THE UNIVERSITY
of EDINBURGH

University of Edinburgh

Doctoral Thesis

Regulatory complexity in gene expression

Author:
Sarah Rennie

Supervisor:
Dr. Martin Taylor

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

School of Medicine and Veterinary Medicine
The University of Edinburgh

2016

Declaration of Authorship

I, Sarah Rennie, declare that this thesis titled, 'Regulatory complexity in gene expression' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UNIVERSITY OF EDINBURGH

Abstract

Institute of Genetics and Molecular Medicine

School of Medicine and Veterinary Medicine

The University of Edinburgh

Doctor of Philosophy

Regulatory complexity in gene expression

by Sarah Rennie

The regulation of gene expression is the driver of cellular differentiation in multicellular organisms; the result is a diverse range of cell types each with their own unique profile of expression. Within these cell types the transcriptional product of a gene is up or down regulated in response to intrinsic and extrinsic stimuli according to its own regulatory programme encoded within the cell. The complexity of this regulatory programme depends on the requirements of the gene to change expression states in different cell lineages or temporally in response to a range of conditions. In the case of many housekeeping genes integral to the survival of the cell, this programme is simple - switch on the gene and leave it on, whereas often the required level and precision of regulatory control is much more involved and lends to subtle changes in expression. This raises many questions of precisely where and how that regulatory information is encoded and whether different biological systems encode it in the same way.

This project attempts to answer these questions through the development of novel approaches in quantifying the output of this regulatory programme according to the state changes as observed from the expression profile of a given gene. Measures of complexity in gene expression are calculated over a wide range of cell types and conditions collected using CAGE, which provides a quantitative estimate of gene expression that precisely defines the promoter utilised to initiate that expression. As expected, housekeeping genes were found to be amongst the least complex, as a result of their uniform expression profiles, as well as those genes highly restricted in their expression. The genes most complex in their expression output were those associated with the presence of H3K27me3 repressive marks; genes poised for activation in a specific set of cell types, as well as those enriched in DNAse I hypersensitive sites in their upstream region but not necessarily conserved in that region. Evidence also suggests that different promoters associated with a gene contribute in different ways to its resultant regulatory complexity, suggesting that certain promoters may be more crucial in driving the regulation of some genes. This allows for the targeting of such promoters in the analysis of certain diseases implicated by changes in regulatory regions. Indeed, genes known to be associated with diseases such as leukaemia and Alzheimer's are found to be highly complex in their expression.

Acknowledgements

Martin Taylor for his support and infinite patience, together with his understanding towards all the unfortunate things that happened in my life over the past five years. And for giving me the day off when my rabbit Buster died.

Colin Semple, my thesis committee (Paul McKeigue and Richard Baldock) and the Evogen group for their comments and helpful advice. In particular Robert Young for the discussions on the time-course CAGE data.

I would like to acknowledge the FANTOM consortium for the fascinating data and insights into gene expression regulation, a field I knew nothing about when I started but have developed a strong passion for. In particular, RIKEN for providing funding for me to go to Yokohama twice and giving me a platform to present my work on both occasions, and I would like to acknowledge the FANTOM consortium for their helpful comments and advice on both occasions.

I would like to acknowledge the MRC Capacity Studentship for funding my studies over 4 years.

Duncan Sproul for his advice on polycomb repression.

Xu Gu, Jo Pethick, Chris Armit for being brilliant friends and the IGMM students for their friendship.

My brother Julian Kitagawa for his supportive comments over online chat at 6am Tokyo time!

BunBun, Mango and Buster, my beloved pet rabbits, Freddie the hamster and Betty and Daisy the rats. Definitely could not have done it without you all!

Jason Rennie, for all the cups of tea!

Contents

List of Figures

List of Tables

Abbreviations

CAGE	**C**ap **A**nalysis of **G**ene **E**xpression
DHS	DNase I hypersensitive site
DNA	Deoxyribonucleic acid
CDS	Coding sequence
GWAS	**G**enome **W**ide **A**ssociation **S**tudy
TSS	**T**ranscription **S**tart **S**ite
QTL	Quantitative trait loci
lncRNA	long non-coding RNA
RNA	Ribonucleic acid
mRNA	messenger RNA
eRNA	enhancer RNA
tRNA	transfer RNA
TBP	TATA-binding protein
CGI	CpG island
SHH	Sonic hedgehog gene
snRNA	small nuclear RNA
UTR	Untranslated region
SAGE	Serial analysis of gene expression
RNA-seq	Whole Transcriptome Shotgun Sequencing
CHIP	chromatin immunoprecipitation (ChIP)
PCR	Polymerase chain reaction
DE	Differential expression
SNP	Single nucleotide polymorphism

`lm`	linear model
`glm`	generalised linear model
HGMD	Database of human gene mutation data
GERP	Genomic Evolutionary Rate Profiling
JAGS	Just another Gibbs Sampler
BUGS	Bayesian using Gibbs Sampling
NB	Negative binomial
polII	Polymerase II

Chapter 1

Introduction to complexity

In this chapter I describe how the basic genetic material in a cell is utilized to form protein products and the layers of regulation involved in order to achieve this. I describe the core promoter and discuss its architecture observed on a genomic scale and how regulatory elements interact with this core promoter, both locally (in cis) and off-site (in trans), in order to regulate transcription beyond the basal levels achieved by the core promoter alone.

I describe next generation sequencing technology, in particular cap analysis of gene expression (CAGE), the basic technology currently leading the research in locating and characterising transcriptional start sites.

I then discuss the concept of 'regulatory complexity', based on the idea that the information content in the genome acts in a combinatorial manner to achieve final levels of expression observed across time and space within an organism. Using this idea the regulation of expression is described as a regulatory program and I discuss how this applies in the context of evolution across species as well as what it means for an individual gene within a single eukaryote. This forms the basis for discussing measures of regulatory complexity in gene expression, which will be covered in Chapter 2.

1.1 Information flow: the central dogma of genetics

The central dogma off molecular genetics (Figure 1.1) describes the flow of genetic information from the genetic blue-print, passed down from parent to child cells and organisms, to the traits observed in the organism, via the production of molecular machinery that builds and controls cells. Deoxyribonucleic acid (DNA) is the storage medium for this genetic blue-print and is a polymer made up of a sugar-phosphate backbone and nucleotide base side chains. Genetic information is encoded as a linear string of four bases: adenine (A), cytosine (C), guanine (G) and thymine (T) along the polymer. This quaternary system acts as a code, analogous to the binary encoding of 1's and 0's in digital technology.

DNA typically occurs not as a single strand as described above, but a double stranded structure where the two polymers form a double helix, which resembles a ladder like structure, with a sugar phosphate backbone and base pairs linked from each strand via a hydrogen bond, thus forming the ladder rungs. The pattern of hydrogen bonding between bases is complementary: A pairs with T and C pairs with G. This suggests a mechanism for replication, since each separated strand of the DNA may act as a template for a new strand. This was realised by Watson and Crick [Watson et al., 1953], who first proposed the anti-parallel, double stranded structure for DNA, and confirmed by Mesalson and Sthal [Meselson and Stahl, 1958], proving a mechanistic basis for the inheritance of genetic material.

The information encoded within DNA translates into proteins, linear polymers of amino acids. In this model, groups of three adjacent bases, known as codons, encode a specific amino acid. In the standard genetic code that is common to most organisms, there are twenty possible encoded amino acids. Since there are $4^3 = 64$ possible triplet combinations of the four bases, there is clearly a degeneracy in this code, implying that multiple codons may encode the same amino acid. There are also three possible stop codons, encoding the end of a protein. As the central dogma illustrates, Figure 1.1, DNA sequence does not translate to proteins directly, but is first transcribed into messenger ribonucleic acid (mRNA) molecules. RNA, like DNA, has a sugar-phosphate

backbone and nucleotide base chains and mRNA is synthesised from one of the DNA strands, using complementary base pairing rules as a template, similar to what is observed in DNA replication.

After the mRNA is synthesised from DNA, it is loaded into a ribosome, which is a large protein which sequentially reads the mRNA as a codon code, guiding the production of the encoded protein through amino acid polymerisation. The proteins produced as the enzymes, regulatory switches and structural components of cells.

Proteins are the major functional product of genomes, however some DNA sequences encode other information; there are many molecular mechanisms which require an RNA molecule rather than a protein as the functional output of a stretch of DNA sequence.

DNA sequence also encodes regulatory information, dictating when and in what quantities specific products of the genome should be manufactured, or expressed. Almost all cells in the human body contains the same genomic DNA sequence. However, clear differences between cell types exist; for example, skin cells have a very distinct structure and function to muscle cells and again neurons and liver cells. The differences between cell types represent changes in the genomic products manufactured, typically being referred to as expressed. It is this aspect of regulated information flow from the genome that is the central focus of this thesis - how one genome encodes the expression patterns for hundreds of different cell types [Forrest et al., 2014].

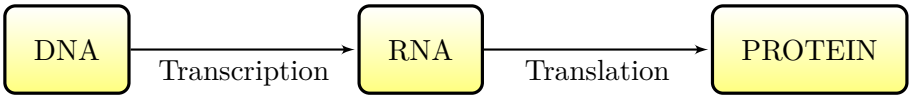


FIGURE 1.1: Central dogma of genetics. DNA becomes RNA through the process of transcription, which becomes a protein through the process of translation.

1.2 Anatomy of the genome

Key aspects of genome structure, content and packaging of DNA are outlined in the subsequent sections that provide context for the discussion of genomic regulation.

1.2.1 Genes

Although the term is very loosely applied across different contexts, the stretch of DNA that encodes a discrete product is referred to as a gene. A protein coding gene is a gene where its transcribed RNA is subsequently translated by the ribosome into a protein. Genes not transcribed into proteins are referred to as non-coding genes.

In a single (haploid) copy of the human genome, there are currently thought to be approximately 20,300 protein coding genes and 24,885 non-coding genes (Ensembl version 79 [Flicek et al., 2013]), although both numbers are subject to regular revision. The transcription of a gene into RNA is performed by RNA-polymerase complexes. All protein coding genes and many of the diverse non-coding RNAs are transcribed by RNA-polymerase II, whereas the main RNA-polymerase I and III are involved in the production of specialist, often high-abundance RNA species such as the ribosomal RNA. This work focusses specifically on the regulation of RNA-polymerase II transcripts.

RNA polymerase II (polII) initiates transcription at the transcription start site (TSS) and extends the RNA polymer along the template DNA until the termination of transcription is triggered, a process often involving transcription across a cleavage and polyadenylation signal (consensus sequence AATAAA) ([Colgan and Manley, 1997, Elkon et al., 2013]).

The resultant RNA is modified co-transcriptionally through the addition of a 7-methylguanosine (m^7G) cap, a Poly-A tail via polyadenylation, and through co-transcriptional splicing to remove non-coding intronic sequences (reviewed in [Bentley, 2014]). The capping process occurs on the 5' end of the nascent molecule, which contains a free triphosphate group, since it is the first nucleotide in the transcript. The enzyme guanyltransferase connects a guanine residual in a reverse orientation via a 5'-5' linkage to this free triphosphate group, which is in turn methylated by a methyltransferase on position 7. [Byszewska et al., 2014, Wei and Moss, 1977]. Capping is generally the first step after the 5' end of the transcript becomes exposed, thus protecting it from degradation and effectively marking them for exportation into the cytoplasm, although not all transcripts are necessarily capped and some may even have their capped removed in decay

associated pathways [Bentley, 2014]. The chemical stucture of the 5' cap is exploited in the CAGE (Cap Analysis of Gene Expression) protecol, which essentially traps the cap structure of the transcript, allowing for sequencing of these transcriptions and the genome-wide identification of transcription start sites [Carninci et al., 2005]. CAGE data forms the basis of the analysis in this project and the technique and datasets will be explained in more detail later in this chapter and in chapter 3.

After capping, polyadenylation and splicing, the result is a mature protein coding transcript, containing a chain of codons encoding amino acids followed by a stop codon to guide termination of translation, referred to as the coding sequence (CDS). The portion of non-coding sequence between the 5' cap and the coding sequences is referred to as the 5' UTR (untranslated region) and the RNA sequence between the coding sequence and the poly-A tail is the 3' UTR. It appears that every step in the production of such a transcript, including the initiation of transcription, transcription elongation, splicing, polyadenylation, transport, translation and degradation, are all mechanisms utilised by cells to regulate the amount of a gene product produced (discussed in later sections) [Jones, 2015]. In addition, alternative initiation sites, splicing and cleavage and poly-A sites utilisation can allow a single gene coding region of DNA to encode multiple functionally distinct products [Elkon et al., 2013]. Indeed, the currently annotated version of the human genome (Ensembl version 79) suggests the 45,185 annotated coding and non-coding genes can be processed into 198,622 distinct transcript species, and this is almost certainly an under-estimate of the total transcript diversity [Flicek et al., 2013].

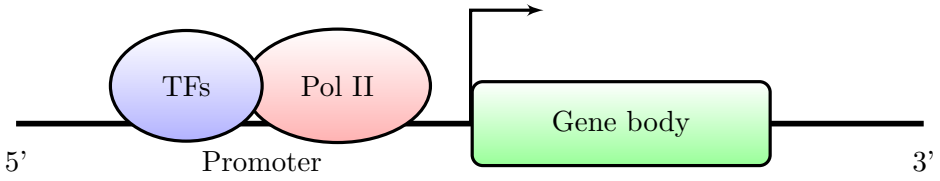


FIGURE 1.2: Simple schematic of transcription. One or more transcription factors (TFs) bind to the promoter region. RNA polymerase II (Pol II) initiates transcription from the transcription start site

The result is a strand of mRNA which can then be translated into a protein. Gene products which are RNAs and not proteins are called non-coding RNAs. For protein-coding genes, after the DNA has been transcribed into mRNA, the mRNA employs

the help of the ribosome to make the required protein. Each set of three nucleotides forms a codon which relates to a particular amino acid which binds to the appropriate locations on the mRNA. When translation is complete, the completed protein falls off of the mRNA strand.

A gene is expressed when its coding information in the DNA is used to create a functional gene product. In the case of protein coding genes, this product is a protein which is synthesised from RNA converted from the DNA code and which provides some kind of function to the cell. Non protein coding genes produce RNA products, such as tRNA, miRNA or snRNA (transfer RNA, microRNA and small nucleur RNA). Gene expression is a tightly regulated process; genes are regulated by a number of different modes - during transcription, post-transcription, during translation, post-translational or through epigenetic factors. Most of the scope of this project focusses on transcriptional initiation and epigenetic factors.

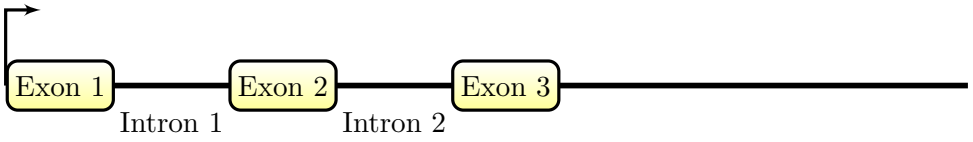


FIGURE 1.3: Simple schematic of the structure of a gene

1.2.2 Promoters and the initiation of transcription

Transcription initiation is the first and foremost step in the expression of a gene. Transcriptional initiation is brought about by the presence of the core promoter of the gene, a short sequence a small distance upstream of the start site. Six general transcription factors bind to the promoter, namely TFIIa,TFIIB,TFIID,TFIIE,TFIIF and TFIIH [Sims et al., 2004]. These assemble to form the pre-initiation complex (PIC). The PIC results in the recruitment of polII to the transcription start site (TSS), so that initiation

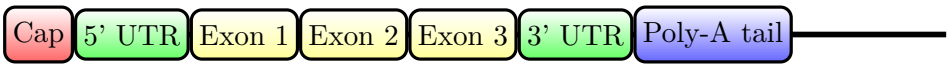


FIGURE 1.4: Simple structure representing an mRNA transcript. The introns have been spliced out, a cap added to the 5' UTR region and a poly(A) tail added to the 3' UTR. Further modifications may occur before it is translated into a protein.

and productive elongation may occur. Evidence is accumulating that the complexes forming the PIC have roles in mediating interactions with the three dimension structure within the nucleus and have important regulatory function in driving cell-specific expression during development [Goodrich and Tjian, 2010].

Even after the completed assembly of the PIC onto the promoter, initiation of transcription by polII does not occur until the template strands are separated. This requires ATP, the unit of energy within the cell, and a subunit of TFIIH [Luse, 2013]. The consensus is that TFIIH, within the PIC, rotates the DNA within the -9 to -2bp region upstream of the actual TSS. Other complexes are further needed to aid the formation of the first bond in the RNA-DNA complex and promote stable elongation and prevent the early abortion of the transcript [Luse, 2013].

As polII moves past the TSS, promoter-proximal pausing, where transcription stops immediately downstream of the TSS, may occur [Jonkers and Lis, 2015], often for long periods of time. This pausing, typically around 30-60 nucleotides downstream of the TSS, allows polII to wait in a 'poised' state for a signal to restart transcription, which may rapidly occur. This mechanism has been commonly observed in the transcriptional cycle of developmental genes, such as *Hsp70* in drosophila, which rely pathways controlled by stimuli [Burgess, 2012, Lis, 1998]. PolII pausing appears to be coupled remodelling the local three dimensional genomic structure, keeping the promoter available for further regulatory cues [Gilchrist and Adelman, 2012, Gilchrist et al., 2010]. Such mechanisms allow for flexibility in the control of transcription, thus improving the cells ability to transcribe according to exact requirements in time and space.

Since this project primarily concerns transcriptional initiation events from the TSS region, the architecture of promoter regions and how they support transcriptional regulation will be discussed in the following section, before describing the regulation of gene expression by regulatory elements outside of the core promoter region.

1.2.3 Promoter architecture

Studies based on CAGE data have classified promoters into different categories [Lenhard et al., 2012] [Carninci et al., 2005]. Type I promoters are 'sharp', generally depending on a single TSS controlling the transcription of the gene and are generally associated with tissue specific expression. Type II promoters are 'broad' and are associated with multiple TSS spread out across the promoter region of the gene. Such genes are generally associated with ubiquitous expression, or 'housekeeping' genes. A third category is Type III promoters, or those genes which are developmentally regulated and are generally associated with polycomb repression marks.

The CpG island (CGI) is associated with ubiquitous expression of genes [Deaton and Bird, 2011]. These are regions where CpG di-nucleotides (C followed directly by a G, linked by a phosphate bond) are overrepresented. Estimates suggest that CGIs are typically found in around 40% of promoters of genes [Fatemi et al., 2005]. According to the promoter classification described above, Type I genes are generally associated with a lack of CpG island, Type II and Type III genes are associated with CpG islands; developmentally regulated genes often have large CpG islands overlapping the main body of the gene [Lenhard et al., 2012].

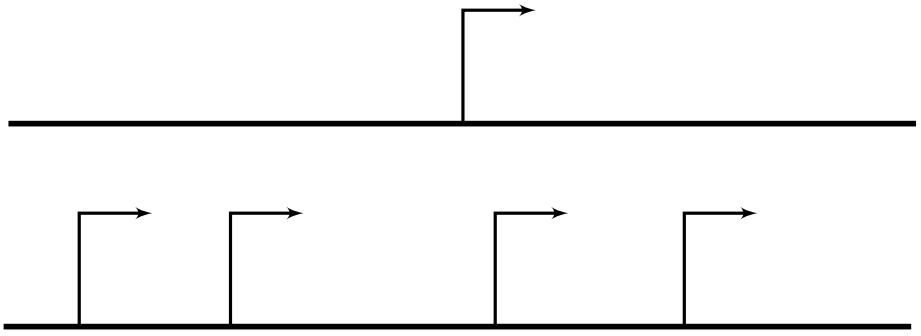


FIGURE 1.5: Focussed transcription vs broad transcription. Focussed or sharp transcription is associated with regulated promoters and is characterized by the existence of a single TSS driving the initiation of transcription. This TSS is generally limited to a small number of nucleotides. Dispersed transcription is commonly associated with constitutive promoters and involves the existence of several weak transcription start sites over a broad range of nucleotides upstream from the gene. Broad or dispersed transcription is often associated with CpG islands, whereas focussed transcription is generally not associated with CpG islands.

The TATA box motif sequence, often associated with cell-specific expression, is the most well-studied. Early studies have argued the essentialness of the TATA box as a requirement for the initiation of transcription [Mathis and Chambon, 1981] although in reality it is associated with only 10-15% of human genes. The TATA box acts as a binding site for TATA-binding protein (TBP). The binding of TBP together with a group of transcription factors and RNA polymerase form the pre-initiation complex, although a recent study suggests TRF2 and not TBP is more important in ribosomal protein coding genes [Wang et al., 2014b].

Whilst TATA is the most well known core promoter element in human, many promoters lack TATA binding sites, and instead contain other elements binding TBP. Common examples include the initiator motif (Inr), TFIIB recognition elements (BRE), downstream promoter element (DPE) and motif ten element (MTE) (see [Roy and Singer, 2015] for a recent review). The BRE consists of elements which may locate either upstream or downstream of the TATA box. They are highly common in eurkaryotic promoters in general, but are over-represented in TATA-less promoters compared to TATA-containing promoters. The downstream promoter element (DPE) was first discovered and characterized in the Drosophila genome and interacts with Inr. MTE (motif ten element), which also interacts with Inr, is often found upstream of the DPE element, but functions independently of DPE [Roy and Singer, 2015].

Studies suggest that different combinations of core promoter elements may be associated with directing the initiation of expression of tissue specific and/or developmental regulation genes [Decker and Hinton, 2013, Müller and Tora, 2014, Roy and Singer, 2015]. In particular, a recent study suggests that rather than a simple promoter architecture, the expression of many ubiquitously expressed genes are actually controlled by multiple overlapping 'selection codes' [Haberle et al., 2014]. This thus raises further questions about the role of the core promoter and exactly how it initiates transcription.

1.2.4 The bidirectional nature of promoters

Bidirectional transcription, whereby transcriptional initiation occurs from both the sense and anti-sense orientation within a tightly spaced region, typically less than 1000bp, has been observed at a large proportion of genes [Trinklein et al., 2004]. Such bidirectional genomic organization often represents the scenario whereby two annotated TSS transcribe in opposite directions, labelled the bidirectional promoter [Adachi and Lieber, 2002]. Such transcriptional coupling allows for potential shared regulation of expression, as observed through their co-expression in a cell-type restricted manner. Supporting this claim, recently [Scruggs et al., 2015] showed that the TSS of bidirectional promoters clearly demarcate a larger than expected nucleosome depleted region (i.e an exposed region of DNA between the TSS), which is highly enriched for transcription factor motifs.

Since the discovery of bidirectional promoters, a class of anti-sense non-coding transcripts called promoter upstream transcripts (PROMPTs) has emerged. These are of low detection rate due to their highly unstable properties, making them quickly degraded by surveillance mechanisms in the cell [Preker et al., 2008]. A recent study of nascent transcriptional initiation events (thus with the ability to detect PROMPTs pre-degradation) by [Duttke et al., 2015] classified promoters according to divergent, where there is an annotated gene in one direction and no annotation in the reverse direction, bidirectional, where there is an annotated gene in both directions, and unidirectional, subsequently suggesting that human promoters generally act unidirectionally. A further study directly challenged this view, suggesting that promoters are generally bidirectional [Andersson et al., 2015], and so further experimental evidence is required to fully understand bidirectional transcription. Furthermore, it does not change that the functional role of anti-sense unstable transcription remains unknown.

Overall, it is clearly apparent from the above that whilst the basic idea of a core promoter is universally understood with the same basic structure in mind, in reality the core promoter is a very flexible and complex piece of transcription machinery, which broadly varies in terms of its architecture and subsequent impact on initiation.

1.3 Cis and trans regulation of gene expression

Promoter binding is essential to the expression of a gene, but other sequences may be required to regulate the levels of transcription beyond their basal levels which may be achieved from binding at the core promoter alone. These include cis-regulatory elements, defined as sequences in the vicinity of the gene promoter which contain transcription factor binding sites which, when bound to by required combinations of transcription factors, produces the desired level of transcription from the gene. These transcription factors, together with other off-location elements such as micro RNAs affecting the expression of the gene, are known as trans-regulatory elements, and both working together are crucial in the regulation of transcription [Dowell, 2010].

The class of cis-regulatory elements could refer to proximal promoters, locus control regions, silencers, enhancers or insulators, the most common cis-regulatory element being the enhancer (next section). Whilst some cis-regulatory elements have been observed to be highly conserved across species, that is they exhibit few changes in sequence compared to surrounding non-functional sequence between species, many of these sequences exhibit high levels of changes, actually resulting in low levels of observed conservation (i.e. they are turned over, a property discussed in the next section and later sections in this chapter) [Meader et al., 2010, Villar et al., 2015].

1.3.1 Enhancers and long range promoter interactions

The main source of regulatory information outside of the promoter is concentrated in genomic regions known as enhancers; transcription factor binding sites that have an up-regulatory effect on the expression levels of the one on which it acts. These enhancer regions are defined in a distinct manner from promoter regions because they are not necessarily present on the immediate upstream region of the gene, but instead may act in a distal manner [Bulger and Groudine, 2011], generally (but not always) located on the same chromosome as the gene and can be upstream, downstream or within the gene body itself [Spilianakis et al., 2005]. Another characterizing feature is their ability to act in an orientation independent manner, with their targets either up- or

downstream of their binding sites. Furthermore, A single enhancer may affect multiple genes and may be located in regulatory regions surrounding a gene on which it does not necessarily act upon. A classic example of a gene undergoing long range enhancer interactions in cis is the *SHH* (sonic hedgehog) gene, with which multiple enhancers have been shown to interact from a distance [Anderson and Hill, 2014].

Much effort has been made to identify and characterize enhancer sequences genome-wide. Early studies have attempt to identify regions of conservation in non-coding sequence; in particular, highly conserved non-coding regions (HCNEs) have been seen to cluster in the vicinity of genes involved in developmental regulation [Nelson and Wardle, 2013, Woolfe et al., 2004]. Enhancers may also detected through the presence of DHS I hypersensitive sites, 'open' regions of DNA exposed for accessibility to transcription factors and cofactors [Thurman et al., 2012]. Whilst conservation and DNase I hypersensitivity has allowed for highly useful genome-wide characterizations of enhancers, it must be noted that the two sets do not necessarily overlap due to the high turnover of regulatory sequences [Meader et al., 2010] and the low specificity of open regions. Combining this information with further evidence such as the presence of coordinated modifications to histones at certain sites has also proved to be a useful avenue in improving the detection of enhancer regions [Rada-Iglesias et al., 2011].

Despite recent studies attempting to address the exact mechanism of how enhancers interact with their targets, the complete mechanism is still not yet fully understood. The most common mechanism by which enhancers are thought to interact with a specific promoter is through looping in the DNA, so that the enhancer and the promoter are brought within proximity of one another within the nucleus [Deng et al., 2012, Doyle et al., 2014, Tolhuis et al., 2002], and appears to involve direct contact between the promoter and enhancer genomic regions [Pombo and Dillon, 2015]. What set of rules determining precisely which pairs of enhancers and promoters interact via looping corresponds to a complex set of parameters and is not simply down to closest proximity [Whalen et al., 2016]. The looping interactions do appear to be restricted to within larger (with an average of 1MB) domains of DNA referred to as topologically associated domains (TADS) [Dixon et al., 2012], which themselves can be proximal within

the nucleus on a higher order level, potentially facilitating longer range interactions [Fraser et al., 2015]. As observed in chromatin conformation data, putative interactions through proximity is more commonly observed within these higher order domains than it is between, and remain stably in place during development [Ghavi-Helm et al., 2014].

Enhancers bind polII to produce a class of RNA commonly referred to as enhancer RNAs (eRNA). Compared with mRNA, eRNA transcripts are, like the so-called PROMPTs produced by anti-sense to promoter regions, typically unstable and quickly degraded by certain 'surveillance' complexes [Andersson et al., 2014a]. The discovery that enhancers may be defined by their bidirectional transcriptional signatures, with eRNA transcripts produced in a constrained sense-antisense configuration has opened further avenues for enhancer detection [Kim et al., 2010]. To this end, recent work by the FANTOM consortium has utilized this observation by mapping enhancers across the genome based on the presence of balanced bidirectional CAGE peaks as a signal of active enhancers and identified around 43,000 candidate enhancers over 808 CAGE libraries in human [Andersson et al., 2014b].

It is important to understand that whilst distinctions are made between what is an enhancer and what is a promoter in research, their contrast in reality is blurred. Indeed, promoters have been seen to act as weak enhancer elements and enhancers may act as alternative promoters. A recent study based on the STARR-seq protocol has identified sequences with enhancer 'potential' [Zabidi et al., 2014], observing that may of these sequences overlap gene promoters, particularly those found activated in the context of a housekeeping-type promoter. Indeed, much work is still being carried out to distinguish actual enhancers from other regulatory elements such as alternate transcription start sites, leading to the view of a 'unified architecture' when studying regulatory elements [Andersson, 2015].

In contrast to an enhancer, a silencer is a transcription factor binding site on the DNA that when bound by a repressor protein, prevents the binding of Polymerase II to the core promoter, whereby reducing the level of or completely silencing transcription for the gene on which the silencer acts on [Kolovos et al., 2012], although the functions

carried out by enhancers and promoters are not always as clearly distinguished as their name suggests [Reynolds et al., 2013]. Co-activators can increase the levels of gene expression by binding to an activator with a DNA binding domain, which binds to the DNA. Co-activators have regulatory roles in transcription, including elongation, RNA splicing, degradation of co-activators-activator complexes. An example of a co-activator is [Chen and Dent, 2014], which works to modify the physical genome structure.

In the next section, the epigenetic regulation of gene expression is discussed.

1.4 Epigenetic regulation of gene expression

Epigenetic factors refer to heritable changes as a result of regulatory signals occurring outside of the DNA itself [Goldberg et al., 2007, Jaenisch and Bird, 2003], which may have consequences in gene expression [Bernstein et al., 2007]. DNA can be thought of as a linear 'string' which is wrapped around protein complexes called nucleosomes. The nucleosomes consist of four histone proteins - two each of H2A, H2B, H3 and H4, forming a histone octomer [Hughes and Rando, 2014]. DNA wraps around each nucleosome 1.7 times with a distance between nucleosomes of approximately 147 bp and the unwrapped DNA between nucleosomes being termed as 'linker' DNA. The resulting 'beads on a string' formation then folds up into higher order structures called chromatin domains. Chromatin domains, recently reviewed in [Chen and Dent, 2014], are important in allowing all of the DNA to compact within the nucleus cell. Once the DNA has folded into a chromatin structure, some parts of the DNA from different chromosomes will become in contact with one another, facilitating long range interactions [Lieberman-Aiden et al., 2009]. Thus, changes in chromatin structure within the cell plays a highly important regulatory role.

Nucleosome binding by transcription factors [Ballaré et al., 2013] and the position and occupancy of the nucleosomes play a crucial role in the regulation of gene expression [Lenhard et al., 2012, Struhl and Segal, 2013]. In particular, the nucleosome directly after the TSS, the +1 nucleosome, has important regulatory functions [Nock et al.,

2012, Rhee and Pugh, 2012] and is typically subject to thousands of combinations of covalent modifications, briefly reviewed in the next section.

The region upstream from the TSS is known as the nucleosome free region (NFR). These regions are formed through the binding of pioneer transcription factors, which result in the rearrangement of nucleosomes, revealing regulatory motifs which become easily accessible to transcriptional-activation associated transcription factor complexes [Iwafuchi-Doi and Zaret, 2014]. Thus, changes in the compactness of chromatin can result in a gene switching between an on and off state of expression; genes present in highly compacted areas of the DNA appear silenced whilst regions sufficiently accessible by transcriptional machinery tend to be more actively transcribed.

1.4.1 Histones

Covalent post-transcription modifications (PTM) occur on histones, which may affect transcription through modifications of chromatin structure or the recruitment of other proteins. These modifications usually occur on one of the N- or C-terminal tails, although they may also occur at globular domains [Campos and Reinberg, 2009]. The first modification discovered was that of acetylation [Phillips, 1963], and further modifications include methylation, phosphorylation and ubuitination [Zentner and Henikoff, 2013]. As mentioned, histone modifications with a direct regulatory impact on transcriptional initiation are typically found on the +1 nucleosome.

Genome-wide chromatin immunoprecipitation (ChIP) based techniques currently form the gold standard for mapping histone modifications across the genome[Consortium et al., 2012, Mikkelsen et al., 2007, Wang et al., 2008]. In particular, trimethylation of H3 lysine 4 (H3K4me3) and acetylation of H3 lysine 27 (H3K27ac) at the promoters of genes is associated with active transcription and trimethylation of H3 lysine 27 (H3K27me3) at the promoter is associated with the repression of transcription [Creyghton et al., 2010, Jenuwein and Allis, 2001, Li et al., 2007]. The trimethylation of H3K27me3 marks is catalysed by the polycomb group proteins (PcG), forming repressive complexes PRC1 and PRC2 [Ku et al., 2008, Margueron and Reinberg, 2011].

Some genes contain both active and repressive marks, known as bivalent genes. They are generally developmental genes and their bivalency is thought to represent genes which are held in a poised chromatin state ready for transcription to take place [Ku et al., 2013, Voigt et al., 2013] are thought to be essential for defining cellular identity and function [Lesch and Page, 2014].

Some marks are found along the body of the gene rather than the promoter of the gene, including trimethylatino of H3 lysine 36 (H3K36me3), whose levels are correlated with active transcription [Barski et al., 2007, Hahn et al., 2011] and trimethylation of H3 lysine 9 (H3K9me3), which is associated with transcriptional silencing [Mikkelsen et al., 2007, Schotta et al., 2004].

Putative enhancers appear to be marked with H3K4me1, often in combination with H3K27ac and H3K27me3 according to transcriptional activity [Rada-Iglesias et al., 2011, Zentner et al., 2011], and are often used to define enhancers [Villar et al., 2015]; in particular a high ratio of H3K4me1 to H3K4me3 is often seen as an enhancer mark [Robertson et al., 2008], although this is far from what can be thought of as a definitive way to distinguish enhancers from promoters [Andersson, 2015].

The scope for possible histone modifications is enormous, both in terms of the diversity of modification sites and the range of modifications available [Tan et al., 2011], making the analysis of the histone modifications very complicated, however certain patterns do exist. Table 1.1 outlines some of the effects that certain histone marks are known to have on expression. Attempts to chart histone modifications on a genome wide scale [Consortium et al., 2012, Zhou et al., 2011] have questioned the existence of a histone ‘code; the idea that groups of histone modifications together act on the transcription in a gene in a predictable manner [Rando, 2012, Wang et al., 2008], however the complexity of the full situation and the number of possible combinations (more than the number of nucleosomes in the genome) make it difficult to determine either way [Keung et al., 2015].

Mark	Transcriptional impact
H3K4me3	Active
H3K27me3	Repressive, poised
H3K27ac	Repressive
H3K36me3	Active
H3K9me3	Repressive

TABLE 1.1: Histone modifications and effect on gene expression [Barski et al., 2007, Benevolenskaya, 2007, Koch et al., 2007]

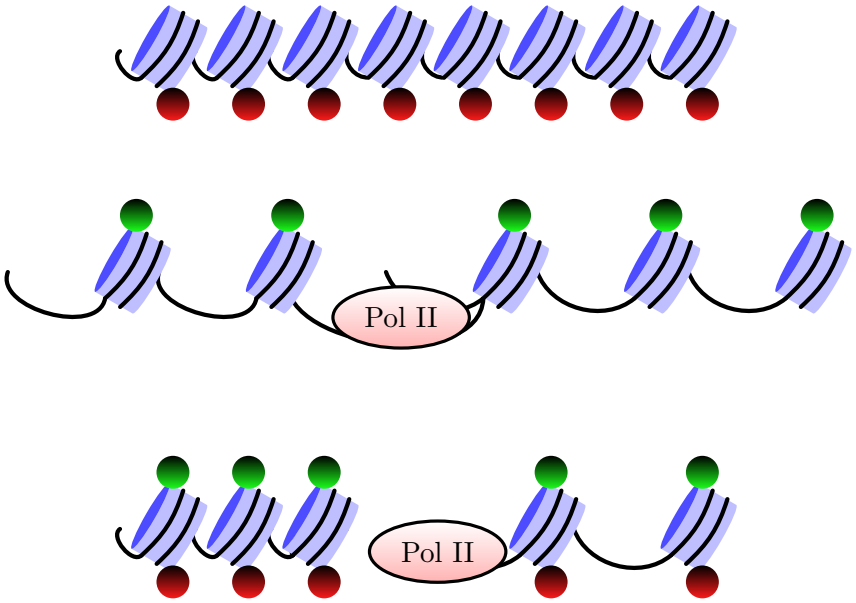


FIGURE 1.6: Diagram showing a bivalent promoter region, active promoter region, and poised promoter region. Green dots represent histone modifications associated with gene activation, such as H3K4me3, whilst red dots represent histone modifications associated with gene repression, such as H3K27me3. (Image adapted from [Kurdistani and Grunstein, 2003], Figure 4)

1.4.2 DNA methylation

DNA methylation refers to the addition of a methyl group to cytosine nucleotides on DNA. It generally acts as a silencing mark, leaving nearby genes in an 'off' state. Methylation occurs at CpG dinucleotides which are typically depleted in the genome [Li et al., 1992] (around 60-90% are methylated in mammals [Jabbari and Bernardi, 2004]), although it generally does not occur at CpG islands.

In adult cell types, CpG methylation is as a general rule fixed, stable and irreversible; marks are laid down in early developmental stages [Smith and Meissner, 2013]. They

are crucial to genomic imprinting, x-chromosome inactivation and chromosome stability [Bird, 2002]. In mice, it has been observed that DNA methylation patterns act to establish the lineage commitment of cell types, but is subsequently erased at around E8.5 and E11 before re-establishment as part of an extensive reprogramming phase [Seki et al., 2005]. It appears that some gene promoters associated with the germline and development have a feed-loop mechanism, which requires DNA methylation to prevent damage by transposable elements and genomic instabilities [Hackett et al., 2012]. Therefore, these genes may not be directly repressed by methylation, but rather methylation is coupled with other regulatory mechanisms. However, it is still not definitive whether the observation of methylation is a feature of silencing, or vice-versa.

Methylation is important in brain development [Lister et al., 2013] and has been linked with the ageing process [Horvath, 2013] and diseases such as cancer, where hypermethylation (the accumulation of methylation in a given region) may occur at CpG islands, causing the silencing of tumour suppressor genes [Esteller, 2007].

1.5 Regulation of gene expression at the post-transcriptional level

During the translation process, mRNA is translated at the ribosome, which involves initiation, elongation and termination of protein synthesis. Regulation can occur at numerous stages [Kong and Lasko, 2012], resulting in differences between the level of transcribed mRNA and the final levels of protein achieved. Post-transcriptional modifications explain a large proportion of the variance in protein, although transcriptional initiation is believed to play the greatest role in this [Li and Biggin, 2015].

Alternative splicing results in the the same gene being able to produce multiple possible gene products, allowing for the scenario where there are approximately 20k genes in the genome but around 80-90k gene products [Brett et al., 2002, Roy et al., 2013]. A significant proportion of genes in the genome can be alternatively spliced. Alternative transcripts are often controlled by separate promoters, and may be distinguished using

whole-transcript level sequencing methods such as RNA-seq, where switches in distributions of isoform usage has been observed in genes comparing tumour samples with controls [Sebestyén et al., 2015].

The complete mRNA transcripts are moved from the nucleus to the cytoplasm where it may be translated into a protein by a process known as nuclear export [Köhler et al., 2007]. Transcripts not properly processed are targeted for degradation. mRNA transcripts are protected from degradation by the capping process. Capping on the 5' end of the mRNA may also occur after it has been exported to the cytoplasm, as well as decapping to aid the degradation of the mRNA [Bentley, 2014]. Degradation occurs when exonucleases gradually shorten the polyA tail. mRNA degradation is known to play crucial roles in post-transcriptional regulation of gene expression; mRNA removed from the area where it is transcribed aids in the process of switching off a gene. It is due to degradation that the levels of protein captured in a cell may not necessarily be the levels of protein translated from mRNA.

Alternative polyadenylation refers to the fact that some protein coding genes have more than one possible site for the poly-A tail to be added, thereby changing the location of the 3' end of the mRNA [Tian et al., 2005]. Sometimes this changes the protein, although usually the result is simply a shortened 3' UTR region [Shen et al., 2008].

From all of the layers upon layers of factors influencing the expression of a gene one can begin to appreciate the extreme difficultly in explaining the regulatory processes involved in each individual gene. Add this to the fact that genes themselves are regulators and many transcribed RNA have multiple regulatory roles.

1.6 Gene expression quantification

1.6.1 Pre next generation methods of capturing gene expression

Prior to the more recent advent of high through-put sequencing technologies, northern blotting was the main method used to determine the relative expression levels of a

specific RNA, by means of electrolysis to separate out strands of RNA and using probes to bind to elements of interest [Alwine et al., 1977]. It is similar to a Southern blot, which detects DNA instead of RNA [Southern, 1975]. The method was time-consuming and heavily limited on the number of genes one could analyse in a single study. Blotting methods were significantly improved upon through the advent of qRT-PCR - reverse transcription (generating a DNA template from the RNA), then quantitative PCR (amplify resulting cDNA, estimate number of original copies. Very sensitive), which resulted in faster and more accurate quantification.

Microarrays quickly became a very attractive method for quantifying the expression of multiple RNAs in parallel [Lashkari et al., 1997, Schena et al., 1995]. Genetic material is placed in a series of probes, one for each expression element. Gene expression intensity is captured by observing the fluorescence of the spot on the chip for a gene. Generally superseded by RNA-seq in recent years, due to RNA-seq's ability to record over entire genome without prior knowledge of genes. Microarray is a hybridization based method of capturing gene expression. A microarray chip consists of a set of spots, one per gene (or other DNA based element), each containing small amounts of DNA sequence for that gene.

1.6.2 Genome wide quantification of transcription

Use of modern high throughput technologies allows for genome wide estimates of transcriptional output. Two distinct categories of transcription sequencing include shotgun based methods, where the mRNA is broken up into fragments, sequenced and mapped back to the reference genome and tag based methods, where short sequences at a specific position of the transcript is sequenced.

1.6.3 RNA-seq

In recent years the most common method employed is RNA-seq, which employs massively parallel sequencing to measure RNA abundance and has largely replaced microarrays due to the lack of required transcriptomic knowledge prior to sequencing

[Mortazavi et al., 2008, Wang et al., 2009]. RNA-seq methods employ shotgun sequencing, so has the advantage of being able to quantify expression across an entire transcript, obtaining information on expression levels for each exon and allowing for the detection of alternative splicing events. The protocol varies, although generally involves fragmentation, followed by conversion of RNA into cDNA, second strand synthesis, ligation of adapter sequences at the 3' and 5' ends and final amplification before the tags are sequenced and mapped back to the reference genome [de Klerk et al., 2014]. A big disadvantage is that RNA-seq often under-represents tags around the 5' and 3' transcript ends [Roberts et al., 2011], thus tag based methods (see below) are often preferred for capturing the locations where transcription starts [Forrest et al., 2014]. Furthermore, RNA-seq is not necessarily strand specific since sequence orientation may be lost during random-primed cDNA synthesis [Roberts et al., 2011] although strand-specific protocols have been developed in recent years [Armour et al., 2009, He et al., 2008, Parkhomchuk et al., 2009], allowing one to distinguish between sense and anti-sense transcription, which may supply important regulatory roles [Faghihi and Wahlestedt, 2009].

1.6.4 Tag-based methods

Tag-based methods capture millions of reads across the genome from a specific part of the transcript, with the aim of generating a fine-scale map of expression regulation at base-pair resolution. They have the feature whereby they are inherently stranded as a result of features in the protocol. The two main types of tag-based methods capture tags from either the 5' end or the 3' of the transcript, the two most common methods, CAGE and SAGE, which will be discussed below.

1.6.5 CAGE sequencing

Cap analysis of gene expression (CAGE) [Forrest et al., 2014, Kodzius et al., 2006, Shiraki et al., 2003] captures the 5' cap structure of the RNA, based on a cap trapping method [Carninci et al., 1996, Takahashi et al., 2012]. The cap trapping protocol

involves the biotinylation of molecules which contain pairs of hydroxide groups on adjacent bonded carbon atoms. Briefly, mRNA is reverse transcribed and then an adapter is ligated to the 3' of the resulting cDNA using random primers, which is used to facilitate the cleavage of the DNA with the restriction endonuclease EcoP15I at a specific recognition site, resulting in a short tag around 25-27 nt long. In the standard CAGE protocol, there is an amplification step using PCR prior to the sequencing of the tags, which are then mapped back to the reference genome. The number of tags overlapping a given nucleotide position then provides a quantitative estimate of expression at that given location.

The Heliscope CAGE method avoids this amplification step altogether and involves only three main steps [de Hoon and Hayashizaki, 2008, Kanamori-Katayama et al., 2011]. These are reverse transcription into cDNA followed by 5'-cap trapping and applying a poly(A) tail to the 3' end, after which sequencing can begin. This method has a distinct advantage over the basic CAGE method, as the removal of the requirement to reverse transcribe a second time, amplification and ligation removes sources of possible bias and therefore provides a more exact quantification of the mRNA levels at the location of the TSS, allowing for more detailed expression analyses.

The first genome-wide sequencing and annotation of full-length cDNAs in mouse was made by the FANTOM Consortium [Okazaki et al., 2002], who have subsequently applied CAGE to a large number of large scale projects, identifying different types genome wide transcriptional initiation in a range of tissues, cell types and species, the data of which forms the basis for the current project [Andersson et al., 2014b, Forrest et al., 2014].

1.6.6 SAGE sequencing

In contrast to CAGE, which captures the 5' end of the transcript, serial analysis of gene expression (SAGE), sequences tags from the 3' end of the transcript and relies on the presence of the poly(A) tail [Hu and Polyak, 2006, Nielsen et al., 2006]. In brief, the protocol works by washing thymine nucleotides attached to magnetic beads over

the cellular contents. The resulting tagged transcriptions are then reverse transcribed, amplified and sequenced. These sequences are mapped back to a reference genome, resulting in genome wide gene expression estimations, which may be thought of as a high-throughput version of microarrays in terms of its expression output [Lin and Li, 2005].

Whilst both CAGE and SAGE produce sequence tags representing RNA fragments present in an mRNA sample, they capture fundamentally different information. Because SAGE only captures tags based on the detection of a recognition site in the polyA tail of an RNA transcript, it does not give any information about alternative splicing, but just the overall expression level for that transcript. Furthermore, if there is no polyA tail present then the expression of that transcript will be missed altogether, affecting approximately 1000 transcripts, or 1% of the genome [Saha et al., 2002]. In conclusion, the main advantage of CAGE over SAGE in the context of of analysing the regulation of expression for a particular gene is the ability to obtain information about the number of and location of alternative promoters contributing the overall expression of that gene.

Whilst CAGE is generally considered an efficient and robust method of capturing 5' capped transcripts, it does have a disadvantage known as 'exon painting' whereby transcripts are sometimes mapped onto the exon junctions, skewing estimates of TSS location and expression levels [Zhao et al., 2011]. This is sometimes been attributed to recapping events [Forrest et al., 2014], although this artefact is discussed and analysed further in Chapter 3 of this thesis, where it is seen that promoter expression is correlated with observed levels of exon painting within a gene.

1.6.7 Methods for capturing nascent transcription

The sequencing of 5' end of RNAs prior to processing steps after polymerase II engagement, known as nascent RNAs, can be achieved through methods called GRO-seq (Global Run-On Sequencing) [Core et al., 2014, 2008] and PRO-seq (a nucleotide resolution version of GRO-seq) [Kwak et al., 2013]. These methods capture those nascent

RNAs which are actually engaged with Pol II at a given moment, including those attached to Pol II that are undergoing different transcriptional steps, i.e. intitation, pausing, productive elongation and termination. Sequencing of nascent transcripts allows for the detection of transcription start sites from unstable transcripts such as eRNAs and PROMPTs, which are often difficult to detect in CAGE as a result of degradation processes.

A variant of GRO-seq is GRO-cap (and PRO-cap), which only captures 5' transcription start sites as a result of its 5' cap enrichment step [Kwak et al., 2013]. GRO-cap has been utilized to compare the initiation rates of thousands of enhancers and promoters genome-wide in GM12878 and K562 cells in a fashion unbiased by degradation complexes, finding striking architectural similarities between them, including similar spacing of divergently initiating promoters, common core promoter element frequencies and tightly regulated nucleosome positioning [Core et al., 2014].

Recently, NET-seq has been developed, which sequences nascent RNA from polymerase that has either backtracked and/or arrested through the identification of the 3' end of the transcript within the Pol II active site. [Churchman and Weissman, 2012] [Nojima et al., 2015]. The method is nucleotide resolution, although may only detect transcripts around 30 nucleotides or greater beyond the TSS [Nojima et al., 2015].

1.6.8 DNase I hypersensitivity

DNase I hypersensitivity sites (DHSs) indicate areas of open chromatin whereby the DNA is accessible to DNase I cleavage enzymes [Thurman et al., 2012]. Mapped sites in the vicinity of the body of a gene, as well as within its exons or introns, may be representative of cis-regulatory modules affecting the regulation and hence expression of the gene. Such sites have been mapped extensively as part of the ENCODE project via a method called DNASE-seq [Thurman et al., 2012]. This study used 125 different human cell types obtaining around 2.9 million distinct DHSs. Another more recent and efficient method for capturing hypersensitive regions is ATAC-seq, where hyperactive Tn5 transposase inserts itself into exposed, nucleosome free regions of DNA, with its

active properties allowing for the cutting of the DNA at these sites [Buenrostro et al., 2013, 2015].

DHS sites mark the locations of actively bound cis-regulatory elements, including promoters, enhancers, insulators, silencers and locus control regions [Thurman et al., 2012], making the resultant maps an important source of gene regulatory information to draw upon. For example, a recent study mapped the gain and loss of DHSs as cells progress from embryonic stem cells to terminal fates [Stergachis et al., 2013], and DHS classifications have been used to characterize RNAs according to their nuclear stabilities [Andersson et al., 2014a]. However, an open question in the area refers to the turnover of such sites and whether many of them represent truly functional sequence, particularly in the absence of a conservation signal, that exhibits high rates of turnover [Meader et al., 2010, Young et al., tted].

1.7 The paradoxes of information content and complexity

How does one define the complexity of an organism? For example, the number of cell types has been alluded to as an indicator of morphological complexity [Chen et al., 2012]) the range of proteins produced within the organism [Schad et al., 2011]. Whilst the size of the genome of an organism in terms of base pairs varies considerably between species, it does not necessarily correlate with the perceived complexity of the phenotype, a phenomenon commonly referred to as the c-value paradox [Gregory, 2001]. Indeed, one of the largest animal genomes ever discovered is that of a locust, which has 17302 predicted genes [Wang et al., 2014a]. However, a lot of this genome size can be explained by the fact that 60% of its genome is made up of repetitive DNA.

A way around this phenomenon is to count the number of genes present in the genome. The g-value paradox is the name given to the fact that organisms vary widely in the number of genes in their genomes [Hahn et al., 2002], from simple prokaryote unicellular organisms such as E-coli, with around 4000 genes, to eukaryote multicellular organisms

such as humans with around 25000 genes. In both of these organisms, genes are heritable units coded within DNA with similar mechanisms by which their expression is controlled.

Many more involved attempts at measuring genome complexity have been attempted. For example, studies in biochemistry look at the 'energy per gene' [Lane and Martin, 2010], which has been shown to be larger in multicellular eukaryotes than 'simplistic' prokaryotes. However the fact remains that it turns out that simply observing simple characteristics of the genome is a very poor indicator of the overall complexity of that organism.

A better perception of genomic complexity may be perceived through observed interplay of regulatory dynamics within an organism [de Mendoza et al., 2016]. Indeed, many signalling pathways and transcription factors are highly conserved and shown to be present through all the way down through to unicellular metazoan organisms [Fairclough et al., 2013, Sebé-Pedrós et al., 2011, 2012]. Furhermore, two mechanisms widely seen to represent shifts in complexity, the evolution of alternative splicing, which allows for a single gene to code for many different products, and the use of long intergenic non-coding RNAs (lincRNA), has also recently been suggested to have been present in *Creolimax*, a unicellular relatives to animals [de Mendoza et al., 2016]. The same study suggests that differences may lie in the differences in how these systems are regulated, for example the increasing dependence on cell type specific expression by these lincRNAs in diverged multicellular organisms [Gaiti et al., 2015]. So, it is likely that the evolution of complexity in the organism lies in the mechanisms involved in the regulation and control of the fundamental tools deeply conserved through lineages, and this 'extra' regulatory complexity may provide an explanation as to why some organisms develop a greater diversity of phenotypes.

1.8 Evolution and the variation of gene regulatory mechanisms

It has long been suggested that phenotypic differences between species are the cause of changes in gene expression as a result of structural changes in the genome. King

and Wilson [King et al., 1975] famously observed that regulatory changes must be responsible for gene expression differences between human and chimpanzee, who share 96% of their DNA sequence, and since then evidence has been accumulating in favour of between lineage gene expression divergence [Gilad et al., 2006, Khaitovich et al., 2006, Tirosh et al., 2006]. Raff and Kaufman [Raff et al., 1991] argued that mutations affecting regulatory regions were less likely to be deleterious than changes in protein coding genes, and it has been seen that cis-regulatory changes are important drivers of diseases such as cancer [Ongen et al., 2014].

A key question often asked is: how do changes in sequence cause changes in gene expression variation, and therefore phenotype and disease? Few studies directly address this question, although one study found that sequence changes at the core promoter may not correlate well with expression divergence [Tirosh et al., 2006]. Another study found that in the case of human mouse macrophage response, divergence in gene expression was found to be negatively correlated with divergence in sequence [Schroder et al., 2012], suggesting that although promoter sequence is evolutionary conserved between species, patterns of gene expression are found to be highly divergent. One explanation is that although the promoter itself may be conserved, there are a number of cis- and trans- acting elements influencing the transcription of the gene. Indeed, it is frequently thought that enhancers appear to be more responsible for divergence in cis-regulation than promoters [Villar et al., 2015, Wittkopp and Kalay, 2012]. Enhancers appear to have lower pleiotropy, where higher pleiotropy means that a single change has an effect on multiple distinct phenotypes [Stearns, 2010]. This is rationalised by the tissue-specific nature of enhancer sequences, as a change in a random tissue is unlikely to affect the phenotype via significant changes in gene expression [Liao and Weng, 2012]. Furthermore, enhancers have been shown to exhibit a degree of redundancy, whereby the loss of function in a given enhancer could be replaced or compensated by another enhancer [Barolo, 2012], often in response to an environmental stimuli [Bothma et al., 2015]. For this reason, it is thought that mutations within them are likely to survive into future generations, giving rise to greater polymorphism/divergence than mutations elsewhere [Wittkopp and Kalay, 2012].

Through the above reasoning, evolutionary forces may be driven through the gain and loss, referred to as the 'turnover', of regulatory elements over time [Frith et al., 2006]. Studies suggest that this turnover is substantial, by observing that as phylogenetic distance from humans increases, there is a dramatic drop off in constraint in sequence, and it is these changes in regulatory features that are driving evolution. [Meader et al., 2010]. One study saw that during early vertebrate evolution, regulatory gains were enriched around transcription factors and developmental genes [Lowe et al., 2011], and a more recent study looked at the fitness consequences of point mutations in the human genome, estimating that since the divergence between human and chimpanzee, 4.2–7.5% of nucleotides in the human genome have influenced fitness [Gulko et al., 2015].

More and more studies are focusing on linking exactly when and how divergences in cis- and trans- regulatory factors contribute to the observed divergence in the resulting expression of a gene, and in turn the effect this has on phenotype [Wittkopp and Kalay, 2012]. For example, changes in cis-regulation have been revealed as a driver of evolution in drosophila, as observed by changes present in wing colour [Gompel et al., 2005]. A classic example is the *HOX* gene paradox [Prince, 2002], which questions how the substantial diversity observed in the anatomical features of body patterning is controlled at the gene expression level, with a recent study revealing a mechanism involving weak interactions of *HOX* gene proteins with transcription factor binding sites [Crocker et al., 2014].

Finally, the advent of next generation sequencing allows for the systematic genome-wide scale identification of eQTLs [Battle et al., 2014, Westra et al., 2013]. An eQTL is a region of the genome containing DNA sequence variants that influence the expression level of one or more genes[Albert and Kruglyak, 2015, Pai et al., 2015]. Many quantitative trait loci have been localised to regions that don't apparently contain protein coding sequence - it is thought that these may be transcriptional regulatory variants; such eQTLs are increasingly being associated with human disease and phenotypic traits [Dimas et al., 2009, Montgomery and Dermitzakis, 2011, Veyrieras et al., 2008]. Furthermore, genome-wide association studies have also concluded the diversity of traits enriched around the core promoter region [Kindt et al., 2013].

It is clear that gene expression evolution is not a linear process, and varies depending on species, organs, lineages and chromosomes [Brawand et al., 2011], and within the genome a high turnover at regulatory elements is observed, providing a mechanism for regulatory innovations and thus a rich source of variation in regulatory complexity.

1.8.1 The upper limit of complexity in the human genome

Studies on the evolution of regulatory complexity speculate on the so called 'upper' limit of complexity in the genome - how much regulation can a single gene potentially undergo? For example, sequence constraints in the genome limit exactly how many cis-binding regulatory elements can lie upstream of the gene, as well as how many splice isoforms a single gene can incorporate and the number of different transcription factors which could potentially bind to the DNA. A study by [Warnefors and Eyre-Walker, 2011b] attempt to characterise sources of regulation influencing genes according to their age, suggesting that complexity is increasing in a continuous manner over evolutionary time and has not yet evolved to its maximum possible level.

Evidence suggests that in practise, however, regulatory complexity does not evolve to some kind of upper limit [Jay, 1996, McShea, 1996, Stewart, 2014], although it necessarily must have some kind of lower limit [Jay, 1996]. There is also evidence that organism complexity also involves phases of reduction and simplification [Wolf and Koonin, 2013], although it does appear clear that the development of multicellular organism and different cell lineages has contritibuted to an overall increase in regulatory complexity through time [Levine and Tjian, 2003, Moore, 2005]. Furthermore, pleiotrophic heterogeniety may causes differences in the evolvability of the genome [Wagner and Zhang, 2011]. For example, older genes (that is, present within highly diverged lineages) are more highly conserved and on average more pleiotropic, whilst less pleiotropic genes have a faster turnover and thus may represent the richest targets for regulatory innovations.

1.9 Towards a measure of regulatory complexity

1.9.1 Defining complexity

As demonstrated with the C-paradox and the G-paradox, genomic complexity is intuitively easy to define but notoriously difficult to quantify [Adami, 2002]. Furthermore, it has been seen that a genome isn't simply what is observed on the sequence, but an orchestra of regulatory layers, both acting on and off-site from the sequence, which interact in a precise and combinatorial manner and vary widely from cell-type to cell-type. This combinatorial approach to the regulatory programming of genes and the contributions of those programmes to their profile of expression output across the range of cell types in the organism raises many questions of precisely where and how that regulatory information is encoded, and whether different biological systems encode it in the same way.

For this reason, the problem of how to accurately measure the regulatory information for any given gene is something that has often been alluded to in genomics [Carninci et al., 2005, González et al., 2015, Hume, 2012, Nagel and Kay, 2012, Schroder et al., 2012]. Thinking of complexity in terms of 'the amount of information an organism stored in its genome' then this is intuitively the information which can be thought of as the accumulated regulation within the cell [Adami, 2002]. This allows us to ask the question: what is the minimum genome that can 'power' an organism? This question is naturally addressed by the concept of Kolmogorov complexity, which in this case is the minimum set of rules, or the regulatory programme, required to achieve the observed gene expression output. Whilst Kolmogorov complexity is believed to be the optimum way of measuring regulatory complexity, it is computationally intractable in practical situations [Kolmogorov, 1963]. Indeed, it is virtually impossible to combine together all of the different regulatory processes into one single complexity measure which describes how regulated an individual gene is in relation to another gene. Furthermore, the question of the exact mechanisms involved in explaining what causes the precise differences in expression between genes and cell types is generally not fully understood, let alone quantified.

Most studies focusing on measuring regulatory complexity in gene expression do so by observing singular aspects of regulation, for example, [Warnefors and Eyre-Walker, 2011b] consider eight 'measures' of complexity: counting transcription factor binding sites, conservation upstream of the gene, the number of TSSs, splicing isoforms, polyadenylation sites, miRNA sites, NMD proportion and RNA editing proportion. Studies have looked at sequence complexity at the promoter [Jin et al., 2014], including across chromosomes and multiple species [Tenreiro Machado, 2012]. A more recent study attempts to connect to the cis regulatory landscape to complexity by assigning DNASE I hypersensitive sites to their nearest gene in the context of understanding expression state transitions in hematopoietic differentiation [González et al., 2015]. A general disadvantage of these described studies in quantifying gene regulation is that they often reflect a single cell type, limit the number of regulatory factors under investigation, and on the whole it is unclear how each mode of regulation should be weighted. This final point underlines the difficulty in summarising the regulation in total over the gene and so reflects the difficulty in defining and measuring biological complexity [Landauer, 1988].

Instead, we could look at how these regulatory processes as a whole impact on the transcriptional initiation of a gene. This is reflected in the patterns of expression we observe, over time and in each specific cell type in the organism. In this project, we commonly refer to this as the output of the 'regulatory programme' acting on a given gene.

1.9.2 Gene expression as a 'regulatory programme'

As has been said, the regulation of gene expression is the key in understanding the process of cell type differentiation in multicellular organisms, and underlies how cells are able to adapt and respond to their environment, as well as precisely maintain their homoeostasis. Despite having the same genetic information, each distinct cell type in the body has its own unique gene expression profile, since each individual gene is controlled by a set of regulatory factors which determine whether the gene is switched on or off in a given cell type, the rate it is transcribed, and how its transcription responds

to a range of biological conditions. This set of regulatory factors can be described as constituting the regulatory programme for that gene.

For housekeeping genes, integral for the survival of the cell, this regulatory programme is conceptually simple - switch on the gene and maintain its transcription at its required level in all cell types; the abundance of mRNA produced is often controlled by the basal strength of the gene's core promoter, without requirement of further regulatory information. The resultant gene expression profile will always appear uniform in its distribution.

Some genes are highly restricted in their expression. For example, beta-globin genes are specifically expressed in erythrocytes [Levings and Bungert, 2002]. Cell specificity is generally determined by the presence of enhancer motifs, which are bound to by required transcription factors in the expressed cell type, but left alone elsewhere. In many such cases of highly specific expression, the complexity of the involved regulatory programme is also theoretically simple - a core promoter and one or more enhancer sequences to direct expression in the required cell type.

Whilst many studies attempt to classify genes according to the housekeeping to tissue restricted axis [Forrest et al., 2014, Frith et al., 2014, Heintzman et al., 2009, Jacox et al., 2010, Schug et al., 2005], the dichotomy of classifying genes in such a fashion is naive, since most genes which are regulatory complex in their expression lie somewhere in the intermediate scale [Jacox et al., 2010, Vinogradov, 2006] (Figure 1.7). Often the required level and precision of regulatory control is much more involved and lends to subtle changes in expression. Modular cis-regulatory elements such as enhancer sequences present in the proximity of the gene as well as at distal locations are responsible for much of this regulation. These enhancers are bound to by a variety of transcription factors which are themselves coded from genes, forming networks of interactions between transcription factors acting in trans- on the target gene. The vast numbers of possible combinations of cis and trans regulatory factors allows for an almost limitless scope for regulatory control in the genome. Promoter architecture is associated with classes of highly regulated genes, for example MHC class I promoter genes, known to be regulatory complex, are associated with Inr and TATAA elements [Lee et al., 2010].

