Universidad Autónoma de Madrid
Facultad de Ciencias
Departamento de Biología Molecular

Computational Approaches to Study Transcriptional Regulation in the Human Genome

PhD THESIS

Juan Manuel Vaquerizas Erdocia

Madrid, 2007

Universidad Autónoma de Madrid
Facultad de Ciencias
Departamento de Biología Molecular

Computational Approaches to Study Transcriptional Regulation in the Human Genome

Dissertation presented by Juan Manuel Vaquerizas Erdocia in candidacy of the degree of Doctor en Ciencias.

The work presented in this dissertation has been performed at the EMBL - European Bioinformatics Institute (Cambridge, UK) under the supervision of Dr. Nicholas M Luscombe and with the tutorship of Dr. Paulino Gómez-Puertas. Part of the work presented here was performed at the Centro Nacional de Investigaciones Oncológicas (Madrid, Spain) under the supervision of Dr. Joaquín Dopazo and has been included with the appropriate permission.

Acknowledgments

I would like to thank first and foremost my supervisor, Nicholas Luscombe, for his encouragement, enthusiasm and invaluable support during my research at the EBI. It has been an enormous privilege and a pleasure to be able to learn from you.

My thanks go also to Paul Bertone, the Luscombe group and my friends at the EBI and Cambridge for the wonderful time I have spent here. Richard Bourgon, Aswin Sai Narain Seshasayee, Annabel Todd and in particular Nicholas Luscombe deserve a mention for helping me to improve immensely the understandability of this dissertation.

Thanks also to Álvaro Mateos who has been an unfailing companion and friend through the highs and lows.

I also want to thank Sarah Teichmann, Asifa Akhtar and Jop Kind for the research collaborations and the long talks. Thanks also to Alberto de la Cruz, Manuel Serrano and Joaquín Dopazo who introduced me to molecular biology and scientific research.

Thanks also to all my friends in and outside the lab in Madrid, Valencia and elsewhere. In particular thanks to Diego, Ester, Míguel, Mario and Sara. I wish you were all around.

A special thanks goes to all my family (including future prospectives) and especially to my parents and brother for their love and unfailing support without whom this would not have been possible.

Finally, my most special thanks go to Lucía. It would have been impossible to enjoy my life so much without such an extraordinary true friend and partner. I will always be in-debt for your love and unconditional support.

Ok, now, from backstops... . . GO!

Cambridge, November 2007.

A mis padres

A Lucía

Abstract

It is essential for an organism's viability to ensure that the correct sets of genes are expressed in the right place and at the right time. There are several mechanisms by which cells regulate the amount of protein produced from genes under different conditions. One of the most basic is transcriptional regulation. By controlling the recruitment of RNA polymerase and associated factors to gene promoters, and the assembly of the transcription initiation complex, transcription factors regulate the transcriptional process, and therefore the expression of particular genes. A large number of human diseases are caused by malfunctions in transcriptional regulation, highlighting the importance of this system.

Here I present a computational study of transcriptional regulation in the human genome. First I identify and analyse the properties of 1,369 sequence-specific DNA-binding transcription factors in the human genome. We show that: (i) 80% of transcription factors belong to just three protein families, with the C_2H_2-Zn finger family being the most common; ii) 40% of factors are spatially clustered in specific chromosomal regions, and as a result may function in a co-ordinated manner; iii) transcription factors either function specifically in one or two tissues or ubiquitously across the whole body, giving rise to a two-tier organisation of global and local regulators; and iv) groups of transcription factors have arisen in the human lineage at key events during evolution (such as the appearance of mammalian organisms).

Secondly, I examine how sequence variation in the human genome, and in particular single nucleotide polymorphisms (SNPs), disrupt the normal function of the transcriptional regulatory system. I predict functional nucleotide sequence motifs (such as transcription factor binding sites and exonic splicing enhancers) inside or in the proximity of genes, and identify SNPs that overlap with them. Despite the simplicity of the approach, many of the predicted disruptive SNPs have been validated experimentally and have been associated with diseases.

Finally, none of the above results could have been obtained without the development of methods and tools required to perform a robust analysis of the data. In the past ten years the tandem development of high-throughput technology along with the sequencing of numerous genomes have produced a flood of data describing biological systems from a global perspective. These new data types often require special statistical or mathematical treatment in order to interpret them. I have devoted a large part of this dissertation towards creating methods and web-tools to analyse genomic data. These include approaches for: (i) cDNA microarray normalisation and quality control; (ii) identifying differentially expressed genes; (iii) building sets of genes with class prediction properties; (iv) performing transcription factor annotation of microarray experiments; (v) assessing the sensitivity and specificity of gene level measurements for Affymetrix GeneChips; (vi) detecting tissue-specific expression from microarray data; and (vii) detecting binding signal for ChIP-chip tiling arrays experiments.

Table of Contents

List of Figures

List of Tables

1. Introduction

"The discovery of regulator and operator genes, and of repressive regulation of the activity of structural genes, reveals that the genome contains not only a series of blue-prints, but a co-ordinated program of protein synthesis and the means of controlling its execution."

- Jacob and Monod 1961

Cellular life must recognise and respond appropriately to diverse internal and external stimuli. By ensuring the correct expression of specific genes, the transcriptional regulatory system plays a central role in controlling many biological processes ranging from cell cycle progression and maintenance of intracellular metabolic and physiological balance, to cellular differentiation and developmental time-courses.

Proteins are produced based on the information contained in the DNA through a series of very tightly controlled phases (Figure 1.1). The first step is called transcription, the process by which a section of the DNA molecule is used as a template to produce a precursor RNA messenger molecule. This is enzymatically processed to a mature mRNA molecule which acts as an intermediary for protein synthesis (Crick et al., 1961; Crick, 1970; reviewed

in Alberts et al., 2002).

The second phase is called translation, a process by which the mRNA is used as a template that determines the sequence of amino acids linked together in a polypeptide protein chain. The amino acid type is inferred from a translation code, in which triplets of nucleotide bases correspond to a particular amino acid or translation stop signal (reviewed in Lodish et al., 2004).

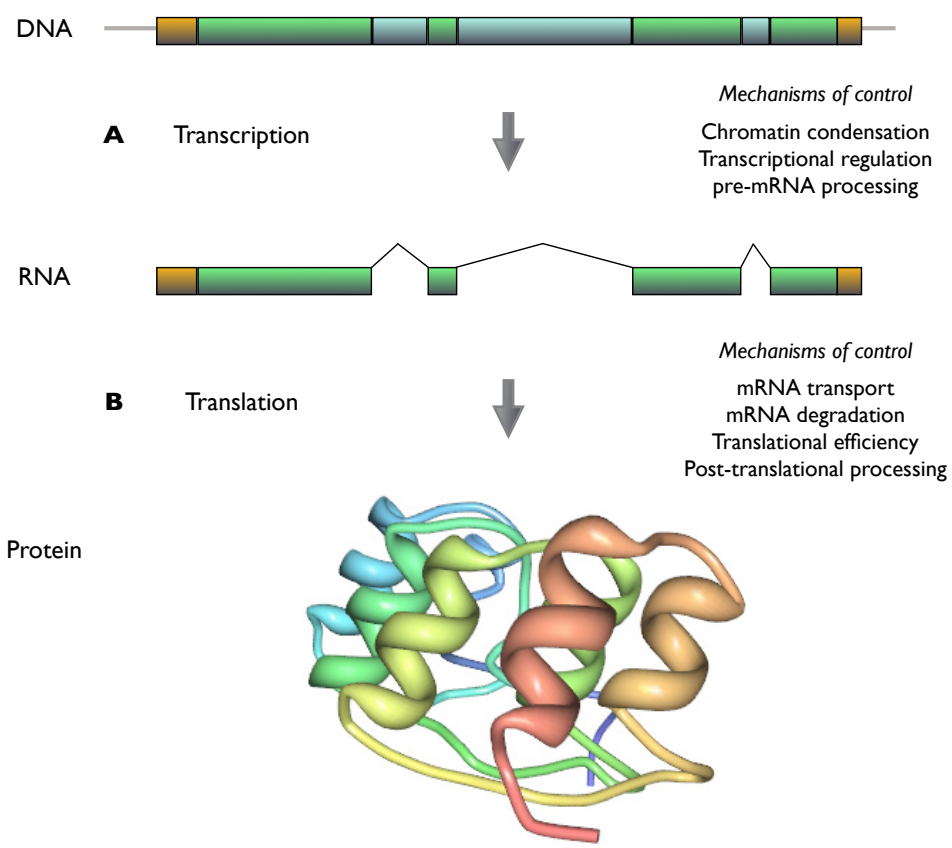

Figure 1.1 | Schematic diagram of protein synthesis from DNA. (A) Genes encoded in genomic DNA are transcribed to a mRNA molecule via transcription. This process is controlled by chromatin condensation, transcriptional regulation mediated by transcription factors and pre-mRNA processing among other mechanisms. (B) Mature mRNA molecules are used as templates to produce amino acid chains via translation. This is regulated by the mRNA transport outside of the nucleus, mRNA degradation, translational efficiency and post-translational processing.

The newly manufactured protein chain folds into a three-dimensional structure, and can be processed further in a third phase via post-translational modifications: different chemical reactions can modify the protein structure and therefore its activity or cellular localisation (reviewed in Lodish et al, 2004; Walsh and Jefferis, 2006).

The amount of protein and its activity are controlled through multiple mechanisms at the transcriptional, translational and post-translational stages. The transcriptional stage includes chromatin condensation, polymerase recruitment or inhibition and RNA processing. During the translational and post-translational phases mechanisms include mRNA export from the nucleus, RNA degradation, translation efficiency, and post-translational modifications (Darnell, 1982). Although all these mechanisms contribute to determining protein concentration and activity, transcriptional regulation is one of the most important. It controls whether or not a gene is expressed and at what level, and therefore the total amount of mRNA that will feed the rest of the process.

In this thesis, I apply genomic and computational biology techniques to further our understanding of transcriptional regulation in humans.

1.1 How transcription is controlled

RNA polymerase is the enzyme responsible for the production of RNA molecules from a DNA template. Although there are variants of this enzyme between prokaryotes and eukaryotes, the molecular mechanisms of transcription are quite conserved and well understood (Bushnell el at., 2004; Cramer et al., 2001; Gnatt et al., 2001; reviewed in Boeger et al., 2005).

Transcription initiates preferentially from specific locations in the upstream region of each gene. These are called gene promoters and are locations where the polymerase and other associated proteins assemble the transcription pre-initiation complex (PIC) (Gross and Oelgeschlager, 2006;

Smale and Kadonaga, 2003). RNA polymerases do not display sequence specificity in their DNA-binding; however in order to control transcription of specific genes they must bind preferentially to certain promoters (Brodsky et al., 2005).

This is of particular importance in higher eukaryotes, as a great majority of their genomes is non-coding (eg, around 98% of the human genome; International Human Genome Sequencing Consortium, 2004). How does the RNA polymerase detect those genes that should be expressed?

Selectivity for certain promoters is achieved through a set of proteins called transcription factors, which are responsible for recruiting the RNA polymerase to specific promoters and activating or repressing the transcriptional activity of genes based on the cell's requirements (Figure 1.2). There are two broad classes of transcription factors:

i) General transcription factors. These are responsible for RNA polymerase assembly in core promoters and are associated with the pre-initiation, initiation, elongation and termination of transcription. The general transcription factors are sufficient for initiating basal levels of transcription in eukaryotic cells (Hampsey, 1998; Thomas and Chiang, 2006).

ii) Non-general or specific transcription factors. These are regulators that activate or repress transcription in different conditions or cell types. These proteins contain a DNA-binding domain that recognises specific binding site sequences or motifs in the DNA. They control the recruitment of the polymerase and general transcription factors thereby promoting or repressing transcription (Kadonaga, 2004). Additionally, they are sometimes associated with the recruitment of chromatin remodelling complexes, which in turn affect expression levels by changing the DNA conformation.

The protein-DNA interaction interface has been extensively studied (Jones et al., 1999; Marmorstein et al., 1992; Luscombe and Thornton, 2002; Suzuki et al., 1995; reviewed in Harrison, 1991; Luscombe et al., 2000) and more than 200 different DNA-binding protein domains have been identified so far (Mulder et al., 2007). The most common mechanism for sequence recognition

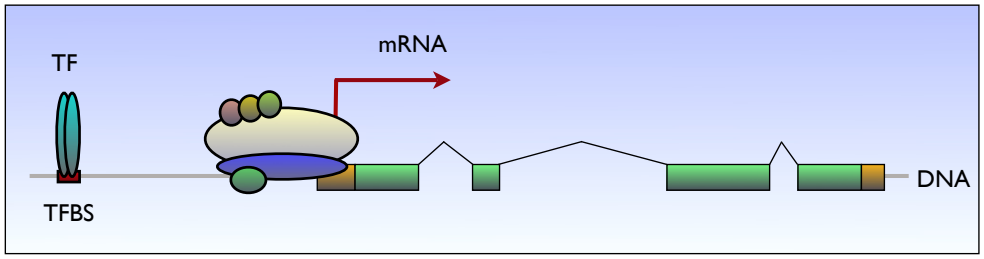

Figure 1.2 | Transcription factor mediated transcriptional regulation. The diagram represents the transcriptional regulatory mechanisms that control gene expression levels in eukaryotic organisms. The RNA polymerase II (light yellow) binds to the promoter of a gene along with the general transcription factors (small sub-units surrounding the polymerase) recruited by sequence-specific transcription factors (light blue) recognising transcription factor binding sites (TFBS).

involves the interaction of an alpha-helix of the DNA-binding domain with the DNA major groove (Luscombe et al., 2000).

Many other proteins also affect transcription without binding directly and specifically to DNA, such as co-factors and histone modifiers. Although many studies group these proteins with direct DNA-binding factors, I have excluded them from my work here, as the mechanism of transcriptional regulation is very different.

The importance of transcriptional regulation is underlined by the large number of diseases caused by a breakdown in the system. These include cancer (Darnell, 2002), and developmental (Boyadjiev and Jabs, 2000) and neurodegenerative diseases (Steffan et al., 2000). Therefore, it is very important to characterise and understand the different mechanisms that cells employ to regulate transcription, as well as the effects resulting from changes in the underlying DNA sequence, through mutations and polymorphisms.

1.2 Genomics

Molecular biology has traditionally been based on a reductionistic approach in which complex biological systems are dissected into their simplest components in order to study them in great detail. Once the individual components are characterised they are then placed together in an attempt to explain the entire system. This method has worked very well in establishing the molecular function of many genes and deciphering the causes of diseases such as phenylketonuria (reviewed in Eisensmith and Woo, 1992), and haemophilia (reviewed in Bowen, 2002). However, the method relies on components retaining the same functionality in isolation and within a system. This is known to be untrue in many cases, as interactions with other components introduce new capabilities (Kitano, 2002). This is highlighted by the limited success of molecular biology in understanding systems such as cancer or development which involve the combined activity of numerous genes.

A complementary approach is to study biological systems as a whole, using a genomic approach. Genomics can be defined as the analysis of the entire genome of an organism and the integration of the functionality of each component to create a global view of the system. This has required the development of new technologies in order to scale molecular biology techniques to work at much larger scale. Genomics has revolutionised the way in which we approach biological research as it allows us to evaluate the behaviour of whole systems, and so describe them using general principles.

An example of the way in which genomics has overturned our previous understanding of biology is the study of mammalian promoters (Carninci et al., 2006). The classical view of promoters derived from studies in prokaryotes, whose genes tipically include an AT-rich region called TATA-box, situated 30 base pairs upstream of the transcription start site, at which the TATA-box binding general transcription factor (TBP) binds and assembles the pre-initiation complex (Alberts et al., 2002). Recent genomic studies, however, have shown that the majority of gene promoters lack a TATA-box and

that this is not a general mechanism for transcriptional initiation in higher eukaryotes. A further surprise is that non-TATA-box promoters appear to have multiple transcription start sites (see Sandelin et al., 2007 for a review). These results clearly highlight the importance of genomic approaches in allowing us to make general conclusions about systems without bias from individual observations.

While powerful, new genomic approaches create challenges not previously faced by biologists: the amount and complexity of the data demands robust mathematical and automatic treatment in order for them to be interpreted sensibly. This associates genomics unequivocally with bioinformatics and computational biology. Purely genomic analyses, however, cannot replace the traditional, careful examination of single components; results generated by computational approaches must be tested experimentally. Therefore, collaborations between experimentalists and bioinformaticians are essential as both approaches complement each other.

1.3 High-throughput data

The sequencing of more than 500 genomes (as of May 2007) including those of human, mouse, chimpanzee or yeast, represents a milestone in our biological knowledge (Lander et al., 2001; The Chimpanzee Sequencing and Analysis Consortium, 2005; Venter et al., 2001; Waterston et al., 2002). However, the availability of these sequences is only the initial step towards understanding the functionality encoded in them and many downstream analyses are still necessary.

First, gene finding algorithms were developed (reviewed in Burge and Karlin, 1998) which allowed us to determine the full set of components of a genome. Next, the identification of genes fuelled the development of high-throughput technologies such as microarrays, yeast-two-hybrid assays and RNAi screenings that allow us to gain information about their functionalities in different systems. Most of these techniques are based on traditional

molecular biology assays for single gene analysis that have been scaled up to analyse thousands of genes at once.

1.3.1 Gene expression microarrays

The level of gene expression for single genes can be measured using the Northern blot. Here, single-stranded radioactive DNA molecules with particular sequences matching the gene of interest are used to measure the amount of transcripts of that gene in nuclear extracts. One of the first high-throughput technologies to be developed was microarrays (Schena et al., 1995), which allowed the simultaneous interrogation of thousands of gene expression levels at once. This is achieved thanks to the availability of the genes' sequences, which allow us to design specific probes assessing the level of expression at a genomic scale. These probes are physically attached to a device that allows us to track their individual signal.

The use of microarrays was pioneered for S. cerevisiae to measure global changes in gene expression for diverse cellular conditions (Cho et al., 1998; DeRisi et al., 1997; Lashkari et al., 1997; Spellman et al., 1998; Wodicka et al., 1997) and to identify abnormally expressed genes in human cancer samples (DeRisi et al., 1996).

The scope of microarray use has greatly expanded since then, and microarrays are now routinely employed to determine the outcome of particular diseases (van 't Veer et al., 2002), classify samples in different groups based on their expression patterns (van de Vijver et al., 2002), explore expression across the whole genome (Bertone et al., 2004), detect sequence variation between individuals of the same species (Janne et al., 2004), or simply measure the expression levels of genes in normal and healthy tissues (Su et al., 2004).

Despite the individual protocol adjustments for different methods, microarrays can be broadly classified into two major groups: (i) printed oligonucleotide microarrays, and (ii) complementary DNA (cDNA) microarrays.

i) Oligonucleotide microarrays. These are manufactured using a technique called photolithography in which single-stranded nucleotide sequences are synthesised directly on the glass slides. The oligonucleotide probes are usually between 25 and 60 bases long. Gene expression levels are measured by extracting and fragmenting mRNA from cellular samples, and then transforming the RNA into biotinylated single-stranded complementary DNA sequences that are hybridised to the corresponding microarray probes. The biotinylated sequences are then stained using a fluorescent streptavidin-phycoerythrin antibody and the expression levels measured using a laser-based scanner. The most commonly used arrays of this type are the Affymetrix GeneChips®.

ii) cDNA microarrays. Instead of using short oligonucleotides, these microarrays use long nucleic acid sequences as probes. The nucleic acids are manufactured by cloning the sequences into bacterial libraries and amplifying them by PCR. Probe lengths in this case vary from several hundred to thousands of base pairs. Once synthesised the probe sequences are physically attached to a glass slide using a robotic spotter or inkjet device. These microarrays are most commonly used for comparing hybridisations between two different mRNA samples, each labelled with a different fluorescent dye attached to the transformed cDNA molecule. The microarrays are then scanned at two different wavelengths to measure the amount of hybridisation from each sample. Therefore, these microarrays are analysed as a comparison between the two samples analysed. These types of microarrays have been widely manufactured in-house using robotic spotters.

Each technology has its benefits and drawbacks: recent studies have compared available technologies and suggested that commercial oligonucleotide microarrays, owing to their standardised manufacturing process and experimental protocols, and the higher probe density provide more reliable results. However it has also been reported that one the biggest reasons for discrepancies among platforms and laboratories arises from differences in post-experimental data processing and analysis (Bammler et

al., 2005; Irizarry et al., 2005; Larkin et al., 2005).

1.3.2 Chromatin immunoprecipitation followed by microarray hybridisation

Chromatin immunoprecipitation is a long-standing technique to detect protein-DNA interactions. Cells are treated with formaldehyde to cross-link the proteins and DNA, so stabilising these complexes in in vivo conditions. The cellular components are extracted and sonicated to break down the chromatin into short fragments of around 200-500 bases in length. The sample is then treated with antibodies against specific transcription factors to immunoprecipitate the protein-DNA complexes. The DNA fragments are then dissociated from the proteins and assayed via Southerns or sequenced to identify the original DNA-binding sites.

The technique is now frequently coupled with microarrays to determine DNA-binding sites on a larger scale (Boyer et al., 2005; Cawley et al., 2004; Horak et al., 2002a,b; Iyer et al., 2001; Martone et al., 2003; Ren et al., 2000). The immunoprecipitated DNA is hybridised to microarrays containing probe sequences for intergenic regions. For smaller genomes it is possible to represent all intergenic regions; however for complex organisms such as mouse and humans, microarrays are usually restricted to promoter regions. The sample is compared against a control of genomic DNA or a mock-immunoprecipitation. The analysis is similar to that for gene expression experiments (see below), except that enrichment is expected only for the immunoprecipitated sample.

1.3.3 Tiling arrays

Continuous improvements to the microarray production process have allowed manufacturers to increase the probe density of slides from several thousand to several millions. This makes it possible to design microarrays that cover not only specific genomic loci but entire genomes. These tiling arrays — so-called because probes "tile" across the whole genome — allow us to interrogate transcription for the genome in an unbiased fashion. Using

tiling arrays, several studies have detected unexpected regions of the human genome that appear to be actively transcribed (Bertone et al., 2004; Cheng et al., 2005; Kapranov et al., 2005). This contributed to the now-accepted view that the human genome contains thousands of non-coding RNA genes, many of them transcribed from the anti-sense strand of known protein coding genes. Tiling microarrays have also been used for chromatin immunoprecipitation experiments revealing widespread protein binding not limited to promoters (Lee et al., 2006).

1.3.4 Re-sequencing techniques

Sequencing technologies have improved in the past 30 years resulting in a dramatic rise in throughput. The increased capacity allows us to sequence many individuals of the same species as well as many other model organisms, in order to measure both intra- and inter-species genomic variation.

i) Intra-species variation. Although we have developed the tendency to conceptualise an organism's genome as having a universal and fixed genomic sequence, individuals in fact never share the exact same DNA sequence (unless they are identical twins). Such differences enable organisms to diversify and therefore adapt to changes in the environment more easily as a population. Different molecular mechanisms lead to intra-species variations, including recombination, segmental duplications, insertions or deletions, translocations and mutations. The most frequent types of variation for humans are point mutations, which consist of single nucleotide changes. These mutations propagate in a sub-population only if they are non-fatal and occur in the germ line. Mutations are officially classified as a single nucleotide polymorphism (SNP) if they arise in more than 1% of the population (Chakravarti, 2001).

More than 12 million SNPs have been identified so far (dbSNP v.126; Sherry et al., 2001) through international collaborations such as the HapMap project (The International HapMap Consortium, 2003) which sequenced diverse human populations. These data are publicly available and enable us to study the contribution of genetic variation to differences in drug response

or disease outcomes.

ii) Inter-species variation. The availability of the genome sequences of other model organisms provides valuable data describing how evolution has progressed over millions of years. Using comparative genomic techniques we can detect species-specific features as well as those that are conserved (Bejerano et al., 2004; reviewed in Boffelli et al., 2004; Ureta-Vidal et al., 2003). Comparative genomics is an important area of research with many applications, but it is not the focus in the current thesis.

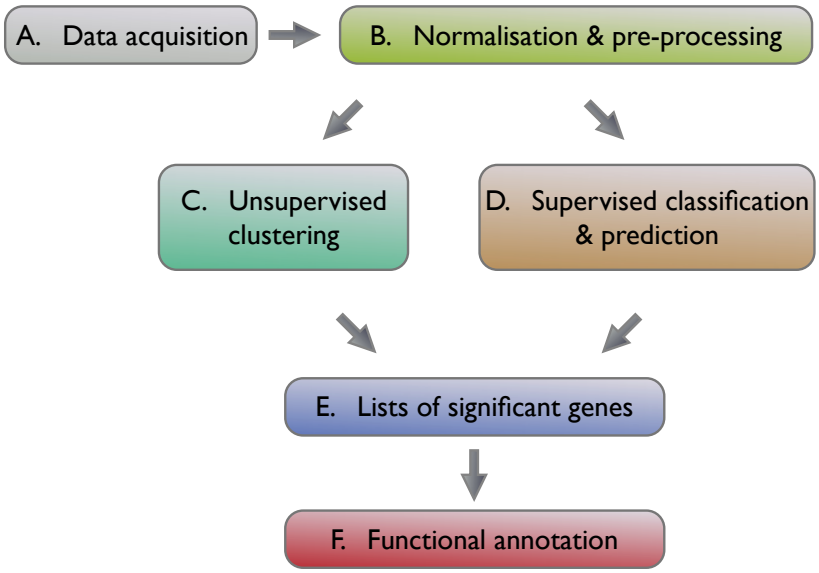

Figure 1.3 | Microarray data analysis flowchart. The diagram shows a common workflow for microarray data analysis. After (A) scanning, the microarray data must be (B) normalised and pre-processed to remove systematic non-biological variation from the data. The next steps consist either of (C) unsupervised clustering to detect genes or samples with similar expression patterns, or (D) perform supervised classification techniques to identify differentially expressed genes or probes between groups of samples. Both approaches return (E) lists of genes that can be subsequently (F) functionally annotated using additional biological information such as genomic location or GO annotations.

1.4 Gene expression microarray data analysis

Standard microarray experiments produce expression data for thousands of genes. Owing to the large quantity of data and the level of noise found in them, we need robust computational tools and statistical techniques to analyse and interpret the results. The most common workflow for analysing microarray data comprises: (i) data acquisition; normalisation and pre-processing; (ii) unsupervised class-discovery; (iii) differential gene expression, supervised classification and prediction; and (iv) functional interpretation of the results (Figure 1.3). Of course the details of the workflow differ depending on the biological question being addressed.

i) Data acquisition, normalisation and pre-processing: This is the most variable step between microarray platforms. A scanner is used to detect the fluorescence for each probe in order to measure the amount of hybridisation: a single wavelength laser is used for oligonucleotide microarrays, or two lasers of different wavelength for cDNA microarrays.

The scanned images are converted into numerical read-outs of fluorescence levels with software such as GenePix. Ideally these values should directly reflect the expression level for genes, however fluorescence values also contain biases because of the manufacturing or handling procedures employed (Quackenbush, 2002). To remove sources of systematic bias, several methods for data normalisation have been developed (Huber et al., 2002; Irizarry et al., 2003; Smyth and Speed, 2003). Methods vary between microarray platforms, although the most common ones are based on the assumption that the majority of genes do not change in expression between samples.

ii) Unsupervised class-discovery or clustering: The aim of this step is to find classes of genes or samples within the same experiment that behave in a similar fashion in their expression. The algorithms do not use any prior knowledge about the class membership of samples and therefore the detection of classes or clusters is based on the similarity in the behaviour of genes. Clustering was among the first analysis techniques employed for

microarray data and has been used extensively, in, for example, a number of seminal research projects which identified genes with similar expression patterns across the yeast cell cycle and sporulation time courses (Chu et al., 1998; Eisen et al., 1998; Spellman et al., 1998).

iii) Differential gene expression, supervised classification and class prediction: This is an alternative approach to identify genes displaying interesting gene expression in which we utilise prior knowledge about the samples being analysed (eg, class membership). Most frequently we apply statistical techniques to highlight features that differ between distinct sets of samples that might be used as classifiers or predictors for class membership. Differential gene expression analysis has been widely employed to select genes behaving differently between various types of cancer, and therefore provide a list of candidate genes for follow-up studies (Clark et al., 2000). Similar approaches have been applied to numerous other biological problems ranging from detecting genes that are responsible for the pluripotency and self-renewing properties of the stem cells (Ivanova et al., 2002; Ramalho-Santos et al., 2002) to the study of co-ordinated gene responses in plants (Schenk et al., 2000). The gene lists —often referred to as the 'signature' for a biological condition— can in turn be used to classify samples or predict the future outcome of the process under investigation (Ramaswamy et al., 2003; van 't Veer et al., 2002). Such gene signatures have been used to develop commercial microarray solutions; for example, the Mammaprint® (`http://www.agendia.com/common.asp?id=80`) has been used to evaluate the risk of breast cancer relapse after surgery.

iv) Functional interpretation of the results: Almost all techniques above output lists of genes related to a particular outcome, behaviour, or temporal expression pattern. To gain insight into the biological implications of these results we can incorporate some of the vast amounts of biological information available from different databases. The most common types of analyses look at the functional annotations from the Gene Ontology database (GO; Ashburner et al., 2000), gene structure and protein domains from InterPro (Mulder et al., 2007), cellular pathways from databases such as

KEGG (Kanehisa et al., 2006), and putative transcription factor-binding sites upstream of differentially expressed genes from databases such as JASPAR (Sandelin et al., 2004), CisRED (Robertson et al., 2006) and TRANSFAC (Wingender et al., 2000).

1.5 Genomics and transcriptional regulation

Transcriptional regulation has been extensively studied from a genomic perspective in recent years. Most experimental studies have focused on yeast, as it is a relatively simple organism that is easy to experiment on, and yet has high levels of conservation relative to more complex organisms. These have used microarrays measuring gene expression under diverse conditions such as stages of the cell cycle, sporulation, diauxic shift, DNA damage, and stress response (Cho et al., 1998; Chu et al., 1998; DeRisi et al., 1997; Gasch et al., 2001; Gasch et al., 2000). There are now over 120 microarray experiments for yeast (as of May 2007) in the ArrayExpress database (Parkinson et al., 2005).

To complement gene expression studies and computational approaches, high-throughput experimental methods have been developed to detect transcription factor binding sites in vivo. In yeast, these studies consisted of ChIP-chip experiments (Horak et al., 2002a; Lee et al., 2002), where the target genes of multiple factors, such as SBF or MBF, the major transcription factors controlling the G1/S phase transition, were determined. This allowed us to recreate the regulatory network and to show which parts of it are active under specific conditions (Luscombe et al., 2004).

In human however, the majority of studies involving genomics have focused on detecting differences between tumours and healthy tissues (Bhattacharjee et al., 2001; DeRisi et al., 1996; Golub et al., 1999; Ramaswamy et al., 2001). As experimental techniques have improved, however, there is now more and more interest in deciphering the logic controlling the transcriptional regulatory processes. The main reason behind this interest is the relationship between the malfunctioning in these control mechanisms and major diseases

such as cancer, and developmental or neurological disorders. Most of current research is directed towards understanding how combinations of multiple transcription factors regulate different processes (Lemon and Tjian, 2000). In order to achieve this goal, studies have identified different locations in the genome where particular transcription factors bind to regulate the expression of particular genes. The computational approaches developed to carry out this task vary from direct sequence analysis, ie, identification of binding site consensus sequences in the genome, to phylogenetic footprinting, where evolutionary information is integrated in the analysis to restrict analysed regions to those conserved between particular species (Tompa et al., 2005). Unfortunately, these computational approaches generate results with false positive rates in the order of thousands for each right prediction (Wasserman and Sandelin, 2004).

These computational studies have been complemented in higher organisms with experimental approaches employing techniques such as ChIP-chip (Martone et al., 2003), DamID (Greil et al., 2006) or the gateway-compatible one-hybrid system (Deplancke et al., 2006), although the outcome of these results in mammalian systems is still unclear mainly due to the quality of the data and the difficulties in their interpretation. It has not been until very recently, coinciding with the improvement on the data quality obtained from ChIP-chip experiments for human, that work has been devoted to understand the transcriptional regulatory network in human. This has involved the analysis of the binding sites for several key developmental transcription factors, such as OCT4, NANOG, or the polycomb group (Boyer et al., 2005; Boyer et al., 2006; Lee et al., 2006). Despite these efforts and interest in understanding transcriptional regulation, our knowledge of this regulatory system in higher organisms is still very basic.

One of the main reasons for our lack of knowledge is that, seven years after the initial publication of the human genome, we still do not have a high-quality, comprehensive list of transcription factors in the human genome. The original human genome papers estimated between 200 and 300 general transcription factors and between 2,000 and 3,000 sequence-specific DNA-

binding transcription factors (Lander et al., 2001; Venter et al., 2001). Since then, only two studies have tried to characterise the repertoire of human transcription factors. In the first, Messina et al. (2004) used the Transfac database and proteins annotated as transcription factors in GO to build a set of transcription factors that they subsequently used as a seed for hidden Markov models searches against the human transcriptome. The results from this study detected a human transcription factor repertoire of 1962 members. A superficial examination of the list reveals that some proteins are mis-annotated as transcription factors. The reason behind this observation is that whereas some DNA-binding domains are very strict and are only found and used by transcription factors, other DNA-binding domains are very promiscuous and occur in many other proteins with different functions.

A similar approach for finding transcription factors is followed by Kummerfeld and Teichmann in the DBD database (Kummerfeld and Teichmann, 2006). There, a set of manually curated DNA-binding domains from Pfam and SUPERFAMILY were used to predict transcription factors for different species. They filtered out those promiscuous DNA-binding domains in order to lower the false positive rate of detection. Although this second approach is much more accurate than the one mentioned above, the stringent DNA-binding filtering, necessary to keep a low number of false positives, decreases the level of coverage for higher organisms.

1.6 Aim of this thesis

The work I present in this thesis aims to increase our level of knowledge about transcriptional regulation in human through the analysis of genomic data using computational and bioinformatic approaches. This is achieved by a global-scale analysis of genomes, gene expression data and sequence variation within and among species.

First I present statistical methods and software tools I developed to analyse diverse genomics data. As new data types emerge, it is necessary to

develop new techniques to analyse them. It is equally important to promote the use of these methods via the publication of software tools that are readily accessible. Many of these tools have been integrated into the Gene Expression Pattern Analysis Suite (GEPAS; Herrero et al., 2003a; Vaquerizas et al., 2005), which is a web-based microarray data analysis platform.

Second I present the identification and characterisation of the human repertoire of transcription factors. As described earlier, the lack of a high-confidence list of transcription factors impedes the analysis of human transcriptional regulation at a global scale. This work aims to fill this gap by providing a gold standard set for future genomic analyses of transcriptional regulation. The functional characterisation of these transcription factors, although performed at a basic level, provides insights regarding their regulatory functions and the global organisation of transcriptional regulation in the human genome.

Third I integrate sequence variation data within humans to detect nucleotide changes that potentially affect transcriptional activity; these are typically SNPs that modify gene promoters or mechanisms associated with mRNA processing. Sequence variation plays an important role in determining the sensitivity of particular individuals in adapting to the environment and understanding the effect on normal cellular regulation is an initial step in this direction. These insights will advance our understanding of multigenic diseases and preventive medicine.

2. Objectives

Genomics aims to examine entire organisms in order to extract general principles governing the function of different biological systems. Here I describe work that utilises genomic data to increase our understanding of transcriptional regulation in the human genome.

Genomic datasets, such as microarray expression data, describe aspects of the cellular regulatory process and they require robust statistical treatment in order to interpret them in a meaningful fashion. Currently much research focus is directed at the analysis of transcription factor binding sites in the human genome. However, we have little knowledge of the transcription factors involved, or the effects that mutations have on their functions.

The objectives of this PhD are:

1. To develop methods and software tools for analysing high-throughput genomic data to increase our understanding of transcriptional regulation in humans.

2. To identify and analyse the function of sequence-specific DNA-binding transcription factors in the human genome.

3. To predict single nucleotide polymorphisms that disrupt regulatory processes.

Objectives

3. Materials and Methods

One of the major tasks in genomics and bioinformatics is the integration of disparate data describing a particular biological system. By incorporating information from different viewpoints, we can construct a global picture of systems in term of their components and functionality, thus allowing us to extract general principles governing their behaviour.

The work presented in this thesis is purely computational, and the data utilised have been collected from numerous public sources. These datasets are very large, complex and diverse, ranging from genome sequences and protein structure information to gene expression measurements.

It is important to use robust methods when integrating and analysing genomic datasets, so that conclusions drawn from them are correct and biological meaningful.

In this chapter I describe the data, methods, programming languages and computational resources used for this thesis.

3.1 Databases and datasets

This section describes the databases and datasets used in this thesis, which are summarised in Table 3.1.

3.1.1 Genome sequence databases

3.1.1.1 Ensembl

Ensembl is an automatic pipeline for genomic annotation developed jointly at the Sanger Institute and the European Bioinformatics Institute (Hubbard et al., 2007). It contains the DNA sequences, annotations, sequence variation,

Table 3.1 | Data types and resources.

Type of data	URL	Reference
Genome sequences		
Ensembl	http://www.ensembl.org	Hubbard et al., 2007
Ensembl Compara	"	"
Ensembl Variation	"	"
Inparanoid	http://inparanoid.sbc.su.se	O'Brien et al., 2005
Protein sequences		
InterPro	http://www.ebi.ac.uk/interpro	Mulder et al., 2007
IPI	http://www.ebi.ac.uk/IPI	Kersey et al., 2004
Gene expression		
ArrayExpress	http://www.ebi.ac.uk/arrayexpress	Parkinson et al., 2005
GNF dataset	http://symatlas.gnf.org/SymAtlas	Su et al., 2004
Unigene	http://www.ncbi.nlm.nih.gov/UniGene	Wheeler et al., 2007
AML/ALL dataset	http://www.broad.mit.edu/cancer/pub/all_aml	Golub et al., 1999
Transcription factors		
Transfac	http://www.gene-regulation.com	Wingender et al., 2000
DBD	http://www.transcriptionfactor.org	Kummerfeld and Teichmann, 2006
Messina et al. 2004	http://www.genome.org/cgi/content/vol14/issue10b	Messina et al., 2004
Sequence motifs		
ESEFinder	http://rulai.cshl.edu/tools/ESE	Cartegni et al., 2003
Gene functions and literature		
Gene Ontology	http://www.geneontology.org	Ashburner et al., 2000
Pubmed	http://www.ncbi.nlm.nih.gov/entrez	Wheeler et al., 2007

and comparative genomic data for 31 different eukaryotic organisms (version 43) including human, chimpanzee and mouse. I have extensively used Ensembl as the main source of genomic information, such as gene, transcript and protein sequences and identifiers, and as a system to visualise them. Currently Ensembl has a continuous data release cycle of two months. Thus I have used different versions of the database as new updates were released; version numbers are indicated where appropriate.

Ensembl is freely available via four online interfaces: (i) the Ensembl web site (`http://www.ensembl.org`); (ii) BioMart, a web-tool that allows users to query Ensembl; (iii) a public MySQL server; and (iv) specific Perl functions that perform automatic database queries. All interfaces provide access to the same data, and I have employed all four options depending on the task.

3.1.1.2 Ensembl Compara

Ensembl Compara is an extension of the main database that contains whole-genome sequence alignments among the 31 genomes using the Blastz-net (Schwartz et al., 2003) and Pecan algorithms. In addition, it provides homologous relationships between genes from different species: (i) orthologous genes that diverged after a speciation event; and (ii) paralogous genes, that duplicated before the speciation event. Homologues are obtained: (a) by identifying Blast reciprocal best-hits or multiple Blast reciprocal best-hits, which allows for many to many relationships; and (b) derived from syntenic genomic regions. Since v41 homologous relationships are also obtained through gene trees (see below). I have used these data in §4.2.7 of the thesis.

Gene trees

Phylogenetic gene trees generated by maximum-likelihood aim to represent the evolutionary history of genes. They provide a better way of detecting evolutionary relationships between genes in different species than approaches based only on sequence similarity searches. Ensembl Compara constructs gene trees by: (i) performing intra- and inter-species all-versus-all pair-wise sequence alignments; (ii) multiple alignments using MUSCLE

(Edgar, 2004); and (iii) reconciling resulting phylogenetic trees with the species tree (NJTREE - `http://treesoft.sourceforge.net/njtree.shtml`). As the gene duplication and speciation events are traceable, Ensembl Compara is able to differentiate between orthologues and paralogues. I have used this approach to assess the evolutionary conservation of human transcription factors in different species (§4.1.7).

3.1.1.3 Ensembl Variation

Ensembl Variation contains sequence variation information such as single nucleotide polymorphisms (SNPs), insertions, deletions, and genomic repeats. The database integrates data from dbSNP and Hapmap — the major sources for sequence variation data — and maps them onto other genomic features contained in Ensembl. I have used this database in §4.3.

3.1.1.4 Inparanoid

Inparanoid contains orthology assignments for 26 eukaryotic species (version 5.1; O'Brien et al., 2005). The database is based on inter- and intra-species pair-wise gene sequence alignments. An out-group is used to calculate a similarity score that is employed as a threshold to determine homologous genes. These are then grouped by species to distinguish between orthologues and paralogues. Versions 4 and 5 of the database were used in the work presented in this thesis (§4.2.7).

3.1.2 Protein databases

3.1.2.1 International Protein Index

The International Protein Index (IPI) database is a database that integrates non-redundant data describing eukaryotic proteomes from primary protein databases (Kersey et al., 2004). Data are collected from: SwissProt, TrEMBL, RefSeq, Ensembl, H-InvDB and Vega. IPI also contains protein domain assignments for InterPro entries (see below). We used the IPI database to identify human transcription factors (§4.2.1).

3.1.2.2 InterPro

InterPro is a protein sequence annotation database (Mulder et al., 2007). It integrates information from external protein sequence and structure databases, such as PFAM (Finn et al., 2006) and SUPERFAMILY (Gough et al., 2001), to define protein domains and motifs. There is a hierarchical classification of the InterPro entries into families, domains, repeats and sites depending on the relationship described; several domains can form part of a family, and domains can contain several types of repeats.

Protein domains — represented by sequence signatures — are often associated to specific molecular functions such as ligand-binding. InterPro provides description and literature citations of the functions of each entry. These are extensively used o predict gene functions through the presence or absence of these domains in the sequence and it is the primary automated functional annotation by the Gene Ontology Annotation project (GOA). InterPro combines motif signatures from multiple databases; the predictions provide greater coverage than the individual data sources alone.

I used InterPro as the source of domains and families to identify sequence-specific DNA-binding proteins (§4.2.1).

3.1.3 Gene expression databases and datasets

3.1.3.1 ArrayExpress

ArrayExpress is a major repository for microarray data hosted at the European Bioinformatics Institute (Parkinson et al., 2005), which provides public access to more than 1,770 user submitted datasets (May 2007). As the database complies with the MIAME standard (Minimum Information About a Microarray Experiment; Brazma et al., 2001), experiments are described with enough detail to reproduce them. Most microarray studies performed before the MIAME standard do not include this information, making them less useful for follow-up studies. The processed microarray data described in this thesis have been submitted to ArrayExpress to make them accessible

to the scientific community.

3.1.3.2 Gene expression data from the SymAtlas Genome Novartis Foundation dataset (GNF dataset)

The SymAtlas dataset from the Genome Novartis Foundation provides gene expression measurements for human, mouse and rat tissues and cell lines, obtained using Affymetrix GeneChips (Su et al., 2004). I have used the human dataset, which contains data for 79 human healthy tissues, cell lines and tumour samples. Each hybridisation was performed using pooled samples from four or more individuals and repeated to provide two biological replicates for each tissue. Two microarray types were used: (i) the commercial HGU133a; and (ii) a custom array covering genes and EST sequences not present on the commercial array. Together they interrogate the expression of 44,775 different transcripts. A full list of tissues can be found at `http://wombat.gnf.org/samples/GeneAtlasv2_sample_info.html`. I use this dataset in §4.2.5 to characterise transcription factor expression.

3.1.3.3 Acute lymphoblastic leukaemia and acute myeloid leukaemia dataset from Golub et al.

The Golub et al. dataset (1999) measures gene expression for 6,817 human genes in 38 samples derived from patients with acute lymphoblastic leukaemia (ALL) or acute myeloid leukaemia (AML) using Affymetrix GeneChips. This was the first study to use microarrays for tumour sample classification. I have used this dataset as a test-set for the validation of analysis methods presented in §4.1.2 and §4.1.3.

3.1.3.4 MSL1, MSL3, MOF, H4K16, Nup153 and Mtor dataset from the Akhtar laboratory (EMBL)

This dataset contains unpublished data for ChIP-chip experiments for the following proteins related to the *Drosophila melanogaster* Dosage Compensation Complex (DCC): MSL1, MSL3, MOF, histone 4 lysine-16 acetylation (H4K16), Nup153 and Mtor. MSL1, MSL3 and MOF are components of the complex (Straub and Becker, 2007). MOF in particular is a histone acetyl-transferase,

whose activity can be assessed by measuring histone-4 lysine-16 acetylation. Nup153 and Mtor form part of the nuclear pore (Akhtar and Gasser, 2007). Chromatin immunoprecipitations were performed in three male and female biological replicates to measure DNA-binding of these proteins. In addition, an antibody against H4K16 was used to measure MOF activity. Chromatin immunoprecipitations were hybridised against the Affymetrix GeneChip Drosophila Tiling 2.0R Array. Genomic input DNA and immunoprecipitated histone-4 were also hybridised as a baseline for comparison. I have used this dataset with kind permission from Dr. Asifa Akhtar, to develop an analysis method for ChIP-chip data that allows us to detect unbiased binding signal for DNA-binding proteins with different protein-DNA affinities (§4.1.7).

3.1.3.5 Unigene

Unigene contains information about expressed sequence tags (EST) in numerous tissues and cell lines (Wheeler et al., 2007). The sequence reads are integrated and collapsed into Unigene clusters based on the overlap between ESTs, and associated to particular genomic loci. The presence of a Unigene cluster is indicative of gene expression. I used Unigene in §4.1.6 as an alternative source of expression data, in order to calibrate Affymetrix GeneChip results.

3.1.4 Transcription factor databases and datasets

3.1.4.1 Transfac

The Transfac database contains literature-extracted descriptions about transcription factors, target genes, binding sites and position weighted matrices for consensus binding sequences in multiple species (Wingender et al., 2000). I have used the position weight matrices (version 7.3 Professional) to identify potential transcription factor binding sites in eukaryotic promoters (§4.1.4).

3.1.4.2 DBD

DBD is a database of predicted DNA-binding transcription factors for 265 prokaryotic and eukaryotic genomes (Kummerfeld and Teichmann, 2006). Transcription factors are predicted using hidden Markov model searches of DNA-binding domains described in PFam and SUPERFAMILY in the coding sequences for each organism. I used DBD to compare the human repertoire of transcription factors obtained in §4.2.1.

3.1.4.3 Transcription factor dataset by Messina et al.

Messina et al. (2004) identified 1,962 human transcription factors by sequence similarity searches against protein domains that were annotated as DNA-binding domains. The authors used Transfac and a set of GO annotated transcription factors to build a seed-set of hidden Markov models, which were then used to search against the human transcriptome. I used the Messina dataset to compare with the transcription factor dataset that I obtained in §4.2.1.

3.1.5 Sequence motifs datasets

3.1.5.1 Motif-scoring matrices for exonic splicing enhancers

The serine/arginine-rich (SR) family of proteins are involved in mRNA splicing by recruiting two small nuclear ribonucleoproteins (U1 and U2AF) that recognise exonic splicing enhancer (ESEs) sequences found in exons. Several binding sites have been identified and used to create motif-scoring matrices describing consensus binding sequences (Cartegni et al., 2003). I have used these matrices to detect ESEs in the human genome (§4.3).

3.1.6 Other databases

3.1.6.1 Gene Ontology

Gene Ontology (GO) provides a defined and structured vocabulary to describe gene and gene products in terms of their molecular functionality, biological processes and cell localisation (Ashburner et al., 2000). The terms are organised in the form of a directed acyclic graph in which more specific terms derive from broader ones and where child terms can have several parents. Gene annotations for different species are manually curated from the literature, or electronically annotated using domain assignments. I used GO to assess the level of knowledge about transcription factors (§4.2.2).

3.1.6.2 PubMed/MEDLINE

MEDLINE (Medical Literature Analysis and Retrieval System Online) is an international database of life sciences literature including biology, biochemistry and medicine journals. PubMed is a web interface for MEDLINE that allows users to perform global searches of the available citation data (Wheeler et al., 2007). In addition to using PubMed as a general source for biological information throughout this thesis, I have employed it specifically to assess the level of knowledge for particular transcription factors in §4.2.2.

3.2 Data analysis methods

The large amount of data generated by high-throughput technologies in the last decade has necessitated the development of new analysis methods. Some of these have become standard procedures and are now used routinely. In this section I summarise the methods and approaches that I have employed throughout this thesis to integrate genomics data. In most cases I have used versions implemented in the R statistical package, or as stand-alone Perl programs.

3.2.1 Statistical methods

3.2.1.1 t-test

The t-test allows us to determine whether two sets of observations are different with a certain level of confidence. The standard Student's t-test compares the mean of two sets of observations under the assumption that data follow a normal distribution and the variance of the two populations are equal. I have used a non-parametric variant of the standard t-test, Welch's t-test, that does not require these assumptions. The t statistic is calculated using the following formula:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_1}}}$$

where X is the average expression, s the estimated variance and N the sample size for each group. This test was implemented as a method for selecting differentially expressed genes between two classes of samples; eg, cancer versus control sample (§4.1.2).

3.2.1.4 Wilcoxon test

The Wilcoxon test is a non-parametric statistical test equivalent to the t-test, to detect differences among variables between two groups by ranking the measurements and using the order to detect significant differences. This method was implemented in §4.1.3 to sort genes based on their differential gene expression between classes of samples.

3.2.1.2 Fisher's exact test

The Fisher's exact test compares the association between two variables in a 2 x 2 contingency table (Table 3.2).

Table 3.2 | 2x2 contingency table.

	condition A	condition A'	total
condition B	a	b	a+ b
condition B'	c	d	c + d
total	a + c	b + d	n

The probability of obtaining any data distribution is given by the formula:

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

and the significance of this probability is assessed by comparing it with all other probabilities that are calculated from the same contingency table. I used this test to assess statistical significance for categorical data, such as transcription factor annotations between sets of genes (§4.1.4).

3.2.1.3 F-ratio

The F-ratio is the ratio of the between group variances against the within group variances (Dudoit et al., 2002). It measures the degree of overlap between two distributions.

$$Fratio = \frac{\sum_i \sum_k I(y_i = k)\left(\bar{x}_{kj} - \bar{x}_{.j}\right)^2}{\sum_i \sum_k I(y_i = k)\left(x_{ij} - \bar{x}_{kj}\right)^2}$$

The method was implemented in §4.1.3 to sort genes based on their differential expression between classes of samples.

3.2.1.5 Permutation test

The permutation test determines the likelihood of an observation given a random sampling of the original distribution. The statistical significance is

calculated by counting the number of times that random samples have a more extreme value than the observed data:

$$pvalue = \frac{N_{extreme} + 1}{N_{permutation} + 1}$$

I have used the permutation test to obtain significance values for different observations, such as the occurrence of transcription factor clusters in certain genomic regions (§4.2.4).

3.2.1.6 Multiple testing correction procedures

Multiple testing correction procedures involves the recalculation of p-values for experiments where a particular test has been repeated multiple times. The statistical significance of each test cannot be used to detect significant observations for the entire dataset as the probability of finding significant observations increases with the number of preformed tests. This phenomenon is particularly important in genomics as often the number of observations is in the order of thousands (eg, microarray experiments).

Several mathematical techniques have been developed to deal with this problem. Here I have used the false discovery rate approach (FDR) (Benjamini and Hochberg, 1995). The method adjusts the expected proportion of incorrectly rejected null hypothesis from the rejected hypothesis and it returns a q-value, which is the minimal FDR of a particular observation at which the test is significant. FDR is less conservative than other multiple testing adjustment approaches, such as the Bonferroni approach that requires a very strong evidence to reject the null hypothesis. I have used FDR instead of other approaches because changes in genomics and gene expression can be subtle and therefore not strong enough to resist other multiple testing corrections. In particular I have used FDR correction for: (i) adjusting p-values in differential expression assignments (§4.1.2); (ii) adjusting p-values in the detection of significant transcription factor clusters (§4.2.4); and (iii)

adjusting p-values for over- or under-representation of transcription factor binding sites in sets of genes (§4.1.4).

3.2.1.7 Hierarchical clustering

Classical hierarchical clustering is a mathematical grouping algorithm that partitions a dataset into sub-groups of similar members. The similarity between entries is calculated using a distance measurement. I have performed hierarchical clustering to represent: (i) groups of genes or tissues that share common expression patterns (§4.2.6); and (ii) to group different species based on the presence or absence of orthologues of human transcription factors (§4.2.7). In both cases I used Euclidean distance as the similarity measure.

3.2.1.8 ROC curves

Receiver operating characteristic (ROC) curves are graphical plots that represent the sensitivity — the proportion of true positives among all true samples — versus the specificity — the proportion of true negatives among all negatives — for a particular analytical approach. When multiple models or methods are available, ROC curves can be used to identify the best approach for analysis. I have used ROC curves to set the best gene expression threshold for Affymetrix data (§4.1.5).

3.2.1.9 Propensity

Propensity values were originally used to detect over- or under-representations of amino acids in particular positions of protein secondary structures, eg, in alpha-helices, using the following equation:

$$propensity = \frac{x_{ij} \sum_i \sum_j x_{ij}}{\sum_i x_{ij} \sum_j x_{ij}}$$

I have adapted this for the calculation of tissue-specificity of transcription factor expression (§4.1.6)

3.2.2 DNA and protein sequence analysis methods

3.2.2.1 BLAST

The Basic Local Alignment Search Tool (BLAST) is a sequence alignment software that allows the pair-wise comparison of protein and nucleotide sequences (Altschul et al., 1990). A typical BLAST search involves the matching of a query sequence in a database containing billions of sequences. In order to maximise the computational speed, the algorithm first finds exact matches of small segments of the query sequence and then extends them using a particular penalty for mismatches. When a high-scoring, ungapped sequence is found BLAST then produces a gapped alignment using a variant of the Smith-Waterman algorithm. In this thesis I have used BLAST to identify human specific transcription factors (§4.2.7).

3.2.2.2 Match

Match is a program accompanying the Transfac database to detect potential transcription factor binding sites (Kel et al., 2003). The program uses position weight matrices derived from experimentally determined binding sites to identify motif in a given stretch of DNA sequence. The user can choose among three predefined thresholds that quantify the significance of a motif: (i) minimising false positives by using similarity values that do not hit in a pool of six million base pair sequences from third exons; (ii) minimising the false negative rate by ensuring detection of 90% of all possible motifs given a position weight matrix; (iii) balancing the two extremes by optimising the number of false positives and negatives. Here, I used option (i) to minimise the false positive rate in predicting transcription factor binding sites in §4.1.4 and §4.3.

3.2.2.3 Triplex forming sequences detection

DNA triplexes are formed between a polypurine or polypyrimidine rich double-stranded nucleic acid and a purine rich single-stranded DNA (for details on the base pairing see Knauert and Glazer, 2001). Goni et al. (2004)

demonstrated that triplex-forming sequences are enriched in gene promoters, suggesting a role in regulating transcription. I predicted a potential triplex-forming region if there are 10 or more consecutive guanines or cytosines in a sequence segment. I used these predictions to map functional SNPs (§4.3).

3.2.2.4 InterProScan

InterProScan identifies InterPro domains in amino acid sequences by combining all the individual domain-finding algorithms from the member databases (Quevillon et al., 2005). These include ProDom (Servant et al., 2002), PRINTS (Attwood et al., 2003), PIR Protein Sequence Database (Wu et al., 2003), Pfam (Finn et al., 2006), SMART (Letunic et al., 2006), TIGRFAMs (Haft et al., 2001), PROSITE (Hulo et al., 2006) and SUPERFAMILY (Gough et al., 2001). I utilised InterProScan to search for DNA-binding domains in the eukaryotic orthologues of human transcription factors (§4.2.7).

3.2.3 Microarray analysis methods

3.2.3.1 Loess print-tip normalisation

Locally weighted scatter-plot smoothing (loess) print-tip normalisation is a popular and effective method for processing two-colour cDNA arrays (Yang et al., 2002). It aims to correct for two major sources of bias in microarray data: (i) dye-dependent fluorescence; and (ii) spatial effects. The first is caused by differences in the fluorescent properties of the two dyes, which causes a non-linear increase in fluorescence intensities. The second is caused by spatial differences in hybridisation in a print-tip group, or because of handling procedures. The method normalises each print-tip group on the array separately by plotting a regression curve for the MA-scatter-plot using loess fitting. It then scales the log-ratio of each probe with the formula:

$$x_{ij} = \log_2 R/G - c_i(A)$$

This normalisation method can be used only if: (i) the number of data points to normalize is large enough to cover the intensity range; (ii) no more than 40% of genes are differentially expressed; and (iii) there is, approximately, an equal number of up- and down-regulated genes. This method was implemented as the preferred normalising algorithm in §4.1.1.

3.2.3.2 Inter-array scale normalisation

Due to differences in experimental procedure (such as total RNA concentration) arrays can display large variations in their range of probe intensity ratios. Differences in scale result in weighting arrays differently to each other during subsequent analyses. Inter-array scale normalisation corrects for severe differences in scale between multiple two-colour cDNA arrays (Yang et al., 2002). The median absolute deviation is used to estimate the scale factor required to correct for inter-array differences:

$$MAD = median\{|x_{ij} - median_i(x_{ij})|\}$$

The factor for each array is subtracted from the log-ratio values of each probe on the array. This method was implemented as the default between arrays normalisation in §4.1.1.

3.2.3.3 GC robust multi-array analysis

GC-Robust Multi-Array analysis (GC-RMA) is an analysis package distributed in Bioconductor for pre-processing Affymetrix GeneChip (Wu et al., 2004). Affymetrix GeneChips® measure expression levels using 10 to 20 pairs of 25-mer probes distributed along each gene. Each pair of probes contains one that matches the nucleotide sequence exactly and another with one mismatch in the central base position. The method is based on three steps: (i) background subtraction; (ii) data normalisation; and (iii) summarization. Background subtraction aims to correct for the effects of

non-specific hybridisation. GC-RMA fits a linear model to the intensity values assuming that the background follows a normal distribution. The background is then subtracted from the intensity values. Data normalisation is performed using quantile normalisation at the probe level. The method ranks the probes according to their intensity values and substitutes them by their mean expression. Summarization of all probes for a gene in a probeset is performed using median polish across all microarrays analysed. This allows us to account both for probe and array effect. I have used GC-RMA in sections §4.1.5 and §4.1.6 to produce high-quality pre-processed data from Affymetrix arrays.

