial mtDNA sequence obtained from the Nucleotide database [163] (see figure A.2). All of the raw read files were aligned using BWA (Burrows-Wheeler Aligner, see section 1.8.2) [103] with default parameters. The aligned read files in BAM format were then merged per individual using SAMtools merge [164]. Next, we removed duplicate reads with SAMtools rmdup (keeping only the read with the highest mapping quality score per set

Transcription factors and histone marks	**Tissue and species**	**Publications**
HNF4A and CEBPA	liver tissue of five species (*H. sapiens, M. musculus, C. familiaris, M. domestica and G. gallus*)	Schmidt et al. (2010) [158]
CTCF, SA1, NRSF/REST and H2AK5ac	liver tissue of six species (*H. sapiens, M. mulatta, M. musculus, R. norvegicus, C. familiaris and M. domestica*)	Schmidt et al. (2012) [40]
CTCF and YY1	LCLs of seven species (*H. sapiens, P. troglodytes, G gorilla, P. pygmaeus, M. mulatta, P. hamadryas and S. oedipus*)	Schwalie et al. (2013) [159]
YY1	liver tissue for two species (*H. sapiens and M. musculus*)	Schwalie et al. (2013) [159]
CEBPA, FOXA1, ONECUT1, and HNF4A	liver tissue of five species (*H. sapiens, M. mulatta, M. musculus, R. norvegicus and C. familiaris*)	Ballester et al. (2014) [160]
H3K4me3 and H3K27ac	liver tissue of 21 species (*H. sapiens, M. mulatta, C. sabeus, C. jacchus, M. musculus, R. norvegicus, C. porcellus, H. glaber, O. cuniculus, T. belangeri, B. taurus, D. delphis, L. albirostris, B. borealis, M. bidens, S. scrofa, C. familiaris, F. catus, M. furo, M. domesticus and S. harrisii*) [161]; and CEBPA, H3K4me1, H3K27ac, and total Histone H3 ChIP-seq data for the liver tissue of 21 species (*M. mulatta, C. sabeus, C. jacchus, M. musculus, R. norvegicus, C. porcellus, H. glaber, O. cuniculus, T. belangeri, B. taurus, D. delphis, L. albirostris, B. borealis, M. bidens, S. scrofa, C. familiaris, F. catus, M. furo, M. domesticus, S. harrisii and O. garnettii*)	*Novel data*

Table 2.1 Mitochondrial heteroplasmy raw data sources.

of coordinates). Finally, reads aligning to the mtDNA with a mapping quality score of at least 20 were extracted using SAMtools view (parameter q=20).

2.3.3 Heteroplasmy detection and data analysis

The heteroplasmy detection algorithm as well as the analysis described in this chapter were implemented in Python 2.7 [165] using the following scientific packages, SciPy [166], Pandas [167], Matplotlib [168] and Pysam (a SAMtools wrapper). Our algorithm processes BAM files, scanning through the entire mitochondrial genome. For each base, it retrieves the set of reads covering that base. This read set is first filtered according to two criteria, 1) reads that have a Phred quality score lower than 23 at the position are discarded and 2) reads for which any of the 5 neighboring base pairs (both directions) has a lower quality than 15 are discarded. Three further criteria are used to call a heteroplasmy on the filtered read set, 1) At least 20 reads should be present in the set, 2) the minor allele (if it exists) should be present on least 15% of the reads and 3) the minor allele should be present on at least two reads of each strand.

2.3.4 Heteroplasmy validation

Sanger sequencing was performed on 34 heteroplasmic positions to confirm the presence of the two sequence variants identified in ChIP-seq datasets.

MtDNA was extracted from 20 mg of flash-frozen liver tissue from each individual of interest, using a protocol adapted from Ahmad et al. 2007 [169]. Tissue samples were homogenised in 1ml homogenisation buffer (100 mM Tris-HCl pH 7.4, 250mM sucrose, 10mM EDTA) in a Precellys 24 homogeniser, with conditions 5000-3x30-30 and tubes CK14 (Bertin Technologies). Nuclei and cellular debris were removed by centrifugation (1500g 10 min 4C), and the supernatant was centrifuged at 10000g for 10 min at 4C to obtain a crude mitochondrial pellet. Mitochondria were suspended in 480 µL of high salt buffer (Tris HCl 10 mM pH 7.6, 10 mM KCl, 10 mM MgCl2, 0.4 M NaCl and 2 mM EDTA) plus 75 µL 10% SDS, and incubated at 55C for 10 min for protein denaturation and solubilisation. Proteins were precipitated by addition of 200 µL 6M NaCl and centrifugation at 11300g for 20 min. Finally, the supernatant containing mtDNA was precipitated with two volumes of 100% ethanol, centrifuged for 10 min at 10000g and 4C, and washed twice with 70% ethanol. The dried mtDNA pellet was resuspended in 100 µL EB buffer (Qiagen), quantified and diluted to a final concentration of 100 ng/µL.

PCR primers were designed to amplify two independent mtDNA fragments (400-1000 bp) spanning each heteroplasmic position. 100 ng of mtDNA were used as template in a 50 µL reaction with Kapa HiFi PCR master mix (Kapa Biosystems) and the following conditions: 95C 3 min; 20 cycles of 98C 20s, 60C 30s, 72C 1 min; 72C 5 min, 4C hold. Sanger sequencing was performed on each amplicon with primers proximal to the heteroplasmy, using at least two different primers per amplicon (see figure 2.11). Heteroplasmies were considered as robustly validated if both alleles could be clearly detected in the chromatograms of more than 50% of successful Sanger sequencing reactions (14 positions). For an additional 6 positions, the minor allele could be detected at lower levels and frequencies, typically in one to three of the reactions (see figure 2.11).

2.3.5 Coverage and genomic context analysis

The mean coverage per individual was calculated using SAMtools depth, which provides the coverage for each base pair. The coverage ratio was calculated as the fraction of mtDNA bases that are covered by at least 20 reads. The resulting data is visible in figure 2.2. To assign the genomic context of each heteroplasmic position, the Ensembl Variant Effect Predictor [170] was used. For three species (*C. jacchus, M. furo and O. garnettii*), annotations were not available and annotations of closely related species were used (respectively: *M. mulatta, C. familiaris and M. mulatta*). The resulting data is visible in figure 2.8.

2.3.6 Heteroplasmy visualization

PRANK [122] with the genomic model was used to generate a multiple alignment of each species' mtDNA. A trimmed species tree extracted from the Ensembl (v. 81) species tree was used as a guide tree. The human mtDNA gene annotation was downloaded from Ensembl (v. 81) BioMart. Jalview [171] was used to visualize heteroplasmic positions for every species displayed on the previously generated alignment (see figure 2.3). Jalview was also used to display heteroplasmies in their sequence context (see figure 2.4).

2.3.7 Low complexity regions

There are five low complexity regions in the human mtDNA (66 to 71, 303 to 309, 514 to 523, 12,418 to 12,425, and 16,184 to 16,193) [147]. To count the number of heteroplasmic positions occurring within these regions across all species, we used the PRANK generated

multiple mtDNA alignment to map the non-human heteroplasmic positions to the orthologous human coordinates.

2.3.8 Disease associated positions

Human positions were directly compared to MITOMAP [172] annotations to identify potential disease associations. For other species, orthologous human positions for each heteroplasmy were first identified using the UCSC Batch Coordinate Conversion (liftover) tool [173], and then compared to the MITOMAP annotations.

2.3.9 ChIP-seq protein binding assay

We performed peak detection with the Model-based analysis of ChIP-seq (MACS) tool [108] for a range of experiments covering nine species (*H. sapiens, M.mulatta, C. jacchus, M. musculus, R. norvegicus, O. cuniculus, C. familiaris, F. catus, and G. gallus*) and four different proteins (CEBPA, CTCF, FOXA1, and YY1). No peaks where reported on the mtDNA genome in all cases.

2.3.10 Human contamination test

To test if the detection method was subject to false positives by cross species contamination, we simulated human contamination in-silico by creating a mixture of *H. sapiens* (human) and *R. norvegicus* (rat) data. By adding random human sequencing reads to a rat sequence file, which did not contain any heteroplasmies, we created mixture files containing 1% and 10% human DNA. No heteroplasmies were detected by our method in these artificially contaminated files.

2.3.11 Supporting data

All the data used in this chapter is accessible on ArrayExpress, under the following accession numbers. Previously published data: HNF4A and CEBPA ChIP-seq data [158]: E-TABM-722 CTCF, SA1, NRSF/REST and H2AK5ac ChIP-seq data [40]: E-MTAB-437 CTCF and YY1 ChIP-seq data [159]: E-MTAB-1511 CEBPA, FOXA1, ONECUT1, and HNF4A ChIP-seq data [160]: E-MTAB-1509 H3K4me3 and H3K27ac ChIP-seq [161]: E-MTAB-

2633 New data from this project: CEBPA, H3K4me1, H3K27ac, and Histone3 total ChIP-seq data: E-MTAB-3933

2.4 Results

2.4.1 Large Mammalian Dataset

We gathered ChIP-seq data from five previously published studies [40, 158–161] and performed new transcription factor and histone modification ChIP-seq experiments (see section 2.3.1) on a selection of samples that were used in the aforementioned papers. We selected published ChIP-seq datasets (signal and input files) from large cross-species comparison studies to mitigate batch effects. The combined data covers a wide range of species spanning the mammalian clade including primates, rodents and domesticated animals such as dogs, cats and cattle, as well as chicken as an out-group vertebrate species. Most of these samples come from liver tissue (see figure A.2), but some consist of lymphoblastoid cell lines. After analysis (see below and section 2.3), we identified a core set of 16 species for comparison — *Homo sapiens* (human), *Macaca mulatta* (macaque), *Chlorocebus aethiops sabaeus* (vervet), *Callithrix jacchus* (marmoset), *Otolemur garnettii* (bushbaby), *Mus musculus domesticus* (mouse), *Rattus norvegicus* (rat), *Heterocephalus glaber* (naked mole-rat), *Oryctolagus cuniculus* (rabbit), *Bos taurus* (cattle), *Delphinus delphis* (dolphin), *Sus scrofa* (pig), *Canis familiaris* (dog), *Mustela putorius furo* (ferret), *Sarcophilus harrisii* (Tasmanian devil), and *Gallus gallus* (chicken).

2.4.2 ChIP-sequencing data for heteroplasmy detection

A ChIP-seq study generally consists of two experiments that each result in short read sequencing data files (see section 1.8.3) [174]. The first is commonly known as the signal file and contains reads resulting from the ChIP experiment, which when mapped to the target genome produces read clusters (peaks) identifying genomic locations where the proteins targeted by the ChIP antibody were bound. The second data file is a control experiment consisting of a similar process but without the immunoprecipitation step, and is used to control for biased genome-wide read coverage arising from preferential sonication of open chromatin [109]. Thus, the control data generally contains reads that map to the entire genome, with few expected enriched regions. Due to the high copy number of the mtDNA,

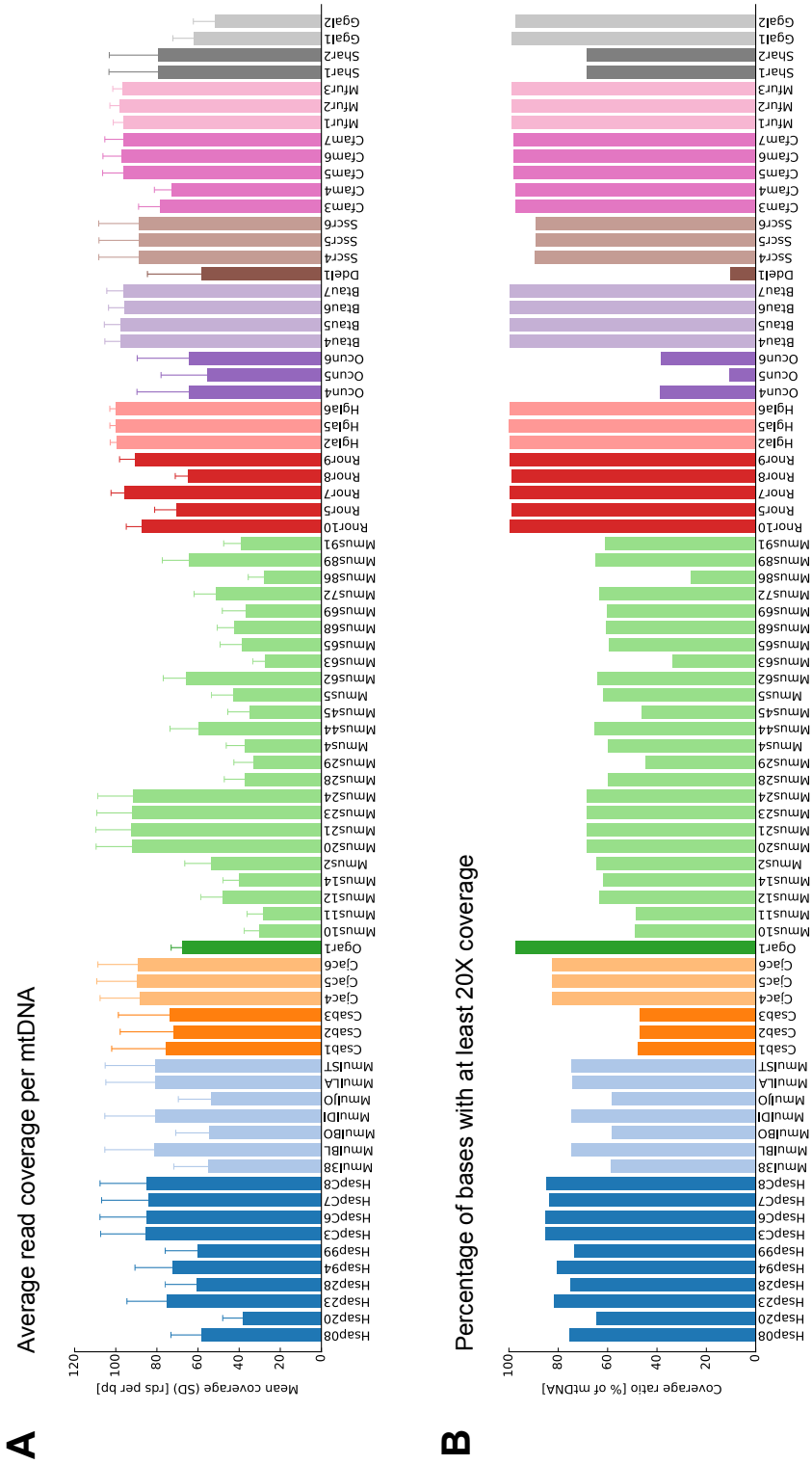


Fig. 2.2 mtDNA read coverage per individual. (A) The mean read coverage per mtDNA basepair for each analyzed individual, colored per species (the error bars represent the standard deviation). (B) The fraction of mtDNA basepairs covered by at least 20 reads (our heteroplasmy detection cutoff) also colored by species.

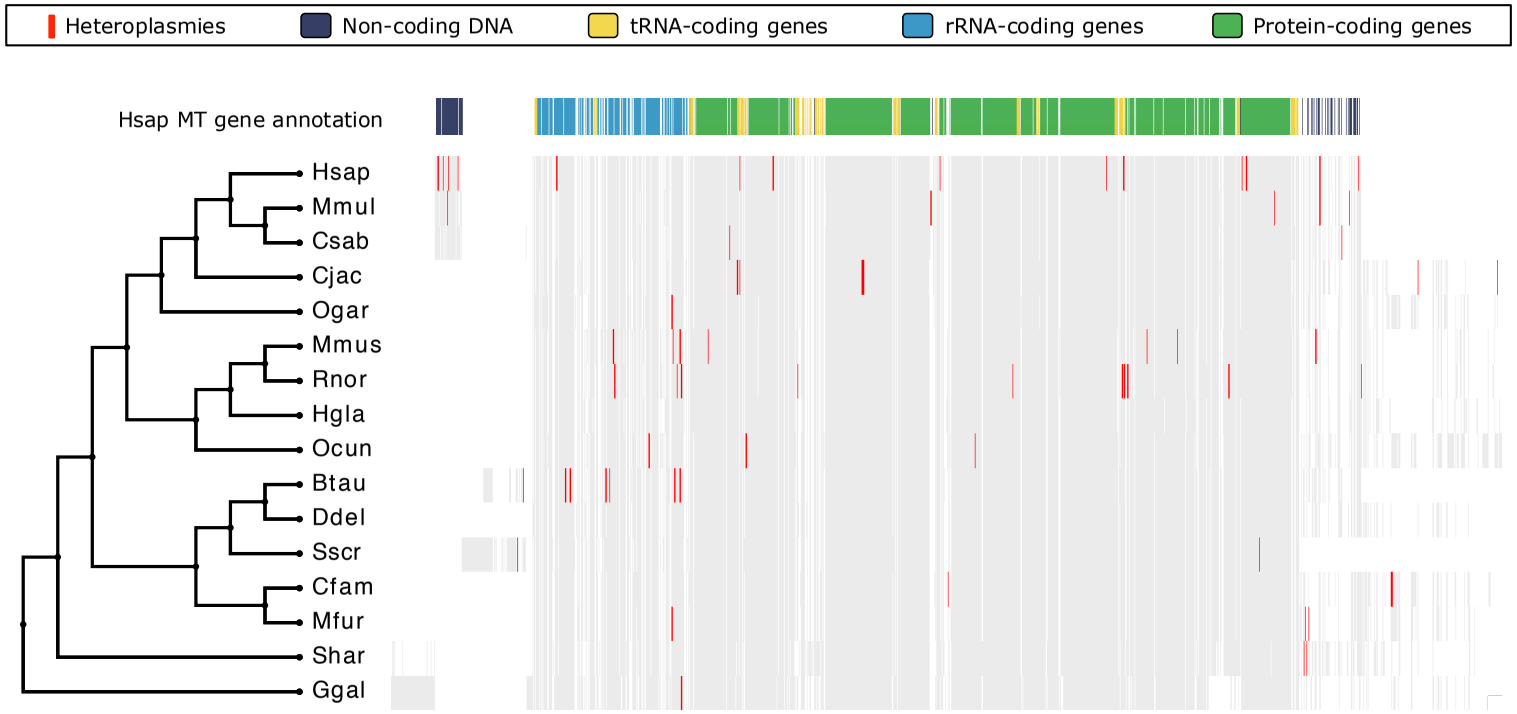


Fig. 2.3 Heteroplasmies in 16 species. Detected heteroplasmy in *H. sapiens* (human), *M. mulatta* (macaque), *C. sabaeus* (vervet), *C. jacchus* (marmoset), *O. garnettii* (bushbaby), *M. musculus* (mouse), *R. norvegicus* (rat), *H. glaber* (naked mole-rat), *O. cuniculus* (rabbit), *B. taurus* (cattle), *D. delphis* (dolphin), *S. scrofa* (pig), *C. familiaris* (dog), *M. putorius furo* (ferret), *S. harrisii* (Tasmanian devil), and *G. gallus* (chicken) displayed in red on the mtDNA multiple alignment with the associated evolutionary tree. The human gene annotation displayed at the top of the figure shows RNA and protein coding genes as well as non-coding regions.

reads within the mitochondrial genome are sequenced many times in both signal and control ChIP-seq experiments. We observed significant read coverage in both data files, even though binding peaks within mtDNA were not detected (see figure 2.2, figure 2.10, and section 2.3.9).

We used ChIP-seq data originally generated to map various histone modifications and transcription factors such as CEBPA and FOXA1, as well as the input/control data for each experiment (see figure 2.1). We merged all experiments corresponding to the same individual, which provides better total coverage and acted as technical replicates (i.e. sequencing of the same biological tissue) for the mtDNA detection experiment. When combined in this way, coverage for each individual was relatively high and homogenous for almost all species, with average mitochondrial coverage above 50X and coverage ratio above 70% (see figure 2.2). As previously reported [148], such coverage levels are adequate to detect high-level heteroplasmies with high specificity. Detailed coverage data for each individual are available in Appendix A (see figure A.1). For six species in the original published studies forming our collected dataset, but not included in the 16 core species (*Monodelphis domestica, Cavia porcellus, Tupia belangeri, Balaenoptera borealis, Mesoplodon bidens, and Lagenorhynchus albirostris*), we observed very low rates of uniquely mapping reads on mtDNA. *C. porcellus* (guinea pig) and *T.belangeri* (tree shrew) have highly fragmented genome assemblies (see section 1.9.2) that may hinder accurate mtDNA read mapping, and *M. domestica* (opossum) is known to have a significantly increased number of NUMTS (Nuclear Mitochondrial DNA sequences) which may also have affected the number of uniquely mtDNA mapping reads [175]. Samples from *Balaenoptera borealis* (sei whale), *Mesoplodon bidens* (Sowerby's beaked whale) and *Lagenorhynchus albirostris* (white-beaked dolphin) were all mapped to the closely related *Tursiops truncatus* (common bottlenose dolphin) species' genome a process that also yielded few uniquely mapping reads to the mtDNA except for *Delphinus delphis* (short-beaked common dolphin), see figure 2.2. Since coverage in these species was insufficient for heteroplasmy detection, we excluded them from further analysis. Finally we also discarded a *Mus musculus* (mouse) individual for which the coverage ratio fell below 10%.

2.4.3 Heteroplasmy detection algorithm

We adapted a previously published heteroplasmy detection methodology [148] for the specific characteristics of ChIP-seq data (see below and section 2.3.3). Briefly, this method is based on a set of criteria to be checked for each mtDNA basepair. In addition to quality thresholds, the algorithm requires a minimum number of reads to be present on each strand. Strand

verification avoids location specific errors that may arise from sequencing errors, since it is uncommon for these to occur at the same location on each strand [148]. Distinguishing characteristics of our method include aligning with BWA [103] (instead of assembling the reads) and a parameter set optimised for ChIP-seq data. This set is generally more stringent than in previous reports [148, 149], and includes a higher base quality threshold (base quality >23) and an added minimum coverage threshold (20 reads). We also increased the minimum heteroplasmy level to 15% (minor allele frequency). Although these changes to the algorithm result in lower expected sensitivity, we do so to optimise specificity from the generally less homogenous sequencing coverage in ChIP-seq samples, compared to that observed in targeted mtDNA resequencing.

Nucleotide repeats present in low complexity regions strongly hinder sequencing quality over those locations. While previous studies have excluded these positions from their analysis, these regions are also more likely to harbour heteroplasmies due to error-prone polymerase activity and limited DNA repair in the mitochondrion. The human mitochondrial annotation database MITOMAP [172] lists several positions as heteroplasmic within these regions, of which we find nine (see section 2.3.7) with our detection method. Since the detection parameters are very stringent, we decided to keep the repetitive regions in our analysis.

In addition to liver samples, we also applied our detection algorithm to ChIP-seq data from several primate species' lymphoblastoid cell lines (LCLs) [159]. We observed that 33% of them expressed more than 25 heteroplasmic positions, which we assume might be due to genomic instability in the immortalised cell lines that could have arisen from a high passage number of the cells [176]. Our results from two *Felis catus* (cat) samples also exhibited a surprisingly high number of heteroplasmies in both individuals (see table A.2). Furthermore, almost all of the positions detected in one cat individual were present in the other, which may be due to low genetic diversity in the source population or the two individuals being siblings. Another possibility is that some contamination occurred in the process. For these reasons, we do not include the primate LCLs or the cat data in the core set of species or the remaining comparative analysis in this chapter.

After the filtering above, our final dataset included comparative heteroplasmy results from liver samples of 79 individuals across 16 species. For these, we found 107 positions in 45 individuals across 14 species (see figure 2.3). A total of 57% of the individuals express heteroplasmy. Our estimate is higher than initial NGS-based reports of human heteroplasmy [148], but consistent with recent reports on high-coverage datasets [152, 153], which also showed that liver tissue expresses more heteroplasmies than other tissues [153]. We find

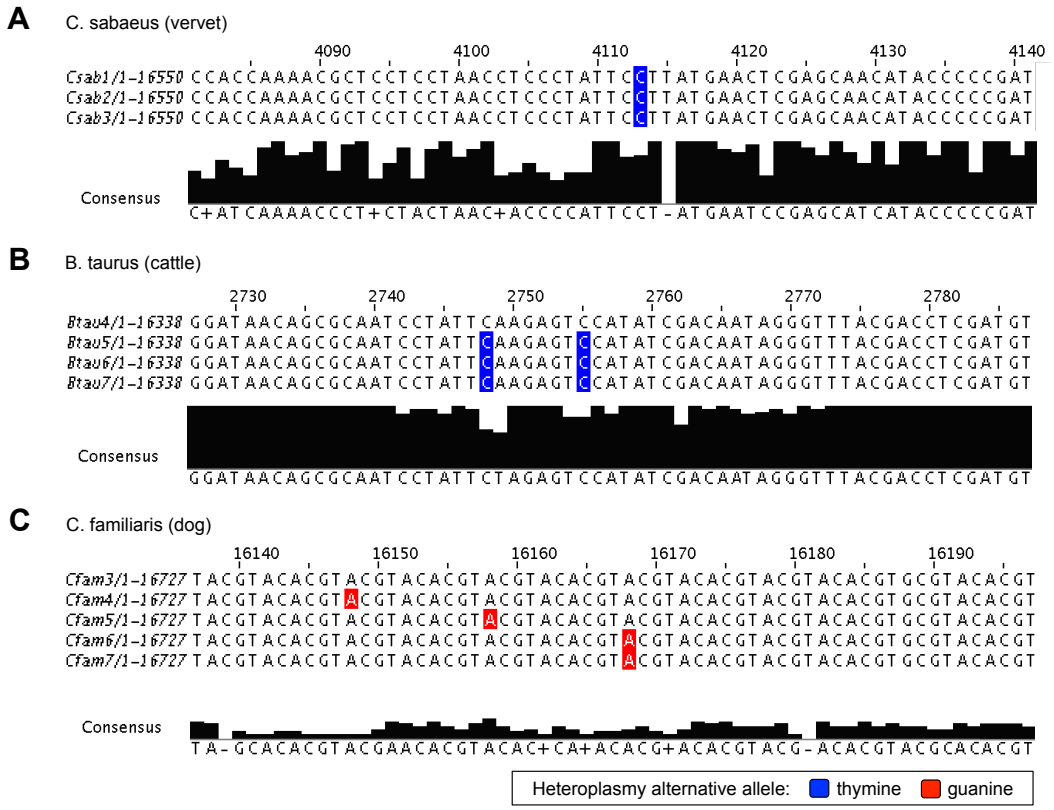


Fig. 2.4 Sequence context of multiple heteroplasmies present in three species (*Chlorocebus sabaeus, Bos taurus and Canis familiaris*). Heteroplasmic positions are coloured according to the alternative allele nucleotide (blue: thymine and red: guanine). The consensus sequence of the multiple alignment is displayed below each species sequences.

heteroplasmic positions in every species except *Heterocephalus glaber* (Naked mole-rat) and *Delphinus delphis* (short-beaked common dolphin).

2.4.4 Heteroplasmic positions present in multiple individuals

Several positions occur in more than one individual of the same species (see figure 2.4 and table A.1), a phenomenon that has previously been observed in human and attributed in part to differential mutation rates across the mtDNA sequence [148]. Since heteroplasmies in human are mostly located at positions with high relative mutation rate [148, 152] and mutation rate patterns are shared across individuals, such positions are thus more likely to exist in a heteroplasmic state in more than one individual. Similar differential mutation rate patterns are likely in other species, and thus the shared positions we observe may also have high mutation rates. Likewise, we asked whether this phenomenon could be linked to sequence conservation, but the heteroplasmic positions occurring in more than one individual do not show evidence of conservation bias (see figure 2.4). That individuals are closely related via their breeding history is another possible explanation.

2.4.5 Read coverage of the heteroplasmic positions

The average number of reads supporting each of our observed heteroplasmic position is 60 (SD 25), which is significantly higher than our 20 read threshold (see figure 2.9). Indeed, there are no observed positions with read coverage of exactly 20 reads and only three positions with coverage of 21 reads. Based on the observed coverage distribution, our detection parameters including the coverage threshold appear to be conservative. For each individual, we compared the average coverage across the mtDNA to the number of observed heteroplasmic positions. We found a minor correlation across the entire distribution (Pearson's r=0.17) (see figure 2.5), but for individuals with high average coverage (>40 reads per mtDNA position), there is essentially no correlation between coverage and number of observed heteroplasmies (Pearson's r=0.05). This result further suggests that our chosen coverage threshold is sufficient for robust detection of high-level heteroplasmies. As expected, the heteroplasmy level distribution is highest at 15% minor allele frequency corresponding to the threshold level (see figure 2.6). Previous studies also report that the majority of heteroplasmic positions occur at the lowest minor allele level [148, 152].

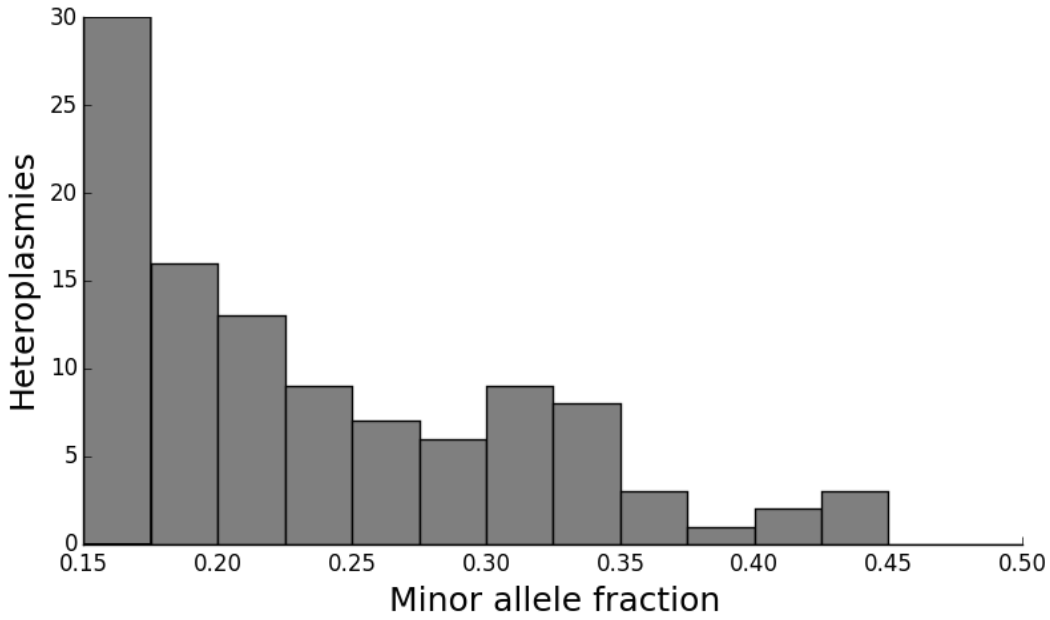


Fig. 2.5 The minor allele fraction of the detected heteroplasmies (heteroplasmy level). This figure shows that most heteroplastic positions are detected close to the detection threshold level of 15%.

2.4.6 Heteroplasmy mutation spectrum analysis

The transition-transversion rate is strongly biased in the mitochondrion [78]. We observe a transition-transversion ratio of Ts/Tv = 3.86 in our results across all species, which is similar to the ratio found in the MITOMAP database (Ts/Tv= 2.95, Fisher's exact test, p=0.31). In fact, the full mutational spectrum we observe in our multi-species dataset (see figure 2.7) is similar to that observed in the MITOMAP database (χ^2 test, p=0.07). Additionally, the fact that the most common Illumina sequencing errors (AC and GT transversions [39]) are rare in our observed multi-species mutation spectrum strongly suggests that we have few false positives due to sequencing errors in our set of heteroplasmic positions.

2.4.7 Genomic location of the detected heteroplasmies

As shown in figure 2.3 and figure 2.8, out of the 107 heteroplasmic positions found, 44 are in intergenic regions with most of these in the hyper-variable D-loop and only five in other mitochondrial intergenic regions. Of the remaining positions, 39 are located in non-protein coding genes (28 in rRNA genes and 11 in tRNA genes) and 24 positions are located in protein coding genes. Of the protein-coding gene changes, 13 are synonymous variations meaning they do not affect the amino acid used in the translated protein, while there are 11 non-synonymous variants in six species. However, most of the observed amino acid changes (7 out of 11) are between residues with similar biochemical properties. In four cases, the observed changes from isoleucine to threonine and from valine to alanine, are modifications between hydrophobic and hydrophilic amino acids.

2.4.8 Heteroplasmic positions associated with disease

For human, about 5% of mtDNA positions are associated with disease [172]. Using the MITOMAP annotations, five positions (15%) among the human heteroplasmies we find are disease associated. This is more than the proportion of positions associated with disease in MITOMAP, but comparable to the previously observed proportion in human from a set of five Eurasian populations (11.8%, Fisher's exact test, p=0.99) [148]. We then asked whether it would be likely that a similar proportion of positions in other species would be considered deleterious. Since there are no MITOMAP-type databases for the other species listing disease associations, we assigned heteroplasmic positions in other species to their orthologous human positions (see section 2.3.8). For the 43 positions in other species that could be confidently

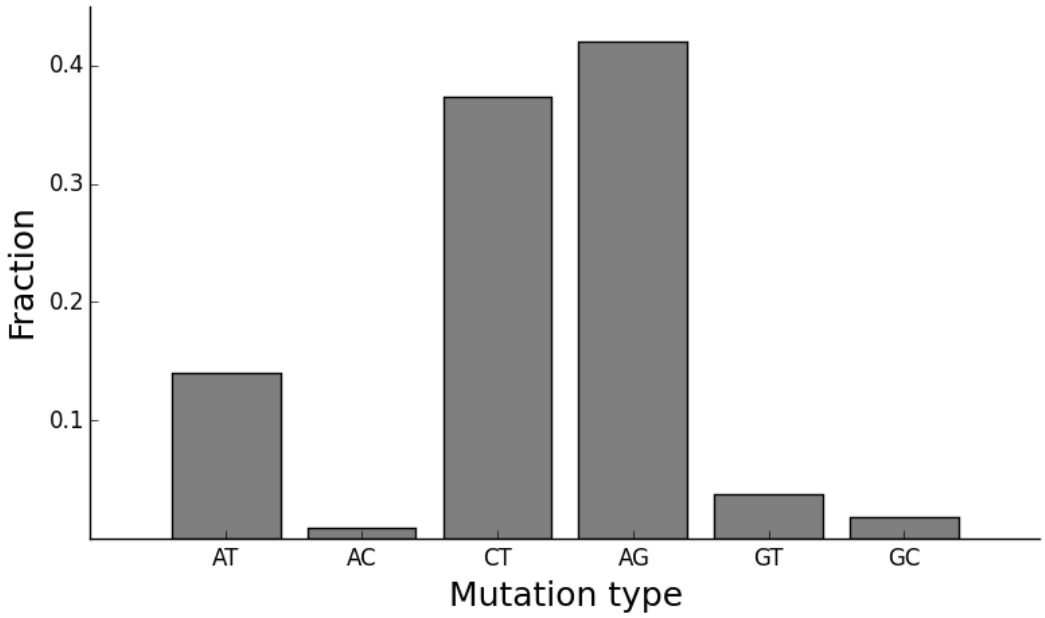


Fig. 2.6 The mutational spectrum of the detected heteroplasmies. The observed mutational spectrum in our results is similar to the spectrum of mutations reported in MITOMAP (χ^2, p=0.07).

assigned an orthologous position on the human mitochondria, two (4.7%) were associated with disease (13882 in Rnor5 — *Rattus norvegicus* and 1068 in Btau4 — *Bos taurus*), a synonymous protein-coding gene mutation in *Rattus norvegicus* and an rRNA mutation in *Bos Taurus*. This is indistinguishable from the baseline disease associated rate in the MITOMAP database (Fisher's exact test, p=0.99), suggesting that mitochondrial disease burden is approximately equal across mammalian species.