Developmental regulators such as *OCT4* and *SOX2* in human are associated with Polycomb Repressive Complex 2 (PCR2) in ES cells [Lee et al., 2006]. In these cases the regulatory program causes the repression of the genes in ES cells, but allows for developmental switches across specific cell lineages, resulting in a complex gene expression profile across adult cell types.

An example of intuitively complex regulation is that of SHH through mammalian development. This key developmental gene is expressed in an ontogenetically, spatially and temporarily diverse set of cells and is crucial to the developmental patterning of limbs, teeth and fore-brain amongst further structures [Lettice et al., 2003]. As with other key developmental genes, the transcriptional regulation of SHH is mediated by a wealth of cis-regulatory enhancers in proximity to the gene, that are highly conserved through vertebrate evolution [Anderson and Hill, 2014, McEwen et al., 2009a]. If excised into a gene expression reporter construct such enhancers can often partially recapitulate the expression pattern of the endogenous gene [Visel et al., 2009], demonstrating the modularity of such cis-regulatory control.

Capturing these expression patterns therefore gives us important information about the overall contribution of the regulatory processes that make up the observed pattern in any given gene. The 'best' expression output from which to measure from would be every cell type at every state of development and in response to every possible biological environment, since this would give a complete picture of how a gene's expression at any one time or location. However, practically, only a selection of cell types or a single time course may be available. This project is concerned with such complexity measures.

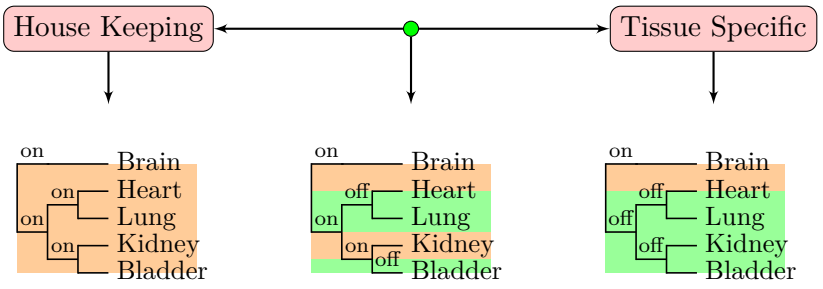


FIGURE 1.7: **Housekeeping vs tissue restricted axis**. Housekeeping genes are tissue specific genes may be explained bi conceptually simple regulatory programmes by observing the combinations of 'on' or 'off' switching through development. The current project hypothesises that genes between these two extremes potentially have highly complex regulatory programmes.

Chapter 2

Aims of the project and overview of thesis

2.1 Aims

1. Develop a justifiable method to quantify the regulatory complexity of a transcriptional programme.

2. Quantify regulatory complexity for all genes and promoters in the human genome.

3. Understand what makes a gene more or less complex.

2.1.1 Aim 1: Develop a justifiable method to quantify the regulatory complexity of a transcriptional programme

- Current ways of measuring complexity do not encompass all of how the genes regulatory program results in its gene expression output
- Aim to come up with a better information based measure on a gene level basis as a way of capturing our defined idea of 'complexity'

2.1.2 Aim 2: Quantify regulatory complexity for all genes and promoters in the human genome

- Calculate differential expression probabilities across FANTOM5 gene level data, TSS level data
- Calculate complexity measures over datasets using calculated probabilities
- Calculate complexity over subsets of the data (monocytes, certain groups of primary cells)

2.1.3 Aim 3: Understand what makes a gene more or less complex

- What regulatory elements contribute the strongest towards complexity in gene expression?
- How does the complexity of individual promoters relate to the complexity of the expression of the gene as a whole?
- What biological interpretations can we deduce from our measure?
- Can we add anything to the 'histone code hypothesis' - do combinations of histone modifications act in a predictive manner to regulatory complexity according to our measures?
- Can we compare cis and trans effects on the complexity in gene expression?

2.2 Brief overview of thesis

- Chapter 1 - Introduction to gene expression, regulation and their evolution
- Chapter 3 - Information theoretic methods which can be applied to gene expression data, introduction to graph theory and introduction to own measures used to estimate gene regulatory complexity..
- Chapter 4 - Introduction to FANTOM5 CAGE data, explain how the data is processed. Describe issues with the data, in particular exon painting and possible strategies around in in downstream analysis. Describe methods for calculating differential expression probabilities and the application to the current dataset.
- Chapter 5 - The analysis complexity scores for primary gene expression cell data, covering large set of transient cell types, and another set containing differentiated CD14+ monocytes. Correlate measures with genomic variables such as gene length, exon count, and distance to nearest genes. Consider cis-regulatory elements such as dnase I hypersensitivity and how sites around the gene correlate with complexity. Consider polycomb regulation by analysing the presence or absence of H3K27me3 marks in the promoter of the gene. Consider other histone marks and what influence they have on the complexity scores.
- Chapter 6 - Discussion of results and further work

Chapter 3

Measuring complexity

3.1 Introduction

The aim of this project is to capture a measure of a gene's regulatory information content, through the quantification of the observed expression states and the state switches that occur between the possible states. Simple stating measures that may be readily applied to gene expression profiles include summary statistics such as mean, median and maximum expression over a collection of cell types. However, it is clear that a more detailed and comprehensive framework will be required, taking into account the structured relationships between the measured cell types

The chapter begins by exploring information theoretic measures that have already been applied to gene expression and explore how these could be applied or adapted to the specific aims of this work. The next section then introduces relevant aspects of graph theory and how measures of differential expression between cell types can be displayed in the format of a graph from which regulatory metrics for a single gene could be extracted. These measures of information form the basis for the rest of the project, where we relate the regulatory complexity of a gene back to its observed regulatory content.

3.2 Information theoretic measures

For the purposes of this project let $X = (x_1, \ldots, x_n)$ represent a gene expression profile, where x_i represents the expression level recorded in sample i, which could represent for example a given time point or time point or differentiation state. We can convert a gene expression profile into a probability distribution:

$$p_i = \frac{x_i}{\sum_{i=1}^{n} x_i} \tag{3.1}$$

where p_i represents the 'mass' of expression in sample i. The entropy H of the profile X is calculated as

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{3.2}$$

Also known as the measure of uncertainty in the profile, H will be at a maximum in the case of uniform expression (ubiquitious) and equal to $\log(n)$, and equal to 0 in the case where all the expression is represented by a single sample (sample specific). In this way genes are classified according to a linear axis between universally expressed and those which are specific. Entropy scores are often normalised to fall between 0 and 1:

$$H = \frac{-\sum_{i=1}^{n} p_i \log_2 p_i}{\log(n)} \tag{3.3}$$

The classification of genes based on their tissue specificity has proven useful in identifying distinct promoter types and associated regulatory strategies [Forrest et al., 2014, Frith et al., 2014, Heintzman et al., 2009, Jacox et al., 2010, Schug et al., 2005]. For example, housekeeping genes in mammals often have CpG island associated promoters and exhibit considerable biological variability in the precise site of transcription initiation. In contrast, tissue specific promoters are less CpG island enriched, often associated with TATA-box transcription factor binding motifs and have less variability

in the site of transcription initiation. Entropy is also used as a measure of structural diversity in gene expression levels [Sherwin, 2010] and modifications of entropy used to understand transcriptome diversity [Martínez and Reyes-Valdés, 2008].

Despite its usefulness in many settings, as a way of measuring transcription regulation it is of less validity. For, the most highly regulated genes are potentially those which are expressed in a range of cell types and in an unpredictable manner. Such genes most often land somewhere arbitrary along the housekeeping - tissue restricted axis. For example, [Jacox et al., 2010] comment that regulatory features appear to be maximised in the central expression breadths as opposed to either end of the scale.

Figure 3.1 illustrates a limitation of the entropy in the context of ubiquitously expressed genes. Whilst entropy is successfully capturing cell-specific vs broadly expressed genes, around 35% of genes are ubiquitously expressed across at least one isoform (reference:analysis from Chapter 5). In Figure 3.1, it is clear that the entropy is not able to distinguish strongly between ubiquitous genes which are potentially simple in their expression, for example *ACTB* is highly uniform in its expression, and ubiquitous genes which exhibit many changes in expression, for example *FOS*. It is postulated that with the range of ubiquitous genes, many different regulatory mechanisms are at play and therefore a study to capture this spectrum is warranted. For example, ubiquitously expressed genes exhibiting profile changes could be under the control of ubiquitously expressed enhancer elements [Zabidi et al., 2014].

A further limitation of the basic entropy is that all samples are consider equally related to each other. This assumption does not hold in the case of biological cell types, as can be observed from intricately derived hierarchical clustering algorithms which groups together similar cell types. A way to overcome this issue may be to introduce some kind of weight structure and redefine the entropy as a weighted entropy.

To account for sample structure a set of weights $w_1 \dots w_n$ can be incorporated to give a weighted version of the Shannon entropy

$$H_w = -\sum_{i=1}^{n} w_i p_i \log_2 p_i \tag{3.4}$$

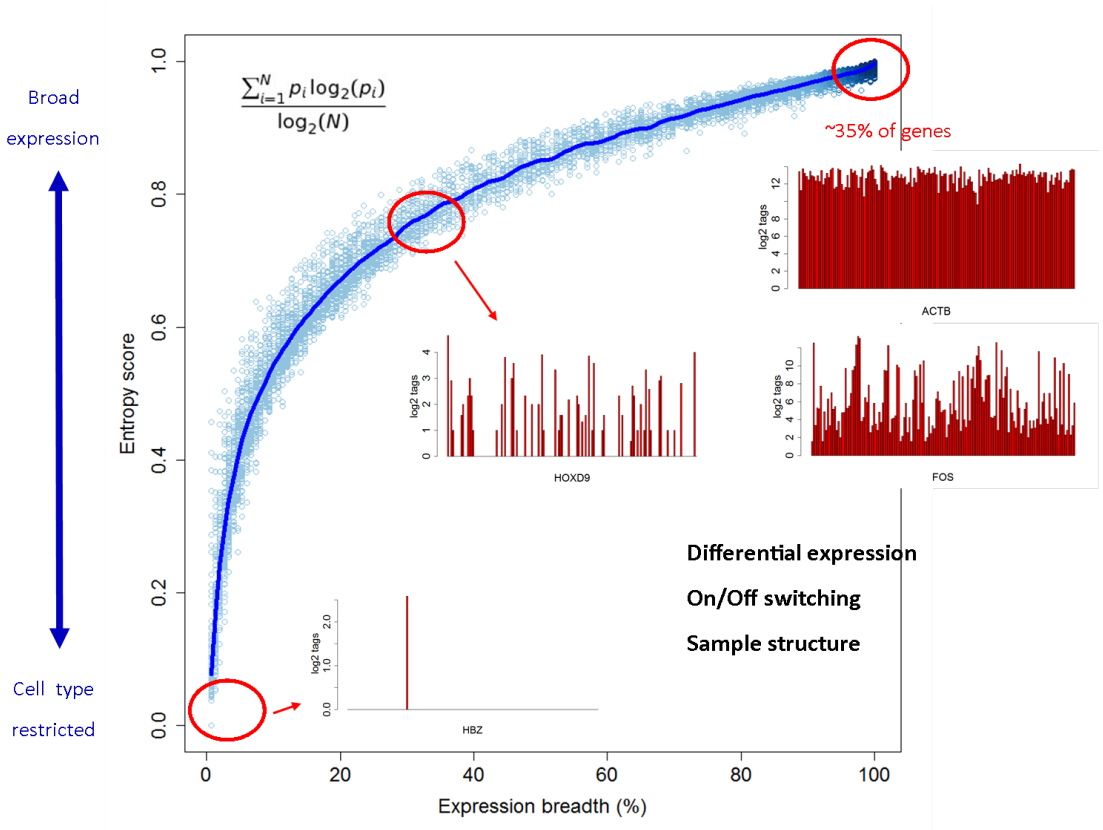


FIGURE 3.1: Breadth of expression (number of expressed primary cell types) against raw complexity scores. Darker blue regions represent regions containing many genes; in particular the dark region at the top generally represents ubiquitously expressed genes across all primary cell types.

The weights w_i can be used to convey information about the structure of the samples within the gene expression profile.

For example, one could attempt to down-weight over-represented samples and up-weight those samples which are more unique in their expression profiles, thus down-weighting similar over-represented sample types so that that resultant combined information of two samples conveying similar information will reduced compared to what is observed in the standard entropy function. Furthermore, weights could be applied to weight for distance from time 0 in a stimulus based time course.

How to define weights is not generally clear in terms of the weighted Shannon entropy, which is typically defined in terms of how samples relate to each other (for example, pairwise correlations, a two dimensional structure), rather than a single weight per cell

type. For example, the *guassian kernal* is often used to define structure between time points:

$$w_{i,j} = \exp\left(\frac{-||t_i - t_j||^2}{\sigma^2}\right) \tag{3.5}$$

where t_i and t_j are the times at points i and j and $sigma^2$ is the variance.

How to define this weighting for a singular time point is more challenging and less intuitive. One option for defining one dimensional weights per sample might be to calculate the diversity in the transcriptome of each sample, as per [Martínez and Reyes-Valdés, 2008].

3.2.1 Mutual information and KL-divergence

The Kullback-Libler divergence is a measure of the difference between two probability distributions.

The mutual information between two discrete random variables X and Y is defined as

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3.6}$$

where $p(x,y)$ is the joint distribution between X and Y. X and Y could refer to the probability distributions of expression across a given set of samples, with probability mass defined as in the definition of entropy above. Mutual information measures the level of dependence between the two given variables.

Mutual information is frequently referred to in gene expression studies when calculating gene regulatory networks [Luo et al., 2008, Steuer et al., 2002, Zhang et al., 2012].

Looking at KL divergence where Pr=1/n looks at the departure from uniformity for the gene expression profile, although this is the inverse of the Shannon entropy.

3.2.2 Kolmogorov complexity

Kolmogorov complexity Kolmogorov [1963] is a measure which assigns an object a complexity value equal to the length of the program needed to encode that object. The length of the program is independent of the coding scheme.

As an example, one might code which samples a gene is switched on or off in as a binary string, encoding a 0 if expression is not present in a sample or a 1 if expression is present. E.g. 011001 for a set of 5 samples, with expression observed in the second, third and sixth. Therefore, the Kolmogorov complexity is at most 6 (the number of samples) in this case. Alternatively, it can be thought of as a way of encoding switches in expression through time. For example, a gene expressed in ES cells might become switched off through methylation in two studies cell types and activated in two other cell types. such a scheme could be coded 10011 where the first digit represents ES cells, the second and third digits represent the cell types with methylation and the final two digits represent the active cell types (the labelling of the scheme is of course permutable). Note that in general Kolmogorov complexity is thought of as the 'ideal' complexity measure, but is NP-computable in most situations (Cannot be computed in polynomial time).

3.2.3 Permutation entropy

The permutation entropy was introduced by [Bandt and Pompe, 2002] and is commonly applicable for measuring the complexity of time series data. This approach has previously been applied to gene expression time series data from Arabidopsis and the resulting complexity measures have been related to the underlying biology of the genes [Sun et al., 2010]. This innovative work represents the closest approximation I am aware of, to the generalised measure of gene regulatory complexity aimed for in this work.

Permutation entropy considers subsets of consecutive time points of a given length and fits the pattern they form (e.g. up, up, down for a subset of length 3, possible combinations of three given in Figures 3.2 and 3.3). The permutation entropy score

(PE) is then given by the Shannon entropy of the relative contributions of each possible type of pattern.

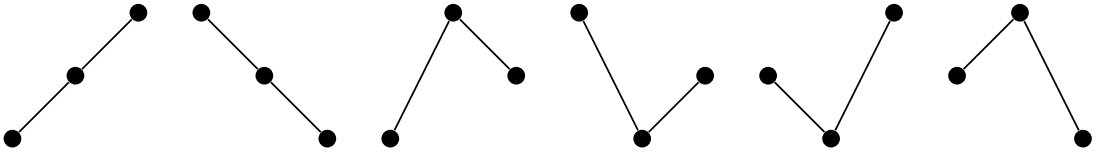


FIGURE 3.2: The ordinal patterns for $n = 3$, assuming all states have their own independent level of expression

For gene expression timecourse data one could add the following patterns

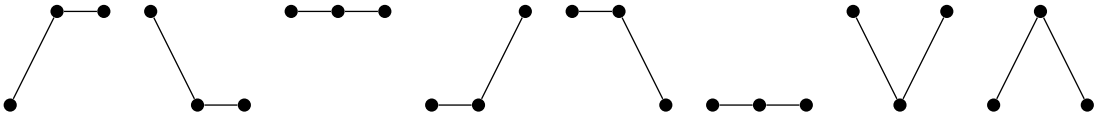


FIGURE 3.3: The ordinal patterns for $n = 3$ that could be added for gene expression data, assuming at least two states have equal expression (no differential expression) with regards to noise in the data

For a choice of $n = 3$, this gives 14 possible patterns, although Sun et al. [2010] consider only 14; they do not distinguish between the case where there are no changes and the gene is expressed and no changes and the gene is not expressed (this is not an issue in time-course data but may be of interest in other data structures). Then, the permutation entropy works be taking all the possible sets of $d = 3$ consecutive time-points (for $n = 7$ as in Sun et al. [2010] this is 5 sets), adding up the number of each 'pattern type' is observed, converting it into a probability distribution (diving the observed number of each pattern by the total number of patterns, i.e. 5 in the $n = 7$ case), and then calculating the Shannon entropy over the possible patterns.

If all of the patterns are the same over all of the time course (e.g. in the case where the gene is constantly expressed at the same level, or the case where expression in the gene is constantly and only going up or down in a monotonic fashion), then the resulting entropy is zero (this is similar to the tissue-specific case of the standard Shannon entropy). If there is a different pattern in every consecutive set of points, then the entropy is maximized (at $\log n = \log 7$), and suggests that the pattern formed by the gene is 'complex'.

The only known study to apply the permutation entropy to gene expression data is that of [Sun et al., 2010], who apply the method to a single 7-time-point Arabidopsis time-course using $d = 3$. They introduced the idea of the 'no-change' pattern, where the time-course remained flat for d time-points in a row. Because of the limited number of time-points, the entropy takes only a few discrete values, making it impossible to distinguish between two genes in the same category. However, as a measure of regulatory complexity, the they did demonstrate potential to capture regulatory information in individual genes, as displayed by GO term analysis [Sun et al., 2010].

Although the permutation entropy is useful in capturing properties of complexity we are looking for, it has a number of disadvantages in terms of describing gene expression patterns across cell types.

1. For small numbers of samples, genes fall into a small, discrete set of categories, making it difficult to distinguish between the relative complexities of genes within the

2. The up-up-up-x n time-points case receives the same score as the 'no-change' throughout the time course. A gene in a time course which is constantly going up over all its time-points is likely to need more regulatory information to sustain up-regulation, as opposed to the constant no-change scenario 'switch on the gene and leave it on at all times'.

3. The permutation entropy assumes a natural ordering of time points (although odes not assume a constant separation between two time-points). This works for time course data but will not necessarily work for primary cell data where the cell types are not ordered in time. A way around this might be to classify each cell types in terms of its euclidean distance from embryonic/stem cell lines. Or to group the cell types into equivalence classes and work within and between the groups without alluding to the idea of time.

3.2.4 Statistical complexity

The concept of statistical complexity was introduced by Lopez-Ruiz et al. [1995] and defines complexity as a trade off between disequilibrium and entropy ('order and disorder').

Define the *disequilibrium* D for a probability distribution as follows:

$$D = \sum_{i=1}^{N} (p_i - \frac{1}{N})^2 \tag{3.7}$$

where N is the length of the probability distribution. Intuitively, this is equal to zero in the case of a uniform distribution. Combining disequilibrium with entropy gives a quantity called *statistical complexity* C:

$$C = H \cdot D = -\left(K \sum_{i=1}^{N} p_i \log p_i\right) \cdot \left(\sum_{i=1}^{N} \left(p_i - \frac{1}{N}\right)^2\right) \tag{3.8}$$

Intuitively, since the Shannon entropy is small for specific distributions and disequilibrium is small for uniform distributions, statistical complexity will be small for both of those scenarios and maximised somewhere in between. Figure 3.4 illustrates this.

3.3 Improving upon current measures

Common disadvantages to all of these methods are follows

- Unclear how to apply the weight structure between samples in every method
- Unclear how to account for within-sample noise by use of replication
- Does not properly model 'changes' from sample to sample, which is essentially what is being driving by gene regulatory mechanisms
- No method appears to be independent of sample structure (e.g. we'd like a method that can be applied to time course as well as across sets of cell types)

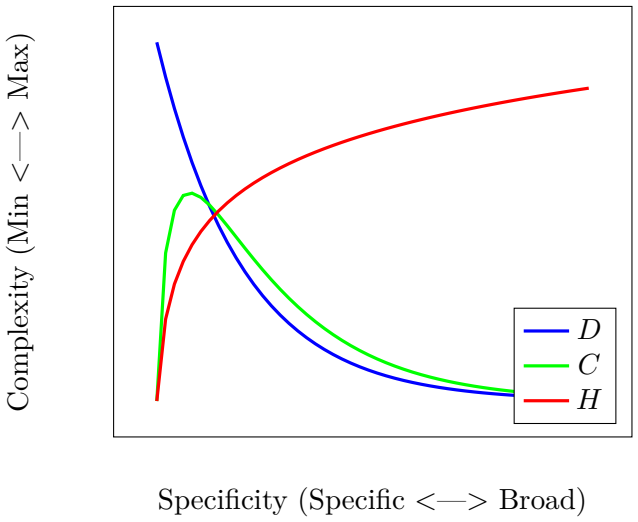


FIGURE 3.4: Statistical complexity example. On the right hand side is the case of a uniform distribution, on the left hand side of the plot is the case where we have all probability mass on one variable. H is entropy (red), D is disequilibrium (blue), C is complexity (as defined by $H.D$) (green).

Moreover many of these methods are two dependent on the breadth of expression - we are interested less in how many of the samples are actually expressed, more in the regulation involved in forming their resultant expression pattern. Whilst it is likely that sample-specificity is the result of more complex regulation than ubiquitous expression, we would like to be able to deduce that as opposed to make assumptions regarding this.

Introducing the concept of modelling changes in expression (differential expression) between samples changes the structure of the problem from one-dimensional to two-dimensional. This is because instead of a single row of data representing a gene, we are left with a matrix, with a measure of differential expression between a given pair of samples filling the data of the matrix.

Changing the problem from one-dimensional to two-dimension also has the distinct advantage of making it clear how apply the weights between the samples (since we are now looking at the problem from the point of view of between samples now). This concept is naturally covered in the field of graph theory.

In the next section we will introduce the basics required to understand graph theory. We will then describe how it works for gene expression profiles and how measures of

complexity can be extracted from the resulting graphs. We will then propose our own methods of complexity, which we will apply to the FANTOM5 data in the next chapter.

3.4 Introduction to graph theory

A graph is a mathematical object made up of a set of *edges* and a set of *vertices*. It is a tool for describing and visualising relationships between pairs of objects. For example, Figure 3.5 shows a simple graph with four vertices, labelled from 1 to 4, with edges between each pair but one.

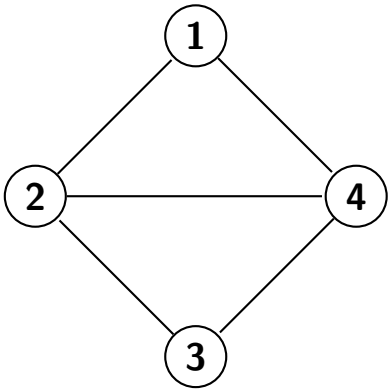


FIGURE 3.5: **Simple connected (but not complete) graph** with 4 vertices, 5 edges. The graph can be completed by the addition of an edge between vertex 1 and vertex 3.

For formally, the graph G is defined by the pair (V,E) where V is the set of vertices and E is the set of edges that make up the graph.

3.4.1 Adjacency matrix of a graph

Let u and v represent two vertices in the set of vertices V. The adjacency matrix A is defined by

$$A_{u,v} = \begin{cases} 1 & \text{if } u, v \in E \\ 0 & \text{if } u, v \notin E \end{cases} \tag{3.9}$$

In other words, it is a matrix with rows and columns representing the vertices and equal to 1 where there exists an edge in the graph connecting two given vertices, and 0 where there does not exist an edge between two given vertices (e.g. vertices 1 and 4 in Figure 3.5. The weighted adjacency matrix is given by

$$A_{u,v} = \begin{cases} w_{u,v} & \text{if } u, v \in E \\ 0 & \text{if } u, v \notin E \end{cases} \tag{3.10}$$

Thus, the weighted adjacency is similar to the adjacency matrix but instead of allocating a 1 between connected pairs of vertices, the edge is given a weight instead. For example, if the vertices of the graph represented cities and the edges represented the existence of a road between two given cities, the weight could represent the distance in km between these two cities.

A graph G is *connected* if for every pair of vertices $(u,v) \in V$ there exists a path between them. A *path* means that it is possible to reach each vertex from another by drawing a line through the available connections, which could involve other vertices in the graph. Figure 3.5 is connected but not complete, where every pair of vertices has a direct edge between them.

3.4.2 Laplacian matrix of a graph

The Laplacian matrix of a graph G is defined by

$$L_{u,v} = \begin{cases} deg(u) & \text{if } u = v \\ -1 & \text{if } u \neq v, (u, v) \in E \end{cases} \tag{3.11}$$

where $deg(u)$ is the degree of the vertex u; the number of edges moving from u to any other vertex. For example, in Figure 3.5, vertex 1 has degree 2 (edges connecting to vertex 2 and vertex 4) and vertex 4 has degree three (edges connecting to all of the other vertices).

A more general weighted version of the Laplacian for a symmetric simple graph G may be defined by

$$L_{u,v} = \begin{cases} \sum_u w_{u,v} - w_{u,u} & \text{if } u = v \\ -w_{u,v} & \text{if } u \neq v \end{cases}$$

where $w_{u,u}$ is the weight of the diagonal and the degree is approximated by the sum of the weights over the appropriate row or column (since G is symmetric) of the matrix representing the graph. For example, it is the sum of all of the distances to other cities from Edinburgh, minus the distance of moving from Edinburgh to Edinburgh itself (zero in this context).

3.4.3 Eigenvalues of graphs

The eigenvector of the n by n matrix A is the vector x such that

$$Ax = \lambda x$$

where λ is known as the eigenvalue for A. The set of eigenvalues is calculated as the roots of the equation $|det(A - \lambda I)|$, where *det* refers to the determinant, of which there are n. The sum of the eigenvalues is known as the trace of, denoted $trace(A)$.

If the matrix A is symmetric then it is called *positive semi definite* if all of its eigenvalues are non-negative.

3.4.4 Properties of eigenvalues of graphs

If the graph G is connected then it has a single largest eigenvalue, λ_{max}. This value represents the average connectivity of G and is always less than or equal to the maximum degree d_{max} of G.

The *chromatic number* of a graph G is defined as the minimum number of colours required such that every pair of connected vertices is coloured distinctly. For example, a graph which is completely connected will need a separate colour for every vertex, so its chromatic number will be the number of vertices n. A cycle of even number of vertices will require only two colours, so its chromatic number would be $n/2$. The chromatic number can be approximated using the Hoffman lower bound

$$\chi G \leq 1 + \frac{\lambda_{max}}{-\lambda_{min}}$$

where λ_{min} and λ_{max} are the smallest and largest eigenvalues of the adjacency matrix respectively.

The difference between the first and second largest eigenvalues is known as the *spectral gap* and represents important information about the connectivity of the graph. For the completely connected graph where there exists an edge between every pair of vertices, as in the case of uniformly expressed genes, the spectral gap is equal to $d_{max} - \lambda_2$ where λ_2 is the second largest eigenvalue. The eigenvalues of the Laplacian are commonly used as indicators of graph connectivity. The smallest eigenvalue is always 0, the second smallest is referred to as the algebraic connectivity.

3.5 Regulatory complexity in gene expression

In this section it is discussed how the graph theory defined above may be applied to gene expression data. We begin by defining the appropriate matrices with respect to gene expression and then looking at how their eigenvalue decomposition can be used to relate to the regulatory information in that gene.

3.5.1 Definitions

Matrix of differential expression

For a given gene g let $p^D_{i,j}$ represent the probability that g is differentially expressed between samples i and j.

For a set of n samples let A be the square n x n matrix with elements

$$A_{i,j} = \begin{cases} p^D_{i,j} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

D is clearly symmetric (since $p^D_{i,j} = p^D_{j,i}$) and satisfies the triangle inequality ($p^D_{i,j} \leq p^D_{i,k} + p^D_{k,j}$ for some $i, j, k \leq n$). In matrix form it looks like

$$\mathbf{D} = \begin{array}{c} \\ sample_1 \\ sample_2 \\ \vdots \\ sample_n \end{array} \begin{array}{c} \begin{array}{cccc} sample_1 & sample_2 & \cdots & sample_n \end{array} \\ \begin{pmatrix} 0 & p^D_{12} & \cdots & p^D_{1n} \\ p^D_{21} & 0 & \cdots & p^D_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p^D_{n1} & p^D_{n2} & \cdots & 0 \end{pmatrix} \end{array}$$

Graphical notations of differentially expressed states

We can define a graph G on the expressed states such that there exists an edge between a pair of vertices i, j samples within the gene g if $p^D_{ij} > 0$.

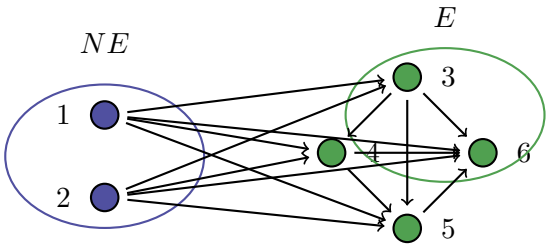


FIGURE 3.6: **Example graph for differential expression** The set NE represents the set of 'off' states and the set E represents the set of 'on' states. The graph is split, referring to NE being an independent set of samples, with no possible connections between them, and E, the set of 'on' states, have all possible connections between them. In the above graph, everything is differentially expressed, suggesting maximum complexity.

Such a graph may be generated for every gene or transcriptional element under study.

Defining the weighted Adjacency matrix for differential expression

$$A_{i,j} = \begin{cases} 0 & \text{if } i = j \\ w_{i,j} p^D_{i,j} & \text{if } i \neq j, i \in E \text{ or } j \in E \\ 0 & \text{if } i \neq j, i \notin E, j \notin E \end{cases}$$

where $0 \leq w^D_{i,j} \leq 1$ is a weight reflecting the similarity between samples i and j.

Intuitively, we wish to up-weight highly correlated pairs of samples, under the assumption that observed differences in expression between such samples would be a highly significant occurrence. On the contrary, the observation of differential expression between distantly related samples are down-weighted. The matrix $A.W$ then represents the information between samples i and j with which we calculate complexity.

$$\mathbf{AW} = \begin{array}{c} \\ sample_1 \\ sample_2 \\ \vdots \\ sample_n \end{array} \begin{array}{c} \begin{array}{cccc} sample_1 & sample_2 & \cdots & sample_n \end{array} \\ \begin{pmatrix} 0 & p^D_{12} w_{12} & \dots & p^D_{1n} w_{1n} \\ p^D_{21} w_{21} & 0 & \dots & p^D_{2n} w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p^D_{n1} w_{n1} & p^D_{n2} w_{n2} & \dots & 0 \end{pmatrix} \end{array}$$

In the case where all the expressed states are differentially expressed, we obtain the following adjacency matrix

$$A_{i,j} = \begin{cases} 0 & \text{if } i = j \\ w^D_{i,j} & \text{if } i \neq j, i \in E \text{ or } j \in E \\ 0 & \text{if } i \neq j, i \notin E, j \notin E \end{cases}$$

This type of graph is referred to as a weighted split graph; a special type of graph which contains a clique (maximally connected set) and an independent set of vertices. In this case the independent set is the set of non-expressed states and the clique is the set of expressed states, with the edges between them refer to the fact that they are

differentially expressed. Since each non-expressed state is differentially expressed with each expressed state, there exist all possible connections between the set of expressed and the set of non-expressed states - say that this type of graph is a complete weighted split graph, denoted $CS^{w}_{n,e}$, where n is the number of samples and e is the number of samples in the clique, that is the number of expressed states.

3.5.2 Defining the Laplacian for differential expression

For the set of expressed states E define the weighted Laplacian as follows

$$L^{D}_{i,j} = \begin{cases} \sum_i w_{i,j} p^{D}_{i,j} & i = j \\ -w_{i,j} p^{D}_{ij} & \text{if } i \neq j, i \in E \text{ or } j \in E \\ 0 & \text{if } i \neq j, i \notin E, j \notin E \end{cases}$$

Properties of $L^{D}_{i,j}$

The properties of $L^{D}_{i,j}$ are as follows:

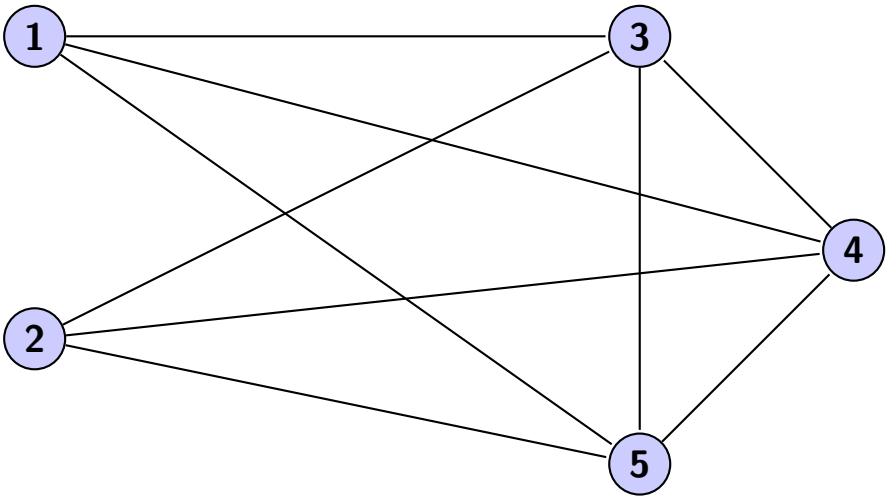
- $L^{D}_{i,j}$ is symmetric.
- $L^{D}_{i,j}$ is positive-semi definite.
- $L^{D}_{i,j}$ represents a graph which is connected and simple.
- Unless the weight between two expressed vertices is zero $L^{D}_{i,j}$ represents a split graph on $|E|$ vertices in its clique.

It follows that the eigenvalues of $L^{D}_{i,j}$ are real and because $L^{D}_{i,j}$ is connected, the set of eigenvalues $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots \lambda_n$ are above 0 and $\lambda_1 = 0$.

λ_2 is commonly referred to as the algebraic connectivity of G and has many interesting properties about the connectivity about the graph, relating to the diameter and mean distance of G. There are many interesting theorems about how λ_2 is bounded, included Cheeger's inequality.

Example

A set of 5 cell types



Dissimilarity matrix

$$\begin{pmatrix} 0 & 0.8 & 0.4 & 0.4 & 0.5 \\ 0.8 & 0 & 0.9 & 0.9 & 0.9 \\ 0.4 & 0.9 & 0.0 & 0.6 & 0.7 \\ 0.4 & 0.9 & 0.6 & 0 & 0.5 \\ 0.5 & 0.9 & 0.7 & 0.5 & 0 \end{pmatrix}$$

Differential expression probabilities example

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0.1 & 0.1 \\ 1 & 0 & 0.1 & 0 & 0 \\ 1 & 1 & 0.1 & 0 & 0 \end{pmatrix}$$

Resulting weighted Laplacian

$$\begin{pmatrix} 2.1 & -0.8 & -0.4 & -0.4 & -0.5 \\ -0.8 & 2.6 & -0.9 & 0 & -0.9 \\ -0.4 & -0.9 & 1.43 & -0.06 & -0.07 \\ -0.4 & 0 & -0.06 & 0.46 & 0 \\ -0.5 & -0.9 & -0.07 & 0 & 1.47 \end{pmatrix}$$

Eigenvalues of the weighted Laplacian are

$$\begin{pmatrix} 3.51 & 2.52 & 1.52 & 0.51 & 0.0 \end{pmatrix}$$

with corresponding eigenvectors

$$\begin{pmatrix} -0.31 & 0.83 & 0.06 & 0.07 & -0.45 \\ 0.85 & 0.06 & -0.01 & 0.26 & -0.45 \\ -0.30 & -0.32 & -0.74 & 0.25 & -0.45 \\ 0.05 & -0.15 & 0.02 & -0.88 & -0.45 \\ -0.29 & -0.42 & 0.67 & 0.30 & -0.45 \end{pmatrix}$$

3.5.3 Regulatory complexity

In the previous section we introduced how graph theoretic concepts can be applied to gene expression data, in particular in displaying information about differential expression between pairwise samples in a gene.

Uniform ubiquitous expression

$$A^D_{i,j} = \begin{cases} 0 & \text{if } i,j \in E \end{cases}$$

$$L^D_{i,j} = \begin{cases} 0 & \text{if } i,j \in E \end{cases}$$

In this case the eigenvalue decomposition is trivial - they are all equal to 0. Since the set of eigenvalues is positive we can see that in the case of completely uniform expression across all samples the eigenvalues will be zero, the minimum achievable score.

Cell type specific expression

For a single expressed state i where $|E| = 1$.

$$A^D_{i,j} = \begin{cases} 0 & \text{if } i = j \\ w_{ij} & \text{if } i \neq j, i \in E \\ 0 & \text{if } i \neq j, i, j \notin E \end{cases}$$

$$L^D_{i,j} = \begin{cases} \sum_i w_{ij} & \text{if } i = j \\ -w_{ij} & \text{if } i \neq j, i \in E \\ 0 & \text{if } i \neq j, i, j \notin E \end{cases}$$

These graphs receive low connectivity scores since the single expressed state is connected with each 'off' state, everywhere else contains no connections. It is dependent on the weight, so its magnitude in relation to other specifically expressed transcription elements depends on the structure of its weights to the other samples.

Differentially expressed everywhere

$p^D_{i,j} = 1$ for all $i, j \in E$:

$$A^D_{i,j} = \begin{cases} 0 & \text{if } i = j \\ w_{ij} & \text{if } i \neq j, i \in E \\ 0 & \text{if } i \neq j, i \text{ and } j \notin E \end{cases}$$

$$L^D_{i,j} = \begin{cases} \sum_i w_{ij} & \text{if } i = j \\ -w_{ij} & \text{if } i \neq j, i \in E \\ 0 & \text{if } i \neq j, i \text{ and } j \notin E \end{cases}$$

For example, in the unweighted case for a ubiquitously expressed gene the differential expression matrix for five samples will look like

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

The eigenvalues of the Laplacian, which will be the matrix containing -1 in the off diagonal and the number of samples n in the diagonal (5 in the example above), will be equal to n with multiplicity $n-1$ (the smallest eigenvalue is equal to 0). Therefore λ_2 has the potential to vary between 0 and n. When weighted the maximum possible value of λ_2 is equal to $\sum_{i,j} w_{i,j}$.

The eigenvalue decomposition of the Adjacency or Laplacian of the matrix of differential expression can therefore be used as a description or measure of information of the overall connectivity in differential expression between weighted samples for a given gene.

3.6 Graph theoretic measures based on eigenvalue decomposition of the differential expression Laplacian or adjacency matrix

Graph theoretic connectivity based methods are generally split into two categories, namely global based measures, returning a number for the whole graph, and local connectivity measures, which returns a value for the connectivity of each node of the

graph, namely the connectivity of each sample measured in the expression profile. Eigenvector centrality is a commonly used method which gives a connectivity score for each vertex.

3.6.1 Eigenvector centrality

For the matrix **AW** described above, eigenvector centrality is defined iteratively using the equations

$$\mathbf{x_i} = \frac{1}{\lambda} \sum_{t \in M(i)} \mathbf{x_i} = \frac{1}{\lambda} \sum_{t \in G} w_{i,j} \alpha_{i,j} x_j \tag{3.12}$$

where $w_{i,j}$ is the weight for sample i with sample j, $\alpha_{i,j}$ is the $(i,j)^{th}$ value of the adjacency matrix $\mathbf{A}$, x_i is the centrality scores for the i^{th} sample, $M(i)$ is the neibourhood of x_i, referring to the set of nodes to which sample i is connected. It is based on the equivalent matrix notation

$$\mathbf{DWx} = \lambda \mathbf{x} \tag{3.13}$$

where $\mathbf{x}$ is the vector of centralities, with one value per sample.

The algorithm to calculate eigenvector centrality starts by assigning a scores of 1 to all samples, so that $x_i = 1$ for all i. It then computes the above equations, the score for sample i is the weighted sum of all of the centralities for the samples in the neighbourhood of sample i. It is then normalised by dividing each by the largest values, and the above equations are again recomputed, in an iterative fashion until x has converged to the required vector of centralities.

3.6.2 Defining a family of complexity measures

Complexity measures may be defined as a family, whereby α and β may be varied according to relative user defined contributions of state switching between 'on' and 'off', and changes in expression between pairs of 'on' states.

$$C_R = \alpha \sum \lambda_{\text{on}} + \beta \sum \lambda_{\text{off}} \tag{3.14}$$

where C_R stands for the measure of complexity (regulatory), λ_{on} is the centrality scores of the 'on' states, and λ_{off} is the centrality scores of the 'off' states. The constants α and β stand for the relative contributions of the 'on' and the 'off' states. These constants make no different to ubiquitously expression complexity scores because there are no 'off' states but one might wish to, for example, up-weight the contribution of differential expression in genes between pairs of samples as opposed to switching between on and off (an 'off' state measures a switch between all 'on' states).

3.7 Normalisation strategies

The most 'complex' genes are differentially expressed between every cell type. However due to a finite, variable number of tags mapped to promoters across the cell types, the power to detect differential expression between every pair substantially increases as expression breadth increases. For example, in order to achieve a $\log_2$(fold-change) in expression between every pair, the ordered expression levels must increase be a factor of 4 from cell type to cell type. If there is, for example, 100 cell types then order to have statistical power to detect changes everywhere, one would need a total of at least $5.35646e + 59$ tags. Even in this scenario, if independent regulatory mechanisms were really causing independent expression levels from cell type to cell type, one would still expect by chance to obtain similar numbers of tags between some cell types. Compare this with expression across three cell types, now we only need 21 tags to see a pairwise fold change greater than 4 between every pair (1 tag in the first, 4 tags in the second

and 16 tags in the third). In order to normalise this, for each gene I simulated new gene expression profiles by permuting tag mappings across expressed cell types according to a multinomial distribution.

If I ask for at least a $\log_2$(fold-change) in expression between every pair then the number of required tags for n cell types is

$$N^{\text{theory}} = 1 + \sum_{i=1}^{n-1} 4^i$$

.

Substituting for N^{actual} and resolving for the number of possible unique expression levels l gives

$$N^{\text{actual}} = 1 + \sum_{i=1}^{l-1} 4^i$$

.

Evaluating the sum and rearranging thus gives

$$l = \frac{\log_2(1 + 3N^{\text{actual}})}{2}$$

.

Thus we can calculate the number of cells types over which it is statistically possible to detect a $\log_2$(fold-change) in expression between every pair on a per gene basis given its mapped tag count.

To generate a probability distribution for tag mapping we defined the probability of a tag mapping to a given cell type according to an inverse power sum distribution:

$$\alpha_k = \frac{1}{\sum_{k=1}^{k-1} 4^k}$$

for $k = 1 \ldots l$. Each cell type was allocated a α_k at random together with a library specific normalisation factor ψ_i:

$$\theta_i = \alpha_{k^{\text{random}}} \psi_i$$

where k^{random} is a random integer from 1 to l.

To simulate permuted profiles we simulate sets of tags from a probability distribution based on the multinomial distribution

$$P(X_1 = x_1, ..., X_n = x_n) = \frac{(N!)}{(\prod_{i=1}^{n} x_i!) \prod_{i=1}^{n} \theta_i^{x_i}}$$

where x_i are integers represents simulated mapped tags to cell type i such that

$$\sum_{i=1}^{n} x_i = N^{\text{actual}}$$

and $\theta_i > 0$ and

$$\sum_{i=1}^{n} \theta_i = 1$$

For each gene we simulated 100 new diversity focused tag distributions and re-calculated complexity scores for that gene, generating a vector π^g. The maximum achieved complexity from these distributions acted as a normalisation factor for a each specific gene

$$q^g = \max(\mathbf{C_r}(\pi^{\mathbf{g}}))$$

with normalised complexity scores given as

$$\mathbf{C}_{\mathbf{r}}^{\text{norm}}(g) = \mathbf{C_r}(g)/q^g$$

Thus, in order to discover the maximum practically possible complexity score for each gene, tags mapped to this gene across cell types are randomly mapped back to the same number of cell types (maintaining expression breadth, but not necessarily the exact same set of cell types). Due to vast numbers of possible random mappings, I derived a function to estimate the potential number of expression levels possible in the data and to attempt to generate expression profiles with at least that number of distinct levels in expression. Using this function allows for a much quicker maximal estimation, as opposed to simple random distribution of tags, which would potentially require huge numbers of iterations.

In order to estimate the maximum observable complexity for a given expression breadth, the maximum complexity obtained by tag redistribution was calculated. Then, a smooth curve may be fitted through the outside edge of the points. Therefore, by applying this strategy, complexity scores may be normalised on a per-gene basis, but also on a breadth level.

3.8 Overview of method

In this chapter, an overview of methods applied to gene expression profiles to capture regulatory information was described. Then a novel approach of capturing information from expression profiles was described by describing the problem in a graph theoretic framework and applied graph theoretic connectivity measures. Whilst graphs are frequently produced to attempt to understand gene-gene interactions, graph theory has not been seen to be applied in the context of gene expression as a tool to understand the regulatory complexity of an individual transcriptional element. An overview of the steps are described and graph framework given in Figure 3.7 and are in general as follows:

Select data

This could include transcriptional elements across

- Tissues across the body

- Cell types or primary cell types across the body (to understand changes in expression of primary cell types within tissues)
- Time courses, to quantify patterns; this has been alluded to in [Sun et al., 2010] but not applied extensively. Furthermore, the current approach allows for the fine tuning of weights according to desired characteristics, for example example late vs early response, sudden vs gradual change or more complex patterns which mix all of these factors.

The data should ideally have replication in order to accurately detect differential expression between pairs of samples (next Chapter).

Measure potential regulatory output

Mine potential information from the expression profile which may explain the regulatory control of that element:

- Pairwise differential expression probabilities
- Determine on/off states
- Determine sample structure

This information can then be used to generate a graph.

Generate weighted graph

Create weighted adjacency and/or Laplacian matrices as defined above.

Calculate graph based connectivity measures

Calculate eigenvector decomposition, eigenvector centrality.

Select α, β and calculate final measure

Or use overall graph connectivity

Inference on final scores

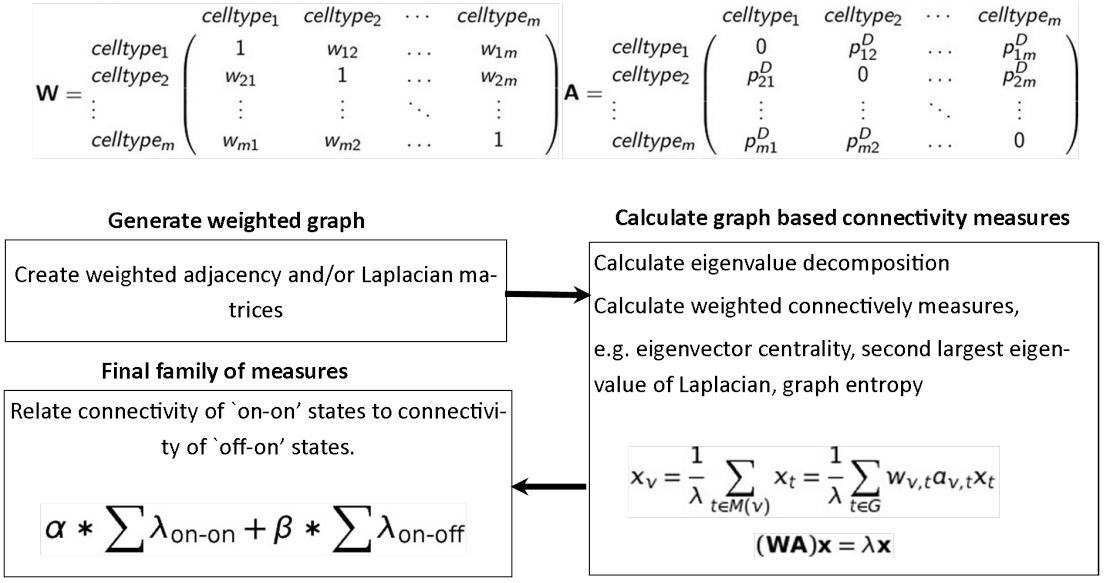


FIGURE 3.7: **Method overview:** Reformulate the problem into a graph theoretic framework and calculate connectivity based measures.

Connect the resulting single measure per transcriptional elements back to the regulatory elements thought to control that elements - for example, cis regulatory sites surround a gene transcription start site, histone modifications or disease SNPs analysis.

3.9 Overview of next chapter

In the next chapter the FANTOM5 project, which maps transcripts of active transcription start sites on a genome wide level across a wide variety of samples, is introduced and described. Methods of calculating differential expression are described. The data used to generate the complexity scores for the inference in Chapter 5 is described and how complexity measures have been applied to this data is given in detail.

Chapter 4

Applying complexity measures to FANTOM5 CAGE

In this chapter the FANTOM5 project is introduced, which generated the wealth of CAGE data analysed in this project. First, the range of samples sequenced and the kinds of normalization and feature clustering procedures that have been applied to quantify genome wide expression levels for transcription start sites and genes are discussed. Next it is explained why the data is particularly suitable for calculating regulatory complexity measures, as discussed in Chapter 3, and how these measures are calculated from the data. Their calculation is highly dependent on accurately quantifying changes in expression between samples for a given transcriptional elements, and methods for calculating these changes are discussed.

4.1 Scope of the project

The FANTOM5 project maps active transcripts and promoters in mammalian genomes [Forrest et al., 2014]. Officially, a total of 573 primary cell samples in human and 128 primary cell samples in mouse were collected. The primary cell data as a whole represents 72,964 peaks of size larger than 10 tags per million (tpm). Of these, 30,517 peaks are unique to the primary cells, making them the largest resource in the FANTOM5

data. Furthermore, there is, overall, at least one promoter covered for more than 95% of annotated protein-coding genes, with only 1225 remaining uncharacterised [Forrest et al., 2014]. Furthermore, of all the peaks unique to the primary cells, around three quarters of them represent novel peaks. This makes this data an important resource for discovering new biologies, although for the current analysis, many of these peaks will not be considered, since only those in the 'robust' set are used, i.e those receiving some kind of additional support from expressed sequence tags (ESTs), H3K4Me3 histone methylation marks and DNase I hypersensitive sites.

In terms of promoter architecture we observed within the FANTOM5 data, it was found that in the robust set of TSS, one or more TSS was annotated to 91% of all protein coding genes, with an average number of TSS per gene of around 4 [Forrest et al., 2014]. Using a threshold between the least and most expressed cell type, 6% of promoters were recorded as housekeeping (>50% of samples with a less than 10 fold change between median and maximum expression). 80% of promoters were found to be cell type restricted and 14% were ubiquitous non-uniform [Forrest et al., 2014].

An accompanying manuscript generated an atlas of enhancers in FANTOM5 libraries [Andersson et al., 2014b], identifying 43,011 putative enhancers based on bidirectional transcription at transcriptional loci across 808 human FANTOM CAGE libraries. Further analysis of phase II of the FANTOM5 project found that bidirectional eRNA-defined enhancers are transcriptionally active before the promoter of a gene itself is active [Arner et al., 2015], suggesting that enhancer RNAs represent the earliest response in differentiation and post-stimulus.

4.2 Normalization, clustering and quality control

4.2.1 Normalization

The RAW sequencing data for the CAGE libraries vary in terms of the total number of tags sequenced, according to the sequencing depth of the sample. The greater the sequencing depth, the more tags uniquely map to the genome. Thus, whilst transcription

events are directly comparable based on the number of tags mapped to regions within the same library, it is more difficult to compare the same transcription event between libraries without first normalising the data.

The libraries for the FANTOM5 CAGE data are first normalized under the assumption that the expected tag count ratio $K_{gj}/K_{gj'}$ for a given gene g between different samples j and j' should be equal to the size ratio $S_j/S_{j'}$ if gene g is not differentially expressed between samples j and j', or if j and j' are not replicates. Assuming that tag read counts are proportional to the expression level and sequencing depth, Anders and Huber [Anders and Huber, 2010] correct for library size by using the median of the ratios of observed counts in order to estimate size factors per library:

$$\hat{s}_j = \underset{i}{\text{median}} \frac{k_{ij}}{(\prod_{v=1}^{m} k_{iv})^{1/m}} \tag{4.1}$$

4.2.2 Clustering

Tags mapped across the genome were clustered into groups representing TSS. The number of tags mapped to the location representing the TSS acts as a proxy for expression levels at that location. For the FANTOM5 data, [Forrest et al., 2014] developed the DPI (decomposition based peak identification) algorithm, illustrated in Figure 4.2. Briefly, groups of tags are first found according to a distance cut off. Larger 'clusters' are then decomposed into non-overlapping sub-clusters, according to compositions of tags observed within it, thus breaking it up into distinct TSS. Whilst the technique allows for the potential to observe multiple TSS within short range for the same gene, often it is unclear whether two clusters of tags nearby too each other should be marked as two distinct transcription start sites (Figure 4.3, top) or whether groups of tags should be separated into two TSS or remain as a single TSS (Figure 4.3, bottom).

A further bias to accurately quantifying initiation events is a concept referred to as exon painting, the subject of the next section.

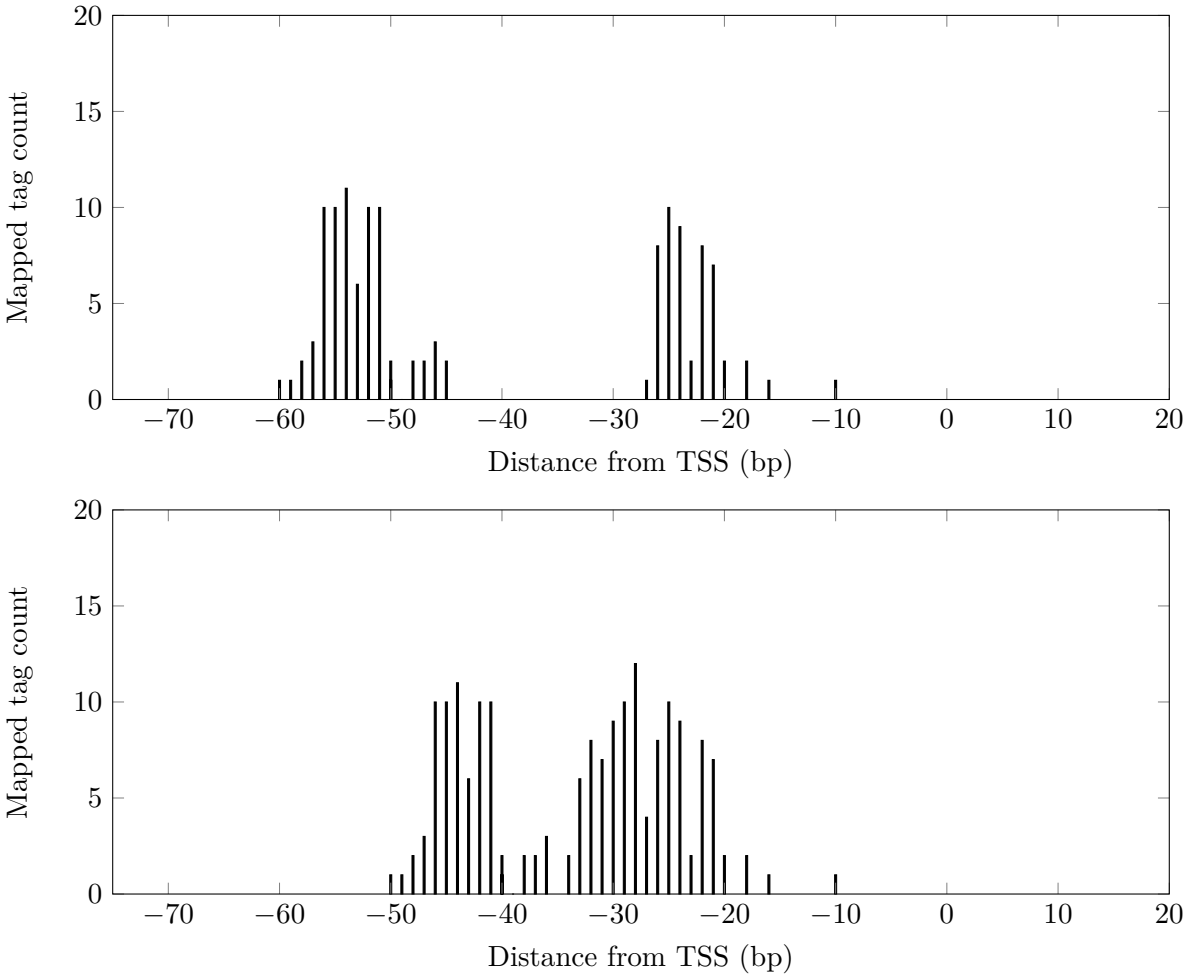


FIGURE 4.1: Example tag counts for TSS. In the top plot, there are two distinct TSS visible, which may be captured by tag clustering techniques. In the second plot, it is less clear whether there is a single TSS, or whether the tags should be split into separate "clusters" representing two transcription start sites.

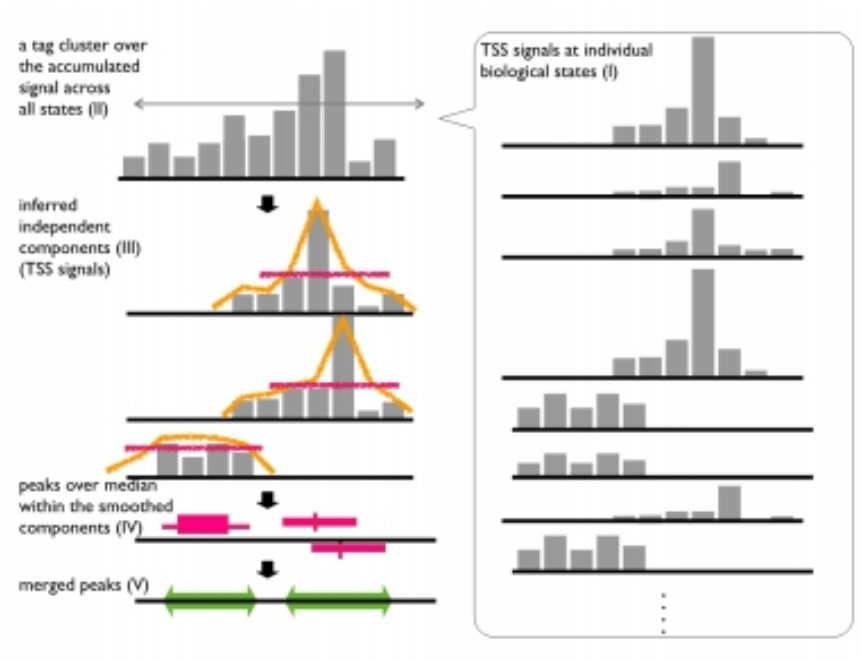


FIGURE 4.2: **DPI clustering algorithm used to detect CAGE peaks in the FANTOM5 data**. Figure adapted from [Forrest et al., 2014]

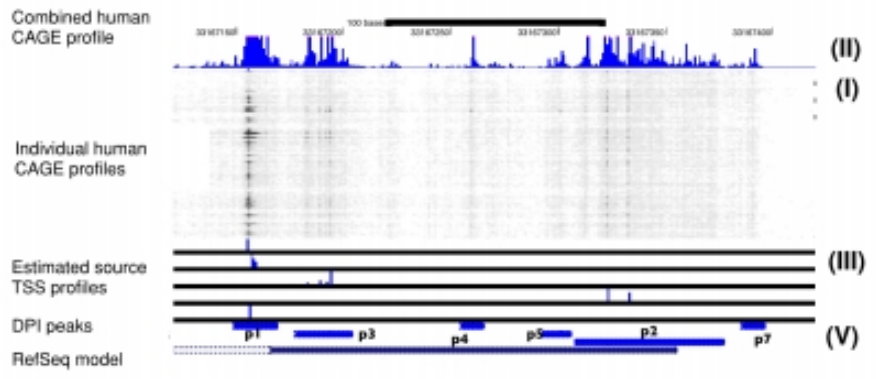


FIGURE 4.3: **Promoters captured through the DPI clustering algorithm**. The DPI peaks row shows the locations of the clustered perks and the top row shows the mapped tag distributions to those locations, based on combined tag counts across all human CAGE libraries. Figure adapted from [Forrest et al., 2014]

4.3 Exon painting

An issue with the CAGE protocol, previously mentioned in Chapter 1, in determining alternative transcription start sites, is the presence of 'exon painting' - whereby multiple cage peaks are detected in exonic regions, adjacent to the location of TSS peaks of genes (illustrated in Figure 4.4, with an example given in Figure 4.5).

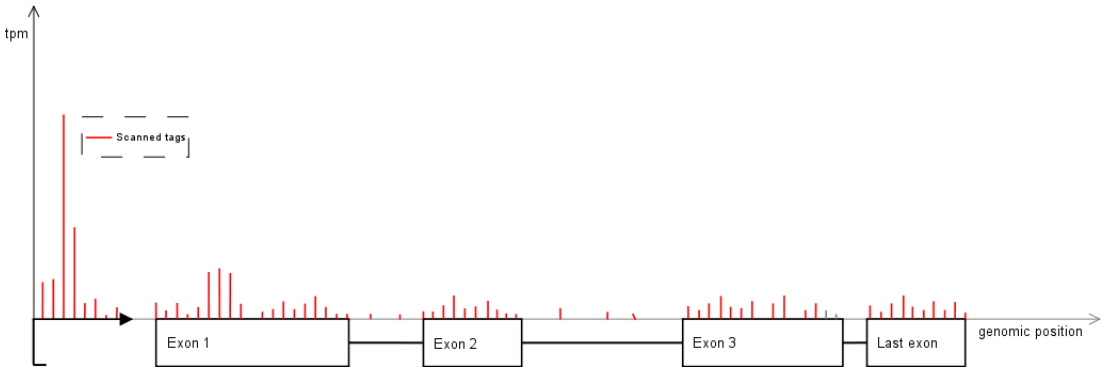


FIGURE 4.4: **Diagram illustrating exon painting artefacts**. Red vertical lines represent number of CAGE tags overlapping that specific nucleotide. Peaks of tags on the far left represent TSS signal, tags on the exons represent painting signal. Tags within the annotated first exon may represent TSS signal, due to different gene isoforms or mis-defined annotation coordinates. Correcting TSS and Exon 1 signal depends on accurately measuring degradation signal across the transcript.

Such tags often contaminate regions with mapped tags representing transcription start sites in or near these genes. The bias was first alluded to in ENCODE [Gingeras, 2009] and expanded upon in the context of CAGE, where exon painting is particularly an issue in the FANTOM5 dataset, due to the large library depths (medium depth 4 million mapped tags [Forrest et al., 2014]). Exon painting in CAGE is thought to relate to the hCAGE protocol, due to the recapping of processed transcripts and is often avoided by only considering TSS in intergenic regions.

An alternative explanation of why exon painting may occur is described in the next section, and is backed up by an analysis of exon painting in the FANTOM5 libraries carried out at the beginning of the current project, showing that levels of painting observed in exons is directly proportional to the transcriptional activity of the core promoter regulating that gene. The analysis also illustrates the power of CAGE in

accurately detecting missed initiation events due to faults in the annotations of protein coding genes.