3.2.3.4 AffyPLM

AffyPLM is a package to perform probe-level analyses of Affymetrix GeneChips (Bolstad et al., 2005). The package enables users to assess relative microarray quality and to determine whether different sets of arrays are of comparable among themselves. In particular, it allows calculations of spatial artefacts, relative log expression values and normalised unscaled standard errors. I have used these measures to assess the quality of the Genome Novartis Foundation dataset (§4.1.5).

3.2.3.5 MAS 5.0

MAS5.0 is a pre-processing method proposed by Affymetrix. It uses the mis-match probes as a control for non-specific binding and compares the signal of these probes against the signal from perfect matches for each probeset to derive present and absent calls based on the significance of these differences. However it has been demonstrated that mis-matches have more signal than perfect-matches in more than 30% of cases, suggesting that this approach might not be appropriate to produce such calls (Naef et al., 2002).

3.2.3.5 PANP

PANP is a method to calculate present and absent calls for Affymetrix arrays. About 300 probesets on the Affymetrix HGU133a GeneChip were

accidentally designed to match the antisense strand of annotated gene loci. As these probes should not hybridise any transcripts, they represent an ideal set of negative controls for the microarray that measures non-specific hybridisation. PANP uses the distribution of signal from these probesets to calculate thresholds with a known false positive rate to assign present/absent calls in each experiment. I used PANP to determine these calls in §4.1.5.

3.2.3.6 Linear models for microarray analysis

Linear models for microarray analysis (limma) is a Bioconductor package for the analysis of microarray data (Smyth and Speed, 2003). It contains normalisation and pre-processing functions for two colour cDNA arrays, and analysis methods based on linear models to perform platform-independent differential expression studies. It implements methods such as loess normalisation, moderated t-statistics, and several multiple testing procedures including FDR. I used limma as the backbone for normalisation procedures implemented in the DNMAD tool (§4.1.1).

3.2.3.7 Classification algorithms

Several classification algorithms were implemented to build two-class predictors for microarray data (§4.1.3): (i) Diagonal Linear Discriminant Analysis (DLDA; Dudoit et al., 2002) is a machine learning method able to find a linear combination of variables that best separate two classes; (ii) k-nearest neighbours (kNN; Dudoit et al., 2002) is a non-parametric classification algorithm that predicts class membership based on the most similar k-samples measured using Euclidean distance; (iii) Random Forest (Breiman, 2001) is a classification and regression tool based on an ensemble of single classification trees; (iv) nearest shrunken centroids (Tibshirani et al., 2002) is a classification algorithm that assigns samples to the class with a nearest centroid; and (v) Support Vector Machines (SVM; Vapnik, 1998; Noble, 2006) are an artificial intelligence technique that allows us to find the optimal separating hyperplane between classes.

All algorithms accept as input a gene expression matrix where rows represent genes and columns represent classes. The algorithm uses the class membership information to train itself. Once the classes have been learnt, a new sample will be assigned according to its similarity with the closest class. In order to maximise the data usage the prediction process is usually accompanied by cross-validation techniques.

3.2.3.8 Cross-validation

Cross-validation is a process by which the predictive accuracy of a model can be assessed based on learning and test datasets, both derived from an initial set of samples. In §4.1.3 a ten-fold cross-validation model was implemented. This method randomly divides a dataset into ten groups maintaining the class proportions, ie, same number of samples per class, whenever this is possible. One of the groups is left out and the prediction algorithm is trained using the other nine groups. The accuracy of the classification algorithm is calculated using the left-out sample. The process is repeated until all groups have been evaluated and the final cross-validated error rates are calculated as the mean error rate of all cross-validation rounds. Different cross-validation techniques are discussed in Ambroise and McLachlan (2002) and Braga-Neto and Dougherty (2004).

3.2.3.9 symp

symp is a Bioconductor package for the analysis of ChIP-chip data. It implements several pre-processing approaches for high-density tiling microarrays such as background subtraction and quantile normalisation provided by GC-RMA. The package also includes functions to obtain significantly enriched regions in ChIP-chip experiments by using data points enriched in the non-immunoprecipitated sample to calculate a symmetric null distribution. This distribution is then used to calculate statistical significance for probes enriched in the immunoprecipitated sample. In §4.1.7 I show the results of the analysis of ChIP-chip data using this approach.

3.2.3.10 FatiGO

FatiGO is a web-tool for the comparison of Gene Ontology annotations between two sets of genes (Al-Shahrour et al., 2004). It calculates p-values for the enrichment of particular GO functions in one gene set over the other. This is done using a Fisher's exact test to compare the proportion of genes with a certain GO annotation. p-values are corrected for multiple testing using FDR. I have used FatiGO in §4.1.6 to validate the usage of the propensity as a tissue-specificity marker.

3.3 Programming languages

The majority of programs developed in this thesis have been written in Perl, R (R Core, 2004, http://www.Rproject.org) and MySQL. In addition I have used Bioconductor extensively as a source of data analysis and visualisation software packages (Gentleman et al., 2004).

I have also used the Ensembl APIs (Core, Compara and Variation) and BioPerl to access and analyse genomic data.

3. 4 Computational facilities

The majority of the computational work presented in this dissertation was run in the research farm of the EMBL - European Bioinformatics Institute in 64-bit architecture machines running CentOS operating system.

The web-tools presented in this thesis included in the GEPAS and Asterias packages run on pound web-balanced apache servers with Debian GNU/Linux operating systems at the Spanish National Cancer Centre and Centro de Investigación Príncipe Felipe (`http://www.gepas.org`; Herrero et

al., 2003a; Vaquerizas et al., 2005; `http://asterias.bioinfo.cnio.es` Diaz-Uriarte et al., 2007).

4. Results

This chapter describes the main results obtained during my PhD research. It is divided into three main sections covering: (i) development of tools and methods for high-throughput data analysis; (ii) functional characterisation of the human transcription factor repertoire; and (iii) identification of potential functional SNPs.

4.1 Development of microarray data analysis methods and tools

This first section of the results presents web-tools and methods I have developed for the analysis of high-throughput gene expression data. These include: (i) two-colour cDNA microarray normalisation; (ii) differential gene expression analysis; (iii) class prediction for microarray data; (iv) functional annotation of transcription-factor regulated genes; (v) sensitivity and specificity measurement for Affymetrix GeneChips; (vi) determination of tissue-specific expression for microarray data; and (vii) analysis of tiling array experiments.

4.1.1 DNMAD: Diagnosis and normalisation for two-colour cDNA arrays

DNMAD is a web-tool for the Diagnosis and Normalisation of MicroArray Data (`http://dnmad.bioinfo.cnio.es`). It allows users to normalise two-colour cDNA microarrays using robust statistical methods and to perform quality assessments of the data (Vaquerizas et al., 2004). The application uses R and Bioconductor for all calculations and provides a web-based Perl CGI interface that communicates with R using the CGIwithR package. By default, it processes the data using print-tip loess normalisation; however, global loess and MAD inter-array normalisation are also provided as user-specified options. In addition to the processed data, the program outputs diagnostic graphs for data quality assessment such as box-plots, MA-plots and slide-location plots.

DNMAD input

Figure 4.1 displays the web interface. Here, the user can select the microarray data files for upload. The server accepts compressed or plain text GenePix files that must be accompanied with a description of the slide layout. Custom files with appropriate headers are also allowed: these must contain a minimal amount of numerical information about each probe (Block, Column, Row, Name, ID, F635 Mean, B635 Median, F532 Mean, B532 Median and Flags).

Once data files have been uploaded, DNMAD offers several user-defined options. At the experimental level, users may select the automated analysis of dye-swaps experiments by indicating the appropriate Cy5/Cy3 comparison for each microarray. At the microarray level, users can decide whether to include flagged spots (ie, low quality spots identified by the user or scanning software). In addition they can choose whether to subtract the background intensity for each spot (which can lead to "negative" intensity values that must be discarded) or to substitute negative values with an arbitrary intensity value of 0.5. Finally, there is an option for performing global loess normalisation, which allows users to normalise the entire microarray in cases

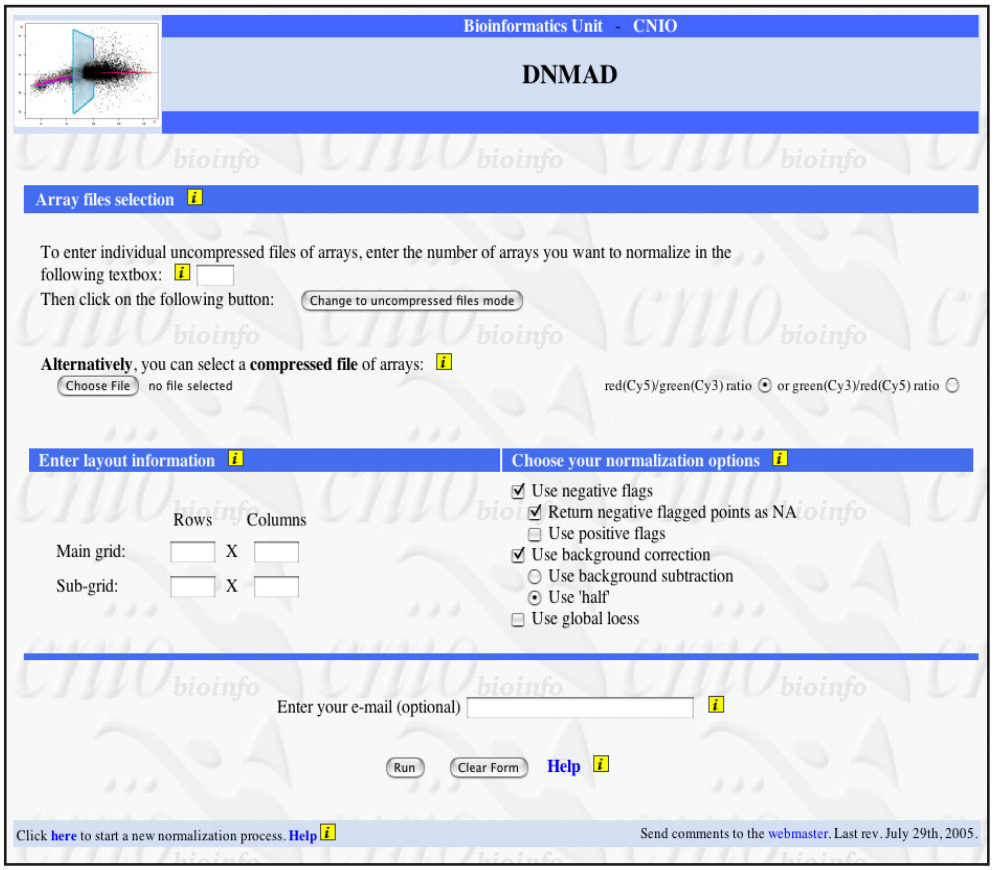

Figure 4.1 | Screenshot of DNMAD web interface. The data upload box and user-defined options are shown.

where the criteria for print-tip loess are not met.

DNMAD output

The web-tool outputs normalised gene expression values for each microarray, described by Cy5/Cy3 $\log_2$ ratios (M-value) and average intensities (A-value). The data files can be downloaded from the output page or can be automatically transferred to other modules in the GEPAS microarray analysis suite for further analysis (Herrero et al., 2003b).

The output also records information about the normalisation procedure employed, errors, warnings, and plots to interpret the results and perform

quality assessments (Figure 4.2, 4.3). These include: (i) box-plots for all microarrays and print-tip groups; (ii) MA-plots with regression curves for each print-tip group; and (iii) diagnostic plots.

Box-plots (Figure 4.2 A, B) represent the overall distribution of the log-ratios. The box at each column indicates the inter-quantile range (IQR), within at which 25% and 75% of the distribution is located. The median is also indicated by a black line inside the box. Extreme intensity values that deviate from the IQR by more than 1.5 times are represented as single dots. Box-plots are computed for pre- and post-normalised data for all the microarrays in the experiment as well as for individual print-tip groups within each microarray. These allow the user to compare log-ratios across all samples and print-tip groups. In a typical experiment raw microarrays may have variable log ratio values between print-tips, which is corrected after normalisation. Extreme differences in raw print-tip behaviour within the same slide might indicate issues with the manufacturing, handling or hybridisation procedures.

The MA-plots in Figure 4.2 C-D show the relationship between the $\log_2$ Cy5/Cy3 ratios (M-values) and average Cy5 and Cy3 intensities (A-values) for a single microarray. The plots are computed for raw and normalised data. In each case, differently coloured regression curves are shown for every print-tip group. Assuming that most genes in the experiment have roughly equal expression levels, significant deviations from zero in the regression curve indicate differences in the detection of dyes across the intensity range. This effect can also be seen if the regression curve is non-linear. These effects should be corrected via the normalisation procedure, resulting in regression curves close to an M-value of zero.

The diagnostic plots (Figure 4.3) allow users to detect hybridisation artefacts on the microarrays (Smyth and Speed, 2003). Plots include: (i) histograms of the Cy5 and Cy3 foreground probe intensities; (ii) probability density plots of the same intensities for individual slides and the entire experiment; and (iii) schematic microarray diagrams of Cy5 and Cy3 foreground and background intensities, and $\log_2$ ratios before and after normalisation. These plots enable users to identify extreme fluorescence intensities or spatial biases that arise

from problems during hybridisation and scanning.

Example of DNMAD usage

Figure 4.2 show pre- and post-normalisation box-plots and MA-plots for a single microarray from Bullinger et al. (2007), in which cDNA microarrays are used to differentiate subclasses of acute myeloid leukaemia. Here we observe how differences between the Cy5 and Cy3 intensity levels are corrected. Regression curves for each print-tip group in the pre-normalised

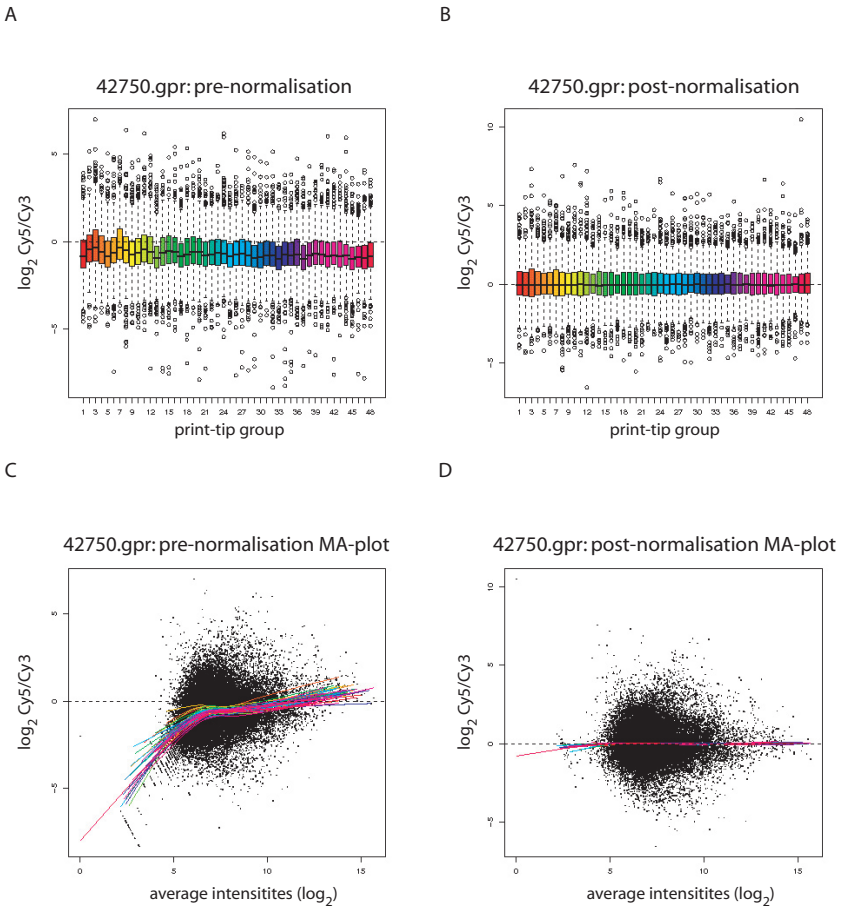

Figure 4.2 | Boxplots and MA-plots for a single cDNA microarray (Bullinger et al., 2007) before and after normalisation. (A, B) Boxplots display the range of $\log_2$ Cy5/Cy3 ratios for each print-tip group (labelled 1-48). (C, D) MA-plots show the breath of $\log_2$ Cy5/Cy3 ratios (M-value) at different fluorescence intensities (A-value).

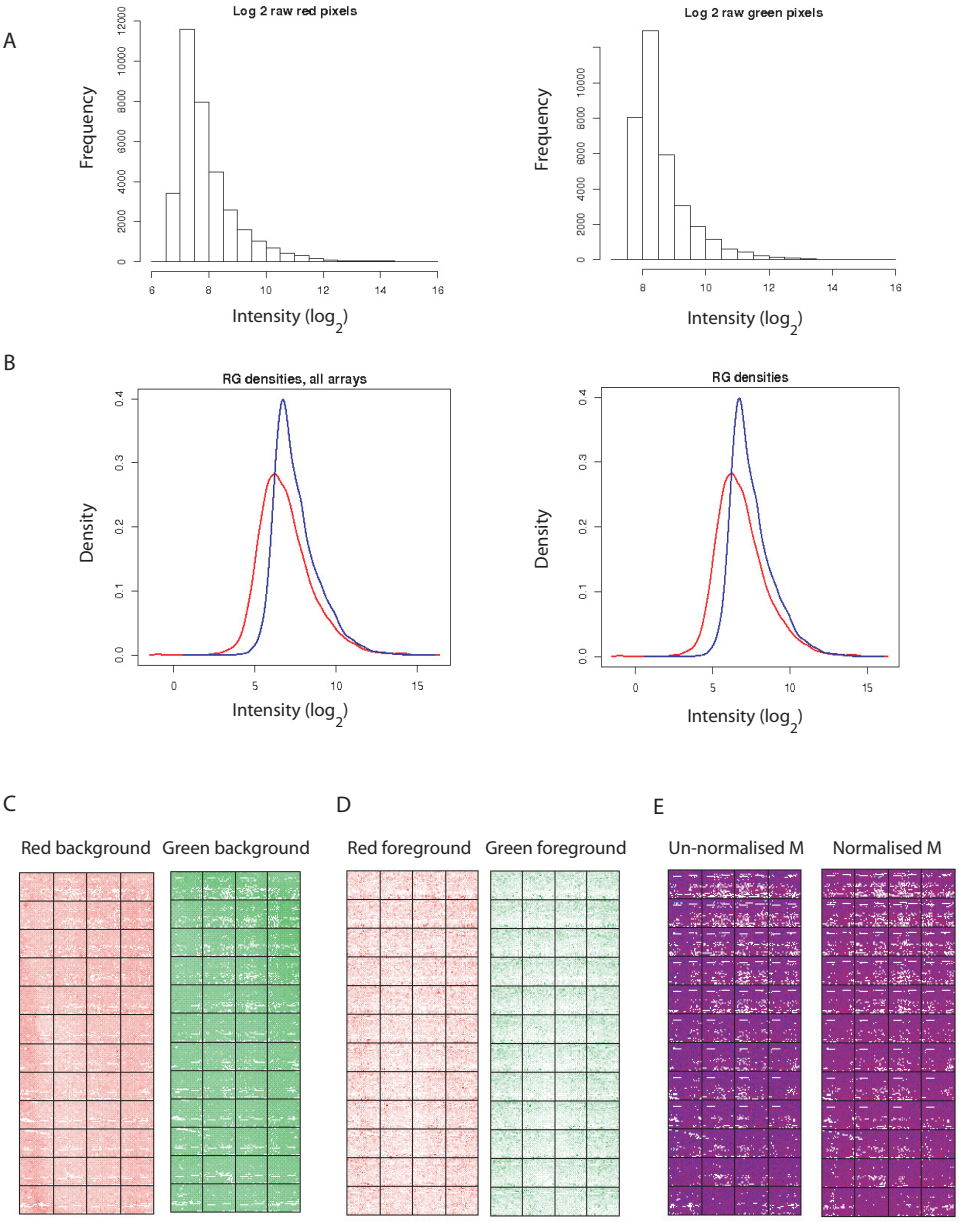

Figure 4.3 | Diagnostic plots for two-colour cDNA microarrays. The diagnostic plots allow users to detect quality problems due to the experimental procedure or the manufacturing process. (A, B) Histograms and probability distributions of Cy5 and Cy3 $\log_2$ intensity values. Comparisons between plots allow users to detect extreme differences between slides. (C-E) Schematic representations of microarrays displaying Cy5 and Cy3 background and foreground intensities, and $\log_2$ ratios before and after normalisation. These diagrams allow users to identify spatial hybridisation artefacts.

MA-plots show how the Cy5 and Cy3 intensities vary non-uniformly across the intensity range. This effect is corrected via the normalisation procedure as shown in the post-normalised MA-plots. The diagnostic plots for this microarray allow us to determine that the slide does not suffer from spatial artefacts (Figure 4.3)

4.1.2 Pomelo: Differential expression for microarray experiments

Pomelo is a web-tool for the detection of differentially expressed genes in microarray experiments. It combines several detection methods including the t-test, ANOVA and Fisher's exact test. The expression of each gene is tested using the selected procedure. The tool then performs a random permutation test to calculate the statistical significance for differential expression. This removes the necessity of assuming normality in the data. p-values obtained are then corrected for multiple-testing using the FDR procedure (see § 3.2.1.6).

Pomelo input

The web-tool is a Perl CGI form that communicates with the main C++ program that runs the statistical tests (Figure 4.4). The server accepts a numerical matrix of expression values as input. Also, users have to provide a text file defining the class membership of the biological samples. By default, the t-test is selected for measuring differential expression, allowing users to compare non-categorical variables between two classes of observations. Analysis of the variance (ANOVA) is also provided for cases when more than two classes are to be compared. The Fisher's exact test allows analysis if non-numerical or categorical variables, such as the presence or absence of a gene. The number of permutations is set by default to 10,000 although this number can be modified depending on particular experiments. Finally, the interface includes several parameters that allow the user to modify the appearance of the graphical output.

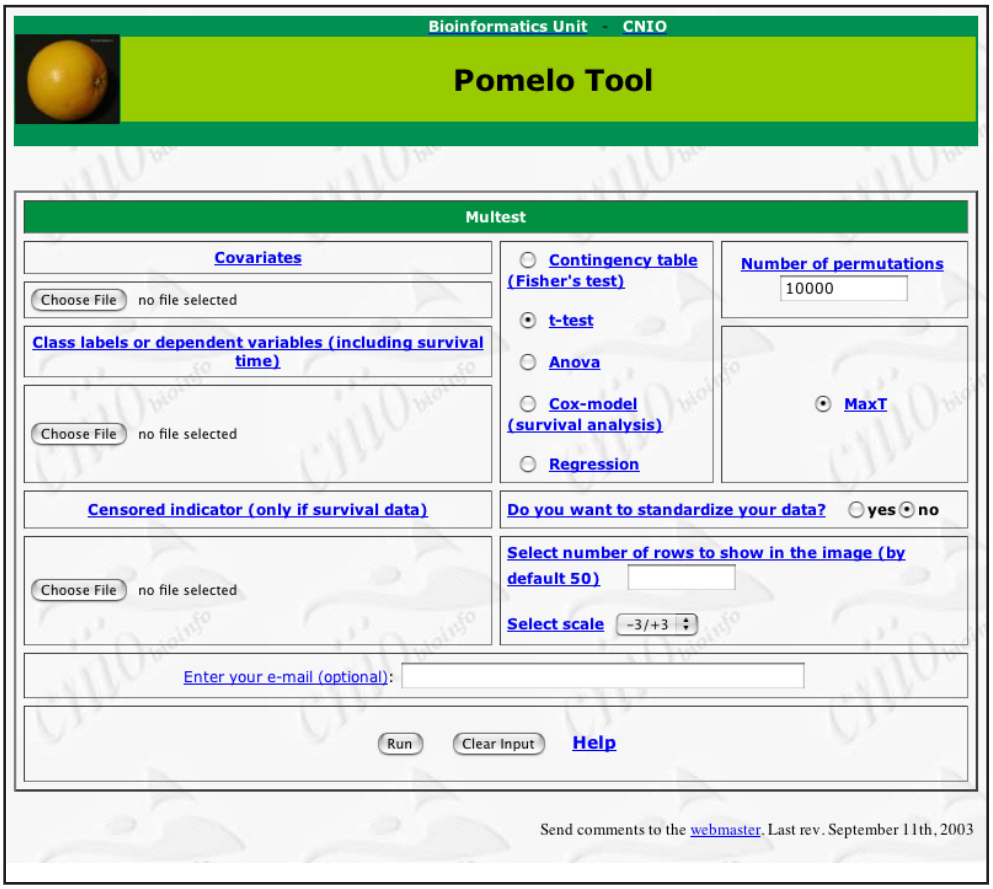

Figure 4.4 | Screenshot of Pomelo web interface. The capture displays the data upload section as well as the user-defined options. The tool is available at `http://pomelo.bioinfo.cnio.es`.

Pomelo output

The output page returns a list of differentially expressed genes, along with the statistical test values, unadjusted and FDR adjusted p-values. The user can use the output as a guide for setting a significance threshold for differential expression. A graphical output for the most differentially expressed genes is also provided (Figure 4.5), which allows users to evaluate visually the results. The output page is connected to other modules of the GEPAS microarray analysis suite so that differentially expressed genes can be analysed for functional annotation.

Example of Pomelo usage

We analysed the data from Golub et al. (1999) as a case study to illustrate the usage of Pomelo. The dataset contains gene expression measurements for 6,817 genes in 38 samples of acute myeloid leukaemia and acute lymphoid leukaemia. The original study demonstrated the use of microarrays as a method for classifying cancer samples. Using the t-test on Pomelo we identified 656 differentially expressed genes at an FDR adjusted p-value threshold of 0.05. Figure 4.5 shows a graphical display of the top 50 most differentially expressed genes.

4.1.3 TNASAS: Class prediction for microarray data

TNASAS is a web-tool for class membership prediction using microarray data. The tool implements simple statistical techniques to output class predictors. These are lists of genes that allow us to distinguish between two different biological samples given their expression profiles.

Building class membership predictors

The procedure involves two steps: (i) selecting the genes to include in the class predictor; and (ii) evaluating the error rates of predictions.

For the first, TNASAS ranks all genes in the dataset by their level of differential expression between classes, which will have the best predicting power. Although any of the approaches described in §4.1.2 could be used for this purpose, the permutation tests for calculating the p-values are computationally expensive. Therefore TNSAS implements gene ranking based on the statistic values themselves. Users can select from the following statistics: (i) the F-ratio, which is the ratio between the variances of the classes; (ii) the Wilcoxon statistic, which is a non-parametric version of the standard t-test; and (iii) variable importance derived from random forests. TNASAS then automatically selects the N best genes based on the ranking procedure, where N is the top 2, 5, 10, 20, 35, 50, 75, 120, 200, 500, 1,000 and 2,000 ranked genes.

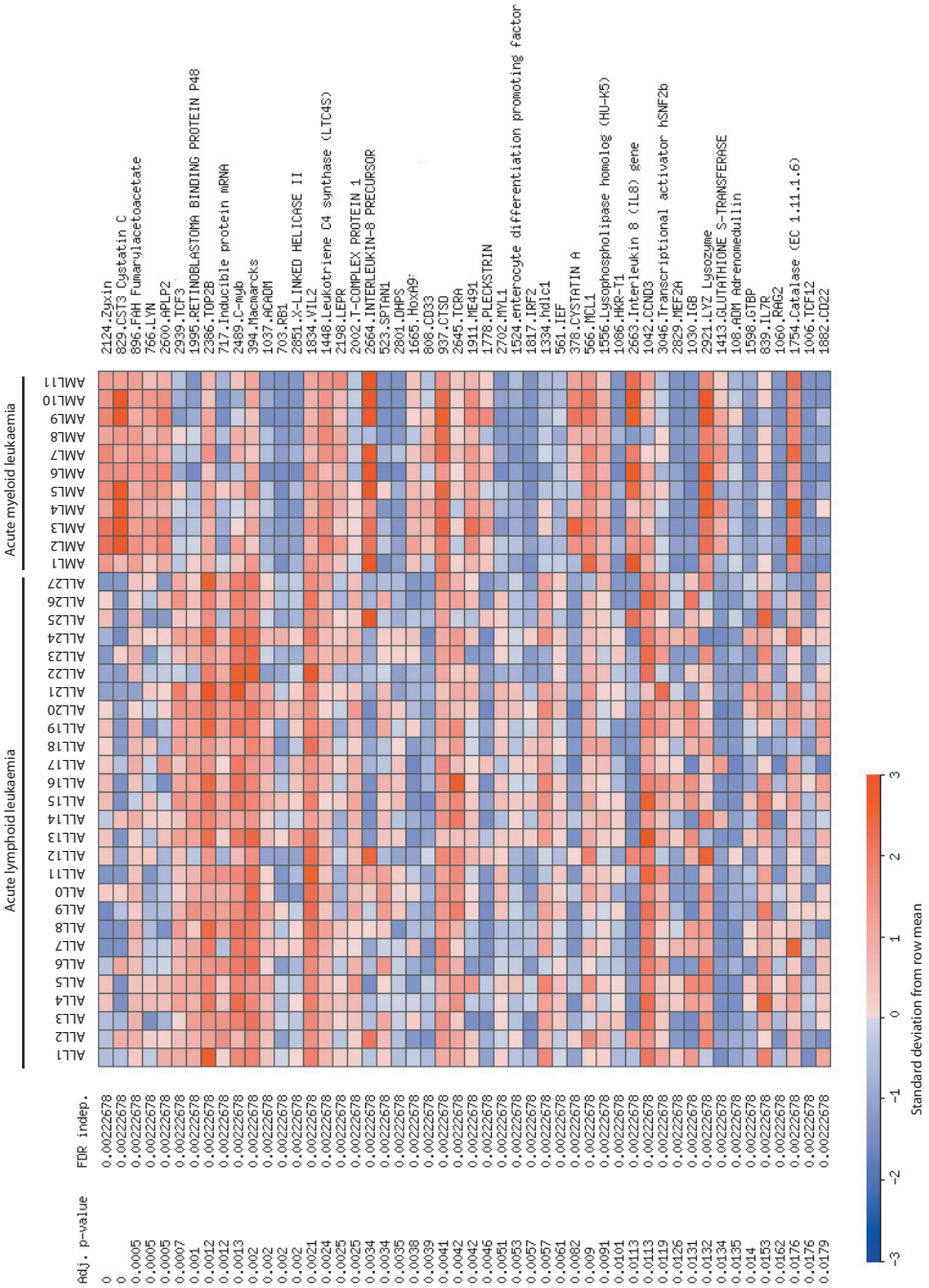

Figure 4.5 | Graphical output from the Pomelo tool. Here, the 50 most differentially expressed genes (rows) between the leukaemia samples (columns) are shown.

For the second, the tool uses one of the user-selected classification algorithms to identify the best set of class predictors. TNASAS implements several simple classification algorithms that have been shown to perform well in microarray data analysis (Diaz-Uriarte and Alvarez de Andres, 2006; Romualdi et al., 2003): (i) diagonal linear discriminant analysis (Barrier et al., 2007); (ii) k-nearest neighbours (Pomeroy et al., 2002); (iii) support vector machines (Vapnik, 1998); (iv) random forests (Svetnik et al., 2003); and (v) shrunken centroids (Tibshirani et al., 2002) (see §3.2.3.7 for details).

The algorithms are trained using a 10-fold cross-validation scheme, and prediction error rates are computed from the left-out sample during training.

Accounting for selection bias

When building class predictors, different biases can severely affect the estimates of the predicted error rates (Ambroise and McLachlan, 2002; Simon et al., 2003).

The first source of bias arises during the filtering procedure, in which genes are ranked by their expression levels. When all samples in the dataset are used for ranking purposes, the classification procedure is biased towards the particular set of genes selected for those samples. This results in more optimistic predicted error rates during the cross-validation procedure as the left-out sample contributed already to the gene selection. TNASAS resolves this by drawing an extra 10-fold cross-validation procedure at the start of the gene ranking process, by which some samples from each class are left out from the ranking procedure and are only used to compute the error rates.

The second source of bias arises as we select the best number of genes to build the predictor 'a posteriori', ie, when we know the error rates of all predictors. To account for this bias, an extra layer of cross-validation is used to determine the error rate of building multiple predictors and then choosing the one that produces the smallest error rates of prediction.

TNASAS input

The tool was implemented using R and Bioconductor and the web-interface is a Perl CGI form that communicates with R using the CGIwithR package (Figure 4.6; Firth, 2005). The server accepts as input a numerical matrix containing $\log_2$ gene expression values of all genes in the experiment and a plain text file indicating the class membership of the samples. The user then selects the ranking and classification methods before submitting the data.

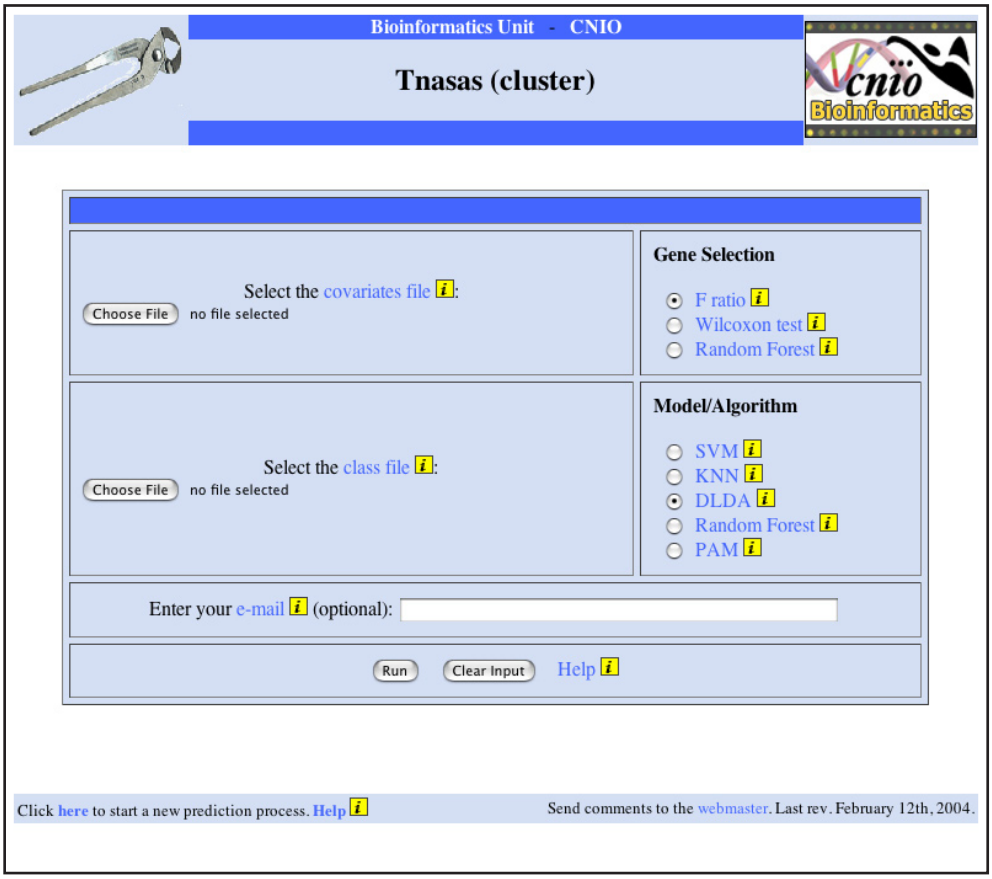

Figure 4.6 | Screenshot of TNASAS web interface. The image shows the data upload boxes and the user-selected algorithms. The tool is available at `http://tnasas.bioinfo.cnio.es`.

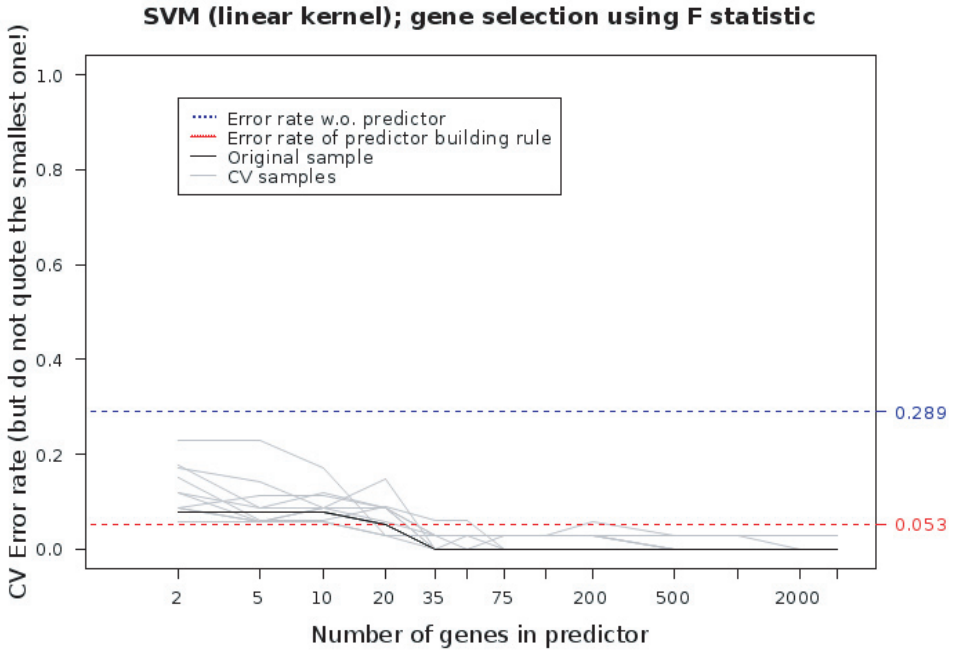

Figure 4.7 | TNASAS cross-validated prediction error rates obtained for the Golub et al. (1999) dataset using Fisher's statistic ranking and Support Vector Machines as classifying algorithm. The plot shows the error rate that would be obtained predicting the most common class (blue dotted line), the cross-validated error rate of the entire process (red dotted line), the error rate for the original sample (solid black line), and the error rates of all the cross-validation sets (grey lines).

TNASAS output

TNASAS outputs text and graphical representations of the best class predictors and the prediction error rates (Figure 4.7). The output also includes the results for the different cross-validation runs, allowing users to evaluate the robustness of predictions.

Example of TNASAS usage

We demonstrated the use of TNASAS by analysing the data from Golub et al. (1999). We selected the F-ratio ranking and SVM classification algorithm to identify the best class predictors to distinguish between acute myeloid

leukaemia and acute lymphoid leukaemia and associated error rates.

The best predictor comprises a set of N=35 genes producing a cross-validated error rate of 5.3%. This means that there is a good separation between the two classes for this dataset. 20 of the 35 genes present in the predictor were selected as predictors in the original study, including the cancer-related retinoblastoma, cyclin D2 and E2A genes. The classification matrix showed that only one of the samples was misclassified and that the different cross-validation rounds were consistent (data not shown). We re-analysed the dataset using different combinations of gene ranking and classification algorithms and observed that the predicted error rates were similar for all combinations (data not shown).

4.1.4 TransFAT: Transcription factor regulation for sets of genes

TransFAT (Transcription Factor Association Test) is a web-tool for identifying transcription factors that potentially regulate sets of human and mouse genes with similar expression profiles. The tool searches for over-represented transcription factor binding sites in the promoter of co-expressed genes. We scanned a 10kb region upstream of all human and mouse genes annotated in Ensembl (v25) using the Match program (Kel et al., 2003) with the following options: (i) vertebrate-specific position weight matrices; and (ii) minimisation of false positives (see §3.2.2.2 for details). This approach resulted in more than 2.5 million putative transcription factor binding sites for 270 different transcription factors in 51,817 human and mouse promoters. We then assigned genes to transcription factors based on the binding site predictions. Using these associations, TransFAT identifies over- and under-representations of transcription factors in sets of genes via Fisher's exact tests. A correction for multiple testing is provided which accounts for testing all human or mouse transcription factors in a given dataset at the same time.

TransFAT input

TransFAT was programmed as a Perl CGI script that communicates with

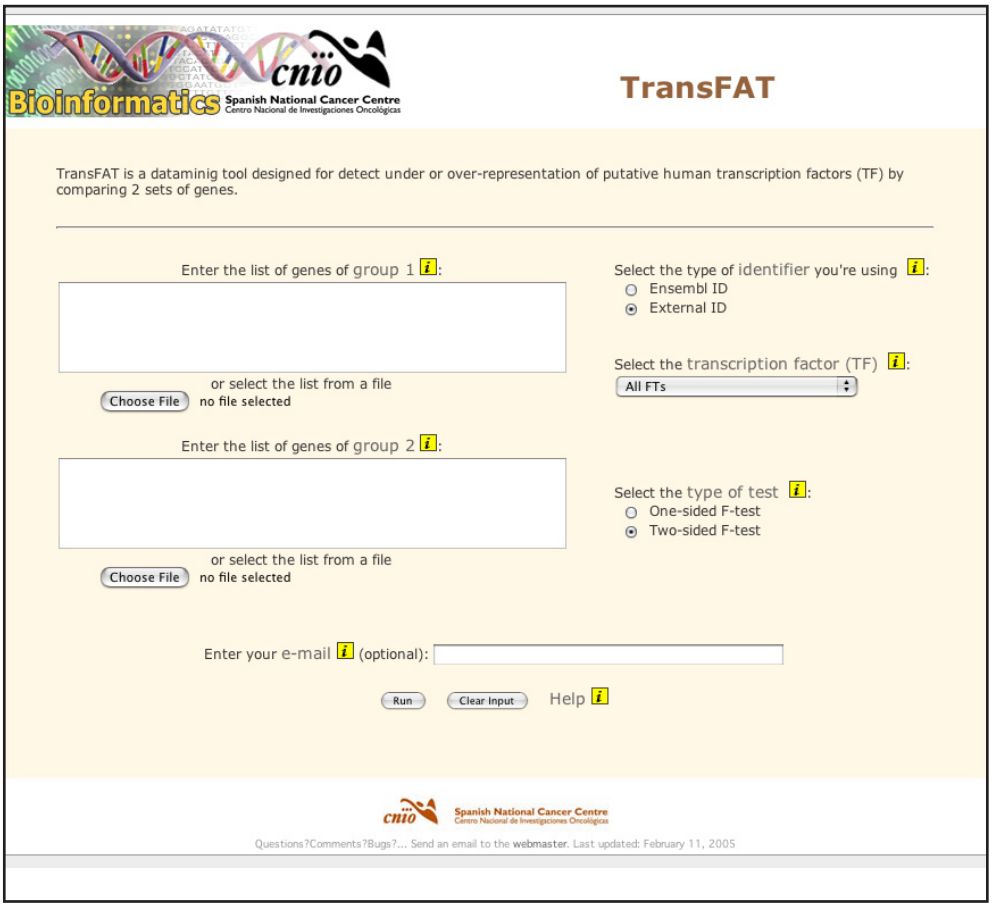

Figure 4.8 | Screenshot of TransFAT web interface. The tool is available at `http://babelomics.bionfo.cipf.es`.

a MySQL relational database, which stores the transcription factor binding site predictions. The user inputs two lists of genes (eg, co-expressed genes and a reference set) that are used to compare the binding site occurrence in promoter regions (Figure 4.8).

TransFAT output

TransFAT outputs the Fisher's exact test results for all evaluated transcription factors, showing the number of genes associated in each group, percentages, test statistic, p-value and FDR adjusted p-value.

Example of TransFAT usage

We tested TransFAT on the set of differentially expressed genes from Golub et al. (1999). Unfortunately there were no significantly over-represented transcription factor binding sites.

4.1.5 Assessing selectivity and specificity in Affymetrix GeneChips

Single-channel oligonucleotide microarrays, such as the Affymetrix GeneChips, allow us to obtain global measurements of mRNA expression levels. However, due to non-specific hybridisation and normalisation procedures, it is difficult to determine whether a gene is expressed or not. Here, I present a method to detect the presence or absence of genes in Affymetrix GeneChips. The method is based on the combined usage of the PANP algorithm and EST sequences to assess the selectivity and specificity of the experiment. We performed the analysis for the GNF dataset and it forms the basis for the analysis of transcription factor expression presented in §4.2.

Quality check and pre-processing of the GNF dataset

We first performed a quality check of the GNF dataset to ensure its quality. We used the Bioconductor package AffyPLM to fit a linear model to the expression measurements across all probes. By analyzing the computed weights and standard errors of the model, we can detect any systematic variance within the data and identify individual arrays that are of poor quality.

The Relative Log Expression (RLE) computes the median expression value across all microarrays, and subtracts value from each probe. The box plots in Figure 4.9A displays the distribution of RLE values for each microarray in the dataset. We expect a large fraction of genes to behave similarly across most cellular conditions, which should result in box plots that are centred on RLE values of 0 and display only a small spread.

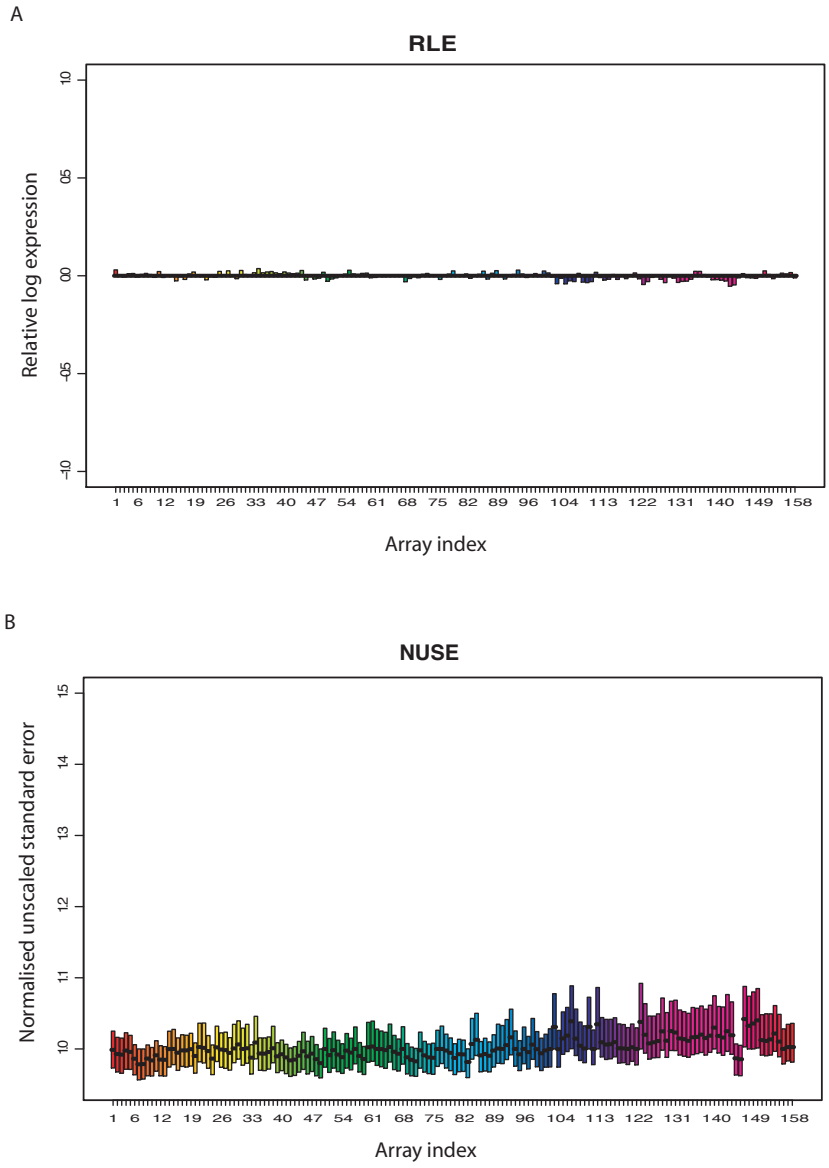

Figure 4.9 | Quality assessment plots for the GNF SymAtlas dataset. (A) Boxplots of the Relative Log Expression (RLE) values show the spread of gene expression levels for each microarray relative to the entire dataset. RLE values are centred about 0 and have small spreads, indicating that all the microarrarys behave similarly. (B) Boxplots of the Normalised Unscaled Standard Error (NUSE) show the standard errors for each microarray. NUSE values are fairly uniform across the dataset indicating that there are no large experimental artefacts affecting a subset of microarrays.

The Normalised Unscaled Standard Error (NUSE) allows us to compare the standard errors across microarrays; similar standard errors within a dataset indicate that it is sensible to compare data between the arrays. The boxplots in Figure 4.9B show that the GNF dataset has similar NUSE values throughout.

The linear model also allows us to draw schematic images of the arrays to identify any spatial artefacts (Figure 4.10). These can show: (i) the raw expression values; (ii) the weights of the model (ie, how much outliers have been down-weighted); (iii) and residual values (ie, the deviation from the expected expression value given the model).

After confirming that the GNF dataset is of high quality and does not suffer from major errors, we then pre-processed the data using the GCRMA algorithm (§3.2.3.3).

Specificity and sensitivity

As described in §3.2.3, PANP uses a set of negative strand-matching probesets to calculate false positive rates at different expression thresholds. The percentage of hits from these probesets gives an accurate estimate of specificity for the list of expressed genes.

To calculate the true positive rate, we compared the microarray data to EST measurements for equivalent tissue types. The Unigene database collects EST reads for different tissues and cell types, and clusters them to particular genomic loci. As an mRNA must be present in a sample to be sequenced, an EST provides proof of gene expression against which we can gauge microarray data. We selected 31 Unigene libraries covering 13 normal adult tissues (Table 4.1). We compare the distribution of microarray expression values for genes that are present or absent as ESTs in the Unigene libraries (Figure 4.11). It is clear that probesets associated with an EST have higher expression values than those that lack an EST. However, it is also notable that there are genes in the EST libraries displaying very low microarray expression values. This shows that although microarrays are able to detect

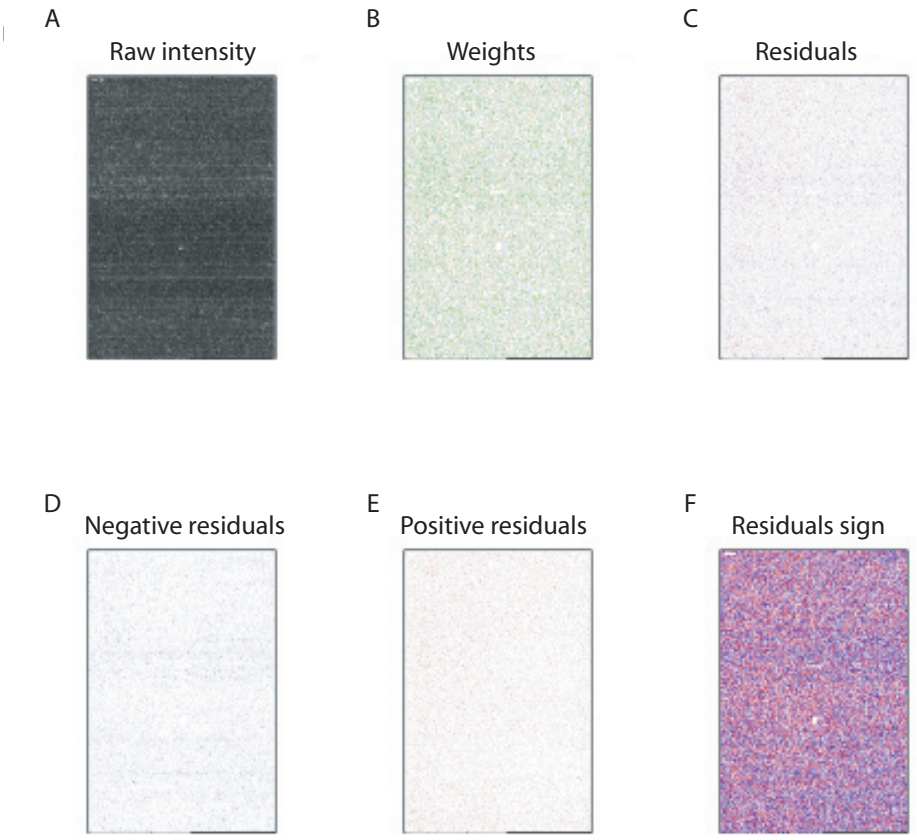

Figure 4.10 | Schematic images of microarrays for the GNF SymAtlas medulla oblongata sample (3AJZ02081479a). The schematics display different values for each probe on the microarray, allowing users to identify visually any spatial artefacts. (A) Raw intensities (colour scale from black for low values to white for high values), (B) linear model weights (white to green), (C) residual values between the raw intensity and the linear model (blue to red), (D) negative residuals (blue to white), (E) positive residuals (white to red), (F) sign of residuals (blue or red).

accurately highly expressed genes, there is a limit to their detection capacity for genes expressed at lower levels.

A major difficulty in the analysis of microarray data is picking a threshold for deciding whether a gene is expressed or not. We plotted ROC curves to compare the performance of several methods for determining thresholds:

Table 4.1 | Unigene libraries for the analysis of specificity of the Affymetrix GeneChip expression data. Mean detected and mean undetected columns correspond to mean expression values for probesets with and without corresponding EST.

Unigene library ID	Tissue	Mean detected	Mean undetected
6759	heart	5.109054	3.941986
6833	kidney	4.750552	3.893576
252	liver	5.582172	3.840463
6989	liver	4.760076	3.86308
249	lung	5.033341	3.890131
6834	lung	4.728005	3.898488
2709	lymph node	5.49709	4.015194
2710	lymph node	5.887549	4.045426
2711	lymph node	5.049647	4.097363
3718	lymph node	4.541702	4.046978
3719	lymph node	4.716034	4.111302
3720	lymph node	4.704832	4.109337
45	muscle	5.093514	3.881305
530	muscle	5.361126	3.871097
6761	muscle	4.515907	3.856882
14414	muscle	4.861265	3.916233
253	ovary	5.391216	3.960875
10196	ovary	4.276806	3.99185
5551	pancreas	5.126557	3.932393
6760	pancreas	5.576891	4.0339
13019	pituitary gland	5.164571	3.847367
250	placenta	6.342932	4.047371
2587	placenta	4.971826	3.903643
6835	placenta	5.278797	3.922187
13000	placenta	4.980353	3.996216
13001	placenta	5.313617	3.976539
6763	prostate	5.14798	3.928227
14129	prostate	5.082174	3.945071
14131	prostate	6.042568	4.099017
14590	spinal cord	5.831633	4.00117
13710	testis	4.959629	3.899426

arbitrary expression value cut-offs that are commonly used ($\log_2(100)$, $\log_2(150)$, $\log_2(200)$); MAS5.0; and PANP. We use the negative strand-matching probesets to calculate the specificity, and the presence or absence of ESTs as a measure of sensitivity.

Figure 4.12 displays the ROC curve for the lymph node. The solid black line represents sensitivity-specificity values for expression value thresholds between $\log_2(1)$ and $\log_2(200)$. There is a dramatic increase in sensitivity at lower expression values, followed by a shoulder in the curve and a slower

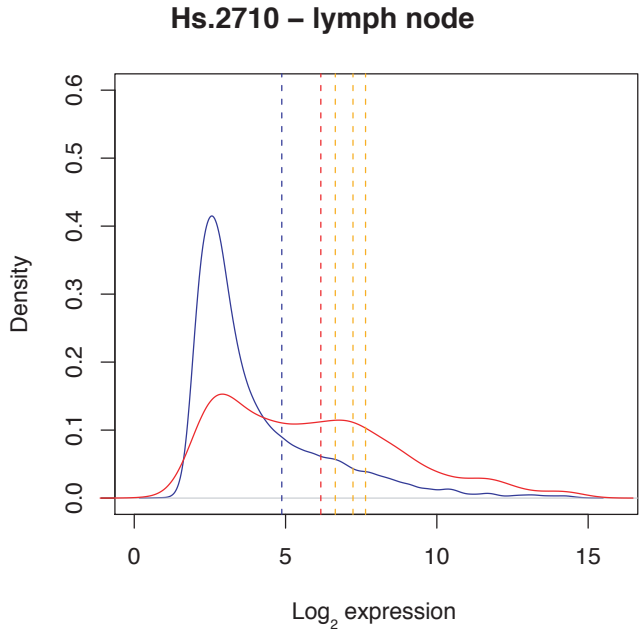

Figure 4.11 | Probability density distributions of microarray expression values of ESTs. Microarray expression values are higher for transcripts that are detected as EST reads (solid red line), compared with those that are undetected (solid blue line). The vertical dotted lines represent the PANP thresholds at 1% (blue) and 5% error rates (red), and commonly used arbitrary thresholds at $\log_2(100)$, $\log_2(150)$, and $\log_2(200)$ expression values (orange).

rise in sensitivity. A good threshold should maximise the sensitivity of detection and minimise the specificity. We clearly see that the commonly used arbitrary thresholds lie at the very left of the curve (orange data points): they produce few false positives, but also miss many expressed genes. The Affymetrix MAS5.0 algorithm gives reasonable false positive control, but returns a fairly low sensitivity (red data point lies right of the curve). At a 1% false positive rate, the PANP algorithm appears to be too stringent (blue low data point). However at a 5% error rate the cut-off lies on the shoulder of the ROC curve, thus maximising sensitivity without compromising much on specificity. Application of the algorithm to all tissues with this cut-off results in detection rates from 45% to 65%.

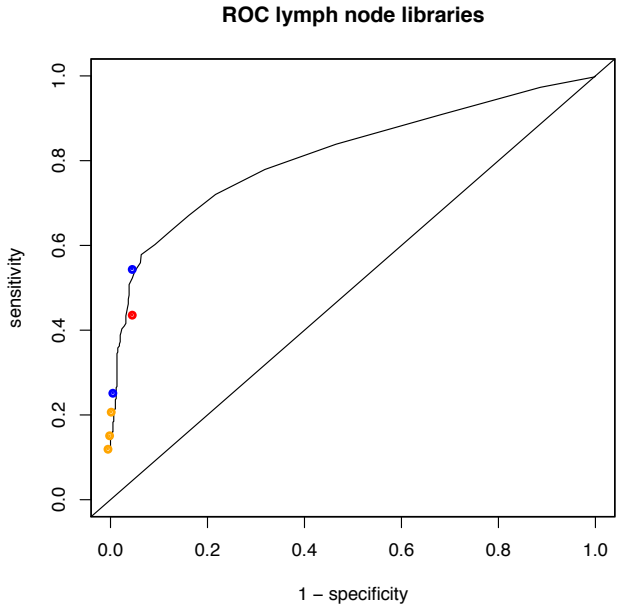

Figure 4.12 | Receiver Operator Characteristic (ROC) curves measuring the sensitivity and specificity of microarray data for the lymph node. The solid black curve represents the quality of gene expression data at different thresholds at expression values between $\log_2(1)$ and $\log_2(200)$. Coloured dots represent different cut-offs: PANP at 1% and 5% error rates (blue), MAS5.0 (red) and commonly used arbitrary thresholds at $\log_2(100)$, $\log_2(150)$ and $\log_2(200)$ (orange). The best balance between sensitivity and specificity is achieved using the PANP 5% cut-off, which occurs at the shoulder of the ROC curve.