2.4.9 Sanger sequencing validation of heteroplasmies

To validate the mitochondrial heteroplasmies identified in ChIP-seq datasets with an independent method, we selected 34 heteroplasmic positions randomly from the total 107 we detected. This validation set includes positions with a range of minor allele frequencies and comprising twelve species. For these, we performed Sanger sequencing on mtDNA amplicons to confirm the presence of two sequence variants at each heteroplasmic position (see section 2.3.4). We validated heteroplasmies with high confidence in 12 positions. A further eight positions showed some evidence for heteroplasmy, while we did not detect any heteroplasmy in the remaining 14 positions (41%). These results indicate that the majority of the heteroplasmies detected in ChIP-seq experiments are true heteroplasmies.

2.5 Discussion

2.5.1 Detected heteroplasmic positions

We find an average number of 2.4 heteroplasmic positions per individual with heteroplasmy. Although this is higher than previously found in human [148, 149], two recent studies show that liver tissue has an increased amount of heteroplasmies compared to other tissues, and most previous studies were performed with whole blood and buccal tissue samples [153, 154]. Our detection method also extends prior approaches by not discarding heteroplasmic positions within low complexity regions (see section 2.4.3), while these were not considered in previous studies. Moreover, we find most heteroplasmic positions with minor allele ratios close to our 15% threshold (see figure 2.6). This is consistent with previous studies, which also found most heteroplasmies at their respective thresholds [148, 152], and suggests that additional heteroplasmies exist beyond the threshold that we used. Finally, since we used relatively stringent detection criteria, it is likely that the heteroplasmic positions we

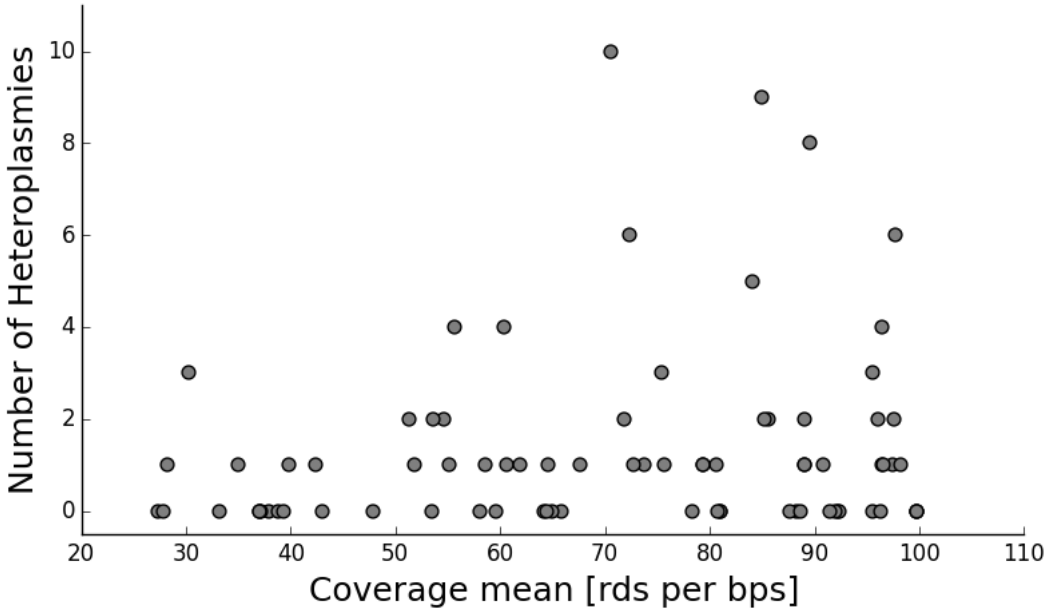


Fig. 2.7 The number of heteroplasmies plotted against the mean coverage for each analysed individual. There is a minor correlation in the data (Pearson's r=0.17), however there is essentially no correlation for individuals with a mean coverage of more than 40 (Pearson's r=0.05).

identified are in fact a subset of all heteroplasmies present in the samples. We hypothesise that the apparent absence of heteroplasmies in *Heterocephalus glaber* (Naked mole-rat) and *Delphinus delphis* (short-beaked common dolphin) is due to the profiling of only a few individuals in these species (one individual of *D. delphis* and three of *H. glaber*).

The genomic location of heteroplasmies across vertebrates is also consistent with previous findings [148, 152, 154] with the majority of positions occurring in the non-coding control region which is the most polymorphic segment of the mtDNA. In fact, most of the heteroplasmies are located in non-coding regions or (structural) RNA genes in which modifications of single nucleotide bases are expected to only have minor effects [177]. Furthermore, heteroplasmies located in protein-coding genes, genome elements that are highly conserved [178], are almost exclusively either synonymous mutations, or result in biochemically similar amino acids.

2.5.2 Sanger sequencing validation

Out of the 34 randomly chosen heteroplasmic positions to be validated, we successfully detected heteroplasmy in 20 (59%). However, for several reasons, the undetected positions may still be true heteroplasmies. First, about a third of these positions have relatively low minor allele fractions (less than 20%) that may be masked during PCR amplification of mtDNA, for example if the major allele is amplified preferentially [179, 180]. Secondly, it is worth noting that the samples for validation came from the same tissue and individual as previously sequenced ChIP samples, but they were from flash-frozen tissue and hence not the exact same sample used for ChIP. Therefore we cannot exclude that some heteroplasmies might exist in localized areas within a tissue [181]. Finally, although the detection algorithm should avoid most sequencing errors, it is possible that a small number of detected heteroplasmies are false positives. Further investigation of these issues would require performing more complicated and sensitive experiments, such as a single base extension assay or ideally PCR-free sequencing. While we deemed this unnecessary in the context of this project, such experiments are likely to improve the validation rate reported here.

2.5.3 Number of different mtDNA genomes

When an individual expresses one heteroplasmic position, it is likely that only two variants of the mtDNA genome exist within its cells. However, when individuals express more than one heteroplasmic position, which is the case for 21 individuals in this study (see table A.1),

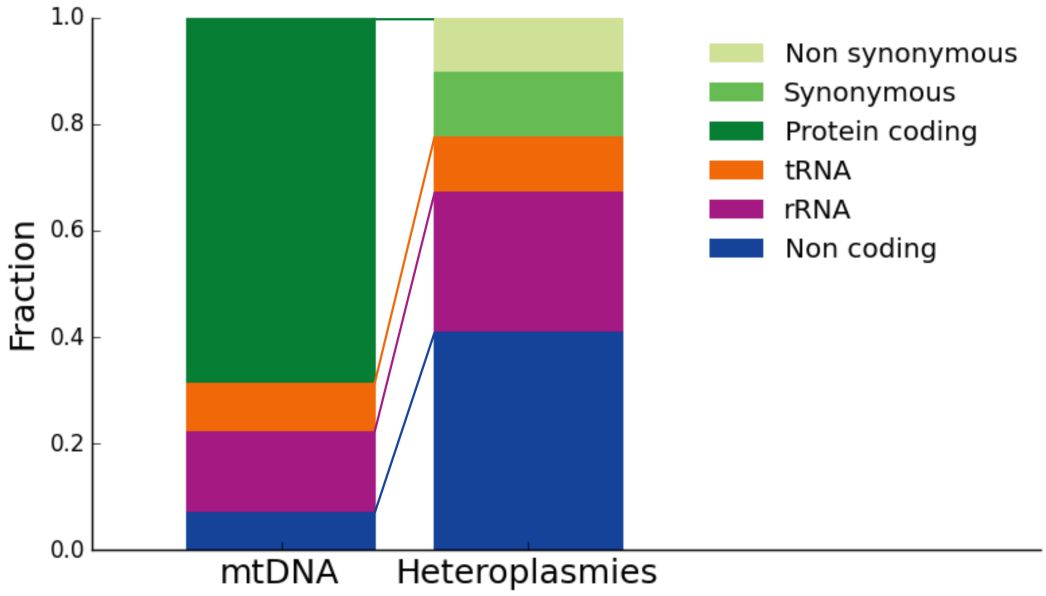


Fig. 2.8 The genomic location of the detected heteroplasmies. We observed a strong bias in the genomic location of the heteroplasmies, which is significantly different to the repartition of mtDNA elements.

we cannot determine the underlying number of mtDNA genome variants. Estimating this number could be possible using the heteroplasmy minor allele fractions (for example if all heteroplasmies have exactly the same level it is likely that there are only two variants and similarly if two different levels are found then there are at least three different types of mtDNA), however this would require a precise evaluation of the heteroplasmy level. This might be feasible with high coverage data but is not realistic with the data presented here. Further, none of the detected heteroplasmic positions have more than 2 alleles. Although our analysis revealed a handful of positions that have one sequencing read containing an additional third base, such potential third alleles never fulfil the criteria to be determined as alternative minor alleles.

2.5.4 MtDNA coverage in ChIP-seq data

It is well known that the mtDNA is sequenced many times due to its high copy number and that significant amounts of mtDNA reads are found in most NGS generated datasets including in ChIP-sequencing data for which mtDNA coverage has even been used as a control for background noise [182]. However, since most ChIP-seq studies focus on binding events in the nuclear DNA and use other control methods, reads mapping to mtDNA have been discarded as mitochondrial contamination. Here we demonstrate that mtDNA coverage in ChIP-seq samples is in fact so high that it may be used similarly to targeted mtDNA shotgun sequencing. We first studied the coverage of each ChIP-seq file independently and observed that it was consistently deeper in the control files compared to the transcription factor binding files, which also had higher mtDNA coverage than the histone ChIP-seq files (see figure 2.10). Low mtDNA coverage in ChIP-seq for histone marks is expected, since histones are not used to pack mtDNA in the mitochondrion [135]. Although the investigated transcription factors do not show signs of binding on the mtDNA (see section 2.3.9), weak binding events could explain the high coverage found in these experiments. Overall, after merging the different files the coverage level and ratio were relatively high and comparable to the data from Li et al. [148], on which we based our detection algorithm. On the other hand, coverage ratio was variable between species, ranging from the entire mtDNA to as little as 10%. This caveat restricts the conclusions we can draw about the genomic location of heteroplasmies within individual species, since some portions of the mtDNA genome are not covered in every individual. However, although we were limited in our ability to conduct detailed analyses, such as cross-species comparisons of heteroplasmic positions, we focused on extracting information from our dataset as a whole to obtain conclusive results.

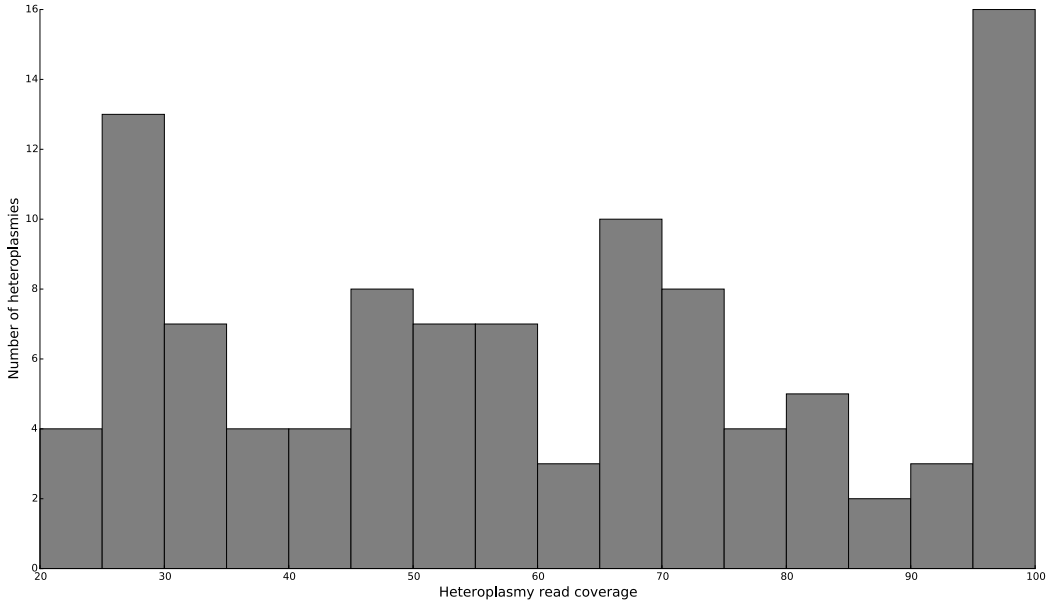


Fig. 2.9 Read coverage of heteroplasmic positions. We observed a significantly higher read coverage of our detected heteroplasmies than the threshold of 20 reads used in our detection method.

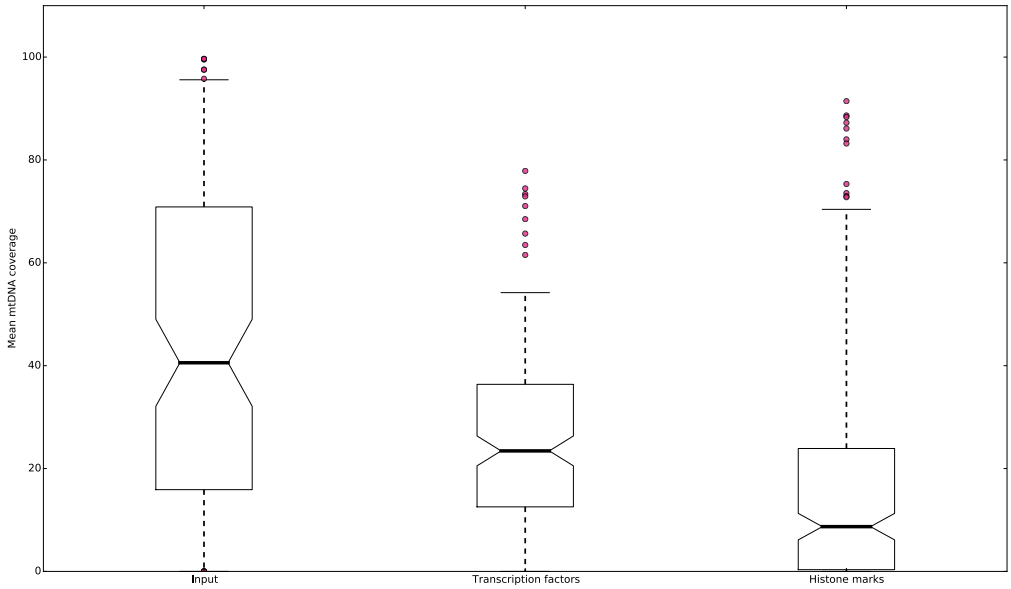


Fig. 2.10 Coverage data of different ChIP-seq data files. This figure shows the differences in coverage level between particular ChIP-seq files. We observed significantly higher coverage in control (input) files, while there was also a difference between histone and transcription factor ChIP-seq data.

2.5.5 ChIP-seq datasets are useful data sources for the study of heteroplasmy

We have shown that it is possible to make use of ChIP-sequencing data that was originally collected for other purposes to explore mitochondrial heteroplasmies across a wide variety of mammalian and vertebrate species by taking advantage of the high number of sequenced mtDNA reads in these experiments. We hope our general approach will encourage the further study of mtDNA and other biological questions using valuable existing datasets. For instance, making use of data arising from other profiling-sequencing methods such as RNA-seq or ATAC-seq might be very useful, as demonstrated by a recent study of the mitochondrial transcriptome [183]. Finally, although we adapted a detection method appropriate for our datasets of relatively low-coverage and variable homogeneity, higher coverage and/or more advanced methods for heteroplasmy detection recently developed [150, 152] could potentially be valuable for a wider variety of existing data types. For example, higher coverage datasets will likely enable new methods to explore lower levels of heteroplasmy in ChIP-seq data. Furthermore, having demonstrated that ChIP-seq coverage of the mtDNA can be substantial and allows for the study of heteroplasmy, it is likely that ChIP-seq data can be used to perform other mtDNA studies in fields such as population genetics, forensics and in some areas of medical research. There is a current trend of creating large genotype-phenotype datasets for ChIP-seq and RNA-seq data, and using these datasets to explore phenotype associations to heteroplasmies could prove extremely valuable to further our knowledge of the phenomenon.

2.6 Conclusions

Our study shows that mitochondrial heteroplasmy displays similar characteristics across vertebrate species; including genomic location and mutation spectrum across the vertebrate species tree. As might be expected, our results strongly suggest that previous heteroplasmy findings established in human are valid for all mammals and possibly all vertebrates. In addition to this, our results also support recent findings that heteroplasmy is more prevalent in liver compared to other tissues. Our results also suggest that any new understanding about heteroplasmies will likely apply across the mammalian clade. It is clear both in our findings and in previous work that mutation rates vary significantly between positions, meaning some positions are more likely to exist as heteroplasmies than others. Although there is limited information, on a molecular level, describing the functional impacts these heteroplasmies may have, our results suggest that those functional impacts are likely to be similar in many

different mammalian species and that many species may be effective model organisms for understanding the biology of heteroplasmy.

Heteroplasmy not detected
Evidence of heteroplasmy
Heteroplasmy detected

Human validation rate	0.61
Non-human validation rate	0.56
Total validation rate	0.59

Individual	Position	Ratio	Alleles	Annotation	Validation
hsa23	72	0.35	TC	NC	6 / 6
hsa23	310	0.28	TC	NC	4 / 6
hsa23	309	0.17	CT	NC	0 / 2
hsaC3	310	0.32	TC	NC	7 / 8
hsaC3	72	0.2	TC	NC	7 / 8
hsaC6	310	0.44	CT	NC	3 / 6
hsaC6	309	0.15	CT	NC	4 / 4
hsaC7	310	0.29	TC	NC	2 / 4
hsaC7	4315	0.23	AT	tRNA	2 / 6
hsaC7	11944	0.26	CT	SYN	4 / 8
hsaC7	16192	0.15	TC	NC	1 / 2
hsaC8	12236	0.32	AG	tRNA	0 / 6
hsaC8	195	0.34	CT	NC	1 / 5
hsaC8	310	0.26	TC	NC	2 / 5
hsaC8	182	0.33	TC	NC	0 / 5
hsaC8	198	0.36	TC	NC	0 / 5
hsaC8	204	0.35	TC	NC	5 / 6
hsaC8	8387	0.16	AG	nSYN	0 / 5
Btau5	1594	0.29	AG	rRNA	2 / 4
Btau6	2755	0.26	CT	rRNA	3 / 4
Cfam4	16148	0.2	GA	NC	3 / 3
Cfam6	16168	0.42	AG	NC	1 / 2
Cjac5	6191	0.39	TC	SYN	0 / 4
Ggal2	3832	0.33	CA (ref: G)	rRNA	1 / 5
Mfur2	15525	0.36	CT	NC	4 / 4
MmulBO	15510	0.32	TC	nSYN	4 / 4
MmulJO	16182	0.4	GA	NC	0 / 4
Ocun4	8499	0.22	AG	SYN	5 / 5
Ocun5	1957	0.25	TA	rRNA	0 / 5
Ocun5	3866	0.18	AT	tRNA	0 / 4
Rnor5	2448	0.18	CT	rRNA	0 / 5
Shar1	15635	0.36	TA	NC	4 / 7
Shar2	15635	0.32	TA	NC	3 / 5
Sscr6	15773	0.22	AG	SYN	0 / 6

Fig. 2.11 Heteroplasmic positions validation. We successfully validated a majority of the detected heteroplasmies using Sanger sequencing. This figure shows the details of the randomly selected 34 positions and the number of experiments in which the two alleles were detected (validation). We defined two categories of validated positions (see section 2.4.9 and section 2.3.4).

Chapter 3

ChIP-seq tissue deconvolution

3.1 Summary

This project investigates the deconvolution potential of complex tissue ChIP-seq data. ChIP-seq studies have mostly been performed on homogenous samples such as cell lines because a considerable amount of homogenous cells is required to obtain reliable, biologically interpretable results. However, most biological tissues are complex and composed of many cell types. I explored a computational approach to model the behaviour of complex tissue ChIP-seq to enable an estimation of cell-type specific DNA-protein binding profiles directly from complex tissue ChIP-seq experiments. I performed in-silico experiments with publicly available datasets in order to simulate complex tissue data and applied several computational methods attempting to deconvolve the simulated data. Although the results were inconclusive, they provide insights into potential deconvolution approaches that could be applicable in the future.

3.2 Background

Almost every cell of complex organisms contains the organism's entire genome (identical DNA). However, their shapes and function vary significantly depending on their context and purpose. For example, neurones have extremely elongated shapes, while muscle cells have higher mitochondrion counts. These differences are mainly driven by which genes are expressed as well as at what level they are expressed (highly or lowly). As previously described in the introduction (see section 1.4.2), proteins play a key role in gene regulation. Transcription factors bind the DNA prior to the polymerases to recruit them and cooperate to initiate the process [184]. Transcription regulators bind to affect the efficiency of the transcription/replication processes (promoting or repressing) [185]. Even histones and how they are biochemically modified [186] have been shown to influence gene regulation [187]. We know that gene expression is the core mechanism behind cellular function (see section 1.4.2), and understanding how this process is regulated is key. Several approaches can be taken to exploreingle protein regulated gene expression. First, from a genomic perspective, one can study the genes that are present in the DNA and how many copies of each exist. Second, analysing the RNA content of the cells (see section 1.9.1) can reveal which genes are expressed and even quantify the expression of those. Finally, studying the binding of regulatory elements in-vivo and analysing their genomic context is a technique, which has provided broad insights into gene expression and regulation (see section 1.8.3). Importantly, in a similar way to the gene expression of different cells, the protein-binding profiles also vary significantly between cell types. In fact, it has been found that the binding profiles of some regulatory elements are invariable across different tissue contexts, while others have tissue specific binding patterns [188]. Proteins, which bind in a similar manner in most situations are said to be ubiquitous. Many studies have investigated the binding profiles of transcription factors, most by analysing ChIP-seq data for the transcription factors directly [40, 160], but others have looked at other proteins which do not bind DNA directly but interact with transcription factors [189]. The recently published (2012) ENCODE project [190], investigated many transcription factors and also demonstrated that ChIP-seq profiles vary significantly depending on context (see section 3.2.1).

3.2.1 The ENCODE project

The ENCyclopedia of DNA Elements (ENCODE) is a very large research project that was initiated in 2003. As the next step to the previously accomplished Human Genome Project, its main aim was to identify all functional elements present in the human genome [191]. To

achieve this, a consortium of scientific institutions and laboratories around the world was established and together they produced more than 1,500 datasets comprising experiments performed in 147 different cell types.

The first phase of the project, consisted in a pilot project, which aimed at testing and comparing analysis methods applied to a defined subset (30 million base pairs or ~1%) of the human genome. As this phase of the project took place before modern NGS based transcription and gene expression analysis methods were developed (see section 1.8.3 and section 1.9.1), older techniques such as ChIP-chip (see section 1.8.3) were used for this initial project phase. Nonetheless, the results, published in 2007 [192], provided novel insights into the function of the human genome. The core findings of the pilot project are:

1. Strong evidence was found that the human genome is almost entirely transcribed. Most of the genomic bases are present in primary transcripts (transcripts before post-transcriptional modifications, see section 1.3.2).
2. Novel insights into transcriptional start sites were discovered. For example, chromatin structure and histone modifications can be used to predict with high confidence the presence and activity of start sites.
3. Significant progress in the understanding of the role of chromatin structure was made, in particular in the context of DNA replication and transcription. For example, it was found that the timing of DNA replication was correlated to chromatin structure.

Following the pilot phase, the production or main project phase was performed, with the target of investigating the function of the human genome in a genome-wide manner. The main experimental methods used were ChIP-seq, DNase Hypersensitivity site mapping, RNA-seq, and several DNA methylation assays. All of the generated data (for the main and the pilot projects) was organised and stored in databases which are publicly available [193]. The research results provided broad insights into how the human genome functions. Briefly, the main results of the project are [190]:

- Most of the human genome is involved in at least one biochemical event (RNA- and/or chromatin-associated) in at least one cell type. Moreover, 99% of the human genome lies within a 1.7kb distance of such an event.
- As a whole, the human genome elements that are primate-specific and the elements for which no mammalian constraint is detectable, show evidence of negative selection.
- The analysis of a classification of chromatin states reveals thousands of states with clear functional properties.

- A correlation was found between the quantity of RNA produced and processed and both chromatin marks and promoter transcription factor binding. This suggests that the majority of RNA expression variation can be explained by promoter functionality.
- There are at least as many non-coding variants located in ENCODE annotated functional elements as in protein-coding regions.
- The majority of disease-associated Single Nucleotide Polymorphisms (SNPs) are located in or close to the ENCODE annotated functional elements (but outside of protein-coding regions). Furthermore, they were able to associate many disease phenotypes with specific cell types or transcription factors.

In the context of this project, in which I investigate the concept of complex tissue ChIP-seq deconvolution, the two most important aspects of the ENCODE project are:

1. The amount of ChIP-seq data produced for the study. ChIP-seq experiments were performed for 119 different DNA-binding proteins in a total of 72 different cell types (although not all combinations of experiments were performed). A specific online platform was created to host and display this data in a matrix [194].
2. A further demonstration of the variation of protein-binding ChIP-seq profiles. An example of this variability can be observed in figure 3.1, which depicts the association between binding profiles for many different proteins in a single cell type.

3.2.2 ChIP-seq to study protein-binding profiles

With the arrival of next generation sequencing (NGS) technology, the ChIP on chip method was improved with the development, in 2007, of the modern chromatin immunoprecipitation followed by sequencing (ChIP-seq) assay (see section 1.8.3). This enabled high definition genome wide studies to be performed, and since then, ChIP-seq has become the standard method to study genome-wide protein-DNA interactions in-vivo [195].

One of the major constraints in performing a ChIP-seq experiment, is the quantity of cells required to obtain reasonably noise-free protein binding profiles. In the early stages, 50 million cells was deemed ideal to produce consistent and reproducible protein binding profiles [174]. However, modern technology is progressively reducing this quantity and a recent study has demonstrated good results can be obtained with as few as 1,000 cells [111]. Furthermore, the first single cell ChIP-seq assay has just been published (September 2015) [89] (see section 1.8.3). Although this study demonstrates the direction in which the technology is

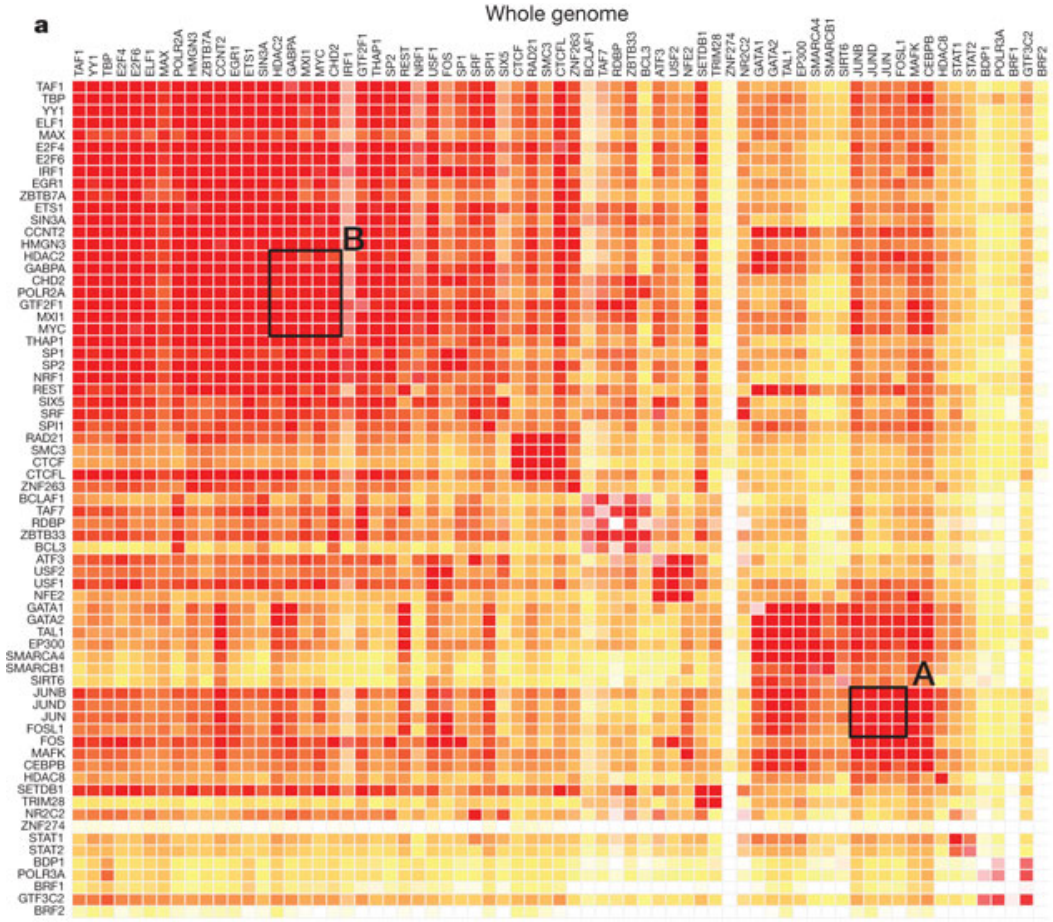


Fig. 3.1 ChIP-seq profile variation of many different proteins in a single cell type. This heat map represents the co-association between many transcription factors in a particular human blood cell line (K562). Each pair is coloured from red (strongest) to yellow (weakest) depending on the strength of the co-association. This illustrates the variation in binding profiles between different transcription factors in the same cellular context. Highlighted section (A) shows pairs for which the association is not dependant on the genomic context, while section (B) encompasses pairs for which the associations vary depending on the genomic context, even though, overall the association is strong. Modified from [190].

evolving, a tiny amount of DNA is precipitated and the sequencing output is extremely sparse (only about a 1,000 reads per cell). This makes the biological interpretation of the results very difficult. Another constraint of this protocol is the usage of advanced microfluidics systems which can be costly and challenging to build. Furthermore, the processing of the cells through the microfluidic chip may trigger cell responses that could alter the transcriptional profiles [196]. Potential future improvements of the technologies used by the ChIP-seq method are further discussed in Chapter 4. At the present time (2016), millions of cells are still necessary to produce reasonably noise-free protein binding profiles with the "standard" ChIP-seq method [195].

As described above (see section 3.2), since protein binding varies between cell types and even across cellular contexts, ChIP-seq samples also need to be highly homogenous in order to obtain reliable and biologically interpretable results (this is also discussed in section 1.6). For these reasons, ChIP-seq studies have mostly been performed on homogenous samples, such as cultured cells from well-characterised cell lines. Some work has also used highly homogenous tissues such as liver with the assumption that the resulting signal corresponds to the majority cell type. However, most biological tissues are composed of many different cell types (see section 1.6) and the biological complexity of whole tissues normally cannot be reproduced in homogenous cell lines. Methods for the isolation of specific cell types from complex tissues such as cell sorting techniques or laser micro-dissection make such experiments possible in some contexts, but these techniques are expensive, time-consuming, and often require cells to be in suspension. For example, cell sorting by FACS (Fluorescence-activated cell sorting) is an experimental procedure that is often performed on blood samples [197]. However, such constraints limit the reproducibility of tissue contexts as well as potentially triggering particular cellular responses. To avoid the tissue disruption and yet to enable complex tissue protein-DNA interactions research, we propose a computational approach to investigate potential deconvolution of complex tissue ChIP-seq signal (see below section 3.2.3).

3.2.3 ChIP-seq deconvolution concept

In a complex tissue ChIP-seq experiment, proteins from each cell type are precipitated. Hence the ChIP-seq peaks in the resulting data could be generated from proteins binding in any or many of the different cells (of potentially different cell types) present in the sample. We have thus formulated the assumption that the individual binding patterns of each cell type act in an additive manner to form the resulting ChIP-seq signal. Under this

hypothesis, deconvolving the separate contributing signals should be possible, at least in some contexts. Evidence supporting the concept of ChIP-seq deconvolution has been found in a previous study that compared binding profiles for several proteins in multiple different cell types (see section 3.2.4). Furthermore, a number of different techniques have been applied to deconvolve gene expression data in similar contexts. For the purpose of this study, we considered the concept of complex tissue ChIP-seq deconvolution to be the process of assigning the ChIP peaks to the contributing cell types. In other words, assessing which binding events occurred in which cell types. We make the strong assumption that we known the amount of cell types present in the heterogenous sample as well as their mixture ratio (for example that the complex tissue is composed of two cell types at ratio 2:1, this information could be estimated by histology). We do not precisely define the input as being one or multiple experiments (for example with different proteins or cell type ratios) and we explore different options. The term "deconvolution" has previously been used in the ChIP-seq context (for example in [198–200]), however these studies all focus on deconvolving the binding signals (peaks) in the sequencing reads from noise. These studies do not attempt to uncover cell type specific binding profiles from heterogeneous sample ChIP-seq data.

3.2.4 Suggestive evidence for ChIP-seq deconvolution

A recent study investigating transcription factor co-binding [201], analysed a large number of ChIP-seq profiles for different proteins in multiple cell types. The scientists integrate the results of 53 different ChIP-seq experiments focusing on transcriptional control in mouse hematopoietic cells. In addition to providing insights into the transcriptional regulation of mouse hematopoiesis, they show that genome-wide binding patterns for transcription factors mostly depend on the cellular environment. After a stringent pre-processing of the peak tracks (calling peaks with multiple algorithms, removing problematic regions, and merging results for analogous experiments), they obtained a total list of binding sites (over all cell types). Following this, they created binary binding profiles for each factor and cell type; marking each potential binding site with a "1" if the protein is bound and a "0" otherwise. They then grouped all these tracks into a matrix and used a hierarchical clustering algorithm to analyse and compare the binding profiles. A heat map (see figure 3.2) was produced highlighting the similarities between binding profiles (using Pearson's correlation, see section 1.10.1). This analysis highlights the preferential clustering of binding profiles by cell type rather than by transcription factor. Only two transcription factors show clusters of profiles across cell types. The first is CTCF and this could be expected as it is well known that CTCF binds in a ubiquitous manner in many different tissues [62]. The second (PU1), however, was

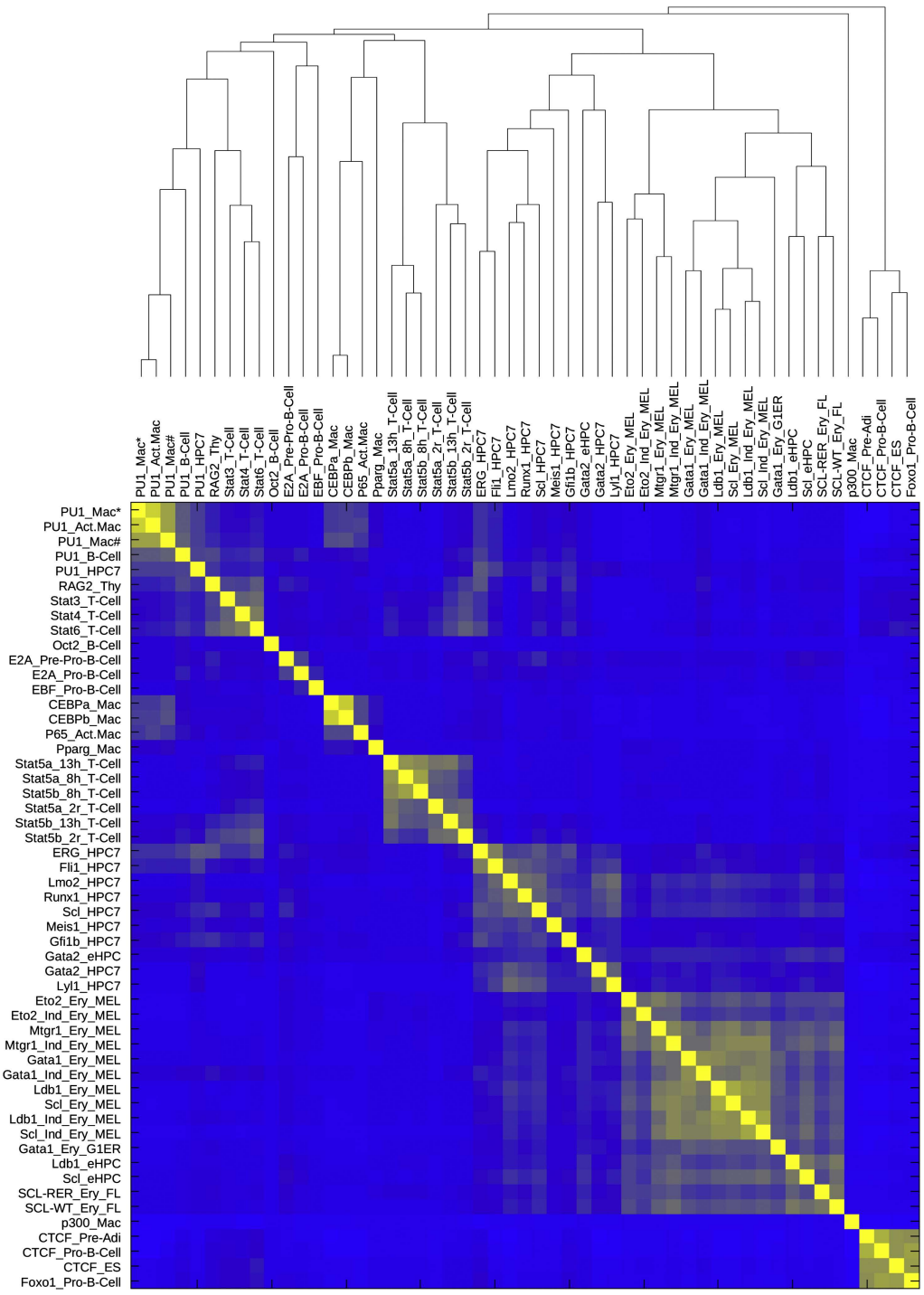


Fig. 3.2 Clustering analysis of 53 different ChIP-seq experiments. This figure depicts a hierarchical clustering performed on binary binding profiles obtained from 53 experiments. This analysis highlights the preferential clustering of binding profiles by cell type rather than by transcription factor, except in two particular cases (there are clear PU1 and CTCF clusters at the top and bottom of the figure respectively). Obtained from [201].

unexpected and they hypothesise that this suggests that there exists variability in binding affinity for some transcription factors. This particular analysis, demonstrating that protein binding profiles cluster by cell type, supports the hypothesis that deconvolving ChIP-seq data from complex tissues could prove successful (and this fact is highlighted in their discussion).