4.3.1 Hypothesised protocol

The hCAGE capture protocol involves the biotinylation of molecules containing pairs of hydroxide groups on adjacent bonded carbon atoms, which also occurs on the ribose sugar (2' and 3' hyroxides). The 5' cap structure and the 3' end of RNA molecules therefore both contain this same structure. cDNA synthesis using random primers will lead to a mixed population of products since reverse transcription sometimes extends right up to the 5' cap, whilst in a huge number of cases reverse transcriptions prematurely terminates. The crucial step of hCAGE is the use of a ribonuclease that will degrade single stranded RNA but not that in heteroduplex with DNA (the RNA is protected by the DNA). Consequently, for any cDNA first strand synthesis reactions that do not extend fully to the cap, the cDNA will be separated from the biotinylated cap by RNase treatment, so not sequenced.

However, the 3' end of RNA molecules will also have been biotinylated in the same reaction. In the rare but possible case where a random primer has annealed and the extreme 3' end and has been extended, the cDNA will protect the 3' RNA and remain associated with a covalently attached biotin group (similar to the cap structure) and thus captured. The 3' end of such cDNAs are then a target for sequencing. As these cDNAs are not then required to have reached the cap structure to enable capture, they are likely to represent a heterogeneous population based on sampling positions along the mRNA transcript where the reverse transcriptase dissociated at random points. This population of tags is likely to be a key source of the exon painting signal and is expected to be proportional to transcript expression level.

4.3.2 Quantifying exon painting

Detecting exon painting is important because it may cause expressin quantification biases in algorithms designed to detect groups of tags representing TSS, where the

FIGURE 4.5: **Example of exon painting**. Example of an exon painting gene with the promoter showing (top). The TSS is the peak of tags in the left, exon painting tags can be seen covering exonic regions. The same plot without the TSS, thus zooming in on the levels of tags mapped to exonic regions. Visualisations are based on a screen-shot from the ZENBU browser [Severin et al., 2014].

algorithm may not be able to distinguish between tags which are representing the genuine expression signal and those that are the result of exon painting across the 5' exon, and therefore the resulting expression value of that TSS will be overstated. At the beginning of the current project, when the FANTOM5 data was newly sequenced, it was unclear how strong the exon painting effect was. Therefore, I began my analysis of the data by conducting investigations into exon painting and how it could potentially be corrected for in downstream analysis. In general, the idea that exon painting represents random degradation suggests that exon painting should be broadly spread across a transcript, whereas promoter signals are likely to peak at a given region where the genuine TSS is. The greater the TSS peak, the more mRNA within the cell whereby the degradation artefact could potentially occur. Thus, the hypothesis tested was whether the larger the promoter signal, the larger the level of 'broad' painting signal across the transcript body, forming a ratio of painting to promoter tags.

To test this hypothesis, tags mapping defined exons were counted together with tags in the promoter region of the gene. Plotting the promoter signal against exon painting (exon tags per nucleotide of transcript), it was seen that for many genes the exon painting correlates pretty well with the promoter TSS signal (Figure 4.6, blue region) and two fairly distinct categories appeared:

1. Genes where the promoter signal is correlated to the painting signal.

2. Genes where the promoter signal is much lower than expected from the painting signal.

Hand-curating some of those genes where the promoter signal is much lower than expected, we find that the annotated gene structure (in this case RefSeq annotations) excludes the obvious promoter for that library. Modifying the gene structure to include the previously missed promoter moves the genes from category 2 above to category 1 (Figure 4.7).

To investigate this further, I removed the annotated 5' exon from the gene structures and re-ran the analysis to obtain a null distribution (Figure 4.6, red points), to represent painted genes but without their real promoters, and found that it overlapped

the genes from category 2 above and excludes category 1 genes. Thus, it would seem that exon painting may have some value as a per-library internal reference for transcript expression. If the promoter signal is weaker than expected for the observed exon painting, it suggests that the correct promoter(s) for that CAGE library have not been assigned to the gene (due to thresholding to find the promoter tags).

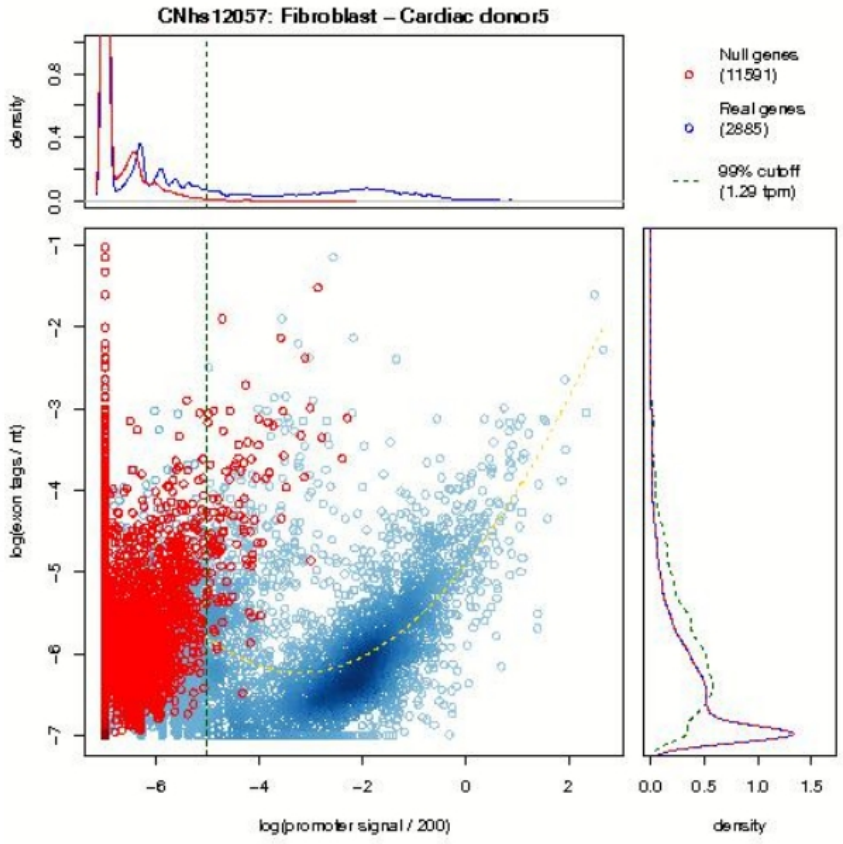


FIGURE 4.6: **Exon painting in an example library (CNhs12057)**. x-axis is the recorded log promoter signal based on gene annotations, y-axis is the log exon painting signal based on averaging tags across non-5' exons. Blue is the actual distribution and red is a generated null distribution based on removing the annotated 5'exon. Cut-offs between promoters which have an under-represented promoter signal are defined taking a 95% threshold on the promoter signal of the null distribution (denoted by dashed green line)

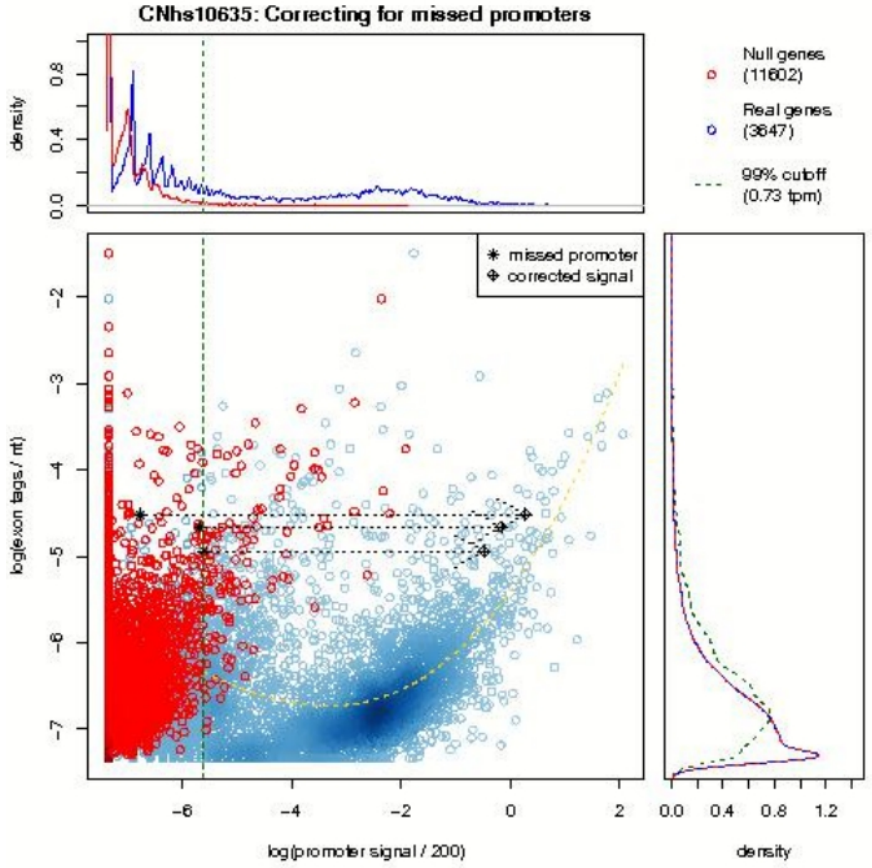


FIGURE 4.7: **Exon painting in an example library (CNhs10635), rescuing the promoters of three genes**. x-axis is the recorded log promoter signal based on gene annotations, y-axis is the log exon painting signal based on averaging tags across non-5' exons. Blue is the actual distribution and red is a generated null distribution based on removing the annotated 5'exon. Cut-offs between promoters which have an under-represented promoter signal are defined taking a 95% threshold on the promoter signal of the null distribution (denoted by dashed green line). Black points represent the locations of three gene before and after 'rescue', these were identified as having under-represented promoter signals, and manual curation found the 'real' promoter based on changing the locations of the annotated 5' exon coordinates.

Gene name	Mappability to 3 d.p.
RPS3	0.748
BTF3	0.748
SSR4	1
SELT	0.834
PPIA	0.629

TABLE 4.1: Mappability values for a sample of five genes with consistently low exon painting values, based on UCSC tracks for 36nt windows. A mappability of 1 indicates that all CAGE tags will map uniquely to the gene, whereas a mappability below 1 indicates the proportion of tags which will be expected to map unique to the gene, with PPIA having the lowest mappability of this sample. Of the reference genome, 83% of genes had a mappability score of 0.95 or greater, and the sample above has lower than expected mappability scores ($p<0.004$). Applying a mappability correction to the exon painting estimations appears to recover some of these genes

4.3.3 Mappability of tags - a further confounding factor

A potential issue with many sequencing based methods is the problem of mapping tags back to the genome. Mappability affects not only CAGE mappings but also for example the mapping of DNase I hypersensitive sites or methylation marks. The average tag size of hCAGE was 36 nt, which although provided sufficiently unique mappings across most promoters, still left from for ambiguity in certain regions. These regions include regions with high repetitivity, as well as regions containing pseudo genes. Whilst these pseudo genes may be silenced epigenetic ally, it is difficult to determine computationally whether a tag should map to the active gene or its pseudo counterpart due to similarities in sequence. The result is that a gene with one or more pseudo-genes will have a lower than expected tag count associated with its transcription start sites.

It was observed that genes with low mappability scores had lower than expected exon painting signals, suggesting that this may be indeed the case - tags which should have mapped onto the first exon of a gene and thus contributing to the painting signal may have actually been mapped to the first exon of their pseudo counterpart. It was noted that applying a mappability correction to the exon painting estimations appears to recover some of these genes.

4.3.4 Discussion

The above analysis suggests that if the promoter signal is weaker than expected for the observed exon painting, then the correct promoter(s) for that CAGE library may not been assigned to the gene. Thus, such analysis as presented above could be used in one of two ways,

1. To identify genes where we may have missed the relevant promoter(s) and target them for manual annotation/RACE.

2. It would be used as a quantitative measure of how well our complete promoterome discovery is adding to current knowledge (i.e. how many genes move from category 2 to category 1 based on TSS finding algorithms).

Furthermore, it was seen that genes with recent processed pseudo-genes have lower than expected exon painting signals, presumably as painting tags will have low mapping scores and applying a mappablilty correction to the promoter region and transcript region seems to recover the extreme outliers (e.g. *PPIA* and *SELT*).

In conclusion, the FANTOM consortium decided to get around the issue of exon painting by creating a threshold based on the ratio of exonic peaks to promoter peaks. For ratios of 0.7 or above they assigned a TSS region as a 'permissive' peaks and a ratio of above approximately 2.0 was assigned as a 'robust' peak, 'corresponding to peaks with a single CTSS in a single experiment supported by 11 or more observations and 1 or more TPM' [Forrest et al., 2014]. Whilst this robust set of peaks provided a well supported and high confidence set of peaks to work with, it was felt that such a threshold removed many of the potential novel peaks, so in many analyses, the permissive set was preferred and in more specific analysis, CAGE tags without thresholding were considered.

4.3.5 Methods

The algorithm is based on a fixed size window at the 5' end of an annotated gene structure to define a promoter, where one TSS per gene is considered. Discrimination

between category 1 and category 2 is based on taking a 95% threshold on the promoter signal of the null distribution (denoted by dashed green line in the attached plots). The presented work is based on RefSeq annotations.

4.4 How to calculate differential expression in CAGE

It is widely believed that changes in the regulation of a gene between samples are reflected in changes in its expression, which observed through changes in its initiation at the core promoter. Therefore, statistically measuring the information from the output of the gene's regulatory programme depends on the ability to accurately detect these expression changes. In this section we discuss methods for detecting differentially expressed genes between samples.

The existence of technical and biological replication allows for estimation of noise, which allows for the estimation of the significance of changes. It has been shown that next generation sequencing methods such as RNA-seq hold up very well in terms of technical variability [Marioni et al., 2008], so studies of differential expression often focus on accounting for biological variability. In the context of mRNA sequencing, noise estimations can be improved in two main ways - one is by increasing the number of replicates and the other is by increasing the depth of coverage at the sequencing level. Although the costs of next generation sequencing are clearly falling, it is still expensive and there is clearly a cost trade-off in terms of sequence depth, replication and the number of different samples required (e.g. more time points with fewer replications per time point, or fewer time points and more replication) [Liu et al., 2014].

A common standard in high throughput sequencing is around 3 biological replications per sample, which is a low level of replication with which to calculate noise estimations. The simplest and most naive way of testing for changes between samples for a specific gene is to use the 2 sample t-test or a non-parametric equivalent. There are a number of reasons why this is not the ideal solution; some of the samples do not have any biological replicates, making it impossible to estimate the level of noise and compare to other samples. However, one might expect the noise level between replicates to be

very similar (with some difference) from one sample to another. So, extrapolating this variance to those samples lacking in replication would be a useful approach since, in a dataset with 20k+ genes and multiple samples, we are essentially discarding the vast majority of the data when we conduct each test of difference [Law et al., 2014].

The presence of a small number of replicates per sample type, with some sample types having a only single donor, together with a large number of TSSs implies that in order to capture differential expression accurately, one could model the data in such a way that noise estimations are shared between TSSs in the same libraries. This is commonly achieved using a Bayesian hierarchical models, where each TSS is given its own noise parameter, but the noise parameters across all TSS are drawn from a common, shared distribution for that sample type [Chung et al., 2013, Law et al., 2014, Vavoulis et al., 2015]. This means that effectively the entire dataset is used in the estimation of each noise parameter and this 'borrowing' effect is particularly useful where only 1 or 2 biological replicates are available (this concept is shown in Figure 4.8).

4.4.1 Techniques of calculating differential expression

For data from microarray experiments, where differential expression analyses were commonly built on, log-transforming the values and assuming a normal distribution was a standard approach [Hoyle et al., 2002]. Even though this log-normal assumption does not hold in count based sequencing data, earlier techniques of calculating differentiation expression in count based data nonetheless extended the log-normal approach in modelling the data to work with counts (e.g. see [Smyth, 2004]). The first truly hierarchical model developed for count data was edgeR [Robinson et al., 2010], with is based on a negative binomial distribution, a common distribution for modelling count data, allowing for large variances in counts, referred to as over-dispersion, which is commonly observing in gene expression data. The negative binomial distribution is given by:

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma^2_{ij}) \tag{4.2}$$

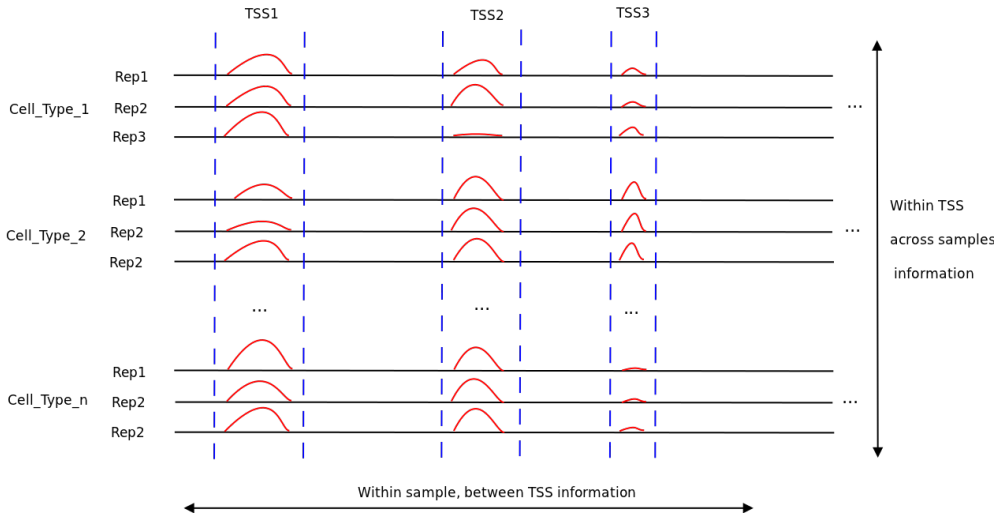


FIGURE 4.8: Comparing CAGE clusters across cell types and replicates. Each horizontal line refers to a single CAGE library (e.g. the library for replicate j within cell type i), with example detected TSSs marked along it. Red densities represent clusters of mapped tags, which corresponds to the expression of its respective TSS within a library. In the example labelled TSS1, differences occur but it is unclear whether differential expression will be detected over noise. In the second example labelled TSS2, there is a potential outlier which in Rep3 of Cell_Type_1 which may affect differential expression analysis. In the example labelled TSS3, a visual inspection of signal vs noise suggests that each cell type may have a (steady state) distinct expression level. The diagram further illustrates how information may be shared (within samples, but also across TSS) in order to aid differential expression detection.

$$\mathrm{NB}(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \tag{4.3}$$

where $p \in [0, 1]$ and $r \sim \{0, 1, 2, 3, \ldots\}$. p can be thought of as the total number of read counts for a given genomic elements.

The approach by Anders and Huber [Anders and Huber, 2010] models the mean count for the j^{th} element of the i^{th} sample as

$$\mu_{ij} = q_{i,\rho(j)s_j} \tag{4.4}$$

where ρj refers to the replicates available for that sample. For the noise estimation, they consider the shot noise, which is the variance observed in read counts assuming equal technical conditions.

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_{i,\rho(j)} \tag{4.5}$$

where $v_{i,\rho(j)} = v_\rho(q_{i,\rho(j)})$.

A further method called *bayseq* is based on an empirical Bayesian approach with the negative binomial distribution, combined with quasi methods, which effectively accounts for over-dispersion and allows for highly conservative estimates (low false-positive rate) [Hardcastle and Kelly, 2010]. Furthermore, it had an `r` package available which efficiently allowed for the set of pair-wise differential expression between all of my samples. I choose this approach due to the fact that I had need to compare many samples pairwise (149 samples in the primary cell set), and therefore a highly conservative approach was required in order to limit the number of false positives.

To illustrate this method for the data used in the analysis differential expression probabilities compared to fold-change for each of the 149 primary cell libraries compared pairwise are given in Appendix D. Each plot is the cumulation of comparing a primary cell type with all other primary cell types. As expected, all plots show high differential expression probabilities for the highest fold change.

Finally, it should be noted that since this project ran many of the calculations for differential expression, many other packages have been published, mostly based on information sharing with the negative binomial distribution [Äijö et al., 2014, Chung et al., 2013, Law et al., 2014, Lee et al., 2011, Vavoulis et al., 2015, Wang et al., 2010], and some greatly improve on accuracy and reduction of false positives by allowing for the down-weighting of observations from samples of poor quality [Law et al., 2014, Liu et al., 2015].

4.5 Applying complexity measures to primary cells data

This section describes the procedure applied to the primary cell data, the focus of the next Chapter.

Select data

Primary cell types provide an excellent platform over which to calculate complexity, since they have minimal ontological heterogeneity; whereas tissues contain mixed primary cell populations, across which differences in expression levels or state may be observed as a result of differential regulation.

Of the primary cells from the FANTOM5 project, a set of 149 distinct cell types were chosen. 138 of these represented normal cells from healthy adults. This was a deliberate decision to understand regulatory complexity from a 'standard' perspective; how genes behave in normal populations.

The remainder represented CD14+ monocytes under a variety of treatments. The aim was to calculate complexity over these two sets separately, to distinguish genes which respond to treatment within the same cell type to those which are observed to be complex across distinct non-treated cells.

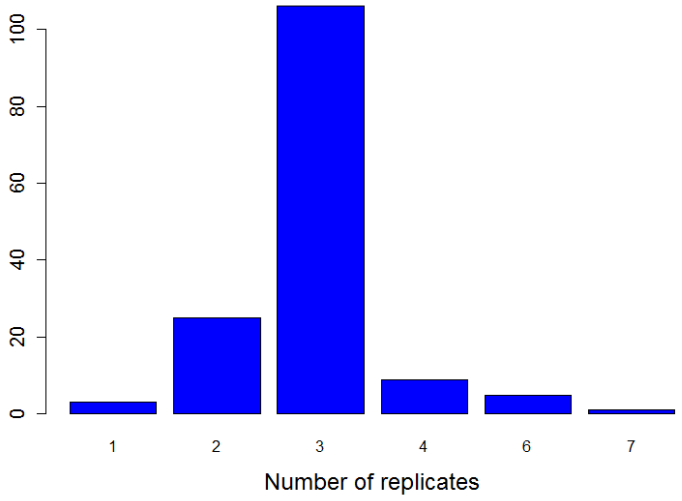


FIGURE 4.9: **Distribution of biological replicates across the 149 primary cells used in this analysis**. Only two had just one replicate, whilst the most common number of replicates was 3.

The distribution of the numbers of replicates per primary cell type is given in Figure 4.9. The names of the primary cell types used in this analysis are given in Appendix C. They

are split by ontological groupings, according to the FANTOM5 resource browser [Forrest et al., 2014] and according to developmental stage - mesodermal or mensochymal. The mesoderm is one of the three early embryonic tissues layers, which, through developmental processes, will differentiate into internal organs, such as the circulatory system and muscles. Much of the mesenchyme is derived from this layer; the mesencymal layer develops into cell types relating to connective tissues, and mesencymal tissues are are characterised by their large extracellular matrix and largely undifferentiated cells [Uccelli et al., 2008]. It is hypothesised that one might expect to observe different reguatory expression patterns in adult termanially differentiated cells. Thus, in order to understand if there is an observed difference in complexity between mesodermal and mensochymal layers, complexity scores were also calculated across the cell types split into these two developmental categories.

$$
\begin{array}{c} \\ gene_1 \\ gene_2 \\ \vdots \\ gene_n \end{array}
\begin{array}{c}
\begin{array}{cccc} sample_1 & sample_2 & \cdots & sample_p \end{array} \\
\begin{pmatrix} N_{11} & N_{12} & \dots & N_{1p} \\ N_{21} & N_{22} & \dots & N_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ N_{n1} & N_{n2} & \dots & Nnp \end{pmatrix}
\end{array}
$$

Measure potential regulatory output

The aim was to mine information of the expression profiles across the chosen set of samples based on:

- Pairwise differential expression probabilities
- Determine on/off states
- Determine sample structure

In order to calculate differential expression, `bayseq` was run pairwise between all samples, which returned the probability of differential expression pairwise for each

gene. For 138 primary cells, this resulting in 15400 possible pairs to test. `bayseq` was run parallel across 24 cores and total calculation time for gene level data took several weeks. In terms of robust TSS level data, total calculation time took several months. In terms of time course data, calculation general took less than a day for either gene or TSS level (due to fewer pairs of libraries). The distribution of mean and median differential expression probabilities is given in Figure 4.10.

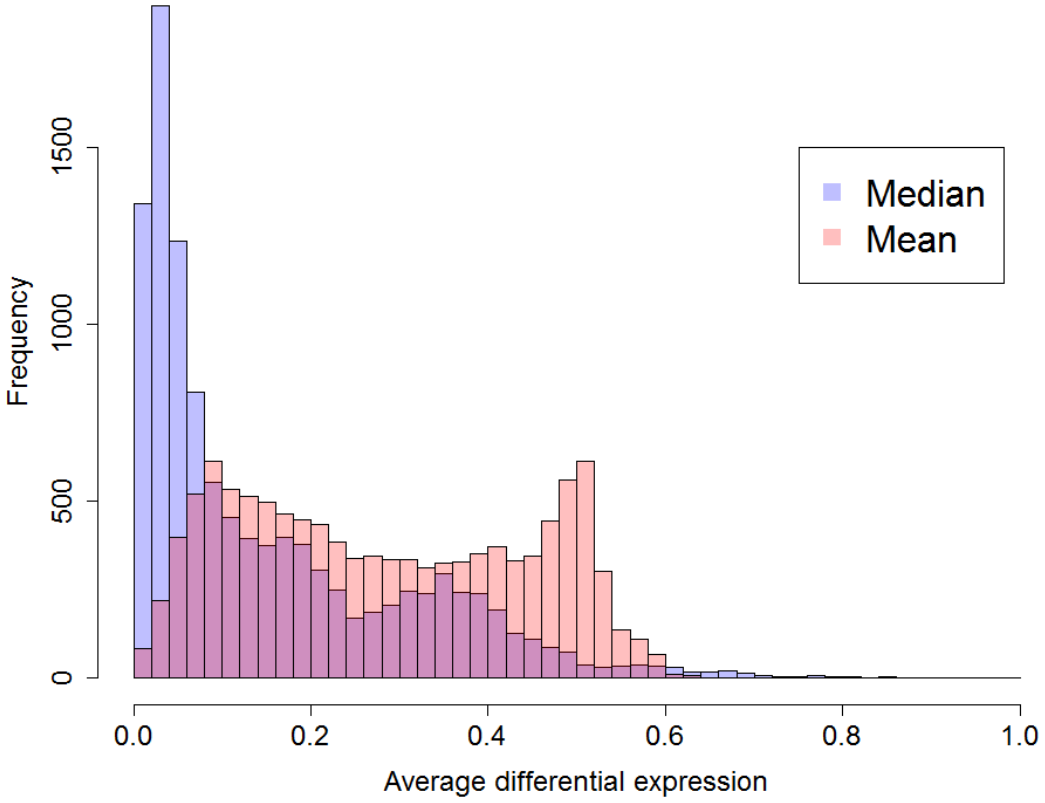


FIGURE 4.10: Distribution of pairwise differential expression probabilities captured used baySeq. The median change is often 0 since this includes 'off-off' probabilities.

Plots of differential expression probabilities against the log fold change in expression are given in Appendix D for all cell types. In order to call whether a pair of states exhibited no change in expression as a result of both states being off, `bayseq` had an option in the code which calculates the probability that both states are null. Furthermore, it was asked that at least two replicates must have a tag count in order to call the presence

of transcription in a cell type. This threshold was used to determine genes which were off in all cell types. These genes were filtered out of the analysis (lack of information about their expression patterns) and not included in the calculations for differential expression (to help speed up processing time). All isoforms for all genes were included in the differential expression calculations, resulting in a total of 27407 rows in the dataset for 138 primary cells, which ranged across 408 libraries (444 libraries for 149 primary cells for the full dataset).

Thus, pairs of states were either recorded as 'off-off', 'on-off' or 'on-on'. In the cases of 'on-off' the probability of differential expression should be 1, in the cases of 'on-on' the probability of differential expression is p^D where p_D is between 0 and 1. Due to noise in the data and low tag counts, 'on-off' states do not always have a probability of differential expression close to 1. These pairwise probabilities are represented in a matrix A:

$$\mathbf{A} = \begin{array}{c} \\ celltype_1 \\ celltype_2 \\ \vdots \\ celltype_m \end{array} \begin{array}{c} \begin{array}{cccc} celltype_1 & celltype_2 & \cdots & celltype_m \end{array} \\ \begin{pmatrix} 0 & p^D_{12} & \dots & p^D_{1m} \\ p^D_{21} & 0 & \dots & p^D_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p^D_{m1} & p^D_{m2} & \dots & 0 \end{pmatrix} \end{array}$$

The decision was made to determine the samples structure from the data itself, by observing pairwise correlations, as opposed to relying in a predetermined model. This was done using a transformation of Pearson's correlation on the log of the counts (plus a pseudocount to account for log not being defined at zero). For two samples, s_x and s_y, the median tag counts across replicates were correlated using the formula

$$\rho_{s_x,s_y} = \frac{\sum_{i=1}^{n}(\log(x_i+1)-\bar{x})(\log(y_i+1)-\bar{y})}{\sqrt{\sum_{i=1}^{n}(\log(x_i+1)-\bar{x})}\sqrt{\sum_{i=1}^{n}(\log(y_i+1)-\bar{y})}} \tag{4.6}$$

where n is the number of transcriptional elements (genes or TSS), x_i is the i^{th} normalised count for sample s_x, y_i is the i^{th} normalised count for sample s_y, $\bar{x} = \frac{1}{n}\sum \log(x_i+1)$, $\bar{y} = \frac{1}{n}\sum \log(y_i+1)$.

Then the weight between the two samples is given as

$$w_{s_x,s_y} = \rho^3_{s_x,s_y} \tag{4.7}$$

This transformation allowed weights to be spread across the 0 to 1 range, making it clearer to distinguish between similar samples in which differential expression occurs in a gene with distantly related samples, which are down-weighted. The weights between all pairs of samples can be put together into a weight matrix

$$\mathbf{W} = \begin{array}{c} \\ celltype_1 \\ celltype_2 \\ \vdots \\ celltype_m \end{array} \begin{array}{c} \begin{array}{cccc} celltype_1 & celltype_2 & \cdots & celltype_m \end{array} \\ \begin{pmatrix} 1 & w_{12} & \dots & w_{1m} \\ w_{21} & 1 & \dots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & 1 \end{pmatrix} \end{array}$$

Generate weighted graph and calculate graph based connectivity measures

The matrix **WA** is then used to create a weighted adjacency matrix over which eigenvector centrality was calculated

$$\mathbf{WAx} = \lambda \mathbf{x} \tag{4.8}$$

Create weighted adjacency and/or Laplacian matrices

Select α, β and calculate final measure

Or use overall graph connectivity λ.

Inference on final scores

Connect with biological information.

4.6 Summary of chapter

First, the FANTOM5 data was described, which has the advantage of quantitatively estimating steady state transcript levels at a single TSS resolution across the whole genome and includes an extensive range of primary cell types, tissues and time courses, with technical and biological replication.

Next, the normalisation and tag clustering (TSS identification) strategies applied across libraries were described.

Following this, methods for estimating differential expression were discussed. I described my own models for calculating differential expression, which are more applicable to time course data (due to memory issues, amongst other things). The methods used to calculate differential expression across genes and clusters in primary cell types was then described.

Finally, the data used in the rest of this thesis was selecting and it was described how it was used to generate the complexity scores and their normalised counterparts that are interrogated extensively in the next chapter.

Chapter 5

Complexity applied to primary cell types

The complexity scores described in Chapter 4 were applied as metrics to assess the observed output of regulatory complexity in the gene expression data for primary cells collected by the FANTOM5 consortium [Forrest et al., 2014]. The data consists of 138 primary distinct primary cells and the scores attempt to take into account both the observed changes and structure of expression observed for a given gene. Details of the method applied is given in Chapter 4. Primary cell types are highly desirable as opposed to tissues or less refined cell types as these represent pure lineages, separating out different cell lineages with the potential to be differentially expression within a given tissue. The FANTOM5 dataset provides a unique selling point in this regard, with its high throughput nature offering a wide variety of primary cells with a median of three replicates.

Presented in the results of the analysis are three main scores; the first is the raw complexity score without normalising, second is the normalised version of complexity which adjusts for the complexity potential across different breadths of expression and the third is the standard entropy score, generally used to measure sample restricted vs ubiquitous expression. All of these scores results in a continuous distribution over

which different categories and types of genes may be compared. Results and Figures are presented, before being discussed in more detail in Chapter 6.

In the first instance, the distributions of scores are observed and compared with expression breadth. The top 10% and bottom 10% of each distribution are checked for functional enrichment, in order to attempt to understand the biological relevance of a gene achieving a low or a high scores. This is followed by an in-depth analysis of the properties of a gene 'complex' in its regulation - including physical gene properties, promoter annotations, hypersensitivity, histone modifications, protein age. Finally presented is an analysis of the kinds of diseases enriched in high complexity scores and a discussion of the possible usefulness of having such a score for the identification of candidate genes from a medical perspective.

5.1 Number of genes in this analysis

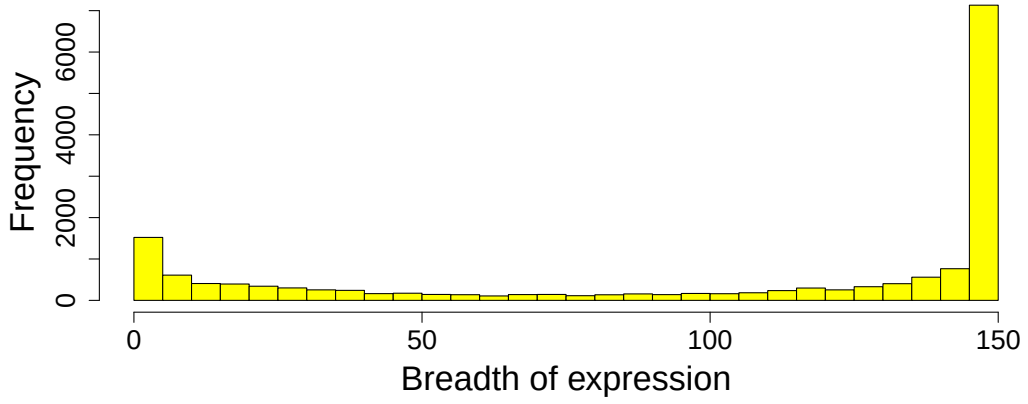


FIGURE 5.1: Histogram showing the **breadth of expression across** set of genes. Breadth of expression runs from 1 (expressed in a single primary cell type) to 149 (expressed in all primary cell types considered in this study). Heights of bars represents the number of genes observed with the given breadth of expression.

This analysis is based on expression estimates for genes in the primary cell data from the FANTOM5 project. Expression levels are captured through summing CAGE tags in the promoter regions of genes based on coordinates from the $refSeq$ database. The

total number of coding transcripts covered in this analysis with expression present in at least one primary cell type are 27407, corresponding to a total of 16111 distinct coding genes, resulting in a mean of 1.7 transcripts per gene.

Genes not expressed in any cell type are not considered in the analysis, due to lack of information of their expression profile - it is assumed that these genes are expressed in at least some cell type under some given conditions, but the FANTOM5 dataset did not capture this pattern of expression. Furthermore, in order to avoid bias from genes which have a large number of isoforms, all highly correlated in their expression, much of the analysis considers only the 16111 distinct genes which are expressed in at least one cell type. When more than one isoform is present, the median complexity score is observed for that gene.

A histogram of the breadth of gene expression, given as the proportion of expressed primary cell types, is given in Figure 5.1. Of these 16111 genes, 5161 (32.0%) are expressed ubiquitously - that is, a positive median expression across biological replicates was observed in all primary cell types. In contrast there are 562 (3.5%) genes with observed expression in just one cell type.

5.2 Distributions of complexity scores and entropy score

Histograms showing the structure of each of the three score distributions is given in Figure 5.2. Both complexity and normalised complexity have a peak at zero for genes which exhibited no complex behaviour. Both scores have a large peak - around 0.4 for complexity and 0.3 for normalised complexity. Complexity has a smaller peak around 0.7 and normalised complexity is right skewed, with very few genes which are 'highly complex'. Entropy is highly concentrated around maximum, due to the large number of ubiquitously expressed genes. Tissue restricted genes form a small peak around zero.

Complexity is plotted against the percentage of expressed cell types and displayed in Figure 5.3. The distribution forms a 'boomerang' shape against breadth, with highly restricted genes and ubiquitous gene achieving the lowest scores and maximised between

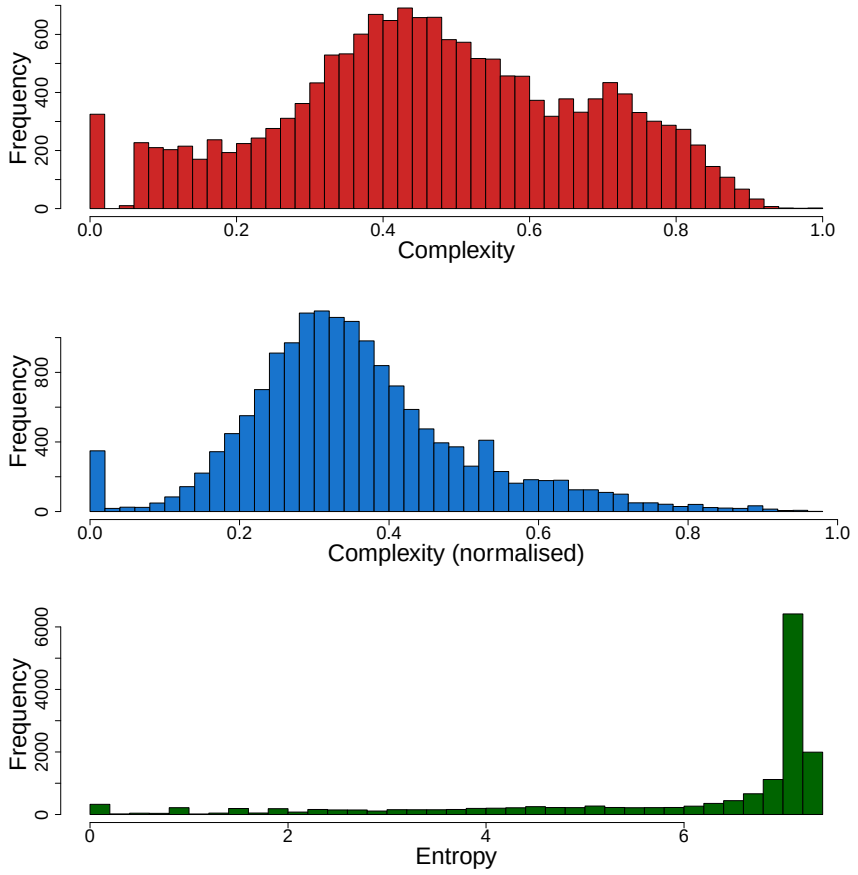


FIGURE 5.2: Histograms displaying **the distribution of each of the three scores** - complexity (red), normalised complexity (blue) and entropy (green).

60 and 100% - these are genes exhibiting complex behaviour of on-off switching, and/or a lot of differential expression between expressed cell types.

Figure 5.4 shows the same plot of breadth of expression against raw complexity scores, including the maximum practical (red) and theoretical (blue) bounds at each given expression breadth, plotted as a smooth curve. The calculation for these bounds was given in Chapter 2.

The complexity score is normalised by the practical complexity score on a per-gene basis and referred to as normalised complexity. Normalised complexity is plotted against expression breadth in Figure 5.5. In the normalised version of the score, highly restricted genes now have a much higher score; these genes form the right hand tail of

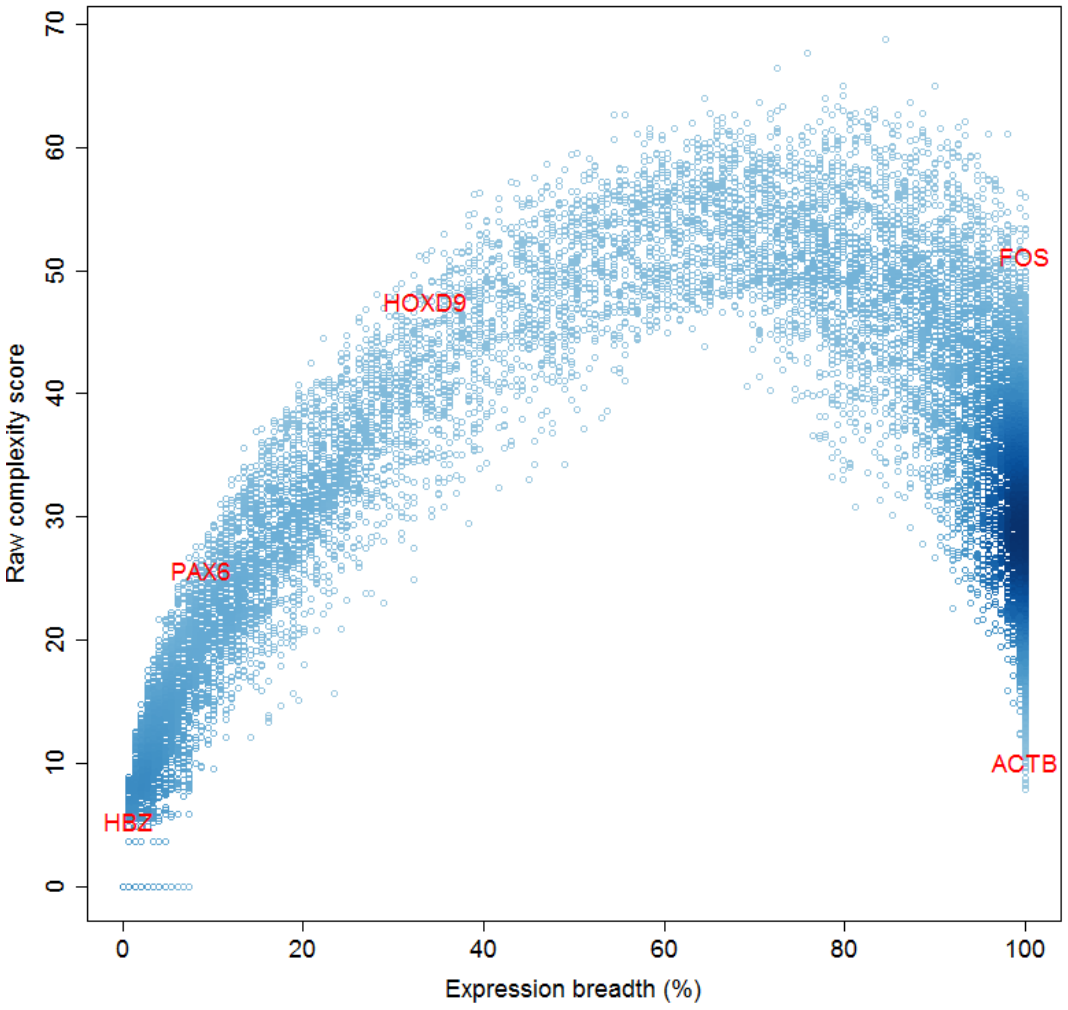


FIGURE 5.3: **Breadth of expression** (percentage of expressed primary cell types) **against raw complexity scores**. Darker blue regions represent regions containing many genes; in particular the dark region on the right hand edge represents broadly/ubiquitously expressed genes across the set of primary cell types.

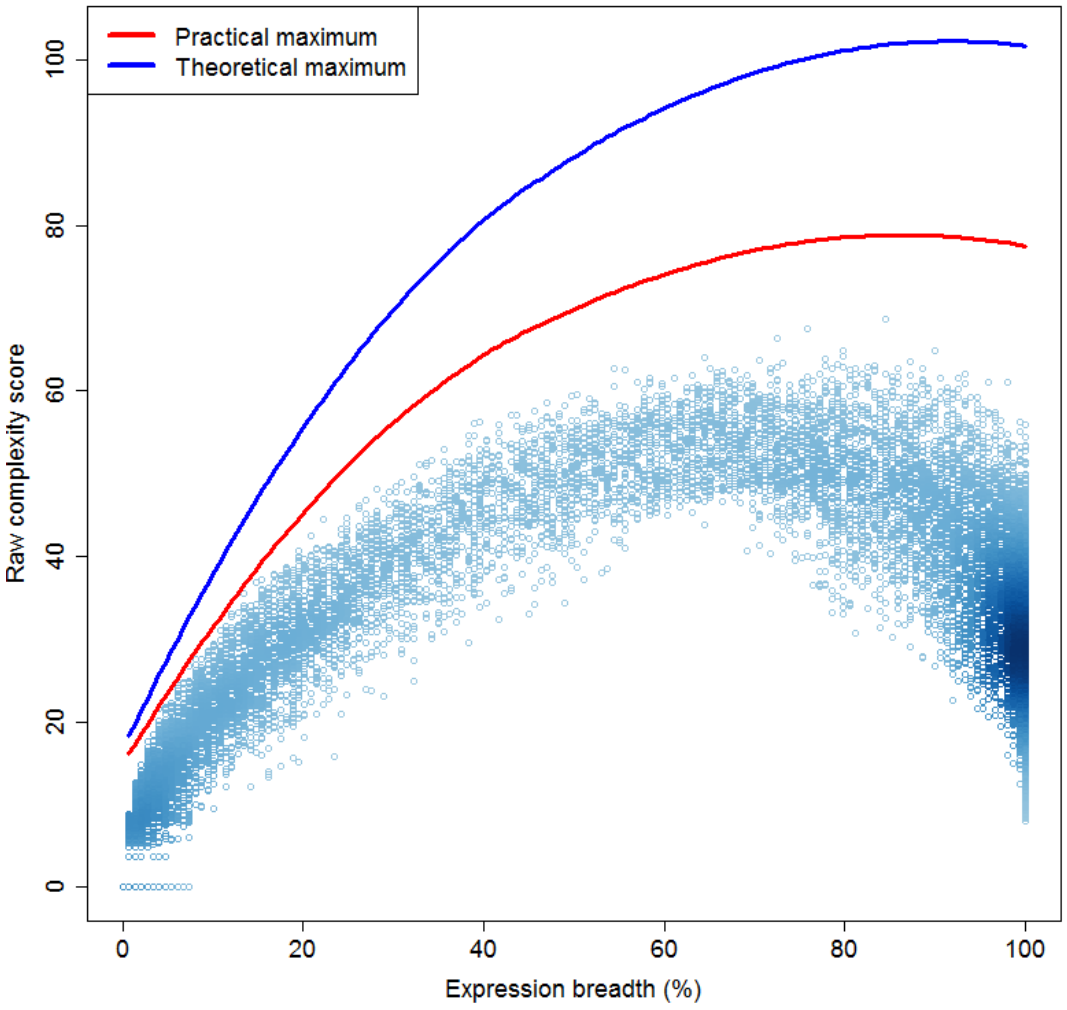


FIGURE 5.4: **Breadth of expression** (percentage of expressed primary cell types) **against complexity scores, illustrating possible normalisation strategies**. The blue line shows the theoretical maximum which may be achieved if all pairs were differentially expressed, the red line shows the practical maximum, maximised across expression breadths. Smooth curves are plotted based on the outside edge of maximum scores.

the distribution observed in Figure 5.2.

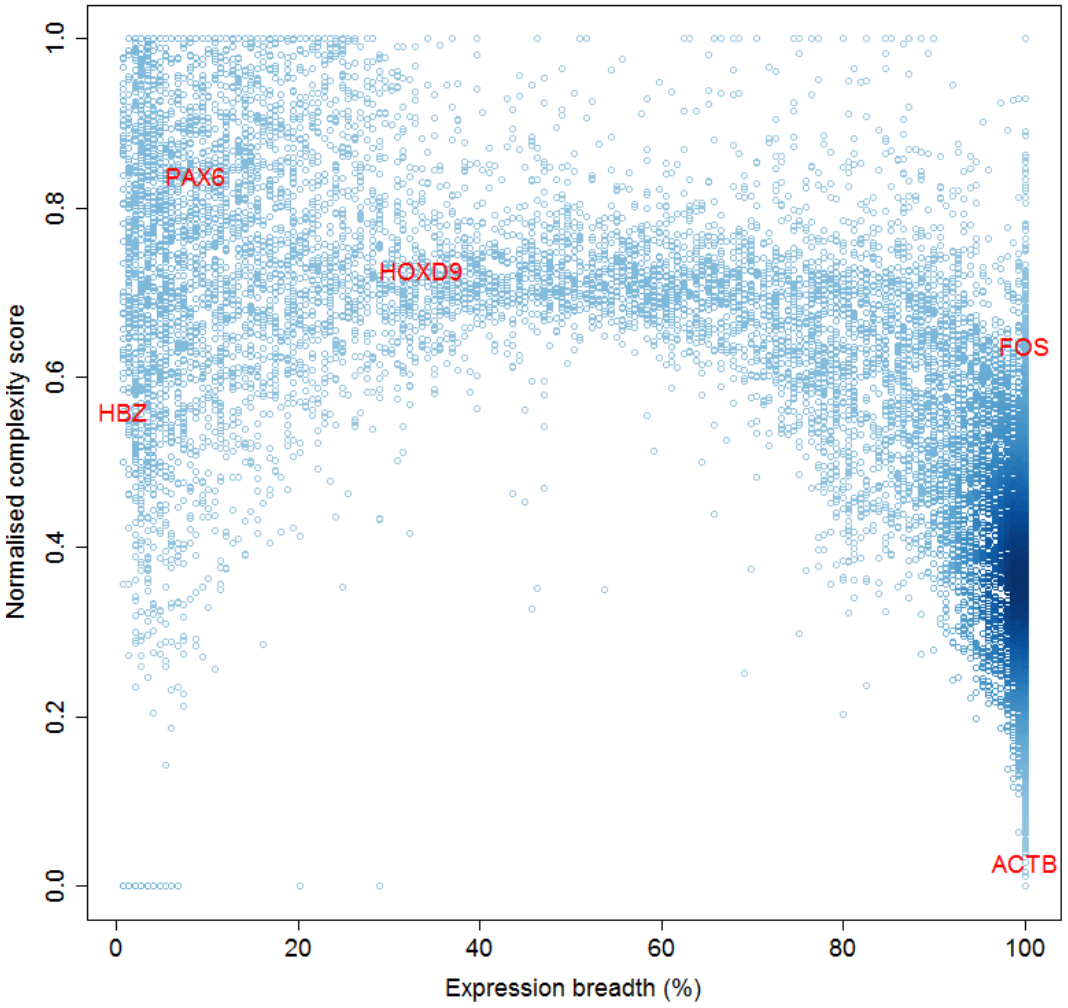


FIGURE 5.5: **Breadth of expression** (percentage of expressed primary cell types) **against normalised complexity scores**. Each gene is normalised by its own maximum possible level based on redistributions of tag counts across its own breadth of expression, as described in Chapter 3.

Another score of interest, but considered less within the context of this analysis, is the 'locally normalised complexity score'. Instead of normalising by all possible complexity potentials, each gene is normalised by the maximum possible complexity for its given range of expressed cell types; this score is essentially a measure of complexity observed

through its differential expression distribution between expression cell types alone, since it removes the effect of on-off switching.

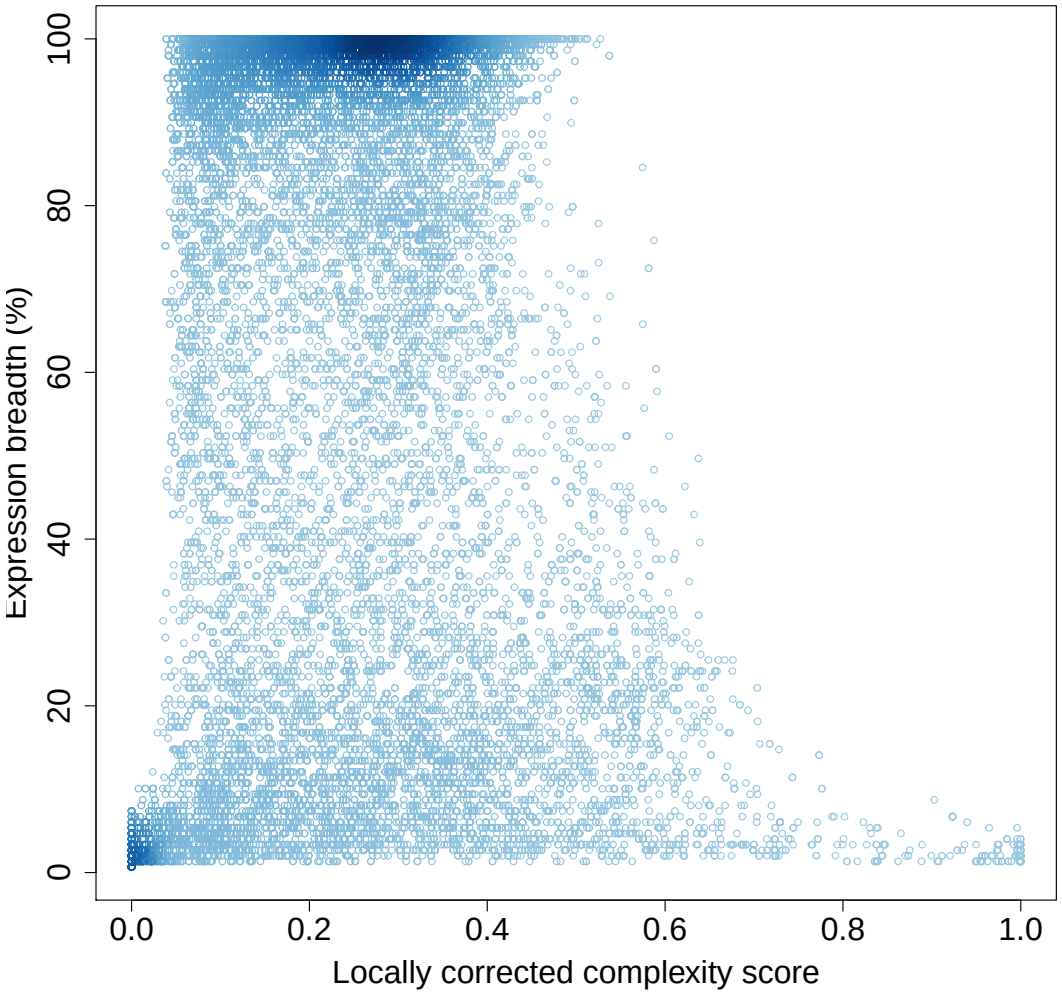


FIGURE 5.6: Breadth of expression (number of expressed primary cell types) against locally normalised complexity scores. These are normalized accounting for changes in expression between expressed cell types, but ignoring the patterns of on and off switching occurring between cell types.

To compare the normalised complexity scores with the locally normalised complexity scores, Figure 5.7 shows the two scores plotted against each other, with a line plotted through the diagonal. Genes which are more strongly complex as a result of observed differential expression as opposed to observed on-off switching are those towards the

right of the plot. Note that ubiquitous genes form the dark patch on the diagonal, since the scores are highly correlated for these genes.

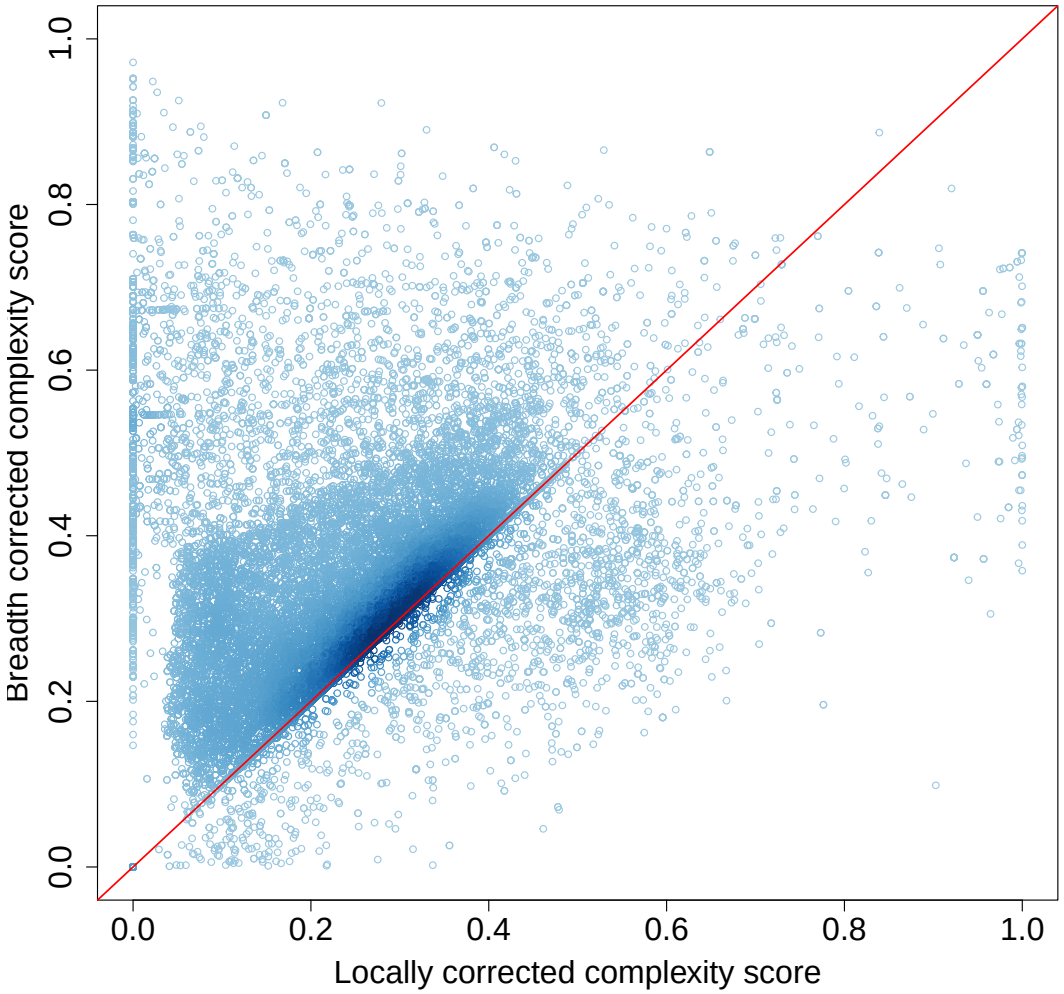


FIGURE 5.7: Complexity corrected globally vs complexity corrected locally. Global scores are corrected by considering the possible combinations of on-off switching, local scores are corrected by considering only the maximal possible differential expression between on-on cell types. A skew towards locally corrected scores suggest that a gene is complex as a result of differential expression between on-on cell-types.

From this section, we can observe that complexity clearly has a non-linear relationship with entropy scores. However, the correlation between them is low, as observed by the large range of complexity scores within each given breadth. Thus, complexity is able to

distinguish between genes of different regulatory capabilities but of the same breadth of expression.

In the next section, an enrichment analysis is carried out, to understand which kinds of genes are highly complex according to functional annotations.

5.3 Functional annotation enrichment and contour plots

The web service *Gorilla* [Eden et al., 2007, 2009] was used in order to carry out a GO term analysis; a query of whether high or low complex genes were statistically enriched in biological processes or function.

5.3.1 Enrichment for complexity scores

The top 20 GO terms for high scoring complexity genes are given in Table 5.1, using the option 'searching for enriched GO terms that appear densely at the top of a ranked list of genes'. The most significant terms relate to developmental processes and the regulation of developmental processes, including the development of anatomical structures. This is expected, since expression changes observed in adult primary cell types are likely to be highly correlated by developmental switches through development (reference).

TABLE 5.1: Top 20 most significant GO terms from the output of *Gorilla*, for the genes with the **highest complexity scores**, based on a single list of **all expressed genes** ranked from high to low complexity.

GO Term	Description	P-value
GO:0032502	developmental process	1.50E-34
GO:0044767	single-organism developmental process	4.35E-32
GO:0051239	regulation of multicellular organismal process	1.98E-31
GO:2000026	regulation of multicellular organismal development	2.47E-30
GO:0030334	regulation of cell migration	6.95E-30
GO:0048856	anatomical structure development	6.95E-30
GO:0030198	extracellular matrix organization	2.04E-29
GO:0043062	extracellular structure organization	2.04E-29
GO:2000145	regulation of cell motility	6.26E-29
GO:0005615	extracellular space	1.45E-29
GO:0032501	multicellular organismal process	1.72E-28
GO:0051270	regulation of cellular component movement	1.93E-27
GO:0044707	single-multicellular organism process	2.02E-27
GO:0009653	anatomical structure morphogenesis	5.21E-27
GO:0040012	regulation of locomotion	5.77E-27
GO:0031012	extracellular matrix	2.42E-26
GO:0050793	regulation of developmental process	3.38E-25
GO:0016477	cell migration	3.78E-24
GO:0042127	regulation of cell proliferation	1.84E-23
GO:0007165	signal transduction	1.70E-22

Table 5.2 shows the top 20 GO terms associated with the lowest rank complexity genes. The highest term is 'extracellular region' and terms relating to plasma membrane components and transporter activity are also high. Membranes function to protect the interior of cells from their outside environment and contain proteins which contribute to cell signalling, adhesion, cellular transport and various regulatory functions. This suggest that genes coding for extracellular components act in a housekeeping type manner, important to every cell.

TABLE 5.2: Top 20 most significant GO terms from the output of *Gorilla*, for the genes with the **lowest complexity scores**, based on a single list of **all expressed genes** ranked from low to high complexity.

GO Term	Description	P-value
GO:0005576	extracellular region	3.98E-28
GO:0031224	intrinsic component of membrane	2.59E-22
GO:0016021	integral component of membrane	4.09E-21
GO:0050907	detection of chemical stimulus involved in sensory perception	1.18E-21
GO:0005887	integral component of plasma membrane	3.65E-19
GO:0031226	intrinsic component of plasma membrane	9.14E-19
GO:0005215	transporter activity	1.54E-19
GO:0022892	substrate-specific transporter activity	5.60E-19
GO:0015075	ion transmembrane transporter activity	1.74E-18
GO:0038023	signaling receptor activity	1.16E-17
GO:0022891	substrate-specific transmembrane transporter activity	2.08E-17
GO:0022857	transmembrane transporter activity	6.03E-17
GO:0004930	G-protein coupled receptor activity	1.52E-16
GO:0004984	olfactory receptor activity	3.48E-16
GO:0007186	G-protein coupled receptor signaling pathway	4.25E-16
GO:0004872	receptor activity	1.66E-15
GO:0008324	cation transmembrane transporter activity	2.24E-15
GO:0044425	membrane part	2.77E-15
GO:0006811	ion transport	1.35E-15
GO:0098655	cation transmembrane transport	1.71E-14

Table 5.3 shows the top GO terms for the most complex genes according to the normalised complexity score. Similar to the non-normalised complexity, the top terms relate to multicellular processes and anatomical development, suggesting that the normalisation strategy does not dramatically change the ranking of genes in terms of their complexity.

5.3.2 Enrichment for normalised complexity scores

TABLE 5.3: Top 20 most significant GO terms from the output of *Gorilla*, for the genes with the **highest normalised complexity scores**, based on a single list of **all expressed genes** ranked from high to low normalised complexity.