4.1.6 Detection of tissue-specific gene expression

Statistical techniques such as those presented in §4.1.2 allow detection of differentially expressed genes among biological samples. However, these approaches become limited – both computationally and conceptually – with increasing numbers of samples, as in the GNF dataset. For example, it is not clear which tissue should be used as a common reference, and even with a list of differentially expressed genes from pairwise comparisons, it is non-trivial to define whether they are specifically expressed in particular samples.

Here we show how a simple statistic, the propensity (§3.2.1), can be used

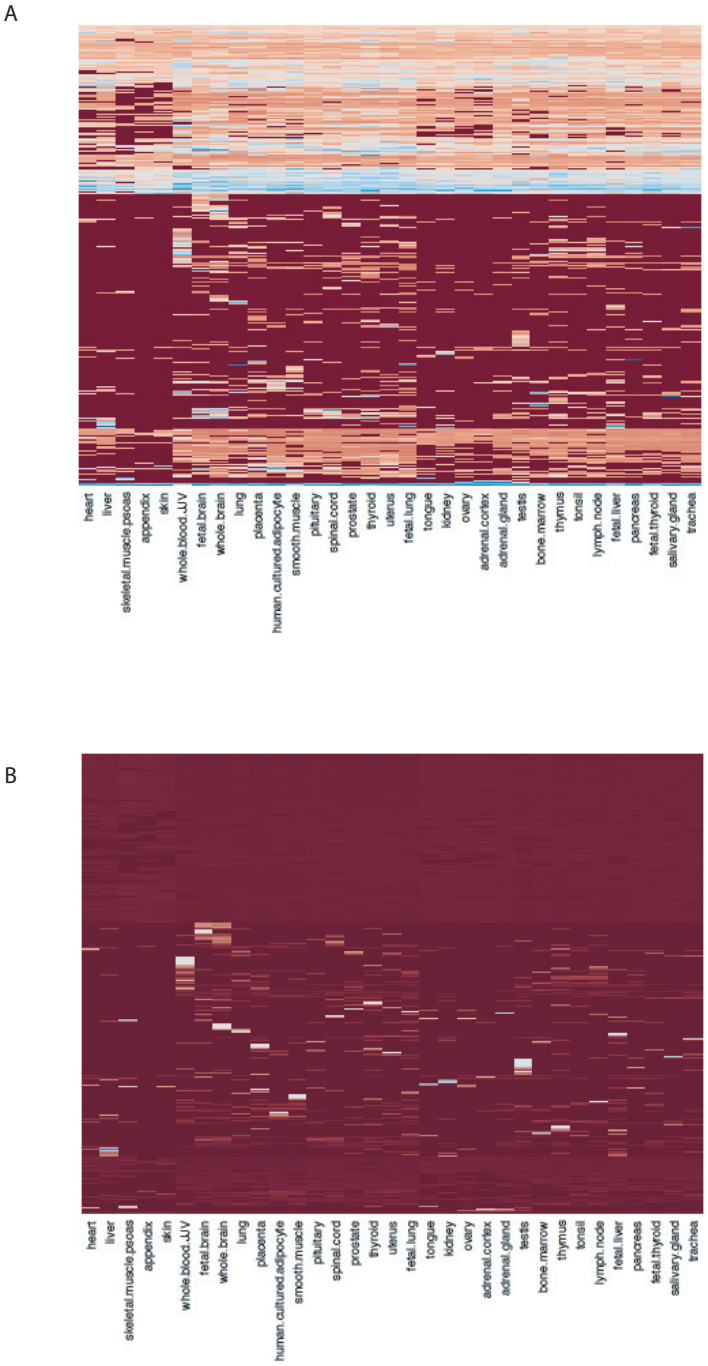

Figure 4.13 | Heatmap of gene expression in 33 major human organs and tissues. Genes (rows) and tissues (columns) are aligned by hierarchical clustering of expression values. Intersecting cells in the heatmaps display (A) expression values (colours range from red to blue respectively for low to high expression) and (B) propensity values (same colour scale).

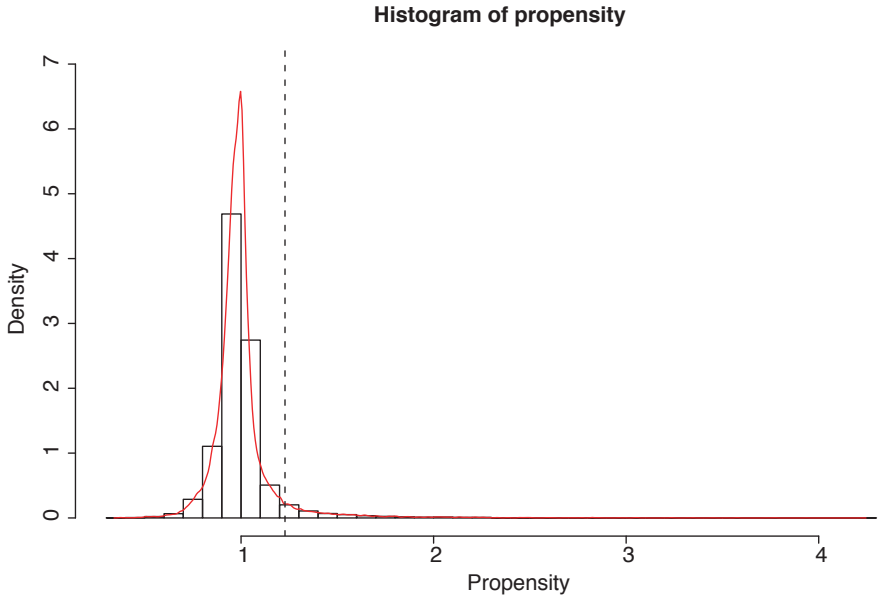

Figure 4.14 | Histogram and probability density distribution for gene propensity values in 33 major human organs and tissues. The dotted line represents the 95th percentile that was used as the threshold for defining tissue-specificity.

to select class or tissue-specific expressed genes based on their expression values. The propensity measures how a gene's expression level compares with the expression of genes in the tissue, and also the gene's expression across all other tissues. A high propensity value indicates that the gene is relatively specific in a given tissue, whereas low propensity suggests that the gene is non-specific. Note that a given gene will have individual propensity values for each tissue that it is expressed in.

We calculated propensities for the GNF dataset, using the mean expression level for biological replicates (Figure 4.13). The distribution of propensity values is centred about 1, indicating that most genes are non-specific in most tissues. There is a long tail of higher propensities, and we select the top 5% of values to represent tissue-specific gene expression (Figure 4.14).

To validate the biological significance of this threshold we analysed the GO functional annotations of expressed genes using the FatiGO web-tool. Though non-specific genes are not enriched in any functional categories, tissue-specific genes display significant enrichments in GO functions that

A

B

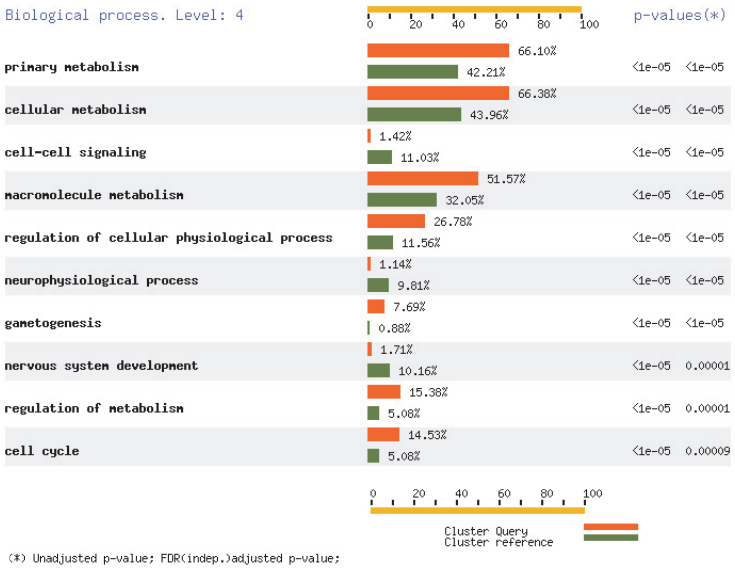

Figure 4.15 | Enriched Gene Ontology functions for tissue-specific genes. (A) Comparison of genes between whole blood (red bars) and whole brain (green) show that immune-related functions are enriched in the former, and neural development in the latter samples. (B) Comparison of genes between testis (red) and whole brain (green). The screenshot is taken from the FatiGO web-tool that was used for the analysis.

are relevant to the tissue in question. For example, a comparison of blood and brain shows that the former expresses many immune response genes, whereas the latter transcribes genes involved in neural processes (Figure 4.15 A). In another comparison, we find that reproductive genes are enriched in the testis compared with the brain (Figure 4.15 B). We observe similar trends for most other tissue-specific genes.

4.1.7 Analysis methods for tiling-array experiments

Tilling microarrays allow us to detect genome-wide gene expression and transcription factor binding in an unbiased manner. In contrast to conventional microarrays that are designed with probes in only the coding or promoter regions, tiling arrays cover the entire genome including non-coding sequences. This means that experiments using tiling arrays output different types of data, which require alternative methods for statistical analysis.

Here I present the analysis of chromatin immunoprecipitation array (ChIP-chip) experiments that were performed in collaboration with Dr Asifa Akhtar's laboratory at EMBL. The study examined the DNA-binding properties of members of the *D. melanogaster* dosage compensation complex (MSL1, MSL3, and MOF) and subunits in the nucleopore complex (Nup153 and Mtor) in male and female cell lines. Precipitates were hybridised to the Affymetrix GeneChip Drosophila Tiling 2.0R Array in triplicate. Genomic DNA was used as the control. In addition, we assessed the effect of histone H4 acetylation on lysine 16 (H4K16Ac) by MOF, using an antibody that recognises the modification.

We pre-processed the raw array output using the GCRMA background correction and quantile normalisation. Genomic regions with enriched hybridisation, representing binding events, were then identified using the Bioconductor package symp (Figure 4.16). For each protein, we calculated the mean intensity of each probe across the three replicates. We then calculated the ratio of mean intensities between the ChIP and control samples. The ratios were then smoothed by averaging the signals of neighbouring probes,

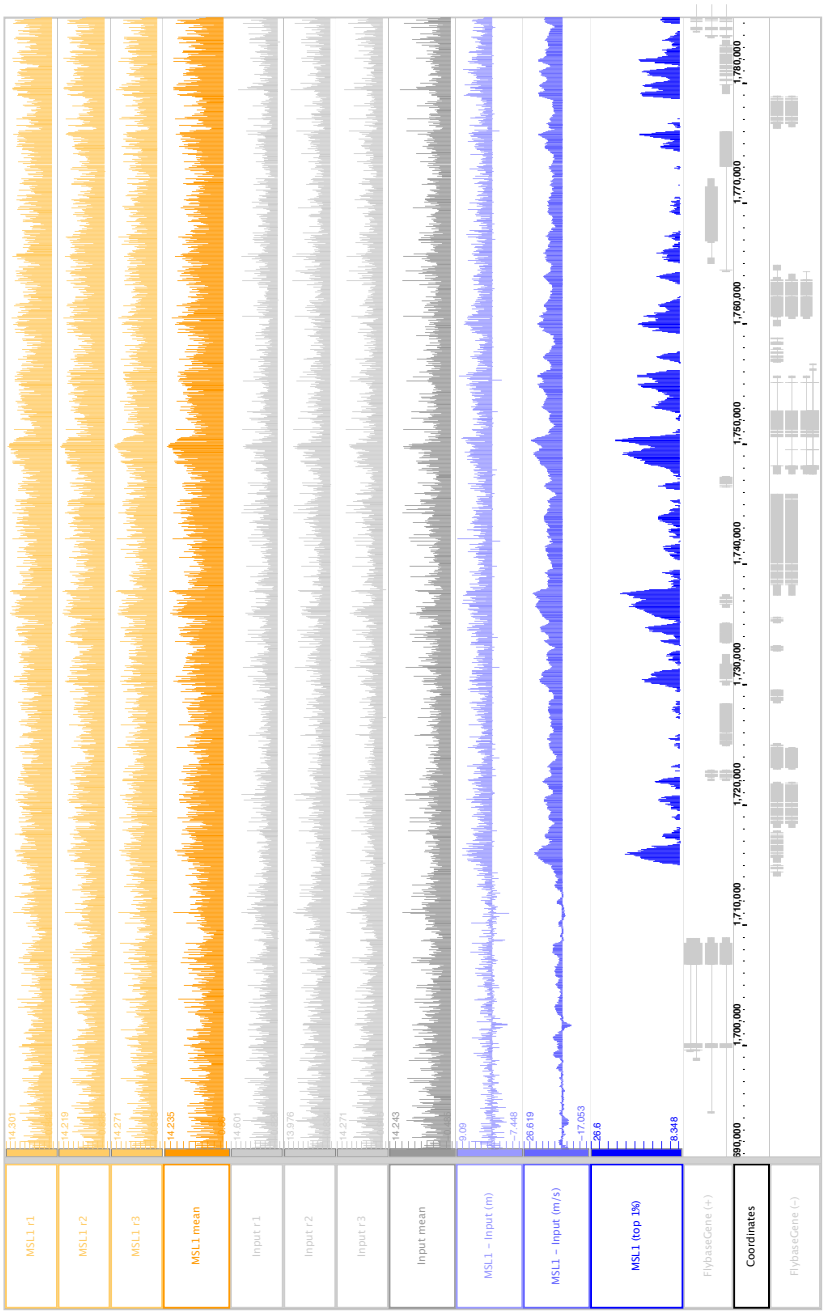

Figure 4.16 | Sample of ChIP-chip signals for MSL1-binding to the X chromosome. (A) GCRMA normalised intensity values for individual probes across three biological replicates (light orange). (B) Mean intensity values of the three biological replicates of MSL1-binding (orange). (C) GCRMA normalised intensity and mean values for the genomic DNA control (light and dark grey). (D) Ratios of MSL-binding and control mean intensity signals (light blue). (E) Smoothed ratios using a 500-bp sliding window (dark blue). (F) Top 1% ratio signal (dark blue).

using a 500bp-sliding window. Finally, we identified genomic regions that are enriched for binding by comparing the ratios on the left side of the distribution (where the signal from the control – ie noise - is larger than the ChIP sample), with the ratios on the right side (where signals from the ChIP sample – ie binding events - is larger). We assumed a symmetrical null distribution based on the ratios on the left side, and defined a binding event as any ratios extending beyond this distribution (FDR adjusted p-value <0.05). The results of these comparisons are presented in Table 4.2.

The binding profiles for MSL1 are consistent with previous reports (Gilfillan et al., 2006). However, our results show substantial differences between in the numbers of binding sites of MSL1, MSL3 and MOF, which contradicts previous findings showing colocalisation by these proteins (Morales at el., 2004). Further, we fail to identify any binding by Nup153 and Mtor, which are known to interact with the DNA, and moreover, associate with the dosage compensation complex (Mendjan et al., 2006).

These anomalies probably occur because the protocols or antibodies were optimised for MSL1. The experiments do not account for differences in DNA-binding affinities or antibody specificities, leading to insufficient binding enrichment of precipitated DNA fragments. This impacts on the analysis in several ways as: (i) all comparisons are made against the same control sample; (ii) equal thresholds are applied for identifying binding sites; and (iii) Nup153 and Mtor, which are structural proteins in the nuclear pore, are more difficult to precipitate. These lead to artefacts in which proteins appear to bind to different genomic locations, with different specificities.

We removed such biases from the analysis by ranking probe intensity ratios. By using the top 1% signal as the cut-off for significant binding, we are able to compare the binding patterns across all proteins (Table 4.3).

We used this approach to assess whether the proteins display chromosomal bias in their binding. As shown in Figure 4.17, binding is overwhelmingly favoured on the X chromosomes compared with the autosomes in male cell lines. However, this difference is abolished for MOF and H4K16 in female cell lines, indicating their activity is altered. That these biases remain when

Table 4.2 | Significant binding sites for DCC and associated proteins using symp symmetric null distribution (FDR adjusted p-value < 0.05).

	Male						Female	
Chromosome	**MSL1**	**MSL3**	**MOF**	**H4K16**	**Nup153**	**Mtor**	**Fem. MOF**	**Fem. H4K16**
2L	0	0	120	3371	1	0	0	2048
2R	4	0	124	3512	1	0	4	2552
3L	2	0	135	3578	1	0	0	2492
3R	1	2	196	4446	4	2	3	2693
4	0	0	6	268	0	0	0	249
X	700	92	707	4313	3	4	2	2254

Table 4.3 | Significant top 1% probes for DCC and associated proteins.

Sample	# tiles	# autosome	# chr. X
Male			
MSL1	29872	265 (0.9%)	29607 (99.1%)
MSL3	"	734 (2.5%)	29138 (97.5%)
MOF	"	10696 (35.8%)	19176 (64.2%)
H4K16	"	4316 (14.5%)	25556 (85.5%)
Nup153	"	16837 (56.4%)	13035 (43.6%)
Mtor	"	15639 (52.4%)	14233 (47.6%)
Female			
Kc MOF	29872	24051 (80.5%)	5821 (19.5%)
Kc H4K16	"	24355 (81.5%)	5517 (18.5%)
Kc Nup153	"	20201 (67.6%)	9671 (32.4%)
Kc Mtor	"	20553 (68.8%)	9319 (31.2%)

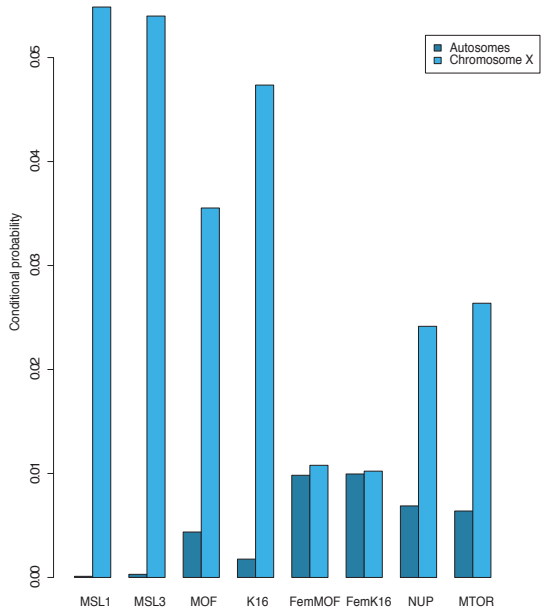

Figure 4.17 | Number of top 1% binding sites on the autosomes and X chromosome. Counts are normalised by the total number of tiles from each class of chromosomes.

we relax the threshold to the top 2, 5, 7, 10 and 15% signal suggests that the observations are robust and are independent of the cut-off (Figure 18).

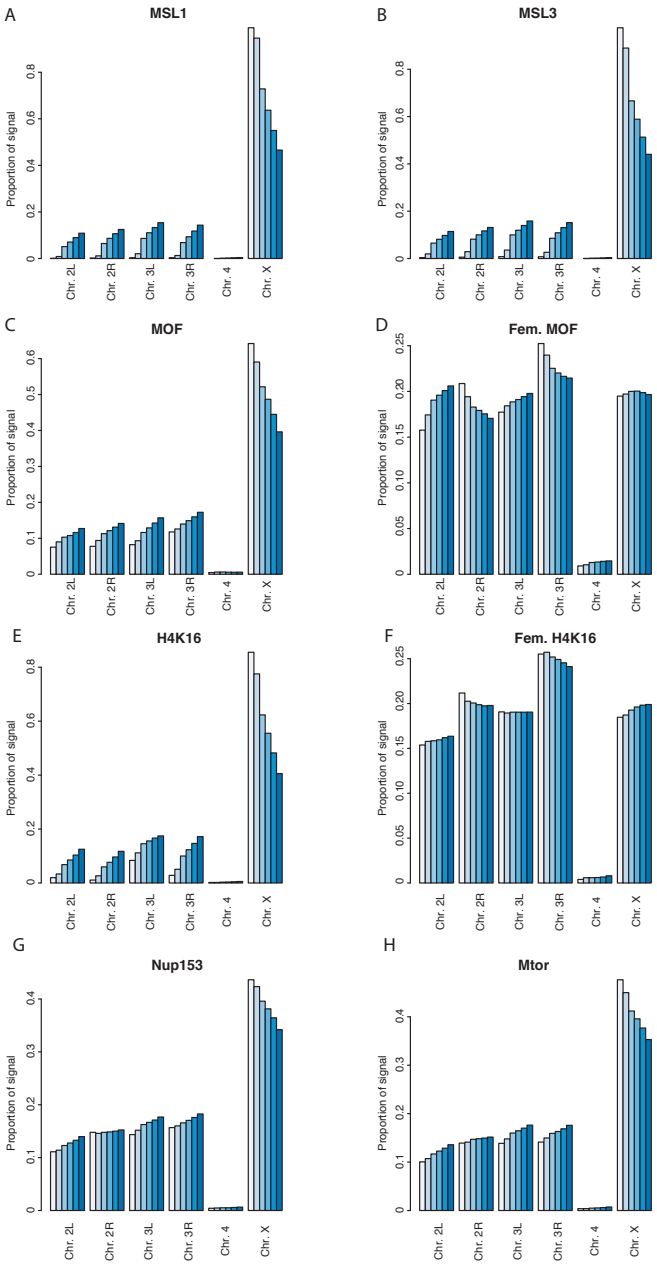

Figure 4.18 | Number of binding sites on autosomes and X chromosome. (A) MSL1, (B) MSL3, (C) MOF, (D) MOF (female cell line), (E) H4K16 acetylation, (F) H4K16 acetylation (female), (G) Nup153, and (H) Mtor. Numbers are shown for the top 1, 2, 5, 7, 10 and 15% of sites (light to dark blue).

As many of the proteins function as part of a complex, many of the binding sites are expected to overlap. Here, we restrict the binding sites to those occurring within a gene locus (Ensembl v41), and we define an overlap if two or more proteins bind to the same gene. Venn diagrams in Figure 4.19 illustrate the high degree of overlap between the MSL proteins, MOF and nuclear pore complex on the X chromosome in male cell lines. This overlap is not apparent for autosomal genes (Figure 4.19C). This confirms that the MSL and MOF proteins function together. Furthermore, it suggests that the nuclear pore complex not only associates with the DNA, but also may be involved in dosage compensation.

We also analysed the location of binding by classifying them by their occurrence in: intergenic, upstream (10kb of the 5′-exon), 5′- and 3′-UTR, intronic and exonic regions (Figure 4.20). Definitions for genomic regions were extracted from Ensembl (v41) and binding sites were mapped using genomic coordinates supplied by Affymetrix. The majority of binding is found in coding sequences and the UTRs. Of particular importance is the difference in the location of MOF binding on autosomal and X-chromosomal genes, when comparing male and female cell lines: MOF binds at both the 5′- and 3′-UTRs in the male X chromosome, but is restricted to just the 5′-UTR in male autosomes, and female cell lines.

These results suggest a mechanism of dosage compensation in which MSL1 and MSL3 on the male X chromosome actively relocate MOF to the 3′-end of genes.

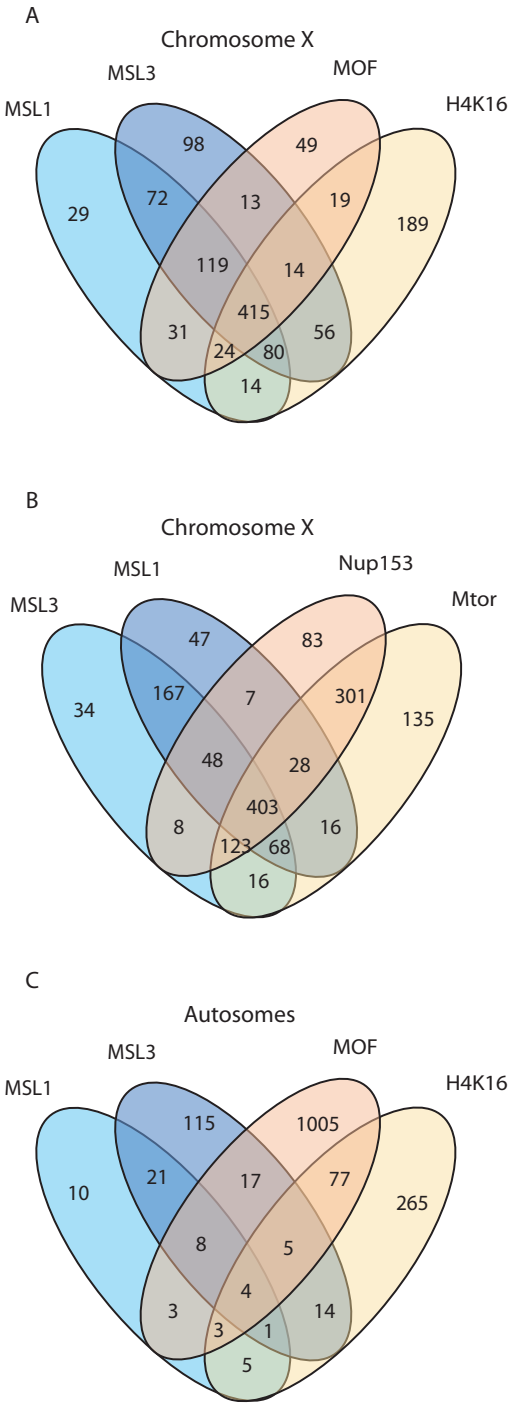

Figure 4.19 | Overlap of bound genes. Number of overlapping genes for: (A) MSL1, MSL3, MOF, H4K14 on the X chromosome; (B) MSL1, MSL3, Nup153, Mtor on the X chromosome; (C) MSL1, MSL3, MOF, H4K16 on the autosomes.

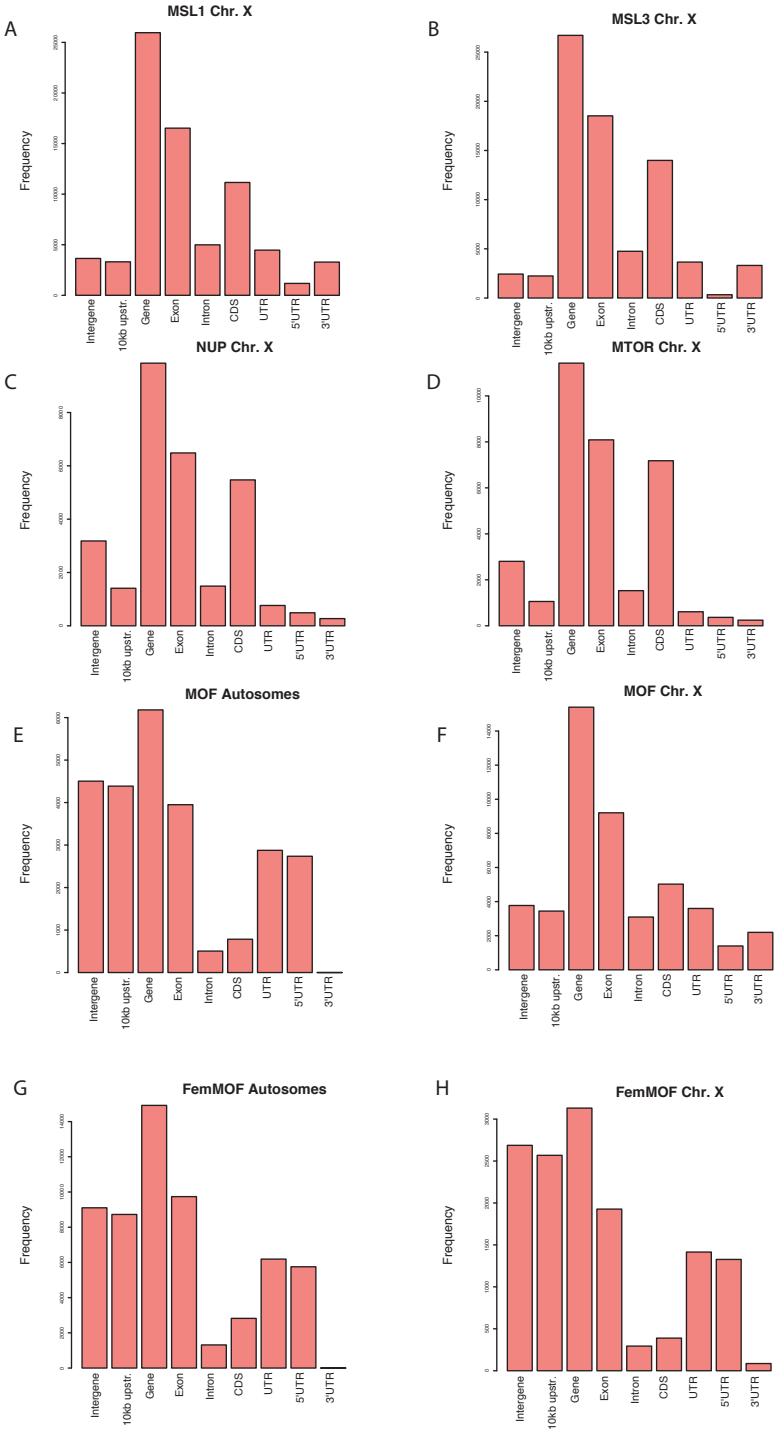

Figure 4.20 | Location of binding sites relative to gene loci. Binding sites are classified by their occurrence in: intergenic, upstream, 5′- and 3′-UTR, intronic and exonic regions.

4.2 Identification and functional characterisation of human transcription factors

The second section of this chapter describes the identification and functional analysis of sequence-specific DNA-binding transcription factors in the human genome. The analysis includes assessment of: (i) GO functional annotation; (ii) structural features; (iii) tissue-specific expression; (iv) evolutionary conservation; and (v) chromosomal location.

4.2.1 Identification of the human transcription factor repertoire

We compiled a high-quality dataset of human transcription factors. First we extracted 256 protein domains and families representing sequence-specific DNA-binding domains from the InterPro database. We manually inspected each InterPro entry using its description and associated literature citations in order to filter out non-DNA-binding domains (Table 4.4).

Next we performed a sequence similarity search using the hidden Markov model for each InterPro domain, and identified all human genes containing DNA-binding domains. This resulted in the selection of 3,848 transcripts associated with 1,932 loci. We also included 525 genes from the DBD (Kummerfeld and Teichmann, 2006) and Messina et al. (2004) datasets that were not detected in our search. This resulted in a list of 2,457 gene loci encoding for potential transcription factors.

Finally, we removed false positives by manually curating every potential transcription factor by examining information available from GeneCards (Safran et al., 2003), Entrez (Maglott et al., 2007) and UniProt (Apweiler et al., 2004). We also assessed the promiscuity of InterPro domains in terms of giving false positive matches and the combination of domains in the genes. We grouped genes into three classes: (i) probable transcription factors, which contain a non-promiscuous DNA-binding domain and have experimental evidence for transcriptional regulatory activity; (ii) potential transcription

Table 4.4 | InterPro entries describing DNA-binding domains found in the human genome. 72 different InterPro domains are found in the human genome. We have have grouped similar domains together according to the parent-child relationships defined in InterPro.

Interpro entry	Family
IPR000232	Heat shock factor (HSF)-type
IPR000327	POU
IPR000418	Ets
IPR000571	ZnF_CCCH
IPR000637	AT_hook_DNA_bd
IPR000679	ZnF_GATA
IPR000910	HMG_1/2_box
IPR000135	"
IPR000967	Znf_NFX1
IPR001275	DM_DNA_bd
IPR001346	IRF
IPR001628	Hrmn_rcpt_DNA_bd
IPR001766	TF_Fork_head
IPR002059	CSP_DNA_bd
IPR002100	TF_MADSbox
IPR002909	IPT_TIG_rcpt
IPR003118	SAM_PNT
IPR003150	RFX_DNA_bd
IPR003350	Hmoeo_CUT
IPR003619	MAD_MH1
IPR013019	"
IPR003656	Znf_BED_prd
IPR003958	CBFA_NFYB_domain
IPR004827	TF_bZIP
IPR004826	"
IPR011616	"
IPR011700	"
IPR005559	CG-1
IPR005612	CBF
IPR006986	NAB_rel
IPR006988	"
IPR006989	"
IPR007087	ZNF_C2H2
IPR007086	"
IPR010921	"
IPR007604	CP2
IPR008967	P53_like_DNA_bd
IPR011539	"
IPR012346	"
IPR011615	"
IPR013524	"
IPR009057	Homeodomain_like
IPR012287	"
IPR000005	"
IPR001005	"
IPR001356	"
IPR000047	"
IPR000747	"
IPR001827	"
IPR007103	"
IPR007104	"
IPR007106	"
IPR007107	"
IPR007108	"
IPR001647	"

Table 4.4 | *Continued.*

Interpro entry	Family
IPR010919	SAND_like
IPR000770	"
IPR011598	HLH_DNA_bd
IPR012345	STAT_DNA_bd
IPR013681	Myelin_TF
IPR002197	Other
IPR007889	"
IPR009021	"
IPR002341	"
IPR010982	"
IPR011526	"
IPR011991	"
IPR008917	"
IPR009061	"

factors, which contain a promiscuous DNA-binding domain and do not have any associated experimental evidence; and (ii) unlikely transcription factors, which have experimental evidence for molecular functions that are not compatible with transcription factor regulatory activity.

Out final dataset comprises 1,369 probable, 239 potential, and 849 unlikely transcription factor loci (Table 4.5). The remainder of the analysis uses only the probable transcription factors (Appendix B).

4.2.2 Transcription factor functional annotations in GO and PubMed

We investigated the current level of knowledge of the regulatory functions for our dataset of human transcription factors by examining: (i) abstract citations in PubMed; and (ii) annotation of biological processes in the Gene Ontology database.

First we queried PubMed to count the number of articles that cite each transcription factor in the title or abstract (Figure 4.21). We separated entries

Table 4.5 | Numbers of human transcription factors.

Transcription factor class	# genes
All hits	2,457
Unlikely	849
Potential	239
Probable	1,369

that focus on the human gene and orthologues in other species. Clearly, a small number on transcription factors dominate, such as the tumour suppressor P53 and the immune response regulator NFkB. However, the vast majority of transcription factors remain uncited.

Second we evaluated the GO "biological process" annotations of each transcription factor. Only 71 (5%) transcription factors have a GO annotation that is supported by experimental evidence (Figure 4.22 A). The percentage increases to 35% (468 genes) when we relaxed the criteria for supporting evidence to include "TAS" (traceable author statement; Figure 4.22 B). However, even when a transcription factor is annotated, very little detail is provided about the actual regulatory function (Figure 4.22 C).

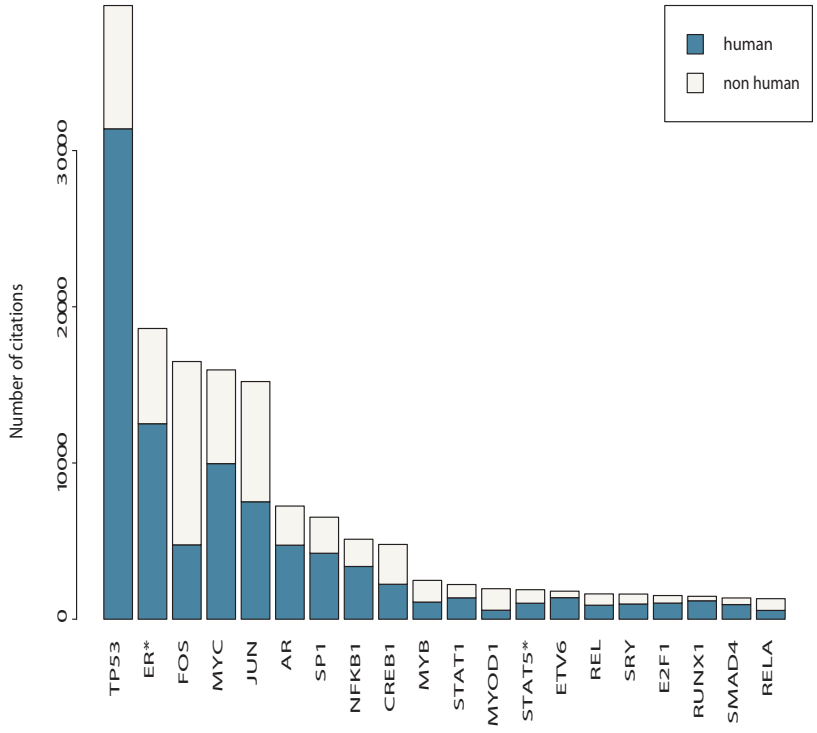

Figure 4.21 | PubMed entries for the top 20 most cited transcription factors. Blue bars represent the number of citations for studies in human and grey bars show the number of citations for all other organisms.

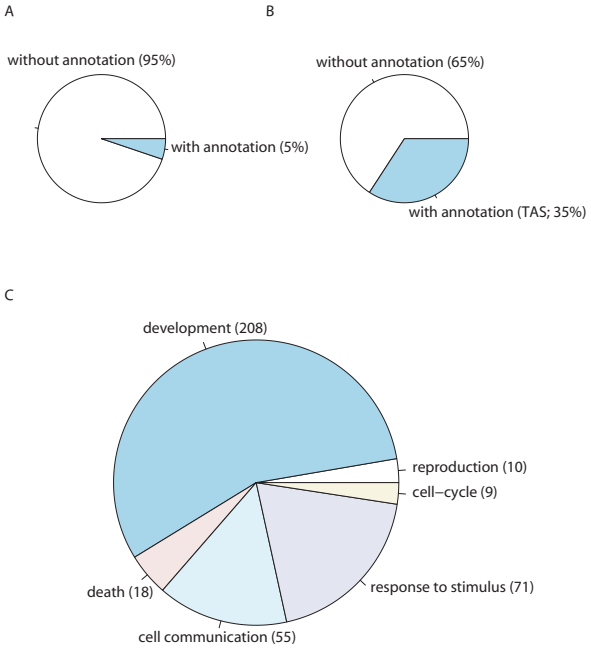

Figure 4.22 | Gene Ontology annotations of biological processes for the human transcription factor dataset. Different sources of evidence are used for annotations: (A) experimental evidence, and (B) traceable author statements (ie, statements in publications that cannot be attributed to a source). (C) The most common Gene Ontology annotations for the transcription factor dataset (number of annotated transcription factors in parenthesis).

These results demonstrate that most human transcription factors are completely uncharacterised and our dataset provides a unique opportunity for discovery by researchers interested in mammalian transcriptional regulation. In the remainder of this section I will describe computational analyses we performed to characterise the regulatory function of these transcription factors.

4.2.3 Structural classification

The most common classification for transcription factors is based on the structure of the DNA-binding domain. In figure 4.23 we show that three types of transcription factors dominate, comprising over 80% of the repertoire:

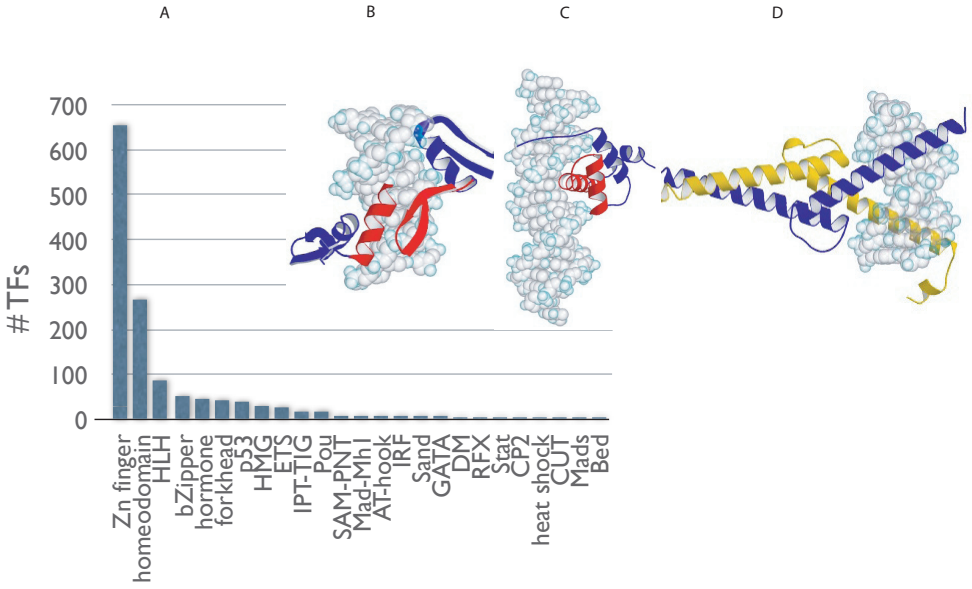

Figure 4.23 | Classification of the human transcription dataset according to the DNA-binding domain. (A) Distribution of DNA-binding domains sorted according to the number of transcription factors. Schematic illustrations of the structures of the most common DNA-binding domains complexed with DNA: (B) C_2H_2-zinc finger (PDB code: 1aay), (C) homeodomain (1fjl), and (D) helix-loop-helix (1am9) (adapted from Luscombe et al., 2000).

C_2H_2 Zn-finger, homeodomain and helix-loop-helix. As observed previously, these transcription factor families bind the DNA by inserting an alpha-helix into the DNA major groove, providing specificity though interactions with nucleotide base edges and stability via interactions with the sugar-phosphate backbone (Luscombe et al., 2000).

4.2.4 Tissue-specific expression of transcription factors

One good indicator of gene function is its expression profile. We examined the distribution of transcription factor expression in the human body by using the Genome Novartis Foundation SymAtlas dataset presented in §4.1.5.

We first examined whether levels of transcription factor expression are different to that of other genes. Figure 4.24 displays mean expression levels across the 79 biological samples, demonstrating that transcription factors

tend to be expressed at lower levels than non-transcription factor genes ($p < 10\text{-}16$).

Next we investigated the tissue-specific usage of transcription factors across 33 major tissues and organs. A transcription factor was defined as "present" if its expression value exceeded the PANP threshold for each array as described in §4.1.5. The histogram in Figure 4.25 (A) shows a large difference in the numbers of transcription factors expressed in each tissue, ranging from 331 in fetal-lung to 155 in the appendix. These differences can in part be explained by the fact that more complex tissues (in terms of cell type content, metabolic and secretory activity) require more transcription factors.

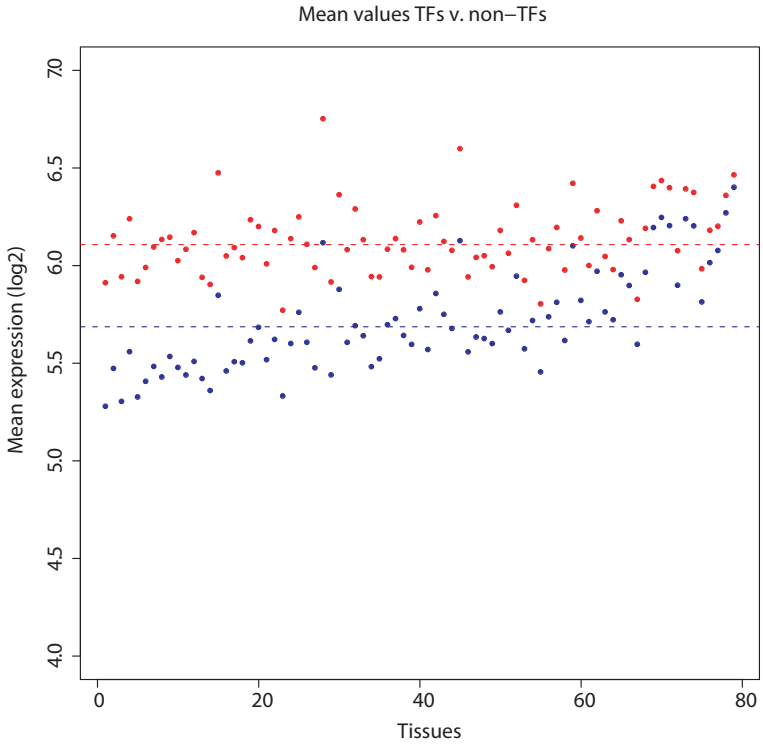

Figure 4.24 | Mean expression levels for transcription factors (blue) and non-transcription factor genes (red) across 79 human organs, tissues and cell lines. Expression data taken from the Genome Novartis Foundation SymAtlas (Su et al., 2004).

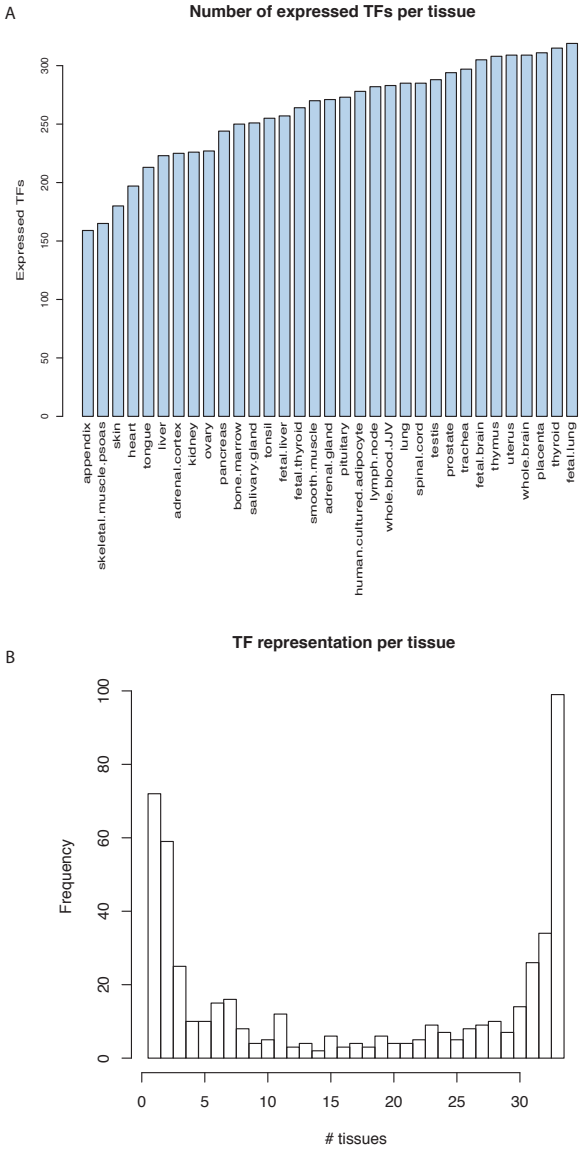

Figure 4.25 | Distributions of transcription factor expression in 33 major human organs and tissues. Histograms display the: (A) number of transcription factors expressed in each tissue, and (B) the number of tissues in which each transcription factor is expressed.

We then assessed how widely transcription factors are expressed. Figure 4.25 B shows the number of tissues in which transcription factors are expressed. The distribution is bimodal, indicating that transcription factors are generally either expressed ubiquitously across all tissues or specifically

in a small number of tissues.

Figure 4.26 displays a heatmap of 508 transcription factors that are expressed in at least one of the 33 tissues examined. We calculated propensity values (§4.1.6) to define formally whether a transcription factor is tissue-specific or general (ie, non-tissue-specific) in its expression.

We classified 165 transcription factors as having non-specific expression. Most of these are ubiquitously expressed although they are sometimes missing from certain tissues. Examples of non-specific transcription factors include well-known regulators such as C/EBP or JUND.

There are also 343 tissue-specific transcription factors, which are expected to play an important role in dictating the cell's identity. Most are expressed uniquely in only one tissue, although some are specifically expressed in several tissues (eg, transcription factors that are specifically expressed in fetal brain, adult brain and spinal cord). Finally, a few transcription factors are expressed ubiquitously but are nonetheless specific as they display higher expression levels in a particular tissue. Different tissues express diverse numbers of specific transcription factors. For example, whole blood, testis or brain contain a larger proportion of specific regulators compared with other tissues such as tongue, heart or appendix.

127 transcription factors are regulators specific to one tissue. Among these are well-known tissue-specific regulators such as the hepatocyte nuclear factor 4 (HNF-4) in liver and the Zn-finger protein ZBTB32 in testis.

Other transcription factors are shared specifically between sets of tissues. In general this is either because the tissues are related, or because there is a contamination between assayed samples. There is often a strong correlation in transcription factor expression between fetal and adult tissues. For example fetal and adult brain, lung and thyroid, share 21, 13 and 8 tissue-specific transcription factors respectively. We also observed similar patterns in the central nervous system (brain, spinal cord and fetal brain), in which transcription factors such as the calmodulin-binding transcription activator CAMTA1 are shared. Here there are also uniquely expressed transcription

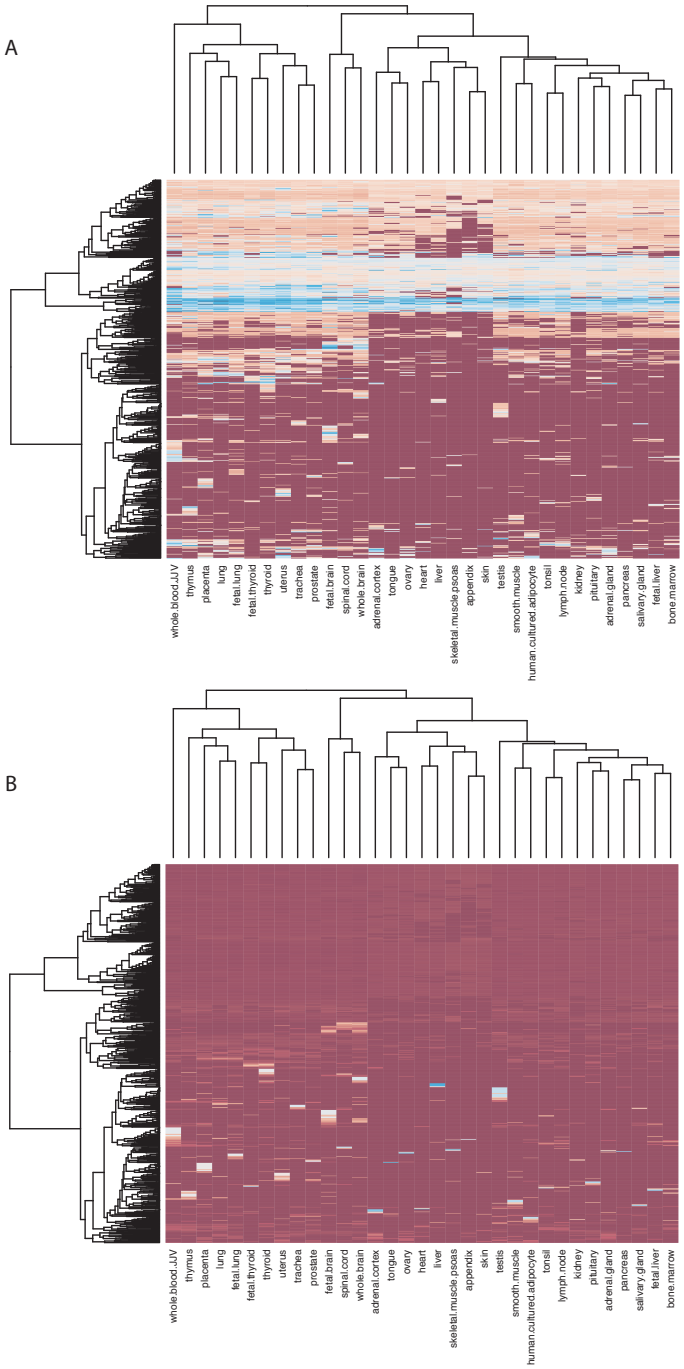

Figure 4.26 | Heatmap of transcription factor expression in 33 major human organs and tissues. Transcription factors (rows) and tissues (columns) are aligned by hierarchical clustering of expression values. Intersecting cells in the heatmaps display (A) expression values (dark red for low expression and dark blue for high expression), and (B) propensity values (same colour scale).

factors such as the proto-oncogene MYCN in fetal brain or the homeobox protein PKNOX2 in brain.

Specific transcription factors in whole blood must be treated with caution as they are likely to display shared expression owing to cross-contamination of tissue samples. Among the 60 specific transcription factors, only 22 are unique to blood and the remaining 38 are shared with 27 other tissues, such as lymph node, thymus, lung, tonsil, bone marrow or thyroid.

Entire classes of transcription factors families sometimes show similar expression profiles. The LIM-homeodomains are expressed in the brain, which agrees with previous studies from the *C. elegans* nervous system (Vermeirssen et al., 2007). Transcription factors containing myelin-like DNA-binding domains are expressed at high levels in fetal and adult brain, and spinal cord. Other families, such as the C_2H_2 Zn-fingers do not have such patterns and are both ubiquitous and specific (Figure 4.27). Notably, a significant proportion of transcription factors are not expressed in any sample, but might become tissue-specific under particular conditions such as development or different stresses.

In summary our results demonstrate a two-tier system of regulation based on global and tissue-specific transcription factors. Most transcription factors lack information regarding their regulatory function. However, our results define the conditions in which transcription factors might act and therefore suggest their function.

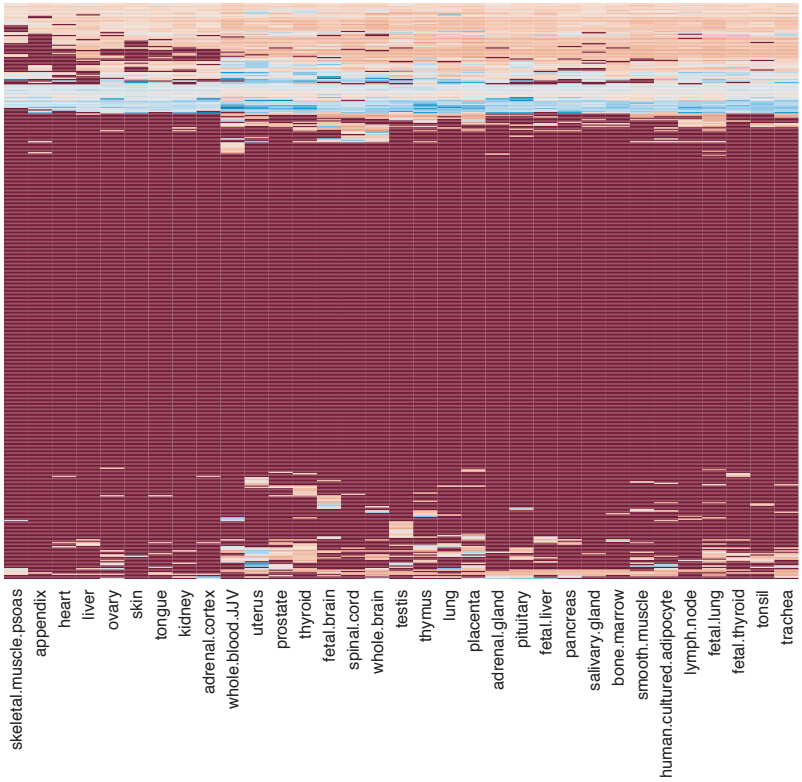

Figure 4.27 | Heatmap of expression for C_2H_2 Zn-finger transcription factor family. Transcription factors (rows) and tissues (columns) are aligned using hierarchical clustering of expression values (dark red for low expression and dark blue for high expression).

4.2.5 Evolutionary conservation of human transcription factors

We examined the evolutionary conservation of human transcription factors across other eukaryotes. We performed this by analysing the gene trees provided by Ensembl Compara (v41) for 1,369 transcription factors. Figure 4.28 shows a heat map depicting the presence or absence of transcription factor orthologues in 25 eukaryotic organisms ranging from yeast to chimpanzee. There are four clear clusters representing major expansions in the transcription factor repertoire, which occurred at important stages of evolution in the human lineage (appearance of animals, vertebrates, mammals and primates). We obtained similar results with other methods for

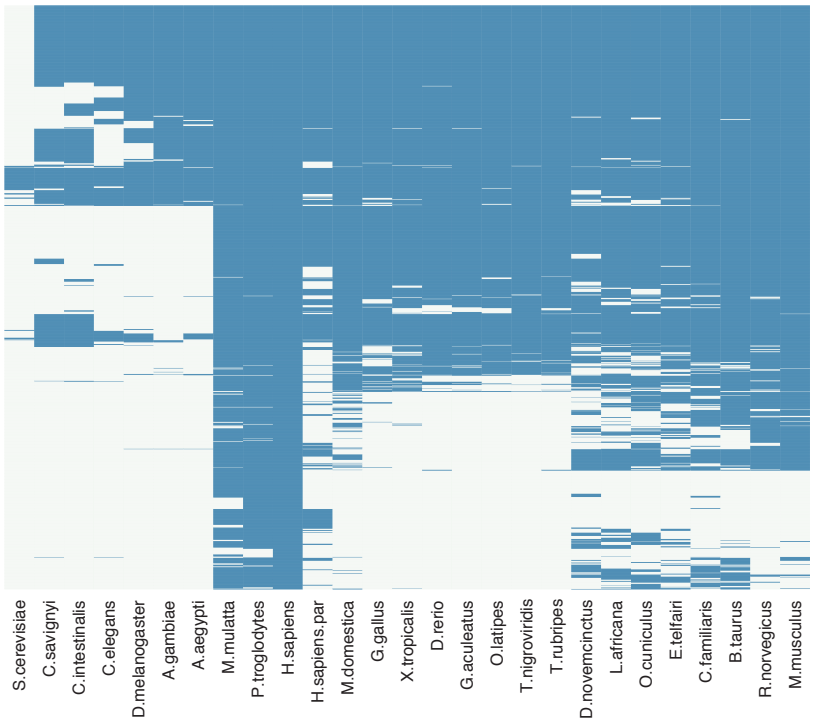

Figure 4.28 | Heatmap of transcription factor orthologues in 25 eukaryotic genomes. Transcription factors (rows) and organisms (columns) are hierarchically clustered according to the presence (blue) or absence (white) of a transcription factor orthologue. Orthologues are identified using automatically generated gene trees from Ensembl Compara. Transcription factors with human paralogues are shown next to the *H. sapiens* column.

predicting orthologues such as BLAST bi-directional best-hits, and orthology assignments by Ensembl Compara (v37) or Inparanoid (v5) (Figure 4.29). 50% of our dataset have paralogues within the human genome. Most of these are also conserved in other vertebrates, which agrees with the hypothesis of a whole genome duplication in early chordates (McLysaght et al., 2002). Other paralogues are conserved in mammals and primates suggesting more recent duplication events (Figure 4.28).