3.2.5 Gene expression deconvolution

Deconvolving data produced from complex tissues is also a topic of interest in the fields of gene expression and cancer genomics (see section 3.6). For gene expression, except for the recent single-cell RNA-seq research (see section 1.9.1), gene expression studies are generally also performed on large amounts of cells (tissues or cell lines). As explained previously (see section 1.9.1), the RNA from all the cells is extracted and sequenced before being aligned to a reference genome. Similarly to ChIP-seq, this ensures that enough RNA is sequenced in order to obtain an interpretable dataset, especially when studies attempt to quantify gene expression for comparative purposes [202]. Another advantage is that by analysing the RNA from a group of cells, the average expression pattern is obtained and the signal from abnormal expression patterns potentially produced by particular individual cells is smoothed out. When studying cultured cell lines, using large amounts of cells is not a problem since they are all identical (or very similar to the least). However, as described above for ChIP-seq studies, when the material of interest is a non-homogenous (complex) tissue, the resulting experimental products are formed by a combination of the individual contributions from each cell. Thus the expression patterns for the various cell types, which are likely different (see section 3.2) are convoluted in the experimental output.

Multiple studies have investigated complex tissue deconvolution in the context of gene expression. Some are based on microarray data (similar method to ChIP chip, see section 1.8.3) [203–205], while more recent work has been done with RNA-seq data [206, 207]. Most methods are based on the same underlying hypothesis that the contribution of the different cell types follows a linear model [208]. As described by Venet et al. [203], the linear model can be defined as following (M_{ij} is the expression level of gene i in sample j, G_{ik} is the "true expression" or signature of gene i in cell type k, C_{jk} is the concentration of cell type k in sample j, N_{ct} is the total number of cell types, and $\mathbf{M}, \mathbf{G}, \mathbf{C}$ are the respective matrices, see table 3.1):

$$M_{ij} = \sum_{k}^{Nct} G_{ik} C_{kj} \iff \mathbf{M} = \mathbf{GC}$$

M	Sample 1	Sample 2	Sample 3
Gene 1	55	70	65
Gene 2	30	18	22
Gene 3	15	12	13

G	Cell type 1	Cell type 2
Gene 1	30	80
Gene 2	50	10
Gene 3	20	10

C	Sample 1	Sample 2	Sample 3
Cell type 1	50%	20%	30%
Cell type 2	50%	80%	70%

Table 3.1 Gene expression deconvolution linear model. This table shows examples of the matrices that constitute the linear model of gene expression deconvolution. **M** is the collection of gene expression experimental measures (normalised). **G** consists of the "true expression" values or expression signatures for each gene and cell type. Finally, **C** contains the cell type mixture fractions of the experimental samples. The example values follow the equation of the model: $\mathbf{M} = \mathbf{GC}$. Modified from [203].

The problem of deconvolving the matrix **M** into two separate matrices **G** and **C** has been tackled with linear regression, factor analysis (a mathematical method used to reduce the number of dimensions in correlated data by exploring underlying variation) and more direct methods such as minimising the error with a least squares approach [203]. Recently, more advanced machine learning methods have also been used [209]. Reasonably good results have been obtained with most of these approaches, suggesting that the initial linearity hypothesis is likely to be correct [208]. Expanding on this a number of different approaches have been developed to deconvolve a restrained version of the problem described above. In some cases, the cell gene expression signatures can be determined experimentally, if the cells are available in pure samples for example. Mathematically, this means the matrix **G** in the model equation is known. In this case, models use deconvolution methods to estimate the cell fractions of complex tissues [204, 207, 210]. Alternatively, since experimental techniques exist to determine (or estimate) the cell fractions within a sample (for example using fluorescence cell sorting), other methods consider matrix **C** to be known, and attempt to recover the cell type gene expression signatures [205]. Logically, since the problem is simplified, the partial deconvolution approaches have provided results of much higher quality [211]. The results shown in figure 3.3 show the partial deconvolution performance from a recent study [206]. They also highlight the difference in deconvolution quality between 50% mixture

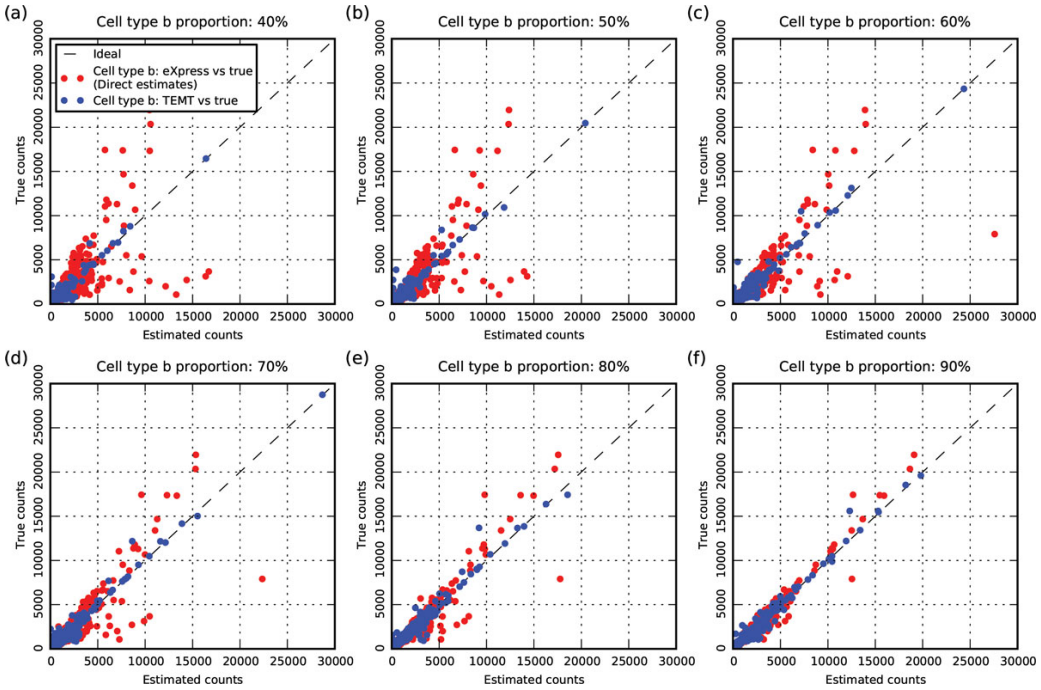


Fig. 3.3 Gene expression deconvolution example results. These figures show the relation between the estimated gene expression (via deconvolution with two different methods) and the true levels of expression. A clear difference in deconvolution quality between 50% mixture ratios and the more biased ratios is observable. Obtained from [206].

ratios compared to more biased ratios, a phenomenon we also observe in this project (see section 3.4.5).

3.2.6 Deconvolution of complex signals with Machine Learning

Machine learning methods (see section 1.10.2) have been applied in many different scientific contexts including genetics and genomics [126]. They are highly efficient in solving classification problems and can thus be used as prediction models. For example, machine learning methods can and have been used to perform gene predictions and annotate genomes with different types of genomic information [212]. As described in section 1.10.2, machine learning algorithms are generally good at finding patterns in complex data, which means they are ideal tools for complex signal deconvolution problems. One of the previously mentioned gene expression deconvolution studies [209] (see section 3.2.5) makes use of an unsupervised learning technique to find hidden patterns to enhance the deconvolution process.

A good example of a deconvolution challenge that has been efficiently solved using modern machine learning models is the "Cocktail party problem". This scenario was first described in a psychological study from 1953 [213]. This attention experiment consisted in a person hearing two messages played through the same loudspeaker (at the time). The participants were then asked to describe both messages separately, essentially performing signal separation. Although this human led separation of signals is still being studied today [214], this problem has now also been investigated from a computational perspective, using algorithms to perform the separation [215]. The "cocktail party problem" has been generalised to a random number of loudspeakers (audio outputs) being listened to by a random number of recording microphones (audio inputs). It has been shown that this problem becomes significantly simpler to solve when more than one input is used. In fact, the problem is solvable by independent component analysis (a computational method used to separate a complex signal into additive sub-signals) as long as there is at least the same amount of microphones (inputs) as there are sources (outputs) emitting at the same time [216]. Machine learning models have also been shown to produce extremely good results when tackling this challenge [217, 218].

The underlying reason why these methods are able to separate the signals, especially in the situation with more inputs than outputs, is that there is a variation between the different signal mixtures recorded. In the very simple case of two outputs and two inputs, output A might be closer to input A, while output B is closer to input B. In such a case, input A would record output A with a higher intensity (higher volume) than output B, and the opposite would happen for input B. This variation between the two recorded signals is what enables

algorithms to deconvolve the complex signal into its respective components (output A & B). In fact, most methods used in gene expression deconvolution also exploit variation between samples to recover the underlying additive components (see section 3.2.5). An approach based on this concept is also investigated in the context of complex ChIP-seq deconvolution and is described below (see section 3.4.5).

3.2.7 ChIP-seq deconvolution approach

To investigate complex tissue ChIP-seq deconvolution, we also hypothesised a linear contribution from each cell type to the resulting signal of a heterogeneous sample (similarly to what has been done for gene expression deconvolution, see section 3.2.5). This means that each binding event (peak) discovered in the mixture signal should be the sum of contributions of each cell types, weighed by their mixture proportion, formally: (P_i is the peak intensity at locus i in the mixed sample, G_{ik} is the theoretical peak intensity at locus i in a pure sample of cell type k, C_k is the concentration of cell type k in the mixed sample, N_{ct} is the total number of cell types present in the mixed sample):

$$P_i = \sum_{k}^{Nct} G_{ik} C_k$$

Based on this hypothesis, we performed a simulation based approach using publicly available ChIP-seq data form the ENCODE project [190], generating mixed samples from different cell-types assayed for the same protein. This enabled us to create an environment in which all variables are known (the mixture fractions, the individual binding profiles as well as the mixture binding profile). In our first approach, we analysed peak parameters under different mixture contexts and with different proteins (see section 3.4.1). Then, we used machine learning classification techniques to attempt deconvolution of the peak signals (see section 3.4.2). The low quality of the deconvolution obtained by this initial method lead us to rethink and simplify our approach. Having observed inconsistencies in our data (for example peaks present in both subsamples but not in the final mixture) and knowing that peak calling algorithms are complex (see section 1.8.3), we decided to use an alternative method to identify protein-bound regions. Instead we used a simpler binning approach, in which 1kb bins were marked as bound or unbound (depending on a read threshold), thus identifying the protein-binding profiles (see section 3.3). After performing different analyses on the generated bins, we applied machine learning methods again to try to deconvolve this newly structured dataset (see section 3.4.4). Following further inconclusive results, we designed another

deconvolution approach based on the variation between binding profiles from different samples, similarly to gene expression deconvolution techniques (see section 3.2.5). The results obtained with this method were promising, in particular in the presence of large variation (see section 3.4.5). Overall, we have investigated several potential approaches to complex tissue ChIP-seq deconvolution and our results demonstrate the challenging nature of this problem. We believe the results obtained with our final approach suggest deconvolution is possible in particular contexts and further investigation in exploiting variation between ChIP-seq samples to deconvolve protein-binding profiles could lead to deconvolution in more generalised contexts (see section 3.5.2).

3.3 Materials and methods

3.3.1 ENCODE Data

In this study we used ChIP-seq data generated for three different proteins (EP300, CEBPB, and CTCF) in the GM12878 lymphoblastoid cell line and the HepG2 hepatocyte cell line. This data was obtained from the ENCODE project [190] and was generated by five different institutions and collaborations: the HudsonAlpha Institute for Biotechnology (wgEncodeHaibTfbs), Stanford/Yale/Davis/Harvard (wgEncodeSydhTfbs), the Broad Institute (wgEncodeBroadHistone), the University of Texas at Austin (wgEncodeOpenChromChip), and the University of Washington (wgEncodeUwTfbs). Additional details regarding the data obtained, including the GEO accession codes are in table 3.2. For each experiment, only the raw unprocessed (FASTQ formatted) sequencing files of the replicates and corresponding control files were downloaded.

3.3.2 Pre-processing and read alignment

The raw sequencing data files (FASTQ) were mixed in different proportions using an in-house developed script. Thresholds were used to ensure that each mixture file contained the same total number of reads (both the control and signal files). Each of these generated mixture FASTQ files were then aligned to the reference human genome obtained from Ensembl [9] (GRch37) using BWA (Burrow-Wheeler Aligner) [103] with the default parameters.

Institution	Protein	Cell Type	GEO
wgEncodeHaibTfbs	P300	GM12878	GSM803387
wgEncodeHaibTfbs	P300	HepG2	GSM803499
wgEncodeHaibTfbs	CEBPB	GM12878	GSM1010850
wgEncodeHaibTfbs	CEBPB	HepG2	GSM1010778
wgEncodeSydhTfbs	P300	GM12878	GSM935559
wgEncodeSydhTfbs	P300	GM12878	GSM935294
wgEncodeSydhTfbs	P300	HepG2	GSM935545
wgEncodeSydhTfbs	CEBPB	HepG2	GSM935493
wgEncodeSydhTfbs	CTCF	GM12878	GSM935611
wgEncodeBroadHistone	CTCF	GM12878	GSM733752
wgEncodeBroadHistone	CTCF	HepG2	GSM733645
wgEncodeOpenChromChip	CTCF	HepG2	GSM822287
wgEncodeUwTfbs	CTCF	GM12878	GSM749706
wgEncodeUwTfbs	CTCF	HepG2	GSM749683

Table 3.2 ENCODE ChIP-seq data used in this project.

3.3.3 ChIP-seq peak calling

The protein binding profiles for each of the mixture and component files were obtained with the Model-based Analysis for ChIP-seq (MACS) program [108]. We provided the software with the appropriate control files (mixed control files were also generated) and used the default parameters.

3.3.4 Peak attribution

To enable the analysis of the complex tissue binding profiles, the peaks from the mixture files were categorised using the peaks of the component files. Using the Bedtools [219] software package, we analysed the overlapping peaks between the component and mixture files. This enabled us to assign peaks from the mixture file to the following categories: “The peak is present in none of the component files”, “The peak is present in the GM12878 component file”, “The peak is present in the HepG2 component file”, and “The peak is present in both component files”.

3.3.5 Bin approach

The bin approach was implemented using the Python programming language [165]. For each 1kb bin, the reads overlapping the corresponding genomic coordinates were retrieved using a Samtools [164] wrapper (Pysam). A threshold was then used to define bins as "bound" or "unbound" in the component files. Similarly to the peak categorisation, we attributed bins depending on the "bound" or "unbound" status of the bins in the component files. Using a threshold of 30 reads resulted in a similar number of bins than previously called peaks, for this reason we used this value for our analysis.

3.3.6 Analysis and plotting

All of the analysis was performed in R [220] and Python [165]. Ggplot2 [221] was used to create plots in R, while Matplotlib [168] was our Python plotting library. The machine learning models (the logistic regression model and the support vector machine) were implemented with the Scikit-learn Python library [222]. Additionally, to train, run and test our models in Python, SciPy [166], a scientific package, and Pandas [167], a data analysis package, were also used. We compared raw read data, ChIP-seq peak annotations and our bin predictions with the Integrative Genomics Viewer (IGV) [223]. To test the hidden Markov model approach, we implemented a HMM in R using the HMM library [224].

3.4 Results

To model complex tissue ChIP-seq, we set up an in-silico experiment. We selected two different, well-studied cell lines for which ChIP-seq data for several proteins was available: the GM12878 lymphoblastoid cell line and the HepG2 human liver hepatocellular carcinoma cell line. These cell types are also distant on the differentiation path which should increase the likelihood of them having different gene expression patterns and thus different protein binding patterns (see section 1.4.2). We also chose three transcription regulators, including one that binds in a cell-type invariant manner: CEBPB, P300 and CTCF (ubiquitous). For each factor, we artificially and randomly mixed the data in given proportions (0%-100%), repeating the process to create replicates. For each replicate, we generated three data files, the previously mentioned mixed file and the two component files that were mixed together. The component files were subsampled to ensure that all of the mixture files (for every mixture ratio) were the same size.

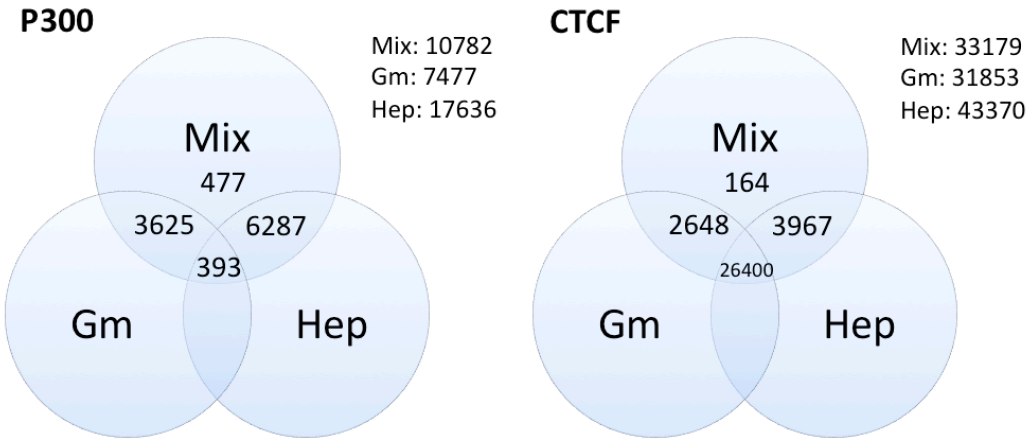


Fig. 3.4 Categorised peaks from CTCF and P300. The above Venn diagrams depict the number of peaks in two representative mixture files for the P300 and CTCF proteins with a 70% GM12878 content. The legends above the diagrams show the total number of peaks in all of the files (mixture and component files). In addition to showing a small number of peaks that appear only in the mixture files, this figure demonstrates the ubiquitous aspect of CTCF with a large number of peaks that are present in both component files (and the mixture file).

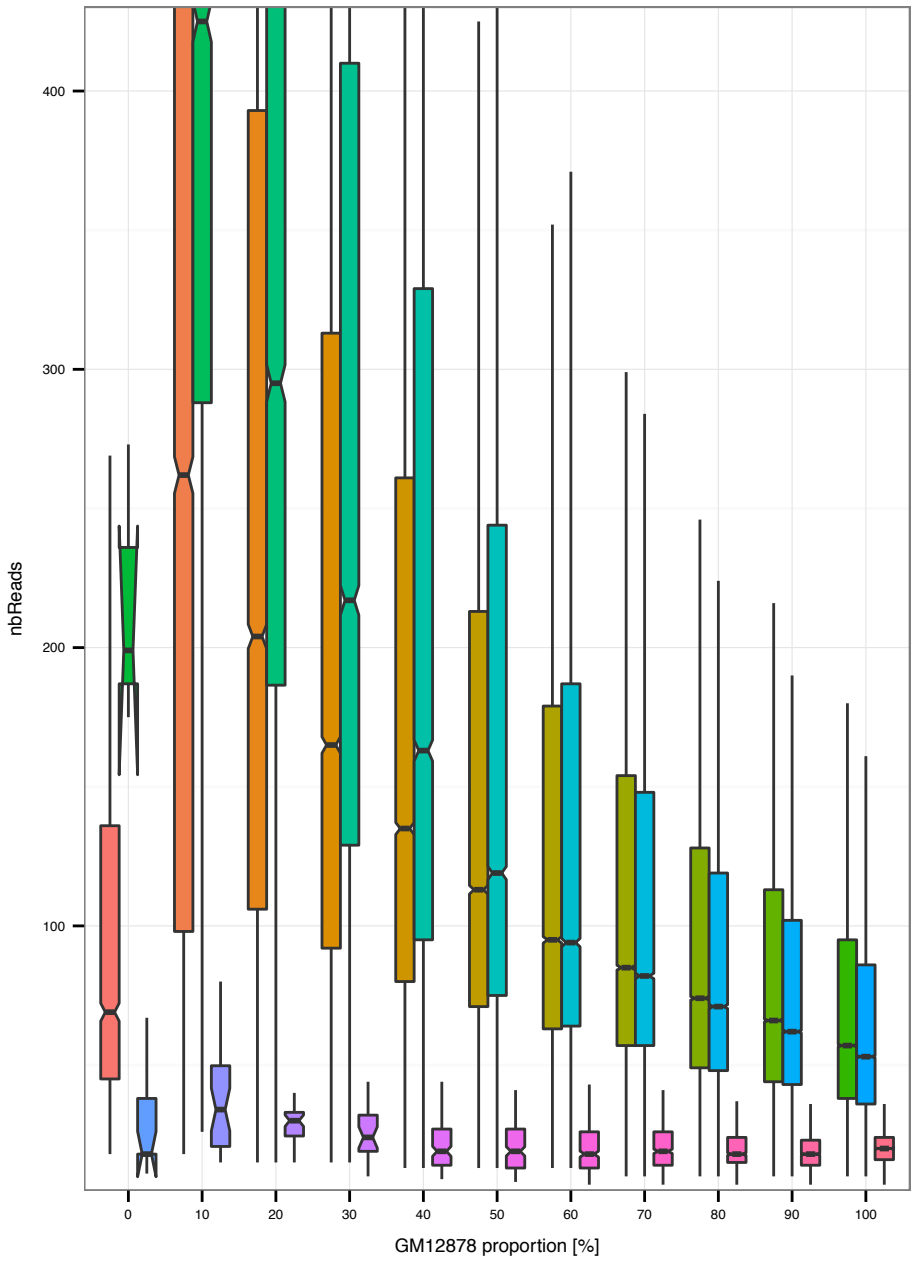


Fig. 3.5 Peak density for different peak categories. This figure depicts the average number of reads per peak in different mixture conditions (from 0% to 100% GM12878 content). The red to green gradient (left-most boxplots) represents all of the mixture peaks that are also present in the GM12878 component file. The second distribution (green to blue), shows the values for peaks that are present in the GM12878 component file but are not detected in the HepG2 component file. Finally, the blue to pink distribution (right-most box plots) depict the average reads per mixture peaks that are not detected in either the GM12878 or the HepG2 subcomponent files. These "new" peaks are very small and likely to be artefacts, so they were ignored in this project.

3.4.1 Analysis of ChIP-seq peak data

Before tackling the complex deconvolution problem, our first approach was to explore our dataset of in silico generated mixture files and assess the potential for deconvolution. We analysed multiple parameters of the protein bound regions and compared data for different proteins. Using a standard peak-calling software (see section 1.8.3 and section 3.3), we determined the protein-binding profiles for the different file types previously generated (mixture and subcomponent files). We then analysed the overlap between component and mixture files (see section 3.3), which enabled us to assign peaks from the mixture files to different categories depending on the presence or absence of these peaks in the subcomponent files. In fact, peaks present in the mixture files can be assigned to four different categories: 1) the peak exists in both subcomponent files, 2) the peak exists in subcomponent file A, 3) the peak exists in subcomponent file B, and 4) the peak does not exist in the subcomponent files.

Analysing the repartition of the mixture files peaks revealed a number of interesting facts. The venn diagram depicted in figure 3.4 shows the comparison of the number of different types of peaks between two mixtures for the P300 and CTCF proteins. The first striking observation was the presence of peaks that did not exist in the subcomponent files but were nonetheless detected in the mixed samples. Further analysis of this peak category suggested they were likely to be artefacts generated by the additive noise from the subcomponent files. In fact, these peaks, when compared to the other categories of peaks, were significantly smaller, both in terms of peak-height and density (see figure 3.5). Since there are also significantly less of these peaks, we hypothesised that they should not have a strong impact on the deconvolution potential of the dataset. Another key observation was that the number of "lost peaks" was important. This was expected and logical since we are subsampling from each of the subcomponent files, thus retrieving only a portion of the peaks from each. However, in particular cases the additive noise of the control files can create an enrichment that the peak-caller will use to ignore a peak in the signal file. This phenomenon can be observed in figure 3.6. The fact that there are many undetected peaks in the mixture file could prove important in the analysis of complex tissue ChIP-seq files, however they should not have any impact on the actual deconvolution challenge in itself. Finally, the variation between the repartition of the peaks in the categories reflects the ubiquitous expression of CTCF: there are significantly more common peaks (present in both cell types) for CTCF than for P300.

In the second step of the peak analysis, we explored several characteristics of the mixed peaks and observed interesting trends. We looked at the overall number of peaks, the height of the peaks and the read density per peak. We observed anti-correlation patterns for most of those

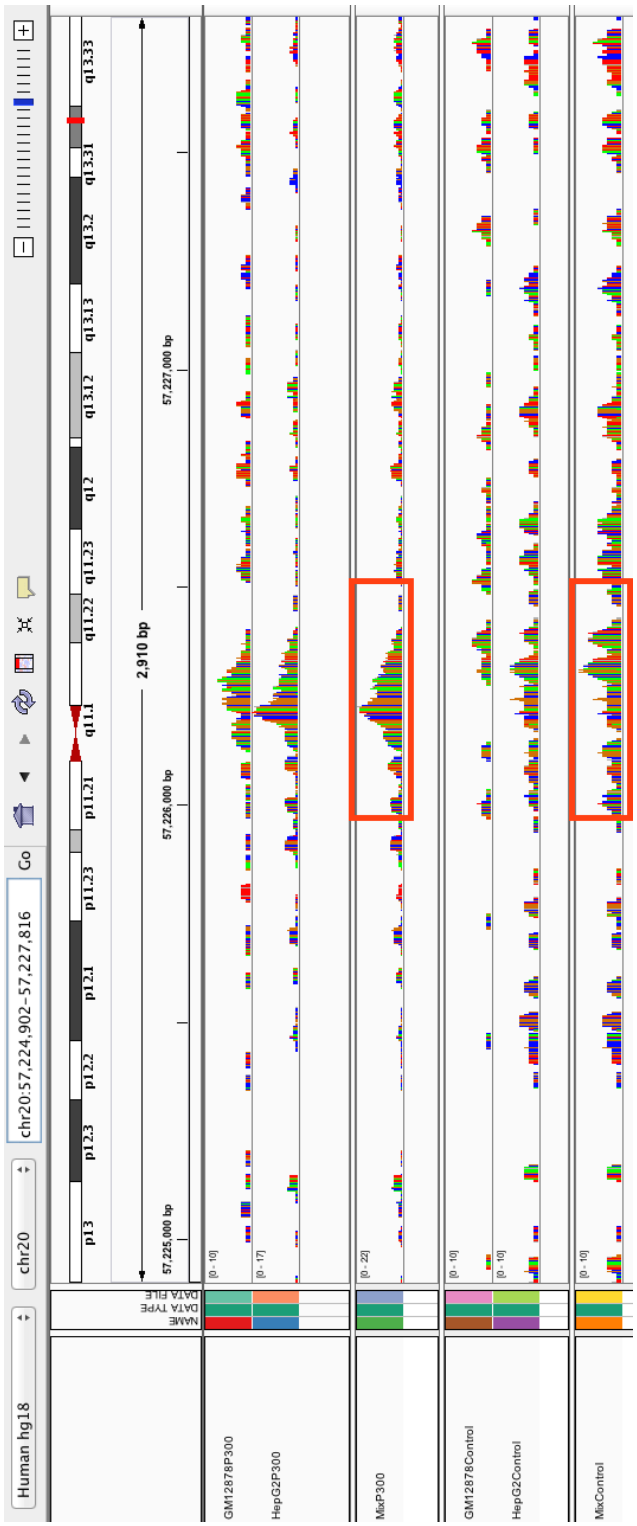


Fig. 3.6 Example of a lost peak in the mixture file. The six tracks depicted on this figure represent the two component signal files along with their respective control files and the mixture signal and control files. The highlighted peak in the mixture signal file was not detected by the peak caller and this is likely due to the increased read density highlighted in the control file.

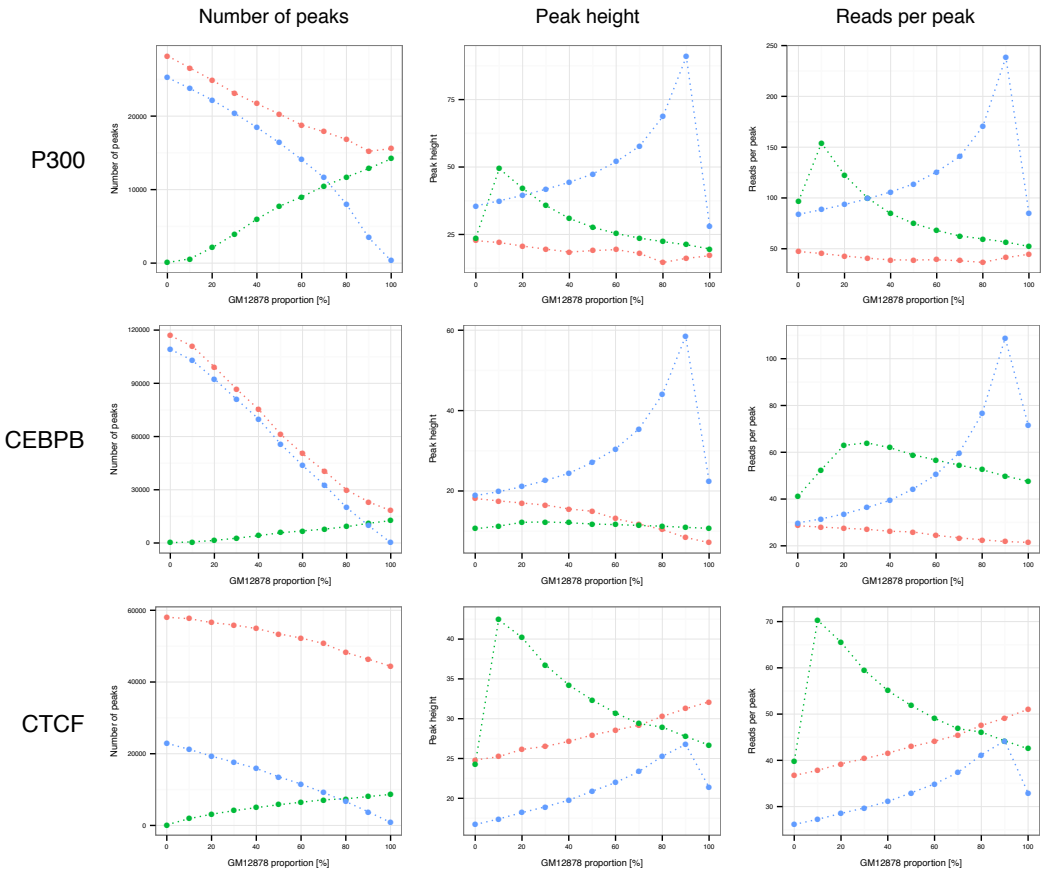


Fig. 3.7 Peak characteristics for the three proteins. For each protein (P300, CEBPB, and CTCF), the plots show the number of peaks that were called in each mixture file along with the mean peak height, and the mean number of reads per peak for different mixture ratios (from 0% to 100% GM12878 content). The green data points correspond to peaks assigned only to GM12878, the blue data points to peaks assigned only to HepG2, while the red data points represent all of the peaks. This figure shows the expected similarity in the patterns of P300 and CEBPB compared to CTCF due to the nature of the expression of the proteins. Additionally, the amount of common CTCF peaks between the cell lines compared to the other proteins, as seen in figure 3.4, is also observable.

parameters, which suggests that there exist differences in the nature of the peaks from the different cell types. The figure 3.7 shows the number of peaks that were called in each sample file along with the mean peak height, and the mean number of reads per peak per sample. It shows the expected similarity in the patterns of P300 and CEBPB compared to CTCF due to the nature of the expression of the proteins. The amount of common CTCF peaks between the cell lines compared to the other proteins, as seen in figure 3.4, is also observable. Since CTCF is ubiquitously expressed its binding pattern should not vary considerably between cell types and thus, the peak parameters should also show low variation. We observe considerable variation both in terms of height and density of the peaks depending on the ratio of cell type mixture and the reasons behind this are unclear. Another aspect for which we have no explanation is the significant drop for both the height and density parameters at the extreme mixture ratios.

However, overall the performed peak analysis, mainly through the observed anti-correlation patterns, suggests that the properties of peaks resulting from different proteins vary. This supports our hypothesis that complex tissue ChIP-seq deconvolution based on peak parameters may be feasible and this is explored further below.

3.4.2 Peak deconvolution using machine learning

Since our peak analysis suggested that peaks from different cell types showed differences in the investigated parameters, we decided to perform a classification approach using machine learning methods, which are known to be efficient for these kind of problems (see section 3.2.6). We focused on supervised learning methods (see section 1.10.2) and used different datasets for training the models before testing them. Using a specific machine learning software package (see section 3.3), we first implemented a logistic regression model and secondly tried a Support Vector Machine model (with a linear kernel). Concretely, I merged detected peaks from multiple in-silico mixture files (of the same mixture ratio), creating a list of peaks defined by their locus, height, and density (number of reads in the peak). Each of these peaks were also marked as binding in one cell type, the other cell type, or both (see section 3.3.4). To perform the classification, 50% of peaks were randomly selected to be part of the training set and the other 50% in the testing set (see section 1.10.2). The prediction results were, however, very similar, between the two models, which could be expected as SVMs (with linear kernels) and logistic regression are both linear approaches (see section 1.10.2) [225]. Overall, the results were not good enough to suggest we had successfully deconvolved the peak profiles. Generally, the prediction performed much better

for extreme mixture ratios. This can be observed on figure 3.8 and on figure 3.9, which depict the respective ROC curves for two different mixture ratios (50% and 90% GM12878), using the SVM model to classify the peaks.