GO Term	Description	P-value
GO:0032501	multicellular organismal process	1.72E-49
GO:0044707	single-multicellular organism process	1.84E-45
GO:0003008	system process	2.64E-40
GO:0044459	plasma membrane part	1.70E-26
GO:0050877	neurological system process	3.64E-24
GO:0048856	anatomical structure development	8.41E-24
GO:0005886	plasma membrane	1.39E-24
GO:0007600	sensory perception	3.90E-21
GO:0044700	single organism signaling	1.15E-20
GO:0023052	signaling	1.60E-20
GO:0032502	developmental process	2.34E-20
GO:0007267	cell-cell signaling	5.45E-19
GO:0044057	regulation of system process	4.04E-18
GO:0031424	keratinization	1.30E-16
GO:0044767	single-organism developmental process	5.43E-16
GO:0005887	integral component of plasma membrane	2.88E-16
GO:0031224	intrinsic component of membrane	2.80E-15
GO:0007154	cell communication	1.54E-15
GO:0009888	tissue development	1.64E-15
GO:0048869	cellular developmental process	8.58E-15

Table 5.4 shows the top GO terms for the lowest ranked genes according to the normalised complexity scores. Whilst not as overall significant in terms of their P-values, the most significant GO terms relate to the membrane and ribosomal processes. These processes are associated with translation and enriched in genes with 'ultra-housekeeping' type behaviour, as whilst translation itself is a highly regulated process, the ribosome represents a key component of non-replicating cells.

TABLE 5.4: Top 20 most significant GO terms from the output of *Gorilla*, for the genes with the **lowest normalised complexity scores**, based on a single list of **all expressed genes** ranked from low to high normalised complexity.

GO term	Description	P-value
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	3.73E-10
GO:0044391	ribosomal subunit	1.32E-10
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	5.19E-10
GO:0006613	cotranslational protein targeting to membrane	6.68E-10
GO:0070972	protein localization to endoplasmic reticulum	9.25E-10
GO:0045047	protein targeting to ER	1.20E-09
GO:0050877	neurological system process	1.73E-09
GO:0072599	establishment of protein localization to endoplasmic reticulum	3.86E-09
GO:0016071	mRNA metabolic process	4.29E-09
GO:0003008	system process	9.17E-09
GO:0022627	cytosolic small ribosomal subunit	1.08E-08
GO:0016021	integral component of membrane	1.15E-08
GO:0030529	ribonucleoprotein complex	1.22E-08
GO:0044445	cytosolic part	1.79E-08
GO:0031224	intrinsic component of membrane	2.40E-08
GO:0006612	protein targeting to membrane	2.23E-08
GO:0051606	detection of stimulus	2.38E-08
GO:0006413	translational initiation	5.79E-08
GO:0006402	mRNA catabolic process	1.14E-07
GO:0007601	visual perception	4.95E-07

5.3.3 Enrichment for entropy scores

Table 5.5 shows the top associated GO terms for genes with the highest entropy scores. The top term is 'poly(A) RNA binding', followed closely by 'RNA binding'. RNA binding as been associated with significant and consistent expression across cell types [Kechavarzi and Janga, 2014].

TABLE 5.5: Top 20 most significant GO terms from the output of *Gorilla*, for the genes with the **highest entropy scores**, based on a single list of **all expressed genes** ranked from high to low entropy.

GO Term	Description	P-value
GO:0044822	poly(A) RNA binding	5.25E-100
GO:0003723	RNA binding	2.60E-99
GO:0044446	intracellular organelle part	1.63E-93
GO:0044422	organelle part	2.79E-89
GO:0032991	macromolecular complex	1.06E-88
GO:0010467	gene expression	1.39E-82
GO:0016071	mRNA metabolic process	7.24E-82
GO:0030529	ribonucleoprotein complex	7.95E-71
GO:0044764	multi-organism cellular process	7.49E-56
GO:0016032	viral process	7.59E-56
GO:0044403	symbiosis, encompassing mutualism through parasitism	7.59E-56
GO:0044424	intracellular part	1.34E-56
GO:0044265	cellular macromolecule catabolic process	7.77E-55
GO:0016482	cytoplasmic transport	1.19E-53
GO:0044428	nuclear part	1.37E-53
GO:0046907	intracellular transport	8.62E-52
GO:0006397	mRNA processing	4.69E-51
GO:0044419	interspecies interaction between organisms	4.32E-50
GO:0008380	RNA splicing	2.24E-48
GO:1902582	single-organism intracellular transport	2.39E-47

Table 5.6 shows the GO terms associated with the lowest entropy scores. The top few terms relate to extracellular region and membrane components. These terms are similar to those observed in Table 5.3, for the highest normalised complexity scores, probably because the normalisation up-weights genes highly restricted in their expression, which have low entropy scores.

In summary, it appears that genes on either end of the complexity access are significantly enriched in biological processes and capture distinct information from entropy

TABLE 5.6: Top 20 most significant GO terms from the output of *Gorilla*, for the genes with the **lowest entropy scores**, based on a single list of **all expressed genes** ranked from low to high entropy.

GO Term	Description	P-value
GO:0005576	extracellular region	2.17E-40
GO:0031224	intrinsic component of membrane	1.00E-29
GO:0016021	integral component of membrane	3.17E-28
GO:0044425	membrane part	4.94E-23
GO:0032501	multicellular organismal process	2.64E-23
GO:0003008	system process	1.36E-22
GO:0044459	plasma membrane part	6.31E-22
GO:0031226	intrinsic component of plasma membrane	1.04E-21
GO:0005887	integral component of plasma membrane	1.39E-21
GO:0005215	transporter activity	2.17E-21
GO:0038023	signaling receptor activity	3.97E-20
GO:0022892	substrate-specific transporter activity	7.00E-20
GO:0015075	ion transmembrane transporter activity	7.71E-20
GO:0044707	single-multicellular organism process	4.68E-20
GO:0050907	detection of chemical stimulus involved in sensory perception	9.03E-20
GO:0007600	sensory perception	1.78E-19
GO:0006811	ion transport	4.03E-19
GO:0022891	substrate-specific transmembrane transporter activity	1.75E-19
GO:0022857	transmembrane transporter activity	7.96E-19
GO:0004872	receptor activity	2.16E-18

scores. In particular, a complex gene is associated with a range of ontology terms relating to developmental processes, cell signaling and mobility.

Next, a similar analysis has been carried out, but focussing only on genes which are ubiquitously expression across all the primary cell types included in the analysis.

5.3.4 Functional enrichment for high scoring ubiquitously expressed genes

In order to gain insight into the differences between genes with the same expression breadth but different complexity scores ranked lists of ubiquitously expressed genes according to their complexity scores were used to perform a GO term analysis. Since these genes are always 'on', their complexity score is driven only by differential expression observed between 'on' states; therefore together with their large sample size

(5161 distinct genes, not account for isoforms) and independence to entropy, they are an interesting set to study.

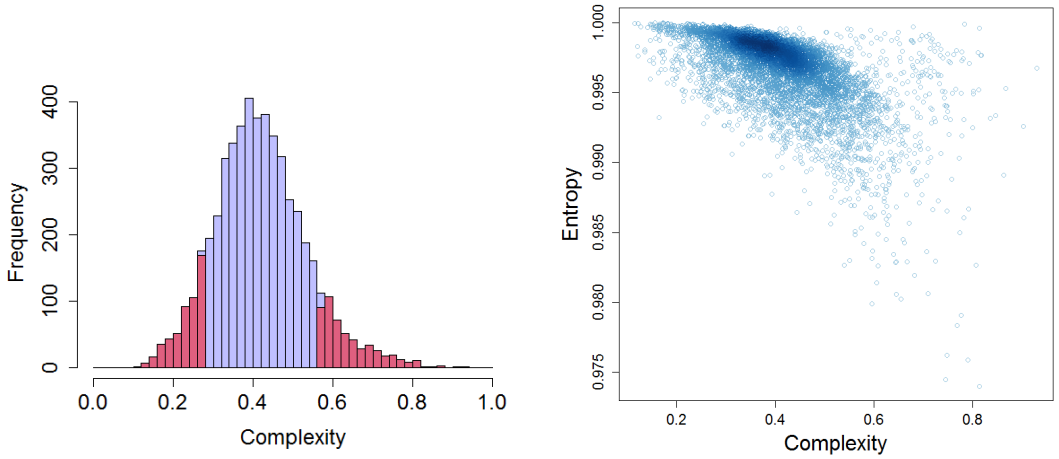


FIGURE 5.8: **Histogram of complexity scores for the ∼35% genes expressed in all primary cell types (left)** (ubiquitous genes). The location of the top and bottom 10% of complex genes are marked in dark pink. Normalisation is generally not required within given expression breadths, therefore only complexity scores are analysed. **Entropy scores against complexity scores for ubiquitous genes (right)**. Whilst complexity scores are spread out across the full range (between 0 and 1), entropy scores are all within 0.975 and 1.000.

The scores for ubiquitous expression complexity form an approximate normal distribution (Figure 5.8), with the lowest 10% reaching around 0.3 and the highest 10% having a complexity from around 0.57. The distribution of ubiquitous expression complexity vs entropy (Figure 5.8) shows that entropy has a very tight range, between 0.975 and 1, whilst complexity covers a broad range, between almost the full potential of between 0 and 1. This implies that whilst entropy is not useful in distinguishing differences between ubiquitously expressed genes, complexity is able to rank them in terms of their observed differential expression distribution.

The most significantly associated GO terms for the highest ranked ubiquitously expressed complex genes are given in Table 5.7. Similar to the complexity results, multicellular organism processes appear at the top of the list. Response to chemical, external stimulus and endogenous stimulus also appear near the top of the list, suggesting that

highly complex ubiquitously expressed genes exhibit up- and down- regulation of expression between cell types as a results of environmental and internal response signals.

TABLE 5.7: Top 20 most significant GO terms from the output of *Gorilla*, for the genes with the **highest complexity scores**, based on a single list of **ubiquitously expressed genes** ranked from high to low complexity.

GO Term	Description	P-value
GO:0032501	multicellular organismal process	2.40E-13
GO:0044707	single-multicellular organism process	3.20E-12
GO:0042221	response to chemical	5.27E-11
GO:0009605	response to external stimulus	3.06E-10
GO:0048523	negative regulation of cellular process	3.39E-10
GO:0009719	response to endogenous stimulus	4.08E-10
GO:0042127	regulation of cell proliferation	5.56E-10
GO:0005886	plasma membrane	2.54E-10
GO:0000786	nucleosome	1.51E-09
GO:0051239	regulation of multicellular organismal process	1.14E-09
GO:0014070	response to organic cyclic compound	1.20E-09
GO:0050896	response to stimulus	1.24E-09
GO:0010033	response to organic substance	1.35E-09
GO:0032502	developmental process	2.27E-09
GO:0044767	single-organism developmental process	2.93E-09
GO:0048856	anatomical structure development	3.46E-09
GO:1901700	response to oxygen-containing compound	3.59E-09
GO:0048518	positive regulation of biological process	7.36E-09
GO:0065007	biological regulation	8.98E-09
GO:0048519	negative regulation of biological process	9.35E-09

Table 5.8 shows the GO terms associated with the least complex ubiquitously expressed genes. These terms involve highly conserved processes required by all cells within an organism, such as translational initiation.

Taken together, we see that ubiquitous complex genes and complex genes are enriched in different GO terms. In particular, developmental process genes are highly complex overall, and genes associated with multicellular processes are enriched in ubiquitous complexity scores. The latter may not be surprising, since these are genes expressed in all cell types and thus would expected to be associated with multi-cellularity.

TABLE 5.8: Top 20 most significant GO terms from the output of *Gorilla*, for the genes with the **lowest complexity scores**, based on a single list of **ubiquitously expressed genes** ranked from low to high complexity.

GO term	Description	P-value
GO:0006413	translational initiation	9.61E-14
GO:0016071	mRNA metabolic process	1.04E-13
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	1.19E-13
GO:0019083	viral transcription	2.22E-13
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	2.29E-13
GO:0006613	cotranslational protein targeting to membrane	3.08E-13
GO:0045047	protein targeting to ER	4.15E-13
GO:0072599	establishment of protein localization to endoplasmic reticulum	1.01E-12
GO:0019058	viral life cycle	1.06E-12
GO:0070972	protein localization to endoplasmic reticulum	1.54E-12
GO:0006612	protein targeting to membrane	1.81E-12
GO:0006412	translation	8.21E-12
GO:0044391	ribosoma subunit	1.81E-14
GO:0030529	ribonucleoprotein complex	1.29E-13
GO:0044445	cytosolic part	9.89E-12
GO:0006412	translation	8.21E-12
GO:0010467	gene expression	1.39E-11
GO:0006414	translational elongation	6.08E-11
GO:0043624	cellular protein complex disassembly	7.99E-11
GO:0032984	macromolecular complex disassembly	8.53E-11

5.3.5 Contour plots

Next it was questioned where the genes related to significant GO terms fell on the complexity vs expression breadth axis. To this end, contours were superimposed onto the distribution (Figures 5.9 to 5.13). Curves are coloured according to the magnitude of density of genes related to a specific GO term; regions within light red circles contain a high density of GO term specific genes and dark regions contain cover locations where they are present but more sparsely located .

Contours relating to the location of known housekeeping related genes (Figure 5.9), as expected, were concentrated around the dense region (dark blue) of all genes with high breadth of expression, although a small number of genes are concentrated around the expression restricted, low complexity region.

For comparison, contours for genes typically thought of as 'master regulators', including the HOX, SOX and PAX gene families, were also generated (Figure 5.10). These genes are known targets for regulation by polycomb, which when combined with activation marks is linked to poised transcription [Stock et al., 2007], and thus exhibit highly regulated gene expression profiles, and so these genes should be enriched in genes with high complexity scores. Whilst master regulators are tightly regulated, some of these genes are nonetheless appear highly concentrated in the ubiquitously expressed, very low complexity region; suggesting that these genes are expressed with little variability in all cell types. This could be due to the lack of sensitivity in the method to detect changes between cell types, or due to the CAGE profiling of genes. Note however that aside from these genes, contours for master regulators are generally accumulated on the right hand edge of the distributions, suggesting that these master regulatory genes are generally complex independent of breadth of expression. When plotting these same contours substituting complexity scores for normalised complexity scores, these genes heavily concentrated on the right hand, more complex, edge of the plot.

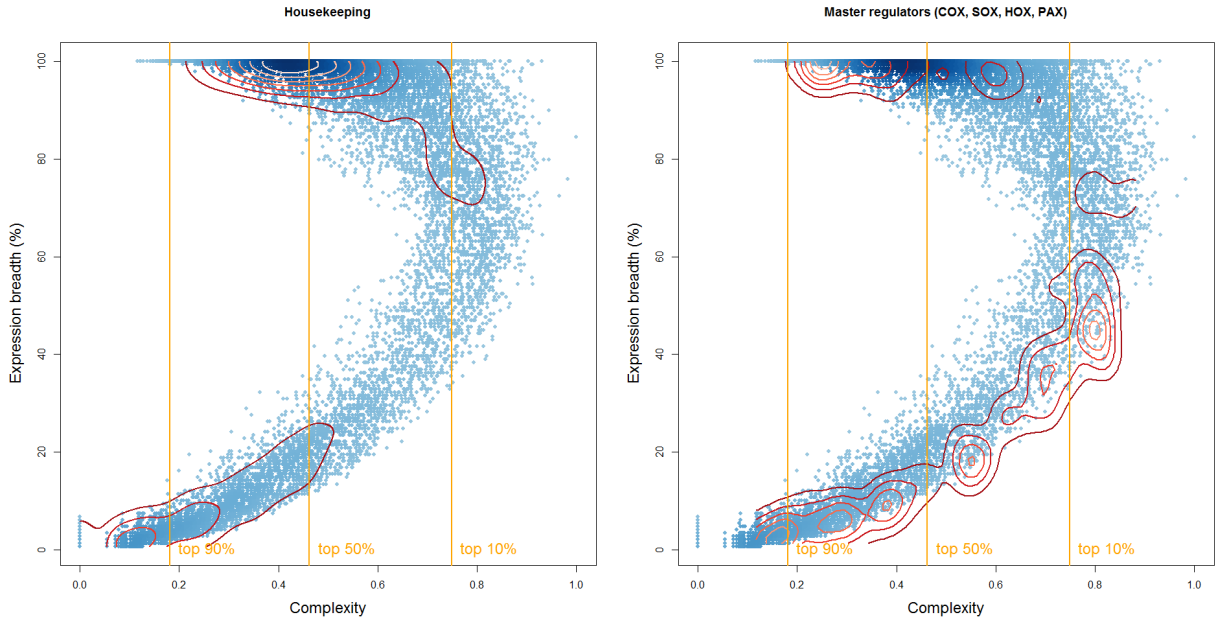


FIGURE 5.9: **Complexity vs breadth with contours for housekeeping genes (left) and master regulatory genes (right)**. Complexity is plotted against the percentage of expressed cell types. Contours showing regions enriched in genes associated with housekeeping tasks or master regulatory genes (defined as HOX, SOX and PAX related genes) are plotted in red. Yellow vertical line proportion areas of the plot according to complexity scores - the highest and lowest 10% most complex genes on the right and left sections respectively, and the centre two regions each containing 40% of the genes. Note that the highest density contours relating to housekeeping genes are generally concentrated within the 10% - 50% region of complexity scores.

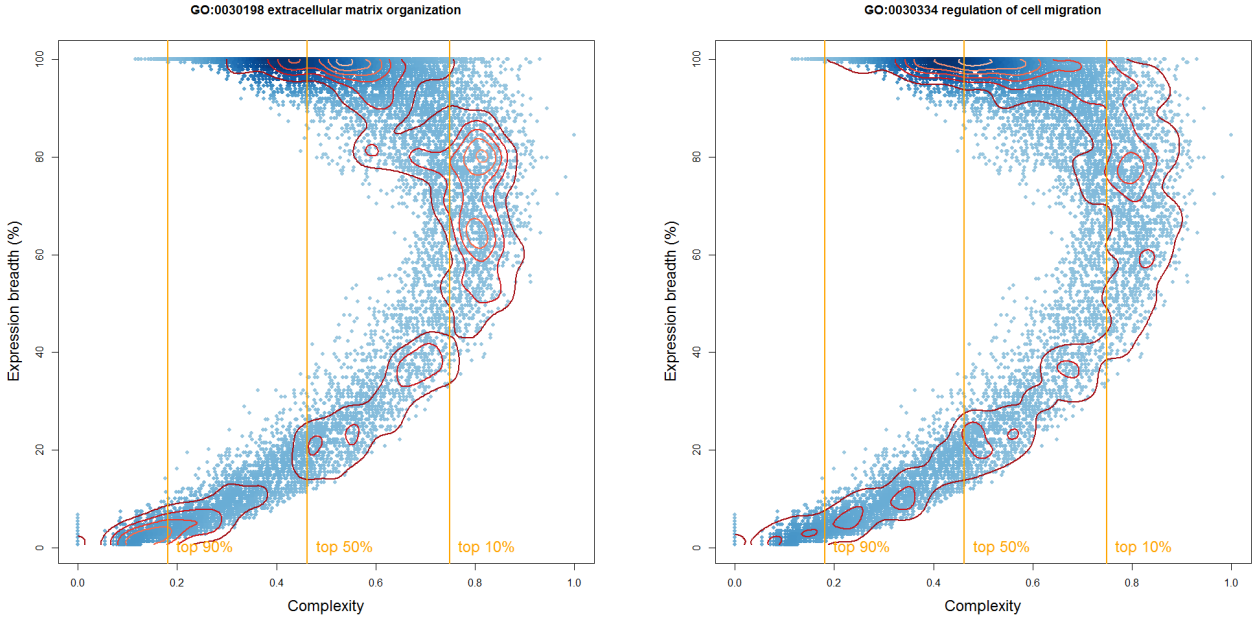


FIGURE 5.10: **Complexity vs breadth with contours for GO terms GO:0030198 extracellular matrix organization (left) and GO:0030334 regulation of cell migration (right)**. Complexity is plotted against the percentage of expressed cell types. Contours showing regions enriched in genes associated with respective GO terms are plotted in red. Yellow vertical line proportion areas of the plot according to complexity scores - the highest and lowest 10% most complex genes on the right and left sections respectively, and the centre two regions each containing 40% of the genes.

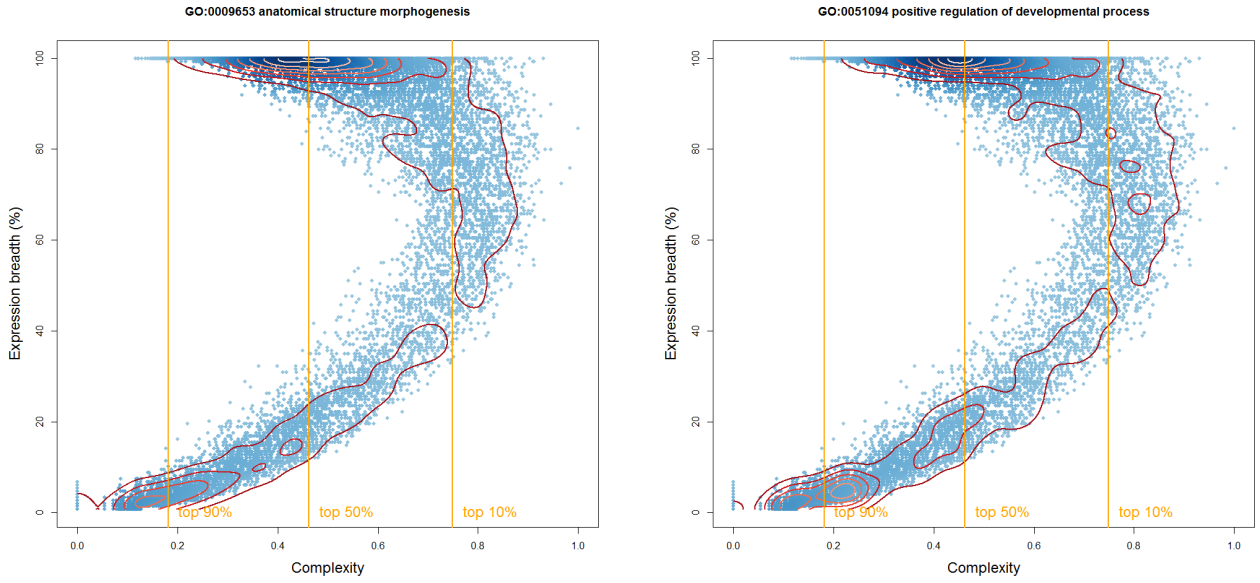


FIGURE 5.11: **Complexity vs breadth with contours for GO terms GO:0009653 anatomical structure morphogenesis (left) and GO:051094 positive regulation of developmental process (right)**. Complexity is plotted against the percentage of expressed cell types. Contours showing regions enriched in genes associated with respective GO terms are plotted in red. Yellow vertical line proportion areas of the plot according to complexity scores - the highest and lowest 10% most complex genes on the right and left sections respectively, and the centre two regions each containing 40% of the genes.

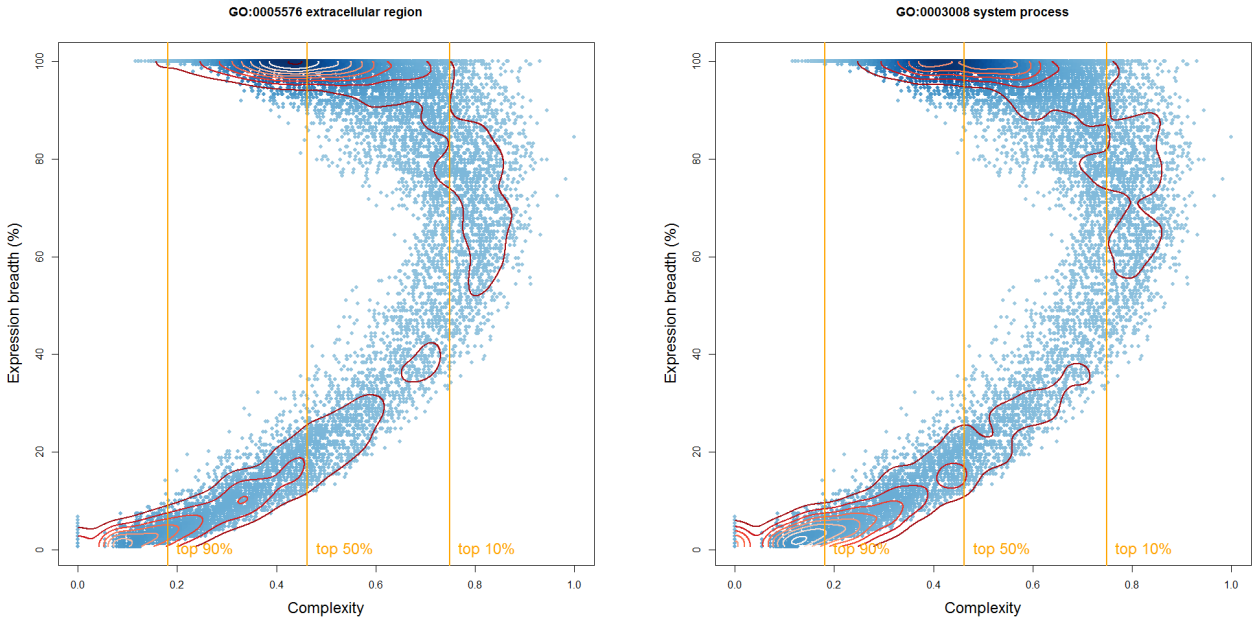


FIGURE 5.12: **Complexity vs breadth with contours for GO terms GO:0005576 extracellular region (left) and GO:0003008 system process (right)**. Complexity is plotted against the percentage of expressed cell types. Contours showing regions enriched in genes associated with respective GO terms are plotted in red. Yellow vertical line proportion areas of the plot according to complexity scores - the highest and lowest 10% most complex genes on the right and left sections respectively, and the centre two regions each containing 40% of the genes.

5.4 Relationship of scores with CpG and TATA

As introduced in Chapter 1, stretches of unmethylated CpG di-nucleotides called CpG islands are often found overlapping the promoters of ubiquitously expressed genes and some regulated genes. CpG island presence around the core promoter therefore holds a natural relationship with expression breadth, so CpG presence is likely to correlate with high entropy scores and CpG absence is likely to correlate with low entropy scores. In this section CpG is correlated with complexity scores; in particular it is of interest to see if CpG presence in non-ubiquitously expression genes are associated with higher complexity than non-ubiquitous and CpG absent genes.

It was seen that the TATA box is a consensus sequence typically found in the core promoter of polII genes; Transcription binding protein complexes bind to the TATA

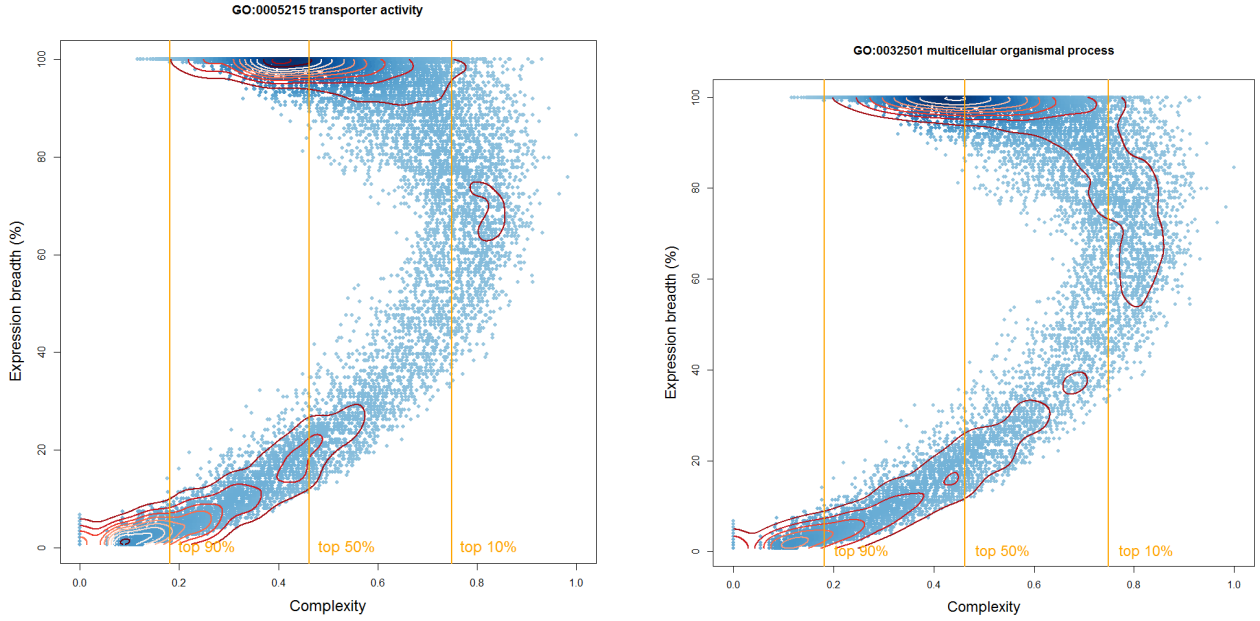


FIGURE 5.13: **Complexity vs breadth with contours for GO terms GO:0005215 transporter activity (left) and GO:0032501 multicellular organismal process (right)**. Complexity is plotted against the percentage of expressed cell types. Contours showing regions enriched in genes associated with respective GO terms are plotted in red. Yellow vertical line proportion areas of the plot according to complexity scores - the highest and lowest 10% most complex genes on the right and left sections respectively, and the centre two regions each containing 40% of the genes.

box in the initiation of transcription. CpG island presence and absence is first analysed in the context of complexity, followed by TATA and then whether the two interact or correlate with scores independently of each other.

5.4.1 Relationships with CpG island presence

Of all genes under analysis, 65.6 % were found to have a CpG island overlapping the core promoter region. Of genes with CAGE expression in every primary cell type under analysis (the defined ubiquitous genes in this study), 85.6 % reported CpG island presence. Of genes expressed in less than two thirds of the primary cell types, 67.1 % report CpG island presence and 33 % of genes expressed in less than one third.

To see how CpG presence and absence changed with increasing complexity scores, a moving average plot was generated (Figure 5.14). CpG (red line) closely follows

expression breadth (entropy levels). A plot of the relative contributions of CpG and TATA box presence to the variability of scores is also given in Figure 5.15

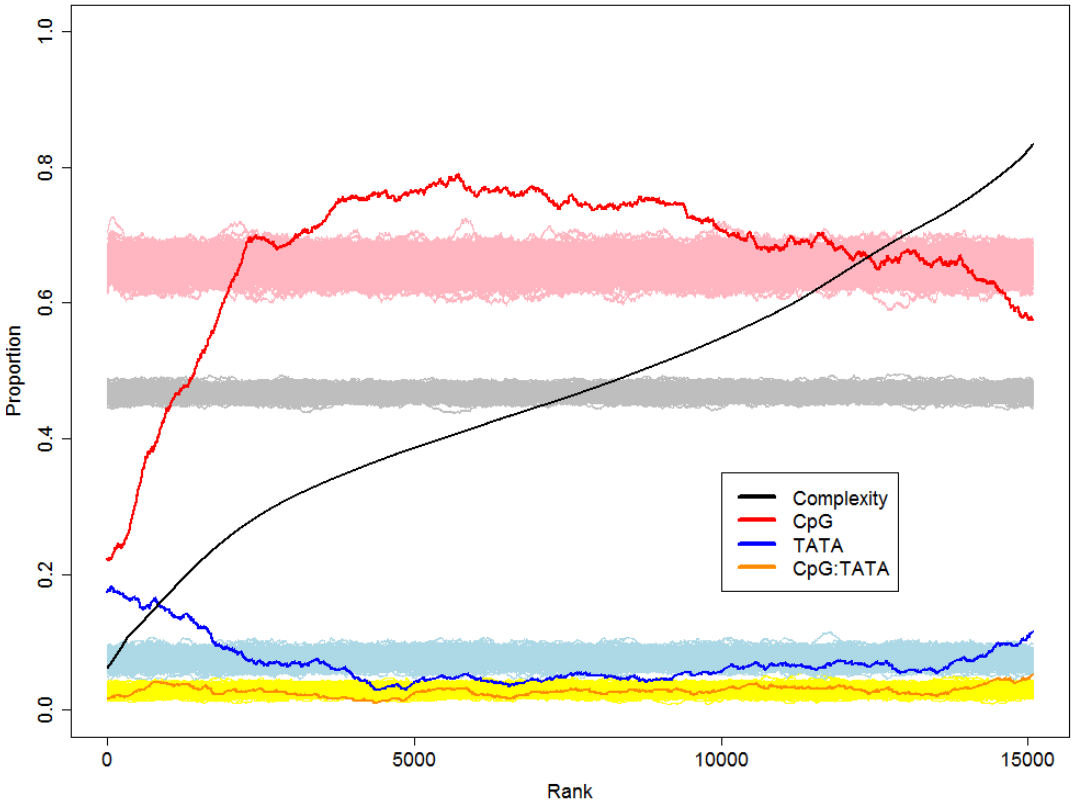


FIGURE 5.14: **Proportions of genes with CpG presence, proportions of genes with TATA presence and proportions of genes with CpG and TATA present together**. Genes are ranked in order of complexity and each data point for each variable is calculated as its averaged values from the current gene and including up to the next 1000 genes. Background distributions are calculated by permuting the ranks of the complexity scores and recalculating the proportions. Proportions of CpG present genes are plotted in red with a pink background distribution, complexity is plotted in black with a grey background distribution, TATA presence proportion is plotted in blue with a light blue background distribution and TATA:CpG interaction is plotted in orange with a yellow background distribution.

5.4.2 Explained variance in complexity scores for CpG and TATA

CpG changes non-linearly with complexity, by increasing from lower to mid- scores and gradually decreasing from mid- to high scores. However, when adjusting for entropy

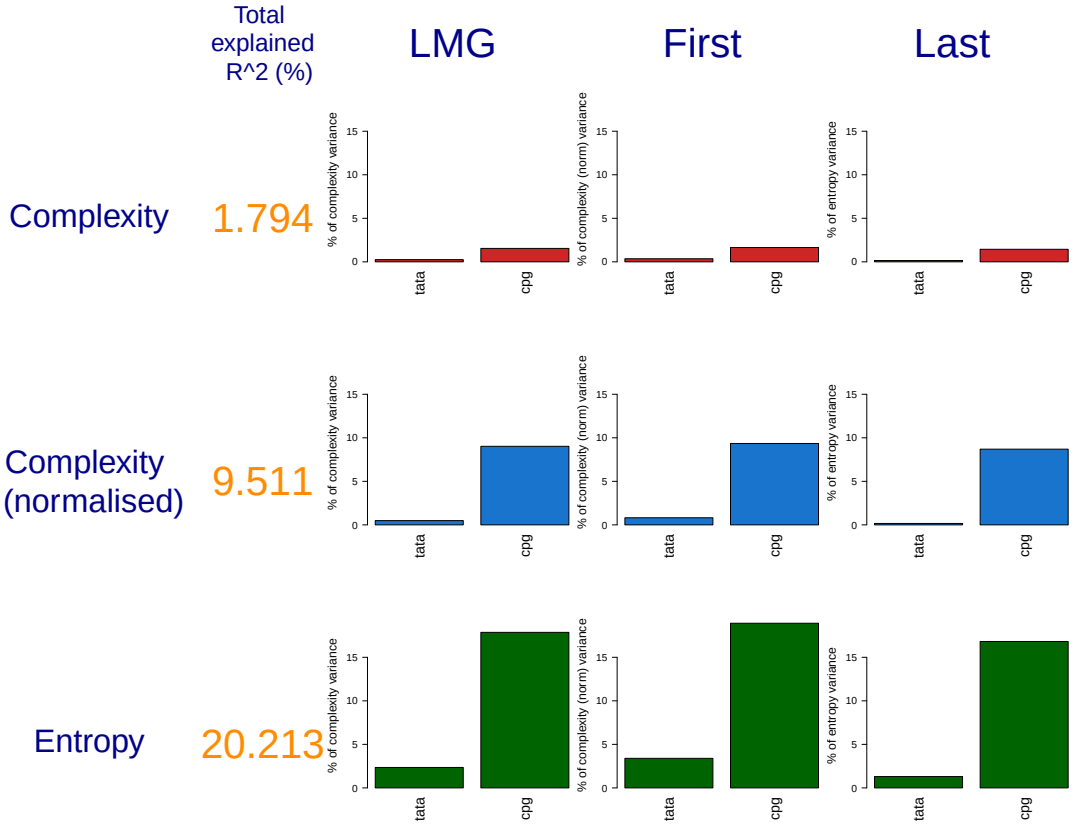


FIGURE 5.15: **Explained percentages of variance in complexity scores** (complexity - red bars, normalised complexity - blue bars and entropy - green bars) for separate regressors: **presence of TATA box in core promoter (TATA) and presence of CpG island in core promoter (CpG). Metrics used are LMG (averaged contributions, e.g. see [Chevan and Sutherland, 1991]), First (before other variables accounted for) and Last (after other variables accounted for). Total explained variance from all regressors is given in orange.**

level, CpG is linearly decreasing with increasing complexity ($p < 2e\text{-}16$), suggesting that highly complex genes are lacking in a CpG island, independent of expression breadth. This is confirmed by considering the subset of genes which are ubiquitously expressed, which shows that the odds of CpG island presence are much reduced in highly complex genes compared to genes with minimal complexity ($p<2e\text{-}16$ based on logistic regression, with log odds of -2.82 from minimum to maximum). This is illustrated in Figure 5.16. The same relationship appears to be generally true when treating all expression breadths individually (e.g. $p = 2.24e\text{-}05$, based on logistic regression for gene restricted to expression in exactly one primary cell type), although sample size is

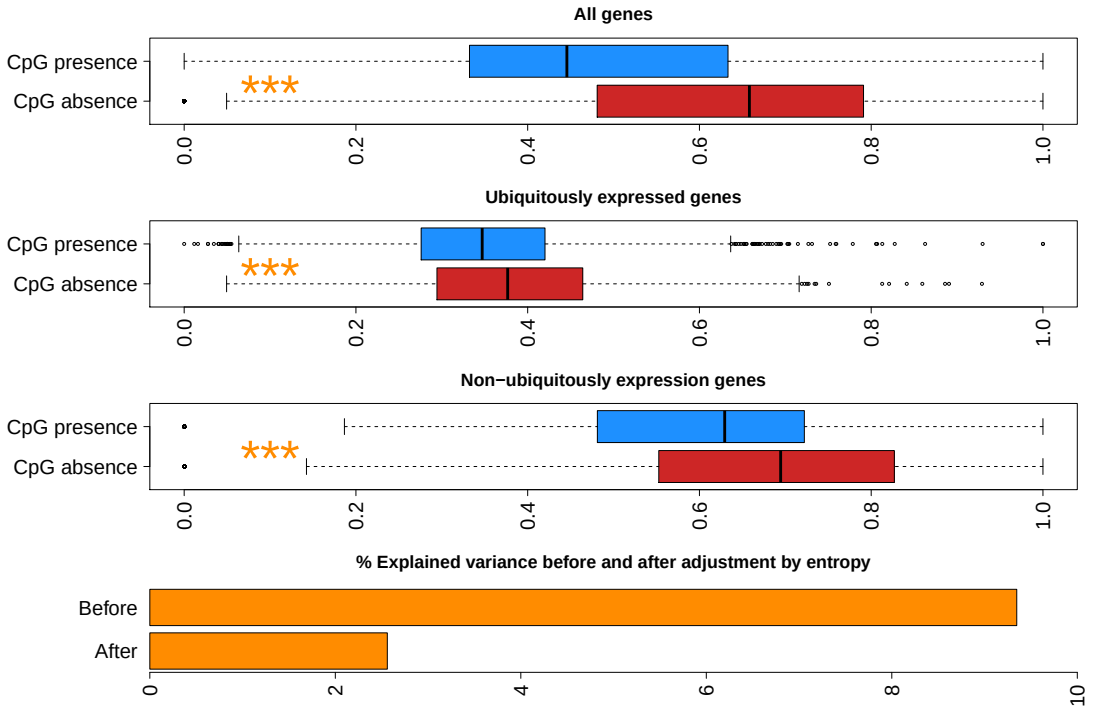


FIGURE 5.16: **Normalised complexity of genes with and without CpG presence (top), and proportions with genes broken down into ubiquitous and non-ubiquitous (middle).** Blue bars indicate presence of a CpG island overlapping the core promoter of the gene and red bars indicate the absence of a CpG island overlapping the core promoter. **Explained variance in normalised complexity scores before and after expression breadth adjustment (bottom).** The `relaimpo` package was used with option "last" in order to obtain explained variance for *CpG* after entropy had been accounted for, using `lm` in *R*.

greatly reduced for many possible breadths of expression. The explained variance after adjustment of entropy is given in (Figure 5.16) and suggests that around 2.5% of the variance post adjustment of entropy is explained by a CpG effect.

5.4.3 Relationships with TATA box presence

Of the expressed genes under analysis, 7.5% were found to have a TATA box overlapping the core promoter region. Only 3.9% of ubiquitously expressed genes had TATA box presence, 12.7% of those genes expressed in under two thirds of primary cell types and 14.2% of genes expressed in less than one third. Figure 5.14 suggests that TATA box

presence has an inverse relationship with entropy scores, as opposed to entropy which approximately follows entropy scores.

Predicted TATA presence increasings with complexity after the adjustment of entropy levels (p<2e-16, logistic regression, change in log odds of 2.07 from minimum to maximum) and within the subset of ubiquitously expressed genes, although the effect is weaker than CpG island presence (p = 9.7e-10, logistic regression) and difficult to detect with smaller sample sizes.

5.4.4 Interactions between CpG and TATA presence

Of the expressed genes used in this analysis, 29.8% contained neither a CpG island or TATA box overlapping the core promoter, 67.4% contained exactly one but not the other, and 2.9% contained both.

When adjusting for entropy and looking at the effects of CpG and TATA together, CpG presence decreases complexity scores by 0.070, independently of TATA presence (p < 2e-16, linear regression), TATA presence increases complexity scores by 0.015, independently of CpG presence (p = 0.0208, linear regression). When TATA and CpG are both present, complexity decreases by 0.019, due a significant interaction effect of 0.036 (p=0.0004, linear regression < 2e-16, linear regression). Therefore, when CpG and TATA are present together, the predicted complexity of a gene is more complex than what would be expected based on the sum of their individual effects.

In conclusion, most of the effect of CpG and TATA on complexity scores are as a result of expression breadth (reflected in entropy levels) and as a result neither associate linearly with changing complexity scores. However, weaker but highly significant relationships do exist between CpG presence, TATA presence and complexity, independent of expression breadth. This is an interesting result in that it answers questions about the kinds of genes which contain these elements above and beyond what is explained using entropy scores.

5.4.5 Methods

CpG presence or absence was calculated using the **cpgIslandExt** downloaded from UCSC [Gardiner-Garden and Frommer, 1987]. The *R* package `genomicRanges` [Lawrence et al., 2013] was used to determine the presence of an overlap between the promoter region of genes with a CpG island in the table, based on RefSeq gene annotations [Dreszer et al., 2012].

TATA presence or absence was estimated by using the JASPAR matrix MA0108.2 for TATA applied to the whole genome [Bryne et al., 2008], in the same method used by the FANTOM5 consortium [Forrest et al., 2014].

Figure 5.14 was calculated from ranked groups of 1000 genes according to lowest to highest complexity scores. Proportions of variables across each set of 1000 genes for CpG presence, TATA presence and CpG:TATA interaction were also plotted. Background distributions were calculated based on permutation ranks where the ranking of the genes was randomly permuted and variables proportions were calculated in the order based on the new ranking. Plots are based on 500 such permutations for each variable. Variable proportions based on the true ranking were plotted over the top in a darker colour. Regions where the true proportions cross outside of the background distribution are potentially significantly associated with complexity scores.

Changes in CpG and TATA presence/absence across complexity was estimated using quantile regression, confirming the observed relationship of decreasing TATA and increasing CpG at lower quantiles, followed by the opposite relationship at the upper quantiles.

To control for entropy in order to test for the effect of CpG and TATA independent of their relationship with expression breadth, F-tests were applied to compare models containing entropy as a covariate with models containing entropy and CpG/TATA. Adjusting for expression breadth in this way is useful for teasing out properties captured by complexity that are not captured by entropy scores (thus providing an argument for choosing complexity over entropy). After adjusting for entropy, the effect of CpG presence/absence become linear (decreasing over all quantiles of complexity with similar

effect sizes). This is confirmed by adjusting for the effect of the entropy when correlating complexity scores with CpG, with CpG significantly decreasing with entropy adjusted complexity.

5.5 Genomic size constraints, isoforms and alternative promoters

5.5.1 Distances between genes and gene length is weakly correlated with complexity scores

Next the physical attributes of genes was analysed;

Firstly the length of the gene and the length of the first intron were calculated (Figures 5.17, and correlated with complexity and entropy scores 5.18 and 5.19).

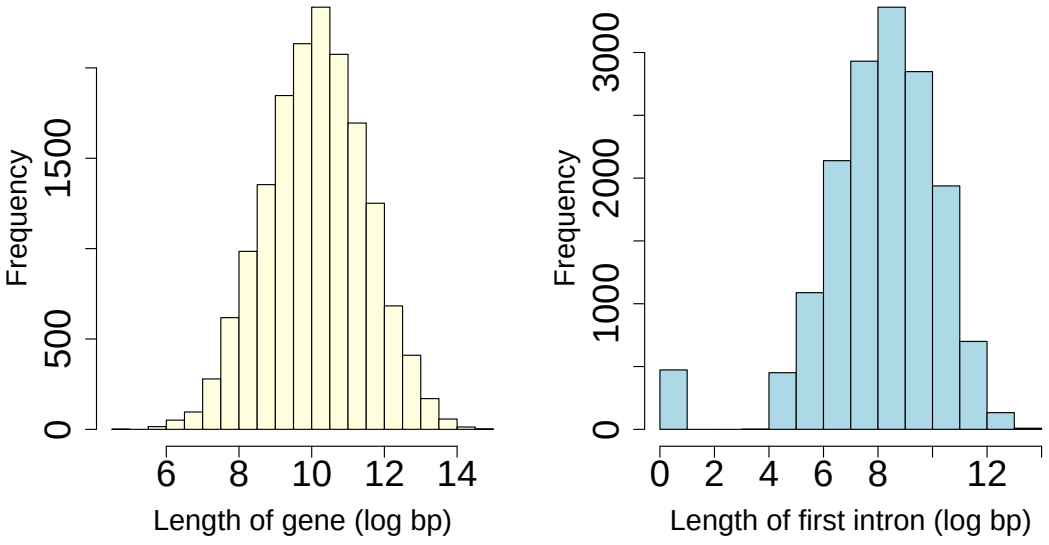


FIGURE 5.17: **Histograms of length of gene (left)**, including exons and introns, and the **length of the first intron of the gene (right)** (0 for single exonic genes), with values given in the log of the number of base pairs. Both distributions treated as approximately normal.

Increasing complexity scores were very weakly but significantly associated with a greater length of the gene and a greater length of the first intron (explaining 1.55% and 1.32% of the variation respectively).

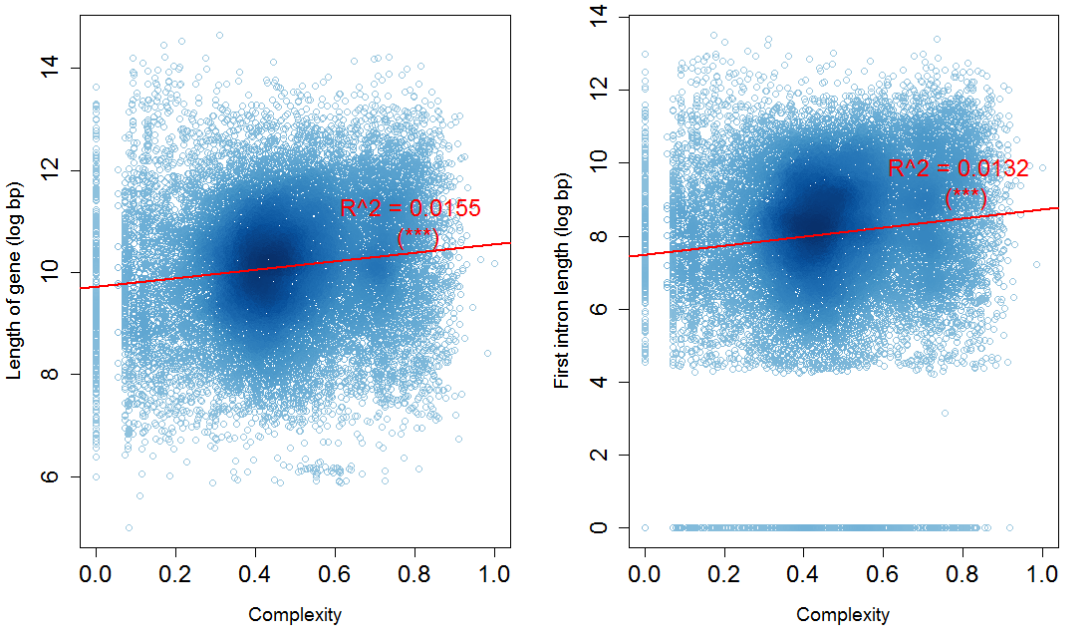


FIGURE 5.18: Scatter plots of **gene length vs complexity**, and the **length of the first intron vs complexity**. Length measurements are given in the log of the number of base pairs. Red lines indicate best fit lines from linear model, with R^2 values calculated from the model. Significance is given as (***), which implies that the slope of the line is highly significant.

A similar but even weaker relationship with observed with entropy scores (0.7% and 0.5% of the respective variation for gene length and first intron length), suggesting that longer genes with longer first introns are more likely to be broadly expressed.

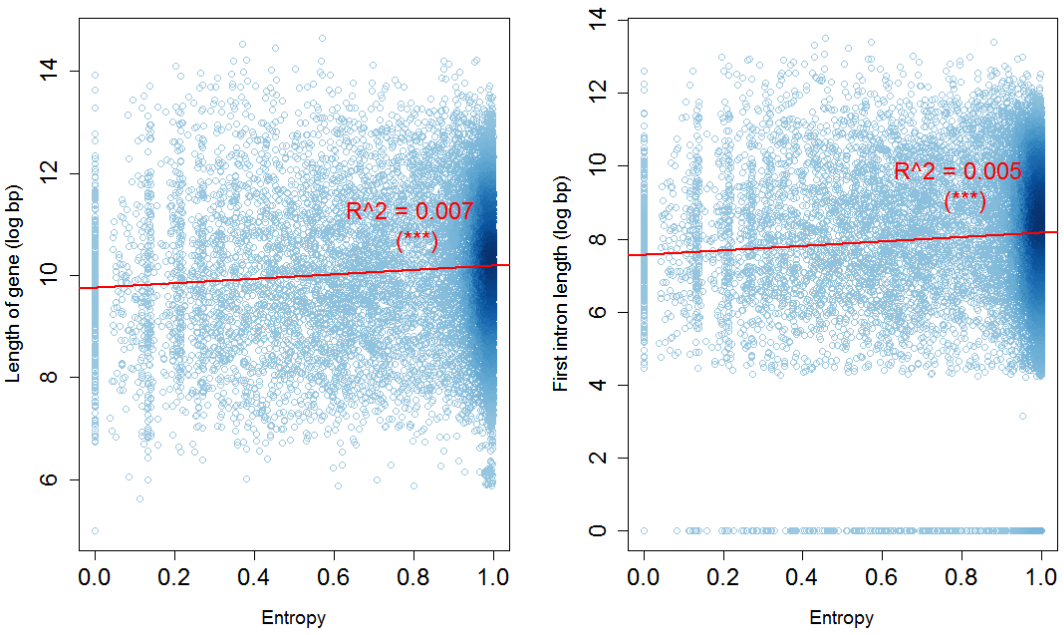


FIGURE 5.19: Scatter plots of **gene length vs entropy**, and the **length of the first intron vs entropy**. Length measurements are given in the log of the number of base pairs. Red lines indicate best fit lines from linear model, with R^2 values calculated from the model. Significance is given as (***), which implies that the slope of the line is highly significant.

The distance to the nearest upstream gene and the distance to nearest downstream gene were calculated (Figure 5.20) and correlated with complexity and entropy scores (Figure 5.21 and 5.22).

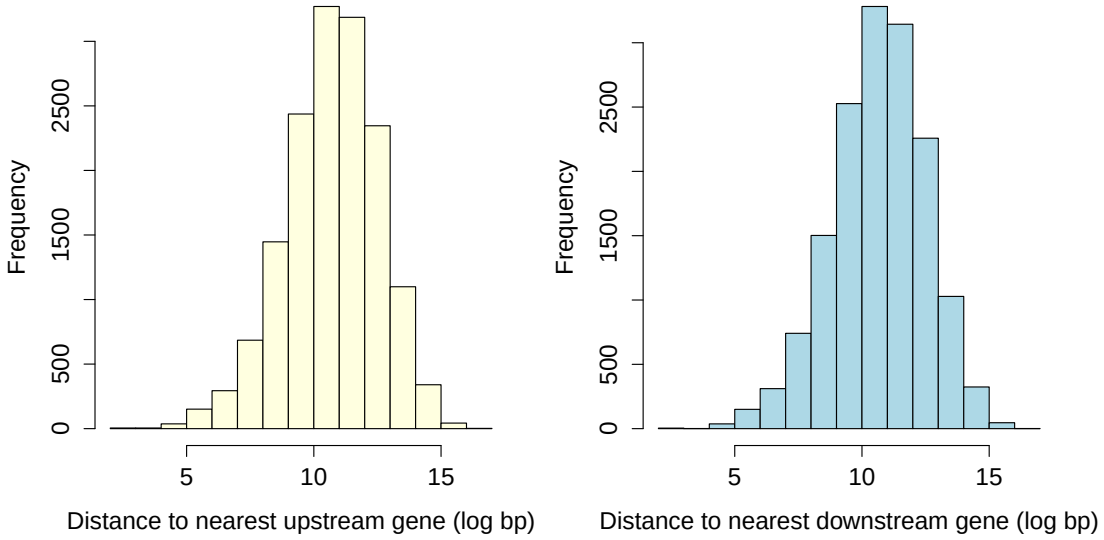


FIGURE 5.20: **Histograms of distance to nearest upstream gene (left)**, and the **distance to nearest downstream gene (right)**, with values given in the log of the number of base pairs. Both distributions treated as approximately normal.

Similar to gene length, the distance to the nearest upstream gene and the distance to the nearest downstream gene were very weakly correlated with complexity (0.4% and 0.2% of variation respectively). Whilst highly significant, this is not a particularly strong finding because the effect size is extremely small.

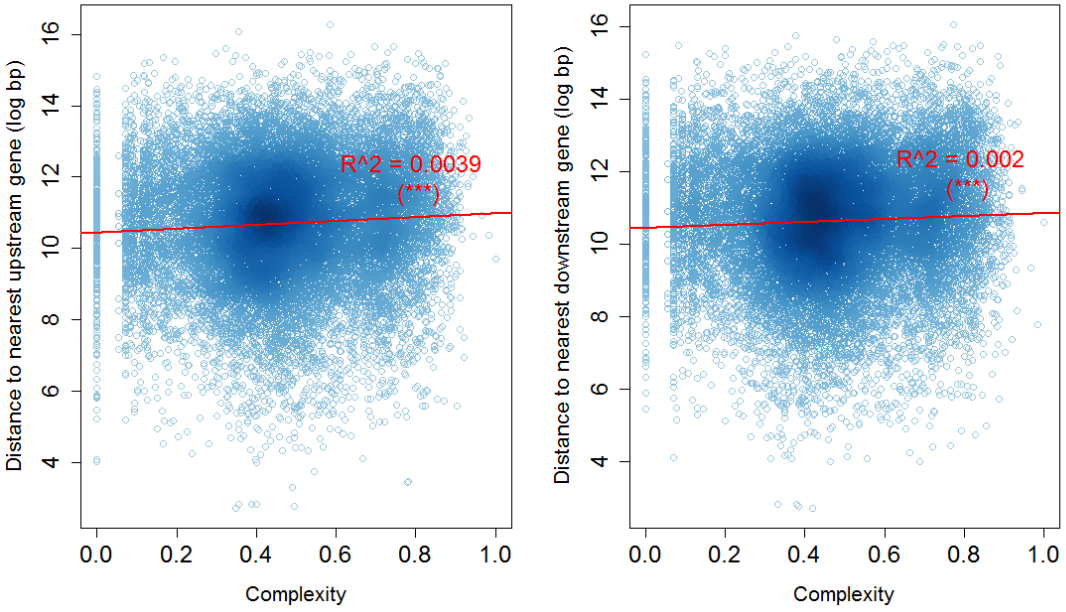


FIGURE 5.21: Scatter plots of the **distance to nearest upstream gene vs complexity** and **distance to nearest downstream genes vs complexity**. Length measurements are given in the log of the number of base pairs. Red lines indicate best fit lines from linear model, with R^2 values calculated from the model. Significance is given as (***), which implies that the slope of the line is highly significant.

Entropy scores were associated with slightly shorter distances between genes (0.005% and 0.4% of up and downstream length), suggesting that cell restricted genes have a preference towards greater space around them. However this effect size is again extremely small.

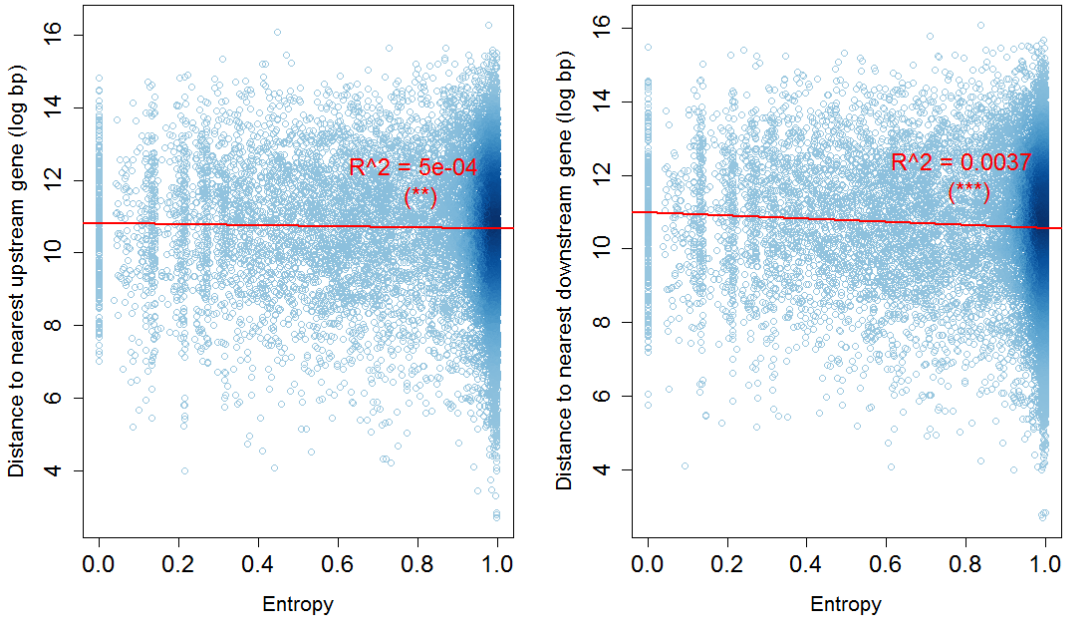


FIGURE 5.22: Scatter plots of the **distance to nearest upstream gene vs entropy** and **distance to nearest downstream genes vs entropy**. Length measurements are given in the log of the number of base pairs. Red lines indicate best fit lines from linear model, with R^2 values calculated from the model. Significance is given as (***), which implies that the slope of the line is highly significant.

In conclusion, it is observed that whilst complex genes appear slightly longer, contain more intronic sequence and space between genes, the effect sizes are very weak. Thus, whilst these factors could be at play in regulating highly complex expression patterns, i.e. through providing more surrounding space from which cis-regulatory elements may act on the gene, they are far from explaining a large proportion of the complexity observed in their expression profiles.

5.5.2 Increased complexity correlates with number of promoters annotated to the gene, exon count and isoforms per gene

The distribution of the number of exons, isoforms and annotated promoters across all genes are given in Figure 5.23.

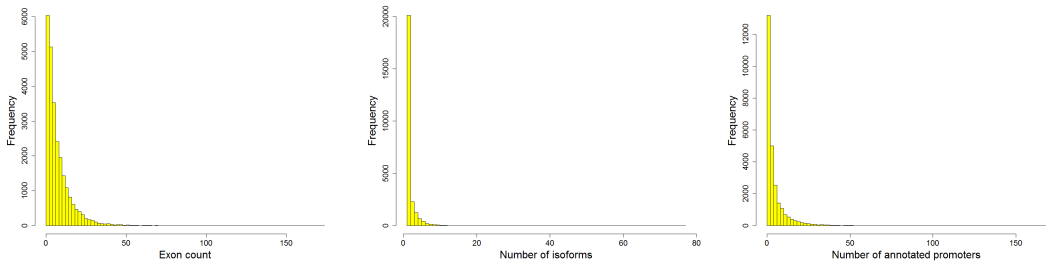


FIGURE 5.23: **Distribution of number of exons per gene (left), number of isoforms per gene (middle), number of annotated promoters per gene (right)**

Complexity scores were group by promoter count and a boxplot drawn for each category (Figure 5.24), clearly showing the significantly increasing relationship between the number of annotated promoters and gene based complexity scores ($p < 1e\text{-}16$).

Figure **??** shows the total variance in complexity scores explained by the total combination of distances, isoforms, exon and promoter numbers. Notice that the total explained variance by all factors is small, only around 1.9% for complexity and 0.9% for normalised complexity scores. This suggests that these factors are not highly important in terms of explaining what makes a gene complex. Notice however that explained percentage for entropy is also low; tissue specificity neither appears to be highly associated with its local physical constraints.

5.5.3 Method

Gene length, distance to nearest upstream and downstream genes, number of exons and number of isoforms were estimated from the refSeq genome table downloaded from UCSC [Dreszer et al., 2012] using the *R* package `GenomicRanges` [Lawrence et al., 2013].