There are 69 (5%) human transcription factor orthologues in *S. cerevisiae*. These transcription factors have functions such as cell cycle control, signalling or stress response in both yeast and human suggesting roles in regulating very basic cellular functions. There are also 50 transcription factors with

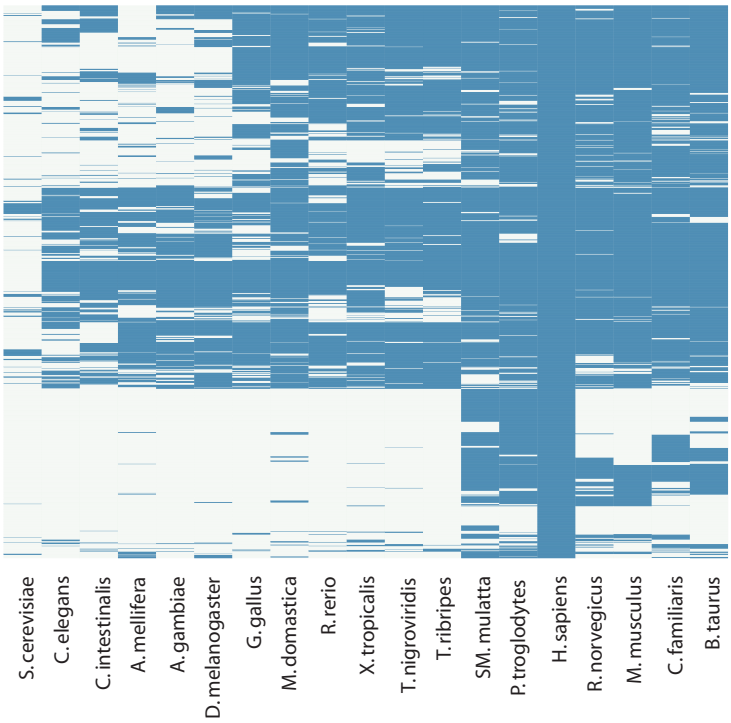

Figure 4.29 | Heatmap of transcription factor orthologues in 19 eukaryotic genomes. Orthogloues are the union of entries from Ensembl Compara and Inparanoid.

no predicted orthologues in chimpanzee or macaque; however, rather than being human-specific regulators, we find that these are artefacts caused by the poor gene annotation of non-human primate genomes.

We next focused on different transcription factor families. Several appeared at different evolutionary points; for example the STAT family is present in all animals, and the myelin-like transcription factors are only found in vertebrates. Of the three most abundant families, homeodomains and helix-loop-helix are well conserved in all animal genomes. Zn-fingers on the other hand show four stages of expansion in animals, vertebrates, mammals and primates (Figure 4.30).

Previous studies have hypothesised that there is a relationship between conservation and expression profile (Freilich et al., 2005); these propose that more evolutionary conserved genes should perform more general cellular functions (such as metabolism) and therefore have a broader expression profile. We compared the tissue specificity of transcription factor expression and the level of conservation by extracting orthology assignments from Ensembl Compara (v37) and Inparanoid (v5). To minimise false positives in the ortholog assignments, we analysed only the union from the two databases. As shown in Figure 4.31, there is no clear relationship between the degree of conservation and expression profile for transcription factors.

We also repeated the analysis by grouping transcription factors into phyletic classes (ie, primate-specific, mammal-specific, vertebrate-specific, animal-specific, and unicellular). Figure 4.32 shows there is no apparent relationship here either.

It is unclear whether this lack of relationship is because of the data quality, or the fact that transcription factors and their functions evolve rapidly and sudden expansions such as the Zn-finger family one, mask the effect.

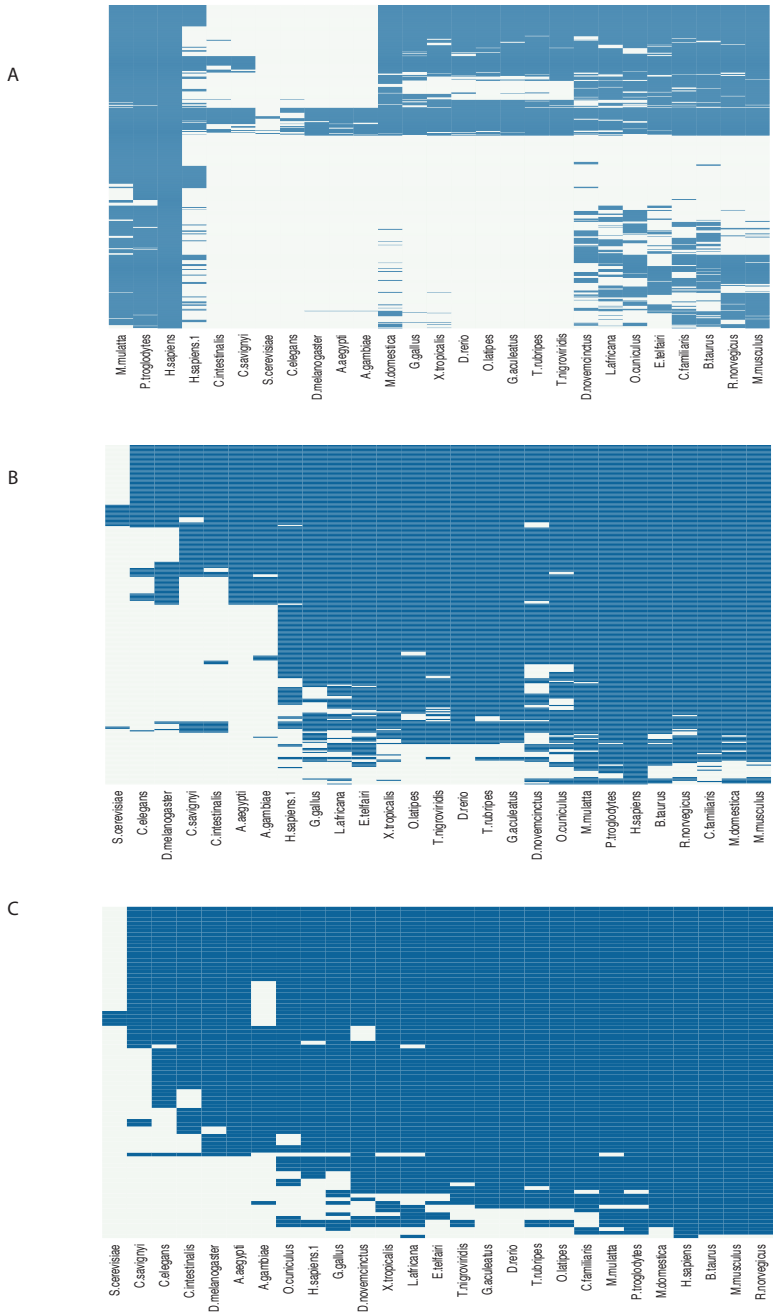

Figure 4.30 | Heatmap of orthologues for transcription factor families. (A) C_2H_2 Zn-finger, (B) homeodomain and (C) helix-loop-helix.

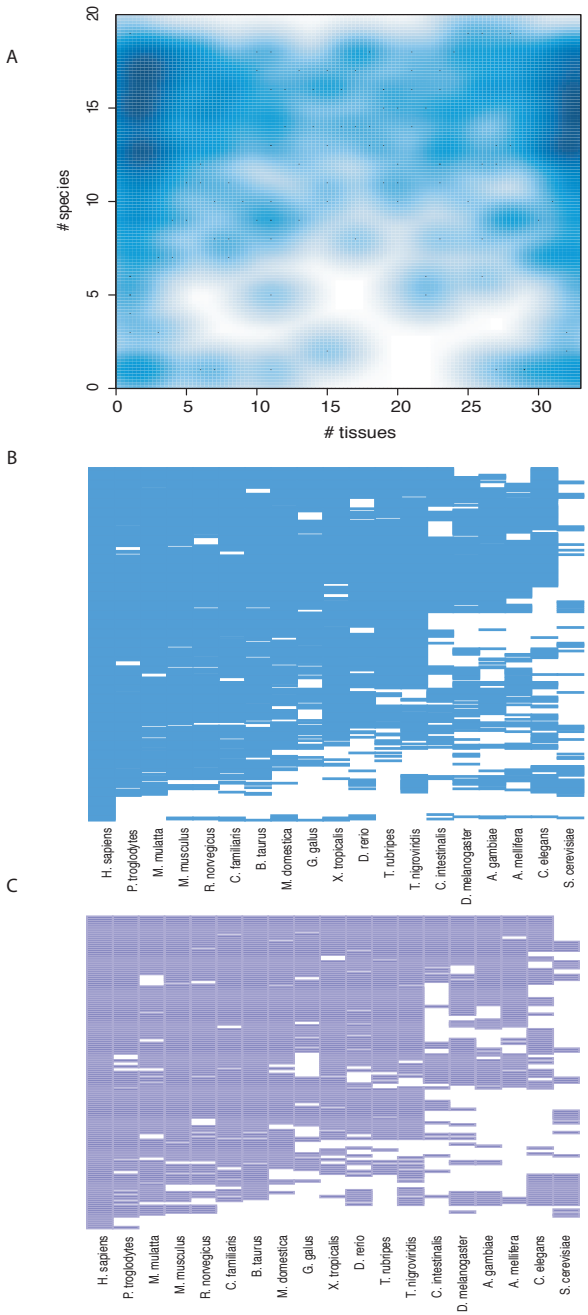

Figure 4.31 | Relationship between transcription factor expression and evolutionary conservation. (A) Scatter plot of the number of tissues in which transcription factors are expressed, and the number of organisms in which orthologues are found. Heatmap of evolutionary conservation for (B) tissue-specific and (C) non-tissue-specific transcription factors.

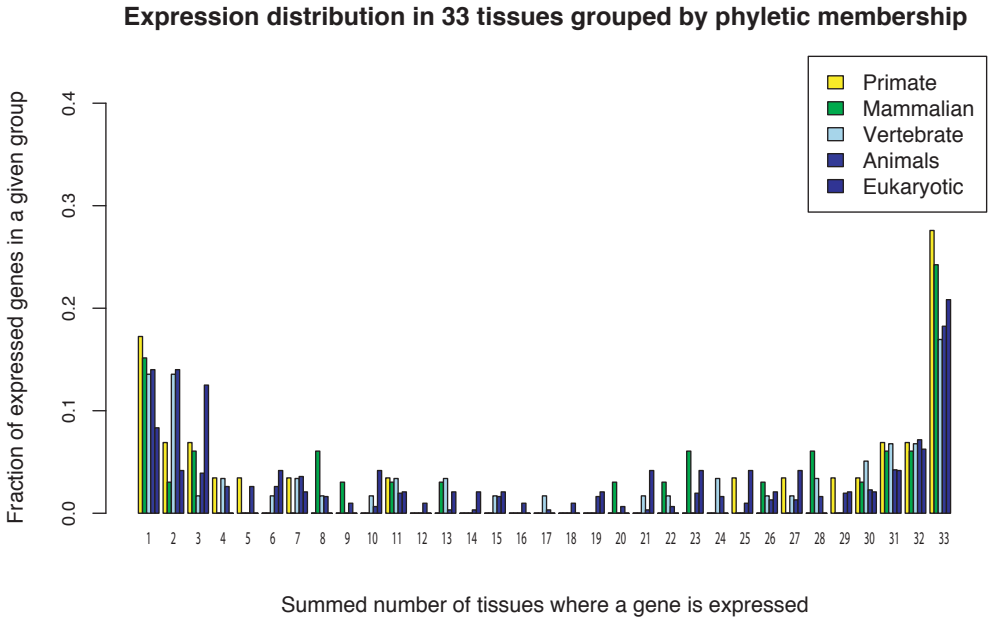

Figure 4.32 | Relationship between transcription factor expression and evolutionary conservation. Plot shows the distribution of the number of tissues in which a transcription factor is expressed, and the most distant evolutionary organism in which there is an ortholog.

4.2.6 Chromosomal location

Chromosomal location is an important characteristic of genes as it can have strong effects on gene regulation and mutation rates (Boutanaev et al., 2002; Chuang and Li, 2004; Pirrotta, 1997).

We first examined whether transcription factor genes are distributed evenly throughout the human genome. We counted the number of transcription factor genes in each chromosome and calculated whether they are over-represented compared with non-transcription factor genes in the same chromosome. Using a random permutation test, we find that only chromosome 19 is enriched with transcription factors ($p < 0.0001$; Table 4.6). In contrast, chromosomes 1-5, 13, 22 and X are depleted ($p < 0.05$; Table 4.6).

Table 4.6 | Transcription factor enriched and depleted chromosomes in the human genome. The p-values were calculated via random permutation tests for each chromosome.

chromosome	# genes	#transcription factors	pvalue
Depleted			
1	3134	113 (3.61%)	0.016
2	2149	77 (3.58%)	0.042
3	1670	57 (3.41%)	0.031
4	1287	31 (2.41%)	0.0002
5	1399	41 (2.93%)	0.004
11	1856	51 (2.75%)	0.0002
13	592	16 (2.70%)	0.024
22	733	19 (2.59%)	0.0096
X	1360	44 (3.24%)	0.022
Enriched			
19	1767	262 (14.83%)	0.0001

There are certain sets of transcription factors that are known to be co-localised in clusters; for example, the Hox A-D developmental genes on chromosomes 2, 7, 10 and 17, and the Zn-finger clusters on chromosome 19 (Eichler et al., 1998). Through clustering, it has been proposed that these transcription factors coordinate their evolution and regulatory activity (Klymenko et al., 2006).

We searched for further transcription factor clusters by sliding a 1Mb window along each chromosome. We counted the number of transcription factors within the window and compared this number against the expected numbers of transcription factors given the chromosomes' gene content (using 10,000 permutations). We applied a threshold of $p < 0.05$ (adjusted for multiple testing using FDR) to identify such clusters. The results were consistent with different window sizes (data not shown).

We detected all the previously reported Hox A-D and Zn-finger clusters. We also identify 40 new clusters, even on chromosomes that are depleted for transcription factors (Figure 4.33).

Overall, 521 (38%) transcription factors are co-localised in clusters. 23 of these clusters consist mostly of Zn-finger transcription factors (14 containing

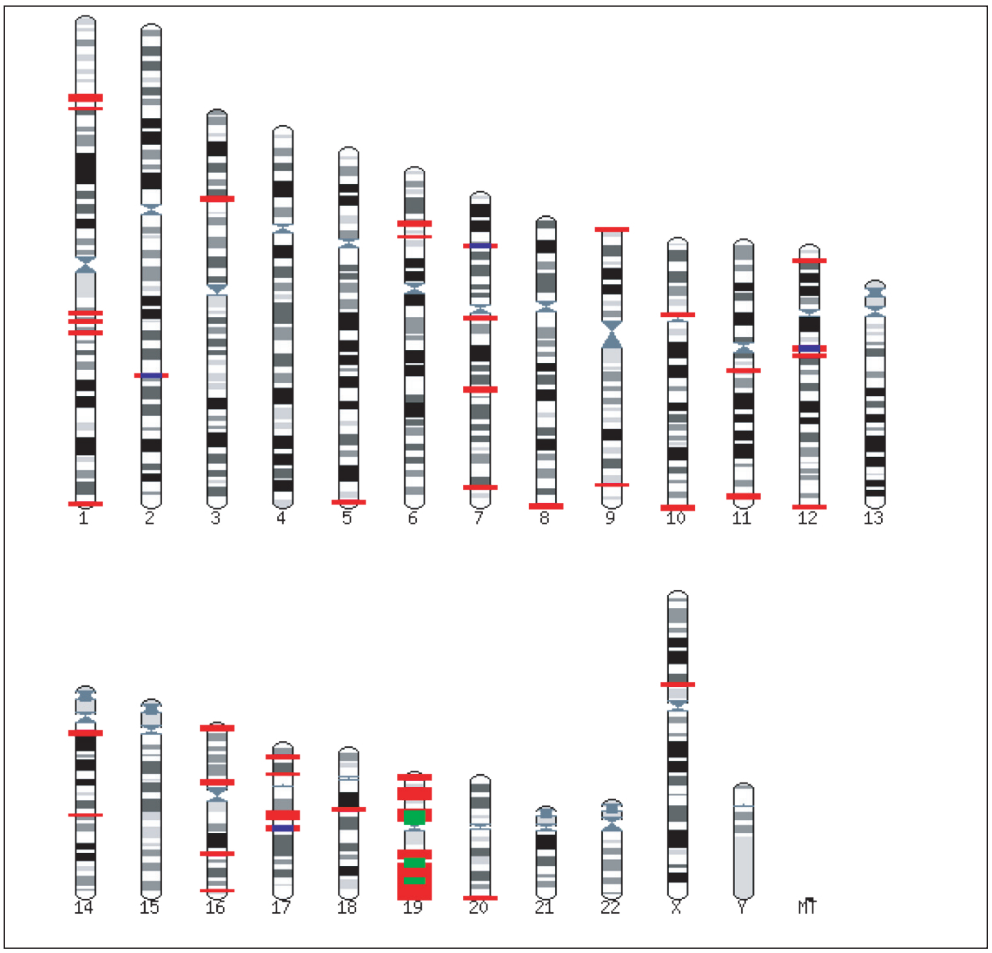

Figure 4.33 | Clusters of transcription factors in the human genome. Previously known clusters of Hox genes (blue boxes) and Zn-fingers (green) are found as well as 40 new clusters (red).

only Zn-fingers), which is the expected frequency given the abundance of this family in the genome. Eight clusters are over-represented by homeodomains, fork-head, or hormone receptor regulators, whereas 21 clusters contained a mixture of multiple transcription factor families.

17 clusters are located in peri-centromeric or sub-telomeric regions. This suggests two possibilities. First, transcription factor clusters may allow for coordinated transcriptional control. Second, these chromosomal regions have been reported as hot-spots for gene duplication, and transcription factor clusters may have formed following recent duplication event.

To elucidate whether clusters enable co-expression we checked their cluster-dependent expression. Although we did not detect significant correlation between transcription factors in the same cluster, we found tentative evidence suggesting coordinated repression of entire clusters, perhaps resulting from chromatin condensation (data not shown).

We then evaluated whether clusters consist of duplicated genes that inserted close to the original sequence. In 5 out the 52 clusters (10%), all members of the cluster are paralogues. In the majority of cases however, most paralogous genes are located in a different region of the genome.

We then assessed when clusters were formed by examining the proportion of transcription factors with ancestors only above the catarrhini and eutheria clades. We find that the clusters in sub-telomeric and peri-centromeric regions have slightly higher proportions of newer transcription factors, although the significance was marginal ($p = 0.0541$).

4.3 Identification and characterisation of functional SNPs

Single nucleotide polymorphisms (SNPs) are the most common type of DNA sequence variation in the human genome. This section examines the potential effects of SNPs on gene expression.

SNPs are formally defined as polymorphisms occurring in at least 1% of a given population, and therefore should not produce a disease phenotype on their own. The effect of SNPs on phenotypes and disease susceptibility is of great interest. Although their effect is likely to be most marked when they occur in protein-coding sequences, they may also be important in introducing changes in the mechanisms for gene expression control. Here, we identify potential functional SNPs located in: (i) transcription factor binding sites; (ii) intron-exon junctions; (iii) exonic splicing enhancers; and (iv) triplex forming regions (Figure 4.34).

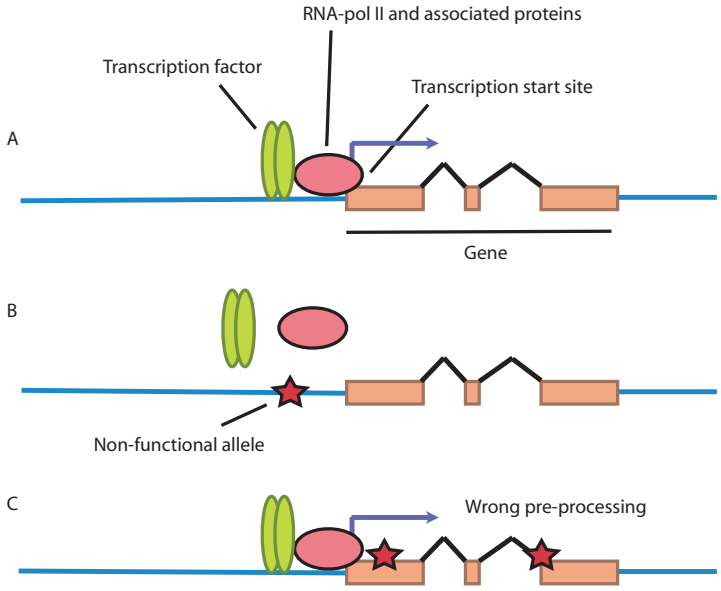

Figure 4.34 | Cartoon representation of possible mechanisms by which SNPs could affect the regulation of gene expression. (A) Representation of a normal regulatory process where the allele of a particular SNP is functional. (B) If the allele disrupts any element of the gene promoter, such as transcription factor (green) binding sites or triplex forming regions, the overall transcriptional activity of the downstream gene can be modified. (C) SNPs located in regions related to splicing, such as ESEs or splicing donor/acceptor sites can alter pre-mRNA processing and therefore the final protein level and/or function.

4.3.1 SNPs affecting transcription factor binding sites

Many transcription factors bind and regulate at the promoter of specific genes by recognising certain DNA sequence motifs. Therefore, changes in the sequence of binding site motifs by mutations or SNPs could alter the binding specificity of transcription factors and so affect the expression of the downstream target genes.

First we predicted the binding sites of 270 transcription factors in the promoter regions of the human genome. We limited the search to a 10kb region upstream of genes, as they are the most likely area for transcription factor binding and regulation. Using 330 high-quality position weight matrices from TRANSFAC, we scanned the promoter regions of 24,037

annotated Ensembl genes (v18.34) using the Match program. We applied settings that minimised the false positive rate for binding site predictions. This resulted in 2,587,478 potential binding sites.

Next we identified all Ensembl annotated SNPs located in those binding sites. 57,412 SNPs overlap with 71,444 binding sites. 19,010 genes contain at least one affected binding site in the promoter.

4.3.2 SNPs affecting splicing boundaries

Splicing is a key mechanism for removing introns from newly transcribed pre-mRNA molecules. The splicesome, which is responsible for this, recognises specific sequences at the start and end of each intron, and uses them to drive the chemical reactions that eventually result in intron excision. Dinucleotide sequences at the 5'-start (GU) and 3'-end (AG) are present in more than 99% of all introns, and are crucial to ensure correct splicing. Thus, SNPs that disrupt those sequences are likely to impact on the correct processing of mRNAs, and may even result in the expression of non-functional proteins.

We identified all SNPs situated in one of the four base position of all gene transcripts annotated in Ensembl (v18.34). This resulted in 884 SNPs located in the intron-exon boundaries of 598 genes.

4.3.3 SNPs affecting exonic splicing enhancers

The serine- and arginine-rich (SR-rich) family of proteins recognise specific exonic sequences to recruit the U1 and U2AF small nuclear ribonucleoproteins that mediate splicing. The sequences bound by SR-rich proteins are termed exonic splicing enhancers and were originally determined by SELEX experiments (Cartegni et al., 2002). SNPs located in these sequences could therefore affect pre-mRNA-processing.

We scanned all human exons using the score matrices of exonic splicing enhancers for the SR-rich proteins SF2/ASF, SC35, SRp40 and Rp55 (Cartegni et el., 2003). We then mapped the predicted enhancers to the human genome and identified SNPs that overlapped with these predictions. This resulted in

138,746 SNPs affecting 17,312 genes.

4.3.4 Triplex target sequences disrupting SNPs

A DNA triplex is a specific nucleic acid conformation produced when a duplex containing a poly-purine sequence is recognised by a triplex-forming oligonucleotide. Genomic surveys have shown that triplex target sequences are significantly over-represented in gene promoters (Goñi et al., 2004). Although their functionality is not clear, they have been proposed to act as stabilisers of promoter conformation. DNA triplexes are formed in regions with continuous tracks of poly-purines and therefore SNPs in these regions can modify the ability of a sequence to form this structure.

We predicted triplex-forming regions by searching for sequences containing at least 10 consecutive purines in the 10kb upstream and the 3'UTRs of all annotated human genes (Ensembl v18.34). We then identified 364,314 SNPs located within the predicted triplex-forming sequences.

4.3.5 PupaSNP

All predicted functional SNPs are stored in a MySQL relational database and the information is distributed through a web-based tool called PupaSNP (`http://www.pupasnp.org`; Conde et al., 2004). The web-tool also highlights SNPs that produce non-synonymous changes (ie, altering the amino acid sequence of the encoded protein), and population frequencies from the HapMap project that can be used in association studies (The International HapMap Consortium, 2003). The database is updated according to the Ensembl release cycle.

5. Discussion

5.1 Development of methods and tools for high-throughput data analysis

In order to allow researchers to interpret genome-scale data in a meaningful manner, the development of new experimental techniques often has to be accompanied by the development of associated analysis techniques. Here I have presented methods and approaches for analysing data from three major genome-scale expression and binding measurement techniques: two-colour cDNA microarrays, Affymetrix Gene Chips, and ChIP-chip coupled with tiling arrays.

5.1.1 Normalisation for two-colour cDNA microarray

Normalisation is the most important step in the microarray data analysis workflow, as one needs to ensure that the data are of high quality before making any inference from them. For two-colour cDNA arrays the main sources of bias arise from the non-linear differences in response between the different dyes, and manufacturing and hybridisation mistakes. The print-tip loess algorithm (Yang et al., 2002) allows us to correct for both errors and is therefore highly recommended.

These techniques are specialised and generally inaccessible to most laboratories without statistical expertise. Therefore I developed DNMAD, a web-tool that allows users to perform a quality check of their microarray data and then normalise them using the print-tip loess method. No previous statistical training is needed to use the tool, making it accessible to anyone who wishes to analyse their own data.

There are two major limitations of most web-based tools: (i) they are often implemented as isolated applications that do not feed into any other software; and (ii) they usually limit the amount of data that can be uploaded. DNMAD takes advantage of its integration with the GEPAS software package (Herrero et al., 2003a; Vaquerizas et al., 2005), which allows users to perform the entire microarray analysis within the same program. It is also designed to cope with a large number of microrrays thus overcoming the data capacity limitation.

5.1.2 Differential gene expression

Identifying differentially expressed genes between distinct cellular conditions has been one of the mayor areas where microarrays have been employed. Many approaches have been used to rank genes according to the amount of expression change between sets of biological samples. These range from direct comparisons using the ratio of expression from two-colours cDNA microarrays (van de Peppel et al., 2003), to more robust methods such as the use of statistical tests (Golub et al., 1999). We developed Pomelo, a tool that implements multiple statistical tests to quantify differential gene expression. We employed simple procedures such as the t-test, and Fisher's exact test for continuous and discrete data respectively. A novel aspect of this tool was the implementation of strategies to deal with the multiple-testing problem, which arises from assessing the expression of many genes at once.

However, the simple statistics implemented in this tool have difficulty in estimating the population variances when dealing with small numbers of replicate microarrays. To solve this, methods have been implemented that estimate the variance from the expression of all probes on the array rather

than individually; for example, different versions of the moderated t-test, and the empirical Bayes estimations (Baldi and Long, 2001; Lonnstedt and Speed, 2002; Smyth, 2004; Tusher et al 2001). These approaches have increased the sensitivity for detecting differentially expressed genes and are now favoured for studies containing only a small number of microarrays.

5.1.3 Building class predictors from microarray data

A major aim of microarray studies is to identify marker genes that are indicative of the cellular states of different samples. Genes grouped in a predictor could be used to predict the class membership of new uncharacterised sample based on its expression profile. Intuitively, one expects the best predictors to be those genes that display greatest differential expression. However, if two genes are significantly differentially expressed but behave similarly, having both as markers would not improve the predictive power. Similarly, including a very large number of genes in the predictor will increase the noise compromising the quality of the predictions. Therefore we require a method that balances these needs to build reliable predictors.

Another aspect to be considered when building predictors is the reported error rate. Several studies have reported predictors for different conditions; for example, van 't Veer and collaborators (2000) published a gene set to determine whether post-surgery breast tumour patients require chemotherapy. The results of this seminal paper were used to produce a commercial microarray — the Mammaprint — which is now used in clinical settings. Unfortunately, this, and other similar studies, have been criticised because the prediction rates reported in the original publications were overestimated, so resulting in not so accurate clinical predictions (Ambroise and McLachlan, 2002; Ransohoff, 2005a; Ransohoff, 2005b).

I have presented a method called TNASAS for building class predictors that combines a gene selection procedure with supervised training classification. The tool computes error rates in an unbiased manner, as it implements a full cross-validation procedure and does not predetermine the number of predictor genes to be outputted.

5.1.4 Functional annotation of co-regulated genes

By integrating functional descriptions of individual genes, we can interpret the expression data from a biologically meaningful standpoint. The most common approach is the use of Gene Ontology information (Al-Shahrour et al., 2004; Beissbarth and Speed, 2004; Dennis et al., 2003), but there are also other sources of information that may be of interest, including InterPro (Mulder et al., 2007), to detect enrichment of particular protein domain classes; and the Kyoto Encyclopaedia of Genes and Genomes (Kanehisa et al., 2006), to identify pathways which contain differentially expressed genes.

In §4.1.4 I presented Transfat, a tool that searches for over- and under-represented transcription factor-target gene relationships. Although this potentially allows us to predict the regulators of genes, it is important to note that most transcription factor binding site predictions return a large number of false positives.

5.1.5 Assessing sensitivity and specificity for Affymetrix data

It is sometimes important to determine the presence or absence of genes in the cell rather than their relative expression. Affymetrix GeneChips are hybridised using only one biological sample, and since there is no competition for probes, the resulting fluorescence can be used a measure for absolute levels of expression. One of the major sources of experimental noise is non-specific cross-hybridisation with probes, which must be considered to determine absolute gene expression measurements.

In §4.1.5 I presented an approach that estimates the true and false positive rates of detection using a set of negative strand matching probe sets and EST measurements from Unigene. By balancing the two, we are able to set objective and reliable thresholds for gene expression levels.

The use of sensible present and absent calls will also benefit differential gene expression measurements. We can filter genes that are absent from all conditions, thereby reducing noise and increasing the sensitivity of multiple-testing procedures (as the number of distinct tests, and thus the degree of correction required, is minimised).

5.1.6 Tissue-specificity selection

In addition to measuring differential expression, it is often of interest to identify genes that are uniquely expressed in a single condition. However this is not trivial when there are more than two biological samples.

In §4.1.6 I demonstrated the use of the propensity value as a measurement for tissue specificity. Propensity calculations were extensively used in structural biology to determine the amino acid composition of alpha-helices (Pace and Scholtz, 1998). Here we adapted the calculation to highlight genes that are specifically expressed in a single tissue. We validated the results by examining the functions of specific and non-specific genes which showed that propensity values allow us to identify automatically genes that contribute to the specialised functions of particular tissues.

5.1.7 Tiling arrays

Tiling arrays allow us to interrogate genomes in an unbiased manner since the entire genome is represented (Bertone et al., 2004). In §4.1.7 we developed an approach to detect DNA-binding from ChIP-chip experiments for which the binding strengths vary widely between the different proteins under consideration. We selected the top 1-15% of binding signals in each ChIP sample. Although this method does not allow us to estimate the absolute number of binding events, it enables us to detect biases in the binding site location between samples.

We detect preferential binding to the X-chromosome for subunits in the *Drosophila melanogaster* dosage compensation complex and structural proteins in the nuclear pore. We also demonstrated that these proteins tend to bind coding sequences rather than introns or intergenic regions, and that the binding occurs preferentially at the 3′-end of genes. By varying the thresholds for the binding signal we showed that the results were robust.

5.1.8 Development of high-throughput data analysis methods

With the continued flood of data from high-throughput experiments, there is a need to develop tools and methods to analyse and infer sensible conclusions from these data. Without appropriate statistical and mathematical treatment of the data, it is difficult to obtain accurate and reproducible results that can be validated in follow-up experiments.

A crucial aspect of methods development is making them available publicly, so they can be rigorously tested by the scientific community and can be developed further (Dudoit et al., 2003). A major success of this model is the Bioconductor (Gentleman et al., 2004) open-source initiative for genomic analysis software using the R statistical package. Bioconductor freely offers a multitude of software tools for analyses ranging from image processing to classification algorithms. As opposed to commercial software such as GeneSpring, the open source model allows everyone to contribute updates and challenge existing methods, leading to a product that is developed more quickly and accurately.

It is also important to take into account the interaction between wet-labs, that perform the experiments, and bioinformaticians, who interpret the results. Although biological research employs more and more genomic approaches, it is clear that most laboratories do not have the necessary expertise to perform all the analyses. These interactions will become increasingly important. Fostering a fluent and bidirectional communication between laboratories will enhance the outcome of projects. In many circumstances, as protocols and experimental techniques become standardised some of the analysis will also become routine. In such cases we should aim to automate the analysis in the form of web tools. These methods can then be used with the appropriate training by non-bioinformaticians (Fox et al., 2006).

This will not always be possible when there are very new techniques or unusual data sets involved, and novel methods will have to be developed on a case-by-case basis by experts. An example of this is described in §4.1.7,

where a specific approach was necessary. It is important that these new methods are tested using suitable benchmarks to demonstrate their validity.

5.2 Human repertoire of transcription factors

A major challenge in the post-genomic era is to understand the functions and usage of genes under different cellular conditions. Transcriptional regulation plays a central role in this as it controls the amount of mRNA produced by cells. Consequently there is a great research interest in deciphering the regulatory interactions between transcription factors and their targets. Most genomic research to date has been directed towards the prediction of transcription factor binding sites using phylogenetic footprinting and integrating the results with gene expression data (Tompa et al., 2005). Although these approaches have been used extensively, the predictions are still very unreliable and we are still at the beginning of understanding how gene regulatory networks are organised. A reliable dataset of human transcription factors that can be used in these studies is something which has been conspicuously lacking.

In §4.2 I presented a high-confidence dataset of human transcription factors identified by the presence of DNA-binding domains. The list has been manually curated to ensure a low number of false positives and a high level of coverage. The results outperform previous attempts to identify mammalian transcription factors.

The regulatory functions of the vast majority of these factors are unknown. Therefore a major aim of the thesis was to functionally characterise the repertoire of human transcription factors by integrating genomic data. We first showed that three families — C_2H_2-zinc finger, homeodomain and helix-loop-helix — dominate, comprising more than 80% of the dataset. Spatially, a large proportion of transcription factors are located in dense clusters on the chromosomes. Some of these have already been reported (Eichler et al., 1998; Lemons and McGinnis, 2006) and are known to be regulated together by chromatin remodelling (Klymenko et al., 2006), suggesting an important

form of control for coordinating transcription factor activity. Surprisingly many clusters reside in sub-telomeric and centromeric regions. These have been recently reported as regions for creating new genes, owing to the large numbers of segmental duplications (Linardopoulou et al., 2005). The presence of clusters in these areas might have contributed to the rapid expansion of C_2H_2-Zn-fingers in the primate lineage.

Gene expression data for 33 human tissues showed that different numbers of transcription factors are expressed depending on tissue complexity: more transcription factors are expressed in tissues that contain diverse cell types, are developing, or are highly active in metabolism and signal transduction. The expression data also revealed a two-tier organisation: about 80 transcription factors are constitutively expressed across all tissues, whereas around 450 factors are only expressed in one or two. These tissue-specific regulators, combined with the ubiquitous ones, specify the exact expression programme required by different cell types.

An analysis of the evolutionary conservation of human transcription factors across eukaryotic species revealed a step-wise introduction of protein families at key stages of evolution. The biggest increase in transcription factors occurred during the emergence of mammals and primates.

Several studies have reported a direct relationship between the level of evolutionary conservation of genes and their range of expression. They suggest that genes arising earlier in evolution should be involved in basic cellular functions and therefore be expressed more broadly (Freilich et al., 2005). This has been disputed in other studies however, where no correlation between tissue origin and expression levels has been reported, suggesting a high rate of evolution for tissues with common ancestral tissue and therefore no conservation of the original expression patterns (Cannon et al., 2004; Yanai et al., 2006). For our transcription factor dataset we were unable to detect any relationship between evolutionary conservation and breadth of expression, and we have yet to determine whether or not this is because the effect is masked by the rapid expansion of the Zn-fingers.

Our dataset is of very high quality and will be invaluable to any researcher interested in human transcriptional regulation. However, the analysis we performed is still basic and much more work is required to complete our understanding of regulatory systems in mammalian organisms.

Follow-up studies include integrating further sources of information such as transcription factor binding sites identified using ChIP-chip and phylogenetic footprinting. This will provide the basis for generating a regulatory network for humans. Medically it will be of interest to assess the usage of these transcription factors in diseased cells, such as tumours.

In the long-term it will be important to examine the transcriptional regulatory system in conjunction with other processes. One possible avenue for exploration is the control of transcription factors themselves, as this impacts all the downstream targets. Transcription factors are controlled by many means, such as chromatin condensation, multiple transcription initiation sites or microRNAs. By integrating the combined effects of these mechanisms we will understand more fully the general mechanism of gene expression control.

5.2.1 Implications of the evolutionary conservation of transcription factors

Due to their major importance in controlling cellular behaviour, the evolutionary conservation of transcription factors plays a fundamental role in understanding the differences between species. For example, the human and chimpanzee genomes have 99% sequence similarity (The Chimpanzee Sequencing and Analysis Consortium, 2005), but gene expression patterns vary considerably between the two organisms, in particular in tissues such as the brain, testis, and immune system (Heissig et al., 2005; Khaitovich et al., 2006a; Khaitovich et al., 2004; Khaitovich et al., 2006b). We have so far not identified any human-specific transcription factor, but it is clear that the regulatory system plays an important role in differentiating the two species.

5.2.2 Collaborations with experimental groups

The impact of the data presented here is highlighted by the fact that they have already been used by experimentalists.

One of the most important characteristics of transcription factors is their binding specificity to particular DNA sequences. Currently, we know very little about DNA sequences recognised by different transcription factors, and consensus binding sequences are available for only 123 transcription factors (Sandelin et al., 2004). Further, as these consensus sequences are obtained from a limited number of experimentally verified sites, they may be biased towards particular types of promoters. Hallikas and collaborators recently developed a method to measure transcription factor-binding affinities in an unbiased screen (Hallikas et al., 2006; Hallikas and Taipale, 2006). The method clones the entire transcription factor or its DNA-binding domain fused to a renilla luciferase protein. This is followed by competitive hybridisation in streptavidin plates of biotin-labelled consensus sequences against all possible unlabelled variants. By comparing the level of binding between the sequences, it is possible to measure the relative binding affinity of a given transcription factor. In collaboration with Prof. Jussi Taipale's group in Helsinki, this method has been extended to our transcription factor dataset. The affinities will provide a fundamental resource for identifying cis-regulatory binding sites, among many other applications.

An alternative approach for identifying binding sites is to perform ChIP-chip experiments. The CRG in Barcelona is currently generating antibodies for 300 transcription factors in our dataset, in preparation for the chromatin immunoprecipitation experiments.

Eukaryotic cells benefit from the extra combinatorial possibilities offered by different heterodimeric associations of transcription factors. This constitutes an extra level of regulation, since the transcription of each associated monomer can be modulated accordingly. Another study being considered is a two-hybrid screen of transcription factors to determine their interaction partners. This will help us to understand combinatorial regulation

at different promoters.

Together, the data from these studies will allow us to begin building a network of the human regulatory system.

5.3 SNPs Analysis

One major task after the genome sequencing is to determine what makes individuals unique. Although the majority of the genome is shared between individuals, there are always variations in the exact nucleotide sequences, except in the case of identical twins. There are several sources of sequence variation including repeats, translocations, transpositions, and single nucleotide polymorphisms. The biggest effects of these changes are observed when they occur in coding regions as they can affect the function of the proteins. In addition, variations in regulatory motifs can impact the regulatory system by affecting gene expression levels or post-transcriptional processing. Most such changes are phenotypically neutral but they can also sometime cause diseases (Walker, 2007).

In §4.3 I presented an analysis of SNPs that might have phenotypic consequences due to changes in: (i) transcription factor binding sites; (ii) triplex forming regions in gene promoters; (iii) splicing donor and acceptor sites; and (iv) spliceosomal SR-rich protein recognition sites.

The degree to which SNPs affect the transcriptional and pre-mRNA processing mechanisms will depend greatly on the location of the polymorphism. For SNPs affecting splicing sites, an alternative transcript of the gene will be produced, although the effect may be attenuated in heterozygotes. The importance of the splicing sites is highlighted by their high level of conservation as well as the low number of SNPs. For the other classes of functional SNPs considered here the effect is less well understood.

5.3.1 Human variation affects gene activity at different levels

The analysis of human variation presented in this thesis is one of the first studies assessing how single nucleotide polymorphisms can affect regulatory mechanisms controlling gene expression. Several of the predictions I made have been already experimentally validated. These include association studies in which putative functional SNPs have been linked to cell-cycle control (Belanger et al., 2005), breast cancer (Barroso et al., 2006; Fernández et al., 2006; Pooley et al., 2006), hypertension (Gong et al., 2007), depression and suicide (Lim et al., 2007) and Alzheimer's disease (Bullido et al., 2007).

Belanger and collaborators (2005) selected 127 SNPs in the promoter region of 16 genes involved in the G1/S transition of the cell cycle, 90 of which where predicted to impact on transcription factor binding sites. They used electrophoresis mobility shift assays to associate eleven of these SNPs with a gain or loss of binding affinity in transcription factor binding sites. Ultimately four promoter haplotypes were confirmed via gene reporter assays.

In another study, Lim et al. (2007) measured allele-specific transcription of the tryptophan hydroxylase isoform 2 (TPH2) in sections of human pons, and detected a functional SNP, predicted to disrupt an exonic splicing enhancer, that occurs with high frequency in normal human subjects. The non-functional version of the SNP is thought to create a truncated protein resulting in a decrease of the global expression level of TPH2. Other laboratories (Barroso et al., 2006; Bullido et al., 2007; Fernández et al., 2006; Gong et al., 2007; Pooley et al., 2006), have used our predictions to explain the significance of SNPs detected in case-control studies.

As I have mentioned previously, predictions for transcription factor binding sites, triplex forming regions or exonic splicing enhancers, suffer from a high false positive rate and this makes it difficult to find truly functional SNPs. The inclusion of evolutionary information, for example through phylogenetic footprinting, may improve predictions. In a similar fashion, the

inclusion of population frequencies in the algorithm would permit selection of polymorphisms under linkage disequilibrium, which would highlight the association of functional SNPs with particular phenotypes. These modifications have been included in later versions of the SNP selection tool (Conde et al., 2005; Conde et al., 2006).

5.4 Future work

The work presented in this thesis aims to further our knowledge of how regulatory systems control gene expression. However the work is limited by its focus on transcription factors. Nevertheless, it is a starting point from which we can progress to broader studies of cellular regulation. In particular, my mid-term goals are to integrate the transcription factor binding affinities with expression data to characterise tissue-specific development in mammalian organisms. The integration of other genomic data, such as polycomb binding, will be vital, as it will introduce other regulatory mechanisms achieved by chromatin condensation or relaxation.

Further, new data types will emerge in the next few years. One of these will be the information generated by the new sequencing technologies, such as 454 or Solexa. In particular, transcriptome-sequencing and ChIP-seq experiments are emerging as serious competitors to microarrays.

Genome-scale RNAi knockout experiments (Ashrafi et al., 2003; Baeg et al., 2005; Boutros et al., 2004) and large-scale imaging data (Neumann et al., 2006) will also contribute to a much more complex view of biological systems. These data will require the development of new analysis methods. In addition, we will face the challenge of integrating these new data types with existing ones, as only by combining measurements of different aspects of the regulatory system will we be able to decipher its full biological function.

Finally I stress that the importance of collaborations between wet and dry laboratories will continue to grow. Early genomic collaborations tended

to involve a one-way interaction in which bioinformaticians analysed the data generated by experimentalists. We are now entering an era where these analyses are generating new hypotheses, which are in turn tested by the original experimental groups, so completing a collaborative cycle.

6. Conclusions

Despite the importance of the transcriptional regulatory system to human viability, its complexity means that we are still yet to understand the system fully. Indeed we continue to make discoveries that were unexpected just a few years ago.

In this thesis, I have presented computational work investigating the nature of transcriptional regulation in the human genome. Although the study falls far short of a full description of the system, my hope is that the findings provide a useful basis for future work.

From the work presented in this PhD dissertation "Computational Approaches to Study Transcriptional Regulation in the Human Genome", we can conclude that:

1. I have identified and analysed the repertoire of DNA-binding transcription factors in the human genome. This is the first thorough study of its kind, and the dataset as well as the analyses, provide invaluable results for further genomic investigations.

We reveal that:

i) three protein families — the C_2H_2-zinc fingers, homeodomain and helix-loop-helix — dominate the repertoire of human transcription factors.

ii) 40% of transcription factors are located in clusters in the genome, which may be indicative of their coordinated activity.

iii) transcription factors are expressed either ubiquitously or specifically giving rise to a two-tier system of global and local regulators.

iv) specific families of transcription factors have expanded in the human lineage at key points during evolution, possibly contributing to our identity as a species.

2. I have predicted SNPs that disrupt the normal function of transcriptional regulation and post-transcriptional processes. In doing so, I have shown that there are thousands of natural variations in the sequence of the human genome which can potentially impact greatly on the regulatory system. The predictions will contribute to our understanding of multigenic diseases by directing experimental studies to the most promising SNPs.

3. I have developed statistical methods and software tools for the analysis of genomic data. Such resources are essential for the robust interpretation of these datasets, and our need for more tools will continue to rise with the increase in complexity and diversity of data types. Further, as more experimentalists employ genomic methods in their work, there will be a greater need to make these methods easily accessible.

References

Akhtar A, Gasser SM (2007) The nuclear envelope and transcriptional control. *Nat Rev Genet* **8**(7): 507-517

Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**(4): 578-580

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Molecular Biology of the Cell, 4th edn. New York and London: Garland Science

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**(3): 403-410

Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* **99**(10): 6562-6566

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**(Database issue): D115-119

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M,

Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**(1): 25-29

Ashrafi K, Chang FY, Watts JL, Fraser AG, Kamath RS, Ahringer J, Ruvkun G (2003) Genome-wide RNAi analysis of Caenorhabditis elegans fat regulatory genes. *Nature* **421**(6920): 268-272

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* **31**(1): 400-402

Baeg GH, Zhou R, Perrimon N (2005) Genome-wide RNAi analysis of JAK/STAT signaling components in Drosophila. *Genes Dev* **19**(16): 1861-1870

Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17**(6): 509-519

Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, Cunningham ML, Deng S, Dressman HK, Fannin RD, Farin FM, Freedman JH, Fry RC, Harper A, Humble MC, Hurban P, Kavanagh TJ, Kaufmann WK, Kerr KF, Jing L, Lapidus JA, Lasarev MR, Li J, Li YJ, Lobenhofer EK, Lu X, Malek RL, Milton S, Nagalla SR, O'Malley J P, Palmer VS, Pattee P, Paules RS, Perou CM, Phillips K, Qin LX, Qiu Y, Quigley SD, Rodland M, Rusyn I, Samson LD, Schwartz DA, Shi Y, Shin JL, Sieber SO, Slifer S, Speer MC, Spencer PS, Sproles DI, Swenberg JA, Suk WA, Sullivan RC, Tian R, Tennant RW, Todd SA, Tucker CJ, Van Houten B, Weis BK, Xuan S, Zarbl H (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* **2**(5): 351-356

Barrier A, Roser F, Boelle PY, Franc B, Tse C, Brault D, Lacaine F, Houry S, Callard P, Penna C, Debuire B, Flahault A, Dudoit S, Lemoine A (2007) Prognosis of stage II colon cancer by non-neoplastic mucosa gene expression profiling. *Oncogene* **26**(18): 2642-2648

Barroso E, Milne RL, Fernandez LP, Zamora P, Arias JI, Benitez J, Ribas G (2006) FANCD2 associated with sporadic breast cancer risk. *Carcinogenesis* **27**(9): 1930-1937

Beissbarth T, Speed TP (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**(9): 1464-1465

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. *Science* **304**(5675): 1321-1325

Belanger H, Beaulieu P, Moreau C, Labuda D, Hudson TJ, Sinnett D (2005) Functional promoter SNPs in cell cycle checkpoint genes. *Hum Mol Genet* **14**(18): 2641-2648

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B* **57**(1): 289-300

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**(5705): 2242-2246

Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* **98**(24): 13790-13795

Boeger H, Bushnell DA, Davis R, Griesenbeck J, Lorch Y, Strattan JS, Westover KD, Kornberg RD (2005) Structural basis of eukaryotic gene transcription. *FEBS Lett* **579**(4): 899-903

Boffelli D, Nobrega MA, Rubin EM (2004) Comparative genomics at the vertebrate extremes. *Nat Rev Genet* **5**(6): 456-465

Bolstad BM, Collin F, Brettschneider J, Simpson K, Cope L, Irizarry RA, Speed TP (2005) Quality Assessment of Affymetrix GeneChip Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S (eds). New York: Springer

Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI (2002) Large clusters of co-expressed genes in the Drosophila genome. *Nature* **420**(6916): 666-669

Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, Koch B, Haas SA, Paro R, Perrimon N (2004) Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science* **303**(5659): 832-835

Bowen DJ (2002) Haemophilia A and haemophilia B: molecular insights. *Mol Pathol* **55**(1): 1-18

Boyadjiev SA, Jabs EW (2000) Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clin Genet* **57**(4): 253-266

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**(6): 947-956

Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK, Bell GW, Otte AP, Vidal M, Gifford DK, Young RA, Jaenisch R (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**(7091): 349-353

Braga-Neto UM, Dougherty ER (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **20**(3): 374-380

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson

A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**(4): 365-371

Breiman L (2001) Random Forests. Machine Learning 45(1): 5-32

Brodsky AS, Meyer CA, Swinburne IA, Hall G, Keenan BJ, Liu XS, Fox EA, Silver PA (2005) Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol* **6**(8): R64

Bullido MJ, Martinez-Garcia A, Tenorio R, Sastre I, Munoz DG, Frank A, Valdivieso F (2007) Double stranded RNA activated EIF2 alpha kinase (EIF2AK2; PKR) is associated with Alzheimer's disease. *Neurobiol Aging*

Bullinger L, Rucker FG, Kurz S, Du J, Scholl C, Sander S, Corbacioglu A, Lottaz C, Krauter J, Frohling S, Ganser A, Schlenk RF, Dohner K, Pollack JR, Dohner H (2007) Gene expression profiling identifies distinct subclasses of core binding factor acute myeloid leukemia. *Blood*

Burge CB, Karlin S (1998) Finding the genes in genomic DNA. *Curr Opin Struct Biol* **8**(3): 346-354

Bushnell DA, Westover KD, Davis RE, Kornberg RD (2004) Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. *Science* **303**(5660): 983-988

Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biol* **4**: 10

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich

S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**(6): 626-635

Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3**(4): 285-298

Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res* **31**(13): 3568-3571

Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**(4): 499-509

Chakravarti A (2001) To a future of genetic medicine. *Nature* **409**(6822): 822-823

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**(5725): 1149-1154

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**(1): 65-73

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I (1998) The transcriptional program of sporulation in budding yeast. *Science* **282**(5389): 699-705

Chuang JH, Li H (2004) Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol* **2**(2): E29

Clark EA, Golub TR, Lander ES, Hynes RO (2000) Genomic analysis of metastasis reveals an essential role for RhoC. *Nature* **406**(6795): 532-535

Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Res* **34**(Web Server issue): W621-625

Conde L, Vaquerizas JM, Ferrer-Costa C, de la Cruz X, Orozco M, Dopazo J (2005) PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes. *Nucleic Acids Res* **33**(Web Server issue): W501-505

Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M, Dopazo J (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res* **32**(Web Server issue): W242-248

Cramer P, Bushnell DA, Kornberg RD (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292**(5523): 1863-1876

Crick F (1970) Central dogma of molecular biology. *Nature* **227**(5258): 561-563

Crick FH, Barnett L, Brenner S, Watts-Tobin RJ (1961) General nature of the genetic code for proteins. *Nature* **192**: 1227-1232

Darnell JE, Jr. (1982) Variety in the level of gene control in eukaryotic cells. *Nature* **297**(5865): 365-371

Darnell JE, Jr. (2002) Transcription factors as targets for cancer therapy. *Nat Rev Cancer* **2**(10): 740-749

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki

RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**(5): P3

Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm L, Reece-Hoyes JS, Hope IA, Tissenbaum HA, Mango SE, Walhout AJ (2006) A gene-centered C. elegans protein-DNA interaction network. *Cell* **125**(6): 1193-1205

DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14**(4): 457-460

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**(5338): 680-686

Diaz-Uriarte R, Alibes A, Morrissey ER, Canada A, Rueda OM, Neves ML (2007) Asterias: integrated analysis of expression and aCGH data using an open-source, web-based, parallelized software suite. *Nucleic Acids Res*

Diaz-Uriarte R, Alvarez de Andres S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**: 3

Dudoit S, Fridlyand J, P. Speed TP (2002) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association* **97**(457): 77-87

Dudoit S, Gentleman RC, Quackenbush J (2003) Open source software for the analysis of microarray data. *Biotechniques* Suppl: 45-51

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113

Eichler EE, Hoffman SM, Adamson AA, Gordon LA, McCready P, Lamerdin JE, Mohrenweiser HW (1998) Complex beta-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. *Genome Res* **8**(8): 791-808

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and

display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**(25): 14863-14868

Eisensmith RC, Woo SL (1992) Molecular basis of phenylketonuria and related hyperphenylalaninemias: mutations and polymorphisms in the human phenylalanine hydroxylase gene. *Hum Mutat* **1**(1): 13-23

Fernandez LP, Milne RL, Barroso E, Cuadros M, Arias JI, Ruibal A, Benitez J, Ribas G (2006) Estrogen and progesterone receptor gene polymorphisms and sporadic breast cancer risk: a Spanish case-control study. *Int J Cancer* **119**(2): 467-471

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* **34**(Database issue): D247-251

Firth D. (2005) CGIwithR: CGI Programming in R.

Fox JA, McMillan S, Ouellette BF (2006) A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Res* **34**(Web Server issue): W3-5

Freilich S, Massingham T, Bhattacharyya S, Ponsting H, Lyons PA, Freeman TC, Thornton JM (2005) Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol* **6**(7): R56

Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO (2001) Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol Biol Cell* **12**(10): 2987-3003

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**(12): 4241-4257

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis

B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**(10): R80

Gilfillan GD, Straub T, de Wit E, Greil F, Lamm R, van Steensel B, Becker PB (2006) Chromosome-wide gene-specific targeting of the Drosophila dosage compensation complex. *Genes Dev* **20**(7): 858-870

Gnatt AL, Cramer P, Fu J, Bushnell DA, Kornberg RD (2001) Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science* **292**(5523): 1876-1882

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**(5439): 531-537

Gong Y, Beitelshees AL, Wessel J, Langaee TY, Schork NJ, Johnson JA (2007) Single nucleotide polymorphism discovery and haplotype analysis of Ca2+-dependent K+ channel beta-1 subunit. *Pharmacogenet Genomics* **17**(4): 267-275

Goni JR, de la Cruz X, Orozco M (2004) Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res* **32**(1): 354-360

Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313**(4): 903-919

Greil F, Moorman C, van Steensel B (2006) DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase. *Methods Enzymol* **410**: 342-359

Gross P, Oelgeschlager T (2006) Core promoter-selective RNA polymerase II transcription. *Biochem Soc Symp*(73): 225-236

Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res* **29**(1): 41-43

Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**(1): 47-59

Hallikas O, Taipale J (2006) High-throughput assay for determining specificity and affinity of protein-DNA binding interactions. *Nat Protoc* **1**(1): 215-222

Hampsey M (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol Mol Biol Rev* **62**(2): 465-503

Harrison SC (1991) A structural taxonomy of DNA-binding domains. *Nature* **353**(6346): 715-719

Heissig F, Krause J, Bryk J, Khaitovich P, Enard W, Paabo S (2005) Functional analysis of human and chimpanzee promoters. *Genome Biol* **6**(7): R57

Herrero J, Al-Shahrour F, Diaz-Uriarte R, Mateos A, Vaquerizas JM, Santoyo J, Dopazo J (2003) GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Res* **31**(13): 3461-3467

Herrero J, Diaz-Uriarte R, Dopazo J (2003) Gene expression data preprocessing. *Bioinformatics* **19**(5): 655-656

Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, Snyder M (2002) Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes Dev* **16**(23): 3017-3033

Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M (2002) GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc Natl Acad Sci U S A* **99**(5): 2924-2929

Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S,

Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E (2007) Ensembl 2007. *Nucleic Acids Res* **35**(Database issue): D610-617

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18** Suppl 1: S96-104

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ (2006) The PROSITE database. *Nucleic Acids Res* **34**(Database issue): D227-230

International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011): 931-945

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**(2): 249-264

Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W (2005) Multiple-laboratory comparison of microarray platforms. *Nat Methods* **2**(5): 345-350

Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, Lemischka IR (2002) A stem cell molecular signature. *Science* **298**(5593): 601-604

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**(6819): 533-538

Janne PA, Li C, Zhao X, Girard L, Chen TH, Minna J, Christiani DC, Johnson BE, Meyerson M (2004) High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. *Oncogene* **23**(15): 2716-2726

Jones S, van Heyningen P, Berman HM, Thornton JM (1999) Protein-DNA interactions: A structural analysis. *J Mol Biol* **287**(5): 877-896

Kadonaga JT (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* **116**(2): 247-257

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34**(Database issue): D354-357

Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* **15**(7): 987-997

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* **31**(13): 3576-3579

Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**(7): 1985-1988

Khaitovich P, Enard W, Lachmann M, Paabo S (2006) Evolution of primate gene expression. *Nat Rev Genet* **7**(9): 693-702

Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, Dietzsch J, Steigele S, Do HH, Weiss G, Enard W, Heissig F, Arendt T, Nieselt-Struwe K, Eichler EE, Paabo S (2004) Regional patterns of gene expression in

human and chimpanzee brains. *Genome Res* **14**(8): 1462-1473

Khaitovich P, Tang K, Franz H, Kelso J, Hellmann I, Enard W, Lachmann M, Paabo S (2006) Positive selection on gene expression in the human brain. *Curr Biol* **16**(10): R356-358

Kitano H (2002) Systems biology: a brief overview. *Science* **295**(5560): 1662-1664

Klymenko T, Papp B, Fischle W, Kocher T, Schelder M, Fritsch C, Wild B, Wilm M, Muller J (2006) A Polycomb group protein complex with sequence-specific DNA-binding and selective methyl-lysine-binding activities. *Genes Dev* **20**(9): 1110-1122

Knauert MP, Glazer PM (2001) Triplex forming oligonucleotides: sequence-specific tools for gene targeting. *Hum Mol Genet* **10**(20): 2243-2251

Kummerfeld SK, Teichmann SA (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res* **34**(Database issue): D74-81

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs

RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921

Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J (2005) Independence and reproducibility across microarray platforms. *Nat Methods* **2**(5): 337-344

Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW (1997) Yeast microarrays for genome wide

parallel genetic and gene expression analysis. *Proc Natl Acad Sci U S A* **94**(24): 13057-13062

Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**(2): 301-313

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**(5594): 799-804

Lemon B, Tjian R (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* **14**(20): 2551-2569

Lemons D, McGinnis W (2006) Genomic evolution of Hox gene clusters. *Science* **313**(5795): 1918-1922

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* **34**(Database issue): D257-260

Lim JE, Pinsonneault J, Sadee W, Saffen D (2007) Tryptophan hydroxylase 2 (TPH2) haplotypes predict levels of TPH2 mRNA expression in human pons. *Mol Psychiatry* **12**(5): 491-501

Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**(7055): 94-100

Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipursky L, Darnell J (2004) Molecular Cell Biology, 5th edn. New York: WH Freeman.