Although not good enough to call this classification approach deconvolution, the ROC curves and respective AUC values (see section 1.10.2) were generally much higher than random. These results thus support the previous hypothesis that the peak characteristics tend to vary according to which cell type they come from.

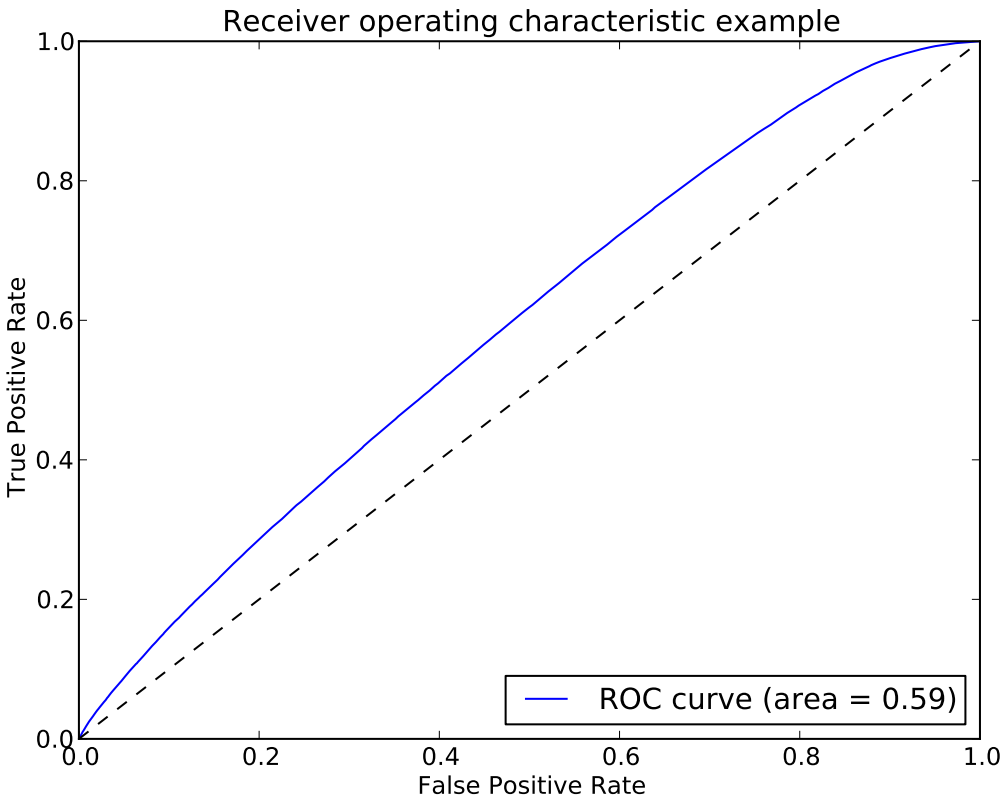


Fig. 3.8 Peak deconvolution (with SVM) ROC curve with mixture ratio 50%. This figure depicts the ROC curve obtained from an SVM classifier used to deconvolve the peaks in a 50% GM12878 content mixture file. The results obtained from this approach were significantly better for extreme mixture ratios (see figure 3.9).

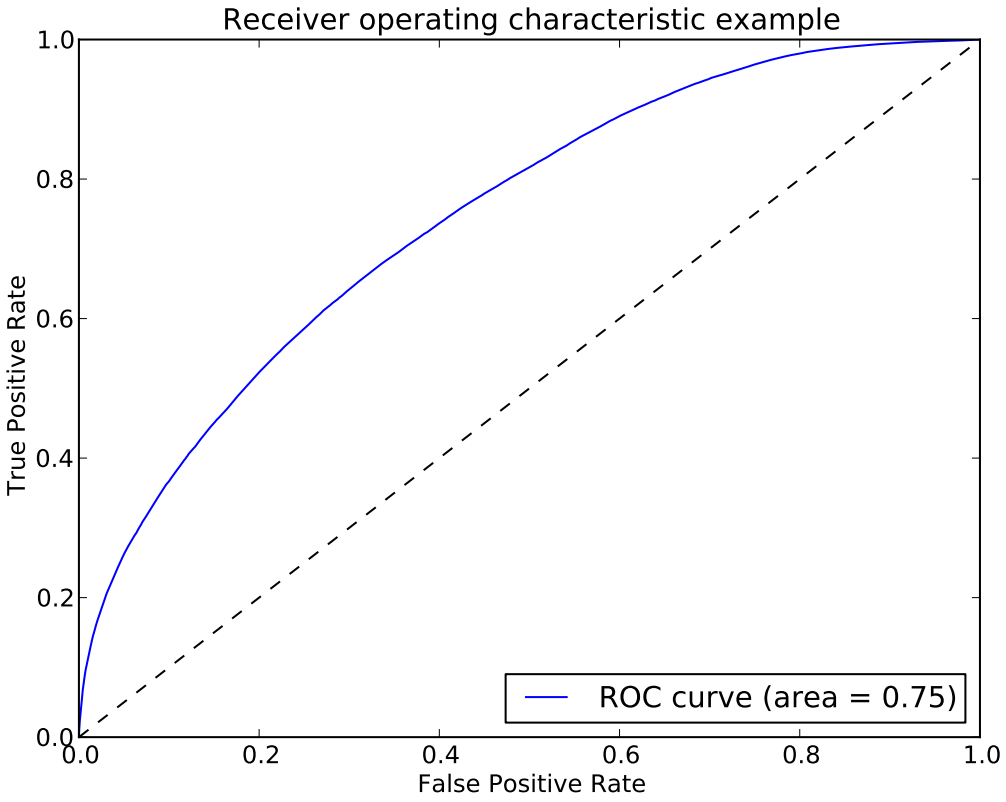


Fig. 3.9 Peak deconvolution (with SVM) ROC curve with mixture ratio 90%. This figure depicts the ROC curve obtained from an SVM classifier used to deconvolve the peaks in a 90% GM12878 content mixture file.

3.4.3 Bin approach

We faced a number of different challenges when working with the previously mentioned ChIP-seq peak dataset. Retrieving peaks from the subcomponent files that would correspond to the mixture file peaks implied defining overlapping thresholds in addition to having to deal with complex corner cases (for example when there were more than one overlapping peaks). Additionally, as previously described (see section 1.8.3), peak calling algorithms are complex which makes it hard to understand what is happening in particular cases such as peaks present only in the mixture file or peaks present in both subcomponent files but not in the mixture file.

To further investigate ChIP-seq binding profiles in complex tissue scenarios, we decided to create a simpler approach to analyse transcription factor binding data, that removed the complex peak-calling phase. The ChIP-seq method generates reads which map to genomic regions that were bound by the protein of interest, creating enriched read regions (see section 1.8.3). By using a binning approach, it is possible to analyse the number of reads per bin and thus assess read enriched regions that likely correspond to protein-bound regions. Such a read-bin model also enables the comparison of multiple files in a straight forward manner, as multiple tracks can be combined in a matrix format (each cell corresponding to the number of reads in that particular bin). We decided that the length of the bins should be 1kb (1,000 base pairs). We investigated several thresholding methods to decide if a bin should be considered "bound" or "unbound" and initially selected a fixed threshold of 30 reads. This threshold yielded numbers of bound bins in the same order of magnitude as the number of (previously called) peaks. A more complex thresholding system based on the specific distribution of reads in each file would probably be more suitable, but the motivation behind this model was simplicity so we decided against a more complex approach. With these fixed criteria, we created binding profiles for our complete dataset. In a similar but simpler manner as previously, we categorised the bins from the mixture files as 1) bins bound in both subcomponent files, 2) bins bound in subcomponent file A, 3) bins bound in subcomponent file B, and 4) bins that were unbound in the subcomponent files. As with the novel-peaks found in the mixture files, very few bins fell into category 4 and for the purpose of classification this category was ignored.

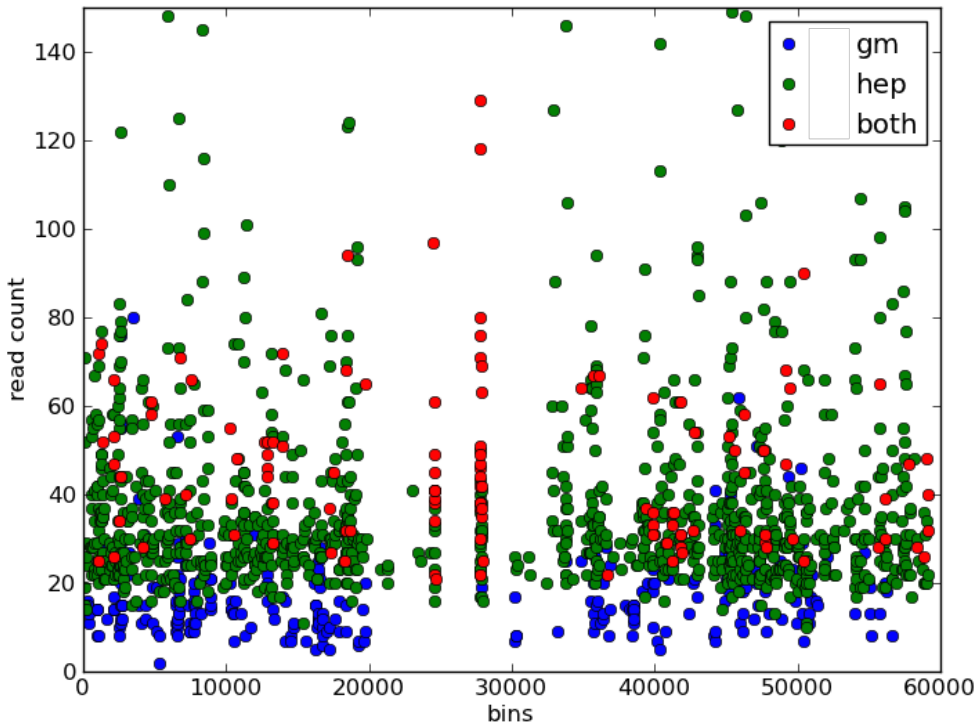


Fig. 3.10 Read bins from a P300 mixture with 20% GM12878 content (chromosome 19). This figure depicts the number of reads per bin along the entire chromosome 19. The bins are coloured depending on the presence of reads in the respective component file bins (using a threshold, see section 3.4.3).

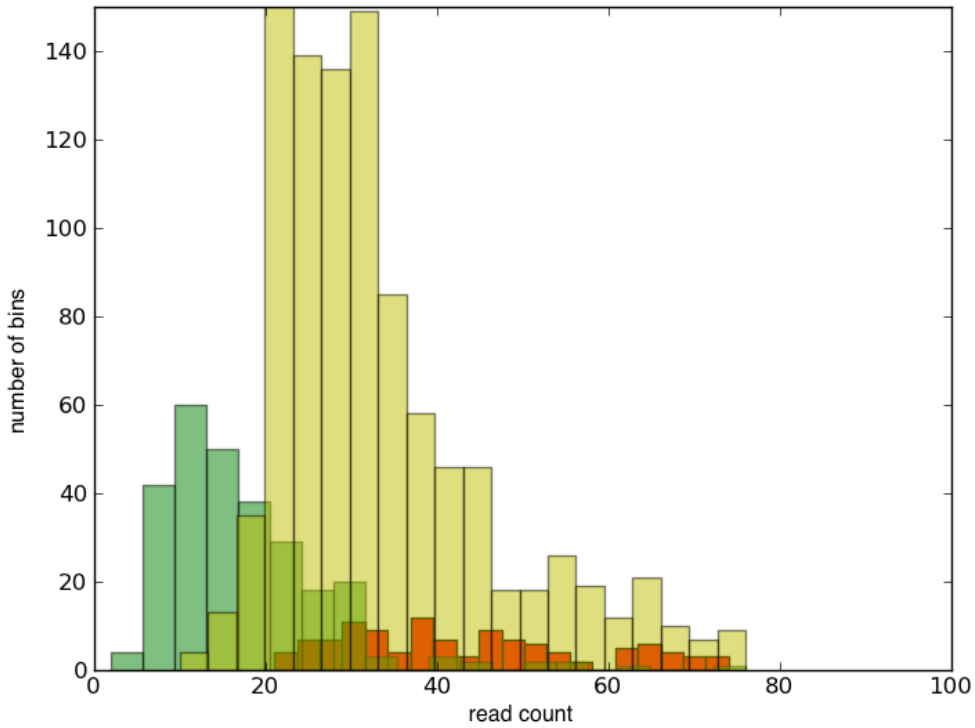


Fig. 3.11 Distribution of bin categories for a P300 mixture with 20% GM12878 content (chromosome 19). The green distribution corresponds to the GM12878 bins, the yellow to the HepG2 bins, and the red to the bins that are "bound" in both component files.

3.4.4 Bin deconvolution with machine learning

Before applying machine learning methods to classify the previously generated read bins, we analysed the bin categories. First, we plotted each bin of mixture files, coloured according to its assigned category. This revealed the layering of each types of bin and although the layers overlap significantly, they are nonetheless different (see figure 3.10). Further plotting these bins using a histogram showed the different bin category distributions (see figure 3.11).

We then proceeded following the same methodology applied when attempting to classify the peaks (see section 3.4.2). We used the same machine learning model (SVM with linear kernel) which we initially trained on a group of datasets including bin data for different cell types and different mixture ratios. We then used the classifier to predict bin categories on different datasets. With this predictive model, as with the previous peak datasets, we obtained inconclusive results. They were only slightly better than random assignment. The bin analysis revealed that in particular cases one bin distribution differs from the others (see figure 3.11). This lead us to try to classify the bins with a simplified two-by-two approach. Using the same machine learning model again, we attempted to only separate the bins into two categories each time. For example, peaks that are only present in cell type A as category one, and all other peaks as category two. Comparing all ROC curves together revealed that the best performance was obtained when predicting the lower proportion cell type bins (see figure 3.12), which is consistent with the previously observed bin distributions (see figure 3.10 and figure 3.11). In conclusion, the information present in the binding signals (reads per bin) is insufficient to properly deconvolve complex tissue ChIP-seq data. However, the information is not null and could prove useful to include in a more complex model potentially combining this with other information sources (for example signature binding profiles, see section 3.5.2).

3.4.5 Bin deconvolution with simple variation model

The two previously described approaches to solve complex tissue ChIP-seq deconvolution, by using machine learning models to classify protein-binding signals (see section 3.4.2 and section 3.4.4) revealed that there exists differentiating information within those signals. However this is not sufficient to extract the two separate signals from the mixture. Inspired by the methods applied in gene expression deconvolution (see section 3.2.5) and by machine learning approaches to the “cocktail party problem” (see section 3.2.6), we decided to investigate the use of variation between samples to deconvolve the complex tissue ChIP-

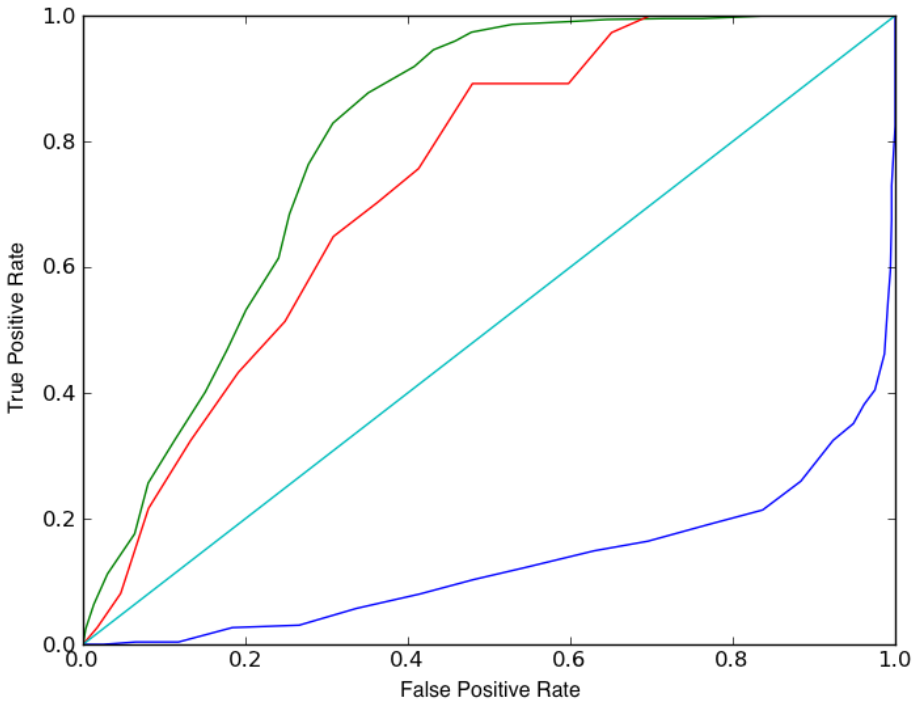


Fig. 3.12 Bin deconvolution (with SVM) ROC curves with mixture ratio 80%. The two-by-two classification ROC curves are respectively, blue for GM12878 bins, green for HepG2 bins, and red for bins "bound" in both.

seq signal. In the "cocktail party problem" scenario, if two outputs are recorded with one input, it is very challenging to deconvolve the two signals. However, when a second input is added to the situation, the variation between the two recorded signals constitutes enough additional information to deconvolve the two signals precisely. The quality of the deconvolution improves as the variation increases (for example if the inputs are placed far from each other). In the analogous ChIP-seq data deconvolution, the outputs correspond to the individual cell type binding signals, while the input is the complex ChIP-seq data (mixture signal). Additional inputs would be different complex tissue ChIP-seq experiments (additional mixture files). The variation that would support the deconvolution would come from the variation in the cell type mixture proportions of the samples in the additional experiments.

The read-bin dataset that was generated previously (see section 3.4.4) contains multiple mixture files with many different cell type proportions. We used this data and created pairs of mixture files with different cell type proportions. As a simple first approach, we selected pairs with opposite mixture proportions (for example pairing a 30% cell type A mixture with a 70% cell type A mixture). We then built a straightforward model based on thresholds to quantify the variation of reads for each bin. The predictive algorithm iterates over every bin of each of the input files. For each pair of bins, it initially compares the read numbers to a bound/unbound threshold to decide whether these bins contain signal. If either of the bins contains enough reads, it then calculates the ratio between the read counts of each bin. This ratio is then compared to a second threshold to assess if the read count has increased, decreased or remained similar in respect to the reference mixture file. Since the cell type proportions of the mixture files are known, it is then possible to assign each bound bin to cell type A or B (read count increasing or decreasing) or both (read count similar). This approach is formally described in Python below (see listing 3.1). Both model parameters (signal threshold and variation threshold) can be optimised for the particular dataset by looping through different values with a training set before applying the model to the test set.

Listing 3.1 Variation model approach in Python.

```
# This function takes as input two values corresponding
# to the number of reads in their respective bins from two
# mixture files, one with a small fraction of cell type A
# (and large of cell type B) and the other a large fraction
# of cell type A (and a small of cell type B)

# The signal threshold is used to determine if a bin is
# bound or unbound
signalT = 30
# The variation threshold is used to determine if a bin
# has varied in content in both mixture files
variationT=0.2

def variationDeconvolution(smallFracBin, largeFracBin):
    # We start by checking that they are not both unbound
    if smallFracBin < signalT and largeFracBin < signalT:
        return "Unbound"

    # We calculate the ratio between the two bins
    ratio = smallFracBin / largeFracBin
    # If the small fraction bin is bigger than the large
    # then we assign this bin to being bound in cell type B
    if ratio > 1 + variationT:
        return "bound in cell type B"
    # If it is the opposite then we assign the bin to being
    # bound in cell type A
    elif ratio < 1 - variationT:
        return "bound in cell type A"
    # If the bin has not varied in size (past the threshold)
    # then we assign that bin as bound in both cell types
    else:
        return "bound in both cell types"
```

This simple approach produced good results for the mixture pairs with most extreme variations (see table 3.3). However the results were more contrasted with the pairs having similar cell type proportions (see figure 3.13). To further visualise these results, we used a genome browser to compare the pure read signal track, with the annotated bins track and the previously generated peak track (from section 3.4.1). An example region in which all three categories of bound signals are present (a GM only signal, a HEP only signal, and a signal present in both) is shown in figure 3.14. Although, this method is trivial and the large variation required to obtain good results means that the application of this approach to real complex tissues is unrealistic, it serves as a proof of concept and demonstrates the potential of this approach. Increasing the number of mixture files should improve results in situations with lower variation between the cell type proportions. Further improvements to this approach are discussed below (see section 3.5.2).

	Both	HepG2	GM12878
Chr 1	0.68 (202)	0.97 (1605)	0.79 (404)
Chr 2	0.64 (92)	0.97 (440)	0.82 (196)
Whole genome	0.75 (2905)	0.97 (20271)	0.77 (4008)

Table 3.3 Variation model prediction results with a 20%-80% pair. This table shows the ratio (and number) of successfully predicted bins for each of the three bin categories and in three different contexts (individual chromosomes and whole genome).

3.5 Discussion

3.5.1 Deconvolution approaches

In the context of this project, we have investigated three different approaches to solve complex tissue ChIP-seq deconvolution. The first two are based on the same concept of attempting to classify peaks/bins according to their characteristics, while the final approach made use of the variation of cell type proportions between different mixture files. The main difference between these approaches is the necessary input to perform deconvolution. The first two classifying methods were designed to be trained on large quantities of data but then applicable to a single complex tissue ChIP-seq data file. This kind of approach is the most practical for "real world" applications since the already trained predictive models could be used in a straightforward manner on any complex tissue ChIP-seq dataset. On the contrary, the

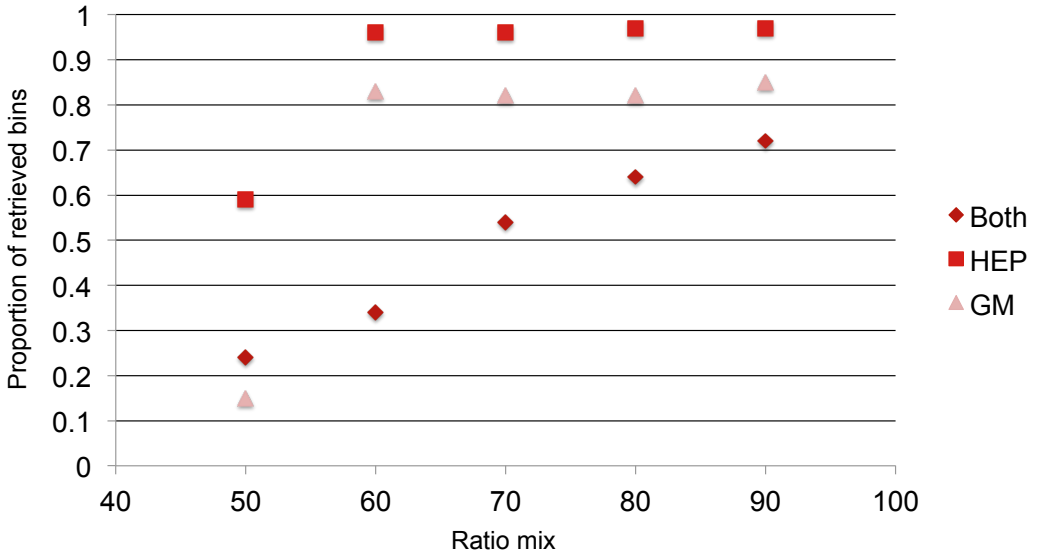


Fig. 3.13 Bin deconvolution with the variation model. This figure shows the ratio of successfully predicted bins (in chromosome 2) for each of the three bin categories and different mixture ratio pairs. For example, the results for the 30%-70% pairs are shown at mixture ratio: 70. The variation prediction model performs better with increasing variation within the pair.

Fig. 3.14 Comparison of the bin predictions with the read profiles. This figure shows our predicted bins along with the raw reads and the peaks that were called (see section 3.4.1). This comparison demonstrates the performance of our variation based deconvolution model (with significant variation) and also suggests that our bin size is appropriate.

variation approach requires the dataset to be generated with deconvolution in mind, including samples with different cell type proportions, so as to make deconvolution feasible. This makes the method the least appealing for "real world" applications as, in addition to the many limiting factors (time, cost or even tissue quantity for example) linked with running multiple ChIP-seq experiments. It will also be extremely challenging, in many cases, to obtain such high variability (of cell type proportions) biological samples.

In terms of deconvolution quality, the binning approach was very similar to the peak based approach. The benefits of the binning approach were mainly on the computational aspect, making many corner cases disappear (for example when multiple peaks overlapping with one). It created a nice mathematical structure to view and process the data and when using appropriate thresholds (~30), the number of bins and peaks were very similar. Furthermore, when we analysed the bin tracks alongside the peak and raw read data tracks, we observed very consistent binding profiles (see figure 3.14). The bin structure was particularly useful for the third (variation based) approach. It enabled the scanning of the entire genome for each track to be done in a clear and simple manner. The main disadvantage of this method is the issue of binding sites overlapping two bins, especially when the read count in both bins falls just below the bound threshold. However, it is unlikely that this case happens often, but further investigation would be required to analyse and optimise this issue.

Overall, our results show that there exists a potential for deconvolution approaches targeting complex tissue ChIP-seq data. Although the deconvolution results of the predictive models were for the most part inconclusive, they were better than random. This suggests that there is some information that we are able to recover, but the models require additional information to perform optimally. We believe that the methods we have applied may constitute a base layer for a more complex approach to complex tissue ChIP-seq deconvolution that would incorporate more information (see section 3.5.2).

3.5.2 Future directions and improvements

As previously mentioned, the deconvolution approaches that we applied require additional information to perform optimally. This new information could come from different sources and the following list contains different potential targets.

- **Gene expression data:** Since transcription factors are involved in gene regulation and expression (see section 1.4.2), it is likely that including gene expression data in a deconvolution model would increase the overall information content, especially if the

gene expression can be deconvolved efficiently (for example with existing methods such as the ones described in section 3.2.5). This would increase the complexity of the experimental procedure as RNA-seq would have to be performed on the same tissue.

- **Combining multiple proteins:** As well as binding DNA, transcription factors also interact between each other. This implies that some transcription factors tend to co-localise with others, and the contrary is true as well (transcription factors "blocking" access to others). This means that a dataset with multiple proteins assayed in the same complex tissue may in certain circumstances (if there exists a dependency between the proteins for example) contain extra information that could improve deconvolution. We discuss this further in section 3.5.3, which describes a different deconvolution methodology using a hidden markov model. Such an approach would involve performing multiple assays for different proteins, however, this is standard practice in many scientific studies.
- **Signature binding profiles:** Although our entire concept of complex tissue ChIP-seq deconvolution is based on the fact that we do not know what the cell type specific binding profiles are, making use of signature profiles would very likely improve the process. High quality cell type specific binding profiles may be unavailable, however, low quality profiles could be obtained in different ways and would nonetheless, likely improve the quality of the results. Such low quality profiles could be obtained by profiling the protein in a similar cell type that can be purified. Another more modern option could be to use the recently developed single cell ChIP-seq assay (see section 1.8.3), which produces very low coverage binding profiles. It may also be possible to purify only one of the cell types contained in the complex tissue (for example the majority cell type) and the addition of this binding profile to a deconvolution model would also likely increase the prediction quality.
- **Combining methods:** The approach with the most promising results was the third method that takes advantage of cell type proportion variation between input samples. However, as discussed above, this is also the most "unrealistic" model as the variation required to obtain good results is unlikely to be attainable in a complex tissue sample. It is likely that with lower variation, a high number of samples would be required to obtain similar results. However, another option that may improve the deconvolution quality with low variation and low number of samples, is to combine this variation based method with the classification approach. Additionally, any of the other three options to increase the information content of the model could be included as well.

Furthermore, in addition to the many improvement options in terms of adding complexity to the deconvolution model, technological advances might also improve the potential for complex tissue deconvolution in the future. As the cost of sequencing continues to drop and the robotisation of experimental assays spreads, generating large datasets will likely become the norm. This would be highly beneficial for the prediction model based on variation. Technological improvements to the experimental protocol itself will likely generate binding profiles of higher quality and precision, which in turn may improve the complex tissue deconvolution process (for example by eliminating open chromatin noise, see section 1.8.3).

3.5.3 Deconvolution using multiple transcription factors with a HMM

As discussed above (see section 3.5.2), a potential option to deconvolve complex tissue ChIP-seq data would be to take advantage of the interaction (dependency) between different transcription factors. Such a model would require as input multiple ChIP-seq experiments with different proteins for the same tissue. The method would then go through all of the bins (in a bin based model) for each of the different ChIP-seq files analysing the combination of bound transcription factors and using this information to deconvolve a particular reference ChIP-seq file. The hypothesis here is that the dependencies between factors varies in different cell types. Since we can observe the bound combination of factors but cannot know if a particular site is bound in cell type A or B or both, this approach seems suitable for a HMM model.

Hidden Markov Models (HMMs) have been widely used in bioinformatics for a number of different purposes. These theoretical tools are used to model Markov processes in which some or all of the states are hidden and the output of those states are observable variables. These mathematical tools are known to be efficient in predicting the context of sequence regions. As an example, ChromHMM is a model that was built to predict different chromatin states and has been successful in doing so [226]. In our complex tissue ChIP-seq mixture environment, two different HMMs can be implemented "on top of each other", such complex HMM structures are called hierarchical HMMs. The first model has two hidden states: "bound" and "unbound", and the observable variables are the number of reads in the bin. To simplify the model, each variable could correspond to a range of read count values (for example with three observations, A: less than 15 reads, B: 15-35 reads, and C: more than 35 reads) and the probabilities of observing A would be much higher in the "unbound" than in the "bound" state. The second is slightly more complex, here the observed variables correspond to the different combinations of ChIP-seq signal for the proteins (for example, with two

proteins, one observation is there is a peak for protein A and for protein B), while the hidden states are (with two cell types): “Proteins only bound in cell type A”, “ proteins only bound in cell type B”, “proteins bound in both cell types”, and “proteins are not bound”, which correspond to the different possibilities for a bound bin in the ChIP-seq file being deconvolved. Each hidden state of the HMM also has three sets of probabilities: transition probabilities, observation probabilities and starting probabilities. The set of transition probabilities defines the probability for each state (including itself) of changing to that state after each step. The set of observation probabilities defines the probability of observing a particular variable from each state. Similarly to machine learning models, these probability sets can be optimised (or “learned”) from training datasets. Finally the set of starting probabilities determines how likely is the process going to start for each state. This second HMM is illustrated in figure 3.15, it has a set of four hidden states $S = \{A, B, AB, U\}$ (A: “only bound in cell type A”, B: “only bound in cell type B”, AB: “bound in both”, U: “bound in neither”), four observable variables $O = \{1, 2, 12, 0\}$ (1: "protein 1 bound", "2 protein 2 bound", "12: both proteins bound", 0: "None bound"), the set of transition probabilities $P_T = \{P_{AA}, P_{AB}, ..., P_{UU}\}$ (P_{AA}: probability of staying in state A, P_{AB}: probability of changing from state A to state B, etc.), the set of observation probabilities $P_O = \{P_{A1}, ..., P_{U0}\}$ (P_{A1}: Probability of observing variable 1 from state A, etc.), and finally the starting probabilities $P_S = \{P_{SA}, ..., P_{SU}\}$ (P_{SA}: probability of starting in state A).

Since we had generated complex tissue ChIP-seq data for three different proteins (see section 3.4.1), we attempted to build and test this model. The theory and the evidence of a successful HMM in a similar scenario [226] suggest that this model is adequate. However, unfortunately, after implementing the HMM, the optimisation of the probability sets using a training dataset did not yield adequate probabilities (probabilities of almost 1 or 0) and the resulting model could not achieve any deconvolution. The reason behind this failure is unclear and further investigation into how to fit and test this model correctly is required. Finally, even though we did not succeed in developing this model, we discuss it here as we believe that this approach provides an additional interesting way of tackling complex tissue ChIP-seq deconvolution.

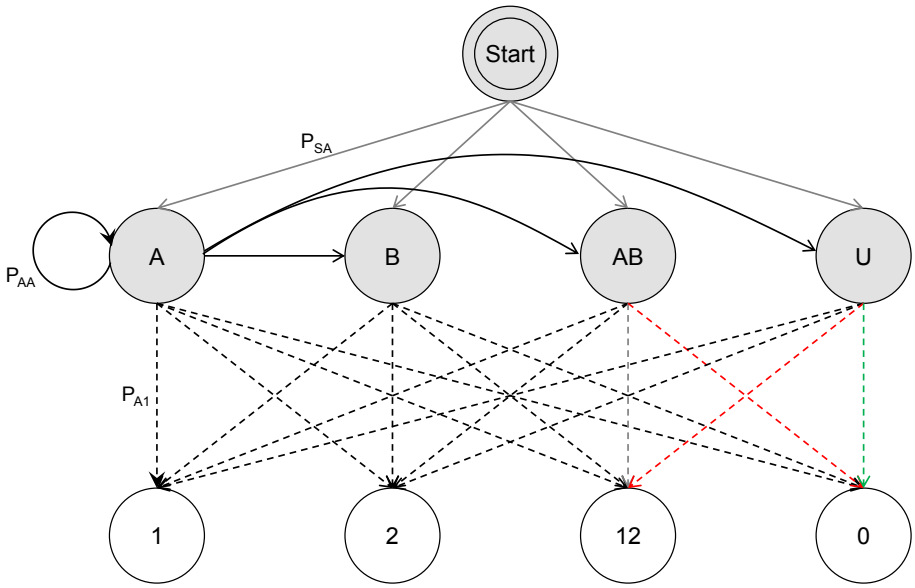


Fig. 3.15 HMM deconvolution approach. This figure depicts the graphical model of the HMM built to deconvolve complex tissue ChIP-seq. It has a set of four hidden states (grey circles) $S = \{A, B, AB, U\}$ (A: “only bound in cell type A”, B: “only bound in cell type B”, AB: “bound in both”, U: “bound in neither”), four observable variables (white circles) $O = \{1, 2, 12, 0\}$ (1: "protein 1 bound", "2 protein 2 bound", "12: both proteins bound", 0: "None bound"), the set of transition probabilities (black arrows) $P_T = \{P_{AA}, P_{AB}, ..., P_{UU}\}$ (P_{AA}: probability of staying in state A, P_{AB}: probability of changing from state A to state B, etc.), the set of observation probabilities (dashed arrows) $P_O = \{P_{A1}, ..., P_{U0}\}$ (P_{A1}: Probability of observing variable 1 from state A, etc.), and finally the starting probabilities (grey arrows) $P_S = \{P_{SA}, ..., P_{SU}\}$ (P_{SA}: probability of starting in state A). For example, we can intuitively expect the probabilities P_{AB0} and P_{U12} to be very small (probability of detecting signal for both proteins when they are unbound in both cell types, and probability of observing no signal when they are both bound), similarly we can expect P_{U0} to be high (probability of observing no signal when nothing is bound).

3.5.4 ChIP-seq data used

In the context of this project, we decided to make use of real ChIP-seq data obtained from the ENCODE project (see section 3.4). Since multiple datasets for several protein-cell type combinations were available, we did not consider other data options at the start of the project. Since our results suggest that performing deconvolution based on the variation in cell type proportions between samples is efficient (with high variation), a logical next step, as discussed above (see section 3.5.2) is to investigate larger datasets with less variation. In this situation, instead of generating or searching for such a large dataset, an alternative would be to make use of simulated ChIP-seq data. A number of different ChIP-seq data simulators are available and using such data to develop and tune predictive models may prove an ideal complementary approach to test and improve ChIP-seq deconvolution methods in the future.