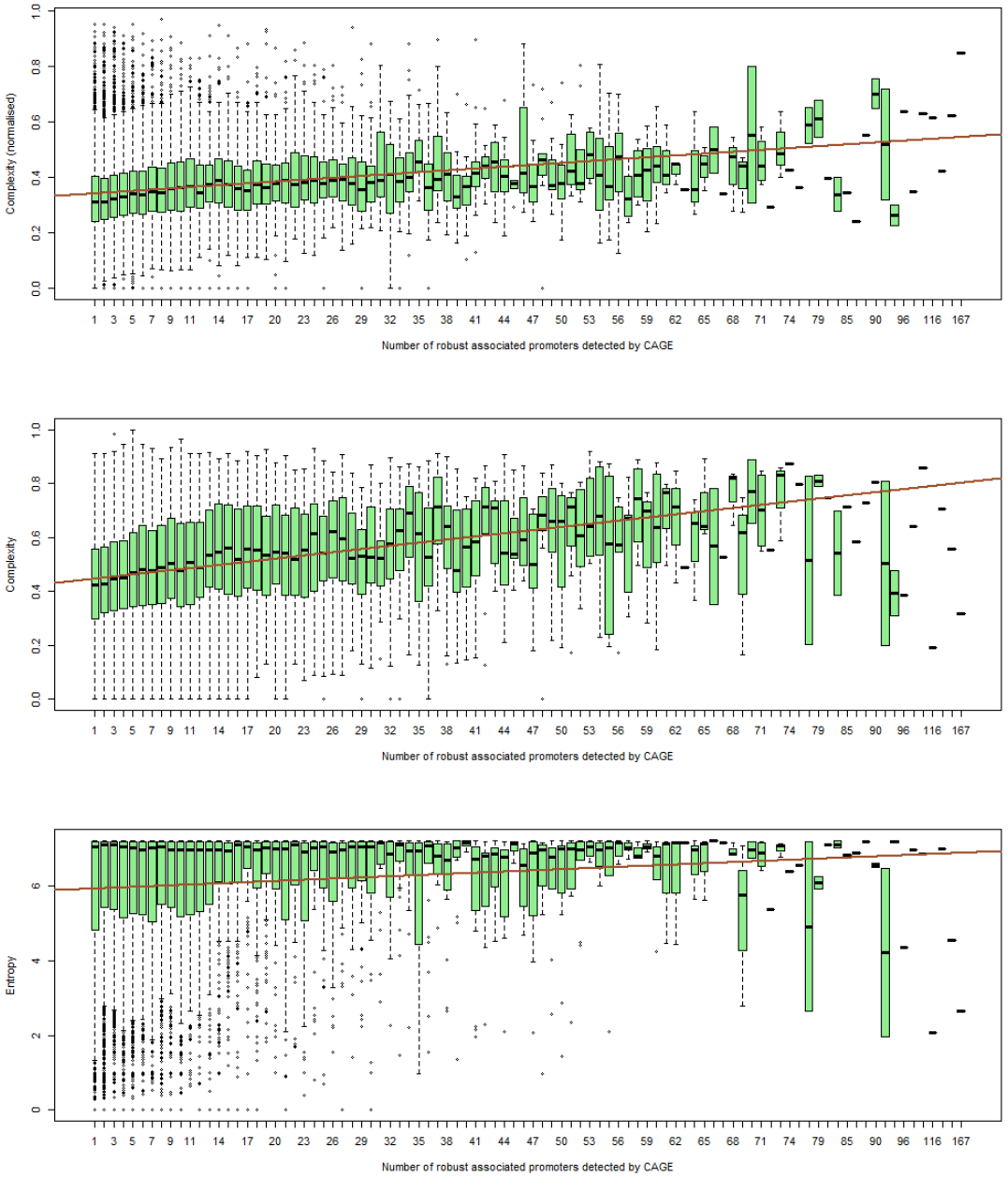


FIGURE 5.24: Boxplots showing the distribution of **complexity scores for each possible number of robust associated promoters per gene detected in FANTOM5 CAGE**. Brown lines represent best fit lines from applying linear model to each of three scores - top: normalized complexity, middle: complexity, bottom: entropy score. Top and middle slopes are highly significant (p<1e-16), entropy slope is weakly significant, according to modelling using the `lm` function for the complexity and normalised complexity, and the `rq` function from the `quantreg` package for the entropy scores

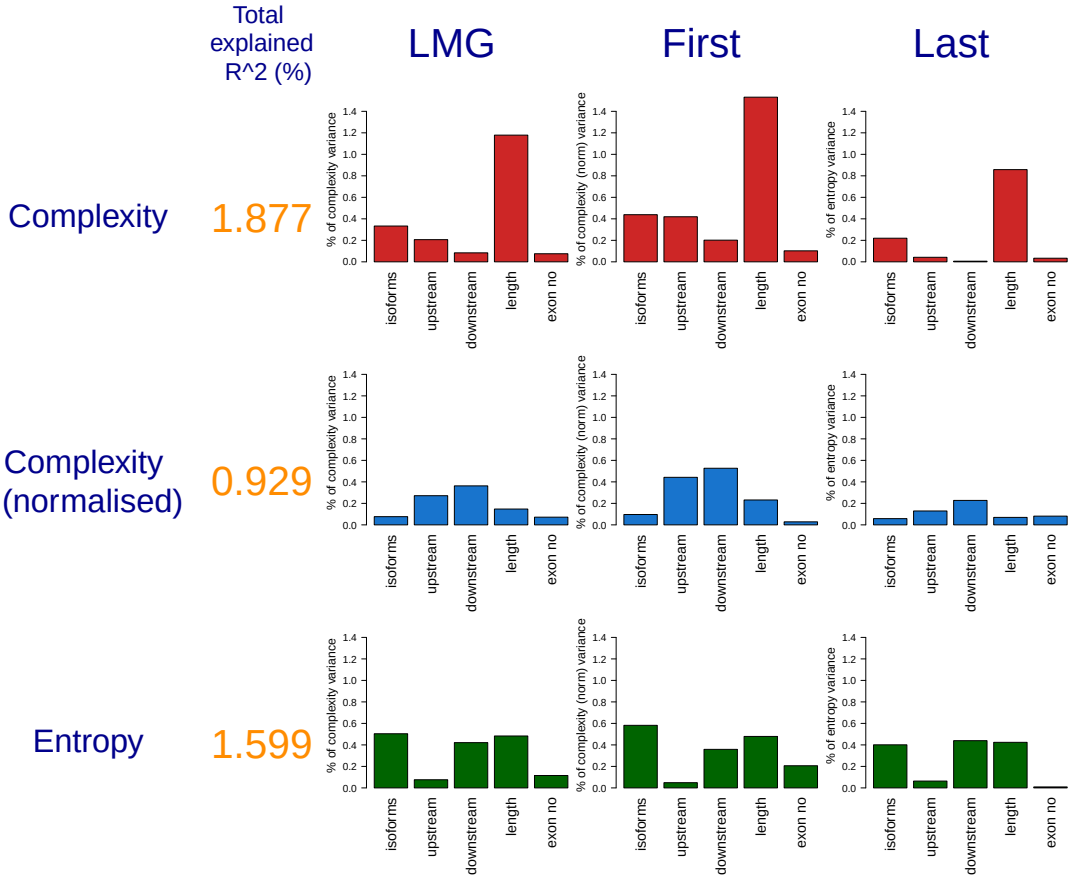


FIGURE 5.25: **Explained percentages of variance in complexity scores** (complexity - red bars, normalised complexity - blue bars and entropy - green bars) for separate regressors: **the number of isoforms (isoforms), distance to nearest upstream gene (upstream), distance to nearest downstream gene (downstream), gene length (length) and number of exons associated with the gene (exon no)** . Metrics used are LMG (averaged contributions, e.g. see [Chevan and Sutherland, 1991]), First (before other variables accounted for) and Last (after other variables accounted for). Total explained variance from all regressors is given in orange.

The number of promoters were estimated by counting the number of robust clusters annotated to a given gene within the FANTOM5 dataset [Forrest et al., 2014]. Clusters were filtered so that expression was present in at least one primary cell type over which complexity was calculated, with a median of at least 1 over the replicates.

A linear model as used to estimate the slope of increase in complexity with increasing promoters. Quantile regression was used to estimate slopes for entropy scores (due to violation of normality assumptions) [Koenker, 2013]. Due to small numbers of genes

in categories with large numbers of promoters, these categories were grouped together into pseudo categories with larger numbers of genes.

5.6 Increased complexity correlates with cis-regulation

As introduced in Chapter 1, cis-regulatory elements (CRE) act to control gene expression, through the recruitment of transcription factors (TF) to binding sites on the sequence within the vicinity of a gene. CREs encompass a wide variety of sequence elements, including promoters and short-range enhancers, and small changes in these elements can have large and unpredictable effects on the resulting expression. Genes under the control of a broad landscape of cis-regulation are on the whole potentially subject to more fine tuned pattern of expression levels according to cell type and biological or environmental needs. Turning this idea around, genes which exhibit complex patterns of expression across the spectrum of cell types are hypothesised to be targeted by more CREs than those which are simple in their expression (which may require only basal levels of transcription achievable through the core promoter alone). This section tests this by correlating measures of DNAse I hypersensitivity and conservation and predicted enhancers in and around the vicinity of genes with measures of complexity and entropy.

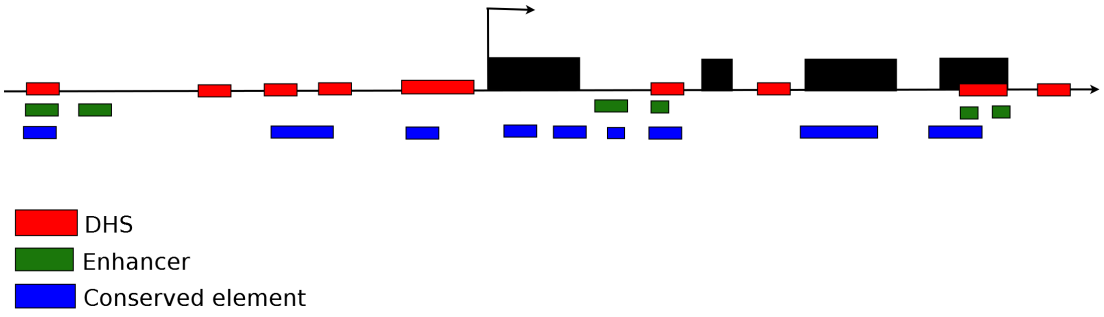


FIGURE 5.26: **Schematic of potential cis-regulatory regions**, based on DNase I hypersensitive sites (DHS) (red), predicted enhancers (green) and conserved GERP++ elements (blue). These regions represent potential transcription factor binding sites with may mediate with the core promoter to regulate the transcription of the gene.

The regions cis-regulatory elements were measured over as as follows

- **Upstream** - 10,000 bp upstream of the upstream core promoter region. Thus, the effect of short-range cis-regulatory elements.
- **Upstream core promoter region** - 250 bp upstream of the transcription start site

- **First intron**
- **Rest of gene** - all parts of gene not including the first intron and including exons
- **Downstream core promoter region** - 250 bp downstream of the end of the last exon

DNAse I hypersensitivity sites indicate areas of open chromatin whereby the DNA is accessible to DNAse I cleavage enzymes. Mapped sites in the vicinity of the body of a gene, as well as within its exons or introns, may be representative of cis-regulatory modules affecting the regulation and hence expression of the gene. Such sites have been mapped extensively [Thurman et al., 2012] and correlated with other markers of regulation e.g. H3K27me3 marks. The number of DNAse I sites acting in a given gene is often used as a proxy for the cis-regulatory effect on the gene, making it a perfect variable to correlate with complexity scores [Thurman et al., 2012]. In terms of transcriptional output, gene expression levels have been liked to the number of hypersensitive sites [Wang et al., 2012].

Conserved sequences in the genome are those which are highly similar or identical between species. Conserved sequence is often observed in genes which appear in multiple species, whilst non-coding sequence is often associated with regulatory regions. Gene expression levels and breadth have been found to be positively correlated with conservation, particularly conservation observed in the introns of the gene [Gorlova et al., 2014, Park and Choi, 2010]. Whilst protein coding genes are generally highly conserved between species, conservation is not necessarily correlated with DNase I hypersensitive sites. However, there exists a significant correlation between the number of hypersensitive sites and the number of GERP conserved elements observed over the same genomic region. The numbers within the first intron are correlated with a Pearson's correlation score of 0.738 ($p < 2.2e-16$). This suggests that many of the observed conserved sites are hypersensitivity sites and vice versa.

Enhancers are short sequences which in general typically act on a gene in cis- , can influence a gene from up or downstream of its promoter, can appear within other genes and act in an orientation independent manner [Shlyueva et al., 2014]. In FANTOM5

CAGE enhancers atlas have been mapped based on the presence of bidirectional transcription [Andersson et al., 2014b]. The enhancer atlas includes the same set of cell types used to calculate complexity scores and therefore removes the issue of a complex gene whereby enhancers are missed due to complexity being calculated over cell types not present in the cell types used to capture the enhancers.

Further note that many regulatory elements affecting gene expression do so from a distance greater than 10kb upstream or downstream of the core promoter region. Furthermore, they do not necessarily act on the gene they are nearest to. In the FANTOM5 project, the presence of correlations between expression vectors across cell types between putative enhancers and promoters has been applied to call an interacting pair [Andersson et al., 2014b]. However, due to the highly specific nature of enhancers, the sensitivity of such an approach is limited. Thus, in the absence of further information about inferred interactions between enhancers and promoters, a cut-off of 10kb is applied.

Figure 5.27 and Figure 5.28 show the distributions of hypersensitivity sites against complexity scores in more detail. The number of overlapping GERP conserved sites for the upstream promoter region against complexity scores is given in Figure 5.29 and for the first intron is given in Figure 5.30.

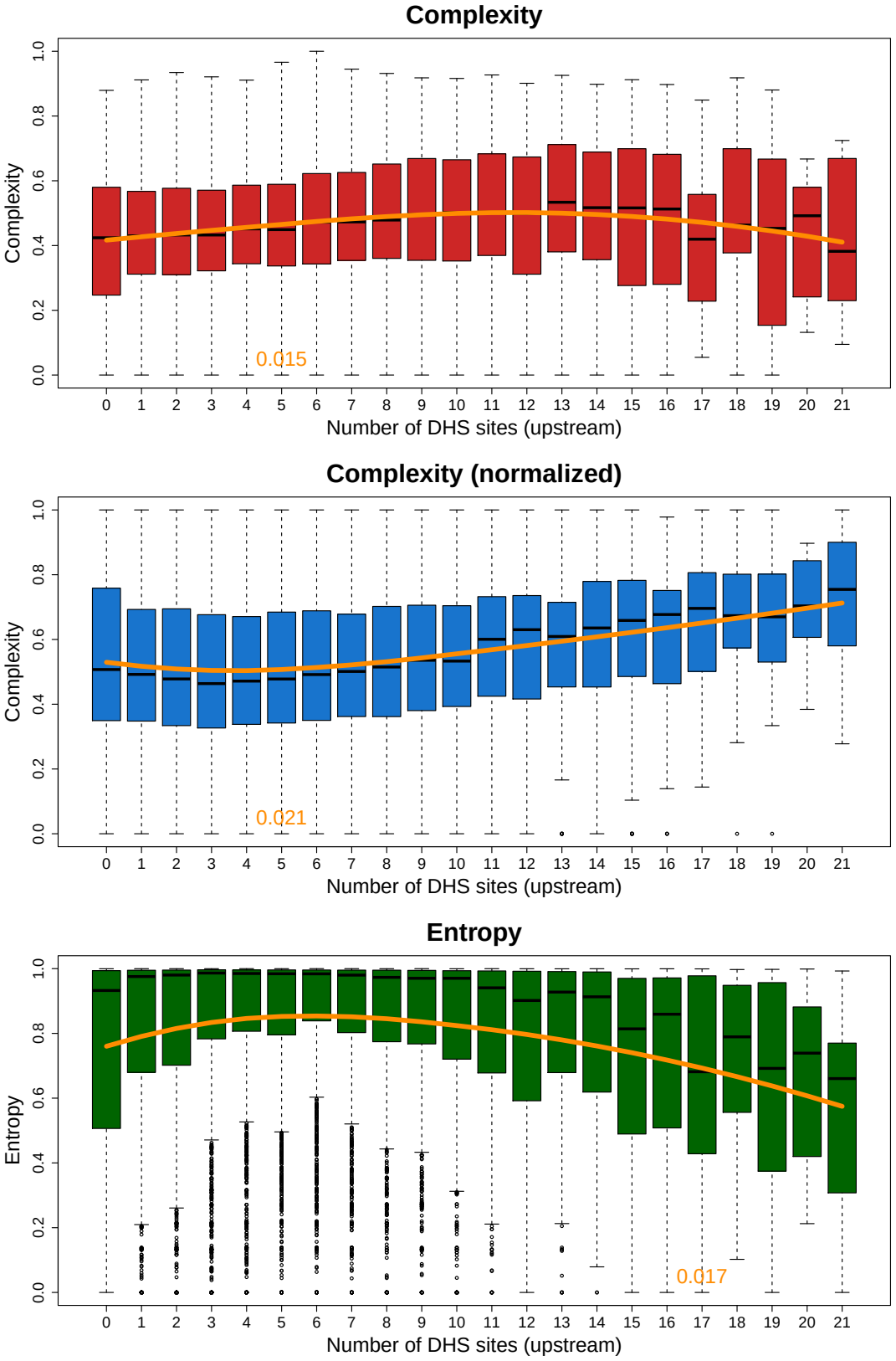


FIGURE 5.27: Boxplots showing the distribution of **scores (x-axis) vs number of DNase I hypersensitive sites within the space of 10k bp upstream of the gene (y-axis)**, excluding the core promoter region and overlap with other genes. Orange lines represent best fit lines from applying `loess()` function to each of three scores - top: complexity, middle: normalised complexity, bottom: entropy score. Orange numbers represent model r^2 values.

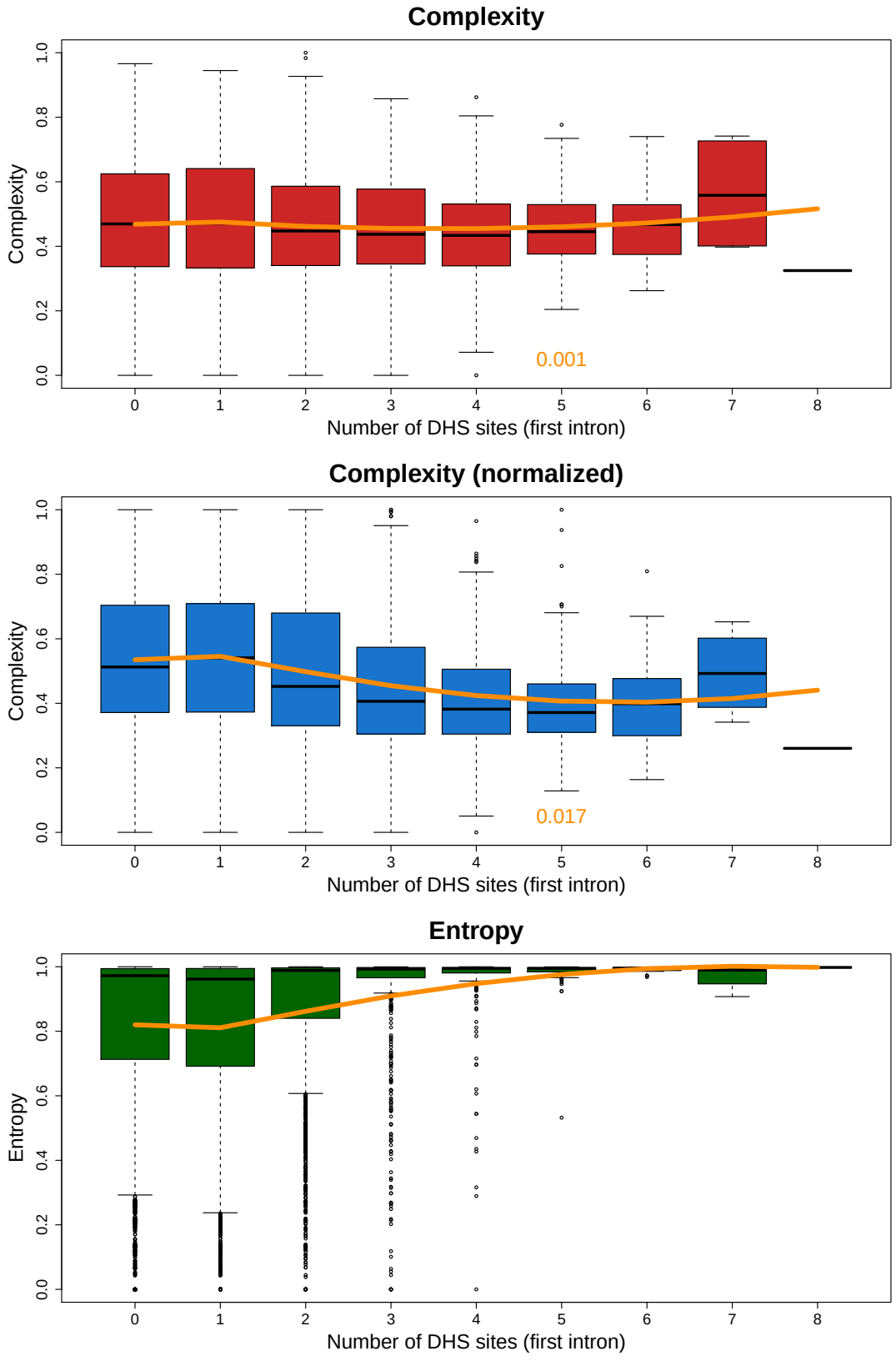


FIGURE 5.28: Boxplots showing the distribution of complexity measures for each possible number of hypersensitivity sites observed within the first intron of the gene. Genes with a single exon are allocated a value of 0. Orange lines represent best fit lines from applying `loess()` function to each of three scores - top: complexity, middle: normalised complexity, bottom: entropy score. Orange numbers represent model r^2 values.

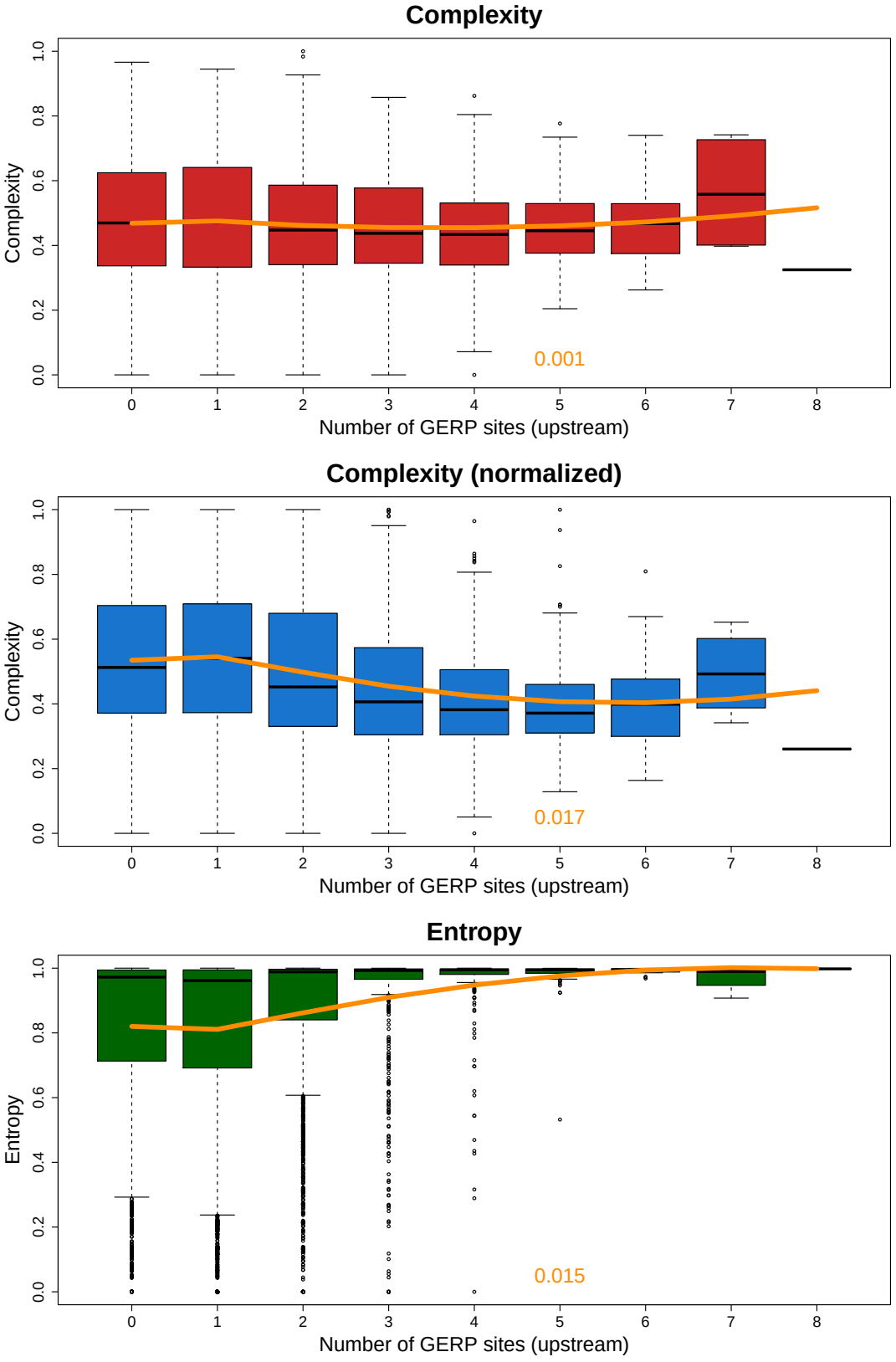


FIGURE 5.29: Boxplots showing the distribution of **scores (x-axis) vs number of GERP conserved elements (y-axis) overlapping the upstream promoter region of the gene**. Orange lines represent best fit lines from applying `loess()` function to each of three scores - top: complexity, middle: normalised complexity, bottom: entropy score. Orange numbers represent model R^2 values.

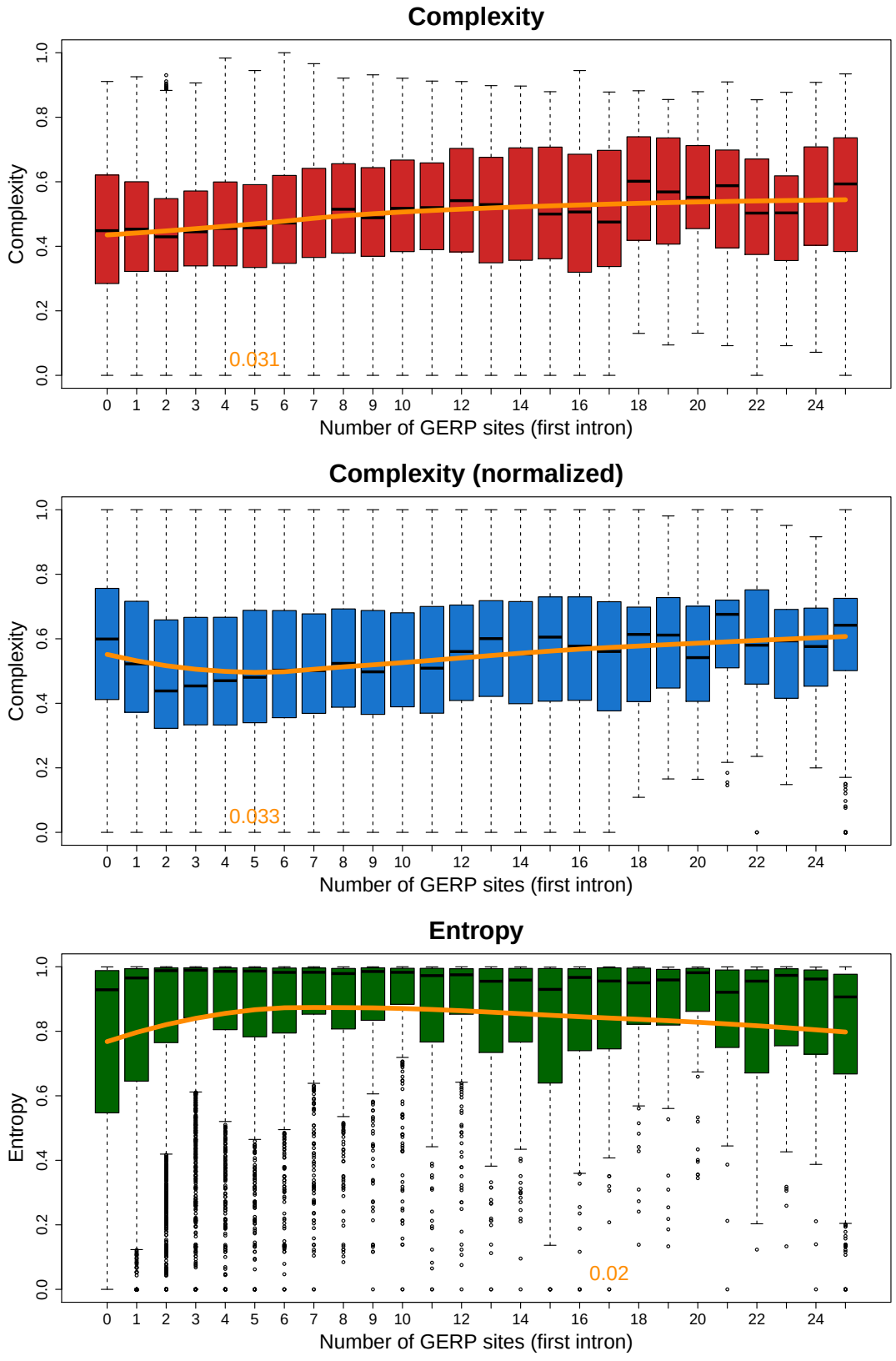


FIGURE 5.30: Boxplots showing the distribution of **scores (x-axis) vs number of GERP conserved elements (y-axis) overlapping the first intron of the gene**. Orange straight lines represent best fit lines from applying linear model (`lm` function) to each of three scores - top: complexity, middle: normalised complexity, bottom: entropy score. Orange numbers represent model R^2 values.

5.6.1 Variance explained by conservation, DNAse I hypersensitivity and predicted enhancers

Figures 5.31,5.32 and 5.33 show the proportion of the explained variance in complexity scores, normalised complexity scores and entropy scores respectively, for each of the three datasets based on counting elements across each of the five regions described. Each variable is given as a singular effect in order to understand their contribution independently of interactions and correlations with other variables.

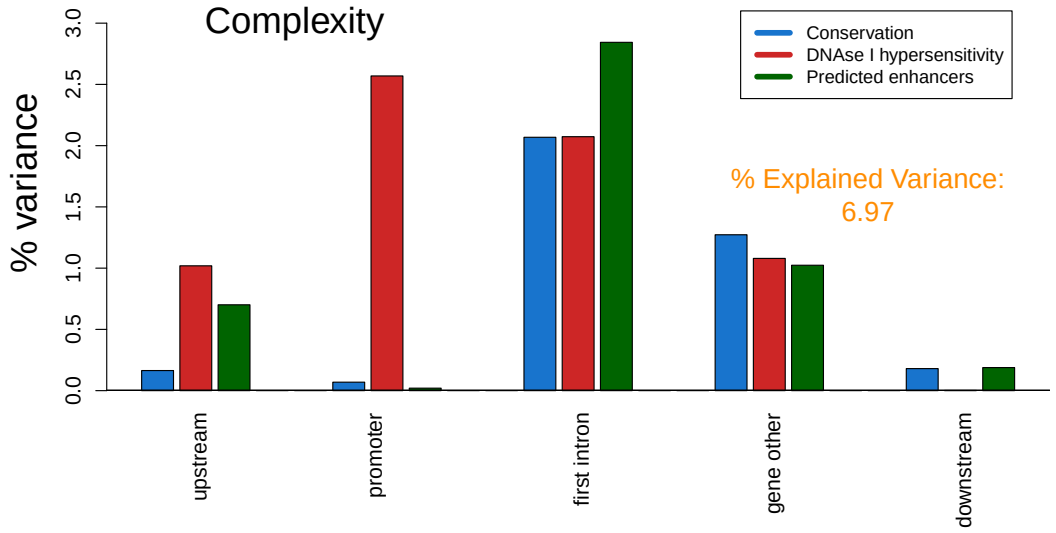


FIGURE 5.31: **Explained percentages of variance in complexity** across five regions of the gene: **upstream, promoter, first intron, gene other and downstream**. Metric used is First (before other variables accounted for), due to correlations between data sources. Total explained variance from all regressors is given in orange.

Counts based on the three datasets appear consistent in the proportion of complexity scores explained based on the first intron and the rest of the gene. In particular, enhancers in the first intron had the strongest effect of all variables. In terms of upstream of the TSS, DHSs in the promoter had a strong influence on complexity, and upstream DHSs and enhancers were also significant.

Including all main effects together in a single model, all of the 15 variables explained 6.97% of the variance in complexity scores. The model including all pairwise interactions explained 8.2% of the variation in complexity scores. The most significant interactions were conservation upstream and enhancers upstream (positive effect, p-value = 0.00014), conservation upstream and conservation in the upstream promoter region (positive effect, p-value = 0.0015), and conservation in the upstream with conservation in the first intron (negative effect, p-value = 0.065). Furthermore, when restricting to ubiquitous genes only, the model including pairwise interactions explained 8.6% of the variation in complexity scores.

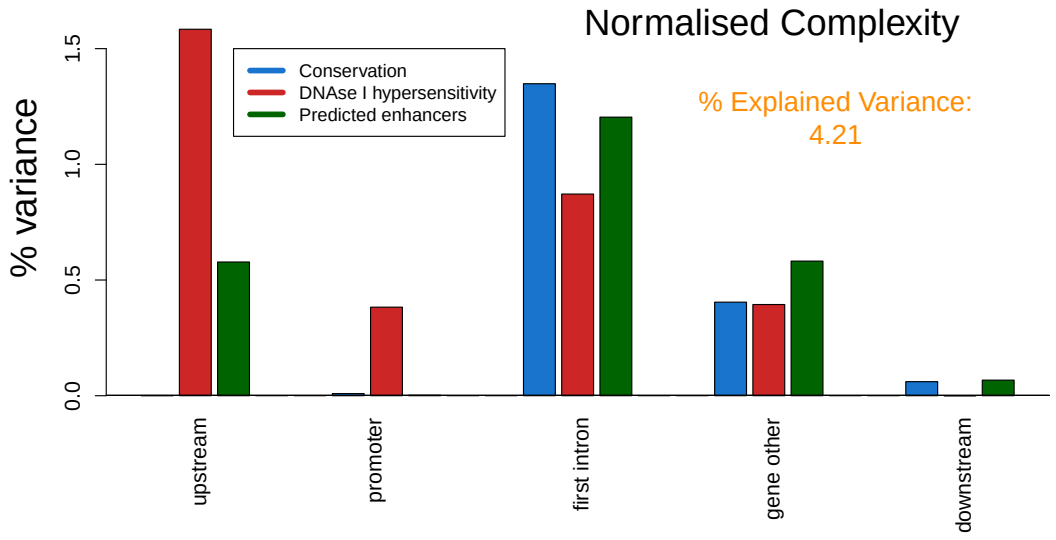


FIGURE 5.32: **Explained percentages of variance in normalised complexity** across five regions of the gene: **upstream, promoter, first intron, gene other and downstream**. Metric used is First (before other variables accounted for), due to correlations between data sources. Total explained variance from all regressors is given in orange.

All variables together in the same model explain 4.21% of the variation in normalised complexity scores and including interactions this increases to 6.6%. When restricting to ubiquitous genes, 8.2% of the variance is explained by the interactions model. Complexity and normalised complexity show similar results across the first intron of the gene, but normalised complexity seems to be influenced to a greater extent by upstream DHSs compared to the presence of DHSs in the core promoter region.

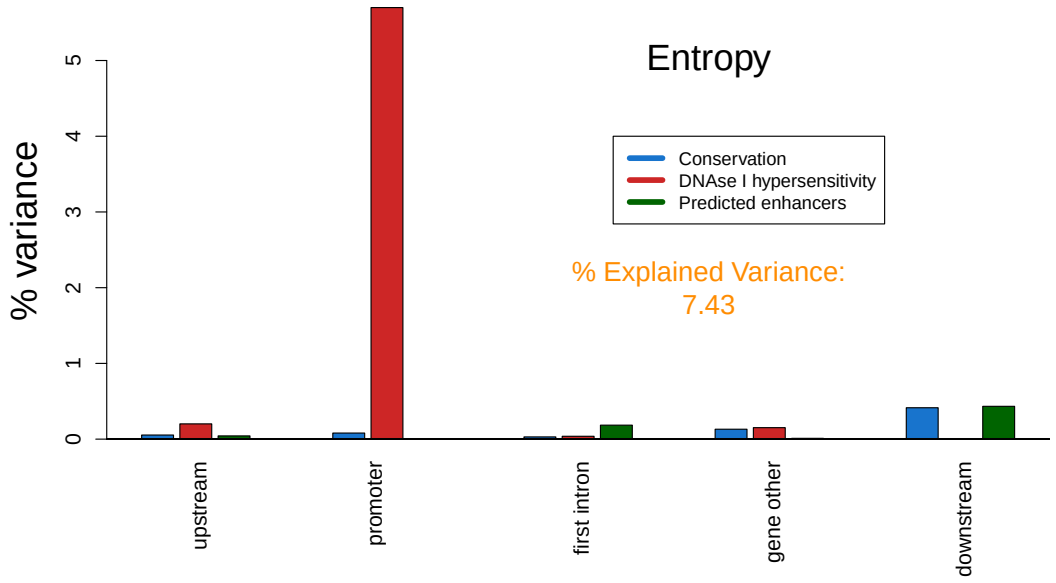


FIGURE 5.33: **Explained percentages of variance in entropy** across five regions of the gene: **upstream, promoter, first intron, gene other and downstream**. Metric used is First (before other variables accounted for), due to correlations between data sources. Total explained variance from all regressors is given in orange.

The all variables model explains 7.43% of the variance in entropy scores, but note this is all based on the effect of DHSs in the promoter region, and none of the other terms were comparatively strong. The interactions model explained 11.0% of the variation

In conclusion, it appears that complex genes (as per the complexity and normalised complexity scores) are enriched in cis regulation in and surrounding the proximity of the gene and first intron and upstream regions appear to be important potential drivers of this. Expression breadth does not appear to be driven by proximal cis- effects (it is not surprising the promoter effect was strong because the more cell types found in ENCODE with accessible chromatin, the more broadly expressed the gene).

However, whilst these effects are highly significant, there is still a large proportion of the variance that is not explained by proximal cis- elements. The results are interpreted further in Chapter 6.

5.6.2 Methods

We referred to ENCODE clustered data available from
http : *//hgdownload.cse.ucsc.edu/goldenpath/hg*19*/encodeDCC/wgEncodeRegDnaseClustered/*,
covering a total of 125 cell types and clustered so that a given cluster covered the presence of any sites across those cell types [John et al., 2011, Thurman et al., 2012].

Enhancer elements were downloaded from the FANTOM5 resource pages, directing to the enhancer atlas [Andersson et al., 2014b]. These enhancers included the same set of cell types over which complexity scores were calculated.

GERP conserved elements were downloaded for each chromosome [Cooper et al., 2005]. Justifications for using these elements were that they were already pre-calculated over appropriate whole genome alignments, and are a widely used dataset with a well established methodology that deals well with gaps in sequence alignments. It is often tricky to deal directly with measuring substitution rates, because when looking at comparisons in coding and non coding sequence (genic/non-genic), there is not necessarily synonymous sites to compare against in non coding regions (makes KS ratios pointless), because mainly interested in non-coding constraint, of which we do not well know the spatial distribution, rather we are looking for clusters of constraint, which is what the GERP constrained elements is measuring.

The *R* package `GenomicRanges` [Koenker, 2013] was used to count the number of sites from each of the three datasets present in the upstream, downstream, upstream promoter, downstream pseudo-promoter, first intron and rest of gene region of the gene. Upstream and downstream distances were calculated at 1000, 10000 and 100000 from the defined 'core promoter', i.e. 250 basepairs from the 3' or 5' of the gene. 10000 bp was used for the final analysis, as this provided the strongest correlation with complexity. Captured genomic regions excluded the promoter regions and exons of neighbouring genes for the DNAse I hypersensitive sites, to avoid bias of recorded sites not necessarily associated with the gene under study. However, enhancers were taken to include the exons of neighbouring genes, to include possible enhancers regulating the given gene from inside other genes.

Counts in the in promoter region is defined as presence or absence of an element, conservation in the gene or first intron is defined as the number of elements overlapping with the entire region. Sites in the first intron of the gene were left both as a raw count and also normalised according to the log of the number of base pairs of the first intron. Since little difference was observed between the two in terms of statistical correlation with complexity, the raw counts were provided in Figure 5.28, but genes with sites above around 60 were grouped into one category for calculating statistical relationships due to sparse data.

Linear models were used to estimate the R^2 effects of the number of sites as covariates and complexity scores as the independent variable. Forward and backward selection were run to find to the most significant two-way interactions.

In order to check for gene length as a confounding factor in the number of overlaps across the gene body, a linear model was calculated, including log gene length as an interaction term with the log of the number of conserved elements across the gene (plus a pseudo count to allow for counts of zero). The interaction was not of significance ($p = 0.056$), nor was the main effect for gene length ($p = 0.079$). This suggests that using the raw number of conserved elements across the gene is better than correcting for gene length. To test gene length correction, the number of conserved elements for MB across the gene was calculated and a linear model calculated to estimate the effect of the log corrected number of sites (plus a small pseudo count) on complexity scores. This term was significant ($p = 1.29e-05$), but much less so than the raw number of counts ($p < 2.2e-16$, adjusted R^2 of 0.001 for the length corrected vs 0.021 for the raw count).

5.7 Histone modifications correlate with complexity scores

As introduced in Chapter 1, histone modifications have been associated with patterns of gene expression. Methylation of histone H3 (H3K4me3) is highly associated with gene activation and is present at the promoters of large numbers of genes [Hussey et al., 2015]. H3K4 tri-methylation is put down by the complex by the methyltransferase

SET1, which associates to a CpG binding protein Cfp1, thus linking H3K4me3 to CpG island associated genes, thus explaining why these genes typically exhibit widespread activation across cell types [Lee et al., 2007, Thomson et al., 2010].

Polycomb proteins are epigenetic regulators of transcription, polycomb mediated methylation of histone H3 (H3K27me3) marked promoters have been commonly associated with repressive marks via the methylation of histones [Di Croce and Helin, 2013]. These marks are generally laid down through development and differentiation but are not required for the initiation of silencing; they appear to have a role in the maintenance of repression in the differential process [Riising et al., 2014]. Moreover, they are not necessary permanent, appearing to act as repression marks in promoters 'poised' for potential activation, particularly when H3K4me3 is also present at the same promoter ([Voigt et al., 2013]). Much evidence has accumulated that the relationships between regulatory elements and expression defined are by histone modifications [Rhie et al., 2014].

The Epigenetics Roadmap project presents large scale mappings of histone modifications across the genome across a variety of tissues and cell types, in what is referred to as 'epigenomes' [Kundaje et al., 2015]. Presented in this section is an attempt to associate signals for commonly studied histone marks in the promoters and across the body of genes with the complexity scores we observed for those genes across the given set of primary cell types.

5.7.1 Complexity in primary cells is highly predicted by combinations of H3K27me3 repressive marks and H3K4me3 activation marks

Figure 5.34 shows the relative distributions of complexity and entropy scores over H3K4me3, H3K27me3 and combinations of cell types containing both marks

the nine combinations of low, medium and high H3K4me3 and H3k27me3 histone marks observed within the core promoter region. As expected, high expression breadth (high entropy) is related to low H3K27me3 marks and high H3K4me3 marks. The highest complexity scores were observed in those genes with high H3K27me3 marks and low

H3K4me3 marks. These genes correspond to the set with the lowest entropy, suggesting that these complex genes are in general more restricted in their expression than other combinations.

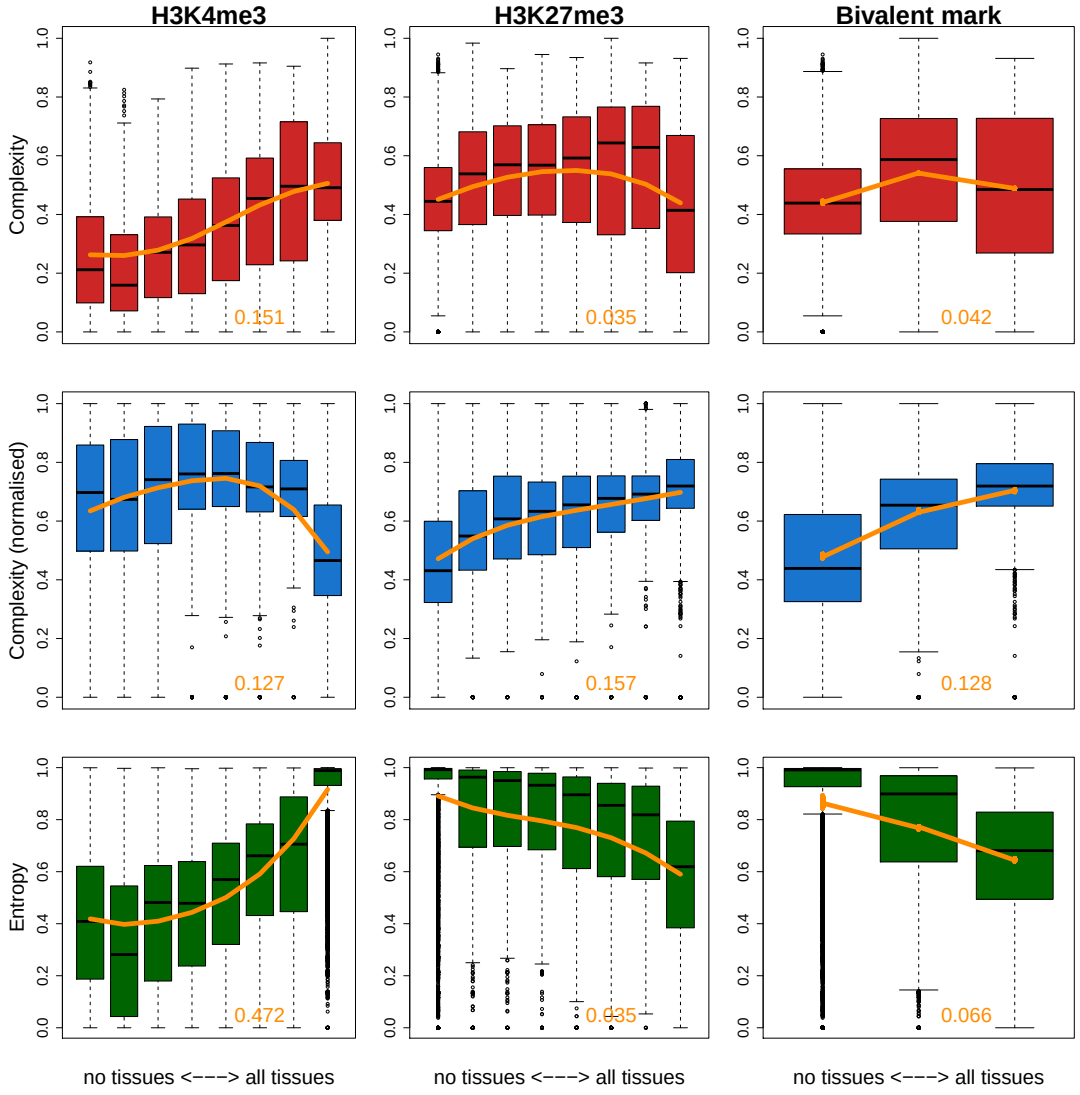


FIGURE 5.34: **Breadth of H3K4me3, H3K27me3 and bivalent marks against complexity scores**: complexity (red), normalised complexity (blue), entropy (green). Left of scale: modifications present in no epigenomes of no tissues at gene promoter, right of scale: modifications present in all epigenomes of all tissues at gene promoter. Orange lines represent smooth best fit lines based on the `loess` function in *R*. Orange numbers represent explained proportion of variance from the `lm` function in *R*, treating modification count as a factor dependent variable.

To observe how histone marks change over complexity scores, Figure 5.25 shows how H3K27me3, H3K27ac and H3K4me3 marks in the core promoter region of genes change

on average as complexity increases. Over the region with high activation signal (H3K4me3), H3K27me3 marks increase with increasing complexity, suggesting that polycomb marks may be a predictor of regulation as opposed to simply a predictor of expression breadth. Both H3K27ac and H3K4me3 marks decrease in the region of high complexity, as a result of highly complex sample restricted genes.

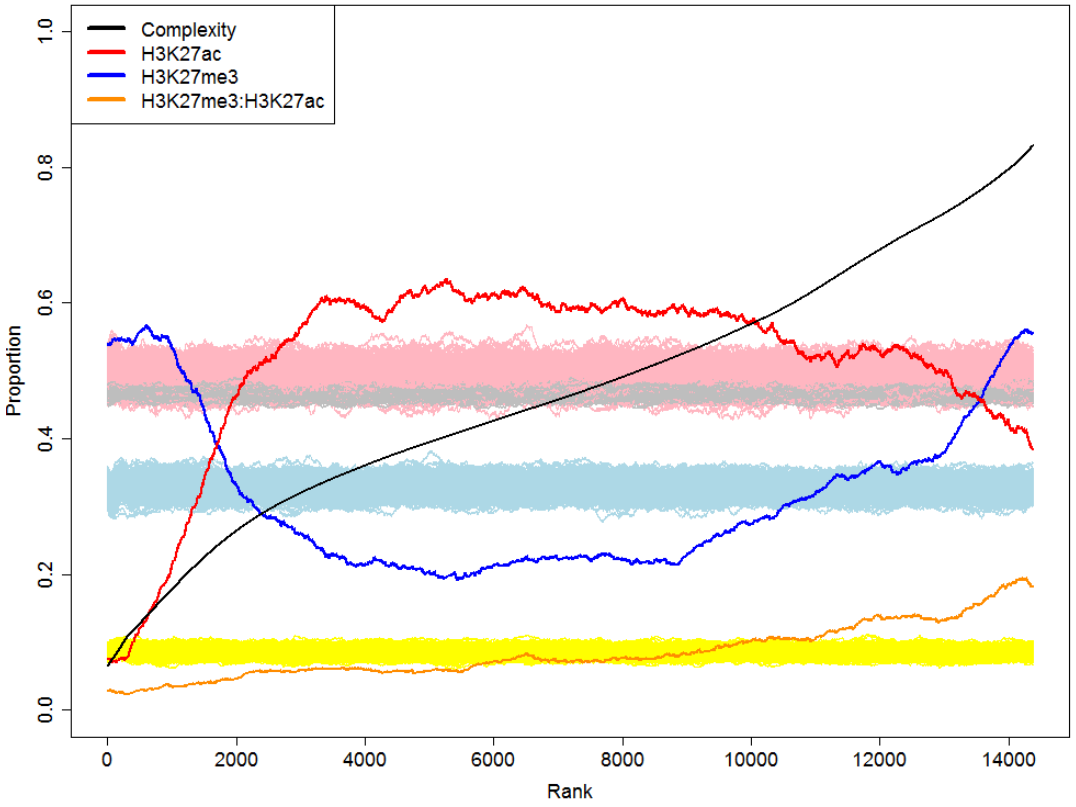


FIGURE 5.35: **Proportion of genes associated with H3K27ac marks in their core promoter, proportion of genes associated with H3K27me3 marks in their core promoter and proportion associated with both H3K27me3 and H3K27ac**. Genes are ranked in order of complexity and each data point for each variables is calculated as its averaged values from the current gene and including up to the next 1000 genes. Background distributions are calculated by permuting the ranks of the complexity scores and recalculating the proportions. Proportions of H3K27ac present genes are plotted in red with a pink background distribution, complexity is plotted in black with a grey background distribution, H3K27me3 proportion is plotted in blue with a light blue background distribution and H3K27ac:H3K27me3 interaction is plotted in orange with a yellow background distribution.

To statistically quantify the effect of histone marks on complexity scores, quantile

regression was run to estimate their effects on low complexity and high complexity genes. As expected, H3K27me3 marks have a strongly increasing effect on complexity overall and particularly in the upper quartiles, relating to high entropy genes.

Thus, it is observed that regulatory complexity scores highly associate with the chromatin signals H3K4me3 and H3K27me3. In particular, the relationship with H3K27me3 suggests that the high complexity scores in genes are due to the on-off switching observed across cell types, potentially due to their poised, silenced state associated with this modification.

5.7.2 Complexity in primary cells is weakly associated with H3K9me3 and H3K36me3 signal recorded over the gene body

As well as histone promoter region marks, epigenetic signals across the gene body have been associated with activation/inactivation across developmentally regulated genes [Dambacher et al., 2010], two being H3K9me3 and H3K36me3. In particular, H3K36 histone residues are methylated co-transcriptionally by the RNA polymerase II SET2, and appear to play a role in maintaining chromatin spacing in yeast during transcriptional elongation [Venkatesh and Workman, 2013]. Thus, it is strongly associated with gene activation and should correlate with entropy scores, although its association with gene complexity scores is less clear.

Figure 5.36 shows the distribution of the the log signal of H3K9me3 and H3K36me3 recorded over the gene body and Figure 5.37 shows the correlation between the two signals and complexity scores.

Complexity scores are significantly positively correlated with both H3K9me3 and H3K36me3 (Figure 5.37), with respective R^2 values of 0.006 and 0.016, corresponding to explained variance percentages of 0.6% and 1.6%. Thus, H3K36me3 is more highly correlated with complexity, although the effect sizes themselves are very small, possibly having small biological significance.

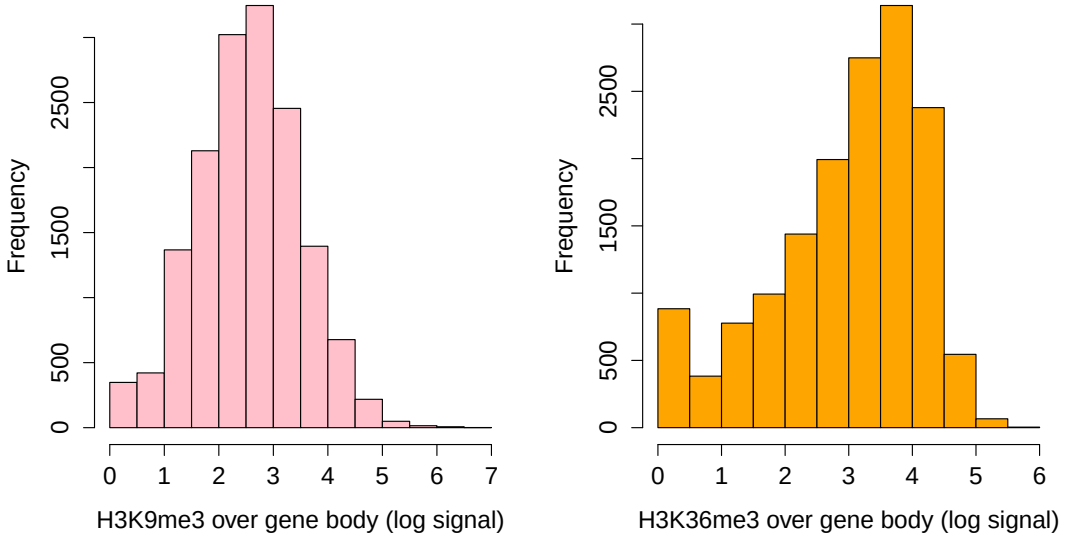


FIGURE 5.36: **Histograms of observed H3K9me3 and H3K36me3 over the body of genes**. Frequencies given in terms of the log of the signal.

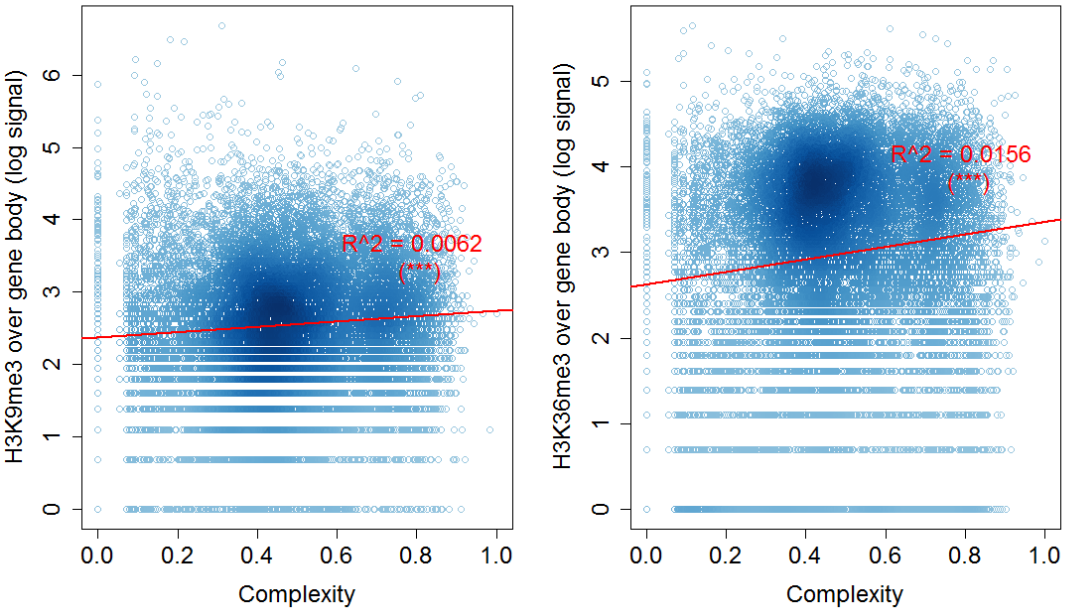


FIGURE 5.37: Scatter plots of log signal of histone mark **H3K9me3 vs complexity**, and the log signal of histone mark **H3K36me3 vs complexity**. Red lines indicate best fit lines from linear model, with R^2 values calculated from the model. Significance is given as (***), which implies that the slope of the line is highly significant.

Figure 5.38 shows the same plot but with entropy scores. Whilst H3K9me3 is weakly correlated with entropy ($R^2 = 0.002$), H3K36me3 is highly correlated ($R^2 = 0.1362$), suggesting that, as expected, H3K36me3 is a signal of activation; the more H3K36me3, the broader the expression of the gene.

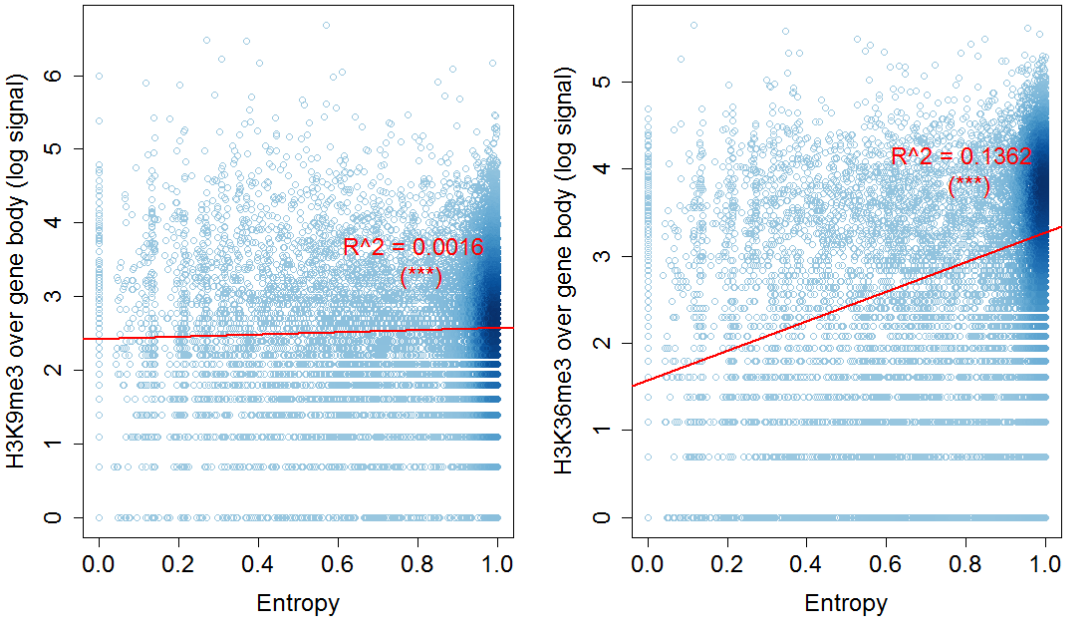


FIGURE 5.38: Scatter plots of log signal of histone mark **H3K9me3 vs entropy**, and the log signal of histone mark **H3K36me3 vs entropy**. Red lines indicate best fit lines from linear model, with R^2 values calculated from the model. Significance is given as (***), which implies that the slope of the line is highly significant.

In conclusion, genes associated with H3K36me3 over the gene body are broadly expressed but also show a moderate increase in complexity, suggesting that this modification associates with genes which undergo more dynamic changes in expression, potentially through its regulatory mechanism of maintaining chromatin states at these genes [Venkatesh and Workman, 2013].

5.7.3 Explained variance from epigenetic modifications

Figure 5.39 shows the percentage of the explained variance of each of 6 histone modifications for complexity scores: H3K4me3, H3K27me3, bivalent (H3K4me4 and H4K37me3

together in the same tissue), H3K27ac, all overlapping the promoter region, and H3K9me3 and H3K36me3 overlapping the gene body.

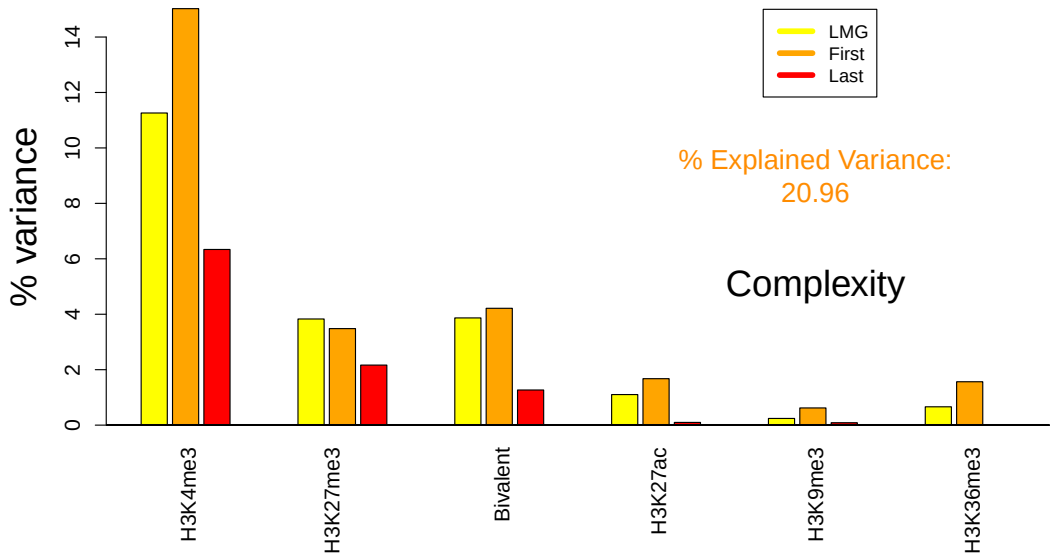


FIGURE 5.39: **Explained percentages of variance in complexity scores for epigenetic variables**. Metrics used are LMG (averaged contributions, e.g. see [Chevan and Sutherland, 1991]), First (before other variables accounted for) and Last (after other variables accounted for). Total explained variance from all regressors is given in orange.

Figure 5.40 shows the percentage of the explained variance of each of 6 histone modifications for normalised complexity scores, based on three metrics (first in model, last in model and LMG, balanced contributions).

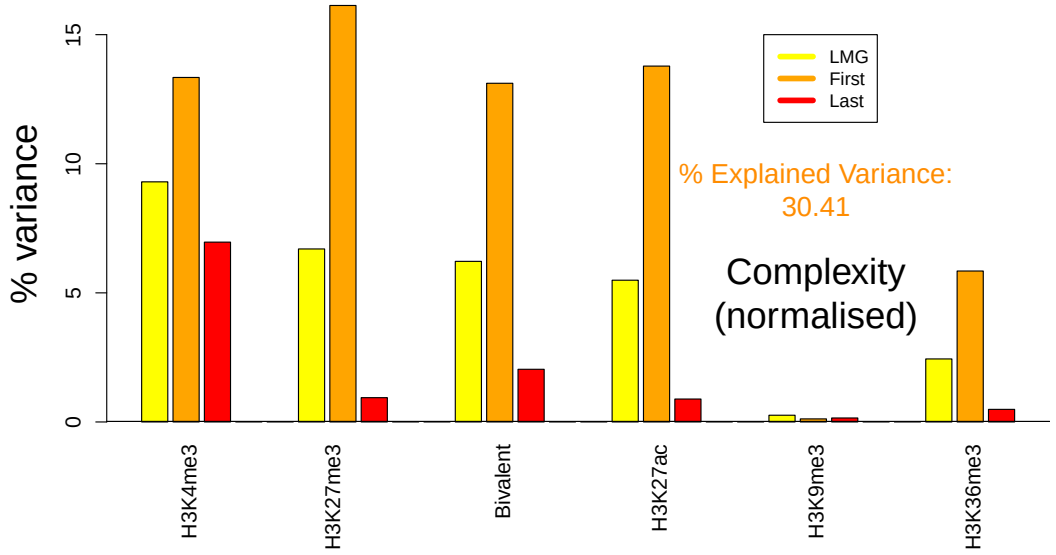


FIGURE 5.40: **Explained percentages of variance in normalised complexity scores for epigenetic variables**. Metrics used are LMG (averaged contributions, e.g. see [Chevan and Sutherland, 1991]), First (before other variables accounted for) and Last (after other variables accounted for). Total explained variance from all regressors is given in orange.