Lonnsted I, Speed TP (2002) Replicated microarray data. *Statistica Sinica* **12**: 31-46

Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. *Genome Biol* **1**(1): REVIEWS001

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**(7006): 308-312

Luscombe NM, Thornton JM (2002) Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol* **320**(5): 991-1009

Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **35**(Database issue): D26-31

Marmorstein R, Carey M, Ptashne M, Harrison SC (1992) DNA recognition by GAL4: structure of a protein-DNA complex. *Nature* **356**(6368): 408-414

Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M (2003) Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A* **100**(21): 12247-12252

McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. *Nat Genet* **31**(2): 200-204

Mendjan S, Taipale M, Kind J, Holz H, Gebhardt P, Schelder M, Vermeulen M, Buscaino A, Duncan K, Mueller J, Wilm M, Stunnenberg HG, Saumweber H, Akhtar A (2006) Nuclear pore components are involved in the transcriptional regulation of dosage compensation in Drosophila. *Mol Cell* **21**(6): 811-823

Messina DN, Glasscock J, Gish W, Lovett M (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res* **14**(10B): 2041-

2047

Morales V, Straub T, Neumann MF, Mengus G, Akhtar A, Becker PB (2004) Functional integration of the histone acetyltransferase MOF into the dosage compensation complex. *EMBO J* **23**(11): 2258-2268

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) New developments in the InterPro database. *Nucleic Acids Res* **35**(Database issue): D224-228

Naef F, Hacker CR, Patil N, Magnasco M (2002) Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biol* **3**(4): RESEARCH0018

Neumann B, Held M, Liebel U, Erfle H, Rogers P, Pepperkok R, Ellenberg J (2006) High-throughput RNAi screening by time-lapse imaging of live human cells. *Nat Methods* **3**(5): 385-390

Noble WS (2006) What is a support vector machine? *Nat Biotechnol* **24**(12): 1565-1567

O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33**(Database issue): D476-480

Pace CN, Scholtz JM (1998) A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J* **75**(1): 422-427

Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, Lilja P, Mukherjee G, Oezcimen A, Rayner T, Rocca-Serra P, Sharma A, Sansone S, Brazma A (2005) ArrayExpress--a public repository for microarray

gene expression data at the EBI. *Nucleic Acids Res* **33**(Database issue): D553-555

Pirrotta V (1997) PcG complexes and chromatin silencing. *Curr Opin Genet Dev* **7**(2): 249-258

Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**(6870): 436-442

Pooley KA, Healey CS, Smith PL, Pharoah PD, Thompson D, Tee L, West J, Jordan C, Easton DF, Ponder BA, Dunning AM (2006) Association of the progesterone receptor gene with breast cancer risk: a single-nucleotide polymorphism tagging approach. *Cancer Epidemiol Biomarkers Prev* **15**(4): 675-682

Quackenbush J (2002) Microarray data normalization and transformation. *Nat Genet* **32** Suppl: 496-501

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. *Nucleic Acids Res* **33**(Web Server issue): W116-120

Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA (2002) "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* **298**(5593): 597-600

Ramaswamy S, Ross KN, Lander ES, Golub TR (2003) A molecular signature of metastasis in primary solid tumors. *Nat Genet* **33**(1): 49-54

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* **98**(26): 15149-15154

Ransohoff DF (2005) Lessons from controversy: ovarian cancer screening and serum proteomics. *J Natl Cancer Inst* **97**(4): 315-319

Ransohoff DF (2005) Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* **5**(2): 142-149

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. *Science* **290**(5500): 2306-2309

Robertson G, Bilenky M, Lin K, He A, Yuen W, Dagpinar M, Varhol R, Teague K, Griffith OL, Zhang X, Pan Y, Hassel M, Sleumer MC, Pan W, Pleasance ED, Chuang M, Hao H, Li YY, Robertson N, Fjell C, Li B, Montgomery SB, Astakhova T, Zhou J, Sander J, Siddiqui AS, Jones SJ (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res* **34**(Database issue): D68-73

Romualdi C, Campanaro S, Campagna D, Celegato B, Cannata N, Toppo S, Valle G, Lanfranchi G (2003) Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. *Hum Mol Genet* **12**(8): 823-836

Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, Shmoish M, Peter Y, Glusman G, Feldmesser E, Adato A, Peter I, Khen M, Atarot T, Groner Y, Lancet D (2003) Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res* **31**(1): 142-146

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**(Database issue): D91-94

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA (2007) Mammalian RNA polymerase II core promoters: insights from

genome-wide studies. *Nat Rev Genet* **8**(6): 424-436

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235): 467-470

Schenk PM, Kazan K, Wilson I, Anderson JP, Richmond T, Somerville SC, Manners JM (2000) Coordinated plant defense responses in Arabidopsis revealed by microarray analysis. *Proc Natl Acad Sci U S A* **97**(21): 11655-11660

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003) Human-mouse alignments with BLASTZ. *Genome Res* **13**(1): 103-107

Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D (2002) ProDom: automated clustering of homologous domains. *Brief Bioinform* **3**(3): 246-251

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**(1): 308-311

Simon R, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* **95**(1): 14-18

Smale ST, Kadonaga JT (2003) The RNA polymerase II core promoter. *Annu Rev Biochem* **72**: 449-479

Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**(1): Article 3

Smyth GK, Speed T (2003) Normalization of cDNA microarray data. *Methods* **31**(4): 265-273

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by

microarray hybridization. *Mol Biol Cell* **9**(12): 3273-3297

Steffan JS, Kazantsev A, Spasic-Boskovic O, Greenwald M, Zhu YZ, Gohler H, Wanker EE, Bates GP, Housman DE, Thompson LM (2000) The Huntington's disease protein interacts with p53 and CREB-binding protein and represses transcription. *Proc Natl Acad Sci U S A* **97**(12): 6763-6768

Straub T, Becker PB (2007) Dosage compensation: the beginning and end of generalization. *Nat Rev Genet* **8**(1): 47-57

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**(16): 6062-6067

Suzuki M, Yagi N, Gerstein M (1995) DNA recognition and superstructure formation by helix-turn-helix proteins. *Protein Eng* **8**(4): 329-338

Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* **43**(6): 1947-1958

Team R Develpment Core. (2006) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

The Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**(7055): 69-87

The International HapMap Consortium. (2003) The International HapMap Project. *Nature* **426**(6968): 789-796

Thomas MC, Chiang CM (2006) The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* **41**(3): 105-178

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* **99**(10): 6567-6572

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**(1): 137-144

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**(9): 5116-5121

Ureta-Vidal A, Ettwiller L, Birney E (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* **4**(4): 251-262

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**(6871): 530-536

van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FC (2003) Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep* **4**(4): 387-393

van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**(25): 1999-2009

Vapnik V (1998) Statistical Learning Theory, New York: John Wiley & Sons, Inc.

Vaquerizas JM, Conde L, Yankilevich P, Cabezon A, Minguez P, Diaz-Uriarte R, Al-Shahrour F, Herrero J, Dopazo J (2005) GEPAS, an experiment-

oriented pipeline for the analysis of microarray gene expression data. *Nucleic Acids Res* **33**(Web Server issue): W616-620

Vaquerizas JM, Dopazo J, Diaz-Uriarte R (2004) DNMAD: web-based diagnosis and normalization for microarray data. *Bioinformatics* **20**(18): 3656-3658

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh

E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. *Science* **291**(5507): 1304-1351

Vermeirssen V, Barrasa MI, Hidalgo CA, Babon JA, Sequerra R, Doucette-Stamm L, Barabasi AL, Walhout AJ (2007) Transcription factor modularity in a gene-centered C. elegans core neuronal protein-DNA interaction network. *Genome Res* **17**(7): 1061-1071

Walker FO (2007) Huntington's disease. *Lancet* **369**(9557): 218-228

Walsh G, Jefferis R (2006) Post-translational modifications in the context of therapeutic proteins. *Nat Biotechnol* **24**(10): 1241-1252

Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**(4): 276-287

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins

FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915): 520-562

Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**(4356): 737-738

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **35**(Database issue): D5-12

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**(1): 316-319

Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ (1997) Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat Biotechnol* **15**(13): 1359-1367

Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ, Ledley RS, Lewis KC, Mewes HW, Orcutt BC, Suzek BE, Tsugita A, Vinayaka CR, Yeh LS, Zhang J, Barker WC (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res* **30**(1): 35-37

Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* **99**: 909-917

Yanai I, Korbel JO, Boue S, McWeeney SK, Bork P, Lercher MJ (2006) Similar gene expression profiles do not imply similar tissue functions. *Trends Genet* **22**(3): 132-138

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method

addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**(4): e15

Appendix A - Resumen

A.1 Introducción

La información genética que cualquier organismo utiliza para crear el conjunto de proteínas que le caracteriza está contenida en su genoma. Estos están formados por largas moléculas de ADN que forman una doble hélice donde la información está codificada mediante secuencias de nucleótidos (Watson and Crick, 1953). En estas moléculas de ADN, podemos encontrar determinados fragmentos, los genes, que contienen la secuencia necesaria de nucleótidos para producir una proteína. Dependiendo de los niveles de expresión de diferentes proteínas cada célula del organismo adquiere una identidad propia.

Las proteínas se producen mediante una serie de procesos que están muy regulados. El primero, denominado trascripción, consiste en la utilización de un segmento determinado de ADN como molde para la creación de una molécula de ARN mensajero, que será procesada y posteriormente traducida por los ribosomas para crear una proteína (revisado en Alberts et al., 2002).

El segundo paso se denomina traducción y es el mecanismo mediante el cual un ARN mensajero es utilizado como molde para concatenar la secuencia de nucleótidos que conforma una proteína. Este proceso es posible gracias a

la correspondencia entre tripletes de nucleótidos y aminoácidos y señales de parada de la traducción (revisado en Lodish et al., 2003).

Esta cadena polipeptídica se pliega en una estructura tridimensional que puede continuar siendo procesada mediante modificaciones post-transduccionales. Estas modificaciones pueden tener un efecto en la actividad de la proteína, si esta ha sido producida en forma de precursor, o en su localización celular (revisado en Lodish et al., 2003).

Los niveles de proteína producidos son controlados mediante diferentes mecanismos en cada una de estas de etapas. De esta manera, el nivel de compactación de la cromatina, la regulación del nivel de expresión, o el procesamiento del ARN afectan a la fase transcripcional. Otros mecanismos, como por ejemplo la exportación del ARN del núcleo al citoplasma, el nivel de degradación del ARN o la eficiencia de la traducción, controlan la etapa transduccional (Darnell, 1982). Aunque todos estos mecanismos tienen importancia respecto al nivel final de proteína activa producida, el control de la trascripción, en la base del proceso, es crucial para el correcto funcionamiento de la célula, ya que de el dependen el resto de fases.

La enzima responsable de transcribir los genes es la ARN polimerasa, de la cual hay varias isoenzimas entre procariotas y eucariotas. Esta enzima no es capaz de reconocer secuencias especificas en el ADN, y por tanto si no hubiese un mecanismo que guiase a partir de donde tiene que transcribir, los ARN producidos serian aleatorios. Sin embargo, los genes son transcritos normalmente a partir de una región situada inmediatamente anterior al gen, el promotor, lo que demuestra que hay mecanismos que permiten a la polimerasa reconocer donde están los genes. Las proteínas que permiten este reconocimiento se denominan factores de trascripción, que se pueden clasificar en: (i) factores de trascripción generales; y (ii) factores de trascripción específicos de secuencia.

Los factores de trascripción generales, aunque alguno puede reconocer secuencias en el DNA, como por ejemplo TBP (de TATA *binding protein*), son los encargados de reclutar a la polimerasa y a otros factores asociados y formar el complejo de pre-iniciación. Posteriormente están implicados en

la iniciación, elongación y terminación de la trascripción. Estos factores son capaces de producir niveles basales de expresión génica (Hampsey, 1998; Thomas and Chiang, 2006).

Los factores de trascripción específicos de secuencia, por el contrario, son los encargados de promover la transcripción o reprimirla en función de las condiciones particulares de cada célula. Estos factores se unen específicamente a determinadas secuencias, lo que les proporciona la posibilidad de reconocer regiones específicas del genoma (Kadonaga, 2004). La interacción entre los factores de trascripción específicos de secuencia y la molécula de ADN ha sido estudiada en detalle. Esto ha permitido determinar una de las maneras mas frecuentes de interacción entre factores de transcripción y el ADN, que consiste en el intercalado de una alfa-hélice proteica en el surco mayor del ADN. De esta manera, los factores de trascripción pueden reconocer secuencias específicas de nucleótidos (revisado en Luscombe et al., 2000).

Existen muchas otras proteínas que se unen específicamente al ADN, como por ejemplo co-factores y enzimas modificadoras de histonas. A pesar de que muchas estudios agrupan estas proteínas en el conjunto de factores de trascripción de un organismo, has sido excluidas del trabajo presentado en esta tesis, ya que el mecanismo por el que ejercen regulación transcripcional es muy diferente.

La importancia de la regulación transcripcional se pone de manifiesto en el gran numero de enfermedades asociadas a fallos en estos mecanismos de control como por ejemplo el cáncer (Darnell, 2000) o enfermedades durante el desarrollo embrionario (Boyadjiev and Jabs, 2000). Por lo tanto, es crucial caracterizar y entender los mecanismos que las células emplean para regular la trascripción.

La genómica

Los avances en biología molecular se han basado tradicionalmente en aproximaciones reduccionistas en las que sistemas biológicos complejos son separados en sus componentes mas simples para ser estudiados en el mayor

detalle posible. Una vez que estos componentes has sido caracterizados, sus características y funciones se intentan integrar para explicar el funcionamiento del sistema. Este método ha sido muy efectivo para determinar la función molecular de muchos genes y las causas de enfermedades como la fenilcetonuria (revisado en Eisensmith and Woo, 1992) y la hemofilia (revisado en Bowen, 2002). No obstante, este método se basa en que los diferentes componentes se comporten de la misma manera cuando actúan en solitario o con el resto de componentes. Se ha demostrado que esto no es cierto en numerosos casos y que las interacciones entre los componentes de un sistema suelen estar acompañadas de nuevas funcionalidades (Kitano, 2002). Este es quizá uno de los motivos por los que las técnicas tradicionales de biología molecular has tenido un éxito limitado a la hora de entender sistemas complejos como el cáncer o las enfermedades neurodegenerativas, que conllevan la actividad combinada de muchos genes.

Una aproximación complementaria es estudiar sistemas biológicos desde un punto de vista global empleando técnicas de genómica. La genómica se puede definir como el análisis de genomas completos de organismos y la integración de la funcionalidad de cada uno de los componentes para crear una visión global del sistema. Debido a la enorme cantidad de datos a integrar en análisis genómicos, es completamente necesario el uso de técnicas de bioinformática y biología computacional. Estas técnicas nos permiten tratar de manera matemática, y, automáticamente, la información disponible para obtener resultados relevantes desde un punto de vista biológico.

La genómica ha revolucionado en los últimos años la manera en la que afrontamos preguntas biológicas ya que nos permite evaluar el comportamiento de sistemas biológicos al completo. Un ejemplo reciente de como aproximaciones genómicas han cambiado radicalmente nuestro conocimiento sobre sistemas biológicos es el estudio de promotores de genes en mamíferos (Carninci et al., 2006). La visión clásica de estos promotores se derivó a partir de estudios en procariotas, cuyos genes presentan normalmente una región rica en AT, llamada caja TATA, que se sitúa alrededor de 30 pares de bases delante del sitio de inicio de trascripción. Carninci y colaboradores demostraron que la mayor parte de genes de eucariotas superiores carecen

de una caja TATA en su promotor. Adicionalmente también demostraron que los genes sin caja TATA poseen múltiples sitios de inicio de la trascripción en vez de un único sitio.

Técnicas de alto rendimiento

Las aproximaciones a escala genómica han sido posibles gracias al desarrollo de diferentes tecnologías de alto rendimiento que nos permiten escalar los experimentos clásicos de biología molecular, como por ejemplo los northern blots, o experimentos de knock-out condicional, de un solo gen por experimento a varias decenas de miles por experimento. La secuenciación de diversos genomas, como por ejemplo el de humano, ratón o chimpancé, ha contribuido enormemente a la consecución de este objetivo, ya que una vez que conocemos la secuencia de ADN para una especie determinada es posible determinar sus genes y diseñar experimentos para investigar su comportamiento bajo determinadas condiciones (Lander et al., 2001; The Chimpanzee Sequencing and Analysis Consortium, 2005; Venter et al., 2001; Waterston et al., 2002).

Matrices de ADN

Una de las primeras tecnologías desarrolladas fueron las matrices de ADN (generalmente conocidos como *microarrays*), que permiten la monitorización de los niveles de expresión de decenas de miles de genes al mismo tiempo. Esto se consigue mediante el diseño de sondas de ADN de cadena simple con secuencias especificas para interrogar determinados genes (Schena et al., 1995). Estas sondas se fijan físicamente a un soporte que nos permite evaluar la presencia de cada gen en determinadas condiciones celulares. Este tipo de tecnología se ha empleado para detectar genes involucrados en el ciclo celular (Spellman et al., 1998), genes que permiten pronosticar el desarrollo de enfermedades (van't Veer et al., 2002), o para determinar los niveles de expresión de genes en tejidos sanos (Su et al., 2004) entre otras muchas

aplicaciones.

Recientemente, y gracias a avances en el proceso de fabricación de los *microarrays*, la técnica ha sido escalada y es capaz de interrogar no solo regiones codificantes del genoma sino también regiones intergénicas, como por ejemplo los promotores de los genes. A su vez, los *microarrays* se han acopado a diferentes procesos, tales como inmunoprecipitación de la cromatina, lo que permite detectar interacciones específicas entre factores de trascripción y el ADN (Boyer et al., 2005; Cawley et al., 2004; Horak et al., 2002a,b; Iyer et al., 2001; Martone et al., 2003; Ren et al., 2000).

Técnicas de secuenciación

La capacidad para secuenciar ácidos nucleicos ha mejorado sustancialmente en los últimos 30 años. Esto ha permitido poder secuenciar diferentes individuos de una determinada especie así como diversos organismos modelo. La disponibilidad de estas secuencias nos permite estudiar la variación entre miembros de la misma especie, así como entre distintas especies. El primer tipo de variación, entre miembros de una misma especie, permite a la especie adaptarse fácilmente a los cambios en el entorno en forma de población. Existen diferentes mecanismos causantes de la variación del ADN intra-especie tales como la recombinación, duplicaciones, inserciones, delecciones y mutaciones. El tipo mas frecuente de variación entre individuos de una misma especie son las mutaciones puntuales, que se propagan en la población si no son fatales y si ocurren en la línea germinal. Cuando estas mutaciones aparecen en mas un 1% de una población se les denomina polimorfismos de cambio único de base (*single nucleotide polymorphism*, SNP; Chakravarti, 2001). Los individuos de una especie pueden tener más predisposición a sufrir una enfermedad o a tolerar de manera diferente distintas condiciones dependiendo de la combinación de alelos de diferentes SNPs.

Genómica y regulación transcripcional

En los últimos años, los mecanismos de regulación transcripcional han sido estudiados extensivamente desde un punto de vista genómico. La mayor parte de estudios experimentales se han realizado en levadura, ya que es un organismo que tiene altos niveles de conservación respecto a otros eucariotas y con el cual es fácil experimentar. Estos estudios han consistido en estudiar la expresión de los genes en diferentes condiciones como por ejemplo las diferentes fases del ciclo celular o daño al ADN (Chu et al., 1998, DeRisi et al., 1997). Estas aproximaciones se han complementado con métodos experimentales para determinar sitios de unión de factores de trascripción in vivo mediante inmunoprecipitación de la cromatina acoplada a *microarrays* de ADN. La combinación de estos datos junto con diferentes técnicas computacionales ha permitido recrear la red de regulación transcripcional de levadura y determinar que partes están activas en diferentes condiciones (Luscombe et al., 2004).

En humano sin embargo, la mayor parte de estudios que han empleados técnicas de genómica se han centrado en comparar el transcriptoma de muestras de determinadas enfermedades y tejidos sanos. No obstante, según las técnicas experimentales han ido mejorando, ha aumentado el interés en descifrar el entramado que controla los procesos de regulación transcripcional. La mayor parte de los esfuerzos actuales se centran en entender como combinaciones de diferentes factores de trascripción regulan diferentes procesos (Lemon and Tjian, 2000), para lo cual se han empleado técnicas computacionales de análisis de secuencias (Tompa et al., 2005) que a su vez han sido complementadas mediante aproximaciones experimentales para determinar estos sitios (Boyer et al., 2005; Boyer et al., 2006; Lee et al., 2006). Pese a estos esfuerzos y al interés general en entender los mecanismos que regulan el control de la trascripción, nuestro conocimiento de estos sistemas en organismos superiores, y en particular en humano, es muy escaso.

Una de las razones detrás de esta falta de conocimiento es que siete años después de la publicación del genoma humano todavía no disponemos de una

lista de calidad de factores de trascripción en humano. Los primeros estudios estimaron un total de entre 200 y 300 factores de trascripción generales y entre 2,000 y 3,000 específicos de secuencia (Lander et al., 2001, Venter et al., 2001). Desde entonces, solo dos estudios se han centrado en determinar este conjunto de genes. El primero, realizado por Messina y colaboradores (2004), se basa en una lista de factores de transcripción anotados en Transfac y GO para predecir, utilizando dominios proteicos de unión a ADN presentes en estos genes, el repertorio completo de factores de trascripción en el genoma. Utilizando esta aproximación, los autores detectaron 1,962 factores de trascripción. Examinado superficialmente la lista de factores propuestos podemos se pone de manifiesto que esta contiene muchos genes anotados erróneamente, posiblemente debido a la deficiente definición de muchos de los dominios de unión a ADN utilizados.

En la segunda aproximación, Kummerfeld y Teichmann (2006) utilizaron una selección de dominios de unión a ADN filtrando aquellos con una baja definición. Esta aproximación resultó en una mejora en la tasa de falsos positivos respecto a Messina et al. (2004), a costa de una perdida en cuanto a la sensibilidad de la detección.

El trabajo que presento en esta tesis esta dirigido a aumentar nuestro nivel de conocimiento sobre regulación transcripcional en humano mediante el análisis de datos a escala genómica utilizando técnicas bioinformáticas. En particular el trabajo se centra en el análisis a escala global de genomas, expresión génica y variación de secuencias de ADN entre miembros de una especie y diferentes especies.

En primer lugar presento métodos estadísticos y herramientas bioinformáticas que he desarrollado y que nos permiten analizar datos a escala genómica. En segundo lugar presento la identificación y caracterización del repertorio de factores de trascripción específicos de secuencia en el genoma humano. Por ultimo presento los resultados del impacto de la variación de secuencia entre individuos de una misma especie en el control de la regulación transcripcional.

A.2 Objetivos

En concreto, los objetivos de la tesis doctoral son:

1. Desarrollar los métodos y herramientas necesarias para poder analizar e integrar los datos procedentes de experimentos de alto rendimiento a estudios sobre el control de la regulación transcripcional.

2. Identificar y caracterizar funcionalmente los factores de trascripción humanos.

3. Identificar y evaluar como los SNPs pueden afectar al proceso de control transcripcional.

A.3 Métodos

Una de las tareas más importantes en genómica y bioinformática consiste en la integración de diferentes tipos de datos que describen un sistema biológico en particular. Incorporando fuentes de información respecto a diferentes aspectos del sistema podemos construir una visión global del mismo que nos permita definir los principios básicos de su funcionamiento.

El trabajo que se presenta en esta tesis es puramente computacional. Los datos analizados se obtuvieron de las principales bases de datos publicas de genomas, proteínas y expresión génica. Estas incluyen:

(i) Bases de datos de genomas. Ensembl y las bases de datos satélite Ensembl Compara y Ensembl Variation se utilizaron como fuente de anotación genómica, conservación evolutiva y datos de SNPs, respectivamente. También se utilizaron otras bases de datos que contienen información evolutiva como por ejemplo Inparanoid.

(ii) Bases de datos de proteínas. Además de la anotación proteica que proporciona Ensembl, se utilizó la base de datos InterPro como fuente de

información sobre dominios proteicos de unión a ADN.

(iii) Bases de datos de expresión génica. Se utilizo ArrayExpress y sobretodo los datos de expresión del Genome Novartis Foundation SymAtlas (Su et al., 2004) como fuente de expresión génica en tejidos sanos en humano. Además se utilizaron otros conjuntos de datos describiendo diferentes condiciones, como por ejemplo el conjunto de datos de leucemia linfoblástica aguda y leucemia mielocítica aguda publicado por Golub et al. (1999), o el conjunto de datos sobre el complejo de compensación de la dosis génica en *D. melanogaster* procedente del laboratorio de Dr. Asifa Akhtar.

(iv) Otras bases de datos y conjuntos de datos. Estos incluyen diversas bases de datos como por ejemplo TRANSFAC o los repertorios de factores de transcripción obtenidos por Messina et al. (2004), o Kummerfeld y Teichmann (2006). También se incluyen aquí bases de datos de anotación génica como Gene Ontology o literatura científica como PubMed.

Estos datos fueron analizados mediante diferentes métodos estadísticos. Estos se pueden clasificar en: (i) métodos de estadística general; (ii) métodos de análisis de secuencias de ADN y proteína; y (iii) métodos de análisis de microarrays.

(i) Métodos de estadística general. Estos incluyen diferentes métodos estadísticos que han sido empleados a lo largo de la tesis tales como el test de la t, test exacto de Fisher, ajuste de p-valores por comparaciones múltiples o agrupamiento jerárquico.

(ii) Análisis de secuencias de ADN y proteína. aquí se han empleado métodos generales para analizar secuencias biológicas tales como BLAST (Altschul et al., 1990), Match (Kel et al., 2003) o InterProScan (Quevillon et al., 2005).

(iii) Métodos de análisis de microarrays. Estos métodos incluyen técnicas de normalización de *microarrays* de ADN, selección de genes diferencialmente expresados entre grupos de muestras, y anotación funcional de resultados.

A.4 Resultados

Desarrollo de métodos y herramientas para el análisis de datos de *microarrays*

La primera sección de los resultados presenta diferentes herramientas de Internet y métodos que he desarrollado para analizar datos procedentes de experimentos de *microarrays*. Estos incluyen: (i) normalización de microarrays de ADNc; (ii) análisis de expresión diferencial de genes; (iii) predicción de clase para datos de *microarrays*; (iv) regulación mediada por factores de trascripción de grupos de genes co-expresados; (v) determinación de la especificidad y sensitividad para Affymetrix GeneChips; (vi) detección de expresión específica de tejido para datos de *microarrays*; y (vii) análisis de tiling *microarrays*.

DNMAD: Normalización de microarrays de ADNc

En esta sección describo DNMAD, una herramienta web diseñada para normalizar datos de expresión procedentes de experimentos de *microarrays* de ADNc. La normalización es un proceso por el cual se elimina variación sistemática en la señal procedente de los *microarrays*. DNMAD usa R y Bioconductor para realizar los cálculos correspondientes y una interfaz programada en Perl CGI desde la que el usuario puede seleccionar los datos a normalizar así como diferentes opciones. DNMAD implementa por defecto el método de normalización denominado *print-tip loess* (*locally weighted scatter-plot smoothing*), mediante el cual se ajustan las intensidades de los dos fluoróforos teniendo en cuenta la posición de las sondas en el *microarray*. DNMAD también implementa otros tipos de normalización tales como loess global y normalización *inter-arrays*. DNMAD proporciona como resultado los valores de expresión normalizados. Además se incluyen diferentes gráficos que proporcionan al usuario una guía para valorar los resultados del proceso

de normalización así como la calidad de los *microarrays* analizados.

Pomelo: expresión diferencial de genes en experimentos de microarrays

Pomelo es una herramienta web para la detección de genes diferencialmente expresados en experimentos de *microarrays*. Le herramienta implementa diversos métodos de análisis que incluyen el test de la t, análisis de la varianza y el test exacto de Fisher. El procedimiento consiste en testar estadísticamente los valores de expresión para cada gen incluido en el microarray entre dos o más grupos de muestras. La herramienta proporciona los p-valores asociados a cada gen ajustándolos para corregir el fenómeno de test múltiple.

TNASAS: Predicción de clases en experimentos de microarrays

TNASAS es una herramienta web que permite obtener predictores de clase basados en datos de expresión génica. La herramienta es un programa en Perl que se comunica con R para realizar las tareas computacionales. TNASAS implementa diferentes métodos para seleccionar genes basándose en su expresión diferencial entre clases, que posteriormente son usados por un algoritmo de predicción de clase. TNASAS devuelve como resultado la lista de genes que forma el mejor predictor probado así como tasas de error de predicción insesgadas basadas en un sistema de validación cruzada.

TransFAT: regulación mediada por factores de trascripción de grupos de gene co-expresados

TransFAT es una herramienta web que permite identificar co-regulación por factores de trascripción en grupos de genes co-expresados. La herramienta es un interfaz web programado en Perl que se comunica con R para realizar tareas de calculo. El programa se basa en predicciones de sitios de unión para 270 factores de trascripción presentes en la base de datos TRANSFAC, que son obtenidas para 10kb en la región 5′ anterior al inicio de trascripción de

todos los genes de humano y ratón mediante el programa Match. TransFAT utiliza el test de Fisher con ajuste para test múltiple para encontrar factores de trascripción con sobre-representaciones de sitios de unión en grupos de genes co-expresados.

Determinación de la especificidad y sensitividad para Affymetrix GeneChips

Los *microarrays* de Affymetrix, al hibridar una sola muestra por *microarray* permiten evaluar los niveles de expresión de manera global. Sin embargo, debido a fenómenos de hibridación no competitiva así como por efecto de las técnicas de normalización, determinar cuando un gen se expresa o no es complicado. En esta sección describo un método que permite detectar la presencia o no de genes en estos experimentos. El método se basa en el uso combinado del algoritmo PANP (§3.2.3) junto a librerías de secuencias expresadas (ESTs) que permite valorar la especificidad y sensitividad del experimento. En §4.1.5 describo el uso de este método con el conjunto de datos SymAtlas de la fundación Genome Novartis. Para ello seleccionamos de la base de datos Unigene 31 librerías de secuencias que son expresadas en diferentes tejidos humanos sanos, y posteriormente evaluamos la diferencia de expresión reportada por los experimentos de *microarrays* para genes detectados como EST frente a genes no detectados. Los resultados de esta comparación muestran que pese a que hay un enriquecimiento de genes expresados con altos niveles de expresión, una parte significativa de los genes se expresan a un nivel que no permite diferenciarlos de genes no expresados.

Posteriormente, utilizando los ESTs como grupo de verdaderos positivos y el grupo de sondas diseñadas contra la cadena negativa que emplea PANP como verdaderos negativos, pudimos determinar mediante curvas ROC los niveles de especificidad y sensitividad para los diferentes *microarrays*, situándose estos entre el 45% y el 65% dependiendo de la muestra.

Expresión especifica de tejido para datos de microarrays

En §4.1.6 describo el uso de un estadístico simple, la propensión, para determinar expresión específica de tejido. La propensión mide la relación entre el nivel de expresión de un gen respecto al nivel de expresión de ese gen en todas las muestras y el nivel de expresión global para el tejido examinado. Altos valores del estadístico indican que el gen esta expresado en el tejido examinado por encima del resto, lo cual indica expresión específica de tejido.

Evaluamos la validez del método analizando los niveles de propensión para el conjunto de datos SymAtlas. Para ello calculamos a partir de los niveles de expresión los valores de propensión para cada gen, y seleccionamos aquellos genes con valores de propensión entre los 5% mas altos como específicos de tejido. Posteriormente validamos esta aproximación mediante el análisis funcional de los genes específicos, que presentaban términos GO relacionados con la función de cada tejido significativamente sobrerrepresentados.

Métodos de análisis para tiling microarrays

Los *tiling microarrays* nos permiten evaluar de manera sistemática la expresión génica o la unión de factores de trascripción al ADN a escala genómica ya que estos *microarrays* contienen sondas para interrogar regiones continuas del genoma. En §4.1.7 presentamos un método para el análisis de *tiling arrays* basado en una selección insesgada del ranking de las sondas mas significativas en diferentes muestras. El método fue testado realizando un análisis del complejo de compensación de la dosis génica en *D. melanogaster* en colaboración con el laboratorio de la Dr. Asifa Akhtar en EMBL. El análisis incluye la determinación de los sitios de unión para cinco proteínas que forman parte o están relacionadas con este complejo: MSL1, MSL3, MOF, Mtor y Nup153. Estas proteínas tienen diferentes afinidades de unión al ADN lo que dificulta el análisis combinado de los sitios de unión de los diferentes miembros del complejo. Este efecto se ve acentuado por la diferente eficiencia de los anticuerpos utilizados para realizar la inmunoprecipitación de la

cromatina para cada una de las proteínas.

Utilizando nuestro método basado en el ranking de las sondas fuimos capaces de determinar la existencia de una preferencia en cuanto a unión al cromosoma X por parte de las proteínas evaluadas, así como un posible mecanismo de acción del complejo basado en una diferente localización en el cuerpo del gen de la proteína MOF.

Identificación y caracterización funcional del repertorio de factores de trascripción específicos de secuencia en humano

En la segunda parte de los resultados se muestra como se identificaron y caracterizaron funcionalmente el grupo de genes que constituye el repertorio de factores de transcripción en humano.

La identificación de los factores de trascripción se llevo a cabo utilizando las secuencias que definen los dominios proteicos de unión a ADN disponible en la base de datos InterPro. Debido a la promiscuidad de alguno de estos dominios se llevo a cabo una comprobación manual de cada uno de los factores para eliminar falsos positivos. Esto resulto en la identificación de 1,369 genes como el repertorio de factores de trascripción de humano.

Posteriormente realizamos la caracterización funcional del mismo. Para ello primero evaluamos el nivel de conocimiento sobre estos genes analizando su anotación en Gene Ontology y el numero de citas por factor de trascripción en PubMed. Los resultados mostraron una muy pobre caracterización para la mayor parte de los factores.

Seguidamente, y utilizando datos genómicos, procedimos a caracterizar diversos aspectos del conjunto de factores de trascripción específicos de secuencia de humano tales como su: (i) clasificación estructural; (ii) expresión y uso en tejidos sanos; (iii) conservación evolutiva; y (iv) localización cromosómica.

Los principales resultados que se pueden extraer del análisis son que: (i) tres familias de factores de trascripción (los dedos de zinc, los homeodominios

y los hélice-bucle-hélice) dominan sobre el resto de familias constituyendo un 80% del total de factores de transcripción en humano; (ii) diferentes tejidos del cuerpo humano utilizan un numero muy diferente de factores de trascripción, existiendo al menos en parte una correlación entre el numero de tipos celulares y actividad metabólica y secretora, con el numero de factores de trascripción activos por tejido. Además, estos factores de trascripción se organizan en dos capas de funcionamiento. Una capa global, con factores de trascripción expresados ubicuamente en todos los tejidos, y una local con factores de trascripción específicos de tejido; (iii) diferentes familias de factores de trascripción han aparecido en diferentes momentos durante la evolución, como por ejemplo la aparición de organismos multicelulares, vertebrados, mamíferos o primates; y (iv) una gran parte de factores de trascripción están localizados en grupos en el genoma, lo que podría concederles algún tipo especial de regulación.

Identificación y caracterización de SNPs funcionales

La tercera parte del trabajo presentado consiste en la detección de aquellos SNPs localizados en regiones del genoma en las que la existencia de un cambio de secuencia pueda tener un efecto en cuanto a la regulación transcripcional o el preprocesado del ARN mensajero.

Para ello se determinaron regiones del genoma con función reguladora, como por ejemplo: (i) sitios de unión de factores de trascripción; (ii) sitios de formación de ADN-triplex; (iii) potenciadores exónicos del proceso de splicing; y (iv) sitios aceptores y donantes de splicing.

Posteriormente se detectaron todos aquellos SNPs localizados en cualquiera de estas zonas. Los resultados de este análisis se hicieron públicos a través de la herramienta PupaSNP (Conde et al., 2004), cuyo contenido es accesible a través de Internet.

A.5 Discusión

El desarrollo de técnicas de alto rendimiento a partir de la secuenciación de múltiples genomas ha producido, y continua produciendo, un flujo de información que requiere el desarrollo de métodos y herramientas para analizarlos. Sin los métodos estadísticos y el tratamiento matemático de los datos es muy difícil obtener resultados fiables y reproducibles que puedan ser posteriormente validados en subsiguientes experimentos. En este disertación he presentado métodos y herramientas para analizar resultados de mediciones de expresión génica o unión de factores de trascripción a escala genómica.

Uno de los mayores retos de la era post-genómica es entender las funciones y el uso de los genes en diferentes condiciones celulares. La regulación transcripcional juega un papel fundamental ya que determina que genes son expresados y en que medida bajo diferentes circunstancias. A pesar del gran interés que suscita esta línea de investigación, la mayor parte de los esfuerzos se han centrado en detectar sitios de unión de factores de trascripción en promotores de genes. Por el contrario, la caracterización del conjunto de factores de trascripción en humano no es suficientemente completa.

En §4.2.1 presentamos un conjunto de factores de trascripción humanos específicos de secuencia de alta calidad que posteriormente caracterizamos funcionalmente. Esta lista ha sido manualmente comprobada para asegurar un bajo numero de falsos positivos y un alto nivel de cobertura. Es importante resaltar la contribución de este conjunto de factores de trascripción al campo, ya que proporciona un conjunto de factores de trascripción que servirán como referencia para cualquier estudio sobre regulación transcripcional en mamíferos. Su importancia se destaca por el hecho de que varios laboratorios experimentales ya están usando los datos en este momento.

Por ultimo, cabe resaltar que otro de los mayores retos después de la secuenciación del genoma humano es determinar que hace a cada individuo único. Aunque la mayor parte del genoma es idéntico entre distintos

individuos de la misma especie, siempre hay variaciones en la secuencia nucleotídica causadas por translocaciones, transposiciones o mutaciones entre otros mecanismos. La mayor fuente de variación entre individuos de una misma especie la constituyen los SNPs.

En §4.3 detectamos aquellos SNPs que son susceptibles de producir un cambio en la correcta regulación de la trascripción o preprocesado de un gen. Algunos de estos SNPs han sido encontrados en regiones con desequilibrio de ligamiento que han asociados posteriormente con diversas enfermedades complejas, como por ejemplo Alzheimer o depresión.

A.6 Conclusiones

En esta tesis he presentado trabajo computacional dedicado a investigar la naturaleza de la regulación transcripcional en el genoma humano. Aunque el estudio no constituye una completa caracterización del sistema, si que forma unos sólidos cimientos que sirvan de base para trabajos posteriores.

Específicamente, del trabajo presentado en esta tesis doctoral titulada "*Computational Approaches to Study Transcripcional Regulation in the Human Genome*" podemos concluir que:

1. He identificado y analizado el repertorio de factores de trascripción específicos de secuencia de humano. Este es el primer estudio de este tipo, y el conjunto de datos, así como los análisis desarrollados son resultados importantes para próximas investigaciones genómicas.

Este estudio ha revelado que:

(i) tres familias de proteínas — dedos de Zn C_2H_2, homeodominos y hélice-bucle-hélice — dominan el repertorio de factores de trascripción humano.

(ii) los factores de trascripción se expresan bien de manera ubicua o bien específicamente en un tejido, dando lugar a un sistema regulación formado por reguladores globales y locales.

(iii) familias específicas de factores de trascripción se han expandido a lo largo del linaje humano coincidiendo con puntos claves en la evolución.

(iv) alrededor de un 40% de los factores de trascripción se localizan en clusters en el genoma, pudiendo esto ser indicativo de su actividad coordinada.

2. He predicho SNPs que modifican el funcionamiento normal de los procesos de regulación transcripcional y post-transcripcional. De esta manera he demostrado que existen miles de variaciones en el ADN que potencialmente pueden tener un impacto en estos sistemas regulatorios. Estas predicciones contribuirán a mejorar nuestro entendimiento de enfermedades complejas, dirigiendo estudios experimentales hacia los SNPs mas prometedores.

3. He desarrollado métodos estadísticos y herramientas para el análisis de datos genómicos. Además he demostrado como estas son esenciales para la correcta interpretación de los conjuntos de datos genómicos.

Appendix A

Appendix B

Transcription factor repertoire
(starts in new page)

Ensembl ID	DB-domain	DB-family	Class	HGNC Symbol	Tissue specific
ENSG00000001167	-	IPR001289	Other	NFYA	
ENSG00000004848	IPR000047 IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	ARX	
ENSG00000005073	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXA11	uterus
ENSG00000005102	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	MEOX1	
ENSG00000005513	IPR000910	-	HMG_1/2_box	SOX8	
ENSG00000005801	IPR007086 IPR007087	-	ZNF_C2H2	ZNF195	
ENSG00000005889	IPR007087	IPR006794	ZNF_C2H2	ZFX	
ENSG00000006194	IPR007086 IPR007087	-	ZNF_C2H2	ZNF263	
ENSG00000006377	IPR000047 IPR001356 IPR009057	-	Homeodomain_like	DLX6	
ENSG00000006468	IPR000418 IPR002341 IPR011991	-	Ets	ETV1	whole.brain, adrenal.cortex
ENSG00000006704	IPR011991	-	Other	GTF2IRD1	
ENSG00000007237	IPR000637	-	AT_hook_DNA_bd	-	whole.brain
ENSG00000007372	IPR001356 IPR007104 IPR009057 IPR011991 IPR012287	-	Homeodomain_like	PAX6	
ENSG00000007866	IPR009057	IPR000818	Homeodomain_like	TEAD3	placenta
ENSG00000007968	IPR011991	IPR003316	Other	E2F2	
ENSG00000008196	-	IPR004979 IPR008122	Other	TFAP2B	
ENSG00000008197	-	IPR004979	Other	TFAP2D	
ENSG00000008441	IPR003619	IPR000647	MAD_MH1	NFIX	
ENSG00000009709	IPR001356 IPR007104 IPR009057 IPR011991 IPR012287	-	Homeodomain_like	PAX7	
ENSG00000009950	IPR011598	-	HLH_DNA_bd	MLXIPL	adrenal.cortex
ENSG00000010030	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	ETV7	
ENSG00000010244	IPR007087	-	ZNF_C2H2	ZNF207	
ENSG00000010539	IPR007086 IPR007087	-	ZNF_C2H2	ZNF200	whole.blood.JJV
ENSG00000010818	IPR007087	-	ZNF_C2H2	HIVEP2	whole.brain, fetal.brain
ENSG00000011332	IPR007087	-	ZNF_C2H2	DPF1	
ENSG00000011451	IPR007087	-	ZNF_C2H2	-	
ENSG00000011590	IPR007087	-	ZNF_C2H2	ZBTB32	testis
ENSG00000012504	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR1H4	fetal.liver, adrenal.cortex

ENSG00000016082	IPR001356 IPR007107 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000018607	IPR007087	-	ZNF_C2H2	ZNF221	
ENSG00000018869	IPR007086 IPR007087	-	ZNF_C2H2	ZNF582	
ENSG00000019549	IPR007086 IPR007087	-	ZNF_C2H2	SNAI2	placenta, human.cultured.adipocyte, smooth.muscle, prostate, uterus, fetal.lung
ENSG00000020256	IPR007086 IPR007087	-	ZNF_C2H2	ZFP64	
ENSG00000020633	IPR008967 IPR012346 IPR013524	IPR000040	P53_like_DNA_bd	RUNX3	lymph.node, bone.marrow, tonsil, salivary.gland, whole.blood.JJV
ENSG00000025156	IPR000232 IPR002341 IPR011991	-	Heat shock factor (HSF)-type	HSF2	
ENSG00000025293	IPR000637 IPR007087	-	AT_hook_DNA_bd ZNF_C2H2	PHF20	
ENSG00000025434	IPR001628	IPR000324 IPR001723 IPR001728	Hrmn_rcpt_DNA_bd	NR1H3	testis, lymph.node, liver, human.cultured.adipocyte
ENSG00000028277	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	IPR000972	Homeodomain_like POU	POU2F2	
ENSG00000029153	IPR001766 IPR011598 IPR011991	IPR001067	HLH_DNA_bd TF_Fork_head	ARNTL2	
ENSG00000029363	-	-	TF_bZIP	BCLAF1	
ENSG00000030419	IPR007087	-	ZNF_C2H2	ZNFN1A2	
ENSG00000031544	IPR001628	IPR000003 IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR2E3	
ENSG00000036549	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	ZZZ3	lymph.node, thyroid
ENSG00000037965	IPR000047 IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXC8	skeletal.muscle.psoas
ENSG00000039600	IPR000910	-	HMG_1/2_box	SOX30	
ENSG00000043039	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	BARX2	
ENSG00000043355	IPR007087	-	ZNF_C2H2	ZIC2	
ENSG00000049768	IPR001766 IPR007087 IPR011991	-	TF_Fork_head ZNF_C2H2	FOXP3	

ENSG00000050344	IPR004827 IPR008917 IPR011616	-	TF_bZIP	NFE2L3	placenta
ENSG00000052835	IPR007087	-	ZNF_C2H2	-	
ENSG00000052850	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	ALX4	
ENSG00000053254	IPR001766 IPR011991	-	TF_Fork_head	CHES1	human.cultured.adipocyte
ENSG00000054598	IPR001766 IPR011991	-	TF_Fork_head	FOXC1	trachea, salivary.gland
ENSG00000056277	IPR007087	-	ZNF_C2H2	SUHW3	
ENSG00000057657	IPR007087	-	ZNF_C2H2	PRDM1	
ENSG00000059728	IPR011598	-	HLH_DNA_bd	MXD1	
ENSG00000060138	IPR002059	-	CSP_DNA_bd	CSDA	skeletal.muscle.psoas, testis,
ENSG00000060566	IPR004827 IPR008917 IPR011616	-	TF_bZIP	CREB3L3	
ENSG00000061455	IPR007087	-	ZNF_C2H2	-	
ENSG00000062370	IPR007086 IPR007087	-	ZNF_C2H2	ZNF228	
ENSG00000063438	IPR011598	-	HLH_DNA_bd	AHRR PDCD6	
ENSG00000063515	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	GSCL	
ENSG00000063587	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000064195	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	DLX3	
ENSG00000064218	IPR001275	-	DM_DNA_bd	DMRT3	
ENSG00000064489	IPR002100	-	TF_MADSbox	MEF2B	
ENSG00000064490	-	-	Other	RFXANK	heart, thymus, lung
ENSG00000064835	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU1F1	pituitary
ENSG00000065029	IPR007087	-	ZNF_C2H2	-	
ENSG00000065970	IPR001766 IPR011991	-	TF_Fork_head	FOXJ2	
ENSG00000065978	IPR002059	-	CSP_DNA_bd	YBX1	
ENSG00000066136	IPR003958	-	CBFA_NFYB_domain	NFYC	
ENSG00000066336	IPR000418 IPR002341 IPR011991	-	Ets	SPI1	lung, whole.blood.JJV
ENSG00000066422	IPR007087	-	ZNF_C2H2	ZBTB11	
ENSG00000066827	IPR007087	-	ZNF_C2H2	ZNF406	
ENSG00000067082	IPR007087	-	ZNF_C2H2	KLF6	lung
ENSG00000067646	IPR007087	IPR006794	ZNF_C2H2	ZFY	
ENSG00000067955	-	IPR003417	Other	CBFB	
ENSG00000068305	IPR002100	-	TF_MADSbox	MEF2A	
ENSG00000068323	IPR011598	-	HLH_DNA_bd	TFE3	

ENSG00000069011	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	PITX1	tongue, pituitary
ENSG00000069667	IPR001628	IPR001723 IPR003079	Hrmn_rcpt_DNA_bd	RORA	
ENSG00000069812	IPR011598	-	HLH_DNA_bd	HES2	
ENSG00000070444	IPR011598	-	HLH_DNA_bd	MNT	
ENSG00000070476	IPR007087	-	ZNF_C2H2	ZXDC	
ENSG00000071564	IPR009057 IPR011598	-	Homeodomain_like	TCF3	
ENSG00000072310	IPR011598	-	HLH_DNA_bd	SREBF1	human.cultured.adipocyte, adrenal.gland, adrenal.cortex
ENSG00000072736	IPR002909 IPR008967 IPR011539	IPR008366	IPT_TIG_rcpt P53_like_DNA_bd	NFATC3	
ENSG00000073282	IPR008967 IPR011615 IPR012346	IPR002117	P53_like_DNA_bd	TP73L	skin, trachea, thymus, tonsil, tongue, salivary.gland
ENSG00000073861	IPR008967	IPR001699	P53_like_DNA_bd	TBX21	whole.blood.JJV
ENSG00000074047	IPR007087	-	ZNF_C2H2	GLI2	
ENSG00000074219	IPR009057	IPR000818	Homeodomain_like	TEAD2	
ENSG00000074657	IPR007087	-	ZNF_C2H2	ZNF532	
ENSG00000075407	IPR007086 IPR007087	-	ZNF_C2H2	ZNF37A	
ENSG00000075426	IPR004827 IPR008917 IPR011616 IPR011700	-	TF_bZIP	FOSL2	ovary, adrenal.cortex
ENSG00000075891	IPR009057 IPR011991	-	Homeodomain_like	PAX2	
ENSG00000077092	IPR001628	IPR001723 IPR003078	Hrmn_rcpt_DNA_bd	-	
ENSG00000077150	IPR002909 IPR008967 IPR011539	IPR000451	IPT_TIG_rcpt P53_like_DNA_bd	NFKB2	
ENSG00000078043	-	-	Other	PIAS2	testis
ENSG00000078399	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXA9	prostate, kidney
ENSG00000078900	IPR008967 IPR011615 IPR012346	IPR002117	P53_like_DNA_bd	TP73	
ENSG00000079432	IPR000910	-	HMG_1/2_box	CIC	
ENSG00000080298	IPR003150 IPR011991	-	RFX_DNA_bd	RFX3	
ENSG00000081059	IPR000910	-	HMG_1/2_box	TCF7	
ENSG00000081189	IPR002100	-	TF_MADSbox	MEF2C	whole.brain, tonsil, fetal.brain
ENSG00000081386	IPR007086 IPR007087	-	ZNF_C2H2	ZNF510	
ENSG00000081665	IPR007086 IPR007087	-	ZNF_C2H2	ZNF93	
ENSG00000082175	IPR001628	IPR000128 IPR001723	Hrmn_rcpt_DNA_bd	PGR	

ENSG00000082641	IPR004827 IPR008917 IPR011616	-	TF_bZIP	NFE2L1	
ENSG00000083307	IPR007604	-	CP2	GRHL2	pancreas, placenta,
ENSG00000083812	IPR007086 IPR007087	-	ZNF_C2H2	ZNF324	
ENSG00000083814	IPR007086 IPR007087	-	ZNF_C2H2	ZNF671	
ENSG00000083817	IPR007086 IPR007087	-	ZNF_C2H2	ZNF416	
ENSG00000083828	IPR007086 IPR007087	-	ZNF_C2H2	ZNF586	
ENSG00000083838	IPR007087 IPR010982	-	ZNF_C2H2	ZNF446	
ENSG00000083842	IPR007086 IPR007087	-	ZNF_C2H2	ZNF8	
ENSG00000083844	IPR007086 IPR007087	-	ZNF_C2H2	ZNF264	
ENSG00000084093	IPR007087	-	ZNF_C2H2	REST	
ENSG00000085274	IPR007087	-	ZNF_C2H2	MYNN	skeletal.muscle.psoas
ENSG00000085276	IPR007087	-	ZNF_C2H2	EVI1	trachea,
ENSG00000085644	IPR007086 IPR007087	-	ZNF_C2H2	ZNF213	
ENSG00000086102	IPR000967	-	Znf_NFX1	NFX1	
ENSG00000087510	-	IPR004979 IPR008123	Other	TFAP2C	skin, placenta, tongue
ENSG00000087903	IPR003150 IPR011991	-	RFX_DNA_bd	RFX2	testis, bone.marrow
ENSG00000088876	IPR007087	-	ZNF_C2H2	ZNF343	
ENSG00000088881	IPR002909	IPR003523	IPT_TIG_rcpt	-	
ENSG00000089116	IPR000047 IPR001356 IPR007107 IPR012287	-	Homeodomain_like	LHX5	
ENSG00000089225	IPR008967	IPR001699 IPR002070	P53_like_DNA_bd	-	
ENSG00000089335	IPR007087	-	ZNF_C2H2	ZNF302	fetal.brain
ENSG00000089775	IPR007087	-	ZNF_C2H2	ZBTB25	
ENSG00000089902	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	RCOR1	
ENSG00000090447	IPR011598	-	HLH_DNA_bd	TFAP4	
ENSG00000090612	IPR007086 IPR007087	-	ZNF_C2H2	ZNF10	
ENSG00000091010	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU4F3	
ENSG00000091656	IPR001356 IPR007087 IPR012287	-	Homeodomain_like ZNF_C2H2	ZFHX4	smooth.muscle, pituitary
ENSG00000091831	IPR001628	IPR000324 IPR001723 IPR012239	Hrmn_rcpt_DNA_bd	ESR1	prostate, uterus
ENSG00000092067	IPR004827 IPR008917 IPR011700	-	TF_bZIP	CEBPE	bone.marrow

ENSG00000092098	IPR001346 IPR011991	-	IRF	ISGF3G RNF31	
ENSG00000092607	IPR008967	IPR001699	P53_like_DNA_bd	TBX15	
ENSG00000095574	IPR007087	-	ZNF_C2H2	ZNFN1A5	
ENSG00000095794	IPR004827 IPR008917 IPR011616 IPR011700	IPR001630	TF_bZIP	CREM	adrenal.gland, adrenal.cortex
ENSG00000095951	IPR007087	-	ZNF_C2H2	HIVEP1	
ENSG00000096401	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	CDC5L	
ENSG00000096654	IPR007086 IPR007087	-	ZNF_C2H2	ZNF184	
ENSG00000099326	IPR007086 IPR007087	-	ZNF_C2H2	ZNF42	thymus, prostate, fetal.thyroid
ENSG00000099949	-	-	Other	LZTR1	
ENSG00000100105	IPR000637 IPR007087	-	AT_hook_DNA_bd ZNF_C2H2	-	
ENSG00000100146	IPR000910	-	HMG_1/2_box	SOX10	spinal.cord, trachea
ENSG00000100207	-	-	Other	TCF20	
ENSG00000100219	IPR004827 IPR008917 IPR011616 IPR011700	-	TF_bZIP	XBP1	trachea, liver, prostate, salivary.gland
ENSG00000100426	IPR003656	-	Znf_BED_prd	ZBED4	
ENSG00000100625	IPR001356 IPR007106 IPR009057 IPR012287	-	Homeodomain_like	SIX4	
ENSG00000100644	IPR011598	IPR001321	HLH_DNA_bd	HIF1A	smooth.muscle
ENSG00000100811	IPR007087	-	ZNF_C2H2	YY1	
ENSG00000100829	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000100968	IPR002909 IPR008967 IPR011539	IPR008366	IPT_TIG_rcpt P53_like_DNA_bd	NFATC4	
ENSG00000100987	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	VSX1	
ENSG00000101057	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	MYBL2	
ENSG00000101076	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	HNF4A	liver
ENSG00000101080	IPR007087	-	ZNF_C2H2	-	
ENSG00000101096	IPR002909 IPR008967 IPR011539	IPR008366	IPT_TIG_rcpt P53_like_DNA_bd	NFATC2	
ENSG00000101115	IPR007087	-	ZNF_C2H2	SALL4	
ENSG00000101126	IPR001356 IPR007087 IPR009057 IPR012287	-	Homeodomain_like ZNF_C2H2	ADNP	fetal.brain
ENSG00000101190	IPR011598	-	HLH_DNA_bd	TCFL5	spinal.cord, testis, whole.brain,

ENSG00000101216	IPR000770 IPR010919	-	SAND_like	GMEB2	testis, thymus
ENSG00000101412	IPR011991	IPR003316	Other	E2F1	
ENSG00000101493	IPR007087	-	ZNF_C2H2	-	uterus
ENSG00000101544	IPR001356 IPR007087 IPR012287	-	Homeodomain_like ZNF_C2H2	-	
ENSG00000101883	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000102034	IPR000418 IPR002341 IPR011991	-	Ets	ELF4	thymus, placenta, bone.marrow,
ENSG00000102145	IPR000679	-	ZnF_GATA	GATA1	
ENSG00000102349	IPR007087	-	ZNF_C2H2	-	
ENSG00000102554	IPR007087	-	ZNF_C2H2	KLF5	skin, pancreas, trachea, placenta, tonsil,
ENSG00000102804	-	IPR000580	Other	TSC22D1	
ENSG00000102870	IPR007087	-	ZNF_C2H2	-	
ENSG00000102878	IPR000232 IPR002341 IPR011991	-	Heat shock factor (HSF)-type	-	
ENSG00000102908	IPR002909 IPR008967 IPR011539	IPR008366	IPT_TIG_rcpt P53_like_DNA_bd	NFAT5	
ENSG00000102935	IPR007087	-	ZNF_C2H2	ZNF423	spinal.cord, whole.brain, uterus
ENSG00000102974	IPR007087	-	ZNF_C2H2	CTCF	thymus
ENSG00000103199	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000103241	IPR001766	-	TF_Fork_head	FOXF1	lung, fetal.lung
ENSG00000103343	IPR007086 IPR007087	-	ZNF_C2H2	ZNF174	
ENSG00000103449	IPR007087	-	ZNF_C2H2	SALL1	spinal.cord,
ENSG00000103495	IPR007087	-	ZNF_C2H2	MAZ	
ENSG00000104447	IPR000679 IPR007087	-	ZNF_C2H2 ZnF_GATA	TRPS1	
ENSG00000104777	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000104856	IPR002909 IPR011539	IPR000451	IPT_TIG_rcpt P53_like_DNA_bd	RELB	
ENSG00000104903	IPR011598	-	HLH_DNA_bd	LYL1	bone.marrow, fetal.liver
ENSG00000105066	IPR007087	-	ZNF_C2H2	-	
ENSG00000105132	IPR007087	-	ZNF_C2H2	-	
ENSG00000105136	IPR007086 IPR007087	-	ZNF_C2H2	ZNF419	
ENSG00000105392	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000105419	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000105497	IPR007086 IPR007087	-	ZNF_C2H2	ZNF175	
ENSG00000105516	IPR004827 IPR011700	-	TF_bZIP	DBP	thyroid