3.6 Conclusions

Through different methods, we have attempted to perform complex tissue ChIP-seq deconvolution. Although successful deconvolution was not achieved, our results suggest that, at least in particular contexts, this process is feasible. The methods we have investigated showed promising aspects, in particular the last cell type proportion variation based deconvolution model. We also discussed a number of improvements that have the potential to significantly improve the predictive power of the models. Different model approaches could be investigated and inspiration for these could be found in other bioinformatics areas; since modern sequencing assays generally require samples containing a high number of cells to obtain meaningful signals and that most real world samples are heterogenous, deconvolving the results of modern assays has been a point of focus in multiple different areas. Gene expression deconvolution has been mentioned previously (see section 3.2.5) but deconvolution methods have been developed in cancer genomics to uncover the fractions of different subpopulations of cells (subpopulations of cells in tumour samples often vary in terms of mutations and copy number variations). Probabilistic approaches developed to infer intra-tumour heterogeneity have shown excellent results and could potentially be adapted to the ChIP-seq deconvolution problem. For example, THetA (a software for Tumor Heterogeneity Analysis) [227] uses a maximum likelihood approach to maximise the probability of observing the experiment sequencing reads given the underlying cancer genomes and the genome mixing fractions. Formally, the genome is segmented into non-overlapping segments $I = (I_1, ..., I_m)$ and the probability that needs to be maximised is $P(\mathbf{r}|\mathbf{C}, \mu)$, where $\mathbf{r}$ is a vector containing the

read depth for each segment in I obtained from aligned sequencing reads, $\mathbf{C}$ is a $m \times n$ matrix representing the counts of each segment for each of the n subpopulations, and μ a vector containing the mixing fractions for each n subpopulations. Finally, despite the recent single-cell ChIP-seq experiment [89], we strongly believe that being able to study cell type specific protein binding in tissues, without having to separate the cells (which may influence the binding profiles, see section 3.2.2), would greatly benefit genomic research. Overall, the research performed in the context of this project has provided valuable insight into the potential for ChIP-seq deconvolution, and will hopefully serve as a basis and motivation for further investigation.

Chapter 4

Conclusion

In this thesis, I have described in detail two separate projects that are based on the same underlying concept: data reusability. As previously described in the introduction (see section 1.7), the arrival of next generation sequencing methods, as well as other modern high throughput experimental technologies, have lead to an increasing amount of publicly available data. Although it is already common practice to examine existing data in particular contexts, such as follow-up studies that build on the previous methodology, it is less common to observe data being reused in totally different contexts. In this work, I focused on data produced by the ChIP-seq experimental method, which is designed to detect genomic regions that are bound by (DNA-binding) proteins (see section 1.8.3). Both projects make use of previously generated ChIP-seq data, firstly to investigate a mitochondrial DNA phenomenon and secondly to model and explore the concept of complex tissue ChIP-seq deconvolution. In addition to showcasing two different applications for ChIP-seq data reusability, both projects provided valuable scientific insights.

The first project investigated the use of ChIP-seq data to detect mitochondrial heteroplasmy (see section 2.2). The first key finding of this work was that in most cases, ChIP-seq data contains relatively high sequencing coverage of the mtDNA (see section 2.4.2), which is sufficient to detect heteroplasmic positions with a "high" minor allele level (see section 2.4.3). In the second phase of the project, we performed a broad analysis of the phenomenon, covering a number of different characteristics in a dataset comprising 79 individuals from 16 species. Our results, which are in the process of being published, provide insights into mitochondrial heteroplasmy over a large portion of the mammalian phylogeny.

Through the second project, I explored the potential for complex tissue ChIP-seq deconvolution. The initial step was to model complex tissue ChIP-seq data by creating in-silico

mixtures of previously generated ChIP-seq data for cell lines (obtained from the ENCODE project). The protein binding profiles of the model complex tissue experiments were first produced using peak calling software, then using a simpler binning approach. The data was then analysed before different classification methods were applied to attempt to deconvolve the individual signals. Despite not having succeeded in performing complex tissue ChIP-seq deconvolution, the results obtained suggest that there are detectable differences between the binding profiles and that there is potential for more complex approaches, which may include additional information, in order to succeed in deconvolving complex tissue ChIP-seq data.

As a whole, this thesis has described novel uses of existing ChIP-seq datasets, enabling us to gain deeper biological understanding of biological systems, as well as providing valuable insight into the assays used to measure them.

4.1 Future work

Below, I list a number of different directions follow-up studies could take in investigating the reuse of ChIP-seq data for heteroplasmy detection as well as for exploring complex tissue deconvolution. Naturally, these suggestions will evolve as the underlying methodology and technology improve and the impact of this is discussed in the next section (see section 4.2).

4.1.1 Mitochondrial Heteroplasmy

The mitochondrial heteroplasmy study that I have performed in the context of this thesis is in itself relatively complete. Using the appropriate experimental validation methods, a further exploration of the lower level heteroplastic positions would be possible. Additionally, investigating the few high level heteroplasmies that were not detected by the experimental validation could provide further insight into the underlying causes of either the validation process not detecting positions or the algorithm misinterpreting positions. Nonetheless, now that the concept of using ChIP-seq data to detect heteroplasmy has been demonstrated, follow-up studies could investigate much larger ChIP-seq datasets within or across species. The Mouse ENCODE project for example has a large quantity of publicly available ChIP-seq data. From the technical side, the detection model could be improved. For example, it could automatically set the optimal parameters depending on the available coverage, which would enable different levels of heteroplasmy detection depending on the data.

4.1.2 Complex Tissue ChIP-seq Deconvolution

Having established that there is potential for complex tissue ChIP-seq deconvolution, further work should be aimed at solving the problem by adding information to the model. A number of potential improvements have already been discussed (see section 3.5.2) and the implementation and testing of these would provide valuable insight into how applicable such an approach might be for a real dataset. Similarly, a large dataset of simulated ChIP-seq data could be used to explore how much data would be required to deconvolve complex tissue ChIP-seq signals using a variation detection model in more realistic, low variation contexts.

4.2 Potential impact of future technology

As described in the introduction (see section 1.4.2), proteins play key roles in cellular function and to understand these roles, it is necessary to locate the genomic regions with which these proteins interact. The use of chromatin immunoprecipitation to study protein binding profiles was first performed nearly thirty years ago [228], before being transformed by the arrival of next generation sequencing technology and adapted to ChIP-seq nearly ten years ago [106]. Moreover, last year was the first time this experimental method was performed on a single cell basis [89] and there is no doubt that this experimental method will further improve and evolve in the future.

In my opinion, the two main steps of the ChIP-seq protocol that will likely improve the most are the immunoprecipitation and the sequencing.

The costs of sequencing are dropping and this will directly impact ChIP-seq studies in two ways. First, the precipitated residues will be sequenced at much higher coverage, which should improve the quality of the binding profiles produced. Second, more replicate experiments will be performed. There are already robotised protocols that are able to perform ChIP-seq experiments on 96 samples at a time [157]. Furthermore, the progress of sequencing methods that do not use PCR is also promising. Such methods would increase the homogeneity of binding profiles, reducing noise and biases due to PCR duplicates. Both my projects would greatly benefit from these advantages; the quantity of replicates is likely to significantly improve the power of deconvolution methods, while PCR free sequencing should increase the precision of heteroplasmy detection methods as well as making them simpler.

Today, the quality of ChIP-seq results depends largely on the specificity of the antibody (how well it binds the protein of interest). Despite the use of methods and benchmarks to validate antibodies, the quality of the antibodies used (and the produced ChIP-seq data) vary. If synthetic biology technology progresses enough, it is possible that, in the future, antibodies will be "made to measure". A significant increase of the binding affinity would considerably reduce the noise currently observable in ChIP-seq data (open chromatin regions for example, see section 1.8.3). Although this will not have any impact on heterolasmy detection (as ChIP signals are not relevant), it may have an impact on deconvolution approaches ensuring only true protein binding signals are considered for classification.

Progress in both these aspects will enable smaller samples to be used to produce high quality results. If the number of required cells (to obtain good results) diminishes enough, obtaining pure samples for some of the major cell types of complex tissues may become possible. Using such data to then deconvolve an experiment performed on a complex tissue sample would likely improve deconvolution, without perturbing the cells with the separation process (see section 1.8.3).

4.3 Final words

As further research is performed and modern technology is developed and applied, it is likely that significant progress will be made both in terms of increasing our knowledge of mitochondrial heteroplasmy and deconvolving protein binding profiles from complex tissue ChIP-seq experiments. This could lead to discoveries that would not only add to the scientific knowledge around cellular and mitochondrial function but also impact medical research. As stated in section 2.2, heteroplasmic positions have been linked to many different human diseases, and tumours are extremely complex tissues with interesting expression and protein binding patterns. It is clear that further understanding of the underlying workings of such medical issues is key to investigate potential solutions.

Finally, to conclude this thesis, I would like to emphasise that science is constantly evolving, and our scientific knowledge is built in small incremental steps. Each individual study, performed by scientists around the world, some of which may not appear to have made a strong impact, but when taken as a whole, have made science what it is today and will likely contribute to what science is tomorrow.

Appendix A

Supplementary material for Chapter 2

A.1 List of detected heteroplasmic positions

Table A.1 Detected heteroplasmic positions. The following table lists the details of all the detected heteroplasmies analysed in this project. The heteroplasmic positions detected in files that were discarded (see section 2.4.2) are listed in Appendix A (see table A.2)

Individual	Position	Major	Minor	Ratio	Coverage	Annotation	Mitomap
Btau4	363	G	C	0.16	68	NC	NoDisease
Btau5	957	G	A	0.17	100	rRNA	NoDisease
Btau7	957	G	A	0.15	100	rRNA	NoDisease
Btau5	968	T	A	0.15	100	rRNA	NoDisease
Btau7	968	T	A	0.15	98	rRNA	NoDisease
Btau4	1068	G	A	0.15	100	rRNA	Disease
Btau5	1594	A	G	0.29	31	rRNA	NoDisease
Btau6	1657	T	C	0.15	99	rRNA	NoDisease
Btau5	2748	C	T	0.19	95	rRNA	NoDisease
Btau6	2748	C	T	0.25	95	rRNA	NoDisease
Btau7	2748	C	T	0.20	95	rRNA	NoDisease
Btau5	2755	C	T	0.19	96	rRNA	NoDisease
Btau6	2755	C	T	0.26	95	rRNA	NoDisease
Btau7	2755	C	T	0.17	96	rRNA	NoDisease
Btau5	2877	T	C	0.15	100	rRNA	NoDisease
Cfam5	7938	G	A	0.31	99	nSYN (G/S)	NoDisease
Cfam4	16148	G	A	0.20	35	NC	Liftover Fail
Cfam5	16158	A	G	0.18	57	NC	Liftover Fail
Cfam6	16168	A	G	0.43	40	NC	Liftover Fail
Cfam7	16168	A	G	0.23	44	NC	Liftover Fail
Cjac5	3692	C	T	0.17	46	nSYN (T/I)	Liftover Fail
Cjac5	3699	G	A	0.26	54	SYN	Liftover Fail
Cjac5	3732	A	G	0.31	52	SYN	Liftover Fail
Cjac5	6164	C	T	0.17	99	SYN	Liftover Fail
Cjac5	6191	T	C	0.39	82	SYN	Liftover Fail
Cjac5	6207	A	T	0.31	54	nSYN (A/S)	Liftover Fail
Cjac6	16293	C	T	0.33	21	NC	Liftover Fail
Cjac5	16478	C	A	0.42	36	NC	Liftover Fail
Cjac5	16479	T	C	0.22	32	NC	Liftover Fail
Cjac6	16479	C	T	0.29	21	NC	Liftover Fail
Csab1	4113	C	T	0.17	41	nSYN (L/F)	Liftover Fail
Csab2	4113	C	T	0.18	39	nSYN (L/F)	Liftover Fail

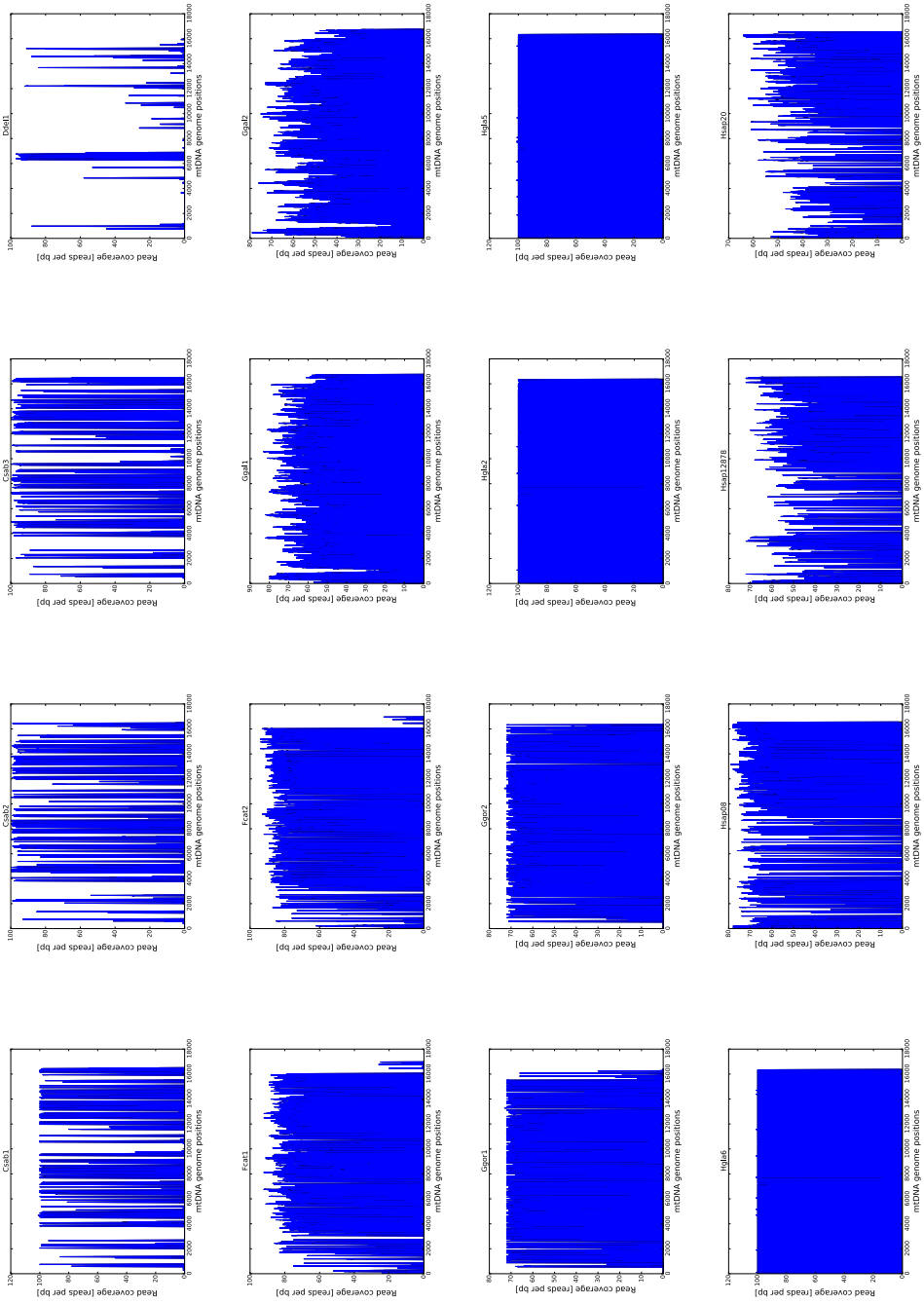
Csab3	4113	C	T	0.17	36	nSYN (L/F)	Liftover Fail
Csab2	16320	T	C	0.21	29	NC	Liftover Fail
Ggal2	3832	C	A	0.33	21	rRNA	Liftover Fail
Ggal1	3837	T	A	0.23	22	rRNA	Liftover Fail
Hsap23	72	T	C	0.35	83	NC	NoDisease
HsapC3	72	T	C	0.20	94	NC	NoDisease
Hsap99	94	G	A	0.21	72	NC	NoDisease
HsapC8	182	T	C	0.33	69	NC	NoDisease
HsapC8	195	C	T	0.34	67	NC	Disease
HsapC8	198	T	C	0.36	69	NC	NoDisease
HsapC8	204	T	C	0.35	78	NC	NoDisease
Hsap94	309	C	T	0.21	72	NC	Disease
Hsap23	309	C	T	0.17	76	NC	NoDisease
Hsap28	309	C	T	0.29	28	NC	NoDisease
Hsap99	309	C	T	0.23	47	NC	NoDisease
HsapC6	309	C	T	0.15	65	NC	NoDisease
Hsap23	310	T	C	0.28	64	NC	Disease
Hsap08	310	T	C	0.35	26	NC	NoDisease
Hsap94	310	T	C	0.31	65	NC	NoDisease
Hsap99	310	T	C	0.43	42	NC	NoDisease
HsapC3	310	T	C	0.32	59	NC	NoDisease
HsapC6	310	C	T	0.44	50	NC	NoDisease
HsapC7	310	T	C	0.29	59	NC	NoDisease
HsapC8	310	T	C	0.26	80	NC	NoDisease
HsapC8	515	A	G	0.15	33	NC	NoDisease
HsapC7	1018	A	G	0.19	26	rRNA	NoDisease
HsapC8	1018	A	G	0.19	27	rRNA	NoDisease
HsapC7	4315	A	T	0.23	31	tRNA	NoDisease
Hsap94	5054	A	G	0.22	59	SYN	NoDisease
HsapC8	8387	A	G	0.16	64	nSYN (V/M)	NoDisease
HsapC7	11944	C	T	0.26	77	SYN	NoDisease
HsapC8	12236	A	G	0.32	69	tRNA	Disease
Hsap94	12308	A	G	0.16	45	tRNA	NoDisease
Hsap94	14766	C	T	0.18	82	nSYN (T/I)	NoDisease
Hsap99	14869	A	G	0.17	29	SYN	NoDisease
HsapC7	16192	T	C	0.15	93	NC	Disease

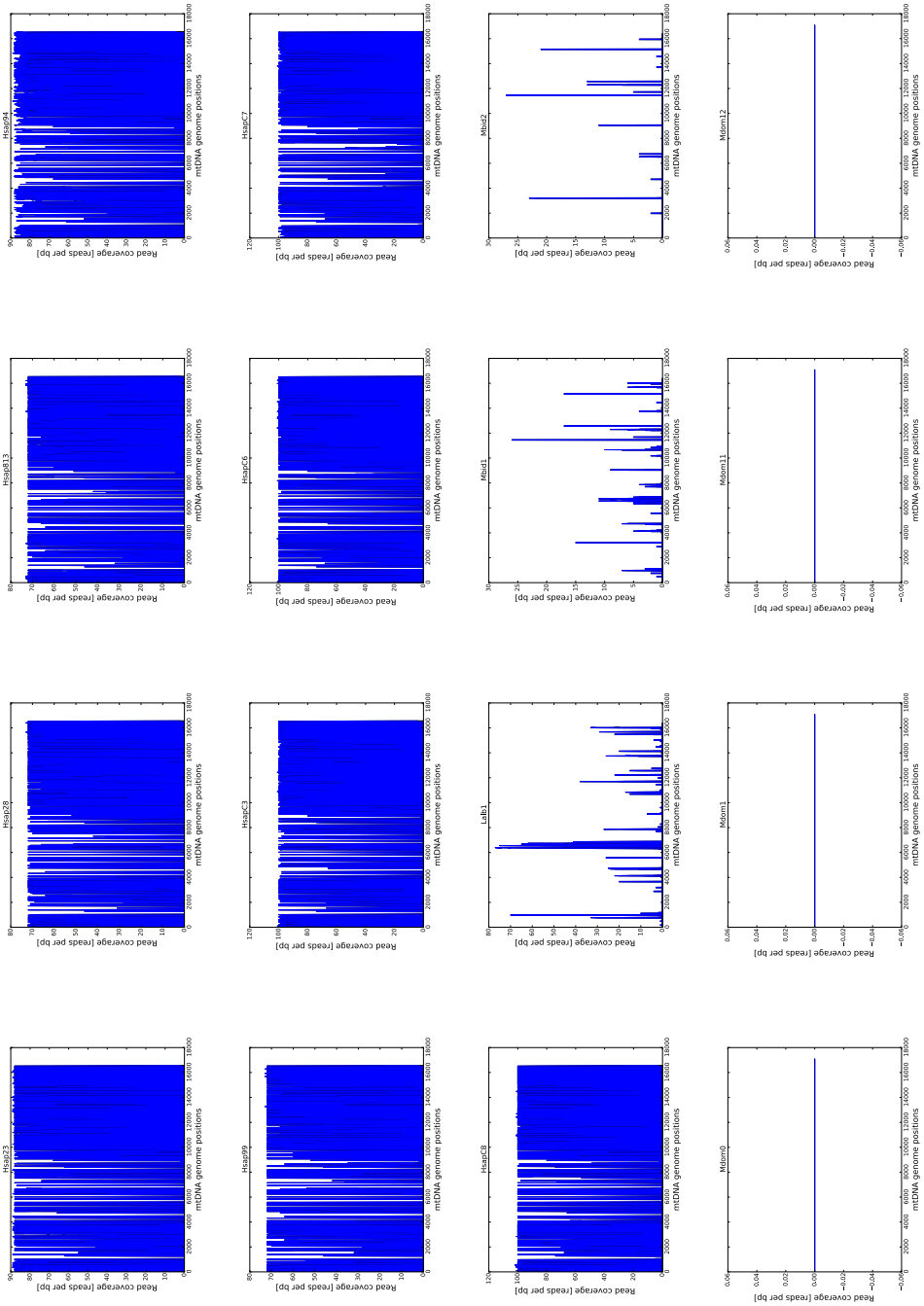
Hsap94	16519	C	T	0.19	85	NC	NoDisease
Mfur3	2326	T	C	0.18	97	rRNA	Liftover Fail
Mfur2	15525	C	T	0.36	91	NC	Liftover Fail
MmulLA	265	C	G	0.26	66	NC	Liftover Fail
MmulBO	8239	T	C	0.21	58	NC	Liftover Fail
MmulBO	15510	T	C	0.32	72	nSYN (I/T)	NoDisease
Mmul38	16182	A	G	0.19	47	NC	Liftover Fail
MmulJO	16182	G	A	0.40	57	NC	Liftover Fail
MmulJO	16354	T	C	0.19	27	NC	Liftover Fail
Mmus45	1364	T	C	0.16	25	rRNA	NoDisease
Mmus10	2360	C	A	0.24	34	rRNA	NoDisease
Mmus10	2525	C	T	0.20	45	rRNA	NoDisease
Mmus11	2525	C	T	0.18	33	rRNA	NoDisease
Mmus68	3072	G	A	0.15	26	nSYN (T/A)	NoDisease
Mmus10	3860	T	A	0.26	27	tRNA	NoDisease
Mmus72	12182	A	G	0.20	60	SYN	NoDisease
Mmus14	12862	T	C	0.23	26	nSYN (V/A)	NoDisease
Mmus72	15519	T	C	0.16	70	NC	NoDisease
Ocun5	1957	T	A	0.32	71	rRNA	NoDisease
Ocun5	3866	A	T	0.33	49	tRNA	NoDisease
Ocun5	3894	T	C	0.28	89	tRNA	NoDisease
Ocun5	3904	T	A	0.24	87	tRNA	NoDisease
Ocun4	8499	A	G	0.22	46	SYN	NoDisease
Ogar1	2358	A	T	0.16	70	rRNA	Liftover Fail
Rnor5	1394	G	A	0.17	72	rRNA	NoDisease
Rnor5	2448	C	T	0.19	75	rRNA	NoDisease
Rnor5	2542	A	G	0.18	74	rRNA	NoDisease
Rnor5	4992	A	C	0.19	67	tRNA	NoDisease
Rnor5	9294	T	C	0.17	69	SYN	NoDisease
Rnor5	11679	G	A	0.23	52	tRNA	NoDisease
Rnor5	11680	A	T	0.24	51	tRNA	NoDisease
Rnor5	11721	C	A	0.18	55	tRNA	NoDisease
Rnor9	11771	G	A	0.17	88	SYN	NoDisease
Rnor5	13882	C	T	0.16	49	SYN	Disease
Rnor5	15856	A	T	0.15	80	NC	Liftover Fail
Shar1	15635	T	A	0.36	25	NC	Liftover Fail

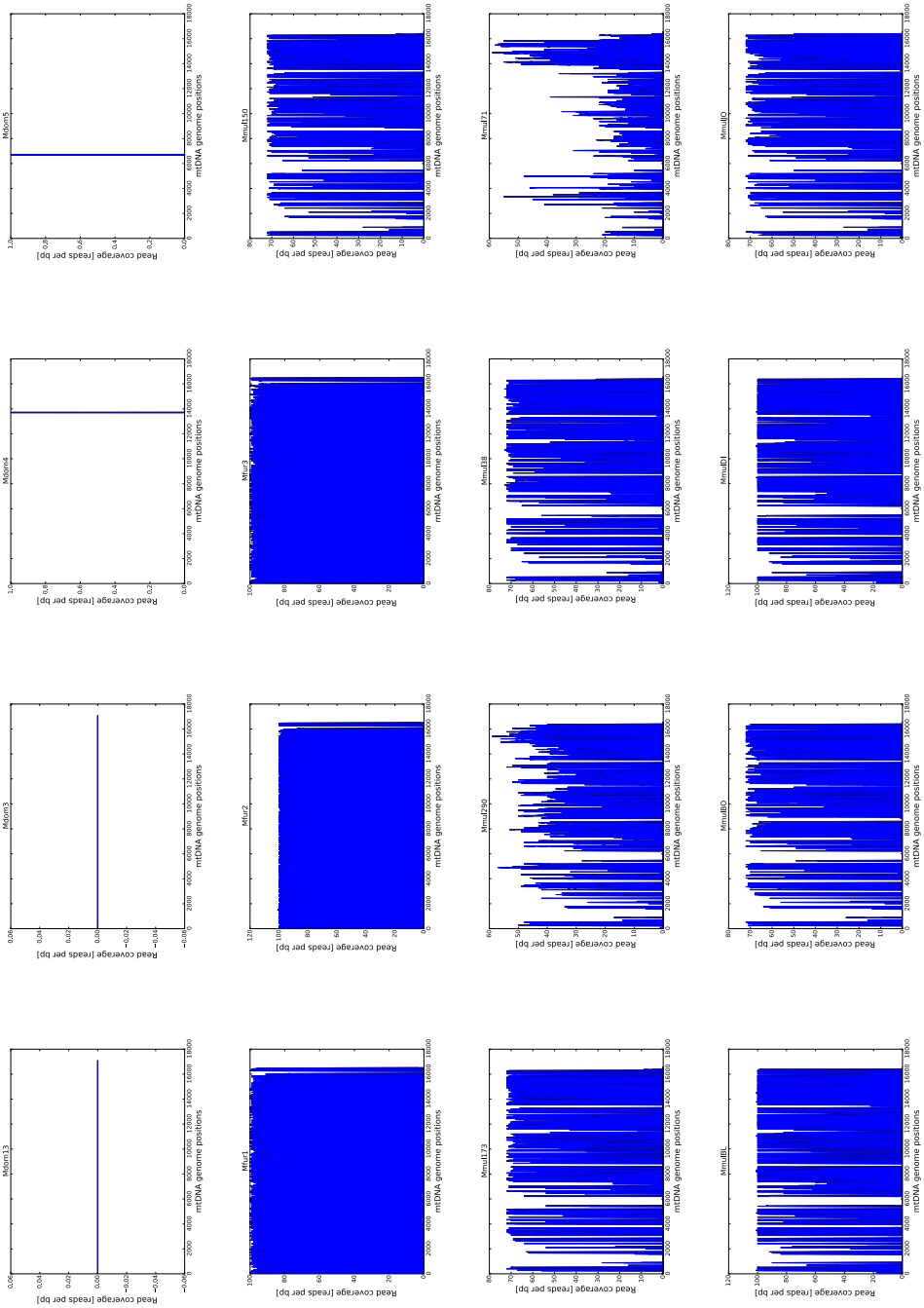
Shar2	15635	T	A	0.32	25	NC	Liftover Fail
Sscr5	1031	T	A	0.16	50	NC	NoDisease
Sscr6	15773	A	G	0.22	32	SYN	NoDisease

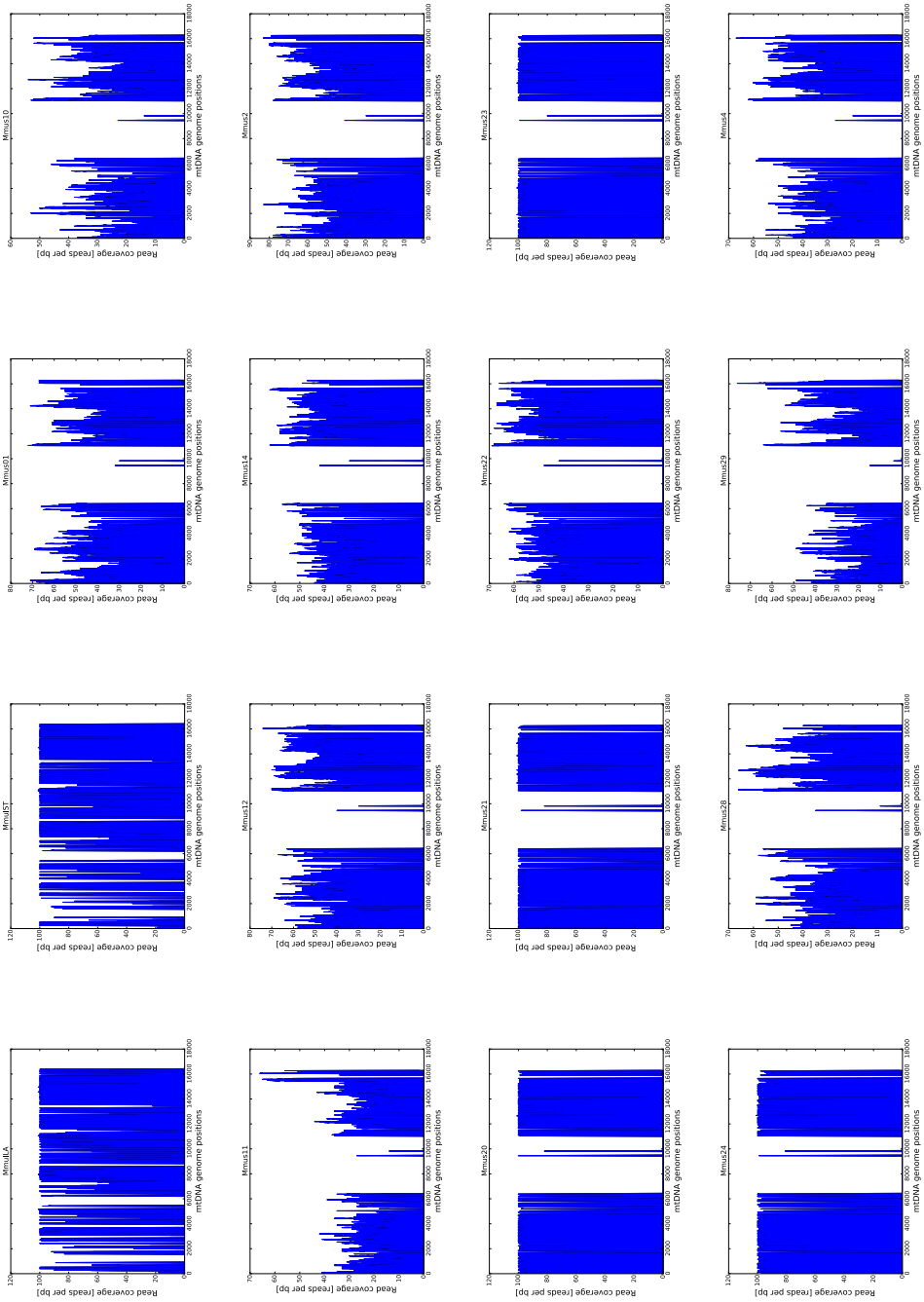
A.2 Detailed coverage data for each individual.

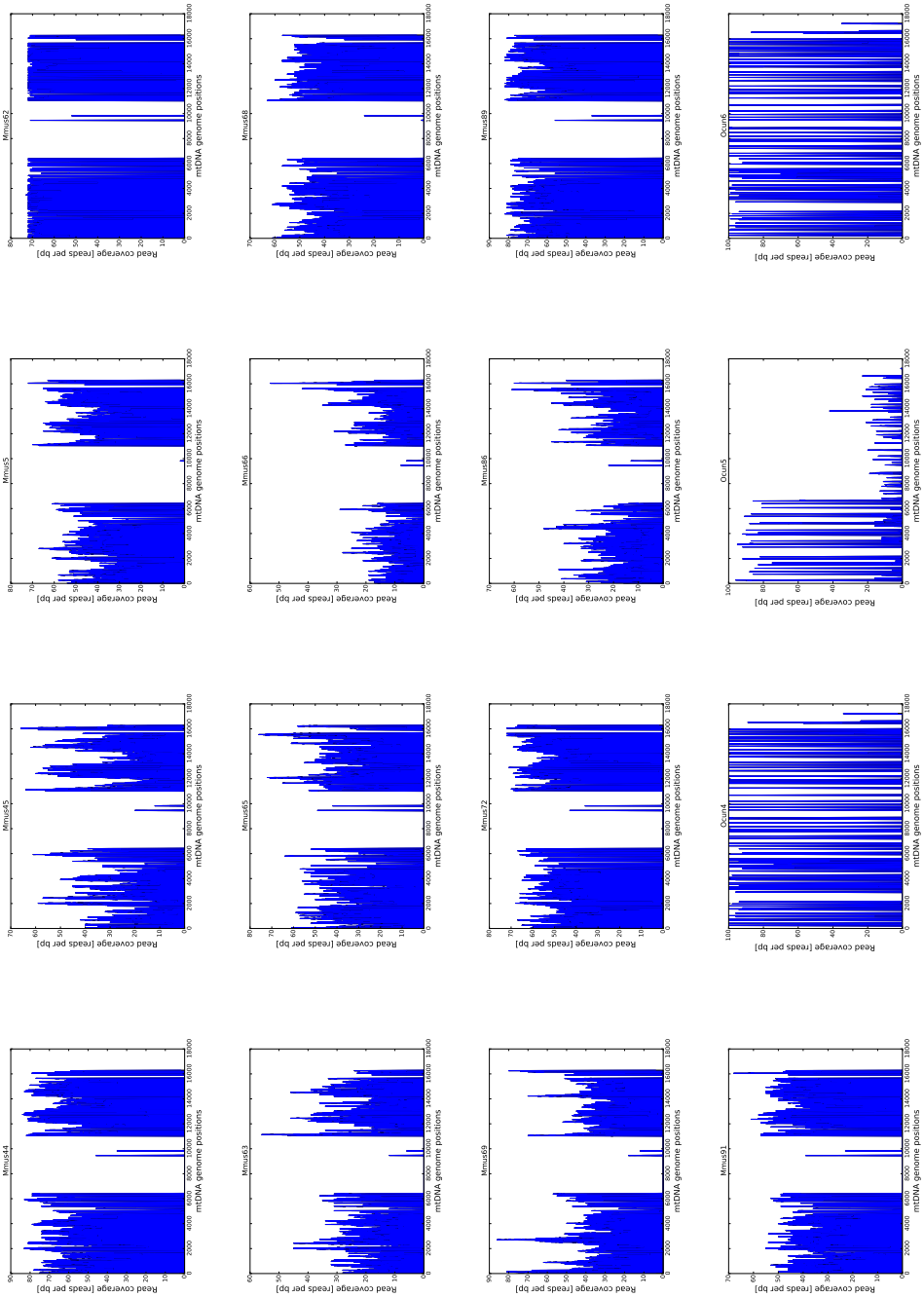
Fig. A.1 Detailed coverage data for all individuals. This set of figures displays the read coverage for each base of the mtDNA for all the studied individuals.