Figure 5.41 shows the percentage of the explained variance of each of 6 histone modifications for entropy scores, based on three metrics (first in model, last in model and lmg, balanced contributions). As was seen in the previous section, H3K4me3, associated with activation, dominates all three metrics. H3K27ac came second in the LMG classifications, followed by H3K27me3. Almost all of the variables other than H3K4me3 are heavily reduced in 'last', suggesting that they are not as important once H3K4me3 has already been accounted for. H3K9me3 does not appear to be important for defining breath of expression.

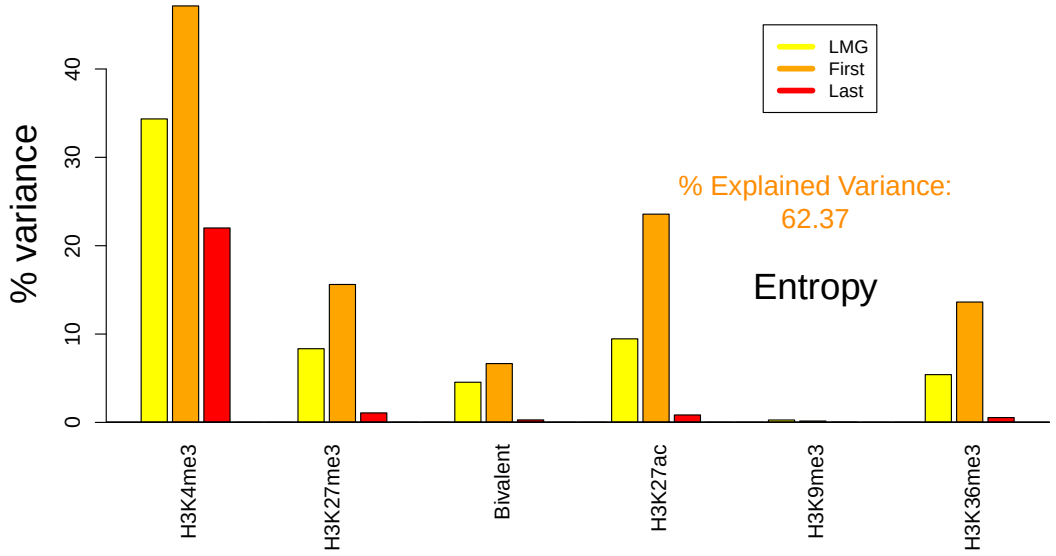


FIGURE 5.41: **Explained percentages of variance in entropy scores for epigenetic variables**. Metrics used are LMG (averaged contributions, e.g. see [Chevan and Sutherland, 1991]), First (before other variables accounted for) and Last (after other variables accounted for). Total explained variance from all regressors is given in orange.

5.7.4 Interactions between CpG genes and histone modifications

The histone mark H3K27me3 interacts with the promoters of untranscribed genes associated with CpG islands, maintaining their silenced state and cell identity [Li et al., 2014, Riising et al., 2014]. To more closely interrogate the strong relationship observed between combinations of active and repressive histone promoter modifications and the

normalised complexity scores, the histone modifications H3K27me3 and H3K4me3 were compared with CpG effects, shown in Figures 5.42 and 5.43.

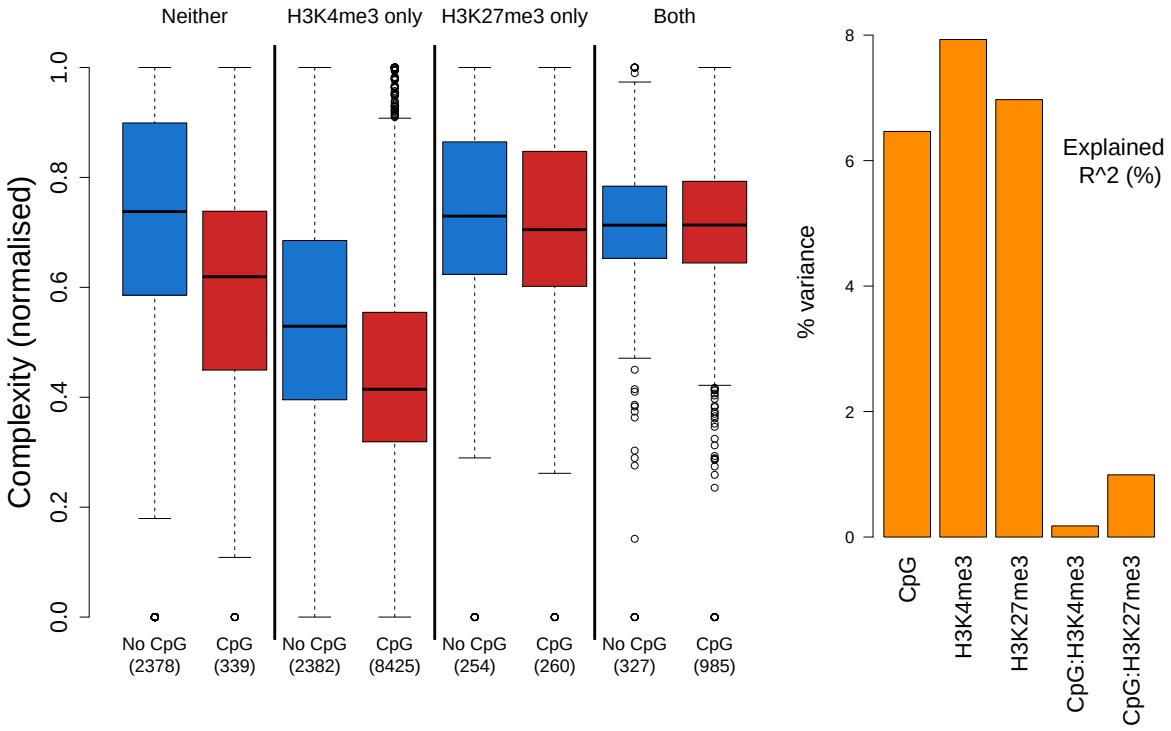


FIGURE 5.42: (Left): **Interactions between CpG, H3K4me3 and H3K27me3 presence/absence at the core promoter** Numbers indicated in brackets are the numbers of genes in each category. Genes containing either mark are broadly associated with that mark in their promoter across all analysed epigenomes, genes with neither are all those not broadly associated with either mark in their promoter. (Right): **Explained proportions of variance in CpG, H3K4me3 and H3K27me3 and CpG interactions**. Relative proportions of explained variance of normalised complexity scores are calculated based on the `lm` (linear model) function in *R*. Total explained variance for model is 22.44%.

Differences between the presence and absence of CpG island were the most prominent in genes which did not display either modification in their promoters across all analysed epigenomes (Figure 5.42, left), followed by those genes which displayed only activation marks in their promoters. Genes not broadly exhibiting either mark and and not associated with a CpG island were highly complex (p-value $<$ 2.2e-16, t.test with mean increase of 0.20 when compared with all other genes), with a mean of 0.70. The least complex genes were those broadly exhibiting H3K4me3 but not H3K27me3

and an associated CpG island. These genes are likely to be ubiquitously expressed housekeeping type genes, which have been shown to have reduced complexity.

Genes exhibiting broad H3K27me3 marks but not H3K4me3, and genes broadly exhibiting both marks showed very little difference between their CpG association (p-values 0.90 and 0.93 based t.test for differences in means,for each case respectively).

The CpG interaction with H3K4me3 and H3K27me3 explains 0.17% and 0.98% of the variation in normalised complexity scores, respectively (Figure 5.42, right), suggesting that CpG associations interact more strongly with H3k27me3 histone modifications than H3K4me3.

Figure 5.43 shows in more detail the interactions between specifically bivalent genes and those genes which exhibit no marks whatsoever in their core promoter across any of the analysed epigenomes. As shown in Figure 5.42, genes exhibiting both marks in all epigenomes showed no significant difference in CpG association, however genes with no potential bivalent marks across any epigenome show dramatically different CpG island presence; those without CpG islands are the least complex (p-value $<$ 2.2e-16, t-test compared with all other genes).

This suggests that not only are genes associated with bivalent marks are more complex, but bivalent genes are complex independent of their CpG association. CpG islands only appear to be influencing the complexity of the gene in those genes were bivalency is not observed, more specific the bivalency marks, the stronger the effect of CpG presence.

5.7.5 Methods

Human 'epigenomes' were download from the Roadmap Epigenomics Consortium ([Kundaje et al., 2015]). A selection of 22 primary tissues were chosen incorporating a range of different tissue groups; the selection of groups and associated IDS are given in Appendix D. For each of these epigenomes, five datasets were downloaded, each containing lists of broad domains enriched for histone ChIP-seq peaks. The five datasets corresponded to five commonly studied histone modifications: H3K4me3, H3K27me3, H3K9me3, H3K36me3 and H3K27ac.

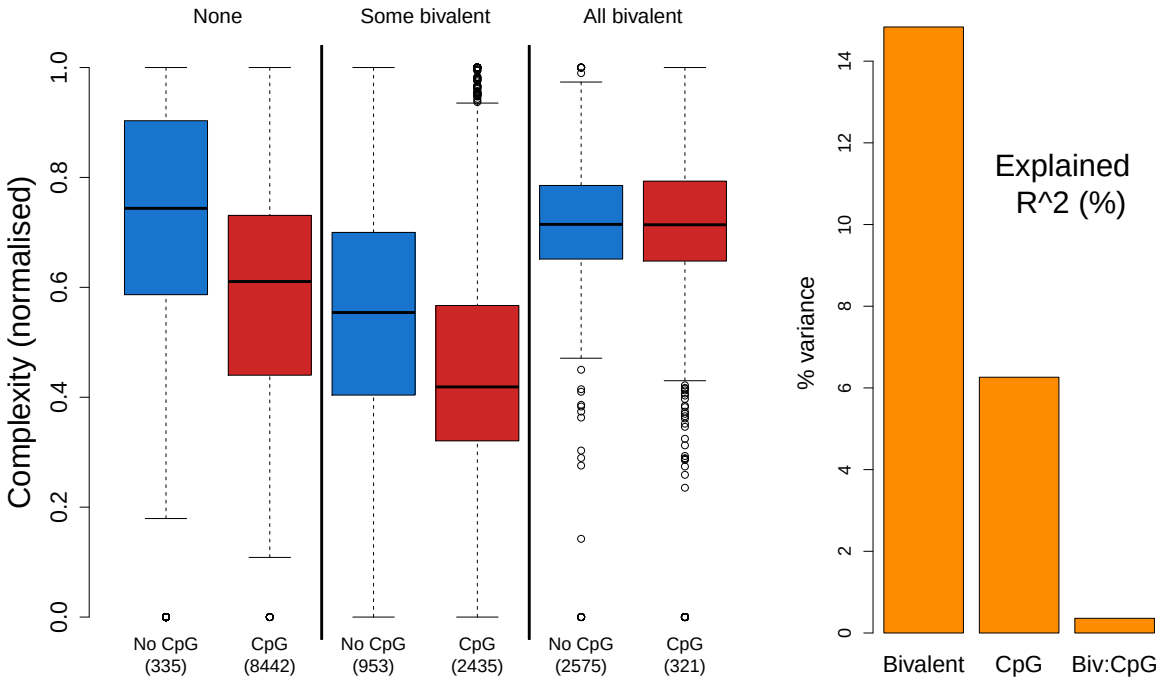


FIGURE 5.43: **Interactions between CpG and bivalency breadth at the core promoter (left)** and **Explained proportions of variance in CpG, bivalency and CpG:bivalent interaction (right)**. "All bivalent" bivalency is defined here as a bivalent mark in all 22 tissues at a gene's promoter, "None" is where none of the 22 tissues have a bivalent mark at the gene's promoter, and "some bivalent" is in between these two states. Numbers indicated in brackets are the numbers of genes in each category.

For H3K4me3, H3K27me3 and H3K27ac modifications, the coordinates for the peaks in each dataset was overlapped with the core promoter region of the gene according to $refseq$ (between 1 and 250 bp upstream of the TSS of the gene), and a count of 0 or 1 allocated to the gene corresponding to presence or absent of the modification in the core promoter. A count from 0 to 22 was then achieved according to how many of the 22 datasets contained the signal in the core promoter for a gene. Due to large numbers of genes in categories 0 and 22 and smaller number within categories 1 to 21 (figure ref), these internal counts were compacted together to achieve a final count from 0 to 6, representing breadth of histone signal for the core promoter of the gene.

In order to determine the bi-valency status of the gene, binary vectors of presence or absence of the modification H3K4me3 over the 22 primary tissues were multiplied with

the corresponding binary vectors for H3K27me3. The result was then summed to give a value from 0 to 22 of how many of the primary tissues contained both marks together for a given gene. As before, these groups were then collapsed to give a breadth of bi-valency from 0 to 6.

For H3K9me3 and H3K36me3 signals, overlaps were measured across the entripy gene body based on gene start and gene end $refseq$ coordinates.

5.8 Protein age

Older genes are more broadly expressed whilst younger genes tend to be restricted in their expression [Hao et al., 2010]. In order to test for the effects of age on gene expression scores, protein age was downloaded from ProteinHistorian [Capra et al., 2012].

Figure 5.44 shows the distribution of complexity, normalised complexity and entropy stratified by gene age (left) and the variance of scores across that gene age (right). In all scores, the variance of the the set of genes generally increases with decreasing age; the newer the gene, the more variety of gene expression patterns are observed.

Complexity scores stratified by protein age gave an approximate R^2 value of 1.4% and 2.7% for the 0.2 and 0.8 quantile respectively. Complex genes increased dramatically in their complexity from unicellular organisms, though the development of multicellularity and peaking at Euteleostomi ($p < 2e\text{-}16$, quantile regression) before reducing to above baseline at the Theria stage ($p < 2e\text{-}16$) and increasing modestly to Human.

Normalised complexity scores stratified by protein age gave an approximate R^2 value of 2.5% and 7.3% for the 0.2 and 0.8 quantile respectively. Complex genes increased dramatically in their complexity from unicellular organisms, and in generally continued to increase with newer genes, thus suggesting an upward trend of complexity as new genes evolve.

For both complexity and normalised complexity, the upper quantile (0.8), or the highly complex genes, is more significantly changing than the lower quantile (0.2), or the genes

of low complexity, suggesting that newer genes appear to evolve with regulation causing more and more complex gene expression patterns, but newer genes also still evolve with patterns of less complexity - for complexity scores, the drift is towards more complex and less complex than older categories of genes evolving together. In normalised complexity scores there is more of an active shift, with both quantiles shifting towards the more complex, but the lower quantile less so than the upper. These findings contradict [Warnefors and Eyre-Walker, 2011a], who find that older genes are more complex in their regulatory mechanisms, since they have had more time to accumulate regulation.

Figures 5.45 and 5.46 agree with [Warnefors and Eyre-Walker, 2011a] - older genes (newer than multicellularity but older than mammalia) have greater numbers of conserved sites overlapping the gene, and greater numbers of DHSs in their first intron.

This suggests that newer genes might be undergoing different regulatory processes than what is captured by DHSs or conservation; perhaps more recently evolved genes are more involved in protein-protein interactions offsite of the DNA. This is discussed further in Chapter 6.

In contrast the lower quantiles of entropy scores (approximate R^2 of 0.11 for 0.2 quantiles) dramatically decrease across time with the increase around the Theria stage mirroring the decrease in complexity observed around the same time, before increasing once again and decreasing slightly in Human. The upper quantile remains constant through time (approximate R^2 of 0.003 for 0.8 quantiles). The earliest genes (those relating to Cellular organisms, Eukaryota and Opisthokonta) have extremely skewed entropy distributions. High entropy corresponds to high expression breadth, suggesting that the oldest genes were enriched in those with housekeeping function. Thus, newer genes appear to evolve more and more specificity, but newer genes may also be broad in their expression.

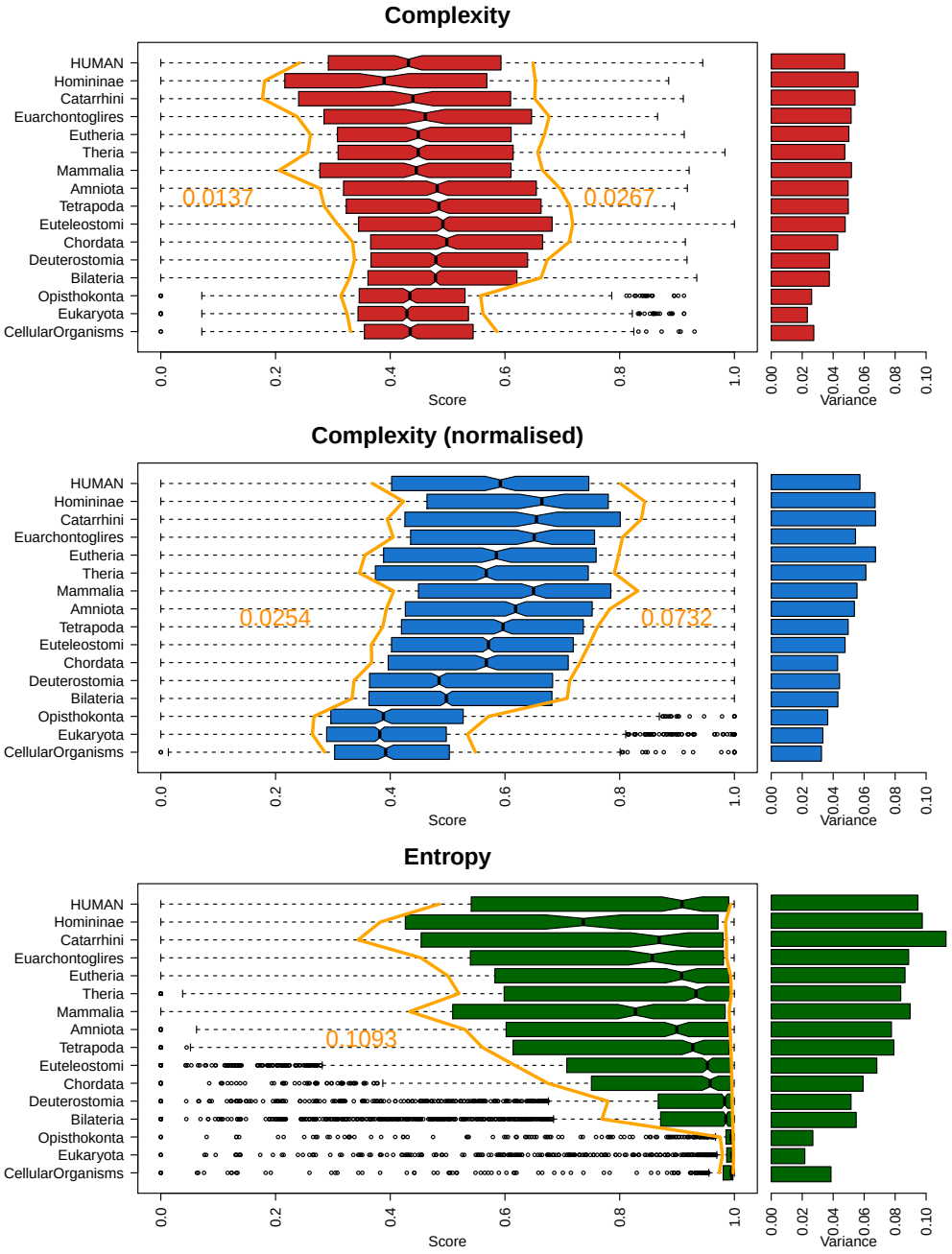


FIGURE 5.44: **Evolution of regulatory complexity**: **Complexity scores (top), normalised complexity scores (centre) and entropy scores (bottom)** across 16 time points for age of related protein, from cellular organisms to human. Left and right orange curves represent fitted estimates from a quantile regression model, fitted at the 0.2 and 0.8 quantiles respectively. Numbers next to the curves represent approximated R^2 values from the entire model. This value is 0.003 for the right hand curve of the entropy score. Entropy normalized to a maximum of 1.

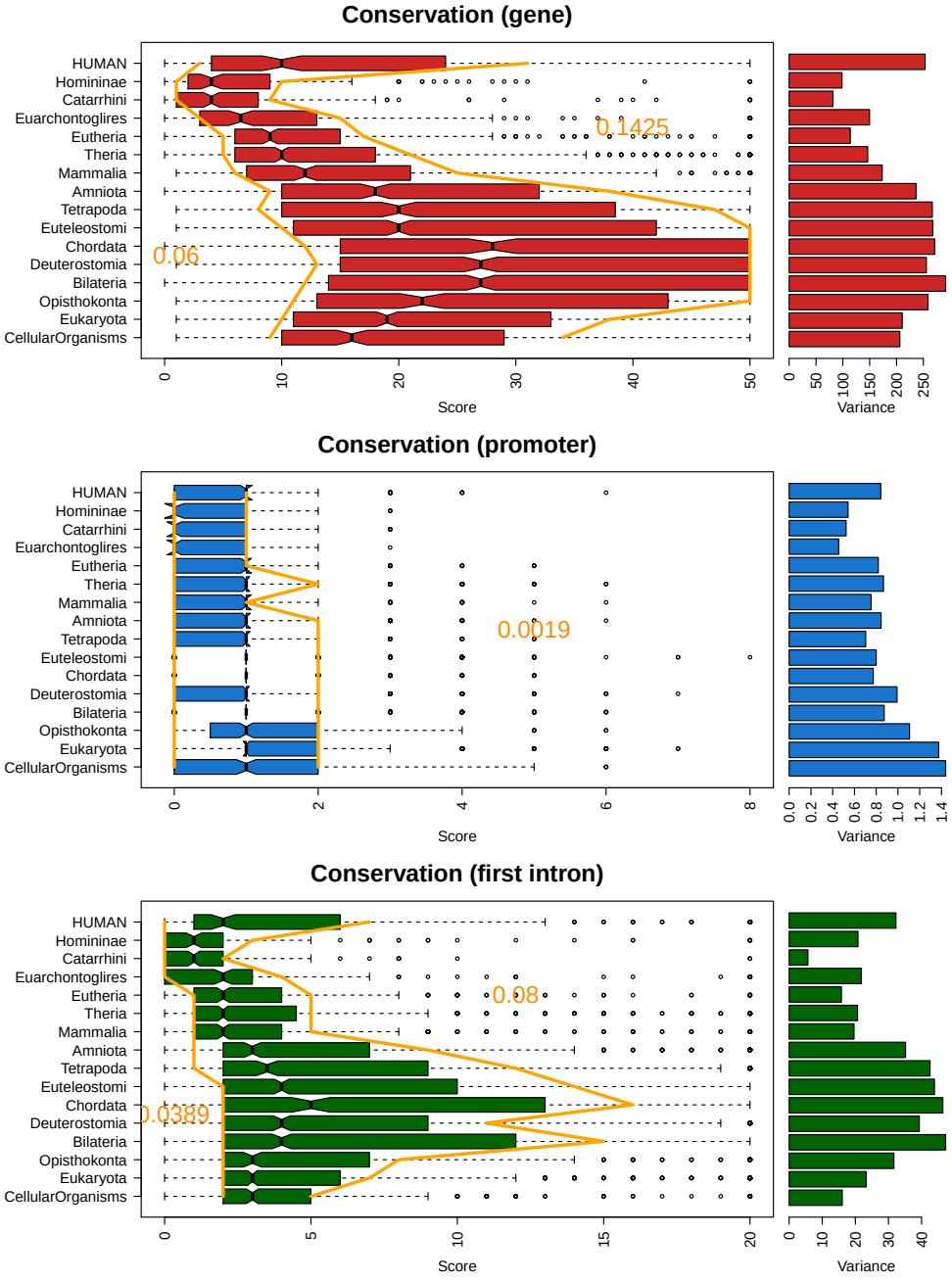


FIGURE 5.45: **Evolution of regulatory complexity**: **Conserved GERP sites across gene (not including first intron) (top), conserved GERP sites in promoter region (centre) and conserved GERP sites in first intron (bottom)** across 16 time points for age of related protein, from cellular organisms to human. Left and right orange curves represent fitted estimates from a quantile regression model, fitted at the 0.2 and 0.8 quantiles respectively. Numbers next to the curves represent approximated R^2 values from the entire model. This value is 0.00 for the left hand curve of the centre plot. GERP site counts are capped to a maximum of 50 for the whole gene and 20 for the first intron.

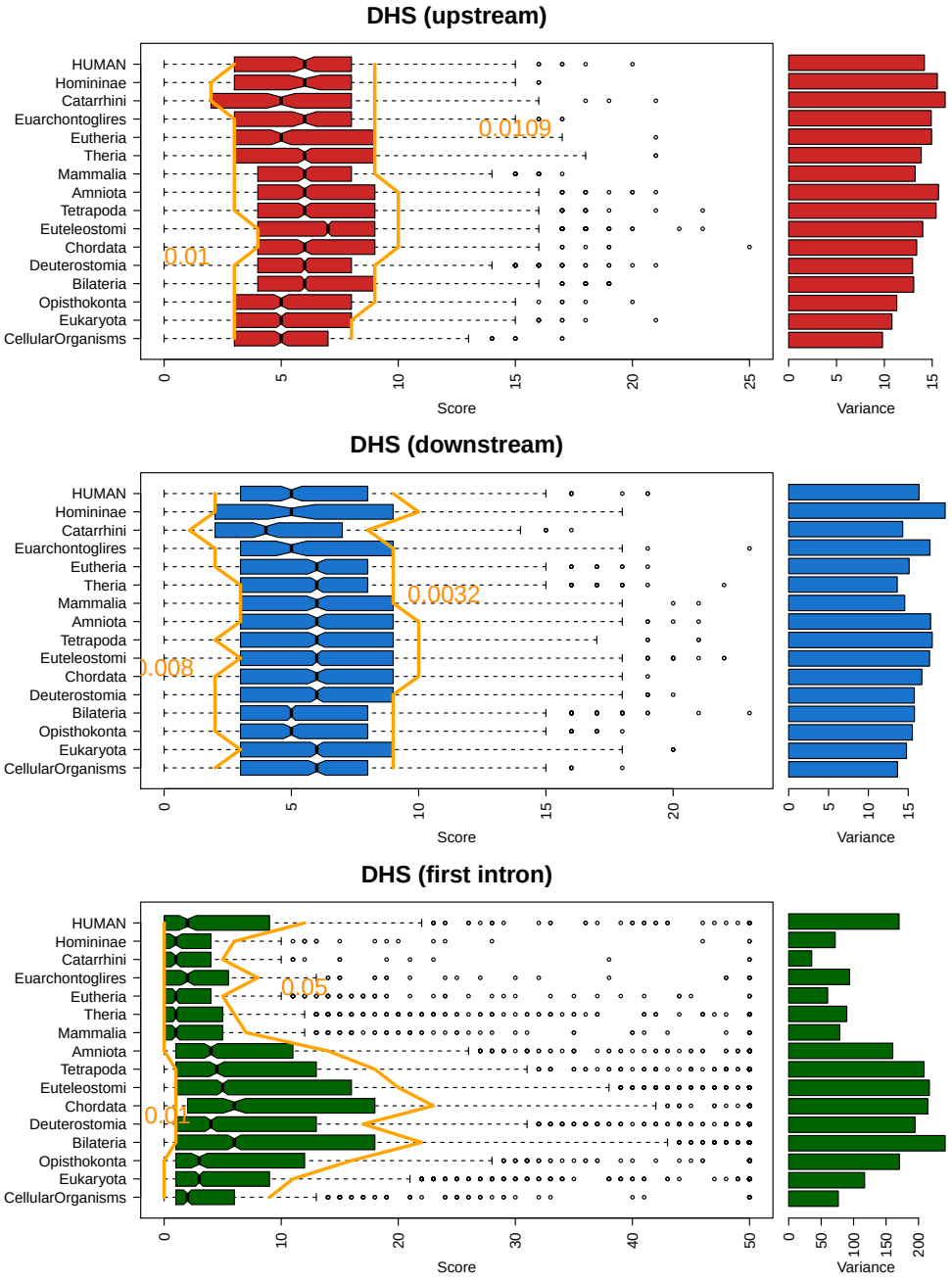


FIGURE 5.46: **Evolution of regulatory complexity**: **Number of DHSs upstream of gene (top), number of DHSs downstream of gene (centre) and number of DHSs in the first intron (bottom)** across 16 time points for age of related protein, from cellular organisms to human. Left and right orange curves represent fitted estimates from a quantile regression model, fitted at the 0.2 and 0.8 quantiles respectively. Numbers next to the curves represent approximated R^2 values from the entire model. This value is 0.01 for the left hand curve of the bottom plot. DHS site counts are capped to a maximum of 50 for the first intron.

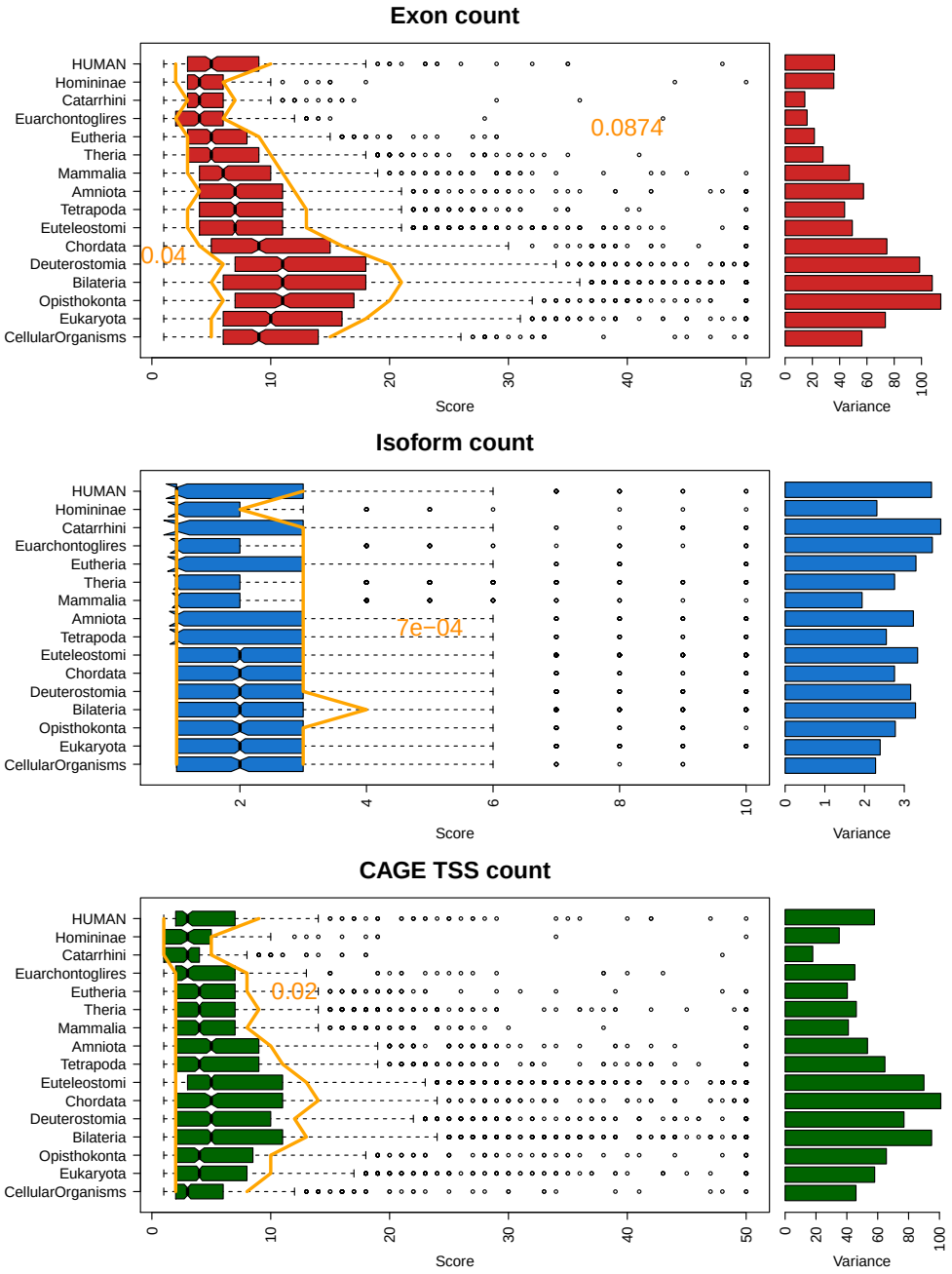


FIGURE 5.47: **Evolution of regulatory complexity**: **Exon count (top), number of isoforms (centre) and number of CAGE annotated TSS (bottom)** across 16 time points for age of related protein, from cellular organisms to human. Left and right orange curves represent fitted estimates from a quantile regression model, fitted at the 0.2 and 0.8 quantiles respectively. Numbers next to the curves represent approximated R^2 values from the entire model. This value is approximately 0 for the left hand curve of the bottom plot and centre plots. TSS and exon counts are capped to a maximum of 50 for illustrative purposes.

5.8.1 Methods

Protein age data was downloaded from ProteinHistorian ([Capra et al., 2012]), categorizing proteins into distinct phylogenetic ages. Data for human involves 16 states, from cellular organisms to human. Protein ids were converted to associated gene ids and compared to scores of gene expression.

Due to the unevenness in the shapes of the distributions across quantiles, quantile regression was used to estimate the significance of changes in scores compared to cellular organisms, with $\tau = 0.2$ for the lower quantiles and $\tau = 0.8$ for the upper quantiles. Quantile regression was based on the `rq` function from the `quantreg` package in *R*.

5.9 What proportion of variation in complexity scores can we explain from our studied variables?

In previous sections, it has been concluded that genomic variables correlate significantly with complexity scores. In particular, the number of hypersensitive sites upstream and within the first intron of the gene, histone modifications in the gene promoter, the number of associated TSS and protein age all appear to have the strongest predictive effect on the regulatory expression complexity.

Figures 5.48 to 5.53 display the variance explained by the main effects of each variable of the complexity scores. They show the variance explained by all of the single effects together and the variance explained including all two-way interactions.

5.9.1 Total variance explained in complexity scores

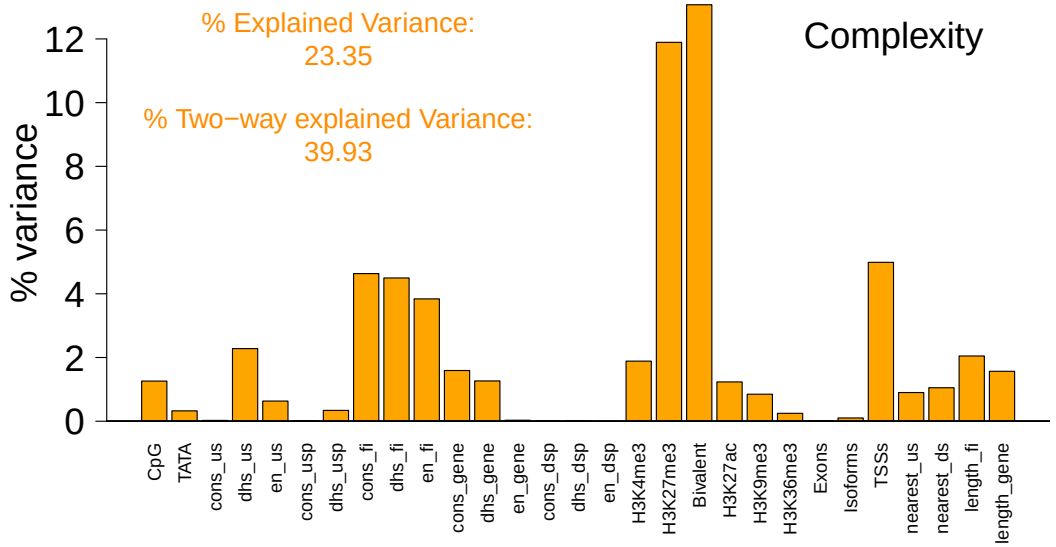


FIGURE 5.48: **Contribution of the variance of the complexity scores explained by each of the 29 variables, based on the "first" metric**. *Abbreviations:* cons = conservation, us = upstream, usp = upstream promoter, dhs = DNase I hypersensitive site, en = enhancer, fi = first intron, gene = gene body minus the first intron, ds = downstream.

Figure 5.48 shows the contributions of the effects of 29 variables on complexity, based on the 'first' metric (individual main effects ignoring the effects of all other variables). As has been seen in previous sections, epigenetic modifications in the core promoter had the strongest effect together with the effects of a CpG island in the core promoter. Cis-regulatory elements also had significant effects and distance based parameters and TSS count had weak but significant effects.

All of the variables together without interactions explain a total of 23.35% of the variation in complexity. With all two way-interactions this increases to 39.93%.

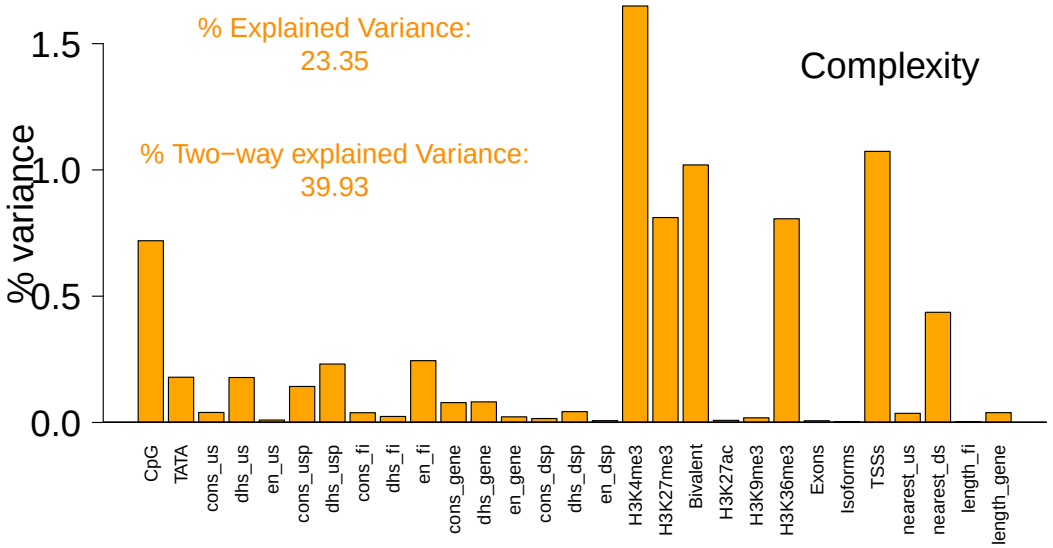


FIGURE 5.49: **Contribution of the variance of the complexity scores explained by each of the 29 variables, based on the "first" metric**. *Abbreviations:* cons = conservation, us = upstream, usp = upstream promoter, dhs = DNase I hypersensitive site, en = enhancer, fi = first intron, gene = gene body minus the first intron, ds = downstream.

Overall, it appears that cis-regulation surrounding the gene are important indicators of regulatory complexity, as well as H3K27me3 and H3K4me3, together with their combinations, which appears to dominant the explained variance. Furthermore, a lot variance is explained by two way interactions between the variables, which reiterates that genes are not regulated by single processes alone; but a combination of multiple factors.

5.9.2 Total variance explained in normalised complexity scores

Figure 5.48 shows the contributions of the effects of 29 variables on normalised complexity, based on the 'first' metric (individual main effects ignoring the effects of all other variables). As has been seen in previous sections, epigenetic modifications in the core promoter had the strongest effect together with the effects of a CpG island in the core promoter. Cis-regulatory elements also had significant effects and distance based parameters and TSS count had weak but significant effects.

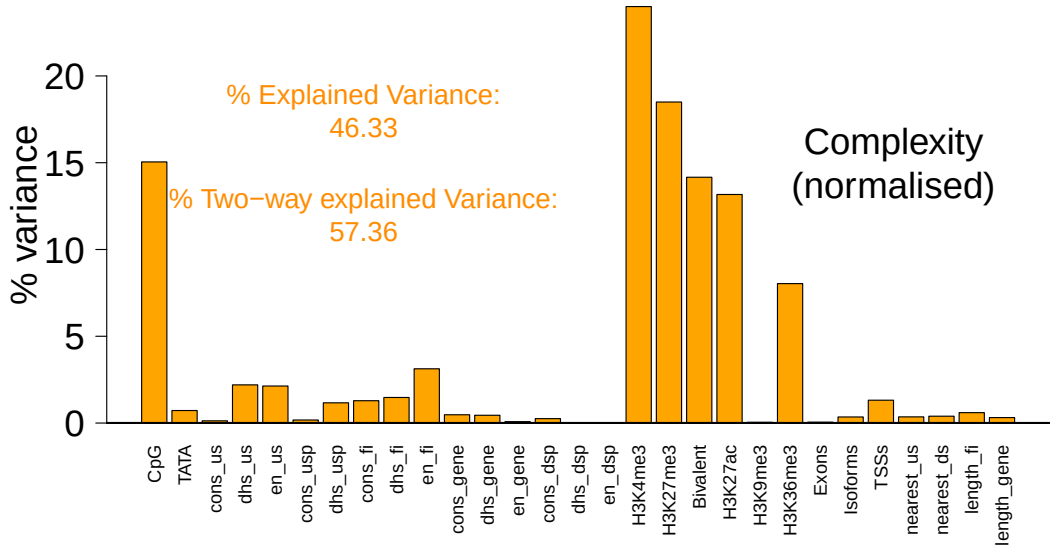


FIGURE 5.50: **Contribution of the variance of the normalised complexity scores explained by each of the 29 variables, based on the "first" metric**. *Abbreviations:* cons = conservation, us = upstream, usp = upstream promoter, dhs = DNase I hypersensitive site, en = enhancer, fi = first intron, gene = gene body minus the first intron, ds = downstream.

All of the variables together without interactions explain a total of 46.33% of the variation in complexity. With all two way-interactions this increases to 57.36%.

Figure 5.49 shows the contributions of the effects of 29 variables on normalised complexity, based on the 'last' metric (individual main effects ignoring the effects of all other variables).

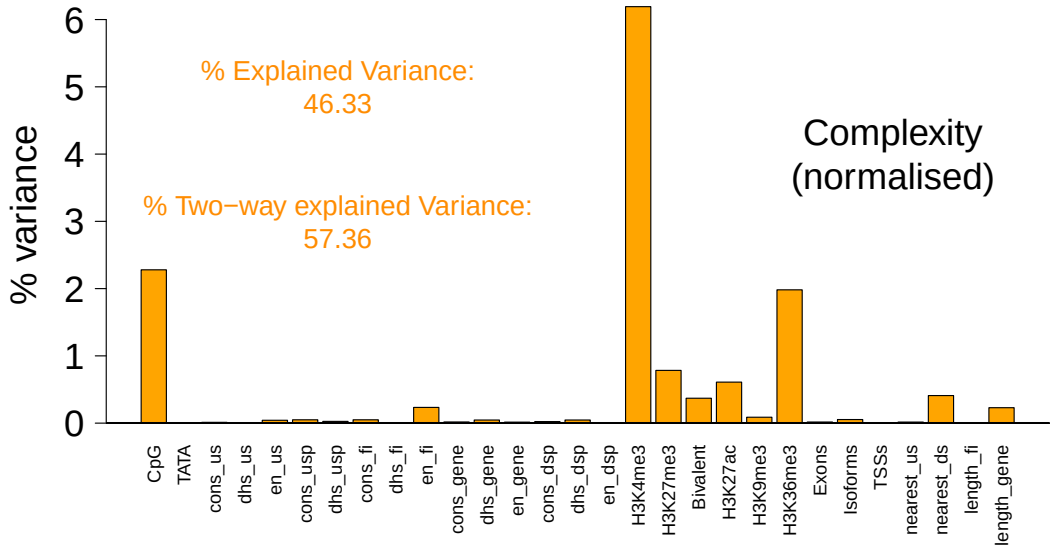


FIGURE 5.51: **Contribution of the variance of the normalised complexity scores explained by each of the 29 variables, based on the "first" metric**. *Abbreviations:* cons = conservation, us = upstream, usp = upstream promoter, dhs = DNase I hypersensitive site, en = enhancer, fi = first intron, gene = gene body minus the first intron, ds = downstream.

It appears that more variation in total is explained by the normalised complexity scores compared to complexity scores, especially when considering the variation explained by the interaction effects. Looking closely, much of this extra variance is due to the added explained variance due to CpG effects and chromatin modifications. It is hardly surprising that CpG effects would increase, since normalised complexity scores are, by design, more tissue restricted, which is highly associated with CpG depletion. Furthermore, since there is less overall activation due to tissue specificity, it is neither surprising that these scores correlate more with H3K4me4. Thus, the added variance explained compared to the complexity scores does not appear to present a case for their overall significance over simply using complexity scores alone in analyses.

5.9.3 Total variance explained in entropy scores

Figure 5.52 shows the contributions of the effects of 29 variables on normalised complexity, based on the 'first' metric (individual main effects ignoring the effects of all other variables).

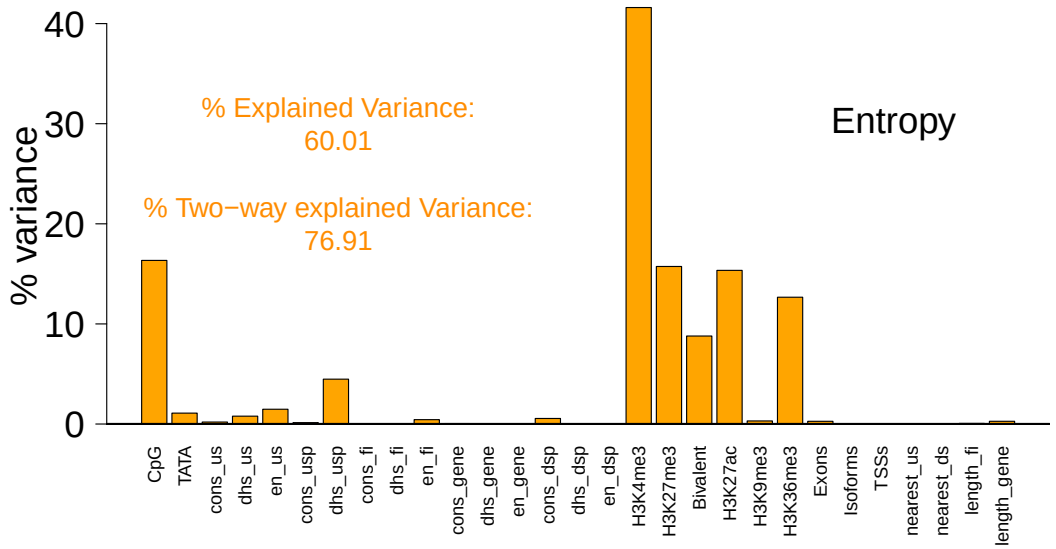


FIGURE 5.52: **Contribution of the variance of the entropy scores explained by each of the 29 variables, based on the "first" metric**. *Abbreviations:* cons = conservation, us = upstream, usp = upstream promoter, dhs = DNase I hypersensitive site, en = enhancer, fi = first intron, gene = gene body minus the first intron, ds = downstream.

All of the variables together without interactions explain a total of 60.1% of the variation in complexity. With all two way-interactions this increases to 76.91%.

Figure 5.53 shows the contributions of the effects of 29 variables on normalised complexity, based on the 'last' metric (individual main effects ignoring the effects of all other variables).

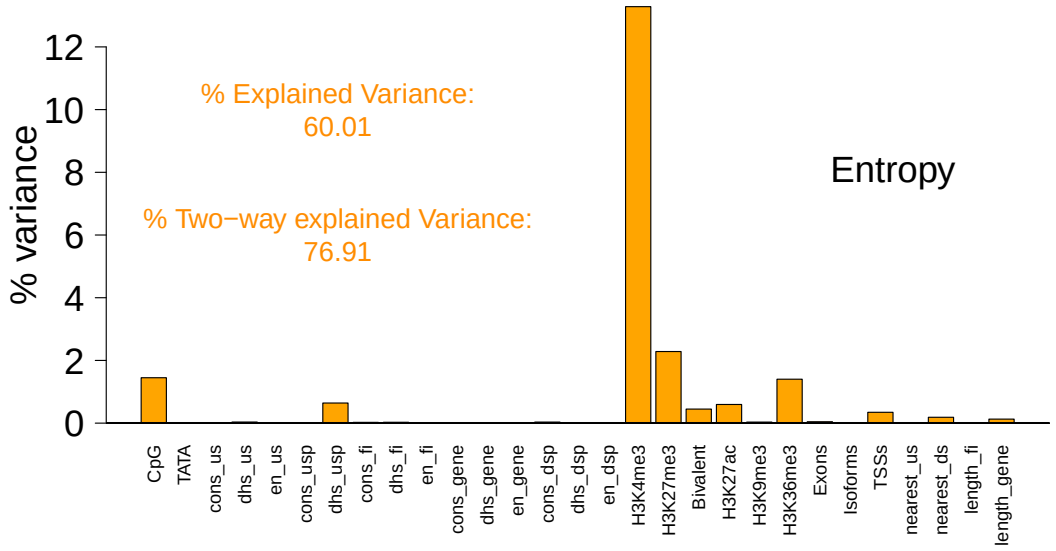


FIGURE 5.53: **Contribution of the variance of the entropy scores explained by each of the 29 variables, based on the "first" metric**. *Abbreviations:* cons = conservation, us = upstream, usp = upstream promoter, dhs = DNase I hypersensitive site, en = enhancer, fi = first intron, gene = gene body minus the first intron, ds = downstream.

Overall, entropy scores are highly dominated by H3K4me3 effects; which is not surprising since they associate with activation and thus breadth of expression. However, the variance explained by cis-regulation surrounding the gene and the effect of H3K27me3 has reduced here compared to complexity scores - these are factors we have defined as important for regulatory complexity, and not a distinguishing property for specificity of expression.

5.10 Disease analysis

Disease is associated with many forms of changes in regulatory mechanisms, including point mutations, splicing regulatory variants and loss of function mutations. Disease causing variants have been mapped to cis-regulatory sequences [Epstein, 2009], particularly in recent years in conjunction with the development of technologies allowing for the screening of genome-wide regulatory elements, such as histone modifications

TABLE 5.9: Primary cell types expressing *HBB* at tpm of at least 1 an their associated tpm values

Primary cell type	tpm
CD34 cells differentiated to erythrocyte lineage	217848.0
Peripheral Blood Mononuclear Cells	703.0
Neutrophils	676.0
chorionic membrane cells	139.0
amniotic membrane cells	122.0
Basophils	58.0
Hepatocyte	51.0
CD19 B Cells	34.0
Anulus Pulposus Cell	19.5
nasal epithelial cells	11.5
Tracheal Epithelial Cells	4.0
CD8 T Cells	3.0
Natural Killer Cells	1.0
CD4 T Cells	1.0

and DNAse I hypersensitivity. Mutations in cis regulatory sequence have been found to cause changes in the expression levels of a gene, predisposing individuals to disease through resulting phenotypic changes [VanderMeer and Ahituv, 2011]. This leads to the question as to whether highly regulated genes affected by a broad cis-regulatory landscape are more susceptible to mutational perturbations than less regulated genes affected by fewer regulatory sequence.

A classic example of disease causing mutations is the beta-golbin (*HBB*) gene; hundreds of mutations of the *HBB* gene are known to cause the disease beta thalassemia, including single nucleotide polymorphisms in the promoter of the gene, resulting in a reduction in beta-golbin production ([Ayub et al., 2010, Galanello and Origa, 2010]).

Of the 138 steady state primary cell types in the current analysis, *HBB* expression appears within 14 at 1 tpm or greater; of which approximately 99.2% of the total tpm-normalised expression levels are within erythrocytes, and the final 0.8% representing low level expression in other cell types (Table 5.9). It receives a complexity score of 0.42 and a normalised complexity score of 1.00, suggesting that this gene may exhibit a greater target for disease causing mutations.

Not just proximal elements, but also distal enhancers represent an important source of mutational targets in human disease ([Kleinjan and van Heyningen, 2005]). In particular, SHH enhancer mutations have been associated with preaxial polydactyly (PPD), which has the phenotype of malformations in limb development. SHH has an expression breadth of 13% normalised complexity score of 0.72 (top 20%) and a complexity score of 0.40 (top 40% for given expression breadth).

Attempting to link exact mutational processes and regulatory changes to disease phenotypes remains an important challenge. Changes are often pleiotropic (affecting one or more genes) and do not necessarily affect the genes they are closest to. Furthermore, diseases are often highly complex in nature, often caused by the accumulated effects of thousands of small changes. Prioritising genes for further study of a disease has received considerable attention and computational methodologies have been employed to with emphasis on looking at gene function, linkage disequilibrium of SNPs and pathway relationships, amongst other things.

Whilst the exact molecular mechanisms linking regulatory perturbation to disease phenotype is not well understood, using complexity and entropy scores as a proxy for the size of the mutational landscape acting on a gene may aid in the selection of potential target genes for closer examination. This section attempts to correlate groups of genes associated with categories of disease with measures of complexity in order to test the validity of this hypothesis.

5.10.1 Complexity and disease associated genes related to anatomical categories

In order to see if complex genes were associated with diseases specific to any anatomical category, genes associated with disease were downloaded from *malacards* [Rappaport et al., 2013] and odds ratios were modelled using logistic regression.

The results of the logistic regression models are given in Table 5.10 and the given odds ratios are plotted with their 95% confidence bounds in Figure 5.54. Visual inspection shows that the normalised complexity scores are clearly the most predictive

of disease status, with all diseases apart from 'smell/taste' significant, with 12 of the categories 'highly significant'. Complexity scores are highly significantly predictive of bone, nephrological and skin diseases. Entropy is associated with an odds ratio smaller than one for blood, cardiovascular gastrointestinal, immune and respiratory diseases, suggesting that expression specific genes may be associated with these disease categories.

TABLE 5.10: **Complexity odds ratios for anatomical categories from disease gene database**. Each point is the result from applying a binomial logistic repression model with logit link and presence and absence of specific disease association as the dependent variable. Results are report by taking the exponential of the resulting log odds ratio from the model.
Key: ( ) = not significant, (.) = p-value<0.1, (*) = p-value<0.05, (**) = p-value < 0.01 and (***) = p-value < 0.001. Odds ratios with confidence bounds are given in Figure 5.54.

	Number of genes	Complexity	Complexity (normalised)	Entropy (inverted)
blood	979	1.33 ( )	2.81 (***)	1.59 (***)
bone	1089	2.72 (***)	1.69 (***)	0.82 ( )
cardio	842	1.43 (*)	2.06 (***)	1.82 (***)
ear	410	1.36 ( )	1.27 ( )	1.31 ( )
endocrine	877	0.995 ( )	1.23 ( )	1.49 (**)
eye	1428	1.19 ( )	1.18 ( )	1.06 ( )
gastrointestinal	862	1.03 ( )	2.12 (***)	1.75 (***)
immune	933	1.4 (*)	3.93 (***)	1.62 (***)
liver	342	0.481 (**)	0.851 ( )	1.54 (*)
mental	738	0.871 ( )	1.03 ( )	1.23 ( )
muscle	394	1.62 ( )	0.899 ( )	0.842 ( )
nephrological	808	1.55 (*)	1.23 ( )	1.22 ( )
neural	2381	1.54 (***)	1.04 ( )	0.846 ( )
oral	336	1.53 ( )	1.34 ( )	1.51 (*)
reproductive	721	1.42 ( )	1.44 (*)	1.28 ( )
respiratory	558	1.18 ( )	3.22 (***)	1.9 (***)
skin	1090	1.74 (***)	2.64 (***)	1.31 (*)
smell/taste	24	2.49 ( )	1.19 ( )	0.914 ( )

Next the gene list associated with 'cancer' was downloaded from *malacards* and odds ratios were calculated using logistic regression. The category as a whole was found to be highly associated with complexity scores (odds ratios of 3.19 and 3.15 for complexity and normalised complexity scores, respectively, $p < 0.05$), and the same cancer category queried for specific cancer types also find significant relationships (Figure 5.55 and

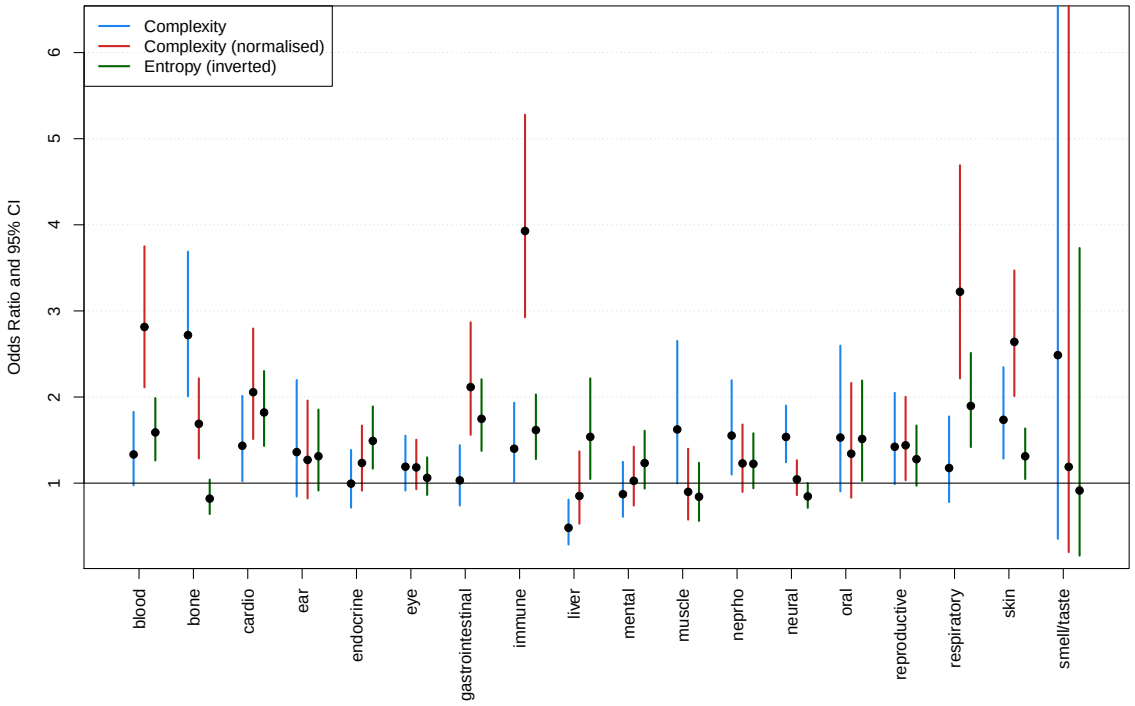


FIGURE 5.54: **Visual plot of odds ratios and 95% confidence intervals across anatomical categories**, for complexity (blue lines), normalised complexity (red) and entropy scores (green), based on logistic regression analysis. Upper limit for entropy scores lies above the top of the plot for eye, nephrological (neprho) and smell/taste (not significant). Data is based on models described in Table 5.10

Table 5.11). The highest category observed was sarcoma (odds ratios of 6.11 ($p < 0.001$) for complexity scores and 2.24 ($p < 0.05$) for normalised complexity scores. Interestingly, the odds ratio for the entropy was not significant for this cancer type (odds ratio 1.34, not significant), suggesting that complexity scores, which were designed to capture highly regulated genes, are able to capture potential cancer targets of genes beyond their specificity.

5.10.2 Methods

Lists of diseases associated with anatomical categories were downloaded from malacards [Rappaport et al., 2013]. Lists of disease associated genes and diseases were downloaded

TABLE 5.11: **Odds ratios for cancer categories**. Each point is the result from applying a binomial logistic repression model with logit link and presence and absence of specific disease association as the dependent variable. Results are report by taking the exponential of the resulting log odds ratio from the model.
Key: ( ) = not significant, (.) = p-value<0.1, (*) = p-value<0.05, (**) = p-value < 0.01 and (***) = p-value < 0.001. Odds ratios with confidence bounds are given in Figure 5.55.

Category	Number of genes	Complexity	Complexity (normalised)	Entropy
cancer	2204	1.99 (***)	1.96 (***)	0.784 (**)
lung cancer	95	3.19 (*)	3.15 (*)	0.892 ( )
carcinoma	414	2.13 (**)	1.64 (*)	0.713 ( )
sarcoma	155	6.11 (***)	2.24 (*)	1.34 ( )
breast cancer	185	2.52 (*)	1 ( )	1.88 ( )
prostate cancer	105	1.96 ( )	2.08 ( )	0.402 (**)
ovarian cancer	70	2.43 ( )	2.4 ( )	0.577 ( )

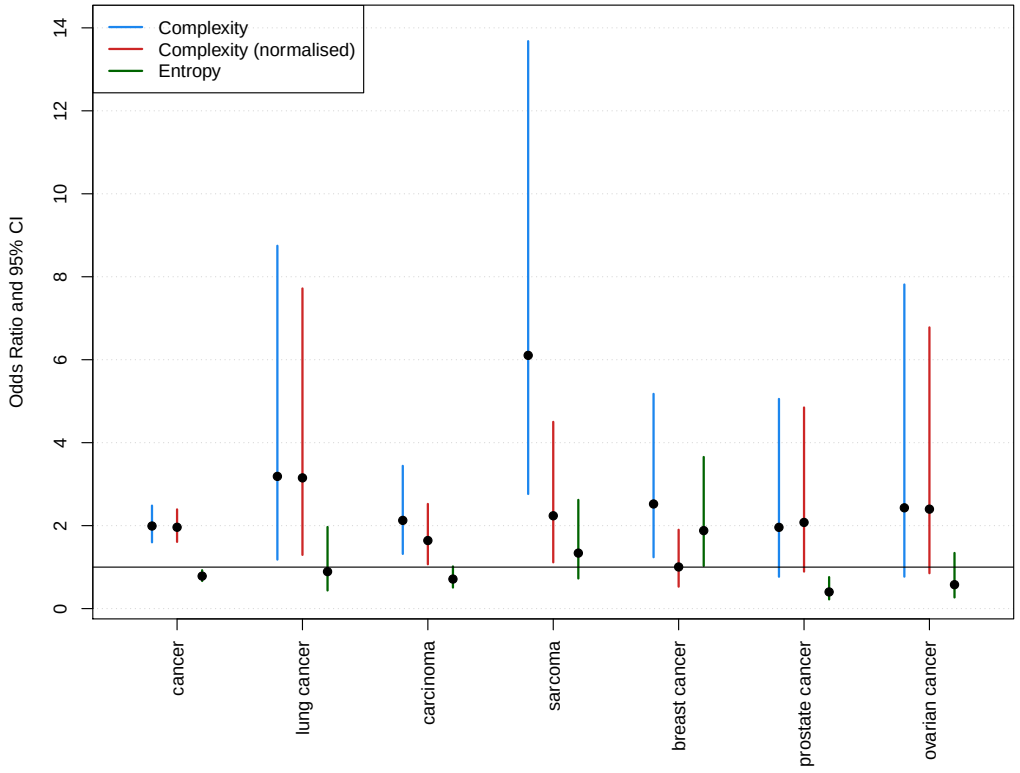


FIGURE 5.55: **Visual plot of odds ratios and 95% confidence intervals across cancer categories**, for complexity (blue lines), normalised complexity (red) and entropy scores (green), based on logistic regression analysis. Data is based on models described in Table 5.11

from genecards [Rebhan et al., 1998, Safran et al., 2010]. For each anatomical categories, diseases were cross-referenced with the disease genes list to see if there was an an associated disease gene. Grouping by anatomical category as opposed to specific diseases allows for greater statistical power in detecting statistical effects.

Cancer associated genes were found by using `grep` to find particular cancers from the genecards disease list. Logistic regression was applied to test whether scores of complexity were predictive of whether or not a given gene is associated with a particular type of disease. Since some of the categories are subsets of other categories (for example, neuronal diseases include the mental disease category), an association spotted within one category may not be a distinct result from observing an association in another category.