ENSG00000105610	IPR007087	-	ZNF_C2H2	KLF1	bone.marrow, fetal.liver
ENSG00000105672	IPR000418 IPR002341 IPR011991	-	Ets	-	
ENSG00000105698	IPR011598	-	HLH_DNA_bd	USF2	
ENSG00000105708	IPR007086 IPR007087	-	ZNF_C2H2	ZNF14	
ENSG00000105717	IPR001356 IPR009057	-	Homeodomain_like	PBX4	
ENSG00000105722	IPR000418 IPR002341 IPR011991	-	Ets	ERF	
ENSG00000105732	IPR007087	-	ZNF_C2H2	ZNF574	
ENSG00000105750	IPR007086 IPR007087	-	ZNF_C2H2	ZNF85	
ENSG00000105856	IPR000910	-	HMG_1/2_box	HBP1	whole.blood.JJV
ENSG00000105866	IPR007087	-	ZNF_C2H2	SP4	
ENSG00000105880	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	DLX5	placenta
ENSG00000105967	IPR011598	-	HLH_DNA_bd	TFEC	
ENSG00000105991	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXA1	
ENSG00000105996	IPR000047 IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXA2	
ENSG00000105997	IPR000047 IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXA3	
ENSG00000106004	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXA5	spinal.cord, adrenal.gland, uterus, lung, fetal.lung, adrenal.cortex
ENSG00000106006	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXA6	
ENSG00000106031	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXA13	
ENSG00000106038	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	EVX1	
ENSG00000106261	IPR007086 IPR007087	-	ZNF_C2H2	ZKSCAN1	pancreas
ENSG00000106331	IPR001356 IPR007104 IPR009057 IPR011991 IPR012287	-	Homeodomain_like	-	

ENSG00000106410	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000106459	IPR002100	-	TF_MADSbox	NRF1	
ENSG00000106462	IPR001005 IPR009057	-	Homeodomain_like	EZH2	testis, thymus
ENSG00000106511	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	MEOX2	placenta, fetal.lung
ENSG00000106536	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU6F2	
ENSG00000106546	IPR011598	-	HLH_DNA_bd	AHR	
ENSG00000106571	IPR007087	-	ZNF_C2H2	GLI3	
ENSG00000106689	IPR001356 IPR007107 IPR009057 IPR012287	-	Homeodomain_like	LHX2	whole.brain, fetal.brain
ENSG00000106852	IPR001356 IPR007107 IPR009057 IPR012287	-	Homeodomain_like	LHX6	whole.brain, fetal.brain
ENSG00000107175	IPR004827 IPR008917 IPR011616	-	TF_bZIP	CREB3	whole.brain, smooth.muscle, prostate
ENSG00000107187	IPR001356 IPR007107 IPR009057 IPR012287	-	Homeodomain_like	LHX3	pituitary
ENSG00000107249	IPR007087	-	ZNF_C2H2	GLIS3	
ENSG00000107485	IPR000679	-	ZnF_GATA	GATA3	skin, thymus, placenta
ENSG00000107807	IPR000747 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000107859	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	PITX3	
ENSG00000108001	IPR002909 IPR011598	IPR003523	HLH_DNA_bd IPT_TIG_rcpt	EBF3	
ENSG00000108064	IPR000910	-	HMG_1/2_box	TFAM	
ENSG00000108312	IPR000910 IPR009057	-	HMG_1/2_box Homeodomain_like	UBTF	
ENSG00000108452	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000108509	IPR002909 IPR005559	-	CG-1 IPT_TIG_rcpt	CAMTA2	
ENSG00000108511	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXB6	pancreas, human.cultured.a dipocyte, kidney
ENSG00000108753	IPR001356 IPR009057 IPR010982 IPR012287	-	Homeodomain_like	TCF2	kidney
ENSG00000108788	IPR011598	-	HLH_DNA_bd	MLX	

ENSG00000108799	IPR001005 IPR009057	-	Homeodomain_like	EZH1	
ENSG00000108813	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	DLX4	placenta
ENSG00000108924	IPR004827 IPR011700	-	TF_bZIP	HLF	placenta, whole.brain, pituitary
ENSG00000109101	IPR001766 IPR011991	-	TF_Fork_head	FOXN1	
ENSG00000109132	IPR000047 IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	PHOX2B	
ENSG00000109320	IPR002909 IPR008967 IPR011539	IPR000451	IPT_TIG_rcpt P53_like_DNA_bd	NFKB1	thymus, smooth.muscle, tonsil, whole.blood.JJV
ENSG00000109381	IPR000418 IPR002341 IPR011991	-	Ets	ELF2	
ENSG00000109685	IPR000910	-	HMG_1/2_box	WHSC1	testis, thymus, fetal.liver
ENSG00000109705	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	BAPX1	
ENSG00000109787	IPR007087	-	ZNF_C2H2	KLF3	fetal.liver
ENSG00000109851	IPR000047 IPR001356 IPR009057	-	Homeodomain_like	-	
ENSG00000109906	IPR007087	-	ZNF_C2H2	ZBTB16	skeletal.muscle.p soas, ovary, human.cultured.a dipocyte, whole.brain, prostate, lung
ENSG00000110693	IPR000910	-	HMG_1/2_box	SOX6	
ENSG00000110851	IPR007087	-	ZNF_C2H2	PRDM4	
ENSG00000111046	IPR011598	-	HLH_DNA_bd	MYF6	skeletal.muscle.p soas
ENSG00000111049	IPR011598	-	HLH_DNA_bd	MYF5	
ENSG00000111087	IPR007087	-	ZNF_C2H2	GLI1	
ENSG00000111145	IPR000418 IPR002341 IPR011991	-	Ets	ELK3	lymph.node, human.cultured.a dipocyte, smooth.muscle, uterus, whole.blood.JJV, thyroid, fetal.thyroid, tonsil, fetal.lung
ENSG00000111206	IPR001766 IPR011991	-	TF_Fork_head	FOXM1	
ENSG00000111249	IPR001356 IPR003350 IPR007108 IPR009057 IPR012287	-	Hmoeo_CUT Homeodomain_like	CUTL2	liver, whole.brain, fetal.brain

ENSG00000111269	IPR008917 IPR011700	-	TF_bZIP	CREBL2	adrenal.gland, pituitary
ENSG00000111424	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	VDR	
ENSG00000111704	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	NANOG	
ENSG00000111783	IPR003150 IPR011991	-	RFX_DNA_bd	RFX4	
ENSG00000112033	IPR001628	IPR001723 IPR003074 IPR003075	Hrmn_rcpt_DNA_bd	PPARD	
ENSG00000112182	IPR004827 IPR008917 IPR011616	-	TF_bZIP	BACH2	tonsil, fetal.brain
ENSG00000112200	IPR007087	-	ZNF_C2H2	ZNF451	
ENSG00000112238	IPR007087	-	ZNF_C2H2	PRDM13	
ENSG00000112242	IPR011991	IPR003316	Other	E2F3	whole.blood.JJV
ENSG00000112246	IPR011598	IPR001067	HLH_DNA_bd	SIM1	
ENSG00000112333	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	NR2E1	
ENSG00000112365	IPR000637 IPR007086 IPR007087	-	AT_hook_DNA_bd ZNF_C2H2	ZBTB24	
ENSG00000112561	IPR011598	-	HLH_DNA_bd	TFEB	whole.blood.JJV
ENSG00000112658	IPR002100	-	TF_MADSbox	SRF	
ENSG00000112837	IPR008967	IPR001699 IPR002070	P53_like_DNA_bd	TBX18	
ENSG00000113196	IPR011598	-	HLH_DNA_bd	HAND1	
ENSG00000113430	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	IRX4	
ENSG00000113580	IPR001628	IPR001409 IPR001723	Hrmn_rcpt_DNA_bd	NR3C1	smooth.muscle, whole.blood.JJV
ENSG00000113658	IPR003619 IPR013019	IPR001132 IPR013790	MAD_MH1	SMAD5	
ENSG00000113722	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	CDX1	
ENSG00000113916	IPR007086 IPR007087	-	ZNF_C2H2	BCL6	skeletal.muscle.psoas, whole.blood.JJV
ENSG00000114126	IPR011991	IPR003316	Other	-	testis, thymus
ENSG00000114315	IPR011598	-	HLH_DNA_bd	HES1	skin, placenta, prostate, fetal.thyroid, tongue, lung, thyroid
ENSG00000114439	IPR000910	-	HMG_1/2_box	BBX	adrenal.gland, prostate, fetal.thyroid, uterus, pituitary
ENSG00000114853	IPR007086 IPR007087	-	ZNF_C2H2	ZNF651	
ENSG00000114861	IPR001766 IPR007087 IPR011991	-	TF_Fork_head ZNF_C2H2	FOXP1	
ENSG00000115112	IPR007604	-	CP2	TFCP2L1	salivary.gland, kidney

ENSG00000115297	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	TLX2	
ENSG00000115415	IPR008967 IPR012345	IPR001217 IPR013799 IPR013801	P53_like_DNA_bd STAT_DNA_bd	STAT1	lymph.node, thymus, smooth.muscle, tonsil, salivary.gland, whole.blood.JJV
ENSG00000115507	IPR001356 IPR007104 IPR009057 IPR012287	IPR003025 IPR003026	Homeodomain_like	OTX1	
ENSG00000115568	IPR007087	-	ZNF_C2H2	ZNF142	
ENSG00000115738	IPR011598	-	HLH_DNA_bd	ID2	liver
ENSG00000115816	IPR005612	-	CBF	CEBPZ	
ENSG00000115844	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	DLX2	fetal.brain
ENSG00000115966	IPR004827 IPR007087 IPR008917 IPR011616 IPR011700	-	TF_bZIP ZNF_C2H2	ATF2	fetal.brain, thyroid
ENSG00000116016	IPR011598	IPR001067	HLH_DNA_bd	EPAS1	placenta, lung, fetal.lung
ENSG00000116017	-	-	Other	ARID3A	placenta
ENSG00000116035	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	VAX2	
ENSG00000116044	IPR004827 IPR011616	-	TF_bZIP	NFE2L2	
ENSG00000116132	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	PRRX1	trachea, human.cultured.a dipocyte, smooth.muscle
ENSG00000116539	IPR000637	-	AT_hook_DNA_bd	ASH1L	
ENSG00000116580	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	GON4L	
ENSG00000116604	IPR002100	-	TF_MADSbox	MEF2D	
ENSG00000116731	IPR007087	IPR009170	ZNF_C2H2	PRDM2	whole.brain, fetal.brain
ENSG00000116793	-	-	Other	PHTF1	
ENSG00000116809	IPR007087	-	ZNF_C2H2	ZBTB17	
ENSG00000116819	-	IPR004979	Other	TFAP2E	
ENSG00000116833	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR5A2	pancreas
ENSG00000116990	IPR011598	IPR002418	HLH_DNA_bd	MYCL1	
ENSG00000117000	IPR007087	-	ZNF_C2H2	RLF	
ENSG00000117010	IPR007086 IPR007087	-	ZNF_C2H2	ZNF684	
ENSG00000117036	IPR000418 IPR002341 IPR011991	-	Ets	ETV3	

ENSG00000117318	IPR011598	-	HLH_DNA_bd	ID3	heart, smooth.muscle, prostate, fetal.thyroid, lung, thyroid
ENSG00000117505	IPR003958	IPR003957	CBFA_NFYB_domain	DR1	testis, whole.blood.JJV
ENSG00000117595	IPR001346 IPR011991	-	IRF	IRF6	
ENSG00000117625	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	RCOR3	
ENSG00000117707	IPR009057	IPR007738	Homeodomain_like	PROX1	
ENSG00000118156	IPR001005 IPR007087	-	Homeodomain_like ZNF_C2H2	ZNF541	
ENSG00000118217	IPR004827 IPR011616	-	TF_bZIP	ATF6	
ENSG00000118260	IPR004827 IPR008917 IPR011616	IPR001630	TF_bZIP	CREB1	
ENSG00000118263	IPR007087	-	ZNF_C2H2	KLF7	
ENSG00000118267	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000118495	IPR007087	-	ZNF_C2H2	PLAGL1	placenta, adrenal.gland, uterus, fetal.lung, adrenal.cortex
ENSG00000118513	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	MYB	thymus, bone.marrow
ENSG00000118526	IPR011598	-	HLH_DNA_bd	TCF21	heart, placenta, smooth.muscle, lung, fetal.lung
ENSG00000118620	IPR007086 IPR007087	-	ZNF_C2H2	ZNF430	
ENSG00000118689	IPR001766 IPR011991	-	TF_Fork_head	FOXO3A	bone.marrow, lung
ENSG00000118922	IPR007087	-	ZNF_C2H2	KLF12	
ENSG00000119042	IPR001356 IPR003350 IPR007108 IPR009057 IPR012287	-	Hmoeo_CUT Homeodomain_like	SATB2	whole.brain, fetal.brain
ENSG00000119138	IPR007087	-	ZNF_C2H2	KLF9	uterus
ENSG00000119508	IPR001628	IPR001723 IPR003070 IPR003072	Hrmn_rcpt_DNA_bd	NR4A3	skeletal.muscle.psoas, adrenal.cortex
ENSG00000119547	IPR001356 IPR003350 IPR007108 IPR009057	-	Hmoeo_CUT Homeodomain_like	ONECUT2	
ENSG00000119574	IPR007087	-	ZNF_C2H2	ZNF499	
ENSG00000119614	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	CHX10	
ENSG00000119715	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	ESRRB	
ENSG00000119725	IPR007087	-	ZNF_C2H2	ZNF410	testis
ENSG00000119866	IPR007087	-	ZNF_C2H2	BCL11A	tonsil, fetal.brain

ENSG00000119919	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	NKX2-3	
ENSG00000119950	IPR011598	-	HLH_DNA_bd	MXI1	whole.brain, bone.marrow
ENSG00000120068	IPR000047 IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXB8	
ENSG00000120075	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXB5	
ENSG00000120087	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXB7	pancreas
ENSG00000120093	IPR000047 IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXB3	
ENSG00000120094	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXB1	
ENSG00000120149	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	MSX2	placenta
ENSG00000120690	IPR000418 IPR002341 IPR011991	-	Ets	ELF1	pancreas, thymus, whole.blood.JJV
ENSG00000120693	IPR003619 IPR013019	IPR001132 IPR013790	MAD_MH1	SMAD9	
ENSG00000120738	IPR007087	-	ZNF_C2H2	EGR1	pancreas, trachea, thymus, adrenal.gland, pituitary, kidney, adrenal.cortex
ENSG00000120784	IPR007086 IPR007087	-	ZNF_C2H2	ZFP30	
ENSG00000120798	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR2C1	
ENSG00000120837	IPR003958	IPR003956 IPR003957	CBFA_NFYB_domain	NFYB	
ENSG00000120963	IPR007087	-	ZNF_C2H2	ZNF706	
ENSG00000121068	IPR008967	IPR001699 IPR002070	P53_like_DNA_bd	TBX2	placenta, lung, fetal.lung
ENSG00000121075	IPR008967	IPR001699 IPR002070	P53_like_DNA_bd	TBX4	
ENSG00000121297	IPR001356 IPR007087 IPR009057 IPR012287	-	Homeodomain_like ZNF_C2H2	-	
ENSG00000121406	IPR007087	-	ZNF_C2H2	ZNF549	
ENSG00000121413	IPR007087	-	ZNF_C2H2	ZNF447	whole.brain, pituitary, thyroid
ENSG00000121417	IPR007086 IPR007087	-	ZNF_C2H2	ZNF134 ZNF211	

ENSG00000121454	IPR001356 IPR007107 IPR009057 IPR012287	-	Homeodomain_like	LHX4	
ENSG00000121864	IPR007087	-	ZNF_C2H2	ZNF639	
ENSG00000121903	IPR001005 IPR007086 IPR007087 IPR009057	-	Homeodomain_like ZNF_C2H2	ZNF31	
ENSG00000122145	IPR008967	IPR001699	P53_like_DNA_bd	TBX22	
ENSG00000122180	IPR011598	-	HLH_DNA_bd	MYOG	
ENSG00000122386	IPR007086 IPR007087	-	ZNF_C2H2	ZNF205	
ENSG00000122482	IPR007087	-	ZNF_C2H2	ZNF644	
ENSG00000122592	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXA7	
ENSG00000122691	IPR011598	-	HLH_DNA_bd	TWIST1	skin, placenta, human.cultured.adipocyte, uterus
ENSG00000122859	IPR011598	-	HLH_DNA_bd	NEUROG3	
ENSG00000122877	IPR007087	-	ZNF_C2H2	EGR2	thyroid
ENSG00000123095	IPR011598	-	HLH_DNA_bd	BHLHB3	spinal.cord, whole.brain, pituitary, thyroid
ENSG00000123268	IPR004827 IPR008917 IPR011616	IPR001630	TF_bZIP	ATF1	whole.blood.JJV, thyroid
ENSG00000123307	IPR011598	-	HLH_DNA_bd	NEUROD4	
ENSG00000123358	IPR001628	IPR001723 IPR003070 IPR003071	Hrmn_rcpt_DNA_bd	NR4A1	
ENSG00000123364	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXC13	
ENSG00000123388	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXC11	
ENSG00000123405	IPR004827 IPR008917 IPR011616	-	TF_bZIP	-	bone.marrow, whole.blood.JJV, fetal.liver
ENSG00000123407	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXC12	
ENSG00000123411	IPR007087	-	ZNF_C2H2	ZNFN1A4	
ENSG00000123576	IPR000047 IPR000327 IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like POU	ESX1	
ENSG00000123685	IPR004827 IPR008917 IPR011616	IPR002112	TF_bZIP	-	
ENSG00000123870	IPR007087	-	ZNF_C2H2	-	
ENSG00000123933	IPR011598	-	HLH_DNA_bd	MXD4	thymus, lung
ENSG00000124092	IPR007087	-	ZNF_C2H2	CTCFL	
ENSG00000124201	IPR000967	-	Znf_NFX1	ZNFX1	

ENSG00000124216	IPR007086 IPR007087	-	ZNF_C2H2	SNAI1	
ENSG00000124232	IPR002909	-	IPT_TIG_rcpt	RBPSUHL	
ENSG00000124440	IPR011598	IPR001067	HLH_DNA_bd	HIF3A	fetal.lung
ENSG00000124444	IPR007087	-	ZNF_C2H2	-	
ENSG00000124459	IPR007086 IPR007087	-	ZNF_C2H2	ZNF45	
ENSG00000124496	IPR001005 IPR007087 IPR012287	-	Homeodomain_like ZNF_C2H2	TRERF1	
ENSG00000124601	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000124664	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	SPDEF	
ENSG00000124766	IPR000910	-	HMG_1/2_box	SOX4	thymus, smooth.muscle, fetal.brain
ENSG00000124782	IPR007087	-	ZNF_C2H2	RREB1	fetal.thyroid, fetal.liver
ENSG00000124813	IPR008967 IPR012346 IPR013524	IPR000040	P53_like_DNA_bd	RUNX2	
ENSG00000124827	-	IPR003902	Other	GCM2	
ENSG00000125285	IPR000910	-	HMG_1/2_box	SOX21	
ENSG00000125347	IPR001346 IPR011991	-	IRF	IRF1	lymph.node, thymus, placenta, lung, whole.blood.JJV, fetal.lung
ENSG00000125398	IPR000910	-	HMG_1/2_box	SOX9	spinal.cord, pancreas, trachea, testis, whole.brain,
ENSG00000125482	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	TTF1	fetal.thyroid
ENSG00000125492	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	BARHL1	
ENSG00000125533	IPR011598	-	HLH_DNA_bd	BHLHB4	
ENSG00000125618	IPR009057 IPR011991	-	Homeodomain_like	PAX8	fetal.thyroid, thyroid
ENSG00000125740	IPR004827 IPR008917 IPR011700	-	TF_bZIP	FOSB	trachea, thyroid, adrenal.cortex
ENSG00000125798	IPR001766 IPR011991	-	TF_Fork_head	FOXA2	
ENSG00000125812	IPR007087	-	ZNF_C2H2	ZNF336	
ENSG00000125813	IPR009057 IPR011991	-	Homeodomain_like	PAX1	thymus, tonsil
ENSG00000125816	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	NKX2-4	
ENSG00000125820	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	NKX2-2	spinal.cord, whole.brain
ENSG00000125846	IPR007087	-	ZNF_C2H2	ZNF133	
ENSG00000125850	IPR007087	-	ZNF_C2H2	OVOL2	

ENSG00000125878	IPR011598	-	HLH_DNA_bd	TCF15	
ENSG00000125945	IPR007086 IPR007087	-	ZNF_C2H2	ZNF436	
ENSG00000125952	IPR011598	IPR002418	HLH_DNA_bd	MAX	whole.blood.JJV
ENSG00000125968	IPR011598	-	HLH_DNA_bd	ID1	pancreas, heart, prostate, uterus, lung, thyroid, fetal.lung
ENSG00000126003	IPR007087	-	ZNF_C2H2	PLAGL2	testis
ENSG00000126351	IPR001628	IPR001723 IPR001728	Hrmn_rcpt_DNA_bd	THRA	spinal.cord, whole.brain, fetal.brain
ENSG00000126368	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR1D1	skin
ENSG00000126456	IPR001346 IPR011991	-	IRF	IRF3	lymph.node
ENSG00000126561	IPR008967 IPR012345	IPR001217 IPR013799 IPR013801	P53_like_DNA_bd STAT_DNA_bd	STAT5A	
ENSG00000126603	IPR007087	-	ZNF_C2H2	GLIS2	
ENSG00000126656	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000126733	IPR009061	-	Other	DACH2	
ENSG00000126746	IPR007087	-	ZNF_C2H2	ZNF384	
ENSG00000126767	IPR000418 IPR002341 IPR011991	-	Ets	ELK1	
ENSG00000126778	IPR000047 IPR001356 IPR007106 IPR009057 IPR012287	-	Homeodomain_like	SIX1	
ENSG00000126804	IPR007087	-	ZNF_C2H2	ZBTB1	
ENSG00000127081	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000127124	IPR007087	-	ZNF_C2H2	HIVEP3	
ENSG00000127152	IPR007087	-	ZNF_C2H2	BCL11B	thymus
ENSG00000127528	IPR007087	-	ZNF_C2H2	KLF2	whole.blood.JJV
ENSG00000128272	IPR004827 IPR011616	-	TF_bZIP	ATF4	
ENSG00000128573	IPR001766 IPR007087 IPR011991	-	TF_Fork_head ZNF_C2H2	FOXP2	
ENSG00000128604	IPR001346 IPR011991	-	IRF	IRF5	
ENSG00000128645	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXD1	
ENSG00000128652	IPR000047 IPR001356 IPR001827 IPR009057	-	Homeodomain_like	HOXD3	
ENSG00000128709	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXD9	uterus
ENSG00000128710	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXD10	

ENSG00000128713	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXD11	
ENSG00000128714	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXD13	
ENSG00000129152	IPR011598	-	HLH_DNA_bd	MYOD1	
ENSG00000129194	IPR000135 IPR000910	-	HMG_1/2_box	SOX15	
ENSG00000129514	IPR001766 IPR011991	-	TF_Fork_head	FOXA1	
ENSG00000129535	IPR004826 IPR004827 IPR008917	-	TF_bZIP	NRL	
ENSG00000129654	IPR001766 IPR011991	-	TF_Fork_head	FOXJ1	
ENSG00000129911	IPR007087	-	ZNF_C2H2	KLF16	
ENSG00000130182	IPR007087	-	ZNF_C2H2	ZNF206	
ENSG00000130382	-	IPR005033	Other	MLLT1	
ENSG00000130522	IPR004827 IPR008917 IPR011616	IPR002112	TF_bZIP	JUND	
ENSG00000130544	IPR007086 IPR007087	-	ZNF_C2H2	ZNF557	
ENSG00000130584	IPR007087	-	ZNF_C2H2	BTBD4	
ENSG00000130675	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HLXB9	
ENSG00000130684	IPR007087	-	ZNF_C2H2	ZNF337	fetal.brain
ENSG00000130700	IPR000679	IPR008013	ZnF_GATA	GATA5	
ENSG00000130711	IPR007087	-	ZNF_C2H2	PRDM12	
ENSG00000130751	IPR011598	-	HLH_DNA_bd	NPAS1	
ENSG00000130803	IPR007086 IPR007087	-	ZNF_C2H2	ZNF317	
ENSG00000130818	IPR007086 IPR007087	-	ZNF_C2H2	ZNF426	
ENSG00000130844	IPR007086 IPR007087	-	ZNF_C2H2	ZNF331	ovary, adrenal.gland, pituitary, adrenal.cortex
ENSG00000130856	IPR002197 IPR007087	-	ZNF_C2H2	ZNF236	
ENSG00000130940	IPR007087	-	ZNF_C2H2	CASZ1	
ENSG00000131061	IPR007087	-	ZNF_C2H2	ZNF341	
ENSG00000131115	IPR007086 IPR007087	-	ZNF_C2H2	ZNF227	
ENSG00000131127	IPR007086 IPR007087	-	ZNF_C2H2	ZNF141	
ENSG00000131196	IPR002909 IPR008967 IPR011539	IPR008366	IPT_TIG_rcpt P53_like_DNA_bd	-	
ENSG00000131264	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	CDX4	
ENSG00000131408	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR1H2	heart, lung, whole.blood.JJV
ENSG00000131668	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	BARX1	

ENSG00000131721	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000131759	IPR001628	IPR001723 IPR003078	Hrmn_rcpt_DNA_bd	RARA	whole.blood.JJV
ENSG00000131845	IPR007086 IPR007087	-	ZNF_C2H2	ZNF304	
ENSG00000131848	IPR007087	-	ZNF_C2H2	ZSCAN5	testis
ENSG00000131849	IPR007086 IPR007087	-	ZNF_C2H2	ZNF132	
ENSG00000132005	IPR003150 IPR011991	-	RFX_DNA_bd	RFX1	
ENSG00000132010	IPR007086 IPR007087	-	ZNF_C2H2	ZNF625	
ENSG00000132130	IPR000047 IPR001356 IPR007107 IPR009057 IPR012287	-	Homeodomain_like	LHX1	
ENSG00000132170	IPR001628	IPR001723 IPR003074 IPR003077	Hrmn_rcpt_DNA_bd	PPARG	human.cultured.adipocyte
ENSG00000132773	IPR000571 IPR008967	-	P53_like_DNA_bd ZnF_CCCH	TOE1	
ENSG00000133250	IPR007087	-	ZNF_C2H2	ZNF414	
ENSG00000133740	-	IPR003316	Other	E2F5	
ENSG00000133794	IPR011598	IPR001067	HLH_DNA_bd	ARNTL	
ENSG00000133884	IPR007087	-	ZNF_C2H2	DPF2	testis, whole.blood.JJV
ENSG00000133937	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	GSC	
ENSG00000134025	IPR002909	IPR003523	IPT_TIG_rcpt	EBF2	
ENSG00000134107	IPR011598	-	HLH_DNA_bd	BHLHB2	trachea, lung
ENSG00000134138	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	MEIS2	prostate, uterus, salivary.gland, pituitary,
ENSG00000134317	IPR007604	-	CP2	GRHL1	
ENSG00000134323	IPR011598	IPR002418	HLH_DNA_bd	MYCN	fetal.brain
ENSG00000134438	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	RAX	
ENSG00000134532	IPR000910	-	HMG_1/2_box	SOX5	testis
ENSG00000134595	IPR000910	-	HMG_1/2_box	SOX3	
ENSG00000134852	IPR011598	IPR001067	HLH_DNA_bd	CLOCK	
ENSG00000134954	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	ETS1	
ENSG00000135100	IPR001356 IPR009057 IPR010982 IPR012287	-	Homeodomain_like	TCF1	
ENSG00000135111	IPR008967	IPR001699 IPR002070	P53_like_DNA_bd	-	placenta, adrenal.gland, prostate, thyroid, adrenal.cortex

ENSG00000135164	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	DMTF1	
ENSG00000135373	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	EHF	trachea
ENSG00000135374	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	ELF5	trachea, salivary.gland
ENSG00000135457	IPR007604	-	CP2	TFCP2	
ENSG00000135547	IPR011598	-	HLH_DNA_bd	HEY2	
ENSG00000135625	IPR007087	-	ZNF_C2H2	EGR4	whole.brain
ENSG00000135638	IPR000047 IPR001356 IPR009057	-	Homeodomain_like	EMX1	
ENSG00000135747	IPR007086 IPR007087	-	ZNF_C2H2	ZNF670	
ENSG00000135899	IPR000770 IPR010919	-	SAND_like	SP110	thymus, tonsil, whole.blood.JJV
ENSG00000135903	IPR001356 IPR007104 IPR009057 IPR011991 IPR012287	-	Homeodomain_like	PAX3	
ENSG00000136327	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	NKX2-8	
ENSG00000136352	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	TITF1	fetal.thyroid, lung, thyroid, fetal.lung
ENSG00000136367	IPR001356 IPR007087 IPR009057	-	Homeodomain_like ZNF_C2H2	ZFHX2	
ENSG00000136451	IPR007087	-	ZNF_C2H2	ZNF161	thymus, uterus, whole.blood.JJV
ENSG00000136535	IPR008967	IPR001699	P53_like_DNA_bd	TBR1	
ENSG00000136574	IPR000679	IPR008013	ZnF_GATA	GATA4	pancreas, testis, heart
ENSG00000136603	IPR009061 IPR010919	-	SAND_like	SKIL	
ENSG00000136630	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HLX1	bone.marrow, lung, whole.blood.JJV
ENSG00000136826	IPR007087	-	ZNF_C2H2	KLF4	skin, trachea, placenta, lung
ENSG00000136866	IPR007086 IPR007087	-	ZNF_C2H2	ZFP37	
ENSG00000136870	IPR007086 IPR007087	-	ZNF_C2H2	ZNF189	
ENSG00000136931	IPR001628	IPR000003 IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR5A1	
ENSG00000136944	IPR001356 IPR007107 IPR009057 IPR012287	-	Homeodomain_like	LMX1B	
ENSG00000136997	IPR011598	IPR002418	HLH_DNA_bd	MYC	
ENSG00000137090	IPR001275	-	DM_DNA_bd	DMRT1	testis

ENSG00000137166	IPR001766 IPR007087 IPR011991	-	TF_Fork_head ZNF_C2H2	FOXP4	
ENSG00000137185	IPR007086 IPR007087	-	ZNF_C2H2	ZNF193	
ENSG00000137203	-	IPR004979 IPR008121	Other	TFAP2A	placenta
ENSG00000137265	IPR001346 IPR011991	-	IRF	IRF4	lymph.node, tonsil
ENSG00000137270	-	IPR003902	Other	GCM1	placenta
ENSG00000137273	IPR001766 IPR011991	-	TF_Fork_head	FOXF2	lung, fetal.lung
ENSG00000137310	-	-	Other	TCF19	
ENSG00000137504	IPR011616	-	TF_bZIP	-	
ENSG00000137709	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	IPR000972	Homeodomain_like POU	POU2F3	
ENSG00000137871	IPR007087	-	ZNF_C2H2	-	uterus
ENSG00000138073	-	-	Other	PREB	
ENSG00000138083	IPR000047 IPR001356 IPR007106 IPR009057 IPR012287	-	Homeodomain_like	SIX3	pituitary
ENSG00000138136	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	LBX1	
ENSG00000138311	IPR007087	-	ZNF_C2H2	ZNF365	spinal.cord, whole.brain
ENSG00000138378	IPR008967 IPR012345	IPR001217 IPR013799 IPR013801	P53_like_DNA_bd STAT_DNA_bd	STAT4	lymph.node, pituitary, whole.blood.JJV
ENSG00000138738	IPR007086 IPR007087	-	ZNF_C2H2	PRDM5	
ENSG00000138795	IPR000910	-	HMG_1/2_box	LEF1	lymph.node, thymus
ENSG00000139083	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	ETV6	
ENSG00000139154	IPR007087	-	ZNF_C2H2	AEBP2	
ENSG00000139352	IPR011598	-	HLH_DNA_bd	ASCL1	spinal.cord, whole.brain
ENSG00000139445	IPR001766 IPR011991	-	TF_Fork_head	-	
ENSG00000139515	IPR000047 IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	IPF1	
ENSG00000139651	IPR007087	-	ZNF_C2H2	ZNF740	
ENSG00000139800	IPR007087	-	ZNF_C2H2	ZIC5	
ENSG00000140009	IPR001628	IPR000324 IPR001723 IPR012239	Hrmn_rcpt_DNA_bd	ESR2	
ENSG00000140044	IPR004827 IPR008917 IPR011700	-	TF_bZIP	-	

ENSG00000140262	IPR009057 IPR011598	-	Homeodomain_like	TCF12	
ENSG00000140265	IPR007086 IPR007087	-	ZNF_C2H2	ZNF690	
ENSG00000140548	IPR007087	-	ZNF_C2H2	ZNF710	
ENSG00000140836	IPR001356 IPR007087 IPR009057 IPR012287	-	Homeodomain_like ZNF_C2H2	ATBF1	prostate, fetal.thyroid
ENSG00000140968	IPR001346 IPR011991	-	IRF	IRF8	lymph.node, thymus, bone.marrow, tonsil, salivary.gland, lung, whole.blood.JJV
ENSG00000140987	IPR007086 IPR007087	-	ZNF_C2H2	ZNF434	
ENSG00000141040	IPR007087	-	ZNF_C2H2	ZNF287	
ENSG00000141448	IPR000679	IPR008013	ZnF_GATA	GATA6	ovary, heart, smooth.muscle, adrenal.gland, uterus, lung, adrenal.cortex, fetal.lung
ENSG00000141510	IPR008967 IPR011615 IPR012346	IPR002117	P53_like_DNA_bd	TP53	thymus, smooth.muscle, whole.blood.JJV
ENSG00000141568	IPR001766 IPR011991	-	TF_Fork_head	FOXK2	testis, adrenal.cortex
ENSG00000141646	IPR003619 IPR013019	IPR001132 IPR013790	MAD_MH1	SMAD4	uterus
ENSG00000141905	IPR003619	IPR000647	MAD_MH1	NFIC	skeletal.muscle.p soas, tongue
ENSG00000141946	IPR007086 IPR007087	-	ZNF_C2H2	ZIM3	
ENSG00000141956	IPR007087	-	ZNF_C2H2	PRDM15	
ENSG00000142025	IPR001275	-	DM_DNA_bd	DMRTC2	
ENSG00000142065	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000142396	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000142528	IPR007086 IPR007087	-	ZNF_C2H2	-	testis
ENSG00000142539	IPR000418 IPR002341 IPR011991	-	Ets	SPIB	lymph.node, heart, tonsil
ENSG00000142556	IPR007087	-	ZNF_C2H2	ZNF614	
ENSG00000142599	IPR000679 IPR001005 IPR009057 IPR012287	-	Homeodomain_like ZnF_GATA	RERE	
ENSG00000142611	IPR007087	-	ZNF_C2H2	PRDM16	
ENSG00000142684	IPR007087	-	ZNF_C2H2	ZNF593	liver
ENSG00000142700	IPR001275	-	DM_DNA_bd	-	
ENSG00000143006	IPR001275	-	DM_DNA_bd	DMRTB1	
ENSG00000143032	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	BARHL2	
ENSG00000143067	IPR007087	-	ZNF_C2H2	ZNF697	

ENSG00000143171	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	RXRG	
ENSG00000143178	IPR008967	IPR001699 IPR002070	P53_like_DNA_bd	SFT2D2 TBX19	testis, pituitary
ENSG00000143190	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	IPR000972	Homeodomain_like POU	POU2F1	
ENSG00000143257	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR1I3	liver
ENSG00000143355	IPR001356 IPR007107 IPR009057 IPR012287	-	Homeodomain_like	LHX9	
ENSG00000143365	IPR001628	IPR001723 IPR003079	Hrmn_rcpt_DNA_bd	RORC	skeletal.muscle.psoas, liver, kidney
ENSG00000143373	IPR007087	-	ZNF_C2H2	ZNF687	
ENSG00000143390	IPR003150 IPR011991	-	RFX_DNA_bd	RFX5	lymph.node
ENSG00000143437	IPR011598	IPR001067	HLH_DNA_bd	ARNT	
ENSG00000143578	IPR004827 IPR008917 IPR011616	IPR001630	TF_bZIP	CREB3L4	
ENSG00000143842	IPR000910	-	HMG_1/2_box	SOX13	
ENSG00000143867	IPR007087	-	ZNF_C2H2	OSR1	
ENSG00000143995	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	MEIS1	trachea, ovary, smooth.muscle, appendix, adrenal.gland, uterus, salivary.gland, adrenal.cortex, fetal.lung
ENSG00000144026	IPR007087	-	ZNF_C2H2	ZNF514	
ENSG00000144218	-	-	Other	AFF3	
ENSG00000144331	IPR007087	-	ZNF_C2H2	ZNF533	
ENSG00000144355	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000144747	-	-	Other	TMF1	
ENSG00000144792	IPR007086 IPR007087	-	ZNF_C2H2	ZNF660	
ENSG00000144852	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR1I2	pancreas, liver
ENSG00000145908	IPR007086 IPR007087	-	ZNF_C2H2	ZNF300	
ENSG00000146587	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000146592	IPR004827 IPR007087 IPR008917 IPR011616 IPR011700	-	TF_bZIP ZNF_C2H2	CREB5	
ENSG00000146757	IPR007086 IPR007087	-	ZNF_C2H2	ZNF92	
ENSG00000147117	IPR007086 IPR007087	-	ZNF_C2H2	ZNF157	
ENSG00000147118	IPR007086 IPR007087	-	ZNF_C2H2	ZNF182	

ENSG00000147124	IPR007086 IPR007087	-	ZNF_C2H2	ZNF41	
ENSG00000147180	IPR007087	IPR006794	ZNF_C2H2	ZNF6	
ENSG00000147421	IPR001356 IPR009057 IPR010982 IPR012287	-	Homeodomain_like	HMBOX1	
ENSG00000147488	IPR013681	IPR002515	Myelin_TF	ST18	spinal.cord
ENSG00000147596	IPR007087	-	ZNF_C2H2	PRDM14	
ENSG00000147789	IPR007086 IPR007087	-	ZNF_C2H2	ZNF7	skeletal.muscle.psoas
ENSG00000147862	IPR003619	IPR000647	MAD_MH1	NFIB	smooth.muscle
ENSG00000148143	IPR007087	-	ZNF_C2H2	ZNF462	
ENSG00000148200	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR6A1	
ENSG00000148337	IPR007087	-	ZNF_C2H2	CIZ1	
ENSG00000148516	IPR001356 IPR007087 IPR009057 IPR012287	-	Homeodomain_like ZNF_C2H2	TCF8	spinal.cord, human.cultured.adipocyte, appendix, uterus
ENSG00000148704	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000148737	IPR000910	-	HMG_1/2_box	TCF7L2	uterus, lung
ENSG00000148826	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	NKX6-2	
ENSG00000149050	IPR007086 IPR007087	-	ZNF_C2H2	ZNF214	
ENSG00000149054	IPR007087	-	ZNF_C2H2	ZNF215	
ENSG00000149922	IPR008967	IPR001699 IPR002070	P53_like_DNA_bd	-	
ENSG00000150051	IPR001356 IPR009057	-	Homeodomain_like	MKX	
ENSG00000150347	-	-	Other	-	
ENSG00000150907	IPR001766 IPR011991	-	TF_Fork_head	FOXO1A	ovary
ENSG00000151090	IPR001628	IPR001723 IPR001728	Hrmn_rcpt_DNA_bd	THRB	
ENSG00000151514	IPR007087	-	ZNF_C2H2	SALL3	
ENSG00000151615	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU4F2	
ENSG00000151623	IPR001628	IPR000324	Hrmn_rcpt_DNA_bd	NR3C2	whole.brain, thyroid
ENSG00000151650	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	VENTX	
ENSG00000151702	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	FLI1	whole.blood.JJV
ENSG00000151789	IPR007087	-	ZNF_C2H2	ZNF659	
ENSG00000151963	IPR007086 IPR007087	-	ZNF_C2H2	-	

ENSG00000152192	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU4F1	
ENSG00000152284	IPR000910	-	HMG_1/2_box	TCF7L1	
ENSG00000152433	IPR007086 IPR007087	-	ZNF_C2H2	ZNF547	testis
ENSG00000152454	IPR007086 IPR007087	-	ZNF_C2H2	ZNF256	
ENSG00000152467	IPR007087	-	ZNF_C2H2	ZSCAN1	
ENSG00000152518	IPR000571	-	ZnF_CCCH	ZFP36L2	whole.blood.JJV, thyroid
ENSG00000152784	IPR007087	-	ZNF_C2H2	PRDM8	
ENSG00000152804	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HHEX	tonsil, fetal.thyroid, whole.blood.JJV, thyroid, fetal.liver
ENSG00000152926	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000152977	IPR007087	-	ZNF_C2H2	ZIC1	spinal.cord, whole.brain, fetal.brain
ENSG00000153234	IPR001628	IPR001723 IPR003070 IPR003073	Hrmn_rcpt_DNA_bd	NR4A2	trachea, ovary, adrenal.gland, adrenal.cortex
ENSG00000153560	IPR007604	-	CP2	UBP1	lymph.node, thymus
ENSG00000153779	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	TGIF2LX	
ENSG00000153807	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXA10	prostate, uterus
ENSG00000153814	IPR007087	-	ZNF_C2H2	-	
ENSG00000153879	IPR004827 IPR011700	-	TF_bZIP	CEBPG	
ENSG00000153896	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000154227	IPR001356 IPR009057	-	Homeodomain_like	LASS3	
ENSG00000154727	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	GABPA	
ENSG00000154832	-	-	Other	CXXC1	
ENSG00000154957	IPR007087	-	ZNF_C2H2	-	
ENSG00000155090	IPR007087	-	ZNF_C2H2	KLF10	placenta, lung
ENSG00000155545	IPR001005 IPR009057	-	Homeodomain_like	MIER3	
ENSG00000155592	IPR007087 IPR009057	-	Homeodomain_like ZNF_C2H2	ZNF694	
ENSG00000156030	IPR001005 IPR009057	-	Homeodomain_like	-	
ENSG00000156127	IPR004827 IPR008917 IPR011616	-	TF_bZIP	BATF	

ENSG00000156150	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000156273	IPR004827 IPR008917 IPR011616	-	TF_bZIP	BACH1	whole.blood.JJV
ENSG00000156853	IPR007086 IPR007087	-	ZNF_C2H2	ZNF689	
ENSG00000156925	IPR007087	-	ZNF_C2H2	ZIC3	
ENSG00000157429	IPR007086 IPR007087	-	ZNF_C2H2	ZNF19	
ENSG00000157514	-	IPR000580	Other	TSC22D3	
ENSG00000157554	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	ERG	placenta, fetal.lung
ENSG00000157557	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	ETS2	whole.brain, lung, fetal.lung
ENSG00000157613	IPR004827 IPR008917 IPR011616 IPR011700	IPR001630	TF_bZIP	CREB3L1	
ENSG00000157657	IPR007087	-	ZNF_C2H2	-	
ENSG00000157933	IPR009061 IPR010919	-	SAND_like	SKI	
ENSG00000158055	IPR007604	-	CP2	GRHL3	
ENSG00000158691	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000158711	IPR000418 IPR002341 IPR011991	-	Ets	ELK4	
ENSG00000158773	IPR011598	-	HLH_DNA_bd	USF1	
ENSG00000158805	IPR007087	-	ZNF_C2H2	ZNF276	
ENSG00000159184	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXB13	prostate
ENSG00000159216	IPR008967 IPR012346 IPR013524	IPR000040	P53_like_DNA_bd	RUNX1	thymus
ENSG00000159263	IPR011598	IPR001067	HLH_DNA_bd	SIM2	
ENSG00000159387	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	IRX6	
ENSG00000159556	IPR001356 IPR007107 IPR009057 IPR012287	-	Homeodomain_like	ISL2	
ENSG00000159882	IPR007087	-	ZNF_C2H2	ZNF230	
ENSG00000159885	IPR007087	-	ZNF_C2H2	ZNF222	
ENSG00000159904	IPR007086 IPR007087	-	ZNF_C2H2	ZNF225	
ENSG00000159905	IPR007086 IPR007087	-	ZNF_C2H2	ZNF234	
ENSG00000159915	IPR007086 IPR007087	-	ZNF_C2H2	ZNF233	
ENSG00000159917	IPR007086 IPR007087	-	ZNF_C2H2	ZNF235	
ENSG00000160007	-	-	Other	GRLF1	

ENSG00000160062	IPR007087	-	ZNF_C2H2	-	
ENSG00000160113	IPR001628	IPR001723 IPR003068	Hrmn_rcpt_DNA_bd	NR2F6	liver, placenta
ENSG00000160199	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	PKNOX1	
ENSG00000160224	IPR000770 IPR010919	-	SAND_like	AIRE	
ENSG00000160229	IPR007087	-	ZNF_C2H2	-	
ENSG00000160321	IPR007086 IPR007087	-	ZNF_C2H2	ZNF208	
ENSG00000160336	IPR007086 IPR007087 IPR010921	-	ZNF_C2H2	-	
ENSG00000160352	IPR007086 IPR007087	-	ZNF_C2H2	ZNF714	
ENSG00000160685	IPR007087	-	ZNF_C2H2	ZBTB7B	
ENSG00000160908	IPR007086 IPR007087	-	ZNF_C2H2	ZNF394	whole.blood.JJV
ENSG00000160961	IPR003656 IPR007086 IPR007087	-	ZNF_C2H2 Znf_BED_prd	ZNF333	
ENSG00000160967	IPR001356 IPR003350 IPR007108 IPR009057 IPR010982 IPR012287	-	Hmoeo_CUT Homeodomain_like	CUTL1	
ENSG00000160973	IPR001766 IPR011991	-	TF_Fork_head	FOXH1	
ENSG00000161298	IPR007086 IPR007087	-	ZNF_C2H2	ZNF382	
ENSG00000161405	IPR007087	-	ZNF_C2H2	ZNFN1A3	
ENSG00000161551	IPR007086 IPR007087	-	ZNF_C2H2	ZNF577	
ENSG00000161642	IPR007087	-	ZNF_C2H2	ZNF385	
ENSG00000161853	IPR007087	-	ZNF_C2H2	ZBTB12	
ENSG00000161914	IPR007087	-	ZNF_C2H2	ZNF653	
ENSG00000161940	IPR007087	-	ZNF_C2H2	BCL6B	
ENSG00000162086	IPR007086 IPR007087	-	ZNF_C2H2	ZNF75A	
ENSG00000162367	IPR011598	-	HLH_DNA_bd	TAL1	bone.marrow, whole.blood.JJV, fetal.liver
ENSG00000162419	IPR000770 IPR010919	-	SAND_like	GMEB1	
ENSG00000162599	IPR003619	IPR000647	MAD_MH1	NFIA	
ENSG00000162624	IPR001356 IPR007107 IPR009057	-	Homeodomain_like	LHX8	
ENSG00000162676	IPR007087	-	ZNF_C2H2	GFI1	thymus, bone.marrow
ENSG00000162702	IPR007087	-	ZNF_C2H2	ZNF281	
ENSG00000162714	IPR007087	-	ZNF_C2H2	ZNF496	
ENSG00000162761	IPR001356 IPR007107 IPR009057 IPR012287	-	Homeodomain_like	LMX1A	

ENSG00000162772	IPR004827 IPR008917 IPR011700	-	TF_bZIP	ATF3	pancreas, trachea, placenta, appendix, lung,
ENSG00000162924	IPR002909 IPR008967 IPR011539	IPR000451	IPT_TIG_rcpt P53_like_DNA_bd	REL	tonsil
ENSG00000162992	IPR011598	-	HLH_DNA_bd	NEUROD1	
ENSG00000163064	IPR000047 IPR000747 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	EN1	
ENSG00000163067	IPR007086 IPR007087	-	ZNF_C2H2	ZNF2	
ENSG00000163132	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	MSX1	human.cultured.adipocyte, uterus, pituitary
ENSG00000163435	IPR000418 IPR002341 IPR003118 IPR011991	-	Ets SAM_PNT	ELF3	pancreas, trachea
ENSG00000163497	IPR000418 IPR002341 IPR011991	-	Ets	FEV	
ENSG00000163508	IPR008967	IPR001699	P53_like_DNA_bd	EOMES	
ENSG00000163565	-	-	Other	IFI16	lymph.node, thymus, human.cultured.adipocyte, tonsil, uterus, whole.blood.JJV
ENSG00000163623	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	NKX6-1	
ENSG00000163666	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	HESX1	
ENSG00000163795	IPR007087	-	ZNF_C2H2	ZNF513	
ENSG00000163848	IPR007087	-	ZNF_C2H2	SLC12A8 ZNF148	fetal.thyroid
ENSG00000163884	IPR007087	-	ZNF_C2H2	KLF15	
ENSG00000163909	IPR011598	-	HLH_DNA_bd	HEYL	
ENSG00000164011	IPR007086 IPR007087	-	ZNF_C2H2	ZNF691	
ENSG00000164048	IPR007087	-	ZNF_C2H2	-	
ENSG00000164093	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	PITX2	
ENSG00000164107	IPR011598	-	HLH_DNA_bd	HAND2	
ENSG00000164185	IPR007087	-	ZNF_C2H2	ZNF474	
ENSG00000164256	IPR007086 IPR007087	-	ZNF_C2H2	PRDM7 PRDM9	
ENSG00000164330	IPR002909 IPR011598	IPR003523	HLH_DNA_bd IPT_TIG_rcpt	EBF	
ENSG00000164379	IPR001766 IPR011991	-	TF_Fork_head	FOXQ1	

ENSG00000164438	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	TLX3	
ENSG00000164458	IPR008967	IPR001699 IPR002070	P53_like_DNA_bd	T	
ENSG00000164532	IPR008967	IPR001699	P53_like_DNA_bd	TBX20	
ENSG00000164600	IPR011598	-	HLH_DNA_bd	NEUROD6	whole.brain, fetal.brain
ENSG00000164631	IPR007086 IPR007087	-	ZNF_C2H2	ZNF12	uterus, pituitary
ENSG00000164651	IPR007087	-	ZNF_C2H2	SP8	
ENSG00000164683	IPR011598	-	HLH_DNA_bd	-	spinal.cord, whole.brain, lung, fetal.brain
ENSG00000164684	IPR007087	-	ZNF_C2H2	ZNF704	
ENSG00000164736	IPR000910	-	HMG_1/2_box	SOX17	uterus, lung
ENSG00000164749	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	HNF4G	
ENSG00000164778	IPR000047 IPR000747 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	EN2	
ENSG00000164853	IPR001356 IPR007104 IPR009057	-	Homeodomain_like	-	
ENSG00000164900	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	GBX1	
ENSG00000164920	IPR007087	-	ZNF_C2H2	-	testis, ovary, human.cultured.a dipocyte, uterus
ENSG00000165030	IPR004827 IPR011700	-	TF_bZIP	NFIL3	ovary, human.cultured.a dipocyte, whole.blood.JJV, fetal.lung, adrenal.cortex
ENSG00000165156	IPR001356 IPR007087 IPR009057 IPR012287	-	Homeodomain_like ZNF_C2H2	ZHX1	
ENSG00000165259	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	CXorf43	
ENSG00000165388	IPR007087	-	ZNF_C2H2	-	
ENSG00000165462	IPR000047 IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	PHOX2A	
ENSG00000165495	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	PKNOX2	whole.brain
ENSG00000165512	IPR007086 IPR007087	-	ZNF_C2H2	ZNF22	pancreas, lymph.node, thymus, fetal.brain, whole.blood.JJV, fetal.lung, fetal.liver

ENSG00000165556	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	CDX2	
ENSG00000165588	IPR001356 IPR007104 IPR009057 IPR012287	IPR003022 IPR003025	Homeodomain_like	OTX2	
ENSG00000165606	IPR001356 IPR007104 IPR009057	-	Homeodomain_like	-	
ENSG00000165643	IPR011598	-	HLH_DNA_bd	SOHLH1	
ENSG00000165655	IPR007087	-	ZNF_C2H2	ZNF503	
ENSG00000165659	IPR009061	-	Other	DACH1	
ENSG00000165684	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	SNAPC4	testis
ENSG00000165702	IPR007086 IPR007087	-	ZNF_C2H2	GFI1B	
ENSG00000165804	IPR007087	-	ZNF_C2H2	ZNF219	
ENSG00000165821	IPR007087	-	ZNF_C2H2	SALL2	spinal.cord, whole.brain, pituitary
ENSG00000166080	IPR007087	-	ZNF_C2H2	ZNF123	
ENSG00000166188	IPR007086 IPR007087	-	ZNF_C2H2	ZNF319	
ENSG00000166211	IPR000418 IPR002341 IPR011991	-	Ets	-	
ENSG00000166261	IPR007087	-	ZNF_C2H2	ZNF202	
ENSG00000166402	-	-	Other	-	
ENSG00000166478	IPR007087	-	ZNF_C2H2	ZNF143	whole.blood.JJV
ENSG00000166526	IPR007086 IPR007087	-	ZNF_C2H2	ZNF3	
ENSG00000166529	IPR007086 IPR007087	-	ZNF_C2H2	ZNF38	
ENSG00000166540	IPR007087	-	ZNF_C2H2	ZNF407	
ENSG00000166704	IPR007086 IPR007087	-	ZNF_C2H2	ZNF606	
ENSG00000166716	IPR007087	-	ZNF_C2H2	ZNF592	
ENSG00000166860	IPR007087	-	ZNF_C2H2	ZBTB39	
ENSG00000166888	IPR008967 IPR012345	IPR001217 IPR013799 IPR013801	P53_like_DNA_bd STAT_DNA_bd	STAT6	whole.blood.JJV
ENSG00000166925	-	IPR000580	Other	TSC22D4	spinal.cord
ENSG00000166949	IPR003619 IPR013019	IPR001132 IPR013790	MAD_MH1	SMAD3	smooth.muscle
ENSG00000167034	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	NKX3-1	trachea, testis, prostate
ENSG00000167074	IPR004827 IPR011700	-	TF_bZIP	TEF	
ENSG00000167081	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	PBX3	ovary, adrenal.gland, fetal.thyroid, adrenal.cortex

ENSG00000167157	IPR000047 IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	PRRX2	
ENSG00000167182	IPR007087	-	ZNF_C2H2	SP2	heart, thymus, lung, whole.blood.JJV, thyroid
ENSG00000167232	IPR007086 IPR007087	-	ZNF_C2H2	ZNF91	pituitary
ENSG00000167377	IPR007086 IPR007087	-	ZNF_C2H2	ZNF23	
ENSG00000167380	IPR007086 IPR007087	-	ZNF_C2H2	ZNF226	
ENSG00000167383	IPR007087	-	ZNF_C2H2	ZNF229	
ENSG00000167384	IPR007086 IPR007087	-	ZNF_C2H2	ZNF180	
ENSG00000167394	IPR007086 IPR007087	-	ZNF_C2H2	ZNF668	
ENSG00000167395	IPR007087	-	ZNF_C2H2	ZNF646	
ENSG00000167528	IPR007087	-	ZNF_C2H2	ZNF641	
ENSG00000167554	IPR007086 IPR007087	-	ZNF_C2H2	ZNF610	
ENSG00000167562	IPR007086 IPR007087	-	ZNF_C2H2	ZNF701	
ENSG00000167625	IPR007087	-	ZNF_C2H2	-	
ENSG00000167635	IPR007086 IPR007087	-	ZNF_C2H2	ZNF146	
ENSG00000167637	IPR007086 IPR007087	-	ZNF_C2H2	ZNF345	
ENSG00000167685	IPR007087	-	ZNF_C2H2	ZNF444	
ENSG00000167766	IPR007086 IPR007087	-	ZNF_C2H2	ZNF83	lymph.node, uterus
ENSG00000167771	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	RCOR2	
ENSG00000167785	IPR007086 IPR007087	-	ZNF_C2H2	ZNF558	
ENSG00000167800	IPR008967	IPR001699	P53_like_DNA_bd	TBX10	
ENSG00000167840	IPR007086 IPR007087	-	ZNF_C2H2	ZNF232	
ENSG00000167967	IPR000637 IPR007087	-	AT_hook_DNA_bd ZNF_C2H2	E4F1	
ENSG00000167981	IPR007087	-	ZNF_C2H2	ZNF597	
ENSG00000168062	IPR004827 IPR011616	-	TF_bZIP	BATF2	
ENSG00000168214	IPR002909 IPR008967	-	IPT_TIG_rcpt P53_like_DNA_bd	RBPSUH	uterus, whole.blood.JJV
ENSG00000168267	IPR007086 IPR007087 IPR011598	-	HLH_DNA_bd ZNF_C2H2	PTF1A	
ENSG00000168269	IPR001766	-	TF_Fork_head	-	
ENSG00000168310	IPR001346 IPR011991	-	IRF	IRF2	lymph.node, bone.marrow, adrenal.gland, whole.blood.JJV
ENSG00000168348	IPR007087	-	ZNF_C2H2	INSM2	
ENSG00000168468	IPR004827 IPR008917 IPR011616	IPR001630	TF_bZIP	CREBL1	