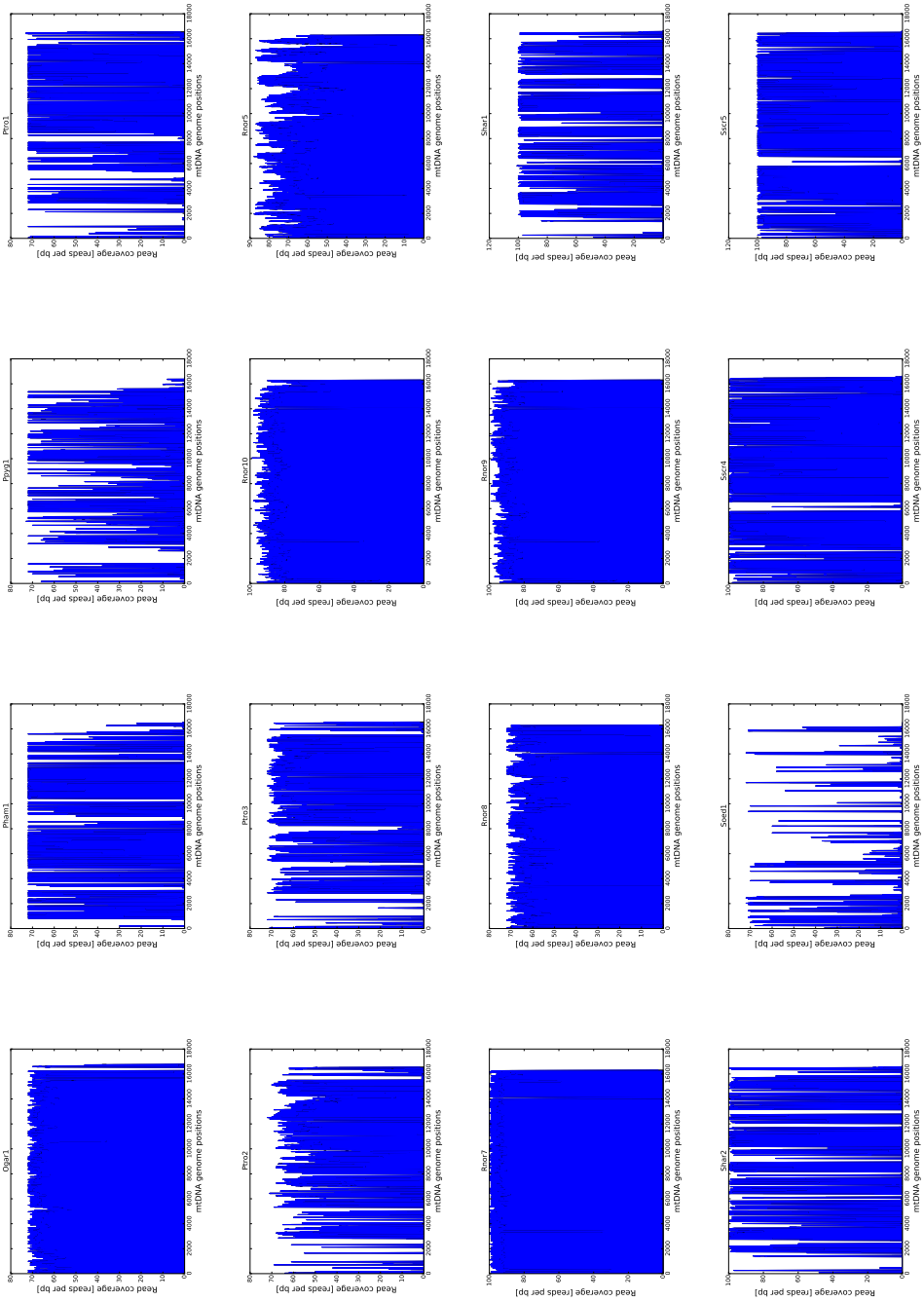

Ddel1
Cgal2
Hgla5
Hsap20
Csab3
Ggal1
Hgla2
Hsap12878
Csab2
Fcat2
Ggor2
Hsap08
Csab1
Fcat1
Ggor1
Hgla6
Read coverage [reads per bp]
mtDNA genome positions

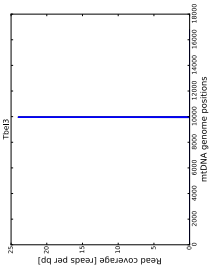

Hsap94
HsapC7
Msd2
Mdom12
Hsap813
HsapC6
Msd1
Mdom11
Hsap28
HsapC3
LaBc1
Mdom1
Hsap923
Hsap99
HsapC8
Mdom9
Read coverage [reads per bp]
mtDNA genome positions

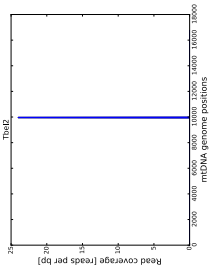

Mdom5
Mdom4
Mdom3
Mdom13
Mfur3
Mfur2
Mfur1
MmuH30
MmuJ71
MmuJ38
MmuJ290
MmuJ173
MmuQ0
MmuBJ
MmuBQ
MmuBL
Read coverage [reads per bp]
mtDNA genome positions

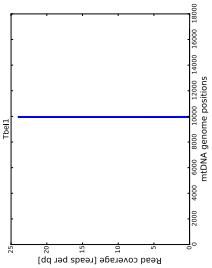

Mmus10
Mmus2
Mmus23
Mmus4
Mmus01
Mmus14
Mmus22
Mmus29
MmusST
Mmus12
Mmus21
Mmus28
MmusLA
Mmus11
Mmus20
Mmus24
Read coverage [reads per bp]
mtDNA genome positions

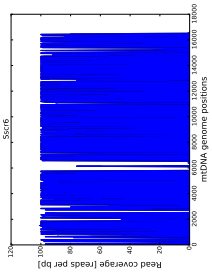

Mmus92
Mmus48
Mmus89
Qcun6
Mmus85
Mmus46
Mmus86
Qcun5
Mmus45
Mmus85
Mmus72
Qcun4
Mmus44
Mmus63
Mmus69
Mmus91
Read coverage [reads per bp]
mtDNA genome positions

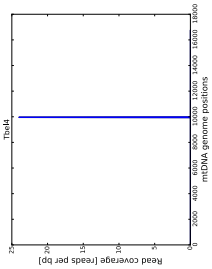

Pcro1
Rnor5
Shar1
Ssci5
Ppyg1
Rnor10
Rnor9
Spar4
Pkan1
Ptro3
Rnor8
Sced1
Ogar1
Ptro2
Rnor7
Slat2
Read coverage [reads per bp]
mtDNA genome positions


Tbet3
mtDNA genome positions
Read coverage [reads per bp]


Tbet2
mtDNA genome positions
Read coverage [reads per bp]


Tbet1
mtDNA genome positions
Read coverage [reads per bp]


Sscr6
mtDNA genome positions
Read coverage [reads per bp]


Tbet4
mtDNA genome positions
Read coverage [reads per bp]

A.3 Heteroplasmic positions detected in the discarded individuals

Table A.2 Heteroplasmies detected in the discarded individuals.

Individual	Position	Major	Minor	Ratio	Coverage
Hsap12878	114	C	A	0.20	64
Hsap12878	115	T	C	0.21	63
Hsap12878	572	C	A	0.22	32
Hsap12878	750	A	G	0.24	70
Hsap12878	1406	T	C	0.39	57
Hsap12878	1414	C	A	0.48	46
Hsap12878	2788	C	T	0.20	44
Hsap12878	2789	C	T	0.23	43
Hsap12878	3144	A	G	0.27	59
Hsap12878	3167	T	C	0.16	51
Hsap12878	3222	C	T	0.33	39
Hsap12878	3275	C	T	0.20	40
Hsap12878	3306	C	T	0.19	53
Hsap12878	3310	C	T	0.17	53
Hsap12878	3372	T	A	0.46	65
Hsap12878	3388	T	C	0.21	71
Hsap12878	3408	T	C	0.25	69
Hsap12878	3423	C	T	0.24	68
Hsap12878	3438	A	G	0.19	68
Hsap12878	3496	G	T	0.45	33
Hsap12878	3498	C	A	0.36	33
Hsap12878	3522	C	T	0.20	46
Hsap12878	3531	G	A	0.17	53
Hsap12878	3559	C	T	0.22	49
Hsap12878	3591	G	A	0.38	58
Hsap12878	3597	C	T	0.23	52
Hsap12878	3618	T	C	0.43	51
Hsap12878	3621	C	T	0.48	54
Hsap12878	3651	C	A	0.16	67
Hsap12878	3669	G	A	0.35	63
Hsap12878	3705	G	A	0.41	59
Hsap12878	3744	A	G	0.28	60
Hsap12878	3801	T	C	0.34	41
Hsap12878	3806	C	T	0.42	33

Hsap12878	5843	G	A	0.49	37
Hsap12878	5846	C	T	0.42	33
Hsap12878	9148	T	C	0.50	28
Hsap12878	9150	A	C	0.44	32
Hsap12878	9175	C	T	0.45	49
Hsap12878	9180	G	A	0.44	39
Hsap12878	9776	C	T	0.25	53
Hsap12878	9785	C	T	0.27	48
Hsap12878	9899	C	T	0.38	65
Hsap12878	9905	C	T	0.49	61
Hsap12878	9989	T	C	0.34	38
Hsap12878	10006	T	A	0.30	50
Hsap12878	10031	T	C	0.28	47
Hsap12878	10094	C	T	0.37	43
Hsap12878	10101	C	T	0.48	52
Hsap12878	10133	A	G	0.38	47
Hsap12878	10139	C	T	0.21	42
Hsap12878	10410	T	C	0.23	44
Hsap12878	10478	C	T	0.37	30
Hsap12878	10532	A	G	0.23	43
Hsap12878	10619	T	C	0.47	43
Hsap12878	10754	A	G	0.41	59
Hsap12878	10776	T	C	0.27	59
Hsap12878	10984	C	T	0.35	48
Hsap12878	11080	T	C	0.44	32
Hsap12878	11320	A	T	0.40	55
Hsap12878	11344	A	G	0.20	59
Hsap12878	12049	C	T	0.19	43
Hsap12878	12063	C	T	0.16	56
Hsap12878	12471	T	C	0.21	61
Hsap12878	12477	T	C	0.22	59
Hsap12878	13566	A	T	0.45	55
Hsap12878	13819	T	C	0.17	54
Hsap12878	13830	T	C	0.26	42
Hsap12878	13857	A	C	0.17	36
Hsap12878	13920	C	T	0.38	63

Hsap12878	13926	T	C	0.49	67
Hsap12878	13947	C	T	0.37	60
Hsap12878	14155	C	A	0.17	35
Hsap12878	14173	T	C	0.30	37
Hsap12878	14383	C	T	0.31	54
Hsap12878	14398	G	A	0.38	65
Hsap12878	14422	A	T	0.24	34
Hsap12878	14512	T	C	0.22	32
Hsap12878	14727	T	C	0.33	21
Hsap12878	14744	C	T	0.31	51
Hsap12878	14755	A	T	0.17	54
Hsap12878	15463	A	G	0.19	47
Hsap12878	15466	G	A	0.30	47
Hsap12878	15499	C	A	0.41	51
Hsap12878	15748	T	C	0.18	56
Hsap12878	15784	T	C	0.17	58
Hsap12878	15826	A	G	0.35	48
Hsap12878	15836	A	T	0.21	47
Hsap12878	16000	G	T	0.45	42
Hsap12878	16023	G	A	0.18	61
Hsap12878	16400	C	A	0.35	63
Hsap12878	16483	G	A	0.48	66
Hsap12878	16497	A	G	0.30	66
Hsap813	310	T	C	0.22	60
Hsap813	494	C	A	0.23	22
Hsap813	567	A	C	0.27	22
Hsap813	574	A	C	0.17	29
Hsap813	13093	A	G	0.47	58
Hsap813	15940	T	C	0.43	28
Mmul173	3561	G	A	0.18	28
Mmul290	12360	T	C	0.35	20
Mmul290	14998	C	T	0.18	51
Mmul290	16214	T	C	0.28	32
Mmul71	2721	G	A	0.31	32
Mmul71	3292	T	C	0.35	23
Mmul71	3315	C	T	0.32	22

Mmul71	3366	A	G	0.34	35
Mmul71	3372	T	C	0.50	30
Mmul71	3579	C	T	0.48	23
Mmul71	3654	T	C	0.32	22
Mmul71	4951	A	G	0.48	27
Mmul71	5002	A	T	0.21	33
Mmul71	5015	A	G	0.42	24
Mmul71	9977	G	A	0.18	22
Mmul71	10119	T	C	0.46	24
Mmul71	10135	T	C	0.28	29
Mmul71	13984	T	C	0.35	20
Mmul71	14118	G	A	0.33	48
Mmul71	14187	A	G	0.50	24
Mmul71	14191	C	T	0.40	20
Mmul71	14275	T	C	0.28	25
Mmul71	14377	T	C	0.35	26
Mmul71	14416	A	C	0.22	41
Mmul71	14524	T	C	0.16	31
Mmul71	14571	T	C	0.17	30
Mmul71	14587	G	A	0.22	37
Mmul71	14599	G	A	0.27	41
Mmul71	14649	T	C	0.45	29
Mmul71	14650	G	A	0.47	30
Mmul71	14668	C	T	0.45	31
Mmul71	14692	A	G	0.26	42
Mmul71	14721	T	C	0.48	29
Mmul71	14771	A	C	0.31	29
Mmul71	14797	T	C	0.21	42
Mmul71	14807	A	G	0.26	43
Mmul71	14830	A	G	0.27	49
Mmul71	14876	G	A	0.18	57
Mmul71	14938	T	C	0.50	26
Mmul71	14944	T	C	0.48	29
Mmul71	14959	G	A	0.37	35
Mmul71	14962	A	G	0.37	35
Mmul71	15004	C	T	0.41	29

Mmul71	15013	C	T	0.24	21
Mmul71	15037	C	T	0.46	24
Mmul71	15040	A	G	0.48	25
Mmul71	15118	G	A	0.38	21
Mmul71	15154	G	A	0.41	32
Mmul71	15166	A	G	0.30	20
Mmul71	15385	G	A	0.26	54
Mmul71	15417	C	T	0.48	23
Mmul71	15451	T	C	0.34	35
Mmul71	15511	C	T	0.47	34
Mmul71	15547	G	A	0.41	37
Mmul71	15553	A	G	0.37	43
Mmul71	15614	T	C	0.45	31
Mmul71	15661	T	C	0.32	28
Mmul71	15664	C	T	0.38	29
Mmul71	15709	T	C	0.20	20
Mmul71	15712	A	G	0.16	25
Mmul71	15805	T	C	0.33	36
Mmul71	15844	A	G	0.30	43
Mmul71	15865	C	G	0.39	23
Fcat1	63	A	T	0.46	24
Fcat1	173	G	A	0.16	44
Fcat1	570	A	G	0.43	23
Fcat1	637	A	T	0.41	54
Fcat1	1442	A	C	0.24	67
Fcat1	1472	G	A	0.29	31
Fcat1	1584	A	T	0.43	30
Fcat1	1708	C	T	0.32	71
Fcat1	2003	C	T	0.32	77
Fcat1	3153	T	C	0.47	47
Fcat1	3181	C	T	0.16	64
Fcat1	3327	C	T	0.21	77
Fcat1	3330	A	T	0.24	78
Fcat1	3383	T	C	0.29	38
Fcat1	3737	G	A	0.33	70
Fcat1	3896	G	A	0.22	65

Fcat1	3941	T	A	0.42	79
Fcat1	4128	A	G	0.25	75
Fcat1	4232	A	G	0.35	40
Fcat1	4508	A	G	0.34	83
Fcat1	4727	A	T	0.22	55
Fcat1	5125	A	G	0.18	51
Fcat1	5149	T	C	0.27	44
Fcat1	5162	T	C	0.36	53
Fcat1	5461	A	G	0.33	75
Fcat1	5644	C	T	0.32	82
Fcat1	5798	G	A	0.20	66
Fcat1	6175	C	G	0.26	57
Fcat1	6224	A	T	0.18	57
Fcat1	6549	T	C	0.38	69
Fcat1	6581	G	A	0.23	69
Fcat1	6656	T	C	0.21	82
Fcat1	6843	T	C	0.24	74
Fcat1	6998	T	C	0.28	25
Fcat1	7424	T	C	0.42	76
Fcat1	7919	G	A	0.27	77
Fcat1	8162	A	G	0.35	68
Fcat1	8240	C	T	0.33	64
Fcat1	8525	C	T	0.31	72
Fcat1	9110	C	T	0.29	82
Fcat1	9211	G	A	0.31	85
Fcat1	9367	C	T	0.35	68
Fcat1	9852	A	G	0.38	71
Fcat1	9978	T	C	0.16	57
Fcat1	10008	A	G	0.41	54
Fcat1	10026	T	C	0.42	57
Fcat1	10406	T	C	0.31	61
Fcat1	10546	C	T	0.36	69
Fcat1	10609	T	C	0.40	70
Fcat1	10851	T	C	0.31	83
Fcat1	11326	A	G	0.20	82
Fcat1	11356	T	C	0.19	74

Fcat1	11480	G	A	0.25	79
Fcat1	11633	C	T	0.25	77
Fcat1	11814	A	G	0.27	78
Fcat1	11921	T	C	0.21	72
Fcat1	12014	C	T	0.37	81
Fcat1	12158	A	G	0.38	78
Fcat1	12197	T	C	0.25	75
Fcat1	12359	T	C	0.29	80
Fcat1	12645	T	C	0.34	79
Fcat1	13032	T	C	0.30	74
Fcat1	13248	G	A	0.29	77
Fcat1	13285	C	T	0.30	79
Fcat1	13590	C	T	0.29	79
Fcat1	14186	A	G	0.35	71
Fcat1	15097	C	T	0.33	81
Fcat1	15104	G	A	0.27	79
Fcat1	15331	G	A	0.25	85
Fcat1	15343	A	G	0.22	85
Fcat1	15478	T	C	0.29	75
Fcat1	15589	C	T	0.36	67
Fcat1	15756	T	C	0.37	76
Fcat1	15904	A	G	0.30	79
Fcat1	16018	C	T	0.32	77
Fcat2	169	C	T	0.20	45
Fcat2	173	G	A	0.29	48
Fcat2	570	G	A	0.47	32
Fcat2	637	A	T	0.34	53
Fcat2	1584	A	T	0.40	25
Fcat2	1708	C	T	0.35	71
Fcat2	2003	C	T	0.43	77
Fcat2	3153	T	C	0.44	52
Fcat2	3181	C	T	0.16	64
Fcat2	3330	A	T	0.15	79
Fcat2	3383	T	C	0.23	44
Fcat2	3737	G	A	0.39	83
Fcat2	3896	G	A	0.25	69

Fcat2	3941	T	A	0.33	88
Fcat2	4128	A	G	0.39	80
Fcat2	4232	A	G	0.46	41
Fcat2	4508	A	G	0.44	85
Fcat2	4727	A	T	0.20	55
Fcat2	5162	T	C	0.26	53
Fcat2	5461	A	G	0.36	84
Fcat2	5644	C	T	0.33	83
Fcat2	5798	G	A	0.27	59
Fcat2	6175	C	G	0.33	58
Fcat2	6224	A	T	0.17	58
Fcat2	6549	T	C	0.42	76
Fcat2	6581	G	A	0.33	78
Fcat2	6656	T	C	0.45	73
Fcat2	6843	T	C	0.46	81
Fcat2	6998	T	C	0.32	31
Fcat2	7424	T	C	0.47	79
Fcat2	7919	A	G	0.47	81
Fcat2	8162	A	G	0.39	67
Fcat2	8240	C	T	0.33	66
Fcat2	8525	C	T	0.30	73
Fcat2	8906	T	C	0.20	76
Fcat2	9110	C	T	0.42	84
Fcat2	9211	G	A	0.28	79
Fcat2	9367	C	T	0.30	74
Fcat2	9852	A	G	0.47	72
Fcat2	9978	T	C	0.16	67
Fcat2	10008	G	A	0.40	65
Fcat2	10026	C	T	0.45	65
Fcat2	10406	C	T	0.46	72
Fcat2	10546	C	T	0.39	75
Fcat2	10609	T	C	0.32	78
Fcat2	10851	T	C	0.45	80
Fcat2	11326	G	A	0.45	86
Fcat2	11356	T	C	0.45	82
Fcat2	11480	G	A	0.33	86

Fcat2	11633	T	C	0.44	81
Fcat2	11814	A	G	0.38	85
Fcat2	11921	T	C	0.39	83
Fcat2	12014	C	T	0.33	79
Fcat2	12158	A	G	0.36	80
Fcat2	12197	T	C	0.27	85
Fcat2	12359	T	C	0.41	83
Fcat2	12645	C	T	0.49	79
Fcat2	13032	T	C	0.32	84
Fcat2	13248	G	A	0.42	78
Fcat2	13285	C	T	0.38	80
Fcat2	13590	C	T	0.36	83
Fcat2	14186	A	G	0.35	83
Fcat2	15097	C	T	0.41	74
Fcat2	15104	G	A	0.38	81
Fcat2	15331	G	A	0.36	85
Fcat2	15343	A	G	0.43	82
Fcat2	15478	T	C	0.43	84
Fcat2	15589	C	T	0.48	80
Fcat2	15756	T	C	0.26	86
Fcat2	15904	A	G	0.34	85
Fcat2	16018	C	T	0.37	90
Ggor1	858	A	G	0.33	42
Ggor1	962	C	T	0.18	66
Ggor1	1074	T	C	0.48	62
Ggor1	1093	A	T	0.36	70
Ggor1	1397	C	T	0.32	66
Ggor1	1398	C	A	0.31	68
Ggor1	1582	T	C	0.16	70
Ggor1	1700	T	C	0.25	71
Ggor1	1714	G	A	0.41	70
Ggor1	1741	C	A	0.24	71
Ggor1	2088	G	A	0.24	67
Ggor1	2447	T	C	0.37	46
Ggor1	2694	C	T	0.32	41
Ggor1	2696	G	A	0.38	40

Ggor1	2807	T	C	0.25	71
Ggor1	2857	A	G	0.21	71
Ggor1	2941	T	C	0.15	72
Ggor1	2950	A	G	0.15	72
Ggor1	3070	C	A	0.15	71
Ggor1	3764	T	C	0.37	60
Ggor1	3804	G	A	0.26	70
Ggor1	3894	C	T	0.33	72
Ggor1	3910	G	A	0.48	71
Ggor1	3930	C	T	0.19	72
Ggor1	3973	G	A	0.18	71
Ggor1	3999	A	G	0.23	71
Ggor1	5023	T	C	0.19	64
Ggor1	5081	T	C	0.29	69
Ggor1	5092	C	T	0.23	69
Ggor1	5232	T	A	0.29	69
Ggor1	5382	T	C	0.28	71
Ggor1	5386	C	T	0.35	71
Ggor1	5421	C	T	0.20	70
Ggor1	5847	C	T	0.26	57
Ggor1	5868	A	G	0.22	64
Ggor1	6093	C	T	0.25	68
Ggor1	6105	T	C	0.29	69
Ggor1	6234	C	A	0.33	36
Ggor1	6237	C	T	0.38	42
Ggor1	6288	C	T	0.26	72
Ggor1	6504	T	C	0.21	71
Ggor1	6537	T	C	0.26	72
Ggor1	6543	T	C	0.18	72
Ggor1	6627	T	C	0.15	71
Ggor1	6633	A	G	0.26	70
Ggor1	6666	T	C	0.17	71
Ggor1	6920	A	G	0.21	71
Ggor1	6924	T	C	0.17	71
Ggor1	7963	T	C	0.35	37
Ggor1	7965	A	G	0.43	37

Ggor1	7999	T	C	0.23	56
Ggor1	8599	T	C	0.22	63
Ggor1	8604	G	A	0.26	54
Ggor1	9236	T	C	0.19	69
Ggor1	9323	C	T	0.33	58
Ggor1	9329	C	T	0.23	56
Ggor1	9430	T	A	0.24	71
Ggor1	9902	T	C	0.24	59
Ggor1	9914	C	T	0.31	59
Ggor1	10043	T	C	0.35	72
Ggor1	10178	G	A	0.15	72
Ggor1	10200	C	T	0.20	70
Ggor1	10246	T	C	0.28	72
Ggor1	10369	G	A	0.25	67
Ggor1	10489	C	A	0.16	67
Ggor1	10504	C	T	0.21	61
Ggor1	10537	C	T	0.31	67
Ggor1	10553	G	A	0.20	69
Ggor1	10768	G	A	0.24	72
Ggor1	11156	C	T	0.17	72
Ggor1	11159	T	C	0.18	72
Ggor1	11266	G	A	0.24	71
Ggor1	11278	C	T	0.33	72
Ggor1	11726	A	C	0.37	70
Ggor1	12328	C	A	0.18	72
Ggor1	12465	G	A	0.35	65
Ggor1	12486	C	A	0.24	59
Ggor1	13041	C	T	0.17	69
Ggor1	13053	G	A	0.20	69
Ggor1	13059	C	T	0.20	70
Ggor1	13281	C	A	0.17	72
Ggor1	13822	G	A	0.17	72
Ggor1	13846	A	T	0.24	70
Ggor1	13855	C	T	0.22	64
Ggor1	13912	C	T	0.18	65
Ggor1	13936	C	T	0.23	70

Ggor1	14038	G	A	0.25	71
Ggor1	14572	G	A	0.31	36
Ggor1	14587	C	A	0.27	45
Ggor1	14596	A	G	0.47	53
Ggor1	14608	C	T	0.20	65
Ggor1	14617	G	A	0.17	72
Ggor1	15250	G	A	0.24	72
Ggor2	2386	C	T	0.41	59
Ggor2	15751	A	C	0.16	45
Lalb1	16031	T	A	0.18	22
Pham1	1607	T	C	0.27	30
Pham1	16248	G	C	0.26	23
Ppyg1	4815	G	A	0.39	28
Ppyg1	5168	G	T	0.16	50
Ppyg1	5169	A	T	0.32	50
Ppyg1	5590	G	T	0.42	31
Ppyg1	8474	C	T	0.36	28
Ppyg1	8489	G	A	0.39	28
Ppyg1	8646	T	C	0.34	35
Ppyg1	8648	A	G	0.42	31
Ppyg1	10572	C	T	0.20	35
Ppyg1	11540	T	C	0.48	25
Ppyg1	11618	T	C	0.27	26
Ppyg1	12489	A	C	0.17	30
Ppyg1	12504	G	A	0.36	39
Ppyg1	12510	A	C	0.26	34
Ppyg1	12516	C	A	0.15	26
Ppyg1	13864	C	G	0.19	37
Ppyg1	13956	T	C	0.49	35
Ppyg1	13966	C	T	0.48	25
Ppyg1	13971	A	G	0.30	23
Ppyg1	13975	C	T	0.17	29
Ppyg1	14263	T	C	0.21	38
Ppyg1	14668	C	T	0.38	21
Ppyg1	14680	G	A	0.32	25
Ppyg1	14689	G	A	0.38	32

Ppyg1	14698	G	A	0.21	38
Ppyg1	14866	C	T	0.32	22
Ppyg1	14872	T	C	0.16	31
Ptro1	137	C	G	0.35	23
Ptro1	377	A	C	0.22	36
Ptro1	9578	T	C	0.18	28
Ptro1	9580	A	C	0.19	27
Ptro1	11708	G	A	0.48	23
Ptro1	12477	T	C	0.45	22
Ptro1	12918	C	T	0.20	45
Ptro1	16433	C	T	0.32	22
Ptro2	14356	A	G	0.21	33
Ptro3	3927	T	C	0.19	54
Ptro3	6820	G	A	0.18	34
Ptro3	8956	C	A	0.18	61
Ptro3	9742	T	C	0.30	33
Ptro3	13443	T	C	0.16	70
Ptro3	13970	A	G	0.17	60
Ptro3	14917	T	C	0.48	33
Ptro3	15926	C	T	0.21	42

A.4 Experimental details for each file analysed in the context of this project.

Fig. A.2 Experimental details for each processed and analysed file. The ID column contains the (lab specific) sequencing ID, the status column states if this file was used within the study or discarded, and the Genome column contains the reference genomes to which the files were aligned.

Novel ChIP-seq data
Redundant files (present in more than one submission)

Individual	Species	Tissue	ID	Protein	Accession	Status	Genome
Bbor1	B. borealis	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Bbor1	B. borealis	liver	merged	input	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Bbor1	B. borealis	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Bbor1	B. borealis	liver	SAN01	H3K4me1	E-MTAB-3933	no	Tursiops_truncatus.turTru1
Btau4	B. taurus	liver	SAN01	input	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau4	B. taurus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau4	B. taurus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau4	B. taurus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Bos_taurus.UMD3.1
Btau5	B. taurus	liver	SAN01	input	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau5	B. taurus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau5	B. taurus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau5	B. taurus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Bos_taurus.UMD3.1
Btau6	B. taurus	liver	SAN01	input	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau6	B. taurus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau6	B. taurus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau6	B. taurus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Bos_taurus.UMD3.1
Btau7	B. taurus	liver	SAN01	input	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau7	B. taurus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau7	B. taurus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Bos_taurus.UMD3.1
Btau7	B. taurus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Bos_taurus.UMD3.1
Cfam3	C. familiaris	liver	CRI01	input DNA	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI01	FOXA1	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI01	HNF4A	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	SAN02	HNF4A	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	SAN03	HNF4A	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI02	HNF6	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	SAN03	HNF6	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	SAN04	HNF6	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI01	CEBPA	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI02	CEBPA	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI03	CEBPA	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI01	input_DNA	E-MTAB-437	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI01	CTCF	E-MTAB-437	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI01	input_DNA	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI01	HNF4a	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	SAN02	HNF4a	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	SAN03	HNF4a	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI01	CEBPA	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI02	CEBPA	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam3	C. familiaris	liver	CRI03	CEBPA	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI01	input DNA	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI01	input DNA	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI01	FOXA1	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI02	FOXA1	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI01	CEBPA	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI02	CEBPA	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI03	CEBPA	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	SAN01	HNF4A	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	SAN02	HNF4A	E-MTAB-1509	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI01	input_DNA	E-MTAB-437	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI01	input_DNA	E-MTAB-437	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI01	CTCF	E-MTAB-437	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI01	input_DNA	E-TABM-722	yes	Canis_familiaris.BROADD2.65

A.4 Experimental details for each file analysed in the context of this project.

Cfam4	C. familiaris	liver	CRI01	input_DNA	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI01	CEBPA	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI02	CEBPA	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	CRI03	CEBPA	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	SAN01	HNF4a	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam4	C. familiaris	liver	SAN02	HNF4a	E-TABM-722	yes	Canis_familiaris.BROADD2.65
Cfam5	C. familiaris	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Canis_familiaris.BROADD2.65
Cfam5	C. familiaris	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Canis_familiaris.BROADD2.65
Cfam5	C. familiaris	liver	SAN01	input	E-MTAB- 2633	yes	Canis_familiaris.BROADD2.65
Cfam5	C. familiaris	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Canis_familiaris.BROADD2.65
Cfam5	C. familiaris	liver	CRI01	CEBPA	E-MTAB-3933	yes	Canis_familiaris.BROADD2.65
Cfam6	C. familiaris	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Canis_familiaris.BROADD2.65
Cfam6	C. familiaris	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Canis_familiaris.BROADD2.65
Cfam6	C. familiaris	liver	SAN01	input	E-MTAB- 2633	yes	Canis_familiaris.BROADD2.65
Cfam6	C. familiaris	liver	CRI01	CEBPA	E-MTAB-3933	yes	Canis_familiaris.BROADD2.65
Cfam6	C. familiaris	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Canis_familiaris.BROADD2.65
Cfam7	C. familiaris	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Canis_familiaris.BROADD2.65
Cfam7	C. familiaris	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Canis_familiaris.BROADD2.65
Cfam7	C. familiaris	liver	SAN01	input	E-MTAB- 2633	yes	Canis_familiaris.BROADD2.65
Cfam7	C. familiaris	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Canis_familiaris.BROADD2.65
Cfam7	C. familiaris	liver	CRI01		E-MTAB-3933	yes	Canis_familiaris.BROADD2.65
Cjac4	C. jacchus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac4	C. jacchus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac4	C. jacchus	liver	SAN01	input	E-MTAB- 2633	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac4	C. jacchus	liver	CRI01	CEBPA	E-MTAB-3933	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac4	C. jacchus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac5	C. jacchus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac5	C. jacchus	liver	SAN01	input	E-MTAB- 2633	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac5	C. jacchus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac5	C. jacchus	liver	SAN01		E-MTAB-3933	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac5	C. jacchus	liver	CRI01	CEBPA	E-MTAB-3933	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac5	C. jacchus	liver	SAN01		E-MTAB-3933	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac6	C. jacchus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac6	C. jacchus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac6	C. jacchus	liver	SAN01	input	E-MTAB- 2633	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac6	C. jacchus	liver	CRI01	CEBPA	E-MTAB-3933	yes	Callithrix_jacchus.C_jacchus3.2.1
Cjac6	C. jacchus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Callithrix_jacchus.C_jacchus3.2.1
Cpor7	C. porcellus	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Cavia_porcellus.cavPor3
Cpor7	C. porcellus	liver	merged	input	E-MTAB- 2633	no	Cavia_porcellus.cavPor3
Cpor7	C. porcellus	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Cavia_porcellus.cavPor3
Cpor7	C. porcellus	liver	CRI01	CEBPA	E-MTAB-3933	no	Cavia_porcellus.cavPor3
Cpor7	C. porcellus	liver	SAN01	H3K4me1	E-MTAB-3933	no	Cavia_porcellus.cavPor3
Cpor8	C. porcellus	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Cavia_porcellus.cavPor3
Cpor8	C. porcellus	liver	merged	input	E-MTAB- 2633	no	Cavia_porcellus.cavPor3
Cpor8	C. porcellus	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Cavia_porcellus.cavPor3
Cpor8	C. porcellus	liver	CRI01	CEBPA	E-MTAB-3933	no	Cavia_porcellus.cavPor3
Cpor8	C. porcellus	liver	SAN01	H3K4me1	E-MTAB-3933	no	Cavia_porcellus.cavPor3
Cpor9	C. porcellus	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Cavia_porcellus.cavPor3
Cpor9	C. porcellus	liver	merged	input	E-MTAB- 2633	no	Cavia_porcellus.cavPor3
Cpor9	C. porcellus	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Cavia_porcellus.cavPor3
Cpor9	C. porcellus	liver	CRI01		E-MTAB-3933	no	Cavia_porcellus.cavPor3
Cpor9	C. porcellus	liver	SAN01	H3K4me1	E-MTAB-3933	no	Cavia_porcellus.cavPor3
Csab1	C. sabaeus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Chlorocebus_sabaeus.ChlSab1.1
Csab1	C. sabaeus	liver	SAN01	input	E-MTAB- 2633	yes	Chlorocebus_sabaeus.ChlSab1.1
Csab1	C. sabaeus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Chlorocebus_sabaeus.ChlSab1.1
Csab1	C. sabaeus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Chlorocebus_sabaeus.ChlSab1.1
Csab2	C. sabaeus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Chlorocebus_sabaeus.ChlSab1.1
Csab2	C. sabaeus	liver	SAN01	input	E-MTAB- 2633	yes	Chlorocebus_sabaeus.ChlSab1.1

A.4 Experimental details for each file analysed in the context of this project.