5.10.3 Complexity is associated with Alzheimer's genes

Alzheimer's is a neurological disease, most associated with ageing and which may affect over 66 million people by 2030 [Wortmann, 2012]. Whilst ageing plays the largest role, the disease has a variety of risk factors and is typically separated according to late and early onset of symptoms. Scientists have not pinned down a single causative gene for the disease, although classic studies of linkage analysis have revealed many genes with weak associated effects and more recent genome-wide association studies have identified vast numbers of candidate genetic variants in important risk genes, such as *CLU* and *PICALM* [Harold et al., 2013].

It is thought that genetic factors could explain as much as 70% of the risk of developing Alzheimer's [Bettens et al., 2013]. The *Alzgene* database [Bertram et al., 2007] is a table of genes based on a meta analysis of genetic studies linking genes to the risk of developing Alzheimer's through the detection of single nucleotide polymorphisms in or around potential risk genes. The most commonly cited risk gene for Alzheimer's is *APOE* [Corder et al., 1993] which *Alzgene* lists as its number one gene, whereby having the *e2/3/4* allele increases the risk of developing Alzheimer's by an odds of 3.685 (data displayed as part of Table 5.12), compared to the next most significant gene, *BIN1*,

with an odds ratio of 1.166. Thus, apart from *APOE*, these are genes with small but significant effect sizes, and so provide an interesting set of targets for comparison with complexity scores.

Table 5.12 gives the top Alzheimer's risk genes from the *Alzgene* database together with commonly associated polymorphisms and relative risk of Alzheimer's development. The *APOE* gene has a complexity score in the top 2% of scores (0.98 quantile) and normalised complexity score in the top 33% (0.66 quantile). Out of the 9 genes in the table, 8 of them have a complexity score in the top 25% of scores, 5 of them have a complexity score in the top 10% of scores, and 3 in the top 5% of scores. In order to formally generalise the hypothesis that Alzheimer's risk genes are more complex, logistic regression was applied to each score, based on 222 genes from the *Alzgene* database (Table 5.13 and displayed graphically in Figure 5.56). The odds of a maximally complex gene being associated with Alzheimer's is 3.18 greater than a non-complex gene for the complexity scores and 2.91 greater for the normalised complexity scores (both highly significant, $p < 0.001$). The consistent relationship between the two scores may be explained by the lack of significant association between entropy and Alzheimer's risk genes, since normalised complexity is an attempt to correct for differences in expression breadth in different genes. Taken together, these results suggest that genes previously associated with Alzheimer's are highly complex and thus complexity scores could potentially be applied as a method for targeting other potential candidate genes for further analysis of their potential link with Alzheimer's.

5.10.4 *HGDM*, *COSMIC* and *GWAS catalogue*

The *HGMD database* is a catalogue of gene mutations linked to disease [Stenson et al., 2003]. All disease references are manually entered from published studies linking mutations in the germline with human associated diseases. The data includes mutations that have occurred in regulatory regions as well as coding regions and relating to splicing events, amongst other things, but does not include somatic mutations or those relating to mitochondria.

TABLE 5.12: **Top Alzheimer's risk associated genes with estimated size of effects and related p-values, with associated complexity scores**. Data is taken from on *http : //www.alzgene/org*. Complexity, normalised complexity and entropy scores are given for each gene, together with the quantile of each score in relation to the full distribution.

Gene	Polymorphism	OR (95% CI)	P-value	Complexity (quantile)	Normalised Complexity (quantile)	Entropy (quantile)
APOE	APOE_e2/3/4	3.685 (3.30-4.12)	<1E-50	0.84 (0.98)	0.64 (0.66)	0.92 (0.37)
BIN1	rs744373	1.166 (1.13-1.20)	1.59E-26	0.62 (0.75)	0.53 (0.52)	0.96 (0.45)
CLU	rs11136000	0.879 (0.86-0.90)	3.37E-23	0.73 (0.87)	0.53 (0.52)	0.96 (0.46)
ABCA7	rs3764650	1.229 (1.18-1.28)	8.17E-22	0.78 (0.93)	0.65 (0.67)	0.97 (0.49)
CR1	rs3818361	1.174 (1.14-1.21)	4.72E-21	0.33 (0.23)	0.96 (0.96)	0.6 (0.18)
PICALM	rs3851179	0.879 (0.86-0.9)	2.85E-20	0.74 (0.89)	0.41 (0.35)	1 (0.9)
MS4A6A	rs610932	0.904 (0.88-0.93)	1.81E-11	0.37 (0.31)	0.86 (0.92)	0.63 (0.2)
CD33	rs3865444	0.893 (0.86-0.93)	2.04E-10	0.39 (0.36)	0.97 (0.97)	0.65 (0.21)
CD2AP	rs9349407	1.117 (1.08-1.16)	2.75E-09	0.54 (0.63)	0.52 (0.52)	1 (0.77)

TABLE 5.13: **Complexity odds ratios for Alzheimer's**. Each point is the result from applying a binomial logistic repression model with logit link and presence and absence of specific disease association as the dependent variable. Results are report by taking the exponential of the resulting log odds ratio from the model.
Key: ( ) = not significant, (.) = p-value<0.1, (*) = p-value<0.05, (**) = p-value < 0.01 and (***) = p-value < 0.001. Odds ratios with confidence bounds are given in Figure 5.56.

Category	Number of genes	Complexity	Complexity (normalised)	Entropy
Alzheimer's	222	3.18 (***)	2.91 (***)	0.651 ( )

The *GWAS catalogue* [Welter et al., 2014], similar to the intentions of the *HGMD database*, catalogues SNPS from the literature based on genome wide association studies (GWAS). Gene associations are reported by the authors in the relevant literature; these are used to associate relevant complexity scores.

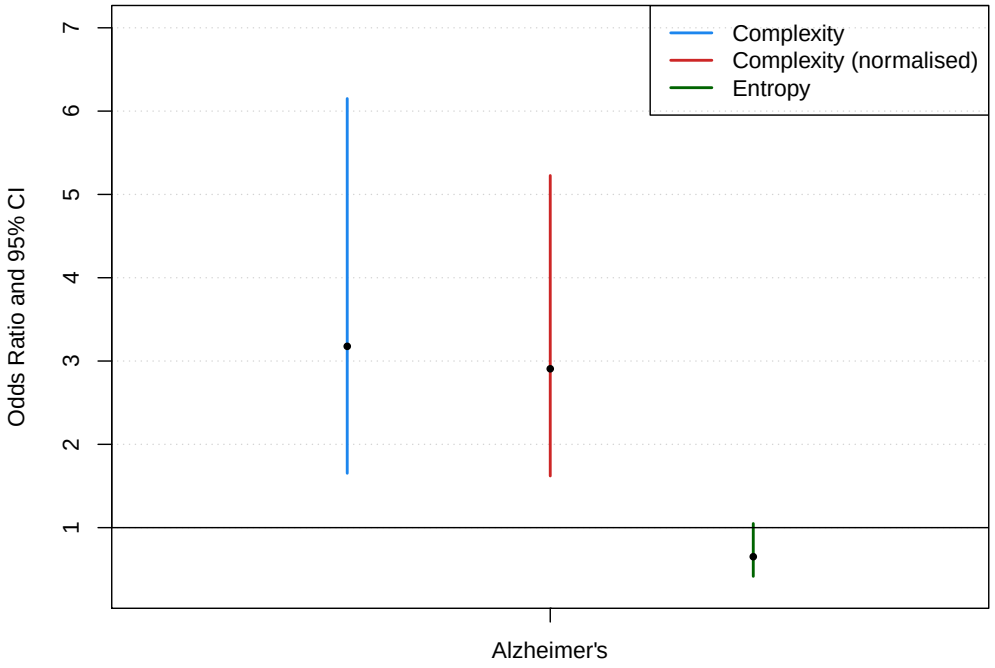


FIGURE 5.56: **Odds ratios for genes associated with Alzheimer's**. Visual plot of odds ratios and 95% confidence intervals across cancer categories, for complexity (blue lines), normalised complexity (red) and entropy scores (green), based on logistic regression analysis. Based on odds ratios from models described in Table 5.13

The catalogue of somatic mutations in cancer (*COSMIC*) from the Sanger Institute ([Forbes et al., 2009]) lists genes associated with somatic mutations which are implicated in cancer based on published literature. At the time of the analysis the list contained 483 genes, including information on histology and tissues.

As a complementary analysis, genes from the three databases described above were downloaded, linked with their appropriate complexity scores and logistic regression applied to test for relative risk of disease associated with highly complex genes. Figure 5.57 shows the overlap in disease associated genes between the three dataset. All three datasets had 128 genes in common. Furthermore, the *GWAS catalogue* and *HGMD* database had 2084 genes in common, probably as a result of studies from the *GWAS catalogue* also being included in the *HGMD database*. Figure 5.58 and Table 5.14

show the separate odds ratios for each of the three databases (with large error bounds for the cancer somatic as a result of a much smaller sample size).

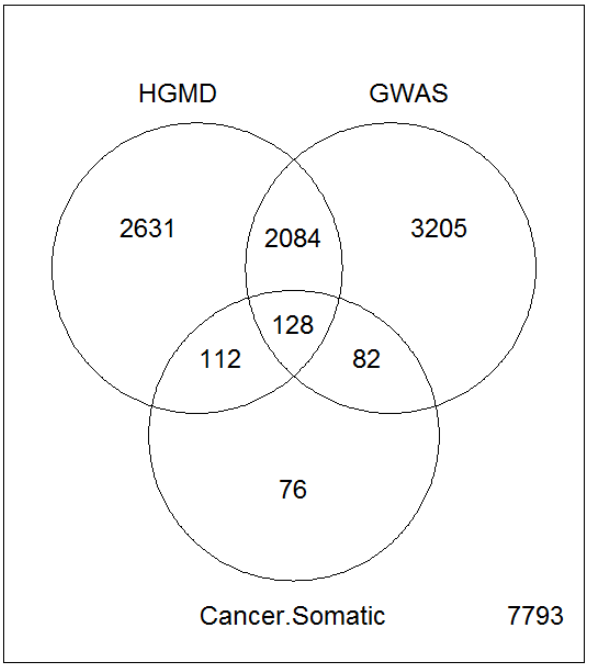


FIGURE 5.57: **Venn diagram illustrating the overlap in genes implicated in disease for the HGMD database, GWAS catalog reported genes and cancer genes with somatic mutations**.

Genes from the GWAS catalogue were associated with high complexity (odds ratio 2.05 for complexity and 2.75 for normalised complexity scores, both highly significant ($p < 0.001$), Table 5.14) and cell type restriction (odds 1.5 for inverted entropy, highly significant ($p < 0.001$)). Genes with high expression breadth were the most highly associated with cancer somatic mutations (odds 0.26 for entropy, highly significant ($p < 0.001$)). Cancer somatic associated genes highly complex (odds 2.67, highly significant ($p < 0.001$)), but not significant according to normalised complexity scores. Taken together, it appears that genes in the *GWAS catalogue* are strongly associated

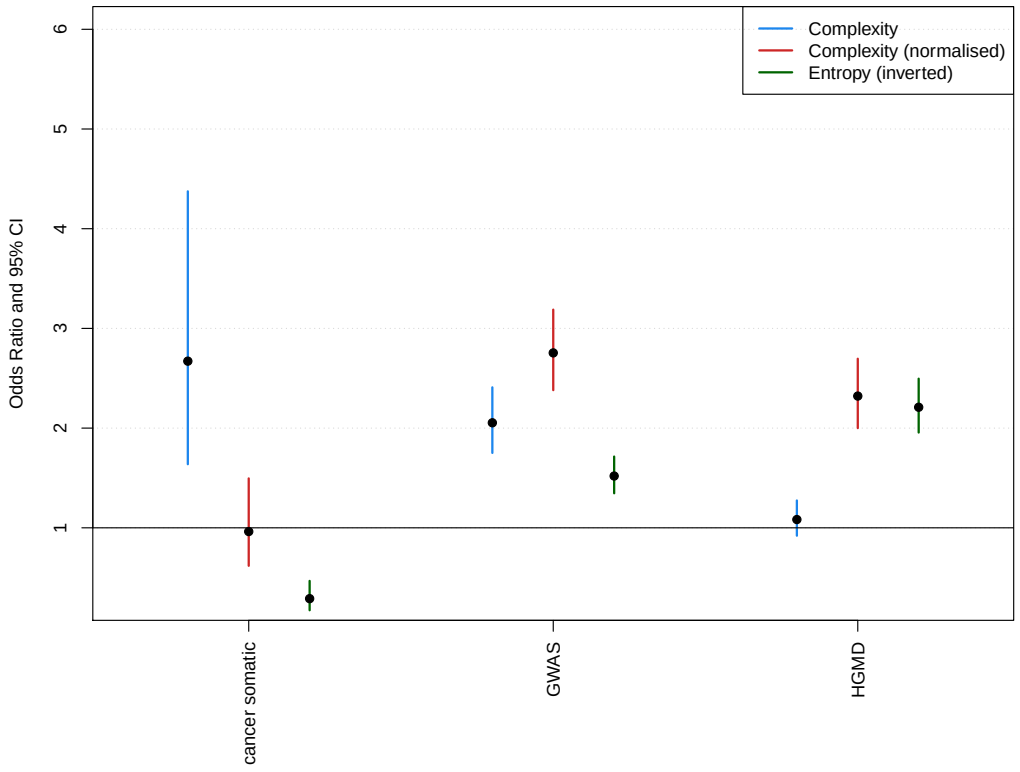


FIGURE 5.58: **Odds ratios for genes associated with cancer somatic mutations, GWAS hits and HGMD genes**. Visual plot of odds ratios and 95% confidence intervals across cancer categories, for complexity (blue lines), normalised complexity (red) and entropy scores (green), based on logistic regression analysis given in Table 5.14.

with both complexity scores and entropy, and genes in the Cancer Somatic and *HGMD* databases were associated with one of the complexity scores and entropy.

TABLE 5.14: **Complexity odds ratios for cancer somatic, GWAS and HGMD associated genes**. Each point is the result from applying a binomial logistic repression model with logit link and presence and absence of specific disease association as the dependent variable. Results are report by taking the exponential of the resulting log odds ratio from the model.
Key: ( ) = not significant, (.) = p-value<0.1, (*) = p-value<0.05, (**) = p-value < 0.01 and (***) = p-value < 0.001. Odds ratios with confidence bounds are given in Figure 5.58.

	Number of genes	Complexity	Complexity (normalised)	Entropy (inverted)
Cancer Somatic	398	2.67 (***)	0.963 ( )	0.29 (***)
GWAS	5499	2.05 (***)	2.75 (***)	1.52 (***)
HGMD	4955	1.08 ( )	2.32 (***)	2.21 (***)

Figure 5.59 shows the relationship between scores and the number of associated SNPs from the GWAS catalogue associated with each reported gene (including all SNP locations). SNP number explains 1.5% of the variation in normalised complexity scores, the strongest relationship of the three scores.

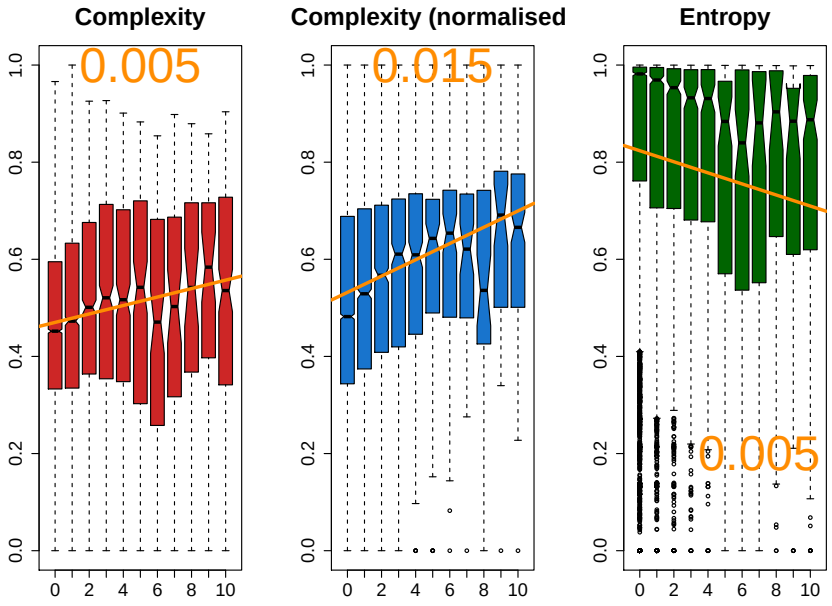


FIGURE 5.59: **Number of associated GWAS SNPs per gene**. SNPs downloaded from the GWAS catalogue and number of SNPs counted per gene, based on author reported genes. Orange lines represent best fit straight line from `lm` function, orange numbers represent associated R^2.

SNPs from the *GWAS catalogue* were broken down by SNP location, based on intron, intergenic, UTR or near gene region. A similar analysis was applied, with the results given in Table 5.60 and Figure 5.60. Disease genes associated with intergenic and intronic SNPs were significantly complexity, the highest being Intergenic complexity (odds 3.13, highly significant, Table 5.60). These SNPs may for example lie in distal enhancer regions which act on the gene. Indeed, disease SNPs have been shown to be over-represented in enhancer regions [Andersson et al., 2014b].

TABLE 5.15: **Odds ratios for reported genes associated with SNPs, broken down by SNP location** SNPs taken from GWAS catalogue with p-value<1.0e-08. Odds ratios based on binomial logistic repression models with logit link and presence and absence of specific disease association as the dependent variable. Results are reported by taking the exponential of the resulting log odds ratio from the model. Inverted entropy is 1 - entropy, where entropy scores are between 0 and 1
Key: ( ) = not significant, (.) = p-value<0.1, (*) = p-value<0.05, (**) = p-value < 0.01 and (***) = p-value < 0.001. Odds ratios with confidence bounds are given in Figure 5.60.

	Number of genes	Complexity	Complexity (normalised)	Entropy (Inverted)
Near gene	631	0.885 ( )	2.09 (***)	1.95 (***)
Intergenic	2718	3.13 (***)	2.33 (***)	1.64 (***)
UTR	269	2.02 (**)	1.67 ( )	1.32 ( )
Intron	3834	2.21 (***)	2.08 (***)	1.24 (**)

The SNPs from the *GWAS catalogue* are then broken down further according to specific diseases. Odds ratios for complexity scores based on diseases with 40 or more genes with associated SNPs are give in Table 5.16 and displayed visually in Figure 5.61.

In order to interrogate the *HGMD database* in more detail, *HGMD* genes were split according to presence of mutation in the protein coding region of the gene, single nucleotide polymorphisms in the regulatory region of the gene and frameshift or truncating variant (FTV), according to the definitions given in the *HGMD* database. It was hypothesised that polymorphisms might be associated with genes which are more complex in their expression, due to a larger mutational target. Table 5.17 and Figure 5.62 show the odds ratios for the three categories of *HGMD* genes. Whilst raw complexity scores were not significant for any category, the normalised complexity scores were significant for all three, with the most significant being polymorphism (odds=4.18, highly significant ($p < 0.001$). Entropy was also highly significant (odds ratio 3.62, $p < 0.001$),

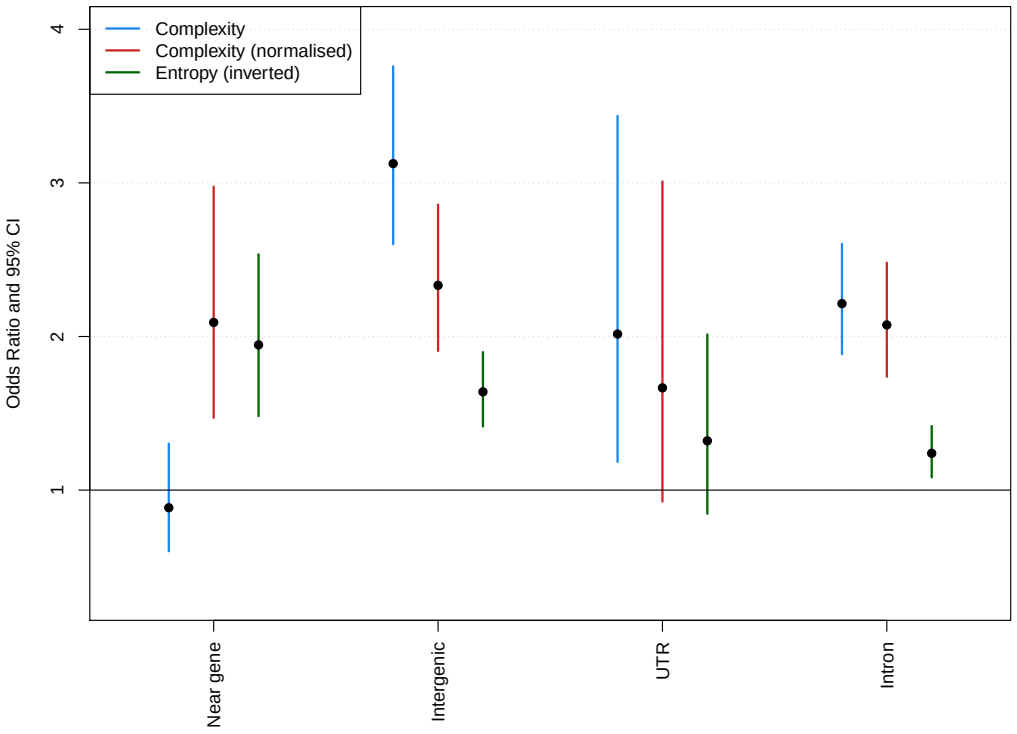


FIGURE 5.60: **Odds ratios for reported genes associated with SNPs, broken down by SNP location and plotted with 95% confidence intervals** Odds based on logistic regression models from Table 5.15. Scores are: complexity (blue lines), normalised complexity (red lines) and inverted entropy scores (green lines), defined as 1 - entropy, where entropy scores are normalised between 0 and 1.

suggesting that genes with these features are more cell-type restricted (odds are significantly less than 1). Since normalisation up-weights cell-type restricted scores, it is not surprising that the normalised complexity scores are more significant. Taken together, SNPs located within regulatory regions are associated with genes which are highly complex according to both complexity scores, whilst genes associated with polymorphisms are highly complex according to normalised complexity scores, as well as highly tissue specific.

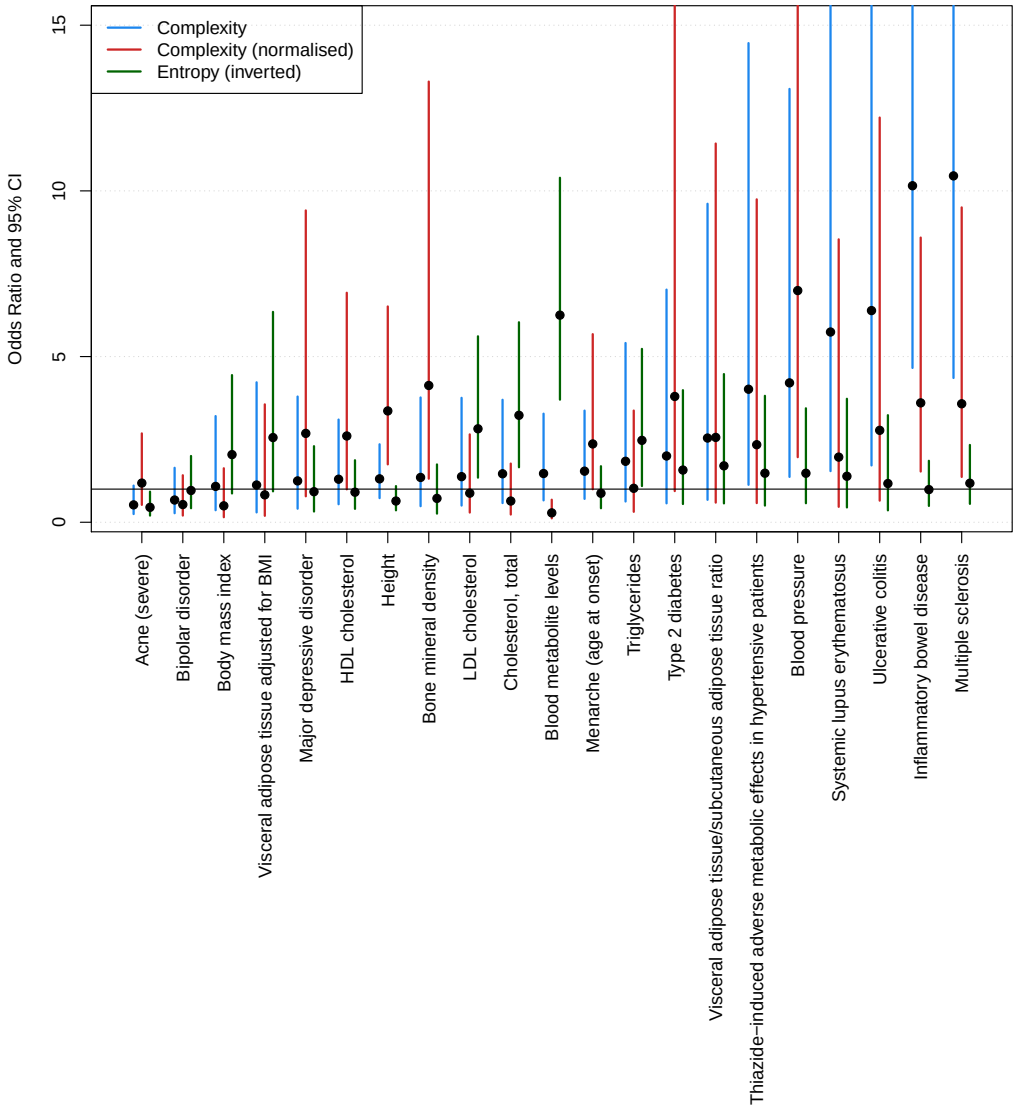


FIGURE 5.61: **Complexity Odds ratios for genes split according to recorded associated SNP location from GWAS category**. Genes are those reported by authors. Visual plot of odds ratios and 95% confidence intervals across cancer categories, for complexity (blue lines), normalised complexity (red) and reversed entropy scores (green), based on logistic regression analysis. Based on results from models described in Table 5.16

TABLE 5.16: **Complexity odds ratios for genes split according to recorded associated SNP location from GWAS category**. Genes are those reported by authors. Each point is the result from applying a binomial logistic repression model with logit link and presence and absence of specific disease association as the dependent variable. Results are report by taking the exponential of the resulting log odds ratio from the model.
Reversed entropy is 1 - entropy, where entropy scores are between 0 and 1
Key: ( ) = not significant, (.) = p-value<0.1, (*) = p-value<0.05, (**) = p-value < 0.01 and (***) = p-value < 0.001. Odds ratios with confidence bounds are given in Figure 5.61.

	Gene count	Complexity	Complexity (normalised)	Entropy (inverted)
Multiple sclerosis	102	10.5 (***)	3.58 (**)	1.18 ( )
Inflammatory bowel disease	129	10.2 (***)	3.6 (**)	0.986 ( )
Ulcerative colitis	44	6.39 (**)	2.77 ( )	1.17 ( )
Systemic lupus erythematosus	44	5.74 (**)	1.97 ( )	1.39 ( )
Blood pressure	60	4.21 (*)	6.99 (**)	1.48 ( )
Thiazide-induced adverse metabolic effects in hypertensive patients	47	4.02 (*)	2.34 ( )	1.48 ( )
Visceral adipose tissue/ subcutaneous adipose tissue ratio	43	2.54 ( )	2.56 ( )	1.7 ( )
Type 2 diabetes	48	2 ( )	3.8 ( )	1.58 ( )
Triglycerides	65	1.84 ( )	1.03 ( )	2.47 (*)
Menarche (age at onset)	124	1.54 ( )	2.36 ( )	0.875 ( )
Blood metabolite levels	118	1.47 ( )	0.284 (**)	6.25 (***)
Cholesterol, total	88	1.46 ( )	0.641 ( )	3.23 (***)
LDL cholesterol	75	1.38 ( )	0.878 ( )	2.82 (**)
Bone mineral density	72	1.35 ( )	4.13 (*)	0.72 ( )
Height	222	1.31 ( )	3.36 (***)	0.641 ( )
HDL cholesterol	100	1.3 ( )	2.6 ( )	0.907 ( )
Major depressive disorder	61	1.25 ( )	2.68 ( )	0.923 ( )
Visceral adipose tissue adjusted for BMI	43	1.12 ( )	0.827 ( )	2.56 ( )
Body mass index	64	1.08 ( )	0.496 ( )	2.04 ( )
Bipolar disorder	95	0.672 ( )	0.534 ( )	0.96 ( )
Acne (severe)	138	0.525 ( )	1.18 ( )	0.45 (*)

5.10.5 Method

Genes implicated in cancer with somatic mutations were downloaded from

http://cancer.sanger.ac.uk/cosmic/census and selecting the list of 'Somatic mutations'.

This list contained 483 genes at the time of the analysis.

GWAS reported genes were obtained by downloading the GWAS catalog from

TABLE 5.17: **Complexity odds ratios for HGMD genes according to presence of mutation, polymorphism and frameshift or truncating variant (FTV)**. Each point is the result from applying a binomial logistic repression model with logit link and presence and absence of specific disease association as the dependent variable. Results are report by taking the exponential of the resulting log odds ratio from the model. Key: ( ) = not significant, (.) = p-value<0.1, (*) = p-value<0.05, (**) = p-value < 0.01 and (***) = p-value < 0.001. Odds ratios with confidence bounds are given in Figure 5.62.

	Number of genes	Complexity	Complexity (normalised)	Entropy (reversed)
Mutation	4333	1.12 ( )	1.59 (***)	1.62 (***)
Polymorphism	2491	0.943 ( )	4.18 (***)	3.62 (***)
FTV	478	0.934 ( )	1.84 (*)	2.13 (***)

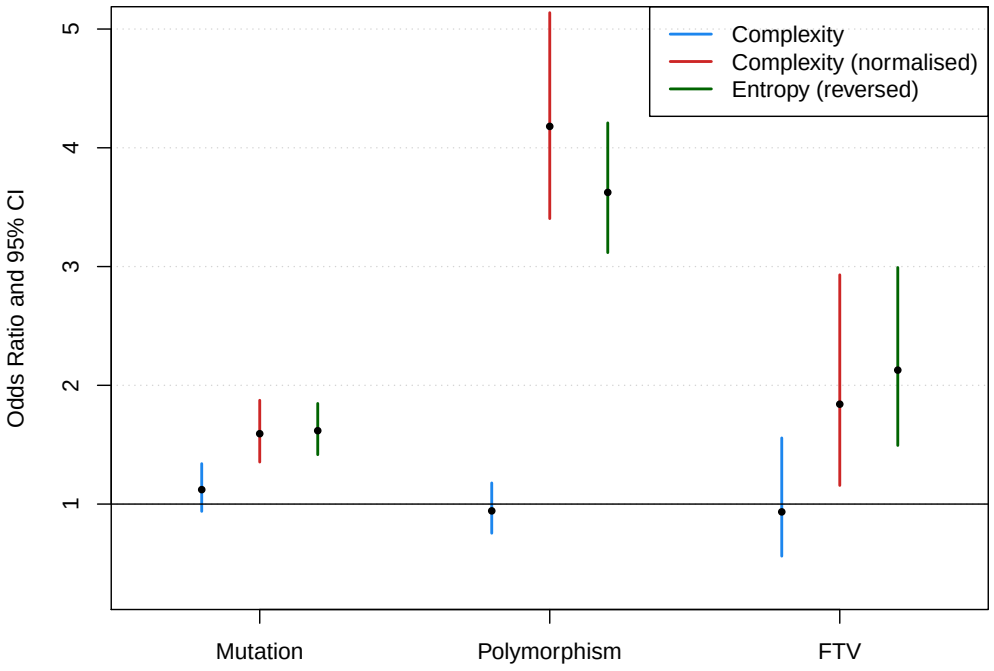


FIGURE 5.62: **Visual plot of odds ratios for genes split according to presence of mutation in the protein coding region of the gene (mutation), single nucleotide polymorphisms in the regulatory region of the gene (polymorphism) and frameshift or truncating variant (FTV)**, according to the definitions given in the HGMD database, including 95% confidence intervals, for complexity (blue lines), normalised complexity (red) and entropy scores (green), based on logistic regression analysis from the models described in Table 5.17.

http://www.genome.gov/gwastudies/ and saving the list of genes from the column labelled 'Reported genes'. At the time of the analysis, a total of 5499 gene names were included in the final calculations for odds ratios.

HGMD disease associated genes were downloaded from http://www.hgmd.cf.ac.uk, including information about coding mutations, polymorphisms and rare variants. At the time of the analysis, a total of 4955 genes were used in the final calculations of odds ratios.

For each gene name in refSeq, a '1' is allocated if that gene is in the reported genes list, and a '0' otherwise. The 'glm' function in R was used to determine significance of the log odds that a complex gene is a disease gene. Since complexity scores are between 0 and 1, odds ratios may be interpreted as the relative risk of a gene of complexity equal to 1 being a disease gene compared to a gene of complexity equal to 0. Confidence intervals were calculated from the same 'glm' models, which were used to create visual plots of the odds for each score.

5.10.6 Haploinsufficient genes

Haploinsufficient genes are those which cause disease as a result of monoallelic loss of function mutations, whereby the single functioning allele of the gene is not sufficient to maintain normal phenotype. Based on the list of 151 genes (see methods), logistic regression was computed to test for association with complexity scores, with the results displayed in Table 5.18 and Figure 5.63. Again, the odds of a haploinsufficient gene being a highly complex gene are significant for both complexity scores (odds ratios 2.52 ($p < 0.05$) and 3.09 ($p < 0.01$) for complexity and normalised complexity scores, respectively, and not significant according to entropy scores (odds ratio 0.579). Whilst the effect is not highly significant due to a low number of tested genes, this suggests a role in which genes exhibiting complex regulatory programmes may be more susceptible to haploinsufficiency.

In all, it has been shown that genes associated with disease are statistically likely to be complex genes. We discuss this section further in the next chapter.

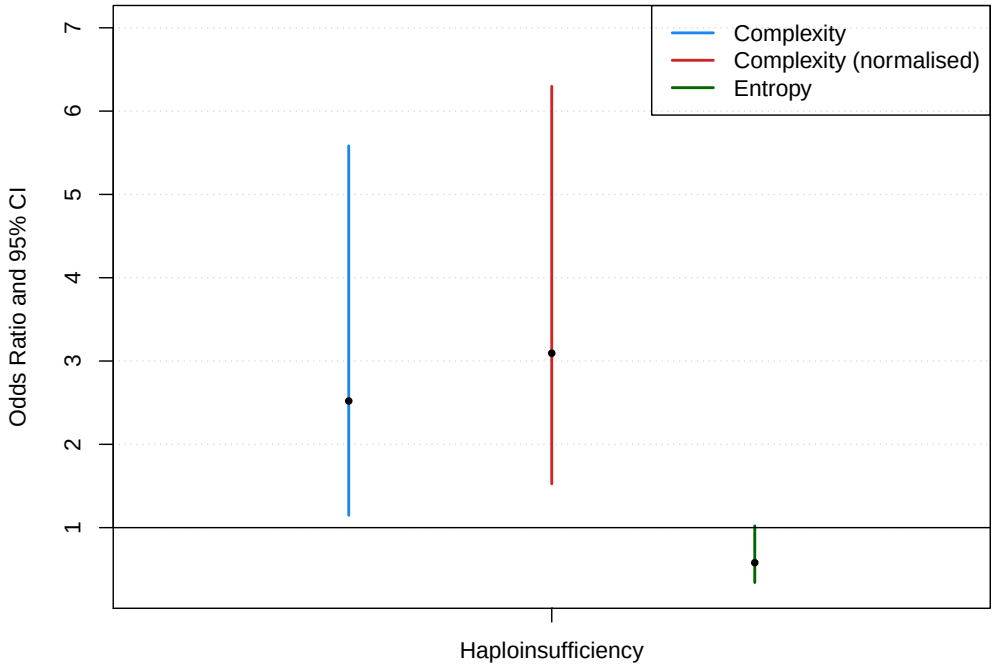


FIGURE 5.63: **Odds ratios for haploinsufficient genes**, displayed visually with 95% confidence intervals, for complexity (blue lines), normalised complexity (red) and entropy scores (green), based on logistic regression analysis from the models described in Table 5.18.

5.10.7 Methods

Analysis based on list of 151 haploinsuffient genes provided by David Fitzpatrick at the IGMM HGU, Edinburgh

TABLE 5.18: **Complexity odds ratios for haploinsufficient genes**. Each point is the result from applying a binomial logistic repression model with logit link and presence and absence of specific disease association as the dependent variable. Results are report by taking the exponential of the resulting log odds ratio from the model. Key: ( ) = not significant, (.) = p-value<0.1, (*) = p-value<0.05, (**) = p-value < 0.01 and (***) = p-value < 0.001. Odds ratios with confidence bounds are given in Figure 5.63.

Category	Number of genes	Complexity	Complexity (normalised)	Entropy
Haploinsufficency	151	2.52 (*)	3.09 (**)	0.579 ( )

Chapter 6

Discussion

The premise of this project was to attempt to understand how the collective effects and interactions of the regulatory architecture affecting a gene relates to its transcriptional output. Transcription is controlled by a number of events, such as transcription factors binding to sequences in and around the gene, made accessible by regions of open chromatin and through interactions with activators and cofactors. Currently, whilst it is well known that transcriptional initiation is a highly regulated process, determining exactly when, where and how a gene is expressed in different cell types, there is no clear understanding of the exact mechanisms involved in this regulation for each gene.

Measuring properties of the effects of regulatory input on transcriptional levels across multiple cell types, tissues or time points allows for the identification of genes potentially controlled by more or less complex regulatory programmes. Measuring the results of these regulatory programmes may in turn be connected back to individual regulatory mechanisms, raising important questions on cause and effect based on circular hypotheses: what effect do different regulatory mechanisms have on gene expression patterns, which are more important in explaining the patterns we observe and how can we effectively measure these patterns based on the hypothesised causative effects of regulatory programmes? Thus, a greater understanding of predicted effects of measuring the information observed in expression can result in more accurate measures, which

may then be fed-back to the biological understanding of gene regulation and how it may affect transcription.

Whilst the concept of measuring gene expression profiles is not new - it has been frequently alluded to in large scale gene expression projects where data exists across multiple samples - the metrics that are frequently applied are not necessarily well equipped for determining regulatory complexity. The most commonly used metric, the Shannon entropy, is effective at capturing the axis between cell-specific and ubiquitous expression and separating out either end of this scale into differences in regulation, for example presence and absence of CpG islands associated with genes [Elango and Soojin, 2011, Schug et al., 2005]. However, in the vast majority of cases the measure does not capture regulatory information beyond this, neither does it relate the important relationships between the samples, for example ontological relatedness. In large data resources such as FANTOM5, such relatedness could readily be mined from the data itself. Thus, a more comprehensive approach to measuring profiles was warranted and presented a yet unexploited opportunity to develop novel metrics which could feed back into the field of transcriptional complexity in multiple contexts.

This project has attempted to do just this, by capturing potential properties of a complex regulatory programme which may be observed in a single transcriptional profile:

- Specificity of expression across samples
- Significant changes in expression observed between samples
- Combinations of on or off states observed across samples
- The above three properties related to the biological (ontological) structure of the samples under study

Measuring biological regulation in this way, as opposed to simply observing sequence information or specific modes of regulation (e.g. transcription factor binding around a gene) has a number of distinct advantages:

- Transcriptional output is capturing the effects of ALL regulatory processes acting on that transcriptional unit and therefore captures regulatory information about mechanisms which are not fully understood. This has the caveat that poor or incomplete measurements of transcriptional levels may miss the effects of some regulatory events.

- Such a metric makes no assumptions about specific regulatory events; whilst it is based on the model of regulatory elements acting together to determine output, it does not say what exactly this model is. For example, although one might intuitively expect genes enriched in cis-regulatory elements to exhibit more changes in expression across a profile, this cause and effect relationship is not explicitly assumed and can therefore be tested.

- Adding in information about sample structure helps to alleviate problems of over-representing certain cell types in studies and helps to pin-point which transcriptional units behave similarly in similar cell types vs those which exhibit highly diverse patterns whilst maintaining similar specificities.

With this in mind, Chapter 2 began by discussing common information theoretic measures for scoring gene expression profiles, their usefulness in the context of collecting regulatory information and their limitations. It then described a novel approach not yet explored in the context of gene expression; that is to convert the problem into a two dimension context, by observing differences occurring between samples, allowing for the inclusion of sample structure information and the exploration of possible metrics in a graph theoretic framework. These metrics were then normalised according to the difference between what can be potentially the upper limit of expression complexity for a given breadth of expression, according to the metrics, and what is actually observed.

Chapter 3 described the FANTOM5 data and why it provides a near perfect platform for such measures to be calculated and studied biologically:

- It quantitatively estimates steady-state transcript levels at a single TSS resolution over the whole genome.

- It includes a wide range of cell types, tissues and time-courses. In particular, a wide range of primary cells sampled from healthy adults worked well in estimating metrics of how a transcriptional elements behaves 'normally' (albeit as a non-transient snap-shot) in a population of cells with minimal ontological heterogeneity. Primary cells allow for the capturing of potentially different regulatory information occurring between cells which may otherwise be grouped together within the same tissue.

- It includes extensive technical and biological replication; allowing for more accurate estimations of differential expression.

This chapter described how the metric developed here was applied to the data and the potential caveats in its application.

In Chapter 4 a subset of complexity scores were interrogated in detail; that is those based on the expression profiles of genes applied across a large set of primary cells, the intention being to capture some of the regulatory complexity of human cellular differentiation. It was found that complex genes were depleted of CpG islands in their core promoter, independent of expression specificity, and that complex genes appear to weakly relate to physical genomic measures of size constraints, numbers of isoforms, numbers of exons and presence of a TATA box sequences in the core promoter. Furthermore, complex genes were significantly enriched in indicators of cis-regulation, namely GERP conservation, DNAse I hypersensitive sites based in Chip-seq data and predicted enhancers based on bidirectional marks in CAGE. Whilst these were found to explain significant variation in complexity scores, supporting the prior hypothesis that genes exhibited more complex regulatory patterns would be enriched for cis-regulatory sequences relative to genes with less complex regulatory. However, searching for the 'missing' explained variance led to the consideration of associations with epigenetic marks in the promoter region of genes. This, surprisingly, was seen to explain a large amount of variation in scores, with putative bivalent promoters across multiple tissues seen to be a highly significant predictor of regulatory complexity. In evolutionary terms, newer genes were also found to be potentially more complex in their expression. Finally,

genes associated with disease collected from a wide variety of sources were found to be more complex than non-disease genes.

In the remainder of the current chapter, the results from Chapter 5 are discussed in more detail, linking them to a biological context. The caveats of methodologies are then explored and potential further work discussed.

6.1 Do complexity scores capture expression patterns in the way we originally hypothesised?

Whilst the Shannon entropy achieves a maximum in the case of ubiquitous expression and a minimum in the case of expression in a single sample, expression complexity was conceived based on the idea that the regulatory programmes causing these two types of expression profiles were conceptually simple. In the case of ubiquitous uniform expression, associated with housekeeping genes, few regulatory elements may be required to achieve this output - switch on the gene in every cell type at a fixed level, remain at that level. Regulatory elements such as enhancers may be present in order to boost and maintain transcription at this level, but in principal the same regulatory elements could be used in all cell types.

The complexity of ubiquitous uniform genes was expected to be zero, since no differential expression or combination of on and off switching between cell types is observed. The result is an empty graph where there are no connections between cell types, which therefore achieves a connectivity score of zero. This is achieved; genes known to be associated with housekeeping tasks integral to cellular function are observed in a dense low complexity, high expression breadth region on the complexity vs breadth scatter plots. In the implemented complexity measure, these genes remain low-complex, independent of normalisation strategies, although normalization does reduce the relative ranking of housekeeping genes due to the up-weighting of non-ubiquitous genes.

Also hypothesised was a separation between ubiquitous-uniform and ubiquitous-non uniform genes. This is barely achieved on the entropy scale, with the 35% of all

genes which were ubiquitous falling within the 0.97-1 range (out of a scale from 0 to 1). The profiles of genes such as, *FOS*, which appear to exhibit many different levels of expression whilst maintaining at least some expression in every cell type appear visually distinct from genes such as *ACTB*, which appears to have a much more uniform profiles. One might suggest there is potential for more regulatory mechanisms involve in the former case, controlling when and what level expression should be in each cell. Complexity scores agree with this idea, as FOS achieves a distinctly higher score than *ACTB*, due to detected differential expression between cell types.

The opposite end of the entropy scale, highly specific expression, may also be conceptually simple and one might hypothesis a scenario where a gene is repressed in every cell type (perhaps through methylation laid down in development), apart from one where a single promoter and/or enhancer is required to achieve transcription. Cell restricted genes are more difficult to hypothesise in terms of their expected complexity than ubiquitous uniform genes, due to a combination of factors leading to a wide variation in regulatory mechanisms. It is often difficult to determine if a gene is truly restricted to a single cell type, or whether it is co-expressed in another, non-sequenced cell type, or whether it is simply off and poised to be triggered by a biological or environmental stimulus, potentially creating a highly complex pattern of expression. Genes with very low expression in cell types may not be detected as expressed due to filtering of CAGE libraries or lack of sensitivity in detecting expression and/or differential expression. Thus there is a distinct lack of information about the potential regulatory landscape of genes where very few CAGE tags are observed in a restricted subset of samples.

For this reason, both complexity scores in their raw form and complexity scores normalised by maximal complexity tag redistributions were presented for analysis. Normalisation clearly up-weights genes which are restricted in the number of primary cell types they are observed in, on the basis that the number of observed levels in expression is potentially fewer than those with high expression breadths, a bias which should be corrected for. This accounts for the lack of information observed when the gene is in an 'off' state in cell types. Using both sets of scores helps to attempt to avoid falling into a circular cycle of corrected for the disparity of what is expected in terms of complex

expression output and what is actually observed in the scores when constructing an informative measure; the measure is built around the idea of capturing regulatory potential based on prior ideas about the effect regulatory mechanisms have on expression output, but should also be usefully applied in inferring biological understandings.

In terms of the actual types of genes expected to be significantly complex, master regulatory genes and genes associated with the regulation of development were expected to score highly, since such genes are reported to be regulated by multiple and often highly conserved enhancers and display temporal and spatially heterogeneous patterns of expression, for example [Elgar, 2009, McEwen et al., 2009b]. GO term analysis confirms this, with developmental regulation scoring as the most significant GO term for the raw complexity scores. Thus, whilst making no explicit assumption about which genes should or should not be complex, the kinds of genes functionally associated with low and high complexity generally appears to match expectation.

6.2 Complexity scores provide useful information over and above what is observed from entropy scores

Entropy scores are based on the first of four factors listed above to determine the complexity of a gene. Since complex clearly encompasses more information than this, it should be expected that one would be able to infer more information about the potential regulatory landscape of the gene. As breadth of expression is often used as a proxy 'complexity' measure for expression patterns, it is useful to measure the extra information which may be observed in complexity scores, and how the two measures differ and complement each other.

An interesting feature of the presented expression complexity metric, which avoids discussion of normalisation strategies entirely, is that it is able to untangle genes within a given breadth of expression by ranking their expression profiles in terms of observed up- and down- regulation. This is particularly useful given the fact that many genes are non-specific in their expression, but not necessarily uniform. The measure therefore

provides a way of partitioning ubiquitously expressed genes between those relating to so called 'ultra-housekeeping' categories on the low complexity end of the scale, and those genes exhibiting more complex regulatory programmes as a result of observed changes in levels of expression across cell types. Ultra-housekeeping has been an important and insightful measure in recent studies ([Forrest et al., 2014, Young et al., tted] where it was crudely calculated based on a simple threshold of the maximum to median (normalised) expression ratio. Consequently, expression complexity represents a more robust and information-richer method with which to define such genes.

6.3 High complexity genes are depleted in CpG islands in their core promoter

Complexity scores and normalised complexity scores were first interrogated for their differences between the presence or absence of a CpG island or TATA box overlapping the core promoters of genes (Section 5.4) . The most complex genes were those depleted of CpG island associations. The presence or absence of CpG island in the promoter is strongly related to expression breadth, and given the relationship of CpG islands with polycomb repression and ubiquitous expression, it is less surprising that complex genes are less enriched in CpG islands [Riising et al., 2014]. However, in order to obtain the extra information obtained by complexity scores, independent of this expression breadth relationship, complexity was first adjusted for by entropy and the remained variation in complexity observed. It was found that, independent of expression breadth, complex genes are depleted of CpG islands in their core promoters. Thus, it would appear that the lack of a CpG island at the core promoter in ubiquitous genes somehow allows for flexible expression profiles which are detected by the complexity score methods.

Indeed, this hypothesis is supported by studies where it has been seen that promoters with CpG depletion have transcription factor binding motifs which appear to be tissue specific [Roider et al., 2009]. Furthermore, it has been seen that genes with and without CpG islands contain different patterns of chromatin modifications associated with transcription [Vavouri and Lehner, 2012]. Thus, since by definition complex genes exhibit

tissue specific changes, it may be that these complex and ubiquitous genes depleted of CpG islands contain a different signature of sequence based core promoter motifs and chromatin organisation which allows them to regulate more changeable expression profiles between different cell types.

6.4 Complexity scores are associated with measures of cis- regulation

Complexity scores were derived based on the idea that the greater the control of gene expression through cis-regulatory binding events, the more likely a gene is to exhibit more complex changes and/or switches between on or off states across the cell types in its profile of transcriptional expression levels.

In order to test this hypothesis, measures of conservation (GERP) and DNase I hypersensitivity in and around each gene were compared to its estimated complexity score (Section 5.6). Whilst the two often overlap, DHS are often not conserved due to high rates of regulatory turnover observed in the human genome [Meader et al., 2010, Villar et al., 2015], hence there is merit in attempting to observe cis- regulation from a variety of datasets.

6.4.1 Hypersensitive I marks at the upstream gene region and promoter region in the absence of conservation

Complexity scores show an enrichment for hypersensitive sites in the proximity of the gene (Section 5.6), but these same regions are not necessarily enriched in conservation. This suggests that whilst many hypersensitive sites overlap conserved regions, complexity might be driven by hypersensitive sites not conserved across species. These sites may be specific to human; or since promoter turnover is high due to a high mutation rate [Taylor et al., 2006, Villar et al., 2015, Young et al., tted], a functionally equivalent element might be present between species but not functionally constrained in its sequence [Dermitzakis and Clark, 2002, Elnitski et al., 2003]. Indeed, the proportion

of shared regulatory sequence between species dramatically decreases as phylogenetic disease increases [Meader et al., 2010], thus it is unsurprising that most regulatory elements acting upstream in cis- may not be under selective constraint over the phylogenetic range measured by GERP. This finding is in some part backed up by a significant enrichment in enhancers in the upstream distance of the gene, suggesting that some of these hypersensitive sites may function as enhancers but lack a conservation enrichment signal.

Entropy scores show a large enrichment for DHSs in the promoter region of the gene (Figures 5.31 and 5.31), suggesting that the presence of accessible chromatin at the promoter is associated with broadly transcribed genes; this is not a surprise since accessible chromatin is a mark of active transcription and the more cell types exhibiting DHSs, the greater the chance of detecting such sites in ENCODE (and similar projects).

6.4.2 Hypersensitive I marks and conservation in first intron of the gene

Enhancers correlated particularly strongly with complexity scores in the first intron of the gene, explaining almost 3% of the variation in non-normalised scores. Interestingly, the same enrichment in first intronic cis-regulation is not observed in entropy scores, suggesting that this source of cis-regulation is due to regulatory complexity but not expression breadth. This effect seems to be most pronounced in genes that evolved in the earlier stages of the evolution of multicellular life, when metazoan bauplans were being defined (see gene age enrichments). Consequently, it is speculated that this enrichment is primarily indicative of developmental regulatory genes.

6.4.3 How much variation is explained in total by cis-regulatory sources

In total, conservation scores, DHSs and predictive enhancers explain 7.0% of the variation in complexity scores and 4.21% of the variation in normalised complexity scores (Figures 5.31 and 5.32). With the inclusion of interaction effects, this rises to 8.4% for the complexity scores and 6.81% for the normalised complexity scores.

Restricting to ubiquitously expressed genes, the explained variance proportion rises to 10.0% (normalisation independent), and when restricted to non-ubiquitous genes the explained variance drops to 3.2% for the normalised scores, and decreases for the complexity scores as breadth is reduced (8.0% for genes expressed in fewer than two thirds of primary cell types). Since ubiquitous genes are separated in terms of their complexity only by their profile of differential expression in relation to their sample structure, it is suggested that genes exhibiting high levels of differential expression between multiple cell types are controlled to a greater extent than genes exhibiting high complexity scores due to on and off switching between cell types. This provides an interesting insight into how regulatory programmes differ according to types of genes.

6.4.4 Conclusions for cis- regulation

In conclusion, the results suggest that expression complexity is indeed associated with proximal cis- regulatory elements surrounding the gene, although much of the variance in complexity scores is still explainable by other sources, which could perhaps refer to long-range interactions, trans- regulatory interactions and chromatin structure, amongst other factors. It must be pointed out that whilst part of the premise of the project was to attempt to understand how and where regulatory information was encoded, it is convenient to attempt to group genes as either 'cis-regulated' or 'trans-regulated' according to their 'preferred' mode of regulation. Although genes are generally regulated through a combination of the two sources and are not exclusive towards one end of the cis- and trans- scale, there is merit in attempting to understand the potential dominance of one mechanism vs another.

6.5 Complexity scores are highly associated with promoter histone marks

It has been seen that histone modifications explain the largest amount of variance in the complexity scores (Section 5.7). As might be expected, H3K4me3 marks, generally

associated with activation, are closely associated with entropy scores, explaining up to 47% of the variance. This is because entropy is a measure of breadth, maximised in ubiquitously expressed genes which are by definition actively expressed in all cell types.

Polycomb group proteins (PcG) catalyse H3K27me3 histone modifications, forming repressive complexes PRC1 and PRC2. These have important functions in determining the identity of stem cells and cellular differentiation. Whilst DNA methylation in general silences the expression of a gene in a cell lineage, polycomb targeted genes observed in somatic cells provide an important potential mechanism for state switching in response to a range of conditions. Promoters of highly regulated genes poised for potential expression are bivalent chromatin domains enriched for both the H3K27me3 mark associated with polycomb repression and the H3K4me3 mark associated with gene activation. Bivalent chromatin at the promoters of genes are poised for potential activation and are thought to be essential for defining cellular identity and function [Lesch and Page, 2014]. Poised chromatin has been seen to be maintained at the promoters of developmental genes at multiple stages of development. Hypotheses for bivalent chromatin in mammalian germ-lines has been thought to act as a prevention mechanism for the locking of the silencing genes by DNA methylation.

6.5.1 Associations with complexity scores and epigenetic marks

Whilst H3K4me3 activation mark was the main determinant of expression breadth, as observed in entropy scores, complex genes exhibited profiles of varying H3K4me3 and H3K27me3 marks. Non-normalised complexity scores showed a positive correlation with H3K4me3 marks, since this score ranks highly genes which are expressed across a broad range of cell types. In terms of repressive marks, raw scores increased with complexity but dropped in the category where H3K27me3 was observed at all promoters of the analysed tissues. Entropy scores suggest that these are relatively specific genes; it could be that since these genes are broadly exhibiting repressive marks, the expression output will suggest that the gene is cell-type restricted.

The above reasoning is why normalised complexity scores, which up-weights restricted genes, shows a strongly increasing relationship with H3K27me3 repressive marks and highly significant relationship with bivalent marks. Genes which are bivalent everywhere generally had the highest complexity scores, an effect that is discussed further in the next section.

6.5.2 Bivalent genes are highly complex

Integrative data from the Epigenetics Roadmap Project [Kundaje et al., 2015] suggest that over 5000 genes from the 16111 under study may potential be associated with bivalent chromatin domains in at least one cell type, and over 1300 may have bivalent chromatin domains in the majority of cell types for a given gene. These values are only potential bivalent genes as the numbers are unlikely to be accurate of true bivalency; as whilst by definition both marks are observed at the same promoter in the same cell type to suggest a promoter is bivalent, there is no indication of whether the marks were observed together in the same cell, or separately in two different cells within the sample taken from that cell type [Voigt et al., 2013].

Genes exhibiting potential for high breadth of bivalency across tissues were in general more complex than those without; this result is highly prominent in complexity scores normalised by expression breadth. In this measure, genes highly associated with activation (H3K3me4) were the least complex and genes highly associated with polycomb repression (H3K27me3) were the most complex; however genes exhibiting these marks together were highly complex. The results suggest a strong association based link between genes held in a poised repressed state, a mark thought to be present in highly regulated developmental genes, and the complexity of its coupled expression pattern. Since these marks were measured in adult somatic tissues, it suggests a model whereby the more tissues where a gene exhibits a bivalent mark in its promoter, the more it is retaining its on-off switching abilities further into development. Loss in bi valency or methylation at this promoter would result in a more stable expression without the flexibility to exhibit subsequent switches.

6.5.3 Poised chromatin interacts with CpG island status

Regions where CpG di-nucleotides are overrepresented, referred to as CpG islands, are seen to overlap a large proportion of promoter regions and are in particular associated with the promoter regions of housekeeping genes, which are expressed in every cell type and maintain and active chromatin state [Bird, 2002]. They appear to play an important role in establishing chromatin state, and are predictive of presence of both H3K4me3 and H3K27me3 together [Deaton and Bird, 2011, Orlando et al., 2012]. Thus, bivalency associated genes which have CpG islands in their core promoter are highly complex, as can be seen when observing interactions between CpG presence and poised chromatin breadth (Figure 5.43). In this Figure, whilst non-cpg genes remain similarly complex independent of epigenetic status at the promoter, genes associated with promoter overlapping CpG islands are highly associated with poised chromatin state, with CpG island associated genes with broad poised chromatin marks exhibiting the highest complexity, together with non-CpG associated genes. Separating this out, bivalency status (0,1 or 2) explains 28.3% of the variation in normalised complexity scores in the subset of genes associated with a CpG island, and only 2.4% of the variation in complexity in genes not associated with a CpG island. Whilst bivalent marks were measured across a set of 22 adult tissues, it appears that this association also holds when measuring across ES cells (8 tissues, correlation of 0.75 in breadth of observed bivalent promoter marks between adult and ES, and ES cell bivalency breadth accounted for 22.9% of the variation in normalised complexity scores). Therefore, what matters is an overall signal of bivalency acting on a gene's promoter as opposed to the developmental stage in which bivalent marks are observed.

Even specifically focussing on genes without CpG islands overlapping their core promoter, combinations of H3K27me3 and H3K4me3 still accounted for 15.9% of the variation in normalised complexity scores, with repressive H3K27me3 status acting the most strongly, suggesting that poised chromatin in absence of bivalency (since non-CpG overlapping promoters are not bivalent) is still associated with regulatory complexity.

6.5.4 Conclusions for epigenetic modifications and their association with complexity scores

The strong association with epigenetic promoter modifications leads to the question of potential mechanisms which may possibly be driving this association.Two possible models for polycomb regulation, suggested by [Voigt et al., 2013] could predominate. The first is that through the differentiation of cell types, there are decision points in which cell lineages diverge and the cell must decide the epigenetic architecture of cells between the two diverged lineages. Genes not requiring expression in a group of downstream differentiated cell types are often silenced through methylation, repressed with polycomb signals or held in a poised bivalent state for future activations. Thus, there is effectively a tree of switching states based on epigenetic modifications occurring at the promoters of genes through development, whereby the cell is making decisions. The final pattern across adult primary cell types is reflected in the final cellular states at the leaves of this tree. Thus, patterns of on and off states in expression are observed across a profile and these are captured using scores of complexity.

The second possible scenario is that the on and off switching observed across a cell type is based on the polycomb-mediated fine-tuning of transcriptional activation based on thresholds of activation and repression [Voigt et al., 2013]. Genes with poised chromatin marks in their promoters may be tightly regulated whereby activation and repressive signals act together to flip states between on or off according to the requirements of the cell. When the signals pass a predetermined threshold, a switch occurs which causes expression changes in a binary manner as opposed to gradual changes in expression. Such a model has been discussed by [Voigt et al., 2013] and it is postulated that bivalent promoters acting in this manner generally exhibit reduced transcriptional noise (testable as further work using CAGE library replicates). Such a model has been observed in HOX clusters ([Montavon and Duboule, 2013]), and also when modelling developmental gene expression patterns in Drosophila [Dupont et al., 2015]. This hypothesis is backed up by observations with haploinsufficient genes; these genes are associated with disease status when one copy of the gene is knocked out due to deletions, so are tightly controlled in their expression levels with only small variation and it is deviations from these

finely tuned levels that are associated with disease. Indeed, disruptions in epigenetic mechanisms have been seen to play a role in haploinsufficciency [Williams et al., 2010], furthermore these genes were shown to be more complex than non-haploinsufficient genes (Subsection 5.10.6).

A further point is that whilst bivalent marks are thought to be representative of poised chromatin states, whereby removal of H3K27me3 repressive marks results in the subsequent activation of the gene, these results are not necessarily fast; recent estimates suggest that histone modifications acting on a given gene may be removed or established within a few minutes [Anink-Groenen et al., 2014]. Thus, whilst poised chromatin allows for expression switching and/or changing, as opposed to methylation patterns which are in general permanent in somatic cells, it is unlikely that it is suitable for genes requiring urgent response by stimulus, such as hypoxia, infection or heat stress responses. Many of these genes scored highly in terms of their complexity scores for ubiquitously expressed genes, suggesting that whilst complexity in regulation appears to be highly associated with epigenetic activity acting on a gene's promoter, other mechanisms are clearly acting to generate expression changes in these sets of genes. Indeed, many of these highly complex genes were seen be associated with a lack of CpG island in their core promoter. Since CpG island mediates polycomb complexes, this suggests that these CpG depleted complex genes may be acting under a variety of protein-protein interactions under a different control to the interplay between H3K4me3 and H3K27me3 promoter modifications.

In conclusion, how the expression pattern of the gene across the range of cell types is influenced by its chromatin structure, according to its pattern of histone modifications, is generally not well understood, and the idea of a histone 'code' determining expression largely debated. These results provide some insight into how combinations of H3K4me3 and H3K27me3 work together in order to generate diverse patterns of transcriptional output. However, since H3K4me3 and H3K27me3 represent only a subset of possible histone modifications, a further, more comprehensive analysis of histone modifications could unravel more intricate relationships, especially as more information across large ranges of cell types becomes more and more available in the future. In particular, whilst

the aforementioned pair of modifications are highly effective in explaining the dynamics of activation and repression in gene expression output, more dynamic modifications may explain expression information not caused by activation and repression.

6.6 Age of gene is associated with complexity scores

The evolution of complexity is often thought of in a variety of terms, such as the morphological complexity observed across different species through time, via the number of cell types (e.g. used as an indicator of morphological complexity in [Chen et al., 2012]) or the size of the genome in an organism. In general, it is often highly debated whether complexity increases in a linear fashion through time (e.g. whether it also includes bursts or trends of simplification), whether it reaches an upper limit, and whether it evolves in a passive or driven manner ([Yaeger et al., 2011]).

There is debate regarding how much non-coding regulatory sequence is functional and how much of this sequence may be employed to the regulation of the expression of any given gene. For example, the amount of available space immediately upstream of a gene provides a physical limit to how many transcription factor binding sites or alternative promoters may potentially fit. The evolution of regulatory complexity has been looked at by [Warnefors and Eyre-Walker, 2011a] and [Lowe et al., 2011]. The former analyses looks at eight separate aspects of regulatory complexity: namely transcription factor binding sites around the gene, conserved bases upstream, the number of TSSs, splicing isoforms, polyadenylation sites, miRNA sites, NMD proportion and RNA editing proportion. In combination these measures actually capture estimates of transcript diversity rather than transcriptional regulatory complexity as defined in this thesis. Comparing the measures across genes divided into 18 age categories, [Warnefors and Eyre-Walker, 2011a] found that older genes appeared to be more enriched in transcript complexity, where genes in the oldest category, eukaryota - the earliest eukaryotes, was in general between 1.3 - 3.1 times more likely to contain more of the mechanism under study. The suggestion is that older genes have had more time to accumulate more regulatory features, which leads to the evidence that complexity has not yet reached

its upper limit as newer genes have not yet accumulated the same amount of features as older genes.