ENSG00000168477	IPR004827 IPR008917 IPR011616	IPR001630	TF_bZIP	CREBL1 TNXB	heart, adrenal.gland, uterus, tongue, lung, thyroid, adrenal.cortex
ENSG00000168505	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	GBX2	
ENSG00000168610	IPR008967 IPR012345	IPR001217 IPR013799 IPR013801	P53_like_DNA_bd STAT_DNA_bd	-	
ENSG00000168661	IPR007086 IPR007087	-	ZNF_C2H2	ZNF30	
ENSG00000168779	IPR000047 IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	SHOX2	
ENSG00000168795	IPR007087	-	ZNF_C2H2	ZBTB5	
ENSG00000168813	IPR007087	-	ZNF_C2H2	ZNF507	
ENSG00000168826	IPR007086 IPR007087	-	ZNF_C2H2	ZNF509	
ENSG00000168875	IPR000135 IPR000910	-	HMG_1/2_box	SOX14	
ENSG00000168916	IPR007087	-	ZNF_C2H2	-	
ENSG00000169016	-	IPR003316	Other	E2F6	
ENSG00000169083	IPR001628	IPR001103	Hrmn_rcpt_DNA_bd	AR	liver, human.cultured.adipocyte, prostate, uterus
ENSG00000169084	IPR003656	-	Znf_BED_prd	DHRSX ZBED1	liver
ENSG00000169131	IPR007086 IPR007087	-	ZNF_C2H2	ZNF354A	
ENSG00000169136	IPR004827 IPR008917 IPR011616	-	TF_bZIP	ATF5	liver, smooth.muscle, fetal.lung, fetal.liver
ENSG00000169155	IPR007087	-	ZNF_C2H2	ZBTB43	
ENSG00000169260	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000169297	-	IPR001723	Other	NR0B1	testis
ENSG00000169548	IPR007087	-	ZNF_C2H2	SUHW1	
ENSG00000169554	IPR001356 IPR007087 IPR012287	-	Homeodomain_like ZNF_C2H2	ZFHX1B	spinal.cord, uterus, fetal.brain,
ENSG00000169594	IPR007087	-	ZNF_C2H2	BNC1	
ENSG00000169635	IPR007087	-	ZNF_C2H2	HIC2	placenta, fetal.brain, fetal.liver
ENSG00000169740	IPR007086 IPR007087	-	ZNF_C2H2	ZNF32	
ENSG00000169840	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	GSH1	
ENSG00000169856	IPR001356 IPR003350 IPR007108 IPR009057 IPR012287	-	Hmoeo_CUT Homeodomain_like	ONECUT1	

ENSG00000169926	IPR007087	-	ZNF_C2H2	KLF13	thymus, human.cultured.adipocyte, lung
ENSG00000169946	IPR007087	-	ZNF_C2H2	ZFPM2	smooth.muscle, uterus
ENSG00000169953	IPR000232 IPR002341 IPR011991	-	Heat shock factor (HSF)-type	HSFY1 HSFY2	
ENSG00000169981	IPR007086 IPR007087	-	ZNF_C2H2	ZNF35	
ENSG00000170122	IPR001766 IPR011991	-	TF_Fork_head	-	
ENSG00000170166	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXD4	
ENSG00000170178	IPR000047 IPR001356 IPR009057	-	Homeodomain_like	HOXD12	
ENSG00000170260	IPR007087	-	ZNF_C2H2	ZNF212	
ENSG00000170265	IPR007087	-	ZNF_C2H2	ZNF282	
ENSG00000170322	-	-	Other	NFRKB	
ENSG00000170325	IPR007087	-	ZNF_C2H2	PRDM10	
ENSG00000170345	IPR004827 IPR008917 IPR011700	-	TF_bZIP	FOS	trachea, lymph.node, bone.marrow, lung, thyroid
ENSG00000170365	IPR003619 IPR013019	IPR001132 IPR013790	MAD_MH1	SMAD1	
ENSG00000170370	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	EMX2	whole.brain, uterus, kidney
ENSG00000170374	IPR007087	-	ZNF_C2H2	SP7	
ENSG00000170485	IPR011598	IPR001067	HLH_DNA_bd	NPAS2	whole.brain, smooth.muscle
ENSG00000170549	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	IRX1	
ENSG00000170561	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	IRX2	
ENSG00000170577	IPR000047 IPR001356 IPR007106 IPR009057 IPR012287	-	Homeodomain_like	SIX2	
ENSG00000170581	IPR008967 IPR012345	IPR001217 IPR013799 IPR013801	P53_like_DNA_bd STAT_DNA_bd	STAT2	
ENSG00000170608	IPR001766 IPR011991	-	TF_Fork_head	FOXA3	
ENSG00000170616	IPR007087	-	ZNF_C2H2	SCRT1	
ENSG00000170631	IPR007086 IPR007087	-	ZNF_C2H2	ZNF16	
ENSG00000170653	IPR004827 IPR007087 IPR008917 IPR011616 IPR011700	-	TF_bZIP ZNF_C2H2	ATF7	
ENSG00000170684	IPR007087	-	ZNF_C2H2	ZNF342	

ENSG00000170689	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXB9	
ENSG00000170802	IPR001766 IPR011991	-	TF_Fork_head	HTLF	
ENSG00000170949	IPR007086 IPR007087	-	ZNF_C2H2	ZNF160	
ENSG00000170954	IPR003656 IPR007086 IPR007087	-	ZNF_C2H2 Znf_BED_prd	ZNF415	
ENSG00000171032	IPR007086 IPR007087	-	ZNF_C2H2	ZNF350	whole.blood.JJV
ENSG00000171056	IPR000910	-	HMG_1/2_box	SOX7	
ENSG00000171161	IPR007086 IPR007087	-	ZNF_C2H2	ZNF672	
ENSG00000171163	IPR007087	-	ZNF_C2H2	ZNF692	
ENSG00000171223	IPR004827 IPR008917 IPR011616	IPR002112	TF_bZIP	JUNB	
ENSG00000171291	IPR007086 IPR007087	-	ZNF_C2H2	ZNF439	
ENSG00000171295	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000171425	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000171443	IPR000637 IPR007087	-	AT_hook_DNA_bd ZNF_C2H2	ZNF524	
ENSG00000171448	IPR007087	-	ZNF_C2H2	ZBTB26	
ENSG00000171466	IPR007086 IPR007087	-	ZNF_C2H2	ZNF562	
ENSG00000171467	IPR007087	-	ZNF_C2H2	ZNF318	
ENSG00000171469	IPR007086 IPR007087	-	ZNF_C2H2	ZNF561	
ENSG00000171532	IPR011598	-	HLH_DNA_bd	NEUROD2	skeletal.muscle.psoas, whole.brain,
ENSG00000171540	IPR000047 IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	OTP	
ENSG00000171574	IPR007086 IPR007087	-	ZNF_C2H2	ZNF584	
ENSG00000171606	IPR007086 IPR007087	-	ZNF_C2H2	ZNF274	lymph.node, whole.blood.JJV
ENSG00000171634	IPR000637	-	AT_hook_DNA_bd	-	tonsil,
ENSG00000171649	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000171656	IPR000418 IPR002341 IPR011991	-	Ets	ETV5	placenta, whole.brain, adrenal.gland, pituitary
ENSG00000171735	IPR002909 IPR005559	-	CG-1 IPT_TIG_rcpt	CAMTA1	spinal.cord, whole.brain, fetal.brain
ENSG00000171786	IPR011598	-	HLH_DNA_bd	NHLH1	
ENSG00000171794	IPR009057	-	Homeodomain_like	UTF1	
ENSG00000171817	IPR007086 IPR007087	-	ZNF_C2H2	ZNF540	
ENSG00000171827	IPR007086 IPR007087	-	ZNF_C2H2	ZNF570	

ENSG00000171843	-	IPR005033	Other	MLLT3	fetal.brain
ENSG00000171872	IPR007087	-	ZNF_C2H2	KLF17	
ENSG00000171940	IPR007087	-	ZNF_C2H2	ZNF217	lymph.node, thymus, placenta,
ENSG00000171956	IPR001766	-	TF_Fork_head	FOXB1	
ENSG00000171970	IPR007086 IPR007087	-	ZNF_C2H2	ZNF57	
ENSG00000172000	IPR007086 IPR007087	-	ZNF_C2H2	ZNF556	
ENSG00000172006	IPR007086 IPR007087	-	ZNF_C2H2	ZNF554	
ENSG00000172018	IPR000418 IPR002341 IPR011991	-	Ets	-	
ENSG00000172059	IPR007087	-	ZNF_C2H2	KLF11	testis, uterus,
ENSG00000172070	IPR007087	-	ZNF_C2H2	SCRT2 SRXN1	
ENSG00000172216	IPR004827 IPR011700	-	TF_bZIP	CEBPB	
ENSG00000172238	IPR011598	-	HLH_DNA_bd	ATOH1	
ENSG00000172262	IPR007087	-	ZNF_C2H2	-	
ENSG00000172273	IPR007087	-	ZNF_C2H2	MIZF	
ENSG00000172379	IPR011598	IPR001067	HLH_DNA_bd	ARNT2	spinal.cord, whole.brain, fetal.brain,
ENSG00000172466	IPR007086 IPR007087	-	ZNF_C2H2	ZNF24	thymus
ENSG00000172468	IPR000232 IPR002341 IPR011991	-	Heat shock factor (HSF)-type	HSFY1 HSFY2	
ENSG00000172667	IPR007087	-	ZNF_C2H2	-	human.cultured.adipocyte, whole.brain, smooth.muscle
ENSG00000172748	IPR007086 IPR007087	-	ZNF_C2H2	ZNF596	
ENSG00000172789	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXC5	
ENSG00000172818	IPR007087	-	ZNF_C2H2	OVOL1	
ENSG00000172819	IPR001628	IPR001723 IPR003078	Hrmn_rcpt_DNA_bd	RARG	
ENSG00000172845	IPR007087	-	ZNF_C2H2	-	whole.blood.JJV
ENSG00000172888	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000173039	IPR002909 IPR008967 IPR011539	IPR000451	IPT_TIG_rcpt P53_like_DNA_bd	RELA	
ENSG00000173041	IPR007086 IPR007087	-	ZNF_C2H2	ZNF680	
ENSG00000173068	IPR007087	-	ZNF_C2H2	BNC2	uterus
ENSG00000173153	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	ESRRA	
ENSG00000173245	IPR000327 IPR010982	-	POU	-	
ENSG00000173253	IPR001275	-	DM_DNA_bd	DMRT2	
ENSG00000173258	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000173275	IPR007087	-	ZNF_C2H2	ZNF449	
ENSG00000173276	IPR007087	-	ZNF_C2H2	ZNF295	

ENSG00000173404	IPR007087	-	ZNF_C2H2	INSM1	pituitary,
ENSG00000173480	IPR007086 IPR007087	-	ZNF_C2H2	ZNF417	
ENSG00000173545	IPR007087	-	ZNF_C2H2	ZNF622	
ENSG00000173757	IPR008967 IPR012345	IPR013799 IPR013801	P53_like_DNA_bd STAT_DNA_bd	STAT5B	placenta, whole.blood.JJV
ENSG00000173917	IPR000047 IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXB2	spinal.cord, trachea, adrenal.gland
ENSG00000173976	IPR001356 IPR007104 IPR009057	-	Homeodomain_like	RAXL1	
ENSG00000174197	IPR008967 IPR011598	IPR001699	HLH_DNA_bd P53_like_DNA_bd	MGA	
ENSG00000174255	IPR007086 IPR007087	-	ZNF_C2H2	ZNF80	
ENSG00000174279	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	EVX2	
ENSG00000174282	IPR007087	-	ZNF_C2H2	ZBTB4	
ENSG00000174306	IPR001356 IPR007087 IPR009057 IPR012287	-	Homeodomain_like ZNF_C2H2	ZHX3	kidney
ENSG00000174332	IPR007087	-	ZNF_C2H2	GLIS1	
ENSG00000174586	IPR007086 IPR007087	-	ZNF_C2H2	ZNF497	
ENSG00000174595	IPR007087	-	ZNF_C2H2	KLF14	
ENSG00000174652	IPR003656 IPR007086 IPR007087	-	ZNF_C2H2 Znf_BED_prd	ZNF266	lymph.node, tonsil, fetal.lung
ENSG00000174738	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR1D2	uterus, pituitary, thyroid
ENSG00000174963	IPR007087	-	ZNF_C2H2	ZIC4	
ENSG00000175105	IPR007087	-	ZNF_C2H2	ZNF654	
ENSG00000175197	IPR004827 IPR011700	-	TF_bZIP	DDIT3	
ENSG00000175213	IPR007087	-	ZNF_C2H2	ZNF408	
ENSG00000175322	IPR007086 IPR007087	-	ZNF_C2H2	ZNF519	
ENSG00000175325	IPR000047 IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	PROP1	
ENSG00000175329	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000175387	IPR003619 IPR013019	IPR001132 IPR013790	MAD_MH1	SMAD2	
ENSG00000175395	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000175550	IPR003958	-	CBFA_NFYB_domain	DRAP1	smooth.muscle
ENSG00000175592	IPR004827 IPR008917 IPR011616	-	TF_bZIP	FOSL1	trachea
ENSG00000175691	IPR007086 IPR007087	-	ZNF_C2H2	ZNF77	

ENSG00000175727	IPR011598	-	HLH_DNA_bd	MLXIP	
ENSG00000175745	IPR001628	IPR001723 IPR003068	Hrmn_rcpt_DNA_bd	NR2F1	whole.brain, fetal.brain, fetal.lung
ENSG00000175787	IPR007086 IPR007087	-	ZNF_C2H2	ZNF169	
ENSG00000175809	IPR007087	-	ZNF_C2H2	ZNF645	
ENSG00000175832	IPR000418 IPR002341 IPR011991	-	Ets	ETV4	
ENSG00000175879	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXD8	
ENSG00000175885	IPR007086 IPR007087	-	ZNF_C2H2	ZNF611	
ENSG00000176009	IPR011598	-	HLH_DNA_bd	ASCL3	
ENSG00000176024	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000176083	IPR007087	-	ZNF_C2H2	ZNF683	
ENSG00000176160	IPR000232 IPR002341 IPR011991	-	Heat shock factor (HSF)-type	HSF5	
ENSG00000176165	IPR001766 IPR011991	-	TF_Fork_head	FOXG1B FOXG1C	testis, whole.brain,
ENSG00000176222	IPR007086 IPR007087	-	ZNF_C2H2	ZNF404	
ENSG00000176232	IPR007087	-	ZNF_C2H2	-	
ENSG00000176293	IPR007086 IPR007087	-	ZNF_C2H2	ZNF135	
ENSG00000176302	IPR001766 IPR011991	-	TF_Fork_head	FOXR1	
ENSG00000176371	IPR001356 IPR007086 IPR007087	-	Homeodomain_like ZNF_C2H2	ZSCAN2	
ENSG00000176399	IPR001275	-	DM_DNA_bd	DMRTA1	
ENSG00000176407	IPR007087	-	ZNF_C2H2	-	testis
ENSG00000176472	IPR007087	-	ZNF_C2H2	ZNF575	
ENSG00000176678	IPR001766 IPR011991	-	TF_Fork_head	FOXL1	
ENSG00000176679	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	TGIF2LY	
ENSG00000176692	IPR001766 IPR011991	-	TF_Fork_head	FOXC2	
ENSG00000176842	IPR001356 IPR009057	-	Homeodomain_like	IRX5	
ENSG00000176887	IPR000910	-	HMG_1/2_box	SOX11	fetal.brain
ENSG00000177030	IPR000770 IPR010919	-	SAND_like	DEAF1	whole.brain, fetal.brain
ENSG00000177045	IPR001356 IPR007106 IPR009057 IPR012287	-	Homeodomain_like	SIX5	
ENSG00000177125	IPR007087	-	ZNF_C2H2	ZBTB34	
ENSG00000177311	IPR007087	-	ZNF_C2H2	ZBTB38	smooth.muscle, uterus
ENSG00000177374	IPR007087	-	ZNF_C2H2	HIC1	
ENSG00000177426	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	TGIF	placenta, smooth.muscle, salivary.gland

ENSG00000177463	IPR001628	IPR000324 IPR001723	Hrmn_rcpt_DNA_bd	NR2C2	skeletal.muscle.psoas
ENSG00000177468	IPR011598	-	HLH_DNA_bd	OLIG3	
ENSG00000177485	IPR007087	-	ZNF_C2H2	ZBTB33	
ENSG00000177508	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	IRX3	
ENSG00000177551	IPR011598	-	HLH_DNA_bd	NHLH2	
ENSG00000177599	IPR007086 IPR007087	-	ZNF_C2H2	ZNF491	
ENSG00000177606	IPR004827 IPR008917 IPR011616	IPR002112	TF_bZIP	JUN	uterus, lung, thyroid
ENSG00000177732	IPR000910	-	HMG_1/2_box	-	fetal.brain
ENSG00000177835	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000177842	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000177853	IPR007087	-	ZNF_C2H2	ZNF518	
ENSG00000177873	IPR007086 IPR007087	-	ZNF_C2H2	ZNF619	
ENSG00000177888	IPR007087	-	ZNF_C2H2	ZBTB41	
ENSG00000177932	IPR007086 IPR007087	-	ZNF_C2H2	ZNF354C	
ENSG00000178042	IPR007087	-	ZNF_C2H2	-	
ENSG00000178150	IPR007087	-	ZNF_C2H2	ZNF114	
ENSG00000178175	IPR007087	-	ZNF_C2H2	ZNF366	
ENSG00000178177	IPR007889 IPR009057 IPR011526	-	Homeodomain_like	-	
ENSG00000178187	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000178229	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000178338	IPR007086 IPR007087	-	ZNF_C2H2	ZNF354B	
ENSG00000178386	IPR007086 IPR007087	-	ZNF_C2H2	ZNF223	
ENSG00000178573	IPR004826 IPR004827 IPR008917	-	TF_bZIP	MAF	liver, prostate, tongue, kidney
ENSG00000178665	IPR007086 IPR007087	-	ZNF_C2H2	ZNF713	
ENSG00000178764	IPR001356 IPR007087 IPR009057 IPR012287	-	Homeodomain_like ZNF_C2H2	ZHX2	
ENSG00000178860	IPR011598	-	HLH_DNA_bd	MSC	
ENSG00000178919	IPR001766 IPR011991	-	TF_Fork_head	FOXE1	fetal.thyroid, thyroid
ENSG00000178928	IPR001356 IPR009057	-	Homeodomain_like	TPRX1	
ENSG00000178935	IPR007086 IPR007087	-	ZNF_C2H2	ZNF552	
ENSG00000178951	IPR007087	-	ZNF_C2H2	ZBTB7A	
ENSG00000179059	IPR007087	-	ZNF_C2H2	ZFP42	
ENSG00000179111	IPR011598	-	HLH_DNA_bd	HES7	
ENSG00000179195	IPR007086 IPR007087	-	ZNF_C2H2	ZNF664	
ENSG00000179348	IPR000679	-	ZnF_GATA	GATA2	placenta,

ENSG00000179388	IPR007087	-	ZNF_C2H2	EGR3	whole.brain, adrenal.cortex
ENSG00000179437	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000179456	IPR007087	-	ZNF_C2H2	ZNF238	whole.brain, fetal.brain
ENSG00000179528	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	LBX2	
ENSG00000179588	IPR007087	-	ZNF_C2H2	ZFPM1	
ENSG00000179772	IPR001766 IPR011991	-	TF_Fork_head	FKHL18	
ENSG00000179774	IPR011598	-	HLH_DNA_bd	ATOH7	
ENSG00000179909	IPR007086 IPR007087	-	ZNF_C2H2	ZNF154	
ENSG00000179930	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000179943	IPR007087	-	ZNF_C2H2	ZNF579	
ENSG00000179965	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000179981	IPR001356 IPR007087 IPR009057 IPR012287	-	Homeodomain_like ZNF_C2H2	TSHZ1	
ENSG00000180035	IPR007086 IPR007087	-	ZNF_C2H2	ZNF553	
ENSG00000180053	IPR001356 IPR009057	-	Homeodomain_like	-	
ENSG00000180318	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	CART1	
ENSG00000180357	IPR007087	-	ZNF_C2H2	-	
ENSG00000180479	IPR007087	-	ZNF_C2H2	ZNF571	
ENSG00000180532	IPR007087	-	ZNF_C2H2	ZSCAN4	
ENSG00000180535	IPR011598	-	HLH_DNA_bd	BHLHB8	
ENSG00000180613	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000180733	IPR004827 IPR011700	-	TF_bZIP	-	lung, fetal.lung
ENSG00000180787	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000180806	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXC9	
ENSG00000180818	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	HOXC10	human.cultured.adipocyte, kidney
ENSG00000180828	IPR011598	-	HLH_DNA_bd	BHLHB5	
ENSG00000180855	IPR007086 IPR007087	-	ZNF_C2H2	ZNF443	
ENSG00000180938	IPR007086 IPR007087	-	ZNF_C2H2	ZNF572	
ENSG00000181007	IPR007087	-	ZNF_C2H2	ZNF545	
ENSG00000181135	IPR007087	-	ZNF_C2H2	-	
ENSG00000181220	IPR007087	-	ZNF_C2H2	ZNF746	

ENSG00000181315	IPR007087	-	ZNF_C2H2	ZNF322A	
ENSG00000181444	IPR007086 IPR007087	-	ZNF_C2H2	ZNF467	whole.blood.JJV
ENSG00000181449	IPR000910	-	HMG_1/2_box	SOX2	
ENSG00000181450	IPR007086 IPR007087	-	ZNF_C2H2	ZNF678	
ENSG00000181472	IPR007087	-	ZNF_C2H2	ZBTB2	
ENSG00000181638	IPR007086 IPR007087	-	ZNF_C2H2	GLI4	
ENSG00000181666	IPR007087	-	ZNF_C2H2	HKR1	
ENSG00000181690	IPR007087	-	ZNF_C2H2	PLAG1	
ENSG00000181722	IPR007087	-	ZNF_C2H2	ZBTB20	prostate, uterus, pituitary
ENSG00000181827	IPR003150 IPR011991	-	RFX_DNA_bd	RFXDC2	
ENSG00000181894	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000181896	IPR007086 IPR007087	-	ZNF_C2H2	ZNF101	
ENSG00000181965	IPR011598	-	HLH_DNA_bd	NEUROG1	
ENSG00000182141	IPR007086 IPR007087	-	ZNF_C2H2	ZNF708	
ENSG00000182158	IPR004827 IPR008917 IPR011616 IPR011700	-	TF_bZIP	CREB3L2	placenta, smooth.muscle, adrenal.gland, fetal.thyroid, pituitary, thyroid
ENSG00000182318	IPR007086 IPR007087	-	ZNF_C2H2	HKR2	
ENSG00000182463	IPR001356 IPR007087 IPR009057 IPR012287	-	Homeodomain_like ZNF_C2H2	TSHZ2	
ENSG00000182568	IPR001356 IPR003350 IPR007108 IPR009057 IPR012287	-	Hmoeo_CUT Homeodomain_like	SATB1	thymus, whole.brain, fetal.brain
ENSG00000182742	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXB4	
ENSG00000182759	IPR004826 IPR004827 IPR008917	-	TF_bZIP	-	
ENSG00000182903	IPR007086 IPR007087	-	ZNF_C2H2	ZNF721	
ENSG00000182968	IPR000910	-	HMG_1/2_box	SOX1	
ENSG00000182983	IPR007086 IPR007087	-	ZNF_C2H2	ZNF662	
ENSG00000182986	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000183072	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	NKX2-5	heart
ENSG00000183309	IPR007086 IPR007087	-	ZNF_C2H2	ZNF623	
ENSG00000183434	IPR011991	IPR003316	Other	TFDP3	
ENSG00000183621	IPR007087	-	ZNF_C2H2	ZNF438	
ENSG00000183647	IPR007086 IPR007087	-	ZNF_C2H2	ZNF530	

ENSG00000183733	IPR011598	-	HLH_DNA_bd	-	
ENSG00000183734	IPR011598	-	HLH_DNA_bd	ASCL2	
ENSG00000183770	IPR001766 IPR011991	-	TF_Fork_head	FOXL2	
ENSG00000183779	IPR007087	-	ZNF_C2H2	ZNF703	
ENSG00000183850	IPR007086 IPR007087	-	ZNF_C2H2	ZNF254 ZNF539	
ENSG00000183900	NA	NA	Other	FOXD1	testis, smooth.muscle
ENSG00000184058	IPR008967	IPR001699	P53_like_DNA_bd	TBX1	
ENSG00000184221	IPR011598	-	HLH_DNA_bd	OLIG1	
ENSG00000184271	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU6F1	
ENSG00000184280	IPR007087	-	ZNF_C2H2	ZNF285	
ENSG00000184294	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000184302	IPR000047 IPR001356 IPR007106 IPR009057 IPR012287	-	Homeodomain_like	SIX6	pituitary
ENSG00000184481	IPR001766 IPR011991	-	TF_Fork_head	MLLT7	
ENSG00000184486	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU3F2	
ENSG00000184492	IPR001766 IPR011991	-	TF_Fork_head	FOXD4 FOXD4L1	
ENSG00000184504	IPR007086 IPR007087	-	ZNF_C2H2	ZNF75C	
ENSG00000184517	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000184540	IPR001356 IPR009057	-	Homeodomain_like	-	
ENSG00000184635	IPR007086 IPR007087	-	ZNF_C2H2	ZNF56	
ENSG00000184659	IPR001766 IPR011991	-	TF_Fork_head	FOXD4L4	
ENSG00000184677	IPR007087	-	ZNF_C2H2	ZBTB40	
ENSG00000184828	IPR007087	-	ZNF_C2H2	-	
ENSG00000184895	IPR000910	-	HMG_1/2_box	SRY	
ENSG00000184937	IPR007087	IPR000976	ZNF_C2H2	WT1	testis, ovary, uterus
ENSG00000185022	IPR004826 IPR004827 IPR008917	-	TF_bZIP	MAFF	placenta, lung
ENSG00000185122	IPR000232 IPR002341 IPR011991	-	Heat shock factor (HSF)-type	HSF1	
ENSG00000185155	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	MIXL1	
ENSG00000185177	IPR007086 IPR007087	-	ZNF_C2H2	-	

ENSG00000185219	IPR007087	-	ZNF_C2H2	ZNF445	
ENSG00000185252	IPR007087	-	ZNF_C2H2	ZNF74	
ENSG00000185278	IPR007087	-	ZNF_C2H2	ZBTB37	
ENSG00000185404	IPR000770 IPR010919	-	SAND_like	-	
ENSG00000185507	IPR001346 IPR011991	-	IRF	IRF7	lymph.node, heart, thymus, tonsil, lung, whole.blood.JJV
ENSG00000185551	IPR001628	IPR001723 IPR003068	Hrmn_rcpt_DNA_bd	NR2F2	uterus
ENSG00000185591	IPR007087	-	ZNF_C2H2	SP1	
ENSG00000185610	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000185630	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	PBX1	
ENSG00000185650	IPR000571	-	ZnF_CCCH	ZFP36L1	
ENSG00000185668	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU3F1	
ENSG00000185669	IPR007086 IPR007087	-	ZNF_C2H2	SNAI3	
ENSG00000185670	IPR007087	-	ZNF_C2H2	ZBTB3	
ENSG00000185697	IPR001005 IPR009057 IPR012287	-	Homeodomain_like	MYBL1	
ENSG00000185730	IPR007086 IPR007087	-	ZNF_C2H2	ZNF696	
ENSG00000185811	IPR007087	-	ZNF_C2H2	ZNFN1A1	
ENSG00000185947	IPR007086 IPR007087	-	ZNF_C2H2	ZNF267	whole.blood.JJV
ENSG00000185960	IPR000047 IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	SHOX	
ENSG00000186017	IPR007086 IPR007087	-	ZNF_C2H2	ZNF566	
ENSG00000186019	IPR007086 IPR007087	-	ZNF_C2H2	ZNF224	
ENSG00000186020	IPR007086 IPR007087	-	ZNF_C2H2	ZNF529	
ENSG00000186026	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000186051	IPR011598	-	HLH_DNA_bd	TAL2	
ENSG00000186103	IPR001356 IPR009057	-	Homeodomain_like	ARGFX	
ENSG00000186130	IPR007087	-	ZNF_C2H2	ZBTB6	
ENSG00000186230	IPR007087	-	ZNF_C2H2	ZNF749	
ENSG00000186272	IPR007086 IPR007087	-	ZNF_C2H2	ZNF17	
ENSG00000186300	IPR007087	-	ZNF_C2H2	ZNF555	
ENSG00000186350	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	RXRA	liver, whole.blood.JJV

ENSG00000186416	-	-	Other	NKRF	whole.brain
ENSG00000186376	IPR007086 IPR007087	-	ZNF_C2H2	ZNF75	
ENSG00000186446	IPR007086 IPR007087	-	ZNF_C2H2	ZNF501	
ENSG00000186448	IPR007086 IPR007087	-	ZNF_C2H2	ZNF197	
ENSG00000186487	IPR013681	IPR002515	Myelin_TF	-	whole.brain, fetal.brain
ENSG00000186496	IPR007086 IPR007087	-	ZNF_C2H2	ZNF396	
ENSG00000186564	IPR001766 IPR011991	-	TF_Fork_head	FOXD2	
ENSG00000186660	IPR007087	-	ZNF_C2H2	ZFP91	
ENSG00000186790	IPR001766 IPR011991	-	TF_Fork_head	FOXE3	
ENSG00000186812	IPR007086 IPR007087	-	ZNF_C2H2	ZNF397	
ENSG00000186918	IPR007087	-	ZNF_C2H2	FBXO16 ZNF395	placenta, prostate,
ENSG00000186951	IPR001628	IPR001723 IPR003074 IPR003076	Hrmn_rcpt_DNA_bd	PPARA	
ENSG00000187079	IPR009057	IPR000818	Homeodomain_like	TEAD1	
ENSG00000187098	IPR011598	-	HLH_DNA_bd	MITF	uterus
ENSG00000187140	IPR001766 IPR011991	-	TF_Fork_head	FOXD3	
ENSG00000187187	IPR007086 IPR007087	-	ZNF_C2H2	ZNF546	
ENSG00000187559	IPR001766 IPR011991	-	TF_Fork_head	FOXD4L3	
ENSG00000187607	IPR007086 IPR007087	-	ZNF_C2H2	ZNF286	
ENSG00000187626	IPR007086 IPR007087	-	ZNF_C2H2	ZNF307	testis
ENSG00000187792	IPR007086 IPR007087	-	ZNF_C2H2	ZNF70	
ENSG00000187801	IPR007086 IPR007087	-	ZNF_C2H2	ZNF643	
ENSG00000187815	IPR007086 IPR007087	-	ZNF_C2H2	ZNF642	
ENSG00000187855	IPR011598	-	HLH_DNA_bd	ASCL4	
ENSG00000187987	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000188013	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	-	
ENSG00000188033	IPR007086 IPR007087	-	ZNF_C2H2	ZNF490	
ENSG00000188095	IPR011598	-	HLH_DNA_bd	-	
ENSG00000188171	IPR007086 IPR007087	-	ZNF_C2H2	ZNF626	
ENSG00000188283	IPR007086 IPR007087	-	ZNF_C2H2	ZNF383	
ENSG00000188290	IPR011598	-	HLH_DNA_bd	HES4	
ENSG00000188295	IPR007086 IPR007087	-	ZNF_C2H2	ZNF669	
ENSG00000188321	IPR007086 IPR007087	-	ZNF_C2H2	ZNF559	
ENSG00000188620	IPR001356 IPR009057	-	Homeodomain_like	HMX3	

ENSG00000188629	IPR007086 IPR007087	-	ZNF_C2H2	ZNF177	
ENSG00000188779	IPR009061 IPR010919	-	SAND_like	-	
ENSG00000188785	IPR007087	-	ZNF_C2H2	-	
ENSG00000188786	IPR007087	-	ZNF_C2H2	MTF1	
ENSG00000188801	IPR007087	-	ZNF_C2H2	ZNF322B	
ENSG00000188841	-	IPR003316	Other	-	
ENSG00000188868	IPR007087	-	ZNF_C2H2	ZNF563	
ENSG00000188994	IPR007087	-	ZNF_C2H2	ZNF292	uterus
ENSG00000189042	IPR007086 IPR007087	-	ZNF_C2H2	ZNF567	
ENSG00000189120	IPR007087	-	ZNF_C2H2	SP6	
ENSG00000189144	IPR007086 IPR007087	-	ZNF_C2H2	ZNF573	
ENSG00000189164	IPR007086 IPR007087 IPR009057	-	Homeodomain_like ZNF_C2H2	-	
ENSG00000189180	IPR007086 IPR007087	-	ZNF_C2H2	ZNF33A	
ENSG00000189190	IPR007086 IPR007087	-	ZNF_C2H2	ZNF600	
ENSG00000189298	IPR007086 IPR007087	-	ZNF_C2H2	ZNF306	
ENSG00000189299	IPR001766 IPR011991	-	TF_Fork_head	FOXR2	
ENSG00000196092	IPR009057 IPR011991	-	Homeodomain_like	PAX5	tonsil
ENSG00000196110	IPR007086 IPR007087	-	ZNF_C2H2	ZNF699	
ENSG00000196132	IPR013681	IPR002515	Myelin_TF	MYT1	
ENSG00000196150	IPR007086 IPR007087	-	ZNF_C2H2	ZNF250	
ENSG00000196152	IPR007086 IPR007087	-	ZNF_C2H2	ZNF79	
ENSG00000196172	IPR007086 IPR007087	-	ZNF_C2H2	ZNF681	
ENSG00000196233	IPR007889 IPR011526	-	Other	-	
ENSG00000196247	IPR007086 IPR007087	-	ZNF_C2H2	ZNF588	
ENSG00000196263	IPR007086 IPR007087	-	ZNF_C2H2	ZNF471	
ENSG00000196268	IPR007086 IPR007087	-	ZNF_C2H2	ZNF493	
ENSG00000196293	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	DUX2 DUX4	
ENSG00000196323	IPR007087	-	ZNF_C2H2	BTBD15	
ENSG00000196345	IPR007086 IPR007087	-	ZNF_C2H2	ZNF167	
ENSG00000196350	IPR007087	-	ZNF_C2H2	-	
ENSG00000196357	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000196378	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000196387	IPR007086 IPR007087	-	ZNF_C2H2	ZNF140	
ENSG00000196409	IPR007087	-	ZNF_C2H2	ZNF658B	

ENSG00000196417	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000196418	IPR007086 IPR007087	-	ZNF_C2H2	ZNF124	
ENSG00000196428	-	IPR000580	Other	TSC22D2	testis, placenta, uterus
ENSG00000196437	IPR007086 IPR007087	-	ZNF_C2H2	ZNF569	
ENSG00000196442	IPR007087	-	ZNF_C2H2	ZNF432	prostate, thyroid
ENSG00000196453	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000196458	IPR007086 IPR007087	-	ZNF_C2H2	ZNF605	
ENSG00000196460	IPR003150 IPR011991	-	RFX_DNA_bd	-	
ENSG00000196482	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	ESRRG	placenta
ENSG00000196484	IPR007086 IPR007087	-	ZNF_C2H2	ZNF564	
ENSG00000196628	IPR009057 IPR011598	-	Homeodomain_like	TCF4	fetal.brain
ENSG00000196646	IPR007086 IPR007087	-	ZNF_C2H2	ZNF136	
ENSG00000196652	IPR007087	-	ZNF_C2H2	ZFP95	
ENSG00000196653	IPR007086 IPR007087	-	ZNF_C2H2	ZNF502	
ENSG00000196670	IPR007087	-	ZNF_C2H2	-	
ENSG00000196693	IPR007086 IPR007087	-	ZNF_C2H2	ZNF33B	
ENSG00000196705	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000196724	IPR007086 IPR007087	-	ZNF_C2H2	ZNF418	
ENSG00000196757	IPR007086 IPR007087	-	ZNF_C2H2	ZNF700	
ENSG00000196767	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU3F4	
ENSG00000196793	IPR007087	-	ZNF_C2H2	ZNF239	
ENSG00000196812	IPR007086 IPR007087	-	ZNF_C2H2	ZNF435	
ENSG00000196826	IPR007087	-	ZNF_C2H2	ZNF709	
ENSG00000196867	IPR007086 IPR007087	-	ZNF_C2H2	ZFP28	
ENSG00000196922	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000196931	IPR007087	-	ZNF_C2H2	-	
ENSG00000196946	IPR007086 IPR007087	-	ZNF_C2H2	ZNF705A	
ENSG00000196967	IPR007086 IPR007087	-	ZNF_C2H2	ZNF585A ZNF585B	
ENSG00000197008	IPR007086 IPR007087	-	ZNF_C2H2	ZNF138	
ENSG00000197016	IPR007086 IPR007087	-	ZNF_C2H2	ZNF470	
ENSG00000197020	IPR007086 IPR007087	-	ZNF_C2H2	ZNF100	
ENSG00000197024	IPR007087	-	ZNF_C2H2	ZNF398	

ENSG00000197025	IPR007086 IPR007087	-	ZNF_C2H2	ZNF542	
ENSG00000197037	IPR007086 IPR007087	-	ZNF_C2H2	ZNF498	
ENSG00000197044	IPR007086 IPR007087	-	ZNF_C2H2	ZNF441	
ENSG00000197050	IPR007086 IPR007087	-	ZNF_C2H2	ZNF420	
ENSG00000197062	IPR007086 IPR007087	-	ZNF_C2H2	ZNF187	
ENSG00000197063	IPR004826 IPR004827 IPR008917	-	TF_bZIP	MAFG	
ENSG00000197123	IPR007086 IPR007087	-	ZNF_C2H2	ZNF679	
ENSG00000197124	IPR007086 IPR007087	-	ZNF_C2H2	ZNF682	
ENSG00000197134	IPR007086 IPR007087	-	ZNF_C2H2	ZNF429	
ENSG00000197162	IPR007087	-	ZNF_C2H2	ZNF688	
ENSG00000197279	IPR007087	-	ZNF_C2H2	ZNF165	
ENSG00000197283	IPR007087	-	ZNF_C2H2	ZBTB9	
ENSG00000197294	IPR007087	-	ZNF_C2H2	ZNF72	
ENSG00000197337	-	IPR003316	Other	-	bone.marrow, fetal.liver
ENSG00000197343	IPR007087	-	ZNF_C2H2	ZNF655	
ENSG00000197363	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000197372	IPR007086 IPR007087	-	ZNF_C2H2	ZNF675	
ENSG00000197472	IPR007086 IPR007087	-	ZNF_C2H2	ZNF695	
ENSG00000197483	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000197497	IPR007086 IPR007087	-	ZNF_C2H2	ZNF665	
ENSG00000197550	IPR007087	-	ZNF_C2H2	ZNF72	
ENSG00000197566	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000197576	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXA4	
ENSG00000197587	IPR001356 IPR007104 IPR009057 IPR012287	-	Homeodomain_like	DMBX1	
ENSG00000197619	IPR007086 IPR007087	-	ZNF_C2H2	ZNF615	
ENSG00000197647	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000197657	IPR007086 IPR007087	-	ZNF_C2H2	ZNF323	
ENSG00000197701	IPR007086 IPR007087	-	ZNF_C2H2	ZNF595	
ENSG00000197714	IPR007086 IPR007087	-	ZNF_C2H2	ZNF272	
ENSG00000197757	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXC6	spinal.cord, human.cultured.adipocyte, adrenal.gland, uterus, kidney

ENSG00000197779	IPR007087	-	ZNF_C2H2	-	
ENSG00000197800	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000197808	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000197841	IPR003656 IPR007086 IPR007087	-	ZNF_C2H2 Znf_BED_prd	ZNF181	fetal.brain
ENSG00000197857	IPR007086 IPR007087	-	ZNF_C2H2	ZNF44	
ENSG00000197905	IPR009057	IPR000818	Homeodomain_like	TEAD4	thyroid
ENSG00000197921	IPR011598	-	HLH_DNA_bd	-	
ENSG00000197928	IPR007086 IPR007087	-	ZNF_C2H2	ZNF677	
ENSG00000197933	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000197935	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000197937	IPR007086 IPR007087	-	ZNF_C2H2	ZNF347	
ENSG00000197951	IPR007086 IPR007087	-	ZNF_C2H2	ZNF71	
ENSG00000197961	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000197990	IPR007087	-	ZNF_C2H2	-	
ENSG00000198026	IPR000637 IPR007087	-	AT_hook_DNA_bd ZNF_C2H2	ZNF335	
ENSG00000198028	IPR007087	-	ZNF_C2H2	ZNF560	
ENSG00000198039	IPR007086 IPR007087	-	ZNF_C2H2	ZNF273	
ENSG00000198040	IPR007087	-	ZNF_C2H2	ZNF84	
ENSG00000198046	IPR007086 IPR007087	-	ZNF_C2H2	ZNF667	
ENSG00000198081	IPR007087	-	ZNF_C2H2	ZFP161	
ENSG00000198093	IPR007086 IPR007087	-	ZNF_C2H2	ZNF649	
ENSG00000198105	IPR007086 IPR007087	-	ZNF_C2H2	ZNF248	fetal.brain
ENSG00000198131	IPR007086 IPR007087	-	ZNF_C2H2	ZNF544	
ENSG00000198160	IPR001005 IPR009057	-	Homeodomain_like	MIER1	
ENSG00000198169	IPR007087	-	ZNF_C2H2	ZNF251	
ENSG00000198176	IPR011991	IPR003316	Other	TFDP1	bone.marrow, fetal.liver
ENSG00000198182	IPR007086 IPR007087	-	ZNF_C2H2	ZNF607	
ENSG00000198185	IPR007086 IPR007087	-	ZNF_C2H2	ZNF334	
ENSG00000198205	IPR007087	-	ZNF_C2H2	ZXDA	
ENSG00000198298	IPR007086 IPR007087	-	ZNF_C2H2	ZNF485	
ENSG00000198300	IPR007087	-	ZNF_C2H2	PEG3 ZIM2	ovary, placenta, whole.brain, adrenal.gland, pituitary, fetal.liver,
ENSG00000198304	IPR007086 IPR007087	-	ZNF_C2H2	ZNF528	
ENSG00000198315	IPR007086 IPR007087	-	ZNF_C2H2	ZNF192	

ENSG00000198342	IPR007086 IPR007087	-	ZNF_C2H2	ZNF442	
ENSG00000198353	IPR001356 IPR001827 IPR009057 IPR012287	-	Homeodomain_like	HOXC4	
ENSG00000198393	IPR007086 IPR007087	-	ZNF_C2H2	ZNF26	
ENSG00000198416	IPR007087	-	ZNF_C2H2	ZNF658B	
ENSG00000198440	IPR007086 IPR007087	-	ZNF_C2H2	ZNF583	
ENSG00000198453	IPR007086 IPR007087	-	ZNF_C2H2	ZNF568	
ENSG00000198455	IPR007087	-	ZNF_C2H2	ZXDB	
ENSG00000198464	IPR007086 IPR007087	-	ZNF_C2H2	ZNF480	
ENSG00000198466	IPR007086 IPR007087	-	ZNF_C2H2	ZNF587	
ENSG00000198477	IPR007087	-	ZNF_C2H2	-	testis
ENSG00000198517	IPR004826 IPR004827 IPR008917	-	TF_bZIP	MAFK	
ENSG00000198521	IPR007086 IPR007087	-	ZNF_C2H2	ZNF43	
ENSG00000198522	IPR007087	-	ZNF_C2H2	ZNF512	
ENSG00000198538	IPR007086 IPR007087	-	ZNF_C2H2	ZNF28	
ENSG00000198546	IPR007087	-	ZNF_C2H2	ZNF511	
ENSG00000198551	IPR007086 IPR007087	-	ZNF_C2H2	ZNF627	
ENSG00000198566	IPR007087	-	ZNF_C2H2	ZNF658	
ENSG00000198584	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000198597	IPR007087	-	ZNF_C2H2	ZNF536	spinal.cord, whole.brain, appendix
ENSG00000198694	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	-	
ENSG00000198740	IPR007087	-	ZNF_C2H2	ZNF652	prostate
ENSG00000198767	IPR007087	-	ZNF_C2H2	YY2	
ENSG00000198783	IPR007087	-	ZNF_C2H2	CCDC16	
ENSG00000198795	IPR007087	-	ZNF_C2H2	ZNF521	
ENSG00000198807	IPR009057 IPR011991	-	Homeodomain_like	PAX9	
ENSG00000198815	IPR001766 IPR011991	-	TF_Fork_head	FOXJ3	
ENSG00000198816	IPR007086 IPR007087	-	ZNF_C2H2	ZNF358	
ENSG00000198839	IPR007087	-	ZNF_C2H2	ZNF277	whole.blood.JJV
ENSG00000198911	IPR011598	-	HLH_DNA_bd	SREBF2	
ENSG00000198914	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU3F3	

ENSG00000198939	IPR007086 IPR007087	-	ZNF_C2H2	ZFP2	
ENSG00000198963	IPR001628	IPR001723 IPR003079	Hrmn_rcpt_DNA_bd	RORB	
ENSG00000203766	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	DUX2	
ENSG00000203767	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	DUX2	
ENSG00000203768	IPR000047 IPR001356 IPR009057	-	Homeodomain_like	DUX2	
ENSG00000203769	IPR000047 IPR001356 IPR009057	-	Homeodomain_like	DUX2	
ENSG00000203770	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	DUX2	
ENSG00000203771	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	DUX2	
ENSG00000203883	IPR000910	-	HMG_1/2_box	SOX18	
ENSG00000204053	IPR011598	IPR002418	HLH_DNA_bd	MYCL2	
ENSG00000204103	IPR004826 IPR004827 IPR008917	-	TF_bZIP	MAFB	lymph.node, liver, placenta, human.cultured.adipocyte, whole.brain, lung, whole.blood.JJV,
ENSG00000204210	IPR007087	-	ZNF_C2H2	ZBTB22	
ENSG00000204231	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	RXRB	
ENSG00000204304	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	PBX2	
ENSG00000204335	IPR007087	-	ZNF_C2H2	-	
ENSG00000204366	IPR007087	-	ZNF_C2H2	ZBTB12	
ENSG00000204519	IPR007086 IPR007087	-	ZNF_C2H2	ZNF551	
ENSG00000204527	IPR000047 IPR001356 IPR009057	-	Homeodomain_like	DUXA	
ENSG00000204531	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	POU5F1 POU5F1P1	
ENSG00000204595	IPR001356 IPR009057	-	Homeodomain_like	DPRX DPRXP4	
ENSG00000204604	IPR007086 IPR007087	-	ZNF_C2H2	-	placenta, prostate
ENSG00000204611	IPR007086 IPR007087	-	ZNF_C2H2	-	
ENSG00000204644	IPR007087	-	ZNF_C2H2	ZFP57	

ENSG00000204789	IPR007086 IPR007087	-	ZNF_C2H2	ZNF204	
ENSG00000204793	IPR001766 IPR011991	-	TF_Fork_head	-	
ENSG00000204828	IPR001766 IPR011991	-	TF_Fork_head	FOXD4L4	
ENSG00000204859	IPR007087	-	ZNF_C2H2	HKR3	liver, lung
ENSG00000204920	IPR007086 IPR007087	-	ZNF_C2H2	ZNF155	
ENSG00000204947	IPR007087	-	ZNF_C2H2	-	
ENSG00000205096	IPR000047 IPR001356 IPR009057 IPR012287	-	Homeodomain_like	DUX2	
ENSG00000205189	IPR007087	-	ZNF_C2H2	ZBTB10	
ENSG00000205245	IPR007086 IPR007087	-	ZNF_C2H2	ZNF67	
ENSG00000205250	-	IPR003316	Other	-	
ENSG00000205922	IPR001356 IPR007108 IPR009057 IPR012287	-	Homeodomain_like	ONECUT3	
ENSG00000205927	IPR011598	-	HLH_DNA_bd	OLIG2	spinal.cord, whole.brain
ENSG00000206083	IPR007087	-	ZNF_C2H2	ZNF409	
ENSG00000206207	IPR007087	-	ZNF_C2H2	-	
ENSG00000206218	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	RXRB	
ENSG00000206247	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	PBX2	whole.blood.JJV
ENSG00000206280	IPR007087	-	ZNF_C2H2	-	
ENSG00000206289	IPR001628	IPR000003 IPR001723	Hrmn_rcpt_DNA_bd	RXRB	
ENSG00000206315	IPR001356 IPR009057 IPR012287	-	Homeodomain_like	PBX2	whole.blood.JJV
ENSG00000206366	IPR007087	-	ZNF_C2H2	-	
ENSG00000206454	IPR000327 IPR001356 IPR007103 IPR009057 IPR010982 IPR012287	-	Homeodomain_like POU	-	

Appendix B

Appendix C

List of publications

Work from the following publications have been presented in this dissertation (see Appendix E for reprints):

* indicates joint first authors.

Vaquerizas JM, Kummerfeld SK, Dopazo J, Teichmann SA & Luscombe NM. 2007. A functional census of human transcription factors. In preparation.

Kind J, Vaquerizas JM, Luscombe NM, Bertone P & Akhtar A. 2007. High resolution tiling arrays reveal MOF histone H4 lysine 16 specific histone acetyltransferase as a global regulator of gene expression in Drosophila. In preparation.

Vaquerizas JM, Conde L, Yankilevich P, Cabezon A, Minguez P, Diaz-Uriarte R, Al-Shahrour F, Herrero J & Dopazo J. 2005. Gepas an experiment-oriented pipeline for the analysis of microarray gene expression data. *Nucleic Acids Res.*, **33**:W616-W620.

Vaquerizas JM, Dopazo J & Díaz-Uriarte R. 2004. DNMAD: web-based diagnosis and normalization for microarray data. *Bioinformatics*, **20**:3656-3658.

* Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M & Dopazo J. 2004. PupaSNP finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.*, **32**: W242-W248.

The work presented in Vaquerizas et al., 2005, Vaquerizas et al., 2004, and Conde et al., 2004, was performed between June 2003 and February 2005 at the Bioinformatics Unit of the Spanish National Cancer Centre, Madrid, under the supervision of Dr. Joaquín Dopazo, and has been included here with the appropriate permission.

Other publications during the PhD studentship:

Rico D, Vaquerizas JM, Dopazo H & Bosca L. 2007. Identification of conserved domains in the promoter regions of nitric oxide synthase 2: implications for the species-specific transcription and evolutionary differences. *BMC Genomics*, **8**:271.

Goni JR, Vaquerizas JM, Dopazo J & Orozco M. 2006. Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. *BMC Genomics*, **7**:63.

Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J & Dopazo J. 2006. PupaSuite: finding functional SNPs for large-scale genotyping purposes. *Nucleic Acids Res.*, **34**:W621-W625.

Al-Shahrour F, Minguez P, Tárraga J, Montaner D, Alloza E, Vaquerizas JM, Conde L, Blaschke C, Vera J & Dopazo J. 2006. Babelomics: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res.*, **34**:W472-W476.

Montaner D, Tárraga J, Huerta-Cepas J, Burguet J, Vaquerizas JM, Conde L, Minguez P, Vera J, Mukherjee S, Valls J, Pujana M, Alloza E, Herrero J, Al-Shahrour F & Dopazo J. 2006. Next station in microarray data analysis: GEPAS. *Nucleic Acids Res.*, **34**:W486-W491.

Conde L, Vaquerizas JM, Ferrer-Costa C, Orozco M & Dopazo J. 2005. PupasView: a visual tool for selecting suitable SNPs, with putative pathologic effect in genes, for genotyping purposes. *Nucleic Acids Res.*, **33**:W501-W505.

Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L & Dopazo J. 2005. Babelomics: a suite of web-tools for functional annotation and analysis of group of genes in high-throughput experiments. *Nucleic Acids Res.*, **33**:W460-W464.

Santoyo J, Vaquerizas JM & Dopazo J. 2004. Highly specific and accurate selection of siRNAs for high-throughput functional assays. *Bioinformatics*, **21**:1376-1382.

Herrero J, Vaquerizas JM, Al-Shahrour F, Conde L, Mateos Á, Santoyo J, Díaz-Uriarte R & Dopazo J. 2004. New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res.*, **32**:W485-W491.

Appendix C

Appendix D

Reprints

Vaquerizas JM, Conde L, Yankilevich P, Cabezon A, Minguez P, Diaz-Uriarte R, Al-Shahrour F, Herrero J & Dopazo J. 2005. Gepas an experiment-oriented pipeline for the analysis of microarray gene expression data. *Nucleic Acids Res.*, **33**:W616-W620.

Vaquerizas JM, Dopazo J & Díaz-Uriarte R. 2004. DNMAD: web-based diagnosis and normalization for microarray data. *Bioinformatics*, **20**:3656-3658.

Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M & Dopazo J. 2004. PupaSNP finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.*, **32**:W242-W248.

Appendix D

W616–W620 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki500

GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data

Juan M. Vaquerizas[1], Lucía Conde[1], Patricio Yankilevich[1], Amaya Cabezón[1], Pablo Minguez[1], Ramón Díaz-Uriarte[1], Fátima Al-Shahrour[1], Javier Herrero[1,2] and Joaquín Dopazo[1,3,*]

[1]Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain, [2]Ensembl Team, EMBL-EBI, Hinxton, Cambridge, UK and [3]Functional Genomics Node, INB, Centro de Investigación Príncipe Felipe, Autopista del Saler 16, 46013 Valencia, Spain

Received February 14, 2005; Revised April 9, 2005; Accepted May 3, 2005

ABSTRACT

The Gene Expression Profile Analysis Suite, GEPAS, has been running for more than three years. With >76 000 experiments analysed during the last year and a daily average of almost 300 analyses, GEPAS can be considered a well-established and widely used platform for gene expression microarray data analysis. GEPAS is oriented to the analysis of whole series of experiments. Its design and development have been driven by the demands of the biomedical community, probably the most active collective in the field of microarray users. Although clustering methods have obviously been implemented in GEPAS, our interest has focused more on methods for finding genes differentially expressed among distinct classes of experiments or correlated to diverse clinical outcomes, as well as on building predictors. There is also a great interest in CGH-arrays which fostered the development of the corresponding tool in GEPAS: InSilicoCGH. Much effort has been invested in GEPAS for developing and implementing efficient methods for functional annotation of experiments in the proper statistical framework. Thus, the popular FatiGO has expanded to a suite of programs for functional annotation of experiments, including information on transcription factor binding sites, chromosomal location and tissues. The web-based pipeline for microarray gene expression data, GEPAS, is available at http://www.gepas.org.

INTRODUCTION

GEPAS, which stands for Gene Expression Profile Analysis Suite, is a web tool designed and oriented to the analysis of DNA microarray gene expression experiments. The emphasis in the development of new tools for GEPAS has been driven by the requirements of data analysis in the most active fields using microarray technologies, which are, without doubt, biomedical applications [e.g. (1–4)]. As a consequence, much stress has been put on the implementation of proper methods for gene selection, predictors, CGH-arrays and functional annotation of experiments. More classical data analysis approaches, such as clustering, have also been incorporated into GEPAS, as well as different options for data preprocessing.

GEPAS has been conceived as an integrated web-based pipeline for the analysis of gene expression patterns where different methods can be used within an integrated interface that provides a user-friendly environment to end users. The way in which the methods are connected has been designed to guide the user by suggesting all the available possibilities to continue with the analysis and to prevent possible inappropriate uses of the tools.

GEPAS, which was originally the backbone of the pipeline of microarray data analysis of the CNIO, was made public three years ago and first published in 2003 (5,6). In the years since, GEPAS has become a *de facto* standard for many researchers and its use has undergone a spectacular growth. In terms of the scope of analysis, GEPAS is the most complete web-based resource that can be found nowadays.

Our aim is to keep GEPAS 'living' by the continuous addition of new algorithms. Here we report the new modules, some trends observed in its use and some novelties.

*To whom correspondence should be addressed. Tel: +34 96 328 96 80; Fax: +34 96 328 97 01; Email: jdopazo@ochoa.fib.es
Present address:
Juan M. Vaquerizas, Lucía Conde, Pablo Minguez, Fátima Al-Shahrour and Joaquín Dopazo, Bioinformatics Department, Centro de Investigación Príncipe Felipe, Autopista del Saler 16, 46013, Valencia, Spain

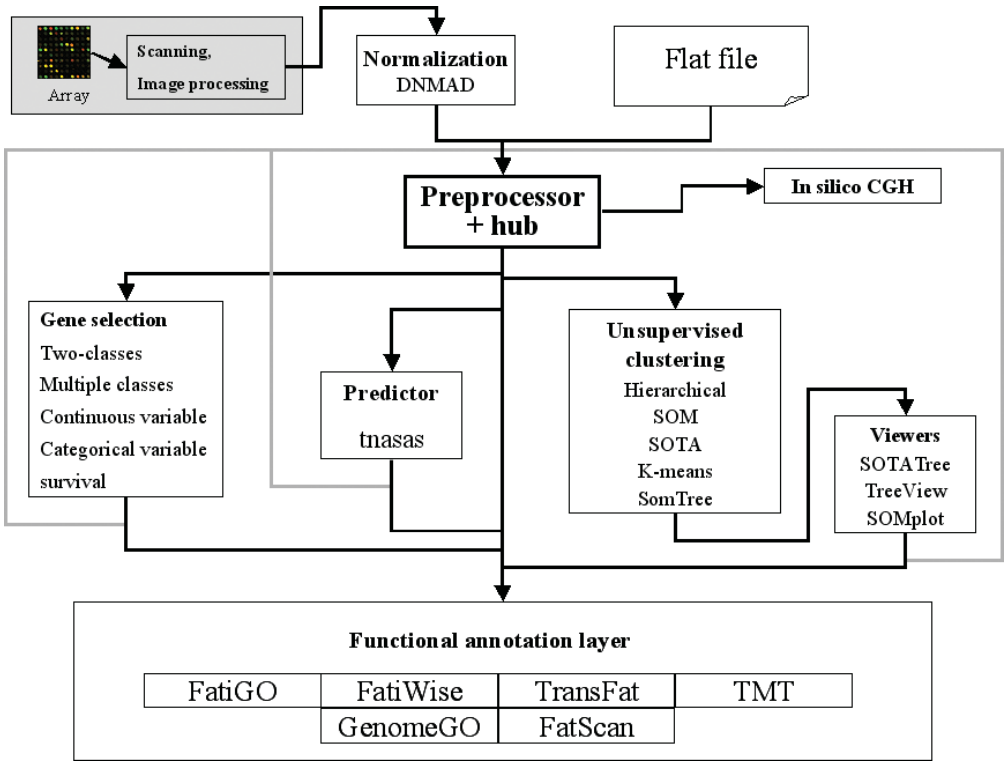

Figure 1. The GEPAS pipeline. The figure summarizes the most important features of the GEPAS pipeline. Black arrows show the flow of information from the raw data to the three main types of analysis: CGH-array, unsupervised clustering and supervised analysis (gene selection or predictors). Functional annotation is possible from the latter two options. Grey arrows represent the possibility to re-analyse parts of the experiments.