Csab2	C. sabaeus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Chlorocebus_sabaeus.ChlSab1.1
Csab2	C. sabaeus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Chlorocebus_sabaeus.ChlSab1.1
Csab3	C. sabaeus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Chlorocebus_sabaeus.ChlSab1.1
Csab3	C. sabaeus	liver	SAN01	input	E-MTAB- 2633	yes	Chlorocebus_sabaeus.ChlSab1.1
Csab3	C. sabaeus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Chlorocebus_sabaeus.ChlSab1.1
Csab3	C. sabaeus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Chlorocebus_sabaeus.ChlSab1.1
Ddel1	D. delphis	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Tursiops_truncatus.turTru1
Ddel1	D. delphis	liver	SAN01	input	E-MTAB- 2633	yes	Tursiops_truncatus.turTru1
Ddel1	D. delphis	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Tursiops_truncatus.turTru1
Ddel1	D. delphis	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Tursiops_truncatus.turTru1
Ddel1	D. delphis	liver	SAN01	totalH3	E-MTAB-3933	yes	Tursiops_truncatus.turTru1
Fcat1	F. catus	liver	SAN01	input	E-MTAB- 2633	no	Felis_catus.Felis_catus_6.2
Fcat1	F. catus	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Felis_catus.Felis_catus_6.2
Fcat1	F. catus	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Felis_catus.Felis_catus_6.2
Fcat1	F. catus	liver	SAN01	H3K4me1	E-MTAB-3933	no	Felis_catus.Felis_catus_6.2
Fcat1	F. catus	liver	CRI01	CEBPA	E-MTAB-3933	no	Felis_catus.Felis_catus_6.2
Fcat2	F. catus	liver	SAN01	input	E-MTAB- 2633	no	Felis_catus.Felis_catus_6.2
Fcat2	F. catus	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Felis_catus.Felis_catus_6.2
Fcat2	F. catus	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Felis_catus.Felis_catus_6.2
Fcat2	F. catus	liver	SAN01	H3K4me1	E-MTAB-3933	no	Felis_catus.Felis_catus_6.2
Fcat2	F. catus	liver	CRI01	CEBPA	E-MTAB-3933	no	Felis_catus.Felis_catus_6.2
Ggal	G. gallus	liver	CRI01	input_DNA	E-TABM-722	yes	Gallus_gallus.Galgal4
Ggal	G. gallus	liver	CRI01	CEBPA	E-TABM-722	yes	Gallus_gallus.Galgal4
Ggal	G. gallus	liver	CRI02	CEBPA	E-TABM-722	yes	Gallus_gallus.Galgal4
Ggal	G. gallus	liver	CRI03	CEBPA	E-TABM-722	yes	Gallus_gallus.Galgal4
Ggal2	G. gallus	liver	CRI01	CEBPA	E-TABM-722	yes	Gallus_gallus.Galgal4
Ggal2	G. gallus	liver	CRI01	input_DNA	E-TABM-722	yes	Gallus_gallus.Galgal4
Ggor1	G. gorilla	LCL	CRI01	CTCF	E-MTAB-1511	no	Gorilla_gorilla.gorGor3.1.66
Ggor1	G. gorilla	LCL	CRI02	CTCF	E-MTAB-1511	no	Gorilla_gorilla.gorGor3.1.66
Ggor1	G. gorilla	LCL	CRI02	CTCF	E-MTAB-1511	no	Gorilla_gorilla.gorGor3.1.66
Ggor1	G. gorilla	LCL	CRI03	NA	E-MTAB-1511	no	Gorilla_gorilla.gorGor3.1.66
Ggor2	G. gorilla	LCL	CRI01	CTCF	E-MTAB-1511	no	Gorilla_gorilla.gorGor3.1.66
Ggor2	G. gorilla	LCL	CRI01	NA	E-MTAB-1511	no	Gorilla_gorilla.gorGor3.1.66
Hgla2	H. glaber	liver	merged	H3K4me3	E-MTAB- 2633	yes	H_glaber_mtDNA(gi322422826)
Hgla2	H. glaber	liver	SAN01	input	E-MTAB- 2633	yes	H_glaber_mtDNA(gi322422826)
Hgla2	H. glaber	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	H_glaber_mtDNA(gi322422826)
Hgla2	H. glaber	liver	CRI01	CEBPA	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla2	H. glaber	liver	merged	H3K4me1	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla2	H. glaber	liver	merged	H3K27Ac	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla2	H. glaber	liver	merged	totalH3	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla5	H. glaber	liver	merged	H3K4me3	E-MTAB- 2633	yes	H_glaber_mtDNA(gi322422826)
Hgla5	H. glaber	liver	SAN01	input	E-MTAB- 2633	yes	H_glaber_mtDNA(gi322422826)
Hgla5	H. glaber	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	H_glaber_mtDNA(gi322422826)
Hgla5	H. glaber	liver	merged	H3K4me1	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla5	H. glaber	liver	merged	H3K27Ac	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla5	H. glaber	liver	merged	totalH3	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla5	H. glaber	liver	CRI01	CEBPA	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla6	H. glaber	liver	merged	H3K4me3	E-MTAB- 2633	yes	H_glaber_mtDNA(gi322422826)
Hgla6	H. glaber	liver	SAN01	input	E-MTAB- 2633	yes	H_glaber_mtDNA(gi322422826)
Hgla6	H. glaber	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	H_glaber_mtDNA(gi322422826)
Hgla6	H. glaber	liver	merged	H3K4me1	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla6	H. glaber	liver	merged	H3K27Ac	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla6	H. glaber	liver	merged	totalH3	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hgla6	H. glaber	liver	CRI01	CEBPA	E-MTAB-3933	yes	H_glaber_mtDNA(gi322422826)
Hsap08	H. sapiens	liver	CRI01	input DNA	E-MTAB-1509	yes	human_g1k_v37
Hsap08	H. sapiens	liver	CRI01	HNF4A	E-MTAB-1509	yes	human_g1k_v37
Hsap08	H. sapiens	liver	CRI01	HNF4A	E-MTAB-1509	yes	human_g1k_v37
Hsap08	H. sapiens	liver	CRI01	input_DNA	E-TABM-722	yes	human_g1k_v37

A.4 Experimental details for each file analysed in the context of this project.

Hsap08	H. sapiens	liver	CRI01	HNF4a	E-TABM-722	yes	human_g1k_v37
Hsap08	H. sapiens	liver	CRI02	HNF4a	E-TABM-722	yes	human_g1k_v37
Hsap12878	H. sapiens	LCL	CRI01	CTCF	E-MTAB-1511	no	human_g1k_v37
Hsap12878	H. sapiens	LCL	CRI02	CTCF	E-MTAB-1511	no	human_g1k_v37
Hsap12878	H. sapiens	LCL	CRI03	NA	E-MTAB-1511	no	human_g1k_v37
Hsap20	H. sapiens	liver	CRI01	YY1	E-MTAB-1511	yes	human_g1k_v37
Hsap23	H. sapiens	liver	CRI01	input DNA	E-MTAB-1509	yes	human_g1k_v37
Hsap23	H. sapiens	liver	CRI01	CEBPA	E-MTAB-1509	yes	human_g1k_v37
Hsap23	H. sapiens	liver	CRI02	CEBPA	E-MTAB-1509	yes	human_g1k_v37
Hsap23	H. sapiens	liver	CRI01	input_DNA	E-TABM-722	yes	human_g1k_v37
Hsap23	H. sapiens	liver	CRI01	CEBPA	E-TABM-722	yes	human_g1k_v37
Hsap23	H. sapiens	liver	CRI02	CEBPA	E-TABM-722	yes	human_g1k_v37
Hsap28	H. sapiens	liver	CRI01	input_DNA	E-MTAB-1509	yes	human_g1k_v37
Hsap28	H. sapiens	liver	CRI01	HNF6	E-MTAB-1509	yes	human_g1k_v37
Hsap813	H. sapiens	LCL	CRI01	CTCF	E-MTAB-1511	no	human_g1k_v37
Hsap813	H. sapiens	LCL	CRI01	CTCF	E-MTAB-1511	no	human_g1k_v37
Hsap813	H. sapiens	LCL	CRI01	NA	E-MTAB-1511	no	human_g1k_v37
Hsap813	H. sapiens	LCL	CRI02	CTCF	E-MTAB-1511	no	human_g1k_v37
Hsap813	H. sapiens	LCL	CRI01	YY1	E-MTAB-1511	no	human_g1k_v37
Hsap813	H. sapiens	LCL	CRI01	YY1	E-MTAB-1511	no	human_g1k_v37
Hsap94	H. sapiens	liver	CRI01	input DNA	E-MTAB-1509	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI02	input DNA	E-MTAB-1509	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI01	CEBPA	E-MTAB-1509	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI02	CEBPA	E-MTAB-1509	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI01	FOXA1	E-MTAB-1509	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI01	HNF4A	E-MTAB-1509	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI01	HNF6	E-MTAB-1509	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI01	input_DNA	E-MTAB-437	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI02	input_DNA	E-MTAB-437	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI02	CTCF	E-MTAB-437	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI01	input_DNA	E-TABM-722	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI02	input_DNA	E-TABM-722	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI01	CEBPA	E-TABM-722	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI02	CEBPA	E-TABM-722	yes	human_g1k_v37
Hsap94	H. sapiens	liver	CRI01	HNF4a	E-TABM-722	yes	human_g1k_v37
Hsap99	H. sapiens	liver	CRI01	YY1	E-MTAB-1511	yes	human_g1k_v37
HsapC3	H. sapiens	liver	CRI01	H3K4me3	E-MTAB- 2633	yes	human_g1k_v37
HsapC3	H. sapiens	liver	CRI01	H3K27Ac	E-MTAB- 2633	yes	human_g1k_v37
HsapC3	H. sapiens	liver	CRI01	input	E-MTAB- 2633	yes	human_g1k_v37
HsapC3	H. sapiens	liver	CRI01	input DNA	E-MTAB-1509	yes	human_g1k_v37
HsapC3	H. sapiens	liver	CRI01	FOXA1	E-MTAB-1509	yes	human_g1k_v37
HsapC6	H. sapiens	liver	CRI01	H3K4me3	E-MTAB- 2633	yes	human_g1k_v37
HsapC6	H. sapiens	liver	CRI01	H3K27Ac	E-MTAB- 2633	yes	human_g1k_v37
HsapC6	H. sapiens	liver	CRI01	input	E-MTAB- 2633	yes	human_g1k_v37
HsapC7	H. sapiens	liver	CRI01	H3K4me3	E-MTAB- 2633	yes	human_g1k_v37
HsapC7	H. sapiens	liver	CRI01	H3K27Ac	E-MTAB- 2633	yes	human_g1k_v37
HsapC7	H. sapiens	liver	CRI01	input	E-MTAB- 2633	yes	human_g1k_v37
HsapC8	H. sapiens	liver	CRI01	H3K4me3	E-MTAB- 2633	yes	human_g1k_v37
HsapC8	H. sapiens	liver	CRI01	H3K27Ac	E-MTAB- 2633	yes	human_g1k_v37
HsapC8	H. sapiens	liver	CRI01	input	E-MTAB- 2633	yes	human_g1k_v37
HsapC8	H. sapiens	liver	CRI01	input_DNA	E-MTAB-437	yes	human_g1k_v37
HsapC8	H. sapiens	liver	CRI01	CTCF	E-MTAB-437	yes	human_g1k_v37
HsapC8	H. sapiens	liver	CRI01	NRSF	E-MTAB-437	yes	human_g1k_v37
Lalb1	L. albirostris	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Lalb1	L. albirostris	liver	SAN01	input	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Lalb1	L. albirostris	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Lalb1	L. albirostris	liver	SAN01	H3K4me1	E-MTAB-3933	no	Tursiops_truncatus.turTru1
Mbid1	M. bidens	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Tursiops_truncatus.turTru1

A.4 Experimental details for each file analysed in the context of this project.

Mbid1	M. bidens	liver	merged	input	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Mbid1	M. bidens	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Mbid1	M. bidens	liver	SAN01	H3K4me1	E-MTAB-3933	no	Tursiops_truncatus.turTru1
Mbid2	M. bidens	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Mbid2	M. bidens	liver	merged	input	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Mbid2	M. bidens	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Tursiops_truncatus.turTru1
Mbid2	M. bidens	liver	SAN01	H3K4me1	E-MTAB-3933	no	Tursiops_truncatus.turTru1
Mdom0	M. domestica	liver	CRI01	CEBPA	E-TABM-722	no	Monodelphis_domestica.BROADO5
Mdom0	M. domestica	liver	CRI02	CEBPA	E-TABM-722	no	Monodelphis_domestica.BROADO5
Mdom0	M. domestica	liver	CRI03	CEBPA	E-TABM-722	no	Monodelphis_domestica.BROADO5
Mdom1	M. domestica	liver	CRI01	input_DNA	E-TABM-722	no	Monodelphis_domestica.BROADO5
Mdom11	M. domestica	liver	merged	H3K4me3	E-MTAB- 2633	no	Monodelphis_domestica.BROADO5
Mdom11	M. domestica	liver	SAN01	input	E-MTAB- 2633	no	Monodelphis_domestica.BROADO5
Mdom11	M. domestica	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Monodelphis_domestica.BROADO5
Mdom11	M. domestica	liver	merged	H3K4me1	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom11	M. domestica	liver	merged	H3K27Ac	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom11	M. domestica	liver	merged	totalH3	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom11	M. domestica	liver	SAN01	H3K4me1	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom11	M. domestica	liver	CRI01	CEBPA	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom12	M. domestica	liver	merged	H3K4me3	E-MTAB- 2633	no	Monodelphis_domestica.BROADO5
Mdom12	M. domestica	liver	SAN01	input	E-MTAB- 2633	no	Monodelphis_domestica.BROADO5
Mdom12	M. domestica	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Monodelphis_domestica.BROADO5
Mdom12	M. domestica	liver	merged	H3K4me1	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom12	M. domestica	liver	merged	H3K27Ac	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom12	M. domestica	liver	merged	totalH3	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom12	M. domestica	liver	SAN01	H3K4me1	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom12	M. domestica	liver	CRI01	CEBPA	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom13	M. domestica	liver	merged	H3K4me3	E-MTAB- 2633	no	Monodelphis_domestica.BROADO5
Mdom13	M. domestica	liver	SAN01	input	E-MTAB- 2633	no	Monodelphis_domestica.BROADO5
Mdom13	M. domestica	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Monodelphis_domestica.BROADO5
Mdom13	M. domestica	liver	merged	H3K4me1	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom13	M. domestica	liver	merged	H3K27Ac	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom13	M. domestica	liver	merged	totalH3	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom13	M. domestica	liver	CRI01	CEBPA	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom13	M. domestica	liver	SAN01	H3K4me1	E-MTAB-3933	no	Monodelphis_domestica.BROADO5
Mdom3	M. domestica	liver	CRI01	SA1	E-MTAB-437	no	Monodelphis_domestica.BROADO5
Mdom3	M. domestica	liver	CRI02	SA1	E-MTAB-437	no	Monodelphis_domestica.BROADO5
Mdom3	M. domestica	liver	CRI03	SA1	E-MTAB-437	no	Monodelphis_domestica.BROADO5
Mdom3	M. domestica	liver	CRI01	CEBPA	E-TABM-722	no	Monodelphis_domestica.BROADO5
Mdom3	M. domestica	liver	CRI02	CEBPA	E-TABM-722	no	Monodelphis_domestica.BROADO5
Mdom3	M. domestica	liver	CRI01	input_DNA	E-TABM-722	no	Monodelphis_domestica.BROADO5
Mdom4	M. domestica	liver	CRI01	CTCF	E-MTAB-437	no	Monodelphis_domestica.BROADO5
Mdom4	M. domestica	liver	CRI01	input_DNA	E-MTAB-437	no	Monodelphis_domestica.BROADO5
Mdom5	M. domestica	liver	CRI01	CTCF	E-MTAB-437	no	Monodelphis_domestica.BROADO5
Mdom5	M. domestica	liver	CRI01	input_DNA	E-MTAB-437	no	Monodelphis_domestica.BROADO5
Mfur1	M. furo	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	M_putorius_mtDNA(gi470229624)
Mfur1	M. furo	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	M_putorius_mtDNA(gi470229624)
Mfur1	M. furo	liver	SAN01	input	E-MTAB- 2633	yes	M_putorius_mtDNA(gi470229624)
Mfur1	M. furo	liver	SAN02	H3K4me1	E-MTAB-3933	yes	M_putorius_mtDNA(gi470229624)
Mfur1	M. furo	liver	CRI01	CEBPA	E-MTAB-3933	yes	M_putorius_mtDNA(gi470229624)
Mfur2	M. furo	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	M_putorius_mtDNA(gi470229624)
Mfur2	M. furo	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	M_putorius_mtDNA(gi470229624)
Mfur2	M. furo	liver	SAN01	input	E-MTAB- 2633	yes	M_putorius_mtDNA(gi470229624)
Mfur2	M. furo	liver	SAN01	H3K4me1	E-MTAB-3933	yes	M_putorius_mtDNA(gi470229624)
Mfur2	M. furo	liver	CRI01	CEBPA	E-MTAB-3933	yes	M_putorius_mtDNA(gi470229624)
Mfur3	M. furo	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	M_putorius_mtDNA(gi470229624)
Mfur3	M. furo	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	M_putorius_mtDNA(gi470229624)
Mfur3	M. furo	liver	SAN01	input	E-MTAB- 2633	yes	M_putorius_mtDNA(gi470229624)

A.4 Experimental details for each file analysed in the context of this project.

Mfur3	M. furo	liver	SAN01	H3K4me1	E-MTAB-3933	yes	M_putorius_mtDNA(gi470229624)
Mfur3	M. furo	liver	CRI01	CEBPA	E-MTAB-3933	yes	M_putorius_mtDNA(gi470229624)
Mmul150	M. mulatta	LCL	CRI01	CTCF	E-MTAB-1511	no	Macaca_mulatta.MMUL_1.65
Mmul150	M. mulatta	LCL	CRI01	CTCF	E-MTAB-1511	no	Macaca_mulatta.MMUL_1.65
Mmul173	M. mulatta	LCL	CRI01	CTCF	E-MTAB-1511	no	Macaca_mulatta.MMUL_1.65
Mmul173	M. mulatta	LCL	CRI01	CTCF	E-MTAB-1511	no	Macaca_mulatta.MMUL_1.65
Mmul290	M. mulatta	LCL	CRI01	CTCF	E-MTAB-1511	no	Macaca_mulatta.MMUL_1.65
Mmul38	M. mulatta	liver	CRI01	input_DNA	E-MTAB-437	yes	Macaca_mulatta.MMUL_1.65
Mmul38	M. mulatta	liver	CRI02	input_DNA	E-MTAB-437	yes	Macaca_mulatta.MMUL_1.65
Mmul38	M. mulatta	liver	CRI01	CTCF	E-MTAB-437	yes	Macaca_mulatta.MMUL_1.65
Mmul38	M. mulatta	liver	SAN02	CTCF	E-MTAB-437	yes	Macaca_mulatta.MMUL_1.65
Mmul71	M. mulatta	LCL	CRI01	CTCF	E-MTAB-1511	no	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	SAN01	input	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	merged	H3K27Ac	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI01	CEBPA	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI01	input DNA	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI01	input DNA	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI02	input DNA	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI01	CEBPA	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI02	CEBPA	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI03	CEBPA	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI04	CEBPA	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI01	FOXA1	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI01	HNF4A	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI01	HNF4A	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI01	HNF6	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	SAN01	totalH3	E-MTAB-3933	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Macaca_mulatta.MMUL_1.65
MmulBL	M. mulatta	liver	CRI01	CTCF	E-MTAB-437	yes	Macaca_mulatta.MMUL_1.65
MmulBO	M. mulatta	liver	CRI01	HNF6	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulBO	M. mulatta	liver	CRI01	input DNA	E-MTAB-1509	yes	Macaca_mulatta.MMUL_1.65
MmulDI	M. mulatta	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulDI	M. mulatta	liver	SAN01	input	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulDI	M. mulatta	liver	merged	H3K27Ac	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulDI	M. mulatta	liver	SAN01	totalH3	E-MTAB-3933	yes	Macaca_mulatta.MMUL_1.65
MmulDI	M. mulatta	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Macaca_mulatta.MMUL_1.65
MmulJO	M. mulatta	liver	CRI01	CTCF	E-MTAB-437	yes	Macaca_mulatta.MMUL_1.65
MmulJO	M. mulatta	liver	CRI01	input_DNA	E-MTAB-437	yes	Macaca_mulatta.MMUL_1.65
MmulJO	M. mulatta	liver	CRI02	input_DNA	E-MTAB-437	yes	Macaca_mulatta.MMUL_1.65
MmulJO	M. mulatta	liver	CRI03	input_DNA	E-MTAB-437	yes	Macaca_mulatta.MMUL_1.65
MmulLA	M. mulatta	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulLA	M. mulatta	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulLA	M. mulatta	liver	SAN01	input	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulST	M. mulatta	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulST	M. mulatta	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
MmulST	M. mulatta	liver	SAN01	input	E-MTAB- 2633	yes	Macaca_mulatta.MMUL_1.65
Mmus10	M. musculus	liver	CRI01	YY1	E-MTAB-1511	yes	Mus_musculus.NCBIM37.65
Mmus11	M. musculus	liver	CRI01	CTCF	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus12	M. musculus	liver	CRI01	HNF6	E-MTAB-1509	yes	Mus_musculus.NCBIM37.65
Mmus12	M. musculus	liver	CRI01	HNF4a	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus12	M. musculus	liver	SAN02	HNF4a	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus12	M. musculus	liver	SAN03	HNF4a	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus14	M. musculus	liver	CRI01	input_DNA	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus14	M. musculus	liver	CRI02	input_DNA	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus14	M. musculus	liver	CRI03	input_DNA	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus14	M. musculus	liver	CRI01	input_DNA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus14	M. musculus	liver	CRI02	input_DNA	E-TABM-722	yes	Mus_musculus.NCBIM37.65

A.4 Experimental details for each file analysed in the context of this project.

Mmus14	M. musculus	liver	CRI03	input_DNA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus2	M. musculus	liver	CRI02	CTCF	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus2	M. musculus	liver	CRI01	HNF4a	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus2	M. musculus	liver	CRI02	HNF4a	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus2	M. musculus	liver	CRI01	CEBPA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus2	M. musculus	liver	CRI02	CEBPA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus20	M. musculus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus20	M. musculus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus20	M. musculus	liver	SAN01	input	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus21	M. musculus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus21	M. musculus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus21	M. musculus	liver	SAN01	input	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus23	M. musculus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus23	M. musculus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus23	M. musculus	liver	SAN01	input	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus24	M. musculus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus24	M. musculus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus24	M. musculus	liver	SAN01	input	E-MTAB- 2633	yes	Mus_musculus.NCBIM37.65
Mmus28	M. musculus	liver	CRI01	input_DNA	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus28	M. musculus	liver	CRI01	input_DNA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus29	M. musculus	liver	CRI01	HNF4a	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus29	M. musculus	liver	CRI02	HNF4a	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus4	M. musculus	liver	CRI02	CTCF	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus4	M. musculus	liver	CRI01	CEBPA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus44	M. musculus	liver	CRI01	FOXA1	E-MTAB-1509	yes	Mus_musculus.NCBIM37.65
Mmus44	M. musculus	liver	CRI01	CTCF	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus45	M. musculus	liver	CRI01	FOXA1	E-MTAB-1509	yes	Mus_musculus.NCBIM37.65
Mmus5	M. musculus	liver	CRI02	CTCF	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus5	M. musculus	liver	CRI01	CEBPA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus62	M. musculus	liver	CRI01	CEBPA	E-MTAB-1509	yes	Mus_musculus.NCBIM37.65
Mmus62	M. musculus	liver	CRI01	input DNA	E-MTAB-1509	yes	Mus_musculus.NCBIM37.65
Mmus63	M. musculus	liver	CRI01	input_DNA	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus63	M. musculus	liver	CRI02	input_DNA	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus63	M. musculus	liver	CRI01	input_DNA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus63	M. musculus	liver	CRI02	input_DNA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus65	M. musculus	liver	CRI01	input_DNA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus66	M. musculus	liver	CRI01	H2AK5ac	E-MTAB-437	no	Mus_musculus.NCBIM37.65
Mmus68	M. musculus	liver	CRI02	CEBPA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus68	M. musculus	liver	CRI03	CEBPA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus69	M. musculus	liver	CRI01	CEBPA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus69	M. musculus	liver	CRI02	CEBPA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus69	M. musculus	liver	CRI03	CEBPA	E-TABM-722	yes	Mus_musculus.NCBIM37.65
Mmus72	M. musculus	liver	CRI01	YY1	E-MTAB-1511	yes	Mus_musculus.NCBIM37.65
Mmus86	M. musculus	liver	CRI01	H2AK5ac	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus89	M. musculus	liver	CRI01	CEBPA	E-MTAB-1509	yes	Mus_musculus.NCBIM37.65
Mmus89	M. musculus	liver	CRI01	HNF4A	E-MTAB-1509	yes	Mus_musculus.NCBIM37.65
Mmus89	M. musculus	liver	CRI01	input DNA	E-MTAB-1509	yes	Mus_musculus.NCBIM37.65
Mmus89	M. musculus	liver	CRI01	input_DNA	E-MTAB-437	yes	Mus_musculus.NCBIM37.65
Mmus91	M. musculus	liver	CRI01	HNF4A	E-MTAB-1509	yes	Mus_musculus.NCBIM37.65
Mmus91	M. musculus	liver	CRI01	HNF6	E-MTAB-1509	yes	Mus_musculus.NCBIM37.65
Ocun4	O. cuniculus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun4	O. cuniculus	liver	merged	input	E-MTAB- 2633	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun4	O. cuniculus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun4	O. cuniculus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun4	O. cuniculus	liver	CRI01	CEBPA	E-MTAB-3933	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun5	O. cuniculus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun5	O. cuniculus	liver	merged	input	E-MTAB- 2633	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun5	O. cuniculus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Oryctolagus_cuniculus.OryCun2.0

A.4 Experimental details for each file analysed in the context of this project.

Ocun5	O. cuniculus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun5	O. cuniculus	liver	CRI01	CEBPA	E-MTAB-3933	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun6	O. cuniculus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun6	O. cuniculus	liver	merged	input	E-MTAB- 2633	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun6	O. cuniculus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun6	O. cuniculus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Oryctolagus_cuniculus.OryCun2.0
Ocun6	O. cuniculus	liver	CRI01	CEBPA	E-MTAB-3933	yes	Oryctolagus_cuniculus.OryCun2.0
Ogar1	O. garnettii	liver	CRI01	H3K4me3	E-MTAB-3933	yes	O_crassicaudatus_mtDNA(gi238866848)
Ogar1	O. garnettii	liver	CRI01	H3K4me1	E-MTAB-3933	yes	O_crassicaudatus_mtDNA(gi238866848)
Ogar1	O. garnettii	liver	CRI01	H3K27Ac	E-MTAB-3933	yes	O_crassicaudatus_mtDNA(gi238866848)
Ogar1	O. garnettii	liver	CRI01	totalH3	E-MTAB-3933	yes	O_crassicaudatus_mtDNA(gi238866848)
Ogar1	O. garnettii	liver	CRI01	input	E-MTAB-3933	yes	O_crassicaudatus_mtDNA(gi238866848)
Pham1	P. hamadryas	LCL	CRI01	CTCF	E-MTAB-1511	no	Papio_anubis.PapAnu2.0
Pham1	P. hamadryas	LCL	CRI01	NA	E-MTAB-1511	no	Papio_anubis.PapAnu2.0
Pham1	P. hamadryas	LCL	CRI01	CTCF	E-MTAB-1511	no	Papio_anubis.PapAnu2.0
Pham1	P. hamadryas	LCL	CRI01	YY1	E-MTAB-1511	no	Papio_anubis.PapAnu2.0
Pham1	P. hamadryas	LCL	CRI01	YY1	E-MTAB-1511	no	Papio_anubis.PapAnu2.0
Ppyg1	P. pygmaeus	LCL	CRI01	CTCF	E-MTAB-1511	no	Pongo_abelii.PPYG2.66
Ppyg1	P. pygmaeus	LCL	CRI01	NA	E-MTAB-1511	no	Pongo_abelii.PPYG2.66
Ppyg1	P. pygmaeus	LCL	CRI01	CTCF	E-MTAB-1511	no	Pongo_abelii.PPYG2.66
Ppyg1	P. pygmaeus	LCL	CRI01	CTCF	E-MTAB-1511	no	Pongo_abelii.PPYG2.66
Ppyg1	P. pygmaeus	LCL	CRI01	YY1	E-MTAB-1511	no	Pongo_abelii.PPYG2.66
Ppyg1	P. pygmaeus	LCL	CRI01	YY1	E-MTAB-1511	no	Pongo_abelii.PPYG2.66
Ptro1	P. troglodytes	LCL	CRI01	CTCF	E-MTAB-1511	no	Pan_troglodytes.CHIMP2.1.4
Ptro1	P. troglodytes	LCL	CRI01	NA	E-MTAB-1511	no	Pan_troglodytes.CHIMP2.1.4
Ptro1	P. troglodytes	LCL	CRI01	CTCF	E-MTAB-1511	no	Pan_troglodytes.CHIMP2.1.4
Ptro1	P. troglodytes	LCL	CRI01	CTCF	E-MTAB-1511	no	Pan_troglodytes.CHIMP2.1.4
Ptro1	P. troglodytes	LCL	CRI01	YY1	E-MTAB-1511	no	Pan_troglodytes.CHIMP2.1.4
Ptro1	P. troglodytes	LCL	CRI01	YY1	E-MTAB-1511	no	Pan_troglodytes.CHIMP2.1.4
Ptro2	P. troglodytes	LCL	CRI01	CTCF	E-MTAB-1511	no	Pan_troglodytes.CHIMP2.1.4
Ptro3	P. troglodytes	LCL	CRI01	CTCF	E-MTAB-1511	no	Pan_troglodytes.CHIMP2.1.4
Rnor10	R. norvegicus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Rattus_norvegicus.RGSC3.4.65
Rnor10	R. norvegicus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Rattus_norvegicus.RGSC3.4.65
Rnor10	R. norvegicus	liver	SAN01	input	E-MTAB- 2633	yes	Rattus_norvegicus.RGSC3.4.65
Rnor10	R. norvegicus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Rattus_norvegicus.RGSC3.4.65
Rnor5	R. norvegicus	liver	CRI01	HNF6	E-MTAB-1509	yes	Rattus_norvegicus.RGSC3.4.65
Rnor5	R. norvegicus	liver	CRI01	HNF4A	E-MTAB-1509	yes	Rattus_norvegicus.RGSC3.4.65
Rnor5	R. norvegicus	liver	CRI01	FOXA1	E-MTAB-1509	yes	Rattus_norvegicus.RGSC3.4.65
Rnor5	R. norvegicus	liver	CRI01	CEBPA	E-MTAB-1509	yes	Rattus_norvegicus.RGSC3.4.65
Rnor7	R. norvegicus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Rattus_norvegicus.RGSC3.4.65
Rnor7	R. norvegicus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Rattus_norvegicus.RGSC3.4.65
Rnor7	R. norvegicus	liver	SAN01	input	E-MTAB- 2633	yes	Rattus_norvegicus.RGSC3.4.65
Rnor7	R. norvegicus	liver	CRI01	input DNA	E-MTAB-1509	yes	Rattus_norvegicus.RGSC3.4.65
Rnor7	R. norvegicus	liver	CRI01	HNF4A	E-MTAB-1509	yes	Rattus_norvegicus.RGSC3.4.65
Rnor7	R. norvegicus	liver	CRI01	CEBPA	E-MTAB-1509	yes	Rattus_norvegicus.RGSC3.4.65
Rnor7	R. norvegicus	liver	SAN02	H3K4me1	E-MTAB-3933	yes	Rattus_norvegicus.RGSC3.4.65
Rnor7	R. norvegicus	liver	CRI01	CTCF	E-MTAB-437	yes	Rattus_norvegicus.RGSC3.4.65
Rnor7	R. norvegicus	liver	CRI01	input_DNA	E-MTAB-437	yes	Rattus_norvegicus.RGSC3.4.65
Rnor8	R. norvegicus	liver	CRI01	input DNA	E-MTAB-1509	yes	Rattus_norvegicus.RGSC3.4.65
Rnor8	R. norvegicus	liver	CRI02	FOXA1	E-MTAB-1509	yes	Rattus_norvegicus.RGSC3.4.65
Rnor8	R. norvegicus	liver	CRI01	HNF6	E-MTAB-1509	yes	Rattus_norvegicus.RGSC3.4.65
Rnor8	R. norvegicus	liver	CRI01	CTCF	E-MTAB-437	yes	Rattus_norvegicus.RGSC3.4.65
Rnor8	R. norvegicus	liver	CRI01	input_DNA	E-MTAB-437	yes	Rattus_norvegicus.RGSC3.4.65
Rnor9	R. norvegicus	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Rattus_norvegicus.RGSC3.4.65
Rnor9	R. norvegicus	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Rattus_norvegicus.RGSC3.4.65
Rnor9	R. norvegicus	liver	SAN01	input	E-MTAB- 2633	yes	Rattus_norvegicus.RGSC3.4.65
Rnor9	R. norvegicus	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Rattus_norvegicus.RGSC3.4.65
Shar1	S. harrisii	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Sarcophilus_harrisii.DEVIL7.0

A.4 Experimental details for each file analysed in the context of this project.

Shar1	S. harrisii	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Sarcophilus_harrisii.DEVIL7.0
Shar1	S. harrisii	liver	SAN01	input	E-MTAB- 2633	yes	Sarcophilus_harrisii.DEVIL7.0
Shar1	S. harrisii	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Sarcophilus_harrisii.DEVIL7.0
Shar1	S. harrisii	liver	SAN01	totalH3	E-MTAB-3933	yes	Sarcophilus_harrisii.DEVIL7.0
Shar1	S. harrisii	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Sarcophilus_harrisii.DEVIL7.0
Shar2	S. harrisii	liver	SAN01	input	E-MTAB- 2633	yes	Sarcophilus_harrisii.DEVIL7.0
Shar2	S. harrisii	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Sarcophilus_harrisii.DEVIL7.0
Shar2	S. harrisii	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Sarcophilus_harrisii.DEVIL7.0
Shar2	S. harrisii	liver	SAN01	H3K4me3	E-MTAB-3933	yes	Sarcophilus_harrisii.DEVIL7.0
Shar2	S. harrisii	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Sarcophilus_harrisii.DEVIL7.0
Shar2	S. harrisii	liver	SAN01	H3K27Ac	E-MTAB-3933	yes	Sarcophilus_harrisii.DEVIL7.0
Shar2	S. harrisii	liver	SAN01	totalH3	E-MTAB-3933	yes	Sarcophilus_harrisii.DEVIL7.0
Shar2	S. harrisii	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Sarcophilus_harrisii.DEVIL7.0
Soed1	S. oedipus	LCL	CRI01	CTCF	E-MTAB-1511	no	Callithrix_jacchus.C_jacchus3.2.1
Soed1	S. oedipus	LCL	CRI02	CTCF	E-MTAB-1511	no	Callithrix_jacchus.C_jacchus3.2.1
Soed1	S. oedipus	LCL	CRI02	NA	E-MTAB-1511	no	Callithrix_jacchus.C_jacchus3.2.1
Sscr4	S. scrofa	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Sus_scrofa.Sscrofa10.2
Sscr4	S. scrofa	liver	SAN01	input	E-MTAB- 2633	yes	Sus_scrofa.Sscrofa10.2
Sscr4	S. scrofa	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Sus_scrofa.Sscrofa10.2
Sscr4	S. scrofa	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Sus_scrofa.Sscrofa10.2
Sscr4	S. scrofa	liver	SAN01	H3K27Ac	E-MTAB-3933	yes	Sus_scrofa.Sscrofa10.2
Sscr4	S. scrofa	liver	SAN01	totalH3	E-MTAB-3933	yes	Sus_scrofa.Sscrofa10.2
Sscr5	S. scrofa	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Sus_scrofa.Sscrofa10.2
Sscr5	S. scrofa	liver	SAN01	input	E-MTAB- 2633	yes	Sus_scrofa.Sscrofa10.2
Sscr5	S. scrofa	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Sus_scrofa.Sscrofa10.2
Sscr5	S. scrofa	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Sus_scrofa.Sscrofa10.2
Sscr5	S. scrofa	liver	SAN01	H3K27Ac	E-MTAB-3933	yes	Sus_scrofa.Sscrofa10.2
Sscr6	S. scrofa	liver	SAN01	H3K4me3	E-MTAB- 2633	yes	Sus_scrofa.Sscrofa10.2
Sscr6	S. scrofa	liver	SAN01	input	E-MTAB- 2633	yes	Sus_scrofa.Sscrofa10.2
Sscr6	S. scrofa	liver	SAN01	H3K27Ac	E-MTAB- 2633	yes	Sus_scrofa.Sscrofa10.2
Sscr6	S. scrofa	liver	SAN01	H3K4me1	E-MTAB-3933	yes	Sus_scrofa.Sscrofa10.2
Sscr6	S. scrofa	liver	SAN01	H3K27Ac	E-MTAB-3933	yes	Sus_scrofa.Sscrofa10.2
Tbel1	T. belangeri	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Tupaia_belangeri.TREESHREW
Tbel1	T. belangeri	liver	SAN01	input	E-MTAB- 2633	no	Tupaia_belangeri.TREESHREW
Tbel1	T. belangeri	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Tupaia_belangeri.TREESHREW
Tbel1	T. belangeri	liver	SAN01	H3K4me1	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel1	T. belangeri	liver	SAN01	H3K27Ac	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel1	T. belangeri	liver	SAN01	totalH3	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel1	T. belangeri	liver	CRI01	CEBPA	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel2	T. belangeri	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Tupaia_belangeri.TREESHREW
Tbel2	T. belangeri	liver	SAN01	input	E-MTAB- 2633	no	Tupaia_belangeri.TREESHREW
Tbel2	T. belangeri	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Tupaia_belangeri.TREESHREW
Tbel2	T. belangeri	liver	SAN01	H3K4me1	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel2	T. belangeri	liver	SAN01	H3K27Ac	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel2	T. belangeri	liver	SAN01	totalH3	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel2	T. belangeri	liver	CRI01	CEBPA	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel3	T. belangeri	liver	SAN01	input	E-MTAB- 2633	no	Tupaia_belangeri.TREESHREW
Tbel3	T. belangeri	liver	SAN01	H3K27Ac	E-MTAB- 2633	no	Tupaia_belangeri.TREESHREW
Tbel3	T. belangeri	liver	CRI01	CEBPA	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel4	T. belangeri	liver	SAN01	H3K4me3	E-MTAB- 2633	no	Tupaia_belangeri.TREESHREW
Tbel4	T. belangeri	liver	SAN01	H3K4me1	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel4	T. belangeri	liver	SAN01	H3K27Ac	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel4	T. belangeri	liver	SAN01	totalH3	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel4	T. belangeri	liver	SAN01	input	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel4	T. belangeri	liver	CRI01	CEBPA	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW
Tbel4	T. belangeri	liver	SAN01	H3K27Ac	E-MTAB-3933	no	Tupaia_belangeri.TREESHREW

Bibliography

1. Gest, H. The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society. *Notes and records of the Royal Society of London* **58,** 187–201 (2004).
2. Darwin, C. On the origin of species by means of natural selection. 1859. *Murray, London* **502** (1991).
3. Dill, K. A. Annual Review of Biophysics Volume 43, 2014 Introduction. *Annual Review of Biophysics* **43,** V–VI (2014).
4. Riley, M. R. Introducing Journal of Biological Engineering. *Journal of Biological Engineering* **1,** 1–3 (2007).
5. Luscombe, N. M., Greenbaum, D. & Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine* **40,** 346–358 (2001).
6. Hagen, J. B. The origins of bioinformatics. *Nature Reviews Genetics* **1,** 231–236 (2000).
7. Smith, D. R. The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs? *Briefings in Functional Genomics,* elv027 (2015).
8. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

9. Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., *et al.* Ensembl 2015. *Nucleic Acids Research* **43,** D662–D669 (2015).
10. Simmons, D. The use of animal models in studying genetic disease: transgenesis and induced mutation. *Nature education* **1,** 70 (2008).
11. Alföldi, J. & Lindblad-Toh, K. Comparative genomics as a tool to understand evolution and disease. *Genome Research* **23,** 1063–1068 (2013).
12. Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brannstrom, A., *et al.* Genomics and the origin of species. *Nature Reviews Genetics* **15,** 176–192 (2014).
13. Higgs, P. G. & Lehman, N. The RNA World: molecular cooperation at the origins of life. *Nature Reviews Genetics* **16,** 7–17 (2015).
14. Koonin, E. V. Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics* **39,** 309–338 (2005).
15. Kumar, D. & Eng, C. Genomic Medicine: Principles and Practice. *Oxford University Press* **852** (2014).
16. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11,** 415–425 (2010).
17. Eliceiri, K. W., Berthold, M. R., Goldberg, I. G., Ibáñez, L., Manjunath, B. S., Martone, M. E., Murphy, R. F., Peng, H., Plant, A. L., Roysam, B., *et al.* Biological imaging software tools. *Nature Methods* **9,** 697–710 (2012).
18. Dill, K. A. & MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **338,** 1042–1046 (2012).
19. Altman, R. B. The expanding scope of bioinformatics: sequence analysis and beyond. *Heredity* **90,** 345 (2003).
20. Hedges, S. B. The origin and evolution of model organisms. *Nature Reviews Genetics* **3,** 838–49 (2002).

21. Dahm, R. Friedrich Miescher and the discovery of DNA. *Developmental Biology* **278,** 274–288 (2005).
22. Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M, Siepel, A. C., Thomas, P. J., McDowell, J. C., *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793 (2003).
23. Yoder, A. D. & Yang, Z. Estimation of Primate Speciation Dates Using Local Molecular Clocks. *Molecular Biology and Evolution* **17,** 1081–1090 (2000).
24. Pennisi, E. Is It Time to Uproot the Tree of Life? *Science* **284,** 1305–1307 (1999).
25. Huson, D. H. & Scornavacca, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology* **61,** 1061–1067 (2012).
26. Cooper, G. M. The Cell: A Molecular Approach. 2nd edition. *Sunderland (MA): Sinauer Associates* **1656** (2000).
27. Bertram, J. S. The molecular biology of cancer. *Molecular Aspects Of Medicine* **21,** 167–223 (2000).
28. Kumar, A. & Singh, A. A review on Alzheimer's disease pathophysiology and its management: an update. *Pharmacological Reports* **67,** 195–203 (2015).
29. Beitz, J. M. Parkinson's disease: a review. *Frontiers in bioscience (Scholar edition)* **6,** 65–74 (2014).
30. Forterre, P. Defining Life: The Virus Viewpoint. *Origins of Life and Evolution of Biospheres* **40,** 151–160 (2010).
31. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561–563 (1970).
32. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737–738 (1953).
33. Horspool, D. File:Central Dogma of Molecular Biochemistry with Enzymes.jpg. *Wikimedia Commons* (2008).
34. Yadav, S. P. The Wholeness in Suffix -omics, -omes, and the Word Om. *Journal of Biomolecular Techniques : JBT* **18,** 277 (2007).

35. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155,** 27–38 (2013).

36. Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., *et al.* The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications* **5** (2014).

37. Church, D., Schneider, V., Steinberg, K., Schatz, M., Quinlan, A., Chin, C.-S., Kitts, P., Aken, B., Marth, G., Hoffman, M., *et al.* Extending reference assembly models. *Genome Biology* **16,** 13 (2015).

38. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic Mapping in Human Disease. *Science* **322,** 881–888 (2008).

39. McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A. & Hirschhorn, J. N. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9,** 356–369 (2008).

40. Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Goncalves, Â., Kutter, C., Brown, G. D., Marshall, A., Flicek, P. & Odom, D. T. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148,** 335–348 (2012).

41. National Human Genome Research Institute. File:DNA transcription.png. *Wikimedia Commons* (2014).

42. Barre-Sinoussi, F., Ross, A. L. & Delfraissy, J.-F. Past, present and future: 30 years of HIV research. *Nature Reviews Microbiology* **11,** 877–883 (2013).

43. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10,** 57–63 (2009).

44. Boumphreyfr. File:Peptide syn.png. *Wikimedia Commons* (2009).

45. Wiseman, F. K., Alford, K. A., Tybulewicz, V. L. J. & Fisher, E. M. C. Down syndrome—recent progress and future prospects. *Human Molecular Genetics* **18,** R75–R83 (2009).

46. Hardy, J. & Singleton, A. Genomewide Association Studies and Human Disease. *New England Journal of Medicine* **360,** 1759–1768 (2009).
47. Cox, D. B. T., Platt, R. J. & Zhang, F. Therapeutic genome editing: prospects and challenges. *Nature Medicine* **21,** 121–131 (2015).
48. Ball, P. Smallest genome clocks in at 182 genes. *Nature News* (2006).
49. Metcalfe, C. J., Filée, J., Germon, I., Joss, J. & Casane, D. Evolution of the Australian lungfish (neoceratodus forsteri) genome: A major role for CR1 and L2 LINE elements. *Molecular Biology and Evolution* **29,** 3529–3539 (2012).
50. Gregory, T. R. The C-value enigma in plants and animals: A review of parallels and an appeal for partnership. *Annals of Botany* **95,** 133–146 (2005).
51. Electron Microscopy Facility at The National Cancer Institute at Frederick (NCI-Frederick). File:Red White Blood cells.jpg. *Wikimedia Commons* (2005).
52. De Roo, M. File:Neurone pyramidal.jpg. *Wikimedia Commons* (2012).
53. Shapiro, J. A. & von Sternberg, R. Why repetitive DNA is essential to genome function. *Biological reviews of the Cambridge Philosophical Society* **80,** 227–250 (2005).
54. Muotri, A. R., Marchetto, M. C. N., Coufal, N. G. & Gage, F. H. The necessary junk: New functions for transposable: Elements. *Human Molecular Genetics* **16** (2007).
55. Vinogradov, A. E. DNA helix: The importance of being GC-rich. *Nucleic Acids Research* **31,** 1838–1844 (2003).
56. Figuet, E., Ballenghien, M., Romiguier, J. & Galtier, N. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biology and Evolution* **7,** 240–250 (2014).
57. Galtier, N. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics* **39,** 251–6 (2001).
58. Vignal, A., Milan, D., SanCristobal, M. & Eggen, A. A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution* **34,** 275–306 (2002).
59. Friedländer, M. R., Lizano, E., Houben, A. J. S., Bezdan, D., Báñez-Coronel, M., Kudla, G., Mateu-Huertas, E., Kagerbauer, B., González, J., Chen, K. C., *et al.* Ev-

idence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biology* **15,** R57 (2014).

60. Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8 (2012).
61. Tessarz, P. & Kouzarides, T. Histone core modifications regulating nucleosome structure and dynamics. *Nature Reviews Molecular Cell Biology* **15,** 703–708 (2014).
62. Kim, S., Yu, N.-K. & Kaang, B.-K. CTCF as a multifunctional protein in genome regulation and gene expression. *Experimental & Molecular Medicine* **47,** e166 (2015).
63. Jakobsen, J. S., Waage, J., Rapin, N., Bisgaard, H. C., Larsen, F. S. & Porse, B. T. Temporal mapping of CEBPA and CEBPB binding during liver regeneration reveals dynamic occupancy and specific regulatory codes for homeostatic and cell cycle gene batteries. *Genome Research* **23,** 592–603 (2013).
64. Walesky, C., Gunewardena, S., Terwilliger, E. F., Edwards, G., Borude, P. & Apte, U. Hepatocyte-specific deletion of hepatocyte nuclear factor-4α in adult mice results in increased hepatocyte proliferation. *American journal of physiology. Gastrointestinal and liver physiology* **304,** G26–37 (2013).
65. Iyer, N. G., Ozdag, H & Caldas, C. p300/CBP and cancer. *Oncogene* **23,** 4225–4231 (2004).
66. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12,** 363–376 (2011).
67. Sharp, A. J., Cheng, Z. & Eichler, E. E. Structural variation of the human genome. *Annual review of genomics and human genetics* **7,** 407–442 (2006).
68. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
69. Teif, V. B. & Bohinc, K. Condensed DNA: Condensing the concepts. *Progress in Biophysics and Molecular Biology* **105,** 208–222 (2011).
70. Crozier, R. & Crosland, M. W. J. Myrmecia pilosula, an ant with only one pair of chromosomes. *Science* **231,** 1278 (1986).

71. Coluccia, E., Cannas, R., Cau, A., Deiana, A. M. & Salvadori, S. B chromosomes in Crustacea Decapoda. *Cytogenetic and genome research* **106,** 215–221 (2004).

72. Mardis, E. R. A decade's perspective on DNA sequencing technology. *Nature* **470,** 198–203 (2011).

73. Wang, Z., Chen, Y. & Li, Y. A Brief Review of Computational Gene Prediction Methods Gene Prediction Methods. *Genomics Proteomics Bioinformatics* **2,** 216–221 (2004).

74. Wheeler, R. File:Nucleosome 1KX5 2.png. *Wikimedia Commons* (2005).

75. Gray, M. W. Mitochondrial evolution. *Cold Spring Harbor Perspectives in Biology* **4** (2012).

76. Porter, C. & Wall, B. T. Skeletal muscle mitochondrial function: is it quality or quantity that makes the difference in insulin resistance? *The Journal of Physiology* **590,** 5935–5936 (2012).

77. Sato, M. & Sato, K. Maternal inheritance of mitochondrial DNA by diverse mechanisms to eliminate paternal mitochondrial DNA. *Biochimica et Biophysica Acta - Molecular Cell Research* **1833,** 1979–1984 (2013).

78. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10,** 512–526 (1993).

79. Castellana, S., Vicario, S. & Saccone, C. Evolutionary patterns of the mitochondrial genome in Metazoa: Exploring the role of mutation and selection in mitochondrial protein-coding genes. *Genome Biology and Evolution* **3,** 1067–1079 (2011).

80. Villarreal, M. R. File:Animal mitochondrion diagram en (edit).svg. *Wikimedia Commons* (2011).

81. Howard, L. File:Mitochondria, mammalian lung - TEM.jpg. *Wikimedia Commons* (2008).

82. Lang, B. F., Gray, M. W. & Burger, G. Mitochondrial genome evolution and the origin of eukaryotes. *Annual Review of Genetics* **33,** 351–397 (1999).

83. Taylor, R. W. & Turnbull, D. M. Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics* **6,** 389–402 (2005).

84. Hamilton, G. The hidden risks for 'three-person' babies. *Nature News* (2015).

85. Budowle, B., Allard, M. W., Wilson, M. R. & Chakraborty, R. Forensics and mitochondrial DNA: applications, debates, and foundations. *Annual review of genomics and human genetics* **4,** 119–141 (2003).

86. Magnacca, K. N. & Brown, M. J. F. Mitochondrial heteroplasmy and DNA barcoding in Hawaiian Hylaeus (Nesoprosopis) bees (Hymenoptera: Colletidae). *BMC Evolutionary Biology* **10,** 174 (2010).

87. Kalicharan, S. File:Mitochondrial DNA en.svg. *Wikimedia Commons* (2008).

88. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16,** 133–145 (2015).

89. Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A. & Bernstein, B. E. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology* **33,** 1165–1172 (2015).

90. Si-Tayeb, K., Lemaigre, F. P. & Duncan, S. A. Organogenesis and Development of the Liver. *Developmental Cell* **18,** 175–189 (2010).

91. Miyajima, A., Tanaka, M. & Itoh, T. Stem/progenitor cells in liver development, homeostasis, regeneration, and reprogramming. *Cell Stem Cell* **14,** 561–574 (2014).

92. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S. & Robinson, G. E. Big Data: Astronomical or Genomical? *PLoS Biology* **13,** e1002195 (2015).

93. Wu, R. & Taylor, E. Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology* **57,** 491–511 (1971).

94. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94,** 441–448 (1975).

95. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–9 (2008).

96. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* **242,** 84–89 (1996).

97. Metzker, M. L. Sequencing technologies - the next generation. *Nature Reviews Genetics* **11,** 31–46 (2010).

98. Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348–352 (2011).

99. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).

100. Feng, Y., Zhang, Y., Ying, C., Wang, D. & Du, C. Nanopore-based fourth-generation DNA sequencing technology. *Genomics, Proteomics and Bioinformatics* **13,** 4–16 (2015).

101. Venkatesan, B. M. & Bashir, R. Nanopore sensors for nucleic acid analysis. *Nature Nanotechnology* **6,** 615–624 (2011).

102. Fonseca, N. A., Rung, J., Brazma, A. & Marioni, J. C. Tools for mapping high-throughput sequencing data. *Bioinformatics* **28,** 3169–3177 (2012).

103. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

104. Burrows–Wheeler transform. *Wikipedia under CC BY-SA 3.0* (2016).

105. Hawkins, R. D., Hon, G. C. & Ren, B. Next-generation genomics: an integrative approach. *Nature Reviews Genetics* **11,** 476–486 (2010).

106. Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. & Zhao, K. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129,** 823–837 (2007).

107. Wilbanks, E. G. & Facciotti, M. T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* **5** (2010).

108. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9,** R137 (2008).

109. Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. & Gerstein, M. B. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology* **27,** 66–75 (2009).

110. Szalkowski, A. M. & Schmid, C. D. Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Briefings in Bioinformatics* **12,** 626–633 (2011).

111. Brind'Amour, J., Liu, S., Hudson, M., Chen, C., Karimi, M. M. & Lorincz, M. C. An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. *Nature Communications* **6,** 6033 (2015).

112. Pepke, S, Wold, B & Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6,** S22–32 (2009).

113. Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H. & Ecker, J. R. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133,** 523–536 (2008).

114. Proudfoot, N. J. Ending the message : poly (A) signals then and now. *Genes & Development* **25,** 1770–1782 (2011).

115. Chen, Z. & Duan, X. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods in Molecular Biology* **733,** 93–103 (2011).

116. Zhao, S. Assessment of the impact of using a reference transcriptome in mapping short RNA-Seq reads. *PLoS ONE* **9** (2014).

117. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29,** 644–52 (2011).

118. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11,** R106 (2010).

119. Miller, J. R., Koren, S. & Sutton, G. *Assembly algorithms for next-generation sequencing data* 2010.

120. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18,** 821–829 (2008).

121. Daugelaite, J., O'Driscoll, A. & Sleator, R. D. An overview of multiple sequence alignments and cloud computing in bioinformatics. *ISRN Biomathematics* **2013** (2013).

122. Löytynoja, A. & Goldman, N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics* **11,** 579 (2010).

123. Eppstein, D. File: DeBruijn-as-line-digraph.png. *Wikimedia Commons* (2006).

124. Mog9207. File:Example_1seq.pdf. *Wikimedia Commons* (2013).

125. Mog9207. File:Example_2simp.pdf. *Wikimedia Commons* (2013).

126. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* **16,** 321–32 (2015).

127. Häggström, M. File:Chi-square distributionCDF.png. *Wikimedia Commons* (2010).

128. Michaelg2015. File:Exam pass logistic curve.jpeg. *Wikimedia Commons* (2015).

129. Cyc. File:Svm max sep hyperplane with margin.png. *Wikimedia Commons* (2008).

130. Rensch, T., Villar, D., Horvath, J., Odom, D. T. & Flicek, P. Mitochondrial heteroplasmy in vertebrates using ChIP-sequencing data. *To Be Confirmed* (2016).

131. Taanman, J. W. The mitochondrial genome: structure, transcription, translation and replication. *Biochimica et Biophysica Acta* **1410,** 103–123 (1999).

132. Kukat, C., Wurm, C. A., Spåhr, H., Falkenberg, M., Larsson, N.-G. & Jakobs, S. Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtDNA. *Proceedings of the*

National Academy of Sciences of the United States of America **108,** 13534–13539 (2011).

133. Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290,** 457–465 (1981).

134. Pakendorf, B. & Stoneking, M. Mitochondrial DNA and human evolution. *Annual Review of Genomics and Human Genetics* **6,** 165–183 (2005).

135. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nature Reviews Genetics* **16,** 530–542 (2015).

136. Andreu, A. L. & DiMauro, S. Current classification of mitochondrial disorders. *Journal of Neurology* **250,** 1403–1406 (2003).

137. Rossignol, R., Faustin, B., Rocher, C., Malgat, M., Mazat, J.-P. & Letellier, T. Mitochondrial threshold effects. *The Biochemical Journal* **370,** 751–762 (2003).

138. Brandon, M, Baldi, P & Wallace, D. C. Mitochondrial mutations in cancer. *Oncogene* **25,** 4647–4662 (2006).

139. Zhidkov, I., Livneh, E. A., Rubin, E. & Mishmar, D. MtDNA mutation pattern in tumors and human evolution are shaped by similar selective constraints. *Genome Research* **19,** 576–580 (2009).

140. Coto, E., Gómez, J., Alonso, B., Corao, A. I., Díaz, M., Menéndez, M., Martínez, C., Calatayud, M. T., Morís, G. & Álvarez, V. Late-onset Alzheimer's disease is associated with mitochondrial DNA 7028C/haplogroup H and D310 poly-C tract heteroplasmy. *Neurogenetics* **12,** 345–346 (2011).

141. Lightowlers, R. N., Taylor, R. W. & Turnbull, D. M. Mutations causing mitochondrial disease: What is new and what challenges remain? *Science* **349,** 1494–1499 (2015).

142. Chinnery, P. F., Thorburn, D. R., Samuels, D. C., White, S. L., Dahl, H. H. M., Turnbull, D. M., Lightowlers, R. N. & Howell, N. The inheritance of mitochondrial DNA heteroplasmy: Random drift, selection or both? *Trends in Genetics* **16,** 500–505 (2000).

143. Brown, J. R., Beckenbach, K, Beckenbach, A. T. & Smith, M. J. Length variation, heteroplasmy and sequence divergence in the mitochondrial DNA of four species of sturgeon (Acipenser). *Genetics* **142,** 525–535 (1996).

144. Klütsch, C. F. C., Seppälä, E. H., Uhlén, M., Lohi, H. & Savolainen, P. Segregation of point mutation heteroplasmy in the control region of dog mtDNA studied systematically in deep generation pedigrees. *International Journal of Legal Medicine* **125,** 527–535 (2011).

145. Greenberg, B. D., Newbold, J. E. & Sugino, A. Intraspecific nucleotide mitochondrial DNA sequence variability surrounding the origin of replication in human. *Gene* **21,** 33–49 (1983).

146. Irwin, J. A., Saunier, J. L., Niederstätter, H., Strouss, K. M., Sturk, K. A., Diegoli, T. M., Brandstätter, A., Parson, W. & Parsons, T. J. Investigation of heteroplasmy in the human mitochondrial DNA control region: A synthesis of observations from more than 5000 global population samples. *Journal of Molecular Evolution* **68,** 516–527 (2009).

147. White, H. E., Durston, V. J., Seller, A., Fratter, C., Harvey, J. F. & Cross, N. C. P. Accurate detection and quantitation of heteroplasmic mitochondrial point mutations by pyrosequencing. *Genetic Testing* **9,** 190–199 (2005).

148. Li, M., Schönberg, A., Schaefer, M., Schroeder, R., Nasidze, I. & Stoneking, M. Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *American Journal of Human Genetics* **87,** 237–249 (2010).

149. Goto, H., Dickins, B., Afgan, E., Paul, I. M., Taylor, J., Makova, K. D. & Nekrutenko, A. Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome Biology* **12,** R59 (2011).

150. Li, M. & Stoneking, M. A new approach for detecting low-level mutations in next-generation sequence data. *Genome Biology* **13,** R34 (2012).

151. Ding, J., Sidore, C., Butler, T. J., Wing, M. K., Qian, Y., Meirelles, O., Busonero, F., Tsoi, L. C., Maschio, A., Angius, A., *et al.* Assessing Mitochondrial DNA Variation

and Copy Number in Lymphocytes of ˜2,000 Sardinians Using Tailored Sequencing Analysis Tools. *PLoS Genetics* **11,** e1005306 (2015).

152. Ye, K., Lu, J., Ma, F., Keinan, a. & Gu, Z. Reply to Just et al.: Mitochondrial DNA heteroplasmy could be reliably detected with massively parallel sequencing technologies. *Proceedings of the National Academy of Sciences* **111,** E4548–E4550 (2014).
153. Naue, J., Hörer, S., Sänger, T., Strobl, C., Hatzer-Grubwieser, P., Parson, W. & Lutz-Bonengel, S. Evidence for frequent and tissue-specific sequence heteroplasmy in human mitochondrial DNA. *Mitochondrion* **20,** 82–94 (2015).
154. Li, M., Schröder, R., Ni, S., Madea, B. & Stoneking, M. Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. *Proceedings of the National Academy of Sciences* **112,** 201419651 (2015).
155. Taylor, R. W., Taylor, G. A., Durham, S. E. & Turnbull, D. M. The determination of complete human mitochondrial DNA sequences in single cells: implications for the study of somatic mitochondrial DNA point mutations. *Nucleic Acids Research* **29,** e74 (2001).
156. Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., *et al.* ArrayExpress update–simplifying data submissions. *Nucleic Acids Research* **43,** D1113–D1116 (2015).
157. Aldridge, S., Watt, S., Quail, M. A., Rayner, T., Lukk, M., Bimson, M. F., Gaffney, D. & Odom, D. T. AHT-ChIP-seq: a completely automated robotic protocol for high-throughput chromatin immunoprecipitation. *Genome Biology* **14,** R124 (2013).
158. 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., *et al.* Five-Vertebrate ChIP-seq Reveals Transcription Factor Binding. *Science* **328,** 1036–1040 (2010).
159. Schwalie, P. C., Ward, M. C., Cain, C. E., Faure, A. J., Gilad, Y., Odom, D. T. & Flicek, P. Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biology* **14,** R148 (2013).
160. Ballester, B., Medina-Rivera, A., Schmidt, D., Gonzàlez-Porta, M., Carlucci, M., Chen, X., Chessman, K., Faure, A. J., Funnell, A. P., Goncalves, A., *et al.* Multi-

species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *eLife* **3,** 1–29 (2014).

161. Villar, D., Berthelot, C., Flicek, P., Odom, D. T., Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M. & Pignatelli, M. Enhancer Evolution across 20 Mammalian Species Article Enhancer Evolution across 20 Mammalian Species. *Cell* **160,** 554–566 (2015).

162. McVean, G. A., Altshuler (Co-Chair), D. M., Durbin (Co-Chair), R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

163. Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., *et al.* RefSeq: An update on mammalian reference sequences. *Nucleic Acids Research* **42,** 756–763 (2014).

164. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Subgroup, . G. P. D. P. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

165. Van Rossum, G. Python tutorial. *Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI)* (1995).

166. Jones, E., Oliphant, T. & Peterson, P. SciPy: Open source scientific tools for Python (2001).

167. McKinney, W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference,* 51–56 (2010).

168. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering* **9,** 99–104 (2007).

169. Ahmad, S., Ghosh, A., Nair, D. L. & Seshadri, M. Simultaneous extraction of nuclear and mitochondrial DNA from human blood. *Genes & Genetic Systems* **82,** 429–432 (2007).

170. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. & Cunningham, F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).

171. Waterhouse. Jalview Version 2 : a multiple sequence alignement editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).

172. Lott, M. T., Leipzig, J. N., Derbeneva, O., Xie, H. M., Chalkia, D., Sarmady, M., Procaccio, V. & Wallace, D. C. mtDNA Variation and Analysis Using MITOMAP and MITOMASTER. *Current Protocols in Bioinformatics* **1,** 1–26 (2013).

173. Hinrichs, A., Karolchik, D, Baertsch, R, Barber, G., Bejerano, G & Clawson, H. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research* **34,** D590–D598 (2006).

174. Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D. & Odom, D. T. ChIP-seq : using high-throughput sequencing to discover protein-DNA interactions. *Methods* **48,** 240–248 (2009).

175. Hazkani-Covo, E., Zeller, R. M. & Martin, W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genetics* **6,** e1000834 (2010).

176. Oh, J. H., Kim, Y. J., Moon, S., Nam, H.-Y., Jeon, J.-P., Lee, J. H., Lee, J.-Y. & Cho, Y. S. Genotype instability during long-term subculture of lymphoblastoid cell lines. *Journal of Human Genetics* **58,** 16–20 (2013).

177. Geisler, S. & Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature Reviews Molecular Cell Biology* **14,** 699–712 (2013).

178. Lin, M. F., Kheradpour, P., Washietl, S., Parker, B. J., Pedersen, J. S. & Kellis, M. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Research* **21,** 1916–1928 (2011).

179. Sosa, M. X., Sivakumar, I. K. A., Maragh, S., Veeramachaneni, V., Hariharan, R., Parulekar, M., Fredrikson, K. M., Harkins, T. T., Lin, J., Feldman, A. B., *et al.* Next-Generation Sequencing of Human Mitochondrial Reference Genomes Uncovers High Heteroplasmy Frequency. *PLoS Computational Biology* **8,** e1002737 (2012).

180. Sobenin, I. A., Mitrofanov, K. Y., Zhelankin, A. V., Sazonova, M. A., Postnov, A. Y., Revin, V. V., Bobryshev, Y. V. & Orekhov, A. N. Quantitative Assessment of Heteroplasmy of Mitochondrial Genome: Perspectives in Diagnostics and Methodological Pitfalls. *BioMed Research International* **2014,** 1–9 (2014).

181. Burgstaller, J. P., Johnston, I. G., Jones, N. S., Albrechtová, J., Kolbe, T., Vogl, C., Futschik, A., Mayrhofer, C., Klein, D., Sabitzer, S., *et al.* MtDNA Segregation in Heteroplasmic Tissues Is Common InVivo and Modulated by Haplotype Differences and Developmental Stage. *Cell Reports* **7,** 2031–2041 (2014).

182. Flensburg, C., Kinkel, S. A., Keniry, A., Blewitt, M. & Oshlack, A. A comparison of control samples for ChIP-seq of histone modifications. *Frontiers in Genetics* **5,** 1–8 (2014).

183. Hodgkinson, A., Idaghdour, Y., Gbeha, E., Grenier, J.-C., Hip-ki, E., Bruat, V., Goulet, J.-p., Malliard, T. D. & Awadalla, P. High-Resolution Genomic Analysis of Human Mitochondrial RNA Sequence Variation. *Science* **344,** 413–415 (2014).

184. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews Molecular Cell Biology* **16,** 129–143 (2015).

185. Reynolds, N., O'Shaughnessy, A. & Hendrich, B. Transcriptional repressors: multifaceted regulators of gene expression. *Development* **140,** 505–512 (2013).

186. Huang, H., Sabari, B. R., Garcia, B. A., Allis, C. D. & Zhao, Y. SnapShot: Histone Modifications. *Cell* **159,** 458–458.e1 (2015).

187. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* **21,** 381–395 (2011).

188. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. a. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics* **10,** 252–263 (2009).

189. Faure, A. J., Schmidt, D., Watt, S., Schwalie, P. C., Wilson, M. D., Xu, H., Ramsay, R. G., Odom, D. T. & Flicek, P. Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome Research* **22,** 2163–2175 (2012).

190. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2013).

191. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* **306,** 636–640 (2004).

192. 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., *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).

193. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology* **9,** e1001046 (2011).

194. Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research* **22,** 1798–1812 (2012).

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

196. De Nadal, E., Ammerer, G. & Posas, F. Controlling gene expression in response to stress. *Nature Reviews Genetics* **12,** 833–845 (2011).

197. Herzenberg, L. A., Parks, D., Sahaf, B., Perez, O., Roederer, M. & Herzenberg, L. A. The history and future of the Fluorescence Activated Cell Sorter and flow cytometry: A view from Stanford. *Clinical Chemistry* **48,** 1819–1827 (2002).

198. Lun, D. S., Sherrid, A., Weiner, B., Sherman, D. R. & Galagan, J. E. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biology* **10,** R142 (2009).

199. Mendoza-Parra, M.-A., Nowicka, M., Van Gool, W. & Gronemeyer, H. Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. *BMC Genomics* **14,** 834 (2013).

200. Gomes, A. L., Abeel, T., Peterson, M., Azizi, E., Lyubetskaya, A., Carvalho, L. & Galagan, J. Decoding ChIP-Seq peaks with a double-binding signal refines binding

peaks to single-nucleotide and predicts cooperative interaction. *Genome Research,* 1686–1697 (2014).

201. Hannah, R., Joshi, A., Wilson, N. K., Kinston, S. & Gttgens, B. A compendium of genome-wide hematopoietic transcription factor maps supports the identification of gene regulatory control mechanisms. *Experimental Hematology* **39,** 531–541 (2011).
202. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* **12,** 87–98 (2011).
203. Venet, D., Pecasse, F., Maenhaut, C. & Bersini, H. Separation of samples into their constituents using gene expression data. *Bioinformatics* **17 Suppl 1,** S279–87 (2001).
204. Lu, P., Nakorchevskiy, A. & Marcotte, E. M. Expression deconvolution: a reinterpretation of DNA microarray data reveals dynamic changes in cell populations. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 10370–5 (2003).
205. Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., Hastie, T., Sarwal, M. M., Davis, M. M. & Butte, A. J. Cell type-specific gene expression differences in complex tissues. *Nature Methods* **7,** 287–289 (2010).
206. Li, Y. & Xie, X. A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues. *BMC Bioinformatics* **14 Suppl 5,** S11 (2013).
207. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29,** 1083–5 (2013).
208. Zhao, Y. & Simon, R. Gene expression deconvolution in clinical samples. *Genome Medicine* **2,** 93 (2010).
209. Gaujoux, R. & Seoighe, C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study. *Infection, Genetics and Evolution* **12,** 913–921 (2012).
210. Qiao, W., Quon, G., Csaszar, E., Yu, M., Morris, Q. & Zandstra, P. W. PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Mi-

croenvironmental and Developmental Conditions. *PLoS Computational Biology* **8,** e1002838 (2012).

211. Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G. F., Selbig, J., Parida, S. K., Kaufmann, S. H. E. & Jacobsen, M. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics* **11,** 27 (2010).
212. Yip, K. Y., Cheng, C. & Gerstein, M. Machine learning and genome annotation: a match meant to be? *Genome Biology* **14,** 205 (2013).
213. Cherry, E. C. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America* **25,** 975–979 (1953).
214. McDermott, J. H. The cocktail party problem. *Current biology : CB* **19,** R1024–R1027 (2009).
215. Lee, T. W., Ziehe, A., Orglmeister, R. & Sejnowski, T. Combining time-delayed decorrelation and ICA: Towards solving the cocktail party problem. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* **2,** 1249–1252 (1998).
216. Hyvärinen, A., Karhunen, J. & Oja, E. Independent Component Analysis. *Analysis* **26,** 481 (2001).
217. Wang, Y. & Wang, D. Cocktail party processing via structured prediction. *Advances in Neural Information Processing Systems,* 224–232 (2012).
218. Choi, S. & Cichocki, A. Adaptive blind separation of speech signals: Cocktail party problem. *International Conference on Speech Processing,* 617–622 (1997).
219. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
220. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (2012).
221. Wickham, H. ggplot2: elegant graphics for data analysis. *Springer New York* (2009).

222. Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R, Dubourg, V, *et al.* Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

223. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. Integrative genomics viewer. *Nature Biotechnology* **29,** 24–26 (2011).

224. Himmelmann-Scientific Software & Dr. Lin - Development. HMM: HMM - Hidden Markov Models. *R Library* (2010).

225. Westreich, D., Lessler, J. & Funk, M. J. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* **63,** 826–833 (2010).

226. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9,** 215–216 (2012).

227. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA : Inferring intra-tumor heterogeneity from high- throughput DNA sequencing data. *Genome Biology* **14,** 1–41 (2013).

228. Gilmour, D. S. & Lis, J. T. In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. *Molecular and Cellular Biology* **5,** 2009–2018 (1985).