SCOPE OF GEPAS

As previously mentioned, GEPAS is experiment oriented. This means that facilities for data manipulation such as rows and columns management are deliberately absent in its design. With the exception of the module DNMAD, which can take as input Genepix (Axon instruments) GPR files from a scanner (see below), GEPAS accepts as input data already preselected (usually coming from a database) in a very simple format: a tab-delimited text file containing genes in rows and experiments in columns (except the first column, which contains the identifiers for the genes).

Several preprocessing facilities are provided. These are normalization along with different kinds of data transformation such as missing value imputation, filtering of 'flat' patterns and extraction of genes based on functional properties.

GEPAS permits two main types of experimental designs: those oriented towards class discovery, for which different clustering methods are available, and those related to supervised questions, which include mainly gene selection and building predictors. GEPAS includes two tools for dealing with both of these problems.

In addition, there is a great interest now in tools that allow CGH-arrays to be handled. GEPAS includes a module for mapping either genomic or mRNA hybridizations over the corresponding chromosomal locations, with different facilities for data visualization.

Finally, GEPAS provides a module for functional annotation of experiments that includes the popular FatiGO (8), as well as a variety of new tools.

GEPAS AT A GLANCE

GEPAS includes a number of interconnected tools implemented as individual modules that can be used either independently or within the pipeline (Figure 1). Since the previous version (6), GEPAS has undergone a number of technical improvements which have not had much impact on its external aspect but have notably changed its performance. Internal links among modules have been improved and redesigned in order to avoid wrong pathways in the pipeline. Some CPU-intensive modules have been moved to dedicated computers (in particular DNMAD, Pomelo and Tnasas). The structure of GEPAS is as follows.

Preprocessing

DNMAD (9) is for normalization using print-tip loess (10,11) (http://www.bioconductor.org), with different possibilities. Some additional options have been included in this new version: the possibility of using a spot's flags, optional use of background subtraction and the possibility of using global loess (instead of print-tip). We have also included a better management of flagged dots, new diagnostic plots (the density plots for either raw or background-corrected red and green channels) and automatic dye-swap. DNMAD can take as input Genepix (Axon instruments) GPR files.

Preprocessor (12) performs some preprocessing of the data (log-transformations, standardizations, imputation of missing values, etc.). Data can also be filtered on the basis of their functional labels [GO terms (13)] using the Knowledge Filtering module (6).

IDconverter, a new module, maps lists of accession numbers and identifiers among different clone, gene or protein standards. IDconverter includes distinct levels of information such as gene level (gene HUGO name, Ensembl gene, Unigene cluster, LocusLink, RefSeq, gene location, gene description), clone level (Affymetrix, GenBank accession number, IMAGE Clone ID) and protein level (SwissProt, TrEMBL, now UNIPROT). Chromosomal locations are obtained from Ensembl.

Analysis

Unsupervised clustering includes different methods such as aggregative clustering (14), SOTA (15,16), SOM (17), *K*-means (18) (which is a new addition in this version of GEPAS) and SOM-Tree (19).

Supervised analysis includes

(i) *Gene selection*. Analysis of genes differentially expressed between two or more classes, related to a continuous experimental factor (e.g. the concentration of a metabolite) or to survival is performed by the module Pomelo (6). Different methods for multiple testing adjustment are included (20–22).

(ii) *Predictors*. The module Tnasas (for 'This is not a substitute for a statistician') implements a simple, although effective, way of building class predictors from microarray data. The error rate is computed taking into account the effect of gene selection and is not biased downwards by the 'selection bias' problem so common in many microarray studies [e.g. (23,24)].

For the *analysis of CGH-arrays*, given the growing interest in microarray-based CGH (array CGH) (25), we have expanded the capabilities of the InSilicoCGH tool, which allows the mapping of the results of microarray hybridizations onto chromosome coordinates. The InSilicoCGH module has been designed for the simultaneous analysis of genomic and mRNA hybridizations on the same expression array. It can also deal with BAC-arrays. We have added a new option: the zoom. This magnifies the view of the desired chromosomal location in order to facilitate detection of the precise position of chromosomal gains and losses; in general, it allows hybridization values at gene level to be viewed in more detail. Figure 2 is a screenshot of the zoom tool.

Functional analysis of experiments

Functional annotation of microarray experiments is an important aspect of analysis that very few packages incorporate. Several modules for functional annotation of microarray

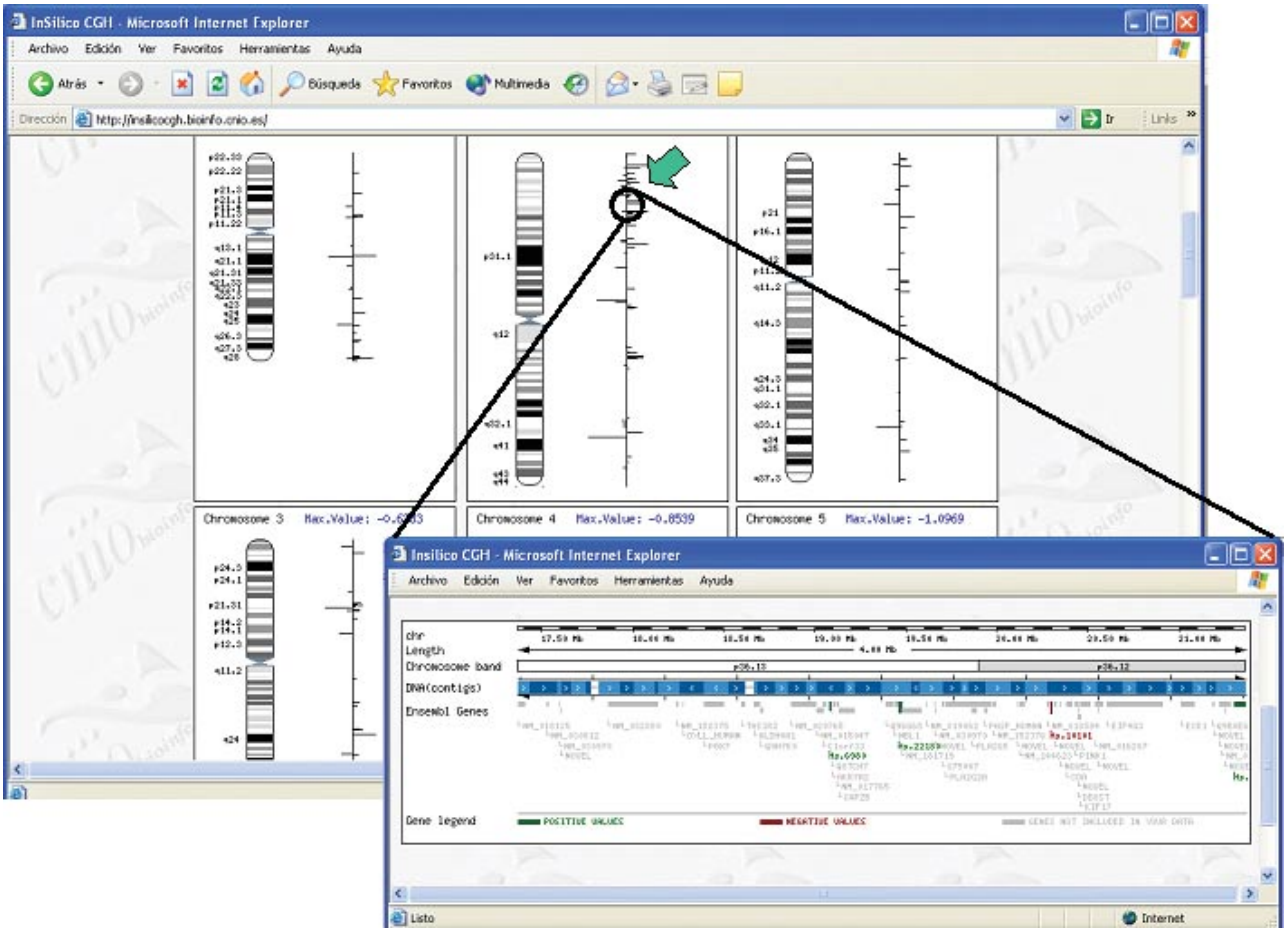

Figure 2. The zoom tool of InSilicoCGH in action. Clicking on the desired chromosomal region produces a pop-up window with a zoom facility. The user can freely move around the point chosen and can easily visualize in detail the hybridization values. Borders of deleted or amplified regions can be precisely defined in this way.

experiments are available. These programs, along with similar ones, are discussed in an accompanying paper by our group (7)

(i) *FatiGO* (8) allows significant asymmetrical distributions of GO terms between groups of genes to be found.
(ii) *FatiWise* (6) does the same with InterPro motifs (26), KEGG (27) pathways and SwissProt keywords, when available.
(iii) *TransFAT* performs the same operations for putative transcription factor binding sites in the promoter regions of genes as predicted by the program Match (28), from the Transfac® database (29).
(iv) *TMT*, the Tissues Mining Tool, is a web application to extract significant information related to the differential expression of two sets of genes in tissues.
(v) *FatiScan* allows the detection of modest but coordinate changes in gene expression values by applying the FatiGO algorithm to lists of genes ordered according to their differences in expression.

All these tools, in addition of being connected to GEPAS (because of its obvious usefulness for the analysis of microarray data), are grouped as an independent resource called Babelomics (7). Babelomics has, at its general purpose, the facilitation of functional annotation in any type of high-throughput experiments (proteomics, interactomics, massive sequencing, etc.).

In terms of its internal architecture, GEPAS is a collection of programs mainly written in C++, although some were written in other programming languages such as R [DNMAD (9)] or PERL [Preprocessor (12)]. These modules are interconnected by PERL wrappers.

A PIPELINE OF MICROARRAY DATA ANALYSIS TOOLS

The efficiency of a modular package such as GEPAS lies largely in its degree of integration of the different data analysis tools. Users can move through a complete pipeline of data analysis in a transparent way, without needing to perform any reformatting operation. In addition, a properly designed workflow can help to prevent possible wrong operations in microarray data analysis owing to misconceptions. Figure 1 illustrates the structure of the GEPAS pipeline. Raw data can be loaded and normalized. Several data transformation options are available through the Preprocessor tool. Depending on the particular problem addressed, data can be directed to any of the three main types of analysis: CGH-array, unsupervised clustering and supervised analysis (gene selection or predictors). A functional annotation is possible from the last two options. GEPAS has been designed in a way that prevents possible misuses of the methods implemented in the package.

TRAINING PROGRAMME AND GEPAS

In addition to the tools, a collection of on-line tutorials that can be used to learn the use of the tools or as a part of a course is available on the GEPAS web page. The structure of the tutorials includes some theory, a guided example and several examples based on publicly available datasets. There are tutorials for (i) normalization using DNMAD, (ii) data preprocessing using the Preprocessor tool, (iii) data clustering using the different algorithms available (UPGMA, SOM, SOTA), (iv) selection of differentially expressed genes using the Pomelo tool and (v) functional annotation using FatiGO.

The tutorials are currently used on different courses, such as a masters in bioinformatics (Spain) and the international FCUL-IGC Post-Graduate Programme in Bioinformatics (http://bioinformatics.fc.ul.pt/).

CONSOLIDATION OF GEPAS AS A WIDELY USED PACKAGE

Our records indicate that, since March 2004, GEPAS has been used to analyse >76 000 experiments, with a daily average of almost 300 uses (statistics can be checked at http://bioinfo.cnio.es/docus/webalizer/ on the different pages for GEPAS, and the particular pages for Pomelo, Tnasas, DNMAD and FatiGO, which are independently monitored). Compared with last year's records (35 000 experiments per year with a daily average of 130) (6), there has been a clear increase in the use of the tool. The distribution of users has also changed. Whereas one year ago it was used more by Spanish researchers (25%), followed by US (.edu and .net domains) (15%), French (10%), UK (5%) and other users (Japanese, German, Dutch, etc.) (6) the profile of users during this last year has changed to 23% US (.edu and .net), 9% French, 6% Spanish, 5% UK and others. These figures suggest that GEPAS seems to be becoming more popular among US-based researchers. Obviously the usage in all countries has increased, since the remainder of the percentages appear to maintain the same level while the absolute number of uses has increased 2-fold.

CONCLUSIONS

Despite the availability of many programs and packages for microarray data analysis, there are still many aspects of the analysis with poor or incomplete coverage. There are a number of options for analysing DNA microarray data (see e.g. http://www.dnamicroarrays.info/software.html). Most of the software available for microarray data analysis focuses on unsupervised cluster methods, which, in many cases, are used for inadequate purposes (23). There are also different initiatives such as BASE (30), Bioconductor (31) and BRB tools (http://linus.nci.nih.gov/BRB-ArrayTools.html), but these are in some cases dependent on a particular computer operating system and usually require from the user previous training in statistics. GEPAS can be considered the most complete web-based resource that can be found nowadays.

Since the first release (5,6), GEPAS has avoided the temptation to become a list of as many methods as possible and evolved really to cope with new challenges that have emerged in the field of microarray data analysis. Much work has been invested in the implementation of a useful workflow. GEPAS provides the user with an integrated environment in which modules can be found for different types of analysis that respond to real analysis demands. Modules are connected in such a way as to avoid improper use of the tools.

From a technical point of view, GEPAS has been designed with the intention of taking full advantage of the properties of

the web: connectivity, cross-platform compatibility and remote usage. The modular architecture allows the addition of new tools and facilitates the connectivity of GEPAS from and to other web-based tools.

With >76 000 experiments analysed during the last year and a daily average of almost 300 uses, GEPAS can be considered a consolidated tool in the field of microarray data analysis.

ACKNOWLEDGEMENTS

J.M.V. is supported by the Formacion del Personal Investigador fellowship program from the Ministerio de Educacion y Ciencia. L.C. is supported by a fellowship from the Fondo de Investigacion Sanitaria (grant PI020919). P.M. is supported by a grant from Genoma España and Canada Genome. This work is partly supported by grants from Fundación Ramón Areces, Fundació La Caixa, Fundación BBVA and RTICCC from the FIS. The Functional Genomics node of the INB is funded by Fundación Genoma España. IBM awarded us a SUR grant. Funding to pay the Open Access publication charges for this article was provided by Fundacion Genoma España.

Conflict of interest statement. None declared.

REFERENCES

1. van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
2. Beer,D.G., Kardia,S.L., Huang,C.C., Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G. *et al*. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med*., **8**, 816–824.
3. Albertson,D.G. and Pinkel,D. (2003) Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet*., **12**, 145R–152R.
4. The Tumor Analysis Best Practices Group (2004) Expression profiling—best practices for data generation and interpretation in clinical trials. *Nature Rev. Genet*., **5**, 229–237.
5. Herrero,J., Al-Shahrour,F., Díaz-Uriarte,R., Mateos,A., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS, a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res*., **31**, 3461–3467.
6. Herrero,J., Vaquerizas,J.M., Al-Shahrour,F., Conde,L., Mateos,Á., Santoyo,J., Díaz-Uriarte,R. and Dopazo,J. (2004) New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res*., **32**, W485–W491.
7. Al-Shahrour,F., Minguez,P., Vaquerizas,J.M., Conde,L. and Dopazo,J. (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucl. Acids Res*., **33**, W460–W464.
8. Al-Shahrour,F., Díaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms to groups of genes. *Bioinformatics*, **20**, 578–580.
9. Vaquerizas,J.M., Dopazo,J. and Díaz-Uriarte,R. (2004) DNMAD: web-based diagnosis and normalization for MicroArray data. *Bioinformatics*, **20**, 3656–3658.
10. Smyth,G.K., Yang,Y.H. and Speed,T.P. (2003) Statistical issues in microarray data analysis. In Brownstein,M.J., Khodursky,A.B. and Totowa,N.J. (eds), *Functional Genomics: Methods and Protocols, Methods in Molecular Biology*. Humana Press, Vol. 224, pp. 111–136.
11. Dudoit,S. and Yang,H.Y. (2003) Documentation of the Bioconductor's marrayPlots package.
12. Herrero,J., Díaz-Uriarte,R. and Dopazo,J. (2003) Gene expression data preprocessing. *Bioinformatics*, **19**, 655–656.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nature Genet*., **25**, 25–29.
14. Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy*. Freeman, San Francisco.
15. Dopazo,J. and Carazo,J.M. (1997) Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol*., **44**, 226–233.
16. Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.
17. Kohonen,T. (1997) *Self-Organizing Maps*. Springer-Verlag, Berlin.
18. Hartigan,J.A. and Wong,M.A. (1979) A *k*-means clustering algorithm. *Appl. Stat*., **28**, 100–108.
19. Herrero,J. and Dopazo,J. (2002) Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. *J. Proteome Res*., **1**, 467–470.
20. Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment*. John Wiley & Sons, New York.
21. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
22. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat*., **29**, 1165–1188.
23. Simon,R., Radmacher,M.D., Dobbin,K. and McShane,L.M. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst*., **95**, 14–18.
24. Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.
25. Snijders,A.M., Nowak,N., Segraves,R., Blackwood,S., Brown,N., Conroy,J., Hamilton,G., Hindle,A.K., Huey,B., Kimura,K. *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genet*., **29**, 263–264.
26. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
27. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
28. Kel,A.E., Gößling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids. Res*., **31**, 3576–3579.
29. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Prüß,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*., **28**, 316–319.
30. Saal,L.H., Troein,C., Vallon-Christersson,J., Gruvberger,S., Borg,Å. and Peterson,C. (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol*., **3**, software0003.1–software0003.6.
31. Gentleman,R.C., Carey,V.J., Douglas,M.B., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 18 2004, pages 3656–3658
doi:10.1093/bioinformatics/bth401

DNMAD: web-based diagnosis and normalization for microarray data

*Juan M. Vaquerizas, Joaquín Dopazo and Ramón Díaz-Uriarte**

Bioinformatics Unit, CNIO, Spanish National Cancer Centre, Melchor Fernández Almagro 3, 28029 Madrid, Spain

Received and revised on May 31, 2004; accepted on July 1, 2004
Advance Access publication July 9, 2004

ABSTRACT

Summary: We present a web server for Diagnosis and Normalization of MicroArray Data (DNMAD). DNMAD includes several common data transformations such as spatial and global robust local regression or multiple slide normalization, and allows for detecting several kinds of errors that result from the manipulation and the image analysis of the arrays. This tool offers a user-friendly interface, and is completely integrated within the Gene Expression Pattern Analysis Suite (GEPAS).

Availability: The tool is accessible on-line at http://dnmad.bioinfo.cnio.es

Contact: rdiaz@cnio.es

1 INTRODUCTION

DNA array technologies (Schena *et al*., 1996) allow the simultaneous monitoring of the expression of thousands of genes. Experiments usually involve measuring the expression levels of genes under several conditions corresponding to different samples, different doses of a certain drug or different stages during an individual's lifetime.

Ideally, DNA microarray techniques allow us to focus on changes in gene expression levels that are solely due to differences between the samples. However, it is widely accepted that other factors such as differences in labelling efficiency, physical properties of the dyes, differences in hybridization, changes in scanner settings, etc. can introduce systematic variation aside from that due to the expression levels of genes in each sample (Yang *et al*., 2002).

Therefore, to obtain accurate and precise results from the analysis of microarray data, normalization of data is essential. Here, we describe a web-based tool for Diagnosis and Normalization of spotted cDNA MicroArray Data (DNMAD).

2 MOTIVATION

Within print-tip group location normalization has proven to be an effective method in normalizing microarray data (Yang *et al*., 2002), but the use of most currently available implementations requires prior training on using a statistical package such as R (http://www.R-project.org), which could represent a relatively high investment especially if the only aim behind such training is to normalize microarray data.

Despite the importance of normalization, there are not many bioinformatic Web tools available to perform normalization procedures, and the few that exist, e.g. SNOMAD (Colantuoni *et al*., 2002) or Multi microarray normalization (http://genome1.beatson.gla.ac.uk/Rweb/anova.html), do not offer certain options that would make the process more efficient such as the use of files coming directly from the scanner or the possibility of entering more than one slide at a time.

In addition, connectivity and data exchange between tools is also important. Web tools integrated within a suite for gene expression pattern analysis, as opposed to individual Web tools, allow users to perform the whole analysis without having to manipulate file formats to make them compatible with the tools used for each individual step. Also, server-based solutions benefit from the server processing capacity (greater than that of personal computers), that allow the user to process more data simultaneously. To allow for efficient handling of requests from many users, the DNMAD service is controlled by a queue system.

The aim of this tool is to provide the scientific community with an easy to use, fully featured web interface for microarray data normalization that is completely integrated within the freely available Gene Expression Pattern Analysis Suite (GEPAS) (Herrero *et al*., 2003a).

3 WEB INTERFACE

This application uses R (R Core, 2004, http://www.R-project.org) and the Bioconductor package limma (Smyth *et al*., 2004, http://www.bioconductor.org), with some custom modifications, to carry out the normalization procedures. The web interface is a Perl CGI script that communicates with R using the CGIwithR (Firth, 2004, http://cran.r-project.org) package. The server accepts a set of GenePix files as input, either compressed as a single file (.tar.gz, .zip or .tar.bz2) or as uncompressed GenePix files coming directly from the scanner. These files must adhere to the original standard file format from GenePix, although customized files containing only the

*To whom correspondence should be addressed.

appropriate columns of data (Block, Column, Row, Name, ID, F635 Mean, B635 Median, F532 Mean, B532 Median and Flags) can also be used if GenePix software has not been used to scan the arrays, or if custom modifications have been made to the data. These columns constitute the minimal set of columns required to perform the normalization. The layout of the array must be introduced into the interface along with the data files. Other options, such as the use of flags (for excluding certain spots from the normalization) or background subtraction, can also be selected in the web interface.

The default normalization method is 'print-tip loess', as implemented in Yang *et al*. (2002). This method consists of a robust local regression for each print-tip group of the array. This method can only be used if the following assumptions are met: (1) the number of points to normalize must be large in each print-tip group; (2) very few genes should be differentially expressed (i.e. the expression of the two cohybridized mRNA samples should be similar for the majority of the spots of the array); and (3) there should be an approximately equal number of up- and downregulated genes in each print-tip group.

If these assumptions cannot be met for each print-tip group, it is possible to use, instead, 'global loess', where the local robust regression is carried out over the whole array instead of for each print-tip group (Yang *et al*., 2002).

4 RESULTS

The output page contains information regarding the type of normalization followed, errors and warnings, plots for interpreting the results and for diagnosis and links to download the normalized data or to enter the hub of GEPAS (Herrero *et al*., 2003a), via the Preprocessor module (Herrero *et al*., 2003b).

A summary of all options selected for normalization is shown at the top of the output page. This information is also attached to the results file that contains the normalized $\log_2$ ratios of expression. Furthermore, a collection of plots and images are provided in order to show plate effect, printing effects and potential problems occurring during the elaboration of the array and the hybridizations, in the sample preparations, in the scanning of the arrays, etc.

These plots include: (1) boxplots for all the arrays and for each array showing individual print-tip groups, to assess the need for normalization for a particular array; (2) MA-plots with the regression curve for each print-tip group, in order to observe the efficiency of the normalization; and (3) diagnostic plots.

Diagnostic plots are compositions of 10 different plots, and they should help the user to detect problems in the arrays (Smyth *et al*., 2003, http://www.bioconductor.org). These plots include: (1) histograms of the raw pixel intensities ($\log_2$) of the red and green mean foreground that help to identify problems in the scanner settings or in the hybridizations; (2) density plots of the $\log_2$ intensities for both channels, displayed in sidewise panels for all the arrays and for each array; and (3) images of the arrays including the red and green background, and the unnormalized and normalized ratio values, which should help to identify spot damaged arrays or spatial patterns. All plots and images can be downloaded along with an HTML document to browse through them.

As the server accepts multiple arrays, and given that these multiples arrays can show differences between their scales, slide-scale normalization is provided. This method, implemented as in Yang *et al*. (2002), reduces differences in the scales of the arrays that are being normalized. After slide-scale normalization, plots and images are provided again, allowing the user to decide if the slide-scale procedure is valuable for the particular data.

Finally, links to download the normalized data [normalized expression ratios in $\log_2$ scale and the A values (the 'average signal' or $0.5 * (\log_2 R + \log_2 G)$)] are displayed, to enable data storage. A direct link to the Pre-analysis module of the Preprocessor (Herrero *et al*., 2003b) is also provided, allowing the user to continue with analysis using the GEPAS suite (Herrero *et al*., 2003a, 2004), such as for clustering genes, or to identify differentially expressed genes, without having to perform any kind of format adjustment to the data.

ACKNOWLEDGEMENTS

We thank A.Lee and an anonymous reviewer for comments on the manuscript, A.Wren for revising the English of the manuscript. We are grateful to S.Dudoit, G.Smyth and P.Kemmeren who provided answers and suggestions about normalization and diagnostics. This tool would not have been possible without the excellent, and free, statistical computing system R and several of its packages available from CRAN or part of Bioconductor, all developed and maintained by volunteers. R.D.-U. was partially supported by a Ramón y Cajal programme from the Spanish Ministry of Science (MCyT). This work is also partly supported by Fundación BBVA and by Project TIC2003-09331-C02-02 of the Spanish MCyT.

REFERENCES

Colantuoni,C., Henry,G., Zeger,S. and Pevsner,J. (2002) SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis. *Bioinformatics*, **18**, 1540–1541.

Firth,D. (2004) CGIwithR: gacilities for the use of R to write CGI scripts.

Herrero,J., Al-Shahrour,F., Díaz-Uriarte,R., Mateos,A., Vaquerizas,J. M., Santoyo,J. and Dopazo,J. (2003a) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.*, **31**, 3461–3467.

Herrero,J., Díaz-Uriarte,R. and Dopazo,J. (2003b) Gene expression data preprocessing. *Bioinformatics*, **19**, 655–656.

Herrero,J., Vaquerizas,J.M., Al-Shahrour,F., Conde,L., Mateos,Á., Santoyo,J., Díaz-Uriarte,R. and Dopazo,J. (2004) New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res.*, **32**, W485–W491.

R Development Core Team (2004) R: a language and environment for statistical computing. Vienna, Austria.

Schena,M., Shalon,D., Heller,R., Chai,A., Brown,P.O. and Davis,R.W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci., USA*, **93**, 10614–10619.

Smyth,G.K., Thorne,N.P. and Wettenhall,J. (2004) Limma: linear models for microarray data version 1.6.6, User's Guide.

Smyth,G.K., Yang,Y.H. and Speed,T. (2003) Statistical issues in cDNA microarray data analysis. *Methods Mol. Biol.*, **224**, 111–136.

Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

W242–W248 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh438

PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level

Lucía Conde, Juan M. Vaquerizas, Javier Santoyo, Fátima Al-Shahrour, Sergio Ruiz-Llorente[1], Mercedes Robledo[1] and Joaquín Dopazo*

Bioinformatics Unit and [1]Hereditary Endocrine Cancer Group, Centro Nacional de Investigaciones Oncológicas (CNIO) Madrid, Spain

Received February 3, 2004; Revised and Accepted April 15, 2004

ABSTRACT

We have developed a web tool, PupaSNP Finder (PupaSNP for short), for high-throughput searching for single nucleotide polymorphisms (SNPs) with potential phenotypic effect. PupaSNP takes as its input lists of genes (or generates them from chromosomal coordinates) and retrieves SNPs that could affect the conserved regions that the cellular machinery uses for the correct processing of genes (intron/exon boundaries or exonic splicing enhancers), predicted transcription factor binding sites (TFBS) and changes in amino acids in the proteins. The program uses the mapping of SNPs in the genome provided by Ensembl. Additionally, user-defined SNPs (not yet mapped in the genome) can be easily provided to the program. Also, additional functional information from Gene Ontology, OMIM and homologies in other model organisms is provided. In contrast to other programs already available, which focus only on SNPs with possible effect in the protein, PupaSNP includes SNPs with possible transcriptional effect. PupaSNP will be of significant help in studies of multifactorial disorders, where the use of functional SNPs will increase the sensitivity of identification of the genes responsible for the disease. The PupaSNP web interface is accessible through http://pupasnp.bioinfo.cnio.es.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the simplest and most frequent type of DNA sequence variation among individuals and they represent one of the most powerful tools for the analysis of genomes (1). Owing to their widespread distribution, SNPs are particularly valuable as genetic markers in the search for disease susceptibility genes, drug response-determining genes, and so on. In the past decades, linkage analysis has been very successful in the identification of genes responsible for mendelian diseases. Nevertheless, direct application of linkage analysis to the case of complex diseases, in which several genes with weaker genotype–phenotype correlations are involved, has resulted in more modest success (2). Now, it is believed that improved genotyping methods in combination with the proper design strategies could bring the genetics of complex diseases to a point of success comparable to where mendelian genetics now firmly resides (3).

There are examples documented in which alleles of more than one gene contribute to the same disease. It is generally believed that multigenic diseases reflect disruptions in the proteins that participate in a protein complex or a pathway (4). Typically, SNPs have been used as markers; that is, the real determinant of the disease was not the SNP itself but some other mutation in linkage disequilibria with it.

The use of functional SNPs could be an important factor for increasing significantly the sensitivity of association tests. In fact, several complex genetic disorders such as Alzheimer's disease (5) and Crohn's disease (6) have been associated with functional SNPs, lending credence to strategies giving priority to candidate markers based on predictable function. The latest build of NCBI's dbSNP (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi) contains 5 772 564 SNPs, with 2 356 957 of them validated. This means that human variation has been screened to an average resolution of 1 SNP for every 566 nt. There is also curated information on SNPs in HGVbase (7). These figures suggest that the possibility of finding the real determinant of a disease among the characterized SNPs can be seriously considered. In fact, dbSNP build 117 contains 24 483 SNPs located in coding regions that produce amino acid change, affecting a total of 9791 different genes. Several estimate suggest that, overall, only 20% of them could damage

*To whom correspondence should be addressed. Tel: +34 912246919; Fax: +34 912246972; Email: jdopazo@cnio.es

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

the protein (8). Much attention has been focused on the possible phenotypic effects of SNPs that cause amino acid changes. The volume of available information together with the development of more sophisticated methods of protein structure prediction has led to different attempts to relate the effect of amino acid changes to structural distortions and, consequently, possible phenotypic effect. Following this, two main different approaches have been taken: on the one hand is the study of conservation of residues in homologous proteins (9) including more sophisticated approaches taking into account the phylogenetic history (10) and, on the other hand, there is the study of changes in the stability (11,12) and other properties of the protein due to changes of amino acids (8,13).

Nevertheless, there are different ways in which the functionality of a gene product can be affected without requiring a amino acid change in the protein. There is increasing evidence that many human disease genes harbour exonic or non-coding mutations that affect pre-mRNA splicing (14). Alternative splicing produced by mutations in intron/exon junctions, or in distinct binding motifs, such as exonic splicing enhancers (ESEs), to which different proteins involved in splicing bind, is the basis of different diseases. In fact, it has been estimated that 15% of point mutations that result in human genetic diseases cause RNA splicing defects (15). For example, a silent mutation in exon 14 of the *APC* gene is associated with exon skipping in a Familial Adenomatous Polyposis (FAP) family (16), and there are many more examples [see Table 2 in (14)]. Also, alterations in the level of expression of gene products can cause diseases. Different SNPs are associated with alterations in gene expression (17) and, in some cases, it is known that they alter some regulatory sequence motif. For example, a regulatory polymorphism in the programmed cell death 1 gene (*PDCD1*), which alters a binding site for the runt-related transcription factor 1 (*RUNX1*) located in an intronic enhancer, is associated with susceptibility to systemic lupus erythematosus in humans (18). It has also been reported that polymorphisms in the gelatinase A promoter region are associated with diminished transcriptional response to estrogen and genetic fitness (19). A recent large-scale screening over a set of 16 chromosomes, found SNPs in the promoters regions of 35% of the genes, and experimental evidence suggested that around one-third of promoter variants may alter gene expression to a functionally relevant extent (20). Therefore, the inclusion of other possible causes of loss of functionality in gene products, beyond the simple estimation of the possible phenotypic effect of an amino acid change, increases considerably the number of SNPs with potential phenotypic effect to be considered for the design of experiments.

Classical statistical linkage tests need a large number of cases if the number of genes to be tested is high. It has only recently been recognized that reliable identification of genetic variants that affect gene regulation is still a challenge in genomics and is expected to play an important role in the molecular characterization of complex traits (21). Another important consideration when analysing multigenic traits is the information available on the genes. Information allows a more targeted approach, by focusing initially on genes whose functionality is related to the disease studied.

Genome surveys based on the information contained in dbSNP show that there are 361 SNPs mapped in splice sites of introns, 1 387 506 in introns and 242 842 in untranslated regions affecting 336 16 306 and 14 198 genes, respectively. A number of these SNPs could be disease determinants.

With the idea of extracting as much information as possible form SNPs with putative phenotypic effect, we have developed PupaSNP Finder (Putative Phenotypic Alterations caused by SNPs; PupaSNP for short). This tool retrieves all the SNPs present in a set of genes of interest that potentially affect the functionality of the gene product. This list is combined with functional information obtained from Gene Ontology (GO) annotations (22). Genes can be directly retrieved from genomic locations or, alternatively, can be taken from a list provided by the user. This corresponds to two typical problems: (i) traits mapped to a given chromosomal region or (ii) traits associated with a given class of genes (e.g. a signalling pathway). Genome coordinates of genes and SNPs are taken from the Ensembl annotation (23).

METHODS

Finding SNPs with potential phenotypic effect

PupaSNP operates with a collection of entries from dbSNP mapped to the Golden Path genome assembly, as implemented in human section of Ensembl (http://www.ensembl.org). As previously mentioned, PupaSNP uses a list of genes and generates a report in which all the SNPs with possible phenotypic effect are listed. The genes can be selected directly by their location in a region of the genome, or just provided as a list (e.g. genes belonging to a given pathway, involved in a particular biological function). Genomic regions can be selected either by defining a range of chromosome coordinates or by directly choosing the cytoband of interest. The engine finds all the genes located within the specified region as well as their promoter regions using Ensembl APIs. In the case of a user-defined list, Ensembl is used to extract their complete intron/exon structure as well as the promoter regions.

The potential effects on the phenotype taken into account are at both transcriptional and gene product levels. These include alterations in (i) transcription factor binding sites, (ii) intron/exon border consensus sequences, (iii) ESE sequences, which are the binding sites for specific serine/arginine-rich (SR) proteins involved in the splicing machinery (24,25) and (iv) the exons that cause an amino acid change. Additionally, the GO terms (22) associated with the genes can be obtained. This is very useful in the case of looking for genes in a chromosomal region, because it can help to discard genes definitively not involved in the disease studied, based on the annotations.

Transcription factor binding sites. In the search for SNPs with potential phenotypic effect, 10 000 bp upstream of the genes, belonging to the promoter region of each gene in the list, are scanned for the presence of possible transcription factor binding sites (TFBSs). The program Match™ (26), version 1.10, from the Transfac® database (27), version professional 7.3, was used for this purpose. SNPs located within these motifs are considered to have a putative phenotypic effect in the expression of the gene. The options used for the program Match™ were (i) group of matrices: vertebrates, (ii) use high quality matrices only and (iii) cutoff selection for

matrix group: to minimize false positives. This cutoff was obtained by exploring the third exon sequences with the weight matrices and was chosen to reduce the number of random putative sites found by the program (26).

Although the scan is done in a region 10 000 bp upstream from the start of the gene, the number of bases to be taken into account in the study is customizable. Obviously, the closer to the start of the gene, the more likely the binding site is to be authentic.

Intron–exon boundaries. Ensembl APIs were used to extract the intron/exon organization of the genes and the corresponding sequences. The two conserved nucleotides at each side of the splicing point, which constitute the splicing signal (14), were then located and all the SNPs altering these signals are recorded.

Exonic splicing enhancers. Mutations that deactivate or activate exonic splicing enhancer sequences may result in exon skipping, malformation, and so on. ESEs also appear to be important in exons that normally undergo alternative splicing. Different classes of ESE consensus motifs have been described, but they are not always easily identified. We have developed a script that scans exon sequences to identify putative ESEs responsive to the human SR proteins SF2/ASF, SC35, SRp40 and SRp55, by using the weight matrices available for them (28). A score is obtained related to the likelihood that the site found is a real ESE. Only ESE sites with scores over the threshold [see (28) and http://exon.cshl.org/ESE/ESEmatrix.html for details] are taken into account in the analysis. Threshold values, above which a score for a given sequence is considered to be significant, are set as the median of the highest score for each sequence in a set of 30 randomly chosen 20 nt sequences (from the starting pool used for functional assays for ESE identification; see http://exon.cshl.org/ESE/ESEmatrix.html). If an SNP disrupts one of these sequences, the new score, corresponding to the mutated sequence, is also calculated. Strong differences between the two score values suggest more drastic effects caused by the SNP.

Changes at amino acid level and functional implications. SNPs that result in a change of amino acid are likely to cause some phenotypic effect and, consequently, are all listed. Since the main purpose of the tool is to cover possible transcriptional effects of the SNPs and there are a number of tools already available for the prediction of phenotypic effects due to mutations in amino acids (see Introduction) PupaSNP only lists them. To help in the identification of possible effects we label SNPs that disrupt any functional motif as listed in Interpro (29), a resource that compiles information on protein families, domains and functional sites. The coordinates of the Interpro motifs within the exons of the genes are extracted from Ensembl and cross-referenced with the SNPs coordinates.

Additional functional information. Since PupaSNP finder works with lists of genes in order to select the best SNP candidates for further use in association analysis, it is very helpful to have functional annotations of the genes. This allows the assignment of priorities based also on the information available on the genes. Information is obtained from (i) Gene Ontology annotations, obtained through the FatiGO engine (30) (available at http://fatigo.bioinfo.cnio.es), (ii) OMIM (Online Mendelian Inheritance in Man), which constitutes a comprehensive, authoritative and timely knowledge base of human genes and genetic disorders (31) and (iii) homologies to other organisms, obtained directly from Ensembl. Gene Ontology is a tree structure (called a directed acyclic graph) in which terms describing three fundamental ontologies (molecular function, biological process and cellular component) have descendants with more detailed descriptions. Thus, descending the hierarchy of GO implies moving towards terms with more detailed descriptions of the ontologies, but, at the same time, there are fewer genes with annotations at such detail. FatiGO works by climbing up the hierarchy to a selected parent level (30) to optimize the number of genes with annotation and the detail of the annotation. Thus, the identification of common parent functions or processes is easier. In this way, the consideration of the SNPs in a functional context can help to understand the potential biological implications of the SNPs and genes studied.

RESULTS

SNPs with possible phenotypic effect

We analysed a total of 24 037 human genes corresponding to the annotations in Ensembl build 34 (version 18.34.1), which contains the mapping of dbSNP 117. By scanning with the Match^{TM} program the 10 000 bp upstream promoter regions of the genes, 2 587 478 transcription factor binding sites, corresponding to 330 different Transfac weight matrices (27), were found. After mapping the SNPs in the promoter regions, 71 444 TFBSs were found to be disrupted by a total of 57 412 SNPs (some SNPs affect more than one TFBS at the same time). A total of 19 010 genes presented at least 1 predicted TFBS disrupted by a SNP, which constitutes a considerable proportion of the total number of genes. The coverage in terms of both SNPs and TFBS predictions was good: only for 54 genes was no single SNP found in the 10 000 bp 5′-upstream region, and only for 2 genes could no predicted TFBS be found (*ENSG00000116119*, or *KV2A_HUMAN*, which is the IG KAPPA CHAIN V-II REGION CUM, and *ENSG00000174994*, or *AK057375*, which seems to be a DNA binding protein). In a number of cases, SNPs affect overlapping TFBSs, which could have a stronger effect still in the phenotype. There are even 2 SNPs that simultaneously affect 15 TFBSs.

The four conserved bases that define intron–exon boundaries were mutated by 844 SNPs, affecting to a total of 598 genes.

Over eight million ESE motifs were found, covering all the genes studied. A total of 138 746 SNPs were found to disrupt ESE sequences. These SNPs affect a total of 17 312 genes.

These results suggest that, in the search for SNPs with potential phenotypic effects, regulatory SNPs or SNPs affecting splicing should not be neglected.

The web interface

Input data. PupaSNP has been designed for high-throughput screening of functional SNPs. Thus, the input consists of a list of genes. The list can be directly provided as a collection of gene identifiers (Ensembl IDs, or external IDs, which include

GenBank, Swissprot/TrEMBL and other gene IDs supported by Ensembl) or can be specified by means of a chromosomal location (cytobands or chromosomal coordinates). In the latter case, PupaSNP extracts all the genes contained in the specified location. Ensembl coordinates are used to extract the genes. Only Ensembl annotated genes, but not predictions, are extracted.

User-defined SNPs. Alternatively, the user can input SNPs not in the database in a very straightforward manner and take advantage of the tools for predicting their potential phenotypic effect. A text file containing the descriptions of the SNPs must be generated. Each line describes one unique SNP with the following tab-delimited data: SNP name, gene (Ensembl ID or external ID), position with respect to the start of the translation and alleles, e.g.

```
MySNP01  ENSG00000000003  -1830  A/G
MySNP02  ENSG00000157873  421    C/G
```

This describes two SNPs: the first in the gene *ENSG00000000003* (*tetraspanin 6*, or *TSPAN6*), 1830 bp away from the transcription start point, with polymorphisms consisting of a change of an A for a G; and the second in gene *ENSG00000157873* (tumor necrosis factor receptor-like 2, *TNFRSF14*), 421 bp within the transcripted region, which corresponds to the first exon of the gene.

The web interface. A web interface to PupaSNP is available at http://pupas.bioinfo.cnio.es/. Lists of genes can be defined by chromosome position, which can be specified in terms of cytoband units or in absolute chromosomal position (as mapped in the corresponding Ensembl assembly). The upstream region makes reference to the number of bases upstream in which TFBSs will be searched for (with a upper limit of 10 000 bp). Also, lists of genes can be uploaded or just pasted into the box. PupaSNP finds all the SNPs mapping to locations that might cause a loss of functionality in the genes. Functional information for the genes can also be obtained from OMIM and from Gene Ontology. Information on homologous genes can also be retrieved. Finally, SNPs do not need to be annotated in the genome to be included in the query tool. The user can specify a list of SNPs using a gene as reference. In this way the use of absolute coordinates, which can easily change between assembly versions, is avoided in favour of the use of coordinates relative to genes, which tend to be more stable. Results include SNPs in a the promoter region of the genes, SNPs located at intron boundaries, SNPs located at exonic splicing enhancers and coding SNPs located at Interpro domains. Figure 1 shows part of the results provided by the program for the SNPs with possible phenotypic effect on genes in the p36.33 cytoband of chromosome 1. Figure 1C is especially interesting because it shows how the scores obtained by the motif scanning method can be used to assess the possible impact of the polymorphism on the recognition of the ESE motif by the cellular machinery.

Both the SNPs and the genes found are linked to the Ensembl Genome Browser.

Experimental validation

The validation status of the SNPs is, in some cases, a much more important factor for their selection than their possible functional role. Such information is scarce: 2 359 534 out of 5 798 183 SNPs in dbSNP build 118 have been validated, which constitutes 40%. However, only 160 466 have estimates of population frequencies and only 94 867 have a phenotype associated. To obtain a sense of the reliability of the SNPs annotated with 'no-info', a set of SNPs was sought for a list of candidate modifier genes related to a phenotype exhibited by *MEN2* (Multiple endocrine neoplasia, type IIA) patients (OMIM, #171400), all of them *RET* mutation carriers. *MEN2* is an autosomal dominant syndrome of multiple endocrine neoplasms, with variable clinical expression even between members of the same family. This fact cannot be explained only by a mutation in a major susceptibility gene, but suggests a role for genetic modifiers, which may also work through quantitative effect.

In most of cases, it was necessary to validate the putative SNPs identified by PupaSNP because there was no information about validation status. To validate SNPs and estimate their allele frequency, 48 non-related individuals from the Spanish population were used. The specific primers used to amplify the fragments of interest by PCR (polymerase chain reaction) were designed using the OLIGO 4.1 program. When possible, the primers were selected and designed to amplify a fragment (200–500 bp) that allowed us to investigate several SNPs at the same time. As a denaturing high-performance liquid chromatograph (dHPLC) system (WAVE, Transgenomics Limited, Crewe, UK) was used for the initial SNP screening, the fragments of interest had a homogeneous GC content across different domains from the DNA fragment to obtain a consistent melting profile. The Navigator software was used for data handling and optimization of the dHPLC system. After normalization, each PCR product that exhibited a change in the chromatogram profile was characterized by sequence analysis. These PCR products were purified using an E.Z.N.A. Cycle-Pure Kit (Omega Bio-tek, USA) according to the manufacturer's instructions, and sequenced using an automatic sequencer ABI PRISM^{TM} 3700 (Applied Biosystems. Perkin Elmer, USA). The reaction was carried out in 4 µl of a Big Dye terminator cycle sequencing Kit (Perkin Elmer, USA), 10 pmol of the sense/antisense primer, 5% DMSO and 6–12 ng of amplified DNA. Although the results obtained here do not pretend to be capable of general extrapolation to the entire database, we have found that 24 out of 28 SNPs assayed proved to be authentic and polymorphic in the Spanish population, which constitutes a good rate.

DISCUSSION

Typically, SNPs have been used as markers to search for the real determinant of a disease in linkage disequilibria with it. As previously mentioned, the use of functional SNPs, which may be the real disease determinants, could be an important factor in increasing the sensitivity of association tests.

Despite the obvious importance that alterations in the regulation, expression level or splicing of genes can have for the phenotype, these have long been ignored in the most common approaches to finding functional SNPs, which have instead focused more on the possible effect of polymorphisms causing amino acid changes. Apart from the databases mentioned above (dbSNP and HGVbase), there are a number of resources

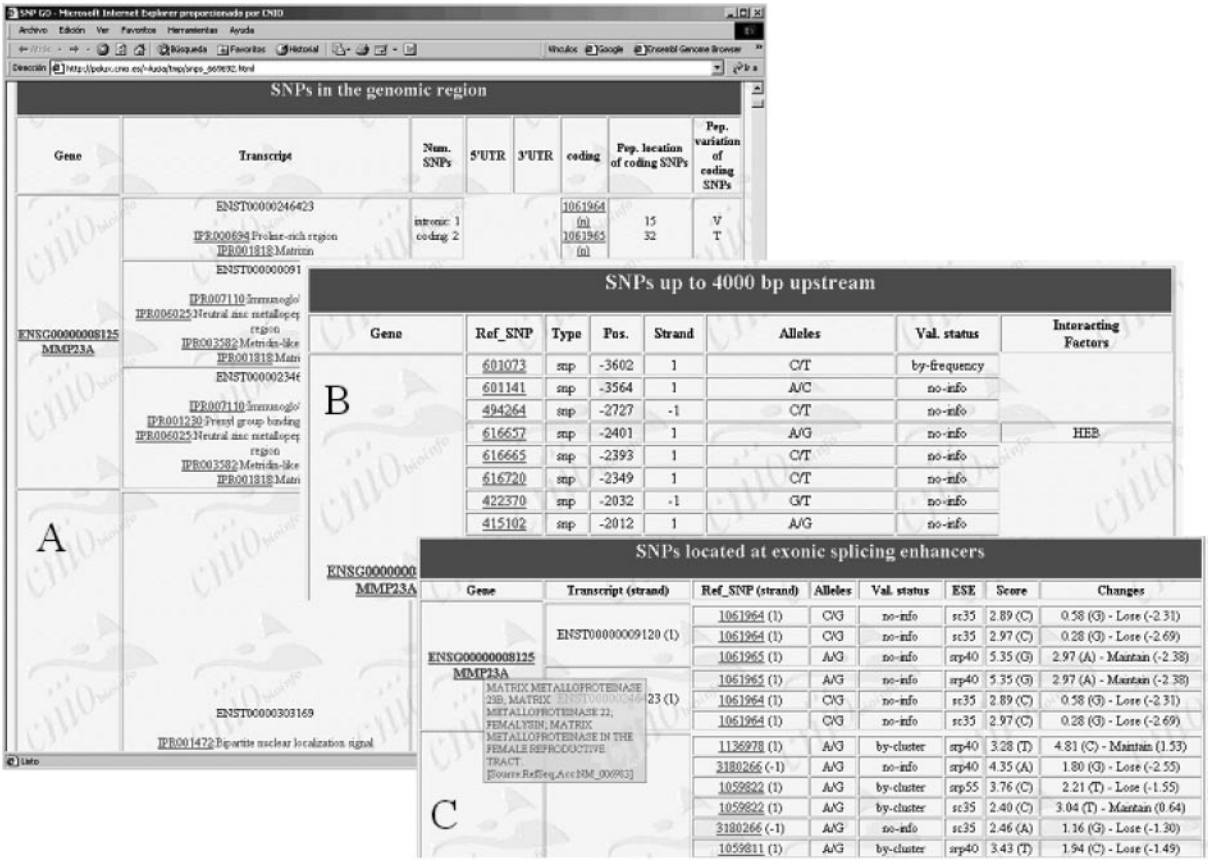

Figure 1. A selection of results from PupaSNP. (**A**) List of genes and the corresponding transcripts with the SNPs mapping to the different regions, which include coding and 5′- and 3′-untranslated regions. For coding SNPs, the position within the transcript and the change produced (if any) is reported. (**B**) SNPs located in the promoter regions (in the example, a limit of 4000 bp was chosen). Disruptions of predicted TFBSs are listed. The validation status of the SNPs ('no-info', 'by-submitter', 'by-frequency', 'by-cluster'; see dbSNP web page) is also provided. (**C**) SNPs located at exonic splice enhancers. The scores make reference to the closeness of the site to the motif. If the polymorphism gives a site with a worst score, this would, generally speaking, probably imply worst recognition of the site by the cellular machinery and, consequently, a putative alteration in the normal splicing process. When the cursor is over the gene name, additional information is displayed.

available over the net collecting information on phenotypes associated with SNPs, such as The Human Gene Mutation Database (http://www.hgmd.org) at the University of Wales, which classifies SNPs according the lesion they cause (missense substitutions, splice variants, and so on) (32) and PicSNP, a catalogue of non-synonymous SNPs obtained from the human genome assembly (33). However, these are mainly specialized catalogues collecting information on SNPs rather than tools for their selection.

PupaSNP constitutes a tool for selecting SNPs with putative phenotypic effects designed for high-throughput experiments. It deals with lists of genes, instead of focusing on individual genes. In addition, more information on different possible motifs with regulatory function has been included. For example, SNPs in ESE had never previously been included in any catalogue.

Multigenic diseases are generally associated with disruptions in proteins that participate in a protein complex or a pathway (4). The inclusion in PupaSNP of information regarding the participation of genes in signalling cascades or in pathways or in protein complexes will be considered in the near future. Databases containing protein interaction data, such as DIP and BIND (see http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-interaction.html), can be an important source of information to be considered in the search for SNPs affecting multigentic traits.

Despite the fact that PupaSNP is more focused on SNPs with possible effects at transcriptional level, the inclusion of an algorithm for improving the predictions of the effect of SNPs in the proteins, such as FoldX (12), would provide, within the same framework, both types of result.

Minimum SNP set selection allows the user to optimize the number of SNPs required to represent haplotype diversity, thus reducing the cost of genotyping by assaying the minimum number of SNPs required. The inclusion of information on linkage disequilibrium or on haplotype blocks can assist in a more efficient selection of SNPs. Some programs, such as HapScope (34), include information on haplotypes and use them to select minimum subsets of SNPs. Another important issue is the reliability of the SNPs. As previously mentioned, only 40% of the SNPs in dbSNP have been validated, and only for 5% are population frequencies are available. This means that most of the SNPs found in any kind of selection will lack information on their possible presence in the population of interest as a manageable polymorphism. Even though our results suggest a high rate of authenticity, even for the SNPs labeled as 'no-info', they must be treated carefully

and cannot be directly extrapolated to the entire database. As population frequencies are included in the database, these data could be of interest for use as part of the selection process of SNPs

PupaSNP will be the tool used in the first step of the pipeline for the study of polymorphisms at the Spanish National Genotyping Centre (CeGen). For this reason it has been developed to cope with high-throughput experimental designs. PupaSNP takes as input lists of genes (or generates them from chromosomal coordinates) and provides results which integrate all the information available as well as obtained by means of predictions of SNPs with possible functional consequences.

ACKNOWLEDGEMENTS

L.C. and this work are supported by grant PI020919 from the Fondo de Investigaciones Sanitarias. F.A.-S. is supported by grant BIO2001-0068 from Ministerio de Ciencia y Tecnología. This work is also partly supported by a grant from Fundació La Caixa and by the Spanish National Genotyping Centre (CeGen), funded by Genoma España, which is using this program for high-throughput SNP selection.

REFERENCES

1. Collins,F.S., Green,E.D., Guttmacher,A.E. and Guyer,M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
2. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
3. Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet*., **33**, 228–237.
4. Badano,J.L. and Katsanis,N. (2002) Human genetics and disease: beyond Mendel: an evolving view of human genetic disease transmission. *Nature Rev. Genet*., **3**, 779–789.
5. Strittmatter,W.J., Saunders,A.M., Schmechel,D., Pericak-Vance,M., Enghild,J., Salvesen,G.S. and Roses,A.D. (1993) Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer's disease. *Proc. Natl Acad. Sci. USA*, **90**, 1977–1981.
6. Hugot,J.P., Chamaillard,M., Zouali,H., Lesage,S., Cezard,J.P., Belaiche,J., Almer,S., Tysk,C., O'Morain,C.A., Gassull,M., Binder,V., Finkel,Y., Cortot,A., Modigliani,R., Laurent-Puig,P., Gower-Rousseau,C., Macry,J., Colombel,J.F., Sahbatou,M. and Thomas,G. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.
7. Brookes,A.J., Lehvaslaiho,H., Siegfried,M., Boehm,J.G., Yuan,Y.P., Sarkar,C.M., Bork,P. and Ortigao,F. (2000) HGBASE: a database of SNPs and other variations in and around human genes. *Nucleic Acids Res*., **28**, 356–360.
8. Sunyaev,S., Ramensky,V., Koch,I., Lathe,W., Kondrashov,A.S. and Bork,P. (2000) Prediction of deleterious human alleles. *Hum. Mol. Genet*., **10**, 591–597.
9. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res*., **11**, 863–874.
10. Miller,M.P. and Kumar,S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet*., **10**, 2319–2328.
11. Chasman,D. and Adams,R.M. (2001) Predicting functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol*., **307**, 683–706.
12. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol*., **320**, 369–387.
13. Ferrer-Costa,C., Orozco,M. and de la Cruz,X. (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol*., **315**, 771–786.
14. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet*., **3**, 285–298.
15. Krawczak,M., Reiss,J. and Cooper,D.N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet*., **90**, 41–54.
16. Montera,M., Piaggio,F., Marchese,C., Gismondi,V., Stella,A., Resta,N., Varesco,L., Guanti,G. and Mareni,C. (2001) A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. *J. Med. Genet*., **38**, 863–867.
17. Yan,H., Yuan, W., Velculescu,V.E., Vogelstein,B. and Kinzler,K.W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.
18. Prokunina,L., Castillejo-Lopez,C., Oberg,F., Gunnarsson,I., Berg,L., Magnusson,V., Brookes,A.J., Tentler,D., Kristjansdottir,H., Grondal,G., Bolstad,A.I., Svenungsson,E., Lundberg,I., Sturfelt,G., Jonssen,A., Truedsson,L., Lima,G., Alcocer-Varela,J., Jonsson,R., Gyllensten,U.B., Harley,J.B., Alarcon-Segovia,D., Steinsson,K. and Alarcon-Riquelme,M.E. (2002) A regulatory polymorphism in *PDCD1* is associated with susceptibility to systemic lupus erythematosus in humans. *Nature Genet*., **32**, 666–669.
19. Harendza,S., Lovett,D.H., Panzer,U., Lukacs,Z., Kuhnl,P. and Stahl,R.A. (2003) Linked common polymorphisms in the gelatinase a promoter are associated with diminished transcriptional response to estrogen and genetic fitness. *J. Biol. Chem*., **278**, 20490–20499.
20. Hoogendoorn,B., Coleman,S.L., Guy,C.A., Smith,K., Bowen,T., Buckland,P.R. and O'Donovan,M.C. (2003) Functional analysis of human promoter polymorphisms. *Hum. Mol. Genet*., **12**, 2249–2254.
21. Hudson,T.J. (2003) Wanted: regulatory SNPs. *Nature Genet*., **33**, 439–440.
22. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M. and Sherlock,G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*., **25**, 25–29.
23. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T., Durbin,R., Eyras,E., Gilbert,J., Hammond,M., Hubbard,T., Kasprzyk,A., Keefe,D., Lehvaslaiho,H., Iyer,V., Melsopp,C., Mongin,E., Pettett,R., Potter,S., Rust,A., Schmidt,E., Searle,S., Slater,G., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Stupka,E., Ureta-Vidal,A., Vastrik,I. and Birney,E. (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res*., **31**, 38–42.
24. Liu,H.X., Zhang,M. and Krainer, A.R. (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev*., **12**, 1998–2012.
25. Schaal,T.D. and Maniatis,T. (1999) Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol. Cell Biol*., **19**, 261–273.
26. Kel,A.E., Gößling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCHTM: a tool for searching transcription factor binding sites in DNA sequences *Nucleic Acids Res*., **31**, 3576–3579.
27. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Prüß,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*., **28**, 316–319.
28. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res*., **31**, 3568–3571.
29. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P., Bucher,P., Copley,R.R., Courcelle,E., Das,U., Durbin,R., Falquet,L., Fleischmann,W., Griffiths-Jones,S., Haft,D., Harte,N., Hulo,N., Kahn,D., Kanapin,A., Krestyaninova,M., Lopez,R., Letunic,I., Lonsdale,D., Silventoinen,V., Orchard,S.E., Pagni,M., Peyruc,D.,

Ponting,C.P., Selengut,J.D., Servant,F., Sigrist,C.J., Vaughan,R. and Zdobnov,E.M. (2003) The InterPro Database brings increased coverage and new features *Nucleic Acids Res.*, **31**, 315–318.

30. Al-Shahrour Díaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
31. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders *Nucleic Acids. Res.*, **30**, 52–55.
32. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
33. Chang,H. and Fujita,T. (2001) PicSNP: a browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome. *Biochem. Biophys. Res. Commun.*, **287**, 288–291.
34. Zhang,J., Rowe,W.L., Struewing,J.P. and Buetow,K.H. (2002) HapScope: a software system for automated and visual analysis of functionally annotated haplotypes *Nucleic Acids Res.*, **30**, 5213–5221.