The present study (Section 5.8) also sees high levels of non-coding sequence conservation in older genes, particularly those around the time of the emergence of multicellularity, according to conservation observed in the first intron and across the gene as a whole (excluding the first intron). Whilst newer genes do not exhibit high levels of non-coding constraint, genes associated with human (the newest category) do appear to contain more conservation on the gene and in its first intron. The same pattern is observed in first intron DNase I hypersensitivity sites (DHSs), suggesting that some or all of these conserved regions are indeed regulatory, but also since the number of DHSs generally outnumbers the observed GERP sites, particularly in older genes, there is perhaps a signal of accumulated regulatory effects occurring over time. Exon count was dramatically also greater in older genes, suggesting that these genes are prone to greater exonization. This is supported by [Corvelo and Eyras, 2008], which suggests that the creation of exons is related to the acquired ability to regulate splicing events.

The results for conservation and hypersensitivity (second and third links above) seem to agree with the idea that older genes are more complex in sequence: older genes (from multicellular organisms onwards) are highly conserved in their regulatory regions, and contain more hypersensitive sites (particularly in their first intron, the results are very similar for the two). In all these plots, human specific genes appear to be outliers - they have are more proximal conserved non-coding sequence, more DHS sites and more exons compared to the six evolutionary groups prior to them.

The present results observe genomic complexity from the perspective of the implied transcriptional regulation of individual coding genes, based on the age of its encoded protein. Entropy scores as a measure of expression breadth show a clear tread towards cell-specificity in newer genes from the advent of multicellularity. The most complex genes by the raw complexity scores coincide with the advent of multicellularity, which includes genes associated with sequence specific DNA binding, allowing for diverse cell types which are able to regulate expression to different levels. This key result is of particular interest when combined with entropy scores; newer genes evolve in an

increasingly cell-specific manner, although whilst providing highly specific functions to an organism, they are not necessarily complex genes.

The evidence presented in Figure 5.44 suggests that entropy evolves in a passive manner, with genes appearing at a variety of breadths at each stage, often accompanied by a new minimum level of complexity and associated with a higher degree of specificity, accompanied by variance levels increasing over time. This is observed by measuring change across the quantiles of scores; the upper quantiles did not significantly change whilst changes across the lower quantiles explained nearly 11% of the variance in entropy scores.

As complexity scores are based on gene expression output rather than direct regulatory information, there is an argument that changes in expression and patterns of on/off switching can be more intricate in newer genes despite evidence that these genes have not had time to evolve mechanisms (or accumulate, as per Warnefors and Eyre-Walker [2011a]). It is possible that the regulatory mechanisms predominating over these newer genes are somehow different those preferred by older genes, for example they may be associated with more long-range interactions whilst having relatively few proximal cis-regulatory elements, or they may be dominated by trans- regulatory effects. Indeed, there is an argument that trans-regulatory processes appear to play a greater role in mammals due to increases in miRNA targets regulating vertebrate genes [Chen et al., 2013]. It was seen in Section 5.6 that genes complex in their expression are enriched in first intronic conservation and DNAse I hypersensitive sites. It appears from Figures 5.45 and 5.46 that these genes are likely to be older genes which may associate with the time period where developmental regulating genes such as *HOX*, *SOX* and *PAX* evolved. It would be of interest to potential future exploration to test whether these newer complex genes are more dominated by trans- based effects.

6.7 Complexity scores are predictive of a variety of categories of disease states

Interruptions in regulatory mechanisms of genes often results in expression changes, silencing or activation, leading to potential disease. Identifying lists of genes susceptible to disease through their perturbation is a crucial step towards prioritization of targets for specific cause/effect relationships. It was hypothesised that regulated genes provided a larger mutational target through binding targets in cis- and the number of trans- acting on the gene, concepts related to the 'transmutational' target size of a gene [Landry et al., 2007]. The application of complexity scores allows for the comparison between regulatory complexity and disease status without having to specify the particular regulatory mechanism involved as a causative factor for any potential associations. Those complex genes highly associated with a specific disease may then be screened for specific regulatory targets, such as mutations in enhancer binding sites or unexpected changes in chromatin. It also makes no pre-assumption of disease status, since only healthy adult primary cells were used to calculate the scores. An analysis of disease associated genes was carried out in Section 5.10

Surprisingly, the scores were found to correlate significantly with a variety of disease states. Grouping potential disease genes by anatomical categories found various significant relationships; in particular, bone, immune, cardiovascular and skin diseases were all found to be highly significant in at least one of complexity or normalised complexity and at least significant in the other. In particular, 1089 genes were associated with bone related diseases and were significantly more complex than non bone disease genes (odds ratios 1.67 for complexity, 2.72 for normalised complexity, both highly significant), but not necessarily restricted to a given expression breadth (odds ratio 1.22 for entropy, not significant). Cancer associated genes in general (2204 genes) were also found to be significantly more complexity (odds ratios 1.99 and 1.96 for complexity and normalised complexity, respectively). Breaking this down into smaller categories, lung cancer, carcinoma and sarcoma were significant, although small number of associated genes for

these categories results in large error bounds and an do not all survive multiple testing corrections.

Scores were highly associated with specific disease states, such as Alzheimer's, cancers and diseases as a results of haploinsufficiency. Genes which highly expression complexity scores were enriched in SNPS associated with disease, in particular SNPs found in the intergenic and intronic regions of the gene (Figure 5.60). Furthermore, effects accumulate - the more associated disease SNPs, the more complex the gene (Figure 5.59). This suggests that disruptions in the regulatory regions of genes with the potential for intricate regulatory control as implied by transcription output are implicated in disease and this can be observed through steady-state gene expression.

6.8 Limitations of the analysis

6.8.1 Technological limitations - speed of processing

In recent years next generation sequencing developments and continued reductions in costs are allowing for the generation of large quantities of cellular states with increasing amounts of biological replication. Furthermore advanced computational methods have been published in recent years for easy and efficient determination of differential expression between states. In this study we have introduced and applied an alternative metric to the Shannon entropy for measuring the complexity of a gene's expression output, which utilises this information and provides a convenient measure for each gene of the observed patterns of state switches between on, off, up or down states. The measure is easily interpretable; all scores lie between 0 and 1, where 1 represents state changes between every pair of states in the study.

6.8.2 The measures do not take into account magnitude or direction of differential expression

A point to make in this study regarding primary cell types is that it is not considered whether there is a significant up or down regulation between pairs of cell types - mainly because it is difficult to determine which cell type to assign as a reference. Thus, only the significance of changes are considered. Whilst not presented in the previous chapter, a more clear structure may be observed in, for example, time course data, or data where there are treatments or controls. In the case of a time course dataset, time may be considered to be based in reference to the lower time point of the pair and differential expression could be reported as increased or decreased relatively. However, complexity scores do not take this into account. Thus a different version may be warranted for some datasets.

6.8.3 Best way to normalise scores

There is clearly a challenge in the normalisation of the data. That is, the complexity scores appear to not work so well in the highly cell restricted cases; these give drastically different scores between the normal score and the normalised score. Parameters like conservation and first intron hypersensitive sites clearly follow aspects of both scores. Therefore, a better way of normalised the tissue restricted cases could be warranted. Although it is unclear on how this may be achieved, note that the normalisation strategy is only an issue comparing across expression breadths, and not within.

6.8.4 The problem of unmatched cell types

There is a very strong relationship with histone modifications, and many complex genes contain a variety of H3K4me3 and H3K3me27 marks in the core promoter region across across tissues. The best way to measure the effects of polycomb regulation on the observed patterns of expression is to measure the presence of polycomb marks in the matching set of primary cells over which the complexity measure was calculated, and

to measure expression over time in all the cell types (and potential all possible conditions!) so that expression switches between poised and activated can be monitored. Similarly, the ENCODE DNAse I hypersensitivity sites data did not use the matched set of cell types with complexity scores, although many of the 125 used cell types potentially overlap. Thus, highly restricted sites will have been missed, but the expression based on its interactions included in the expression complexity scores. The predicted enhancers based on bidirectional expression was useful in this regard, as enhancers were inclusive of the same CAGE libraries used to calculate complexity. This may explain why enhancers had good predictive power, particularly when counted in the first intron of the gene where their effects were stronger than that of hypersensitive site counts.

6.9 Further work

6.9.1 Exploring different connectivity measures and weight structures

The flexibility of this approach is that different measures of connectivity and sample structure can be used over the graph. Some aspects of the measure may be varied:

The connectivity measure used to calculate the final score. This was based on the eigenvalue decomposition of the graph which appeared to provide an average connectivity over the graph which best matched the intuition of the kinds of genes expected to be complex (i.e. developmental regulatory genes were the highest GO term). Some global connectivity measures correlated well with the current approach, and others did not necessarily measure the graph in the expected way and were not considered. However, other measures may be used, for example some which capture more local properties of the graph. Furthermore, the flexibility of applying a complexity measure which captures node connectivity as well as graph connectivity is that one can up or down weight different properties of interest. For example, one could priorities on-on connections over on-off connections, or vice-versa.

The weighting structure between the cell types can also be varied. Weights calculated for the current results were based on a transformation of the Pearson's correlation

coefficient applied to the log of the tags per million (plus a pseudocount to avoid the log of zero), which provided an even distribution of weights across all pairs of cell types. However, other methods could be used, such as branch lengths based on a predicted tree over the samples. In the case of time points, weights based on distance between time points could be used. In the scores currently applied across time course data (not presented in the current work), the weights between time points were varied to distinguish between genes which varied dramatically between time points (for example, a switch on from off to highly expressed between two adjacent time pints) and genes which exhibited gradual change over a number of time points. This letter case was useful in picking up genes whereby the data was perhaps too noisy to pick up a small change between adjacent time points, but was able to pick up slightly larger changes between two or three time points.

Accurately calculating differential expression between pairs of samples is an important aspect of calculating measures of expression complexity. In this project bayseq was chosen as it shared information about the variance between genes in samples, thereby improving accuracy, particularly in cases where there were a very small number of replicates (sometimes only one). Furthermore, bayseq returned probabilities, allowing for the problem to be applied in a probabilistic framework without having to define cutoffs. Furthermore, it was conservative, meaning that it gave fewer false positives (although perhaps some false negatives), which appealed since there were 11026 possible pairs of samples for ubiquitously expressed genes and many false positives may have skewed the score. As protein coding genes are generally more highly expressed than for example enhancers it was felt that this was a good approach. If one were to calculate the scores across enhancer data then a less conservative method should be used, as with fewer tags and a lot of noise a highly conservative method may pick up no differential expression. This was observed from applying bayseq across some of the individual TSS level data (not shown in the thesis). One's own models may also be applied - for example, as part of this project, JAGS was used to model time course expression data.

6.9.2 Improving feedback between regulatory inputs and outputs

Understanding the regulatory architecture of the genome and how it relates to the transcriptional output of specific genes aids in the measurement of gene expression profiles to capture this regulatory information. In turn, complexity scores aid in the understanding of these regulatory processes, potentially guiding researches to important sources of regulation in subsets of genes.

6.9.3 Better understanding of cis- vs trans- effects and their relative contribution to complexity scores

It is highly probable that most complex genes display a range of different mechanisms that contain a variety of cis- and trans- based effects and much of the variation in scores is probably occurring in protein-protein interactions occurring off site from the genome, which were not measured as a parameter. A better understanding of trans- effects and how the influence gene expression would be useful to correlate with complexity scores.

In terms of cis- regulation it was found that proximal elements were significantly correlated with complexity (Section 5.6), but as cis- regulation may potentially occur anywhere on the chromosome, although 1MB is thought to be a potential limit ([Hill and Lettice, 2013]) a stronger understanding of distal-interactions is required. This generally done with Hi-C data, but such data is still very limited in cell types and only measure specific cell lines (with few replications), therefore this was not done in this project as there is not the opportunity to sample the diversity of interactions to warrant a detailed investigation. However this missing data is beginning to be collated and provides a future avenue for tying transcriptional regulation with long range interactions.

6.9.4 Obtain a stronger understanding of ubiquitously expressed genes in terms of comparisons with cis-regulatory effects

It was seen that seen that when restricting to the set of ubiquitously expressed genes, highly complexity genes appear to have a stronger surrounding landscape of cis-regulatory element, furthermore they appeared to be slightly depleted in CpG islands in their core promoter. Highly complex ubiquitous genes appeared to be associated with response based effects (Table 5.7). It is unclear whether these response based genes which are highly complex are those with broad cis-regulatory landscapes, or whether the set of complex ubiquitous genes contain genes which these broad landscapes which are not necessarily response-based (i.e. these complex genes contain a mixed repertoire of regulatory mechanisms).

A further test of this hypothesis could be to split up genes according to different GO terms (associated with for example, signalling response or developmental processes) and compare the architectures of regulatory elements (for example, DHSs or CpG/TATA effects) between the different groups. Therefore, whilst it is generally understood that ubiquitously expressed genes undergo different sets of regulatory mechanisms than ubiquitous genes, it is possible that across the identical breadths of expression it is possible to observe a wide range of regulatory control.

6.9.5 Single cell sequencing

Whilst primary cells are effective at providing a pure lineage to understand changes in expression from cell type to cell type, single cell sequencing is the best way to understand how expression changes on a cellular level. For example, tightly regulated genes probably stay close in expression from cell to cell, others might exhibit a noisy profile. Fully understanding the patterns between individual cells in a spatial and temporal manner is a challenge which technology may make a realistic goal for the future. This will help to fully capture the full extent of the regulatory mechanics occurring within the cell [Levo and Segal, 2014].

6.10 Work not included in this thesis and projects started

6.10.1 Analysis of time-course data

The complexity measure has been applied across the CAGE libraries for the human time courses from the FANTOM5 project. These include a mixture of differentiation based time-courses and some stimulus response based time courses. Patterns for this data have been clustered to identify early response genes, and other patterns [Arner et al., 2015]. Measures of expression complexity complement this data by picking out non-standard patterns, such as multiple responses within a single time course, or responses which exhibit gradual vs sudden changes. Genes which are known regulators such as EGR1 and EGR2 appear to rate highly across all time-courses, suggesting merit in continuing with the analysis.

6.10.2 TSS level data

Analysis of measures applied to TSS data. Scores have been applied to robustly defined TSS associated with genes, with the aim of understanding how individual TSS influence the overall transcription of the gene. From current analysis, the closest TSS to the gene generally the most highly expressed and the subsequent ones more restricted, often provided more complex patterns in expression. However this has not been fully explored and has therefore not been included in this thesis.

6.10.3 Analysis of human-mouse matching cell types

Started an analysis of human-mouse matching cell types. This has been conducted in a probabilistic frame work, measuring changes which appear between matched cell types in either human or mouse across orthologous genes, vs changes which appear in both human and mouse.

Appendix A

Samples used in this analysis

Below is a list of all of the primary cells included in the analysis, with their replicates and the developmental stage (either mesodermal cell or mesenchymal if an annotation exists). Reference: [Forrest et al., 2014]

Group 1	Reps	Developmental
Smooth Muscle Cells - Airway Asthmatic	6	
Skeletal Muscle Satellite Cells	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Colonic	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Uterine	2	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Prostate	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Subclavian Artery	3	CL:0000222 (mesodermal cell)
Skeletal Muscle Cells differentiated into Myotubes - multinucleated	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Aortic	4	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Brain Vascular	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Esophageal	2	CL:0000222 (mesodermal cell)
Skeletal Muscle Cells	6	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Airway Control	4	
smooth Muscle Cells - Bronchial	2	CL:0000222 (mesodermal cell)
Cardiac Myocyte	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Brachiocephalic	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Tracheal	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Carotid	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Umbilical Artery	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Umbilical Vein	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Pulmonary Artery	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Internal Thoracic Artery	3	CL:0000222 (mesodermal cell)
Smooth Muscle Cells - Coronary Artery	3	CL:0000222 (mesodermal cell)

Group 2		
Placental Epithelial Cells	3	
Retinal Pigment Epithelial Cells	4	
Tracheal Epithelial Cells	3	
Corneal Epithelial Cells	3	
Alveolar Epithelial Cells	3	
Ciliary Epithelial Cells	3	
Olfactory Epithelial Cells	4	
Lens Epithelial Cells	3	
Mammary Epithelial Cells	3	
Small Airway Epithelial Cells	3	
Nasal Epithelial Cells	2	
Bronchial Epithelial Cells	7	
Esophageal Epithelial Cells	3	
Gingival Epithelial Cells	3	CL:0000134 (mesenchymal cell)
Mallassez-derived Cells	2	
Prostate Epithelial Cells polarized	3	

Group 3		
Hepatic Stellate Cells Lipocyte	3	CL:0000134 (mesenchymal cell)
Preadipocyte - Omental	3	CL:0000134 (mesenchymal cell)
Preadipocyte - Visceral	3	CL:0000134 (mesenchymal cell)
Preadipocyte - Subcutaneous	3	CL:0000134 (mesenchymal cell)
Preadipocyte - Breast	2	CL:0000134 (mesenchymal cell)

Group 4		
Endothelial Cells - Microvascular	3	CL:0000222 (mesodermal cell)
Endothelial Cells - Aortic	4	CL:0000222 (mesodermal cell)
Endothelial Cells - Umbilical Vein	3	CL:0000222 (mesodermal cell)
Enthothelial Cells - Lymphatic	3	CL:0000222 (mesodermal cell)
Endothelial Cells - Artery	3	CL:0000222 (mesodermal cell)
Endothelial Cells - Thoracic	2	CL:0000222 (mesodermal cell)
Endothelial Cells - Vein	3	CL:0000222 (mesodermal cell)
Hepatic Sinusoidal Endothelial Cells	3	

Group 5		
Melanocyte - Light	3	
Mesothelial Cells	2	CL:0000222 (mesodermal cell)
Melanocyte - Dark	3	

Group 6		
CD8 T Cells	3	CL:0000134 (mesenchymal cell)
CD4 CD25 CD45RA Regulatory T Cells expanded	3	
Macrophage - monocyte derived	3	CL:0000134 (mesenchymal cell)
CD4 CD25 CD45RA Naïve Regulatory T Cells	1	
Natural Killer Cells	3	CL:0000134 (mesenchymal cell)
Basophils	1	CL:0000134 (mesenchymal cell)
Mast Cells	4	CL:0000134 (mesenchymal cell)
CD4 CD25 CD45RA - Memory Regulated T Cells expanded	3	
CD34 Cells Differentiated to Erythrocyte Lineage	2	CL:0000134 (mesenchymal cell)
CD4 CD25 CD45RA - Memory Regulated T Cells	3	CL:0000134 (mesenchymal cell)
CD19 B Cells	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes	3	CL:0000134 (mesenchymal cell)
Gamma Delta Positive T Cells	2	
Migratory Langerhans cells	3	CL:0000134 (mesenchymal cell)
Neutrophils	3	CL:0000134 (mesenchymal cell)
CD4 T Cells	3	CL:0000134 (mesenchymal cell)
CD4 CD25-CD45RA Naïve Conventional T Cells	3	
CD4 CD25-CD45RA Naïve Conventional T Cells expanded	3	
CD4 CD25-CD45RA Memory Conventional T Cells expanded	3	
Immature Langerhans Cells	2	CL:0000134 (mesenchymal cell)
Dendritic Cells - Monocyte Immature Derived	2	CL:0000134 (mesenchymal cell)

Group 7		
Renal Cortical Epithelial Cells	2	
Osteoblast	3	CL:0000134 (mesenchymal cell)
Osteoblast - Differentiated	3	CL:0000134 (mesenchymal cell)
Tenocyte	3	CL:0000134 (mesenchymal cell)
Synoviocyte	3	CL:0000134 (mesenchymal cell)
Renal Proximal Tubular Epithelial Cells	3	CL:0000222 (mesodermal cell)
Renal Mesangial Cells	3	CL:0000134 (mesenchymal cell)
Pericytes	3	
Hair Follicle Dermal Papilla Cells	3	CL:0000134 (mesenchymal cell)
Trabecular Meshwork cells	3	CL:0000222 (mesodermal cell)
Renal Glomerular Endothelial Cells	4	CL:0000222 (mesodermal cell)
Hepatocyte	3	CL:0000134 (mesenchymal cell)
Prostate Stromal Cells	3	CL:0000134 (mesenchymal cell)
Sertoli Cells	2	CL:0000134 (mesenchymal cell)
Renal Epithelial Cells	3	

Group 8		
Myoblast	3	CL:0000222 (mesodermal cell)
Fibroblast - Conjunctival	2	CL:0000134 (mesenchymal cell)
Fibroblast - Dermal	6	CL:0000134 (mesenchymal cell)
Fibroblast - Skin Spinal Muscular Atrophy	3	CL:0000134 (mesenchymal cell)
Fibroblast - Aortic Adventitial	3	CL:0000134 (mesenchymal cell)
Fibroblast - Skin Dystrophia Myotonica	3	CL:0000134 (mesenchymal cell)
Fibroblast - Choroid Plexus	3	CL:0000134 (mesenchymal cell)
Fibroblast - Lymphatic	3	CL:0000134 (mesenchymal cell)
Fibroblast - Mammary	3	CL:0000134 (mesenchymal cell)
Peripheral Blood Mononuclear Cells	3	CL:0000134 (mesenchymal cell)
Fibroblast - Villous Mesenchymal	3	CL:0000134 (mesenchymal cell)
Fibroblast - Lung	3	CL:0000134 (mesenchymal cell)
Fibroblast - Skin Normal	3	CL:0000134 (mesenchymal cell)
Fibroblast - Cardiac	6	CL:0000134 (mesenchymal cell)
Fibroblast - Peridontal Ligament	6	CL:0000134 (mesenchymal cell)
Keratocytes	3	CL:0000134 (mesenchymal cell)

Group 9		
Meningeal cells	3	CL:0000134 (mesenchymal cell)
Astrocyte - Cerebellum	3	
Astrocyte - Cerebral Cortex	3	
Anulus Pulposus Cells	2	
Schwann Cells	3	
Neurons	3	
Neucleus Pulposus Cells	3	

Group 10	
Sebocyte	3

Group 11	
Amniotic Epithelial Cells	3

Group 12		
Mesenchymal Stem Cells - Adipose	3	CL:0000134 (mesenchymal cell)
Mesenchymal Precursor Cells - Adipose	3	CL:0000134 (mesenchymal cell)
Mesenchymal Precursor Cells - Bone Marrow	3	CL:0000134 (mesenchymal cell)
Multipotent Cord Blood Unrestricted Somatic Stem Cells	2	CL:0000134 (mesenchymal cell)
Mesenchymal Stem Cells - Bone Marrow	4	CL:0000134 (mesenchymal cell)
Neural Stem Cells	2	
Mesenchymal Stem Cells - Hepatic	2	CL:0000134 (mesenchymal cell)
Mesenchymal Stem Cells - Umbilical	3	CL:0000134 (mesenchymal cell)
Mesenchymal Precursor Cells - Cardiac	4	CL:0000134 (mesenchymal cell)

Group 13		
Urothelial Cells	3	
Keratinocyte - Epidermal	3	
Adipocyte - Breast	2	CL:0000134 (mesenchymal cell)
Adipocyte - Perirenal	1	CL:0000134 (mesenchymal cell)
Hair Follicle Outer Root Sheath Cells	2	
Adipocyte - Omental	3	CL:0000134 (mesenchymal cell)
Adipocyte - Subcutaneous	3	CL:0000134 (mesenchymal cell)

Group 14		
Chorionic Membrane Cells	3	CL:0000134 (mesenchymal cell)
Mesenchymal Stem Cells - Amniotic Membrane	2	CL:0000134 (mesenchymal cell)
Amniotic Membrane Cells	3	CL:0000134 (mesenchymal cell)

Group 15	
Perineurial Cells	2

Group 16		
Chrondrocyte - de diff	3	CL:0000134 (mesenchymal cell)
Chrondrocyte - re diff	2	CL:0000134 (mesenchymal cell)

Group 17	
Salivary Acinar Cells	3

Group 18		
CD14 Monocytes - treated with Group A Streptococci	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes derived endothelial progenitor cells	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes - treated with Cryptococcus	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes - treated with BCG	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes - mock treated	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes - treated with Trehalose Dimycolate TDM	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes - treated with B-glucan	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes - treated with Lipolysaccharide	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes - treated with IFN N-hexane	3	
CD14 Monocytes - treated with Salmonella	3	CL:0000134 (mesenchymal cell)
CD14 Monocytes - treated with Candida	3	CL:0000134 (mesenchymal cell)

A.0.1 Samples used from the epigenetics roadmap project

Cell type/ tissue group	EID	Epigenome name
Brain	E071	Brain hippocampus middle
Brain	E068	Brain anterior caudate
Brain	E070	Brain germal matrix
Adipose	E063	Adipose nuclei
Muscle	E108	Skeletal muscle female
Muscle	E100	Psoas muscle
Heart	E105	Right ventricle
Heart	E065	Aorta
Smooth muscle	E078	Duodenum smooth muscle
Smooth muscle	E076	Colon smooth muscle
Smooth muscle	E111	Stomach smooth muscle
Digestive	E109	Small intestine
Digestive	E101	Rectal mucosa donor 29
Digestive	E077	Duodenum mucosa
Digestive	E094	Gastric
Other	E097	Ovary
Other	E087	Pancreatic islets
Other	E091	Placenta
Other	E066	Liver
Other	E098	Pancreas
Other	E096	Lung
Other	E113	Spleen

TABLE A.1: 'Epigenomes' used from the epigenetics roadmap project

Appendix B

Differential expression probabilities

Differential expression probabilities vs log fold change, for each of the 149 primary cell types used in the analysis. The probabilities for each cell type gives are those based on a change between that cell type and all genes of all other cell types.

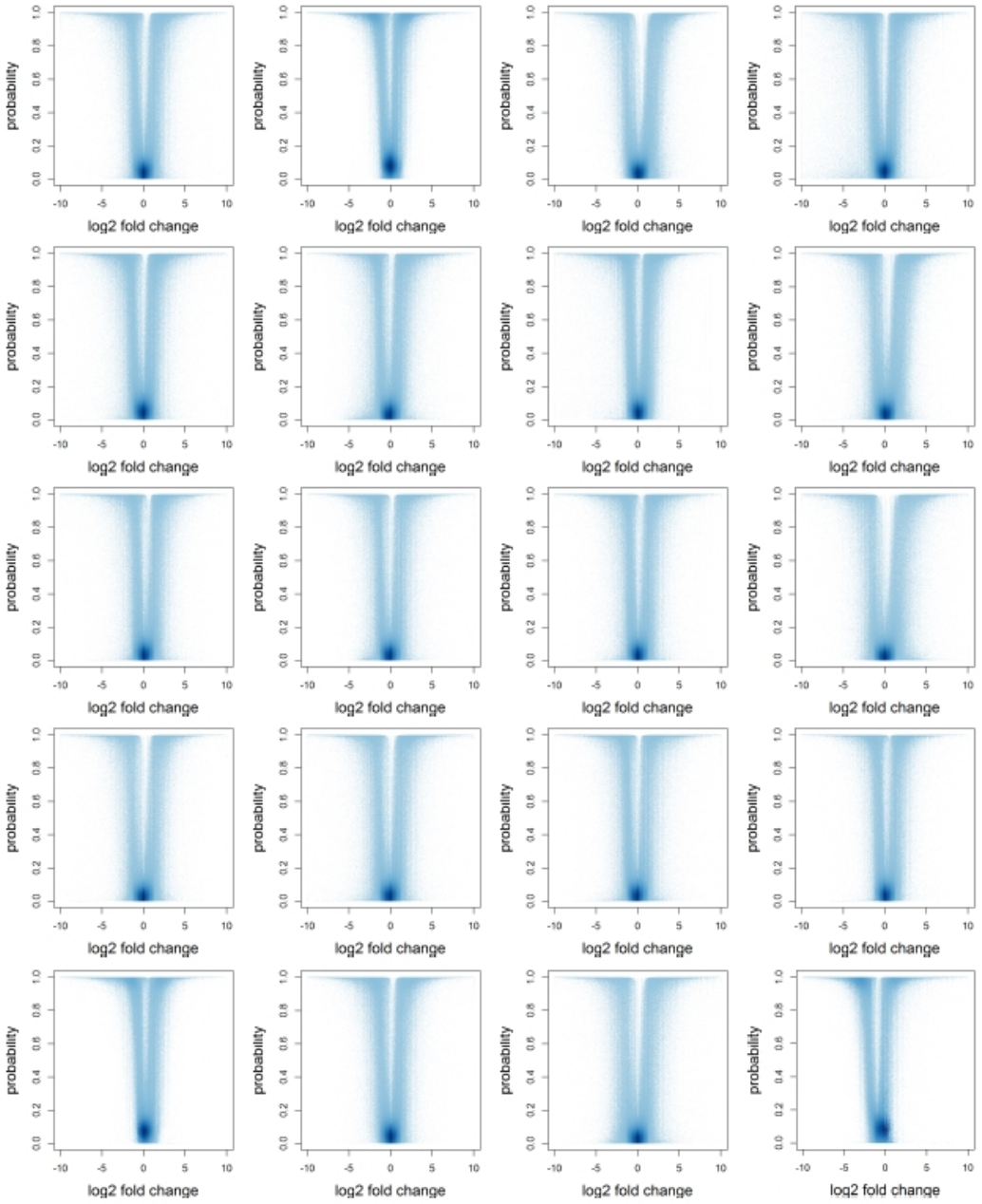


Figure B.1

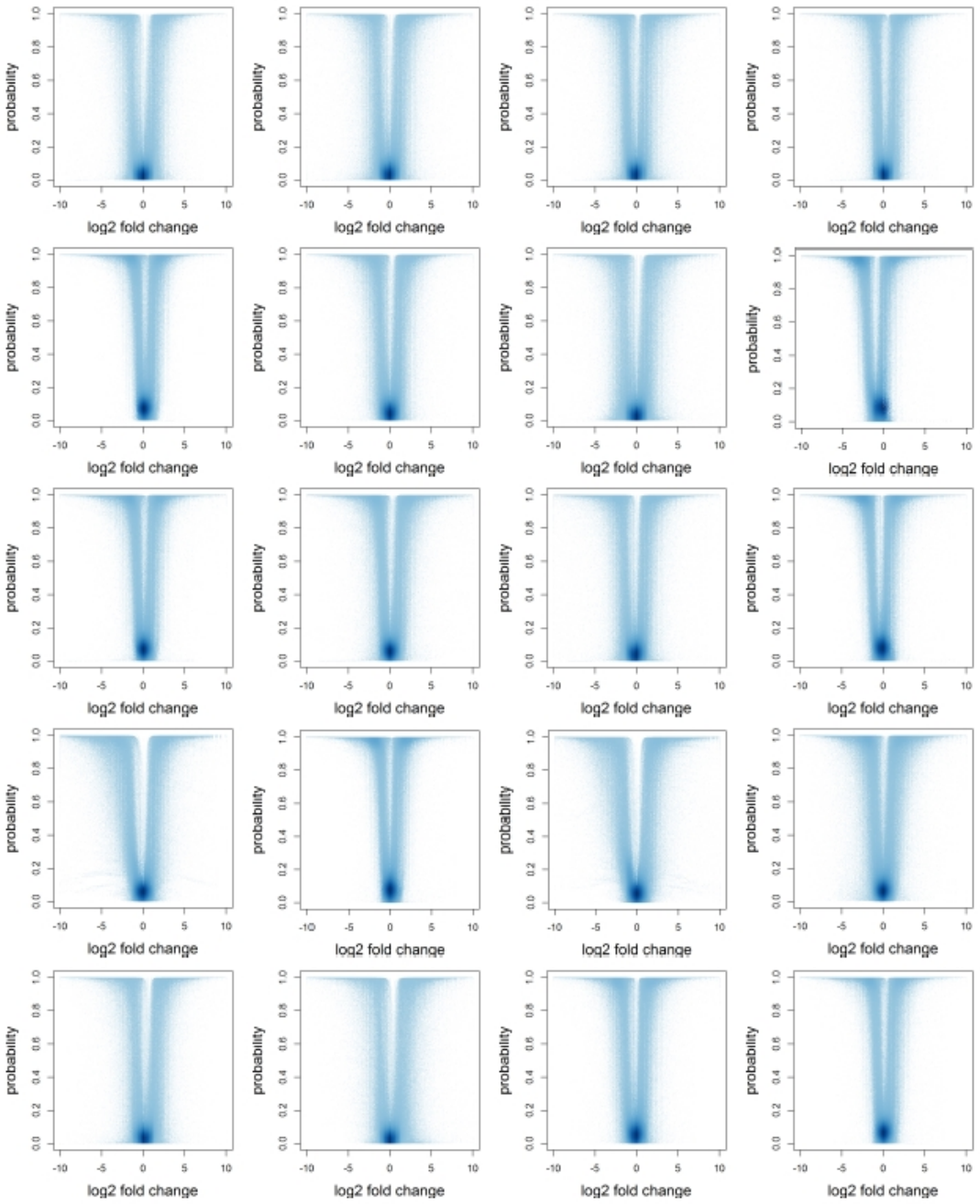


FIGURE B.2

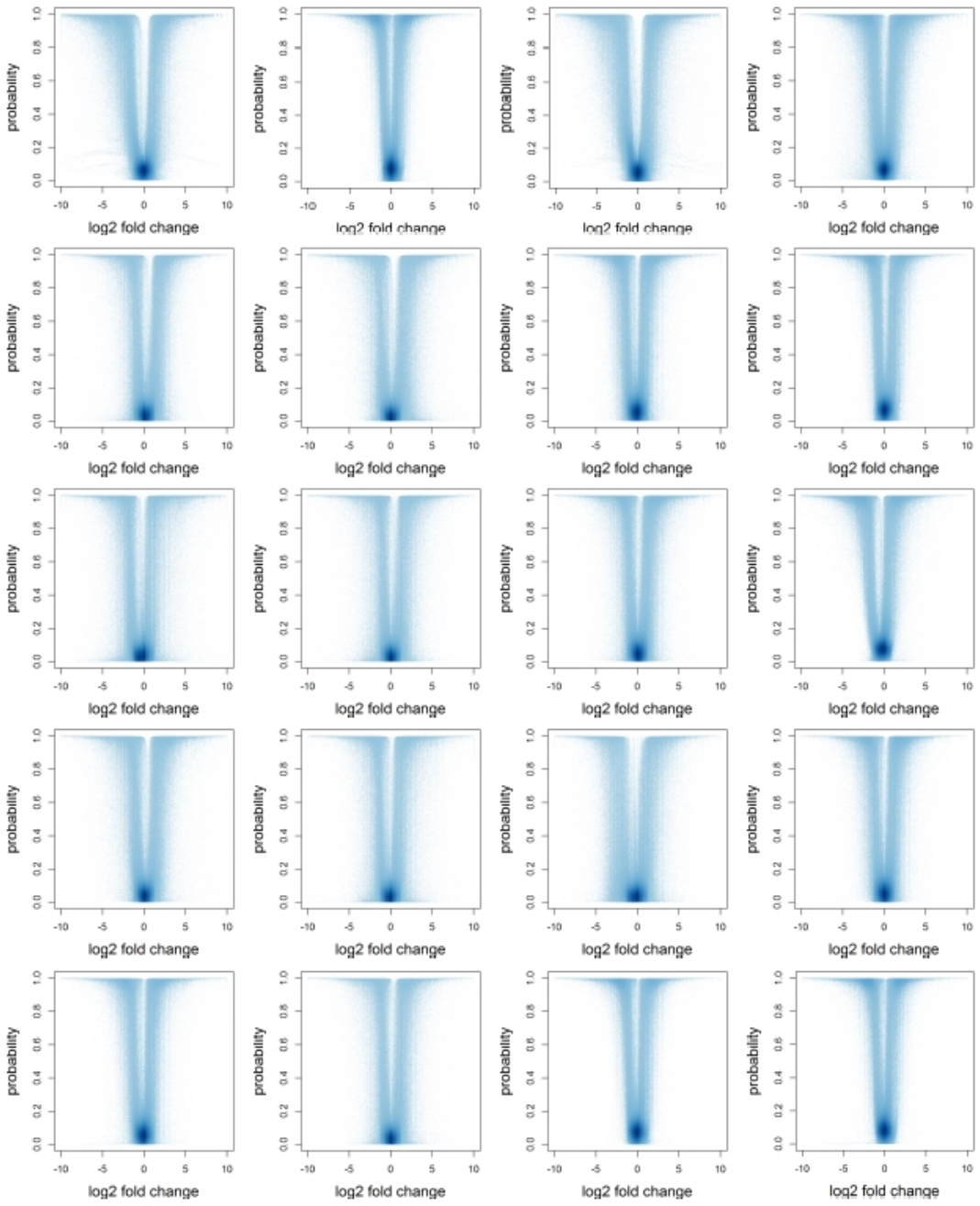


FIGURE B.3

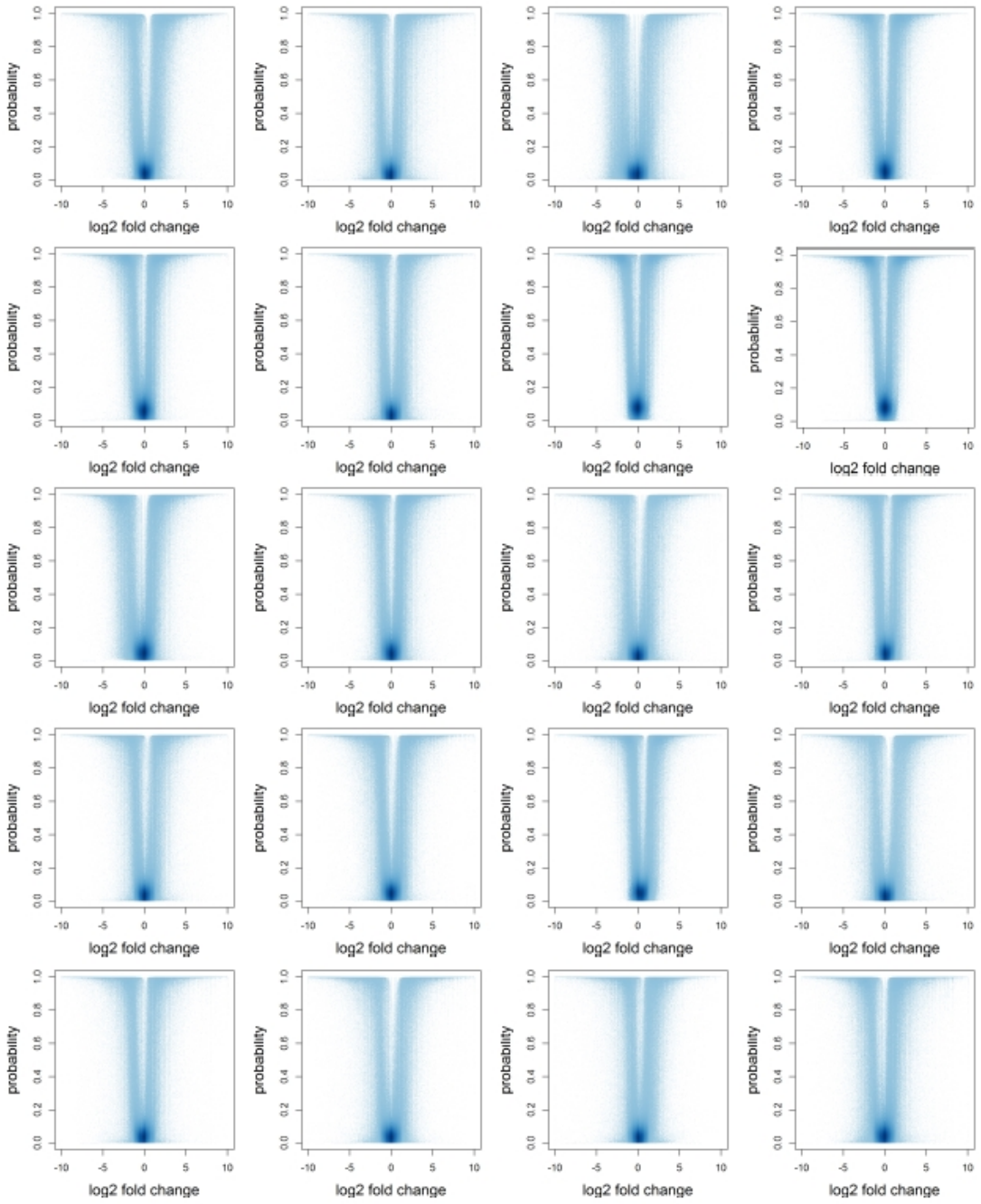


FIGURE B.4

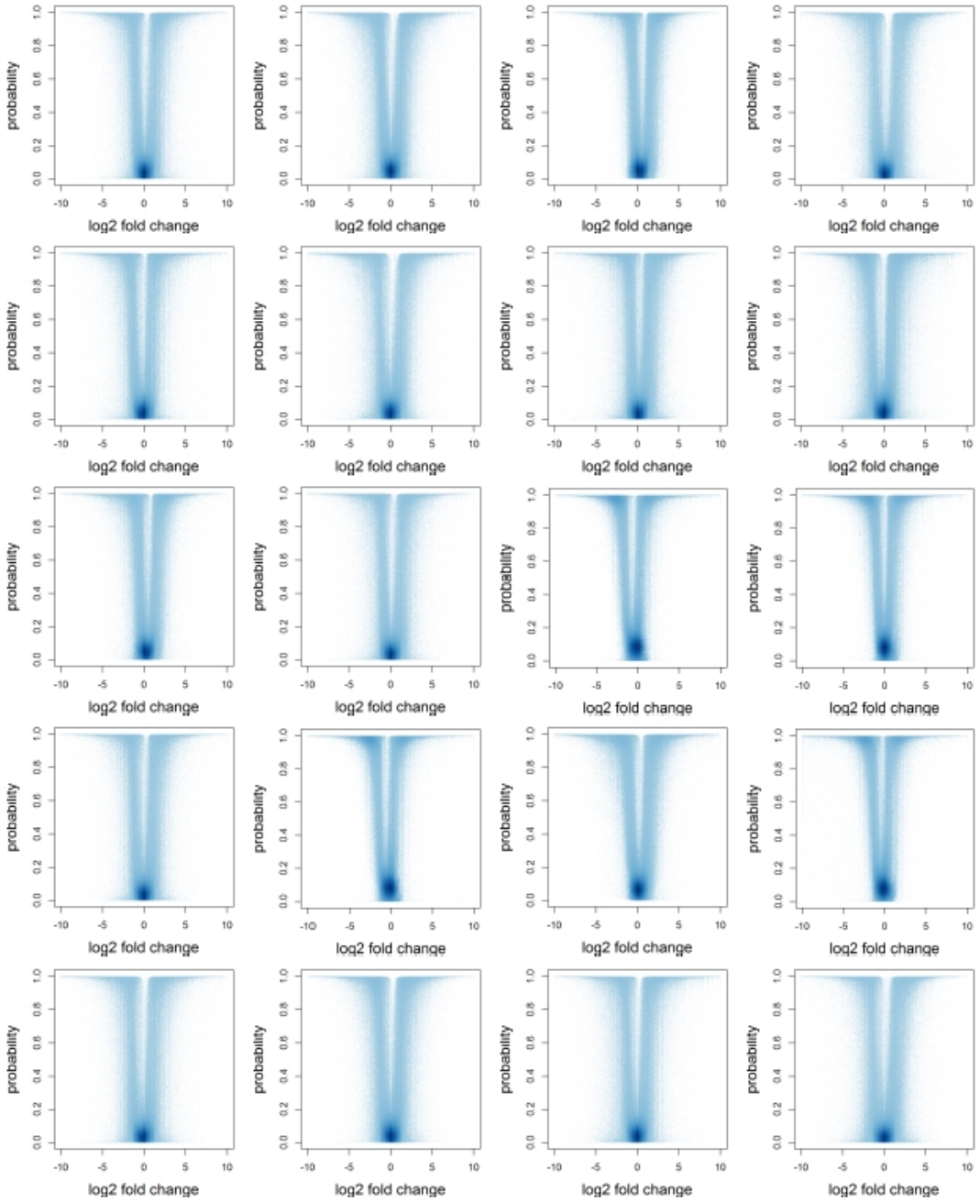


FIGURE B.5

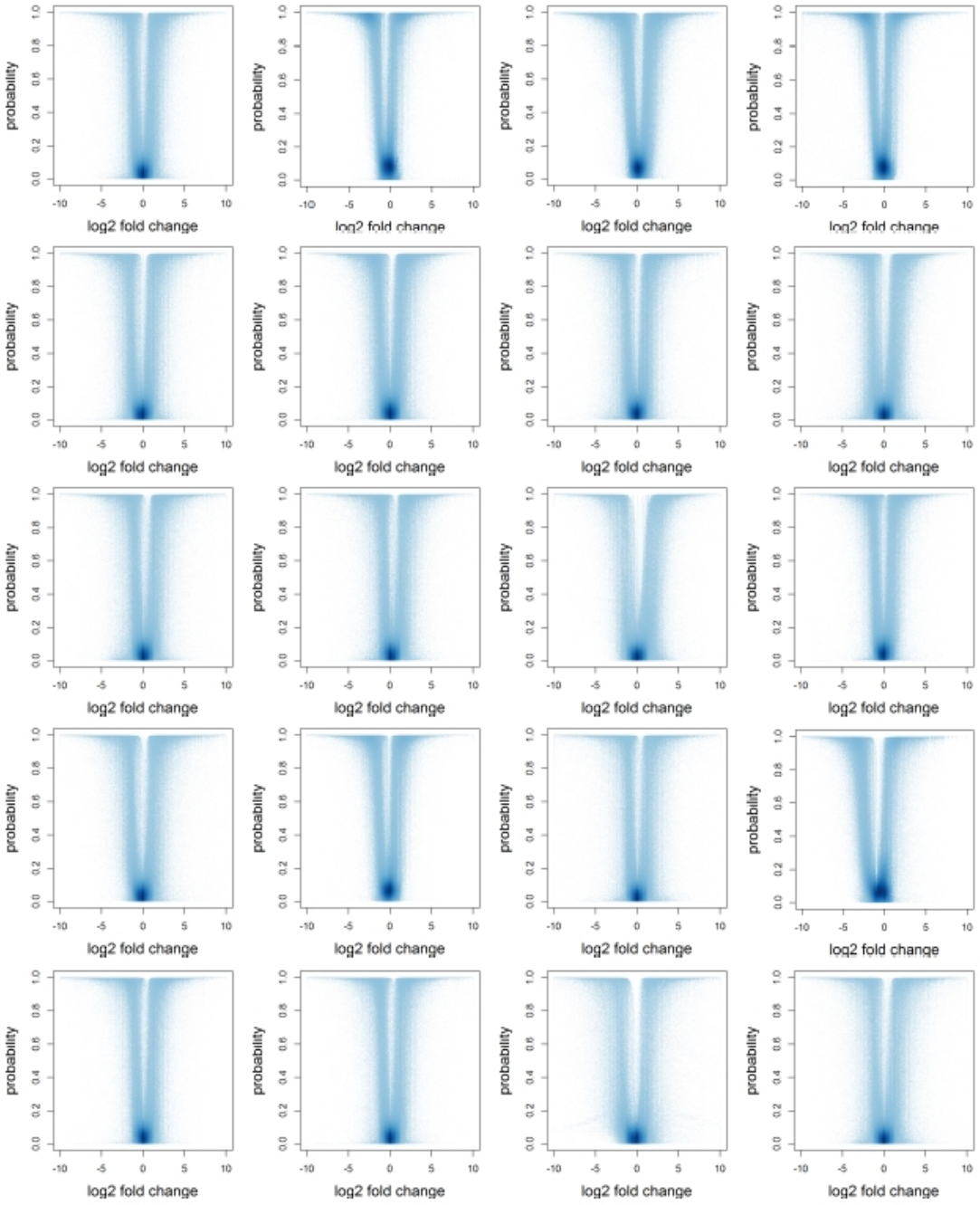


FIGURE B.6

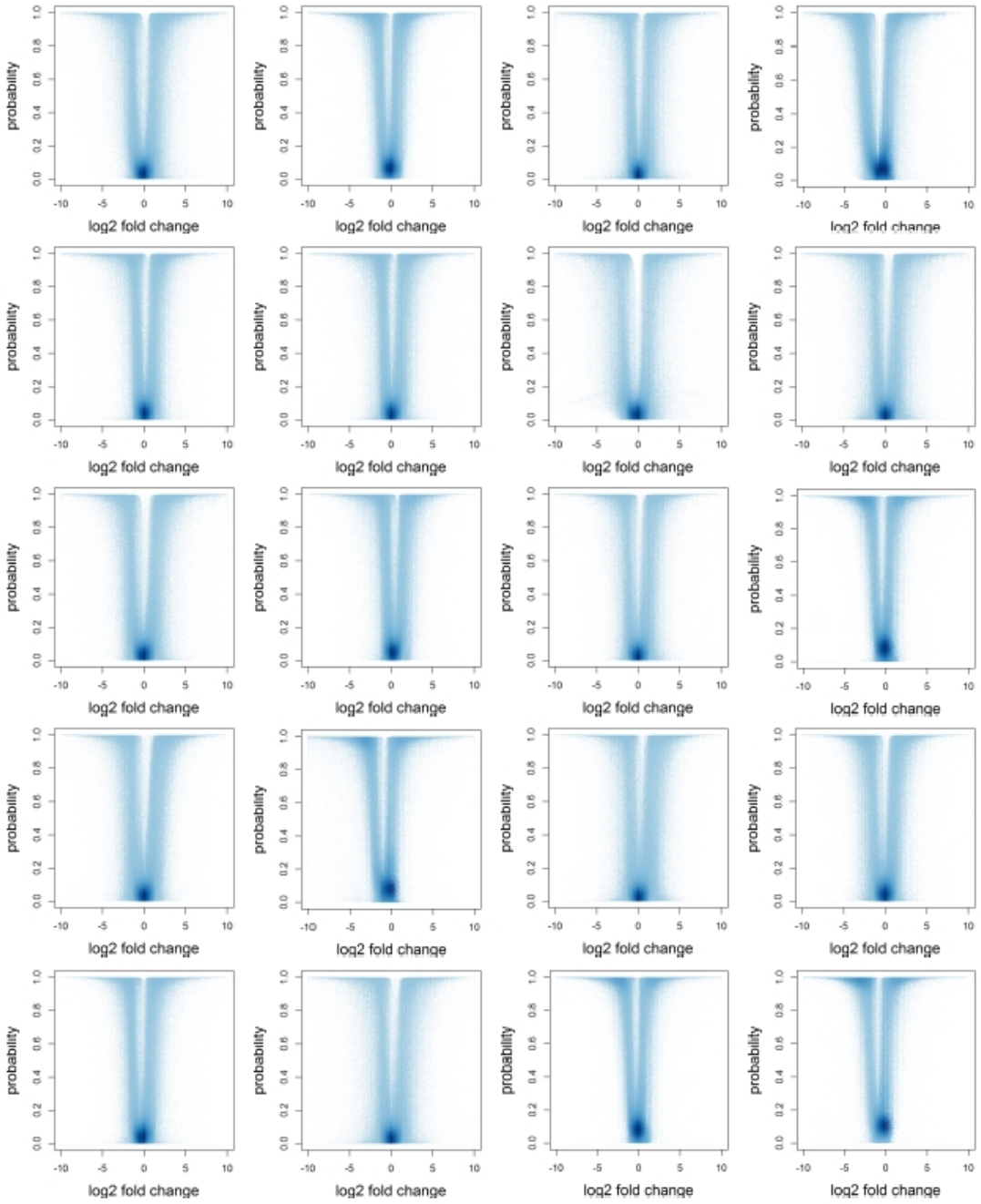


Figure B.7

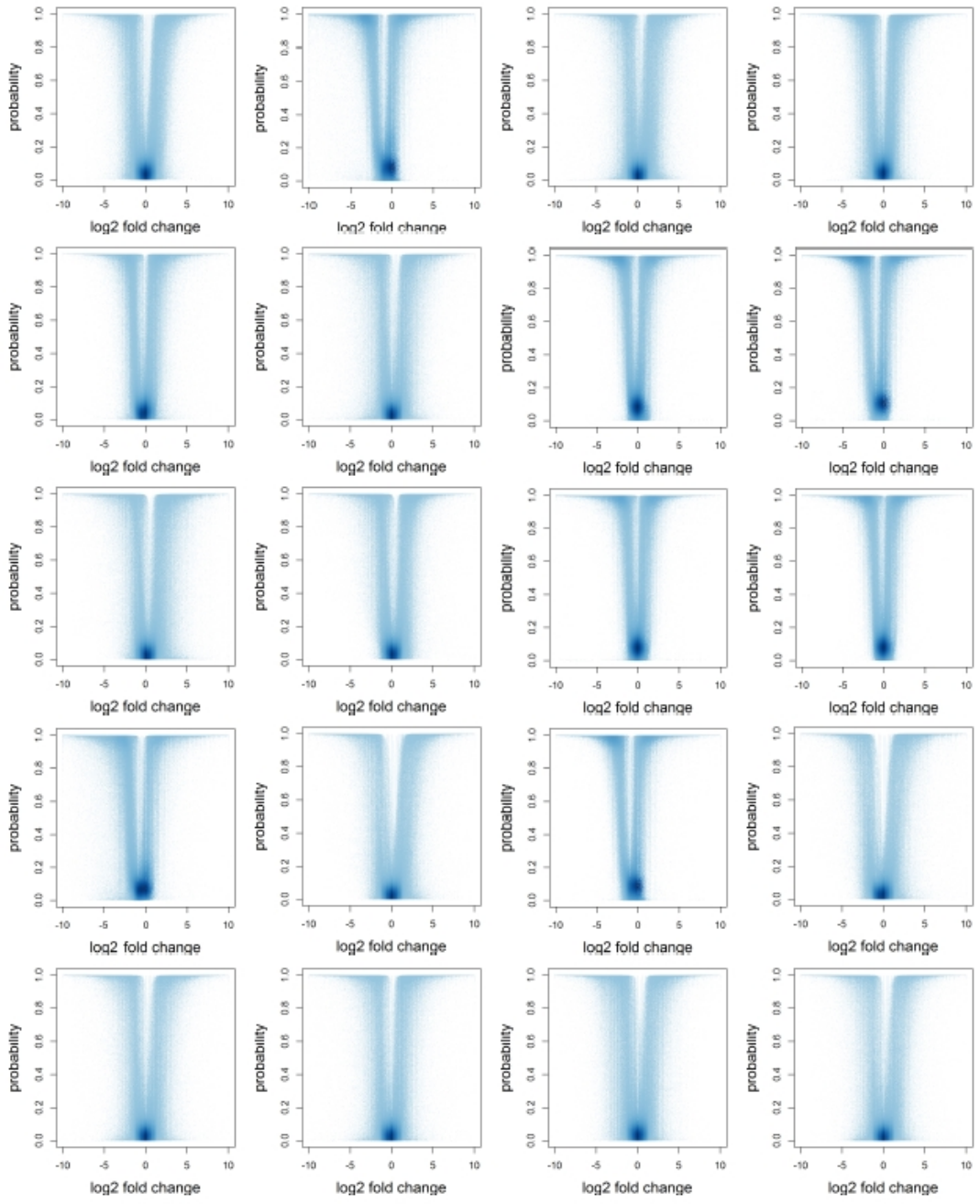


FIGURE B.8

Bibliography

Adachi, N. and Lieber, M. R. (2002). Bidirectional gene organization: a common architectural feature of the human genome. *Cell*, 109(7):807–809.

Adami, C. (2002). What is complexity? *BioEssays*, 24(12):1085–1094.

Äijö, T., Butty, V., Chen, Z., Salo, V., Tripathi, S., Burge, C. B., Lahesmaa, R., and Lähdesmäki, H. (2014). Methods for time series analysis of rna-seq data with application to human th17 cell differentiation. *Bioinformatics*, 30(12):i113–i120.

Albert, F. W. and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*.

Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. *Proceedings of the National Academy of Sciences*, 74(12):5350–5354.

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

Anderson, E. and Hill, R. E. (2014). Long range regulation of the sonic hedgehog gene. *Current opinion in genetics & development*, 27:54–59.

Andersson, R. (2015). Promoter or enhancer, what's the difference? deconstruction of established distinctions and presentation of a unifying model. *BioEssays*, 37(3):314–323.

Andersson, R., Andersen, P. R., Valen, E., Core, L. J., Bornholdt, J., Boyd, M., Jensen, T. H., and Sandelin, A. (2014a). Nuclear stability and transcriptional directionality separate functionally distinct rna species. *Nature communications*, 5.

Andersson, R., Chen, Y., Core, L., Lis, J. T., Sandelin, A., and Jensen, T. H. (2015). Human gene promoters are intrinsically bidirectional. *Molecular cell*, 60(3):346–347.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014b). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Anink-Groenen, L. C., Maarleveld, T. R., Verschure, P. J., and Bruggeman, F. J. (2014). Mechanistic stochastic model of histone modification pattern formation. *Epigenetics & chromatin*, 7(1):30.

Armour, C. D., Castle, J. C., Chen, R., Babak, T., Loerch, P., Jackson, S., Shah, J. K., Dey, J., Rohl, C. A., Johnson, J. M., et al. (2009). Digital transcriptome profiling using selective hexamer priming for cdna synthesis. *Nature methods*, 6(9):647–649.

Arner, E., Daub, C. O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., Lennartsson, A., Rönnerblad, M., Hrydziuszko, O., Vitezic, M., et al. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, 347(6225):1010–1014.

Ayub, M. I., Moosa, M. M., Sarwardi, G., Khan, W., Khan, H., and Yeasmin, S. (2010). Mutation analysis of the hbb gene in selected bangladeshi β-thalassemic individuals: Presence of rare mutations. *Genetic testing and molecular biomarkers*, 14(3):299–302.

Ballaré, C., Castellano, G., Gaveglia, L., Althammer, S., González-Vallinas, J., Eyras, E., Le Dily, F., Zaurin, R., Soronellas, D., Vicent, G. P., et al. (2013). Nucleosome-driven transcription factor binding and gene regulation. *Molecular cell*, 49(1):67–79.

Bandt, C. and Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102.

Barolo, S. (2012). Shadow enhancers: Frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *Bioessays*, 34(2):135–141.

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

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome research*, 24(1):14–24.

Benevolenskaya, E. V. (2007). Histone h3k4 demethylases are essential in development and differentiation this paper is one of a selection of papers published in this special issue, entitled 28th international west coast chromatin and chromosome conference, and has undergone the journal's usual peer review process. *Biochemistry and cell biology*, 85(4):435–443.

Bentley, D. L. (2014). Coupling mrna processing with transcription in time and space. *Nature Reviews Genetics*, 15(3):163–175.

Bernstein, B. E., Meissner, A., and Lander, E. S. (2007). The mammalian epigenome. *Cell*, 128(4):669–681.

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., and Tanzi, R. E. (2007). Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database. *Nature genetics*, 39(1):17–23.

Bettens, K., Sleegers, K., and Van Broeckhoven, C. (2013). Genetic insights in alzheimer's disease. *The Lancet Neurology*, 12(1):92–104.

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

Bothma, J. P., Garcia, H. G., Ng, S., Perry, M. W., Gregor, T., and Levine, M. (2015). Enhancer additivity and non-additivity are determined by enhancer strength in the drosophila embryo. *Elife*, 4:e07956.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich,

P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348.

Brett, D., Pospisil, H., Valcárcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. *Nature genetics*, 30(1):29–30.

Bryne, J. C., Valen, E., Tang, M.-H. E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research*, 36(suppl 1):D102–D106.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). Atac-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, pages 21–29.

Bulger, M. and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144(3):327–339.

Burgess, D. J. (2012). Gene expression: More roles and details for polymerase pausing. *Nature Reviews Genetics*, 13(7):450–451.

Byszewska, M., Śmietański, M., Purta, E., and Bujnicki, J. M. (2014). Rna methyltransferases involved in 5′ cap biosynthesis. *RNA biology*, 11(12):1597–1607.

Campos, E. I. and Reinberg, D. (2009). Histones: annotating chromatin. *Annual review of genetics*, 43:559–599.

Capra, J. A., Williams, A. G., and Pollard, K. S. (2012). Proteinhistorian: tools for the comparative analysis of eukaryote protein origin. *PLoS computational biology*, 8(6):e1002567.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., et al. (1996). High-efficiency full-length cdna cloning by biotinylated cap trapper. *Genomics*, 37(3):327–336.

Chen, C.-Y., Chen, S.-T., Juan, H.-F., and Huang, H.-C. (2012). Lengthening of 3′ utr increases with morphological complexity in animal evolution. *Bioinformatics*, 28(24):3178–3181.

Chen, T. and Dent, S. Y. (2014). Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nature Reviews Genetics*, 15(2):93–106.

Chen, Y.-C., Cheng, J.-H., Tsai, Z. T.-Y., Tsai, H.-K., and Chuang, T.-J. (2013). The impact of trans-regulation on the evolutionary rates of metazoan proteins. *Nucleic acids research*, 41(13):6371–6380.

Chevan, A. and Sutherland, M. (1991). Hierarchical partitioning. *The American Statistician*, 45(2):90–96.

Chung, L. M., Ferguson, J. P., Zheng, W., Qian, F., Bruno, V., Montgomery, R. R., and Zhao, H. (2013). Differential expression analysis for paired rna-seq data. *BMC bioinformatics*, 14(1):110.

Churchman, L. S. and Weissman, J. S. (2012). Native elongating transcript sequencing (net-seq). *Current Protocols in Molecular Biology*, pages 14–4.

Colgan, D. F. and Manley, J. L. (1997). Mechanism and regulation of mrna polyadenylation. *Genes & development*, 11(21):2755–2766.

Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, 15(7):901–913.

Corder, E., Saunders, A., Strittmatter, W., Schmechel, D., Gaskell, P., Small, G., Roses, A., Haines, J., and Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. *Science*, 261(5123):921–923.

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., and Lis, J. T. (2014). Analysis of nascent rna identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics*, 46(12):1311–1320.

Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848.

Corvelo, A. and Eyras, E. (2008). Exon creation and establishment in human genes. *Genome Biol*, 9(9):R141.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al. (2010). Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al. (2014). Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*.

Dambacher, S., Hahn, M., and Schotta, G. (2010). Epigenetic regulation of development by histone lysine methylation. *Heredity*, 105(1):24–37.

de Hoon, M. and Hayashizaki, Y. (2008). Deep cap analysis gene expression (cage): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques*, 44(5):627.

de Klerk, E., den Dunnen, J. T., and AC't Hoen, P. (2014). Rna sequencing: from tag-based profiling to resolving complete transcript structure. *Cellular and Molecular Life Sciences*, 71(18):3537–3551.

de Mendoza, A., Suga, H., Permanyer, J., Irimia, M., and Ruiz-Trillo, I. (2016). Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. *eLife*, 4:e08904.

Deaton, A. M. and Bird, A. (2011). Cpg islands and the regulation of transcription. *Genes & development*, 25(10):1010–1022.

Decker, K. B. and Hinton, D. M. (2013). Transcription regulation at the core: similarities among bacterial, archaeal, and eukaryotic rna polymerases. *Annual review of microbiology*, 67:113–139.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., Dean, A., and Blobel, G. A. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6):1233–1244.

Dermitzakis, E. T. and Clark, A. G. (2002). Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Molecular biology and evolution*, 19(7):1114–1121.

Di Croce, L. and Helin, K. (2013). Transcriptional regulation by polycomb group proteins. *Nature structural & molecular biology*, 20(10):1147–1155.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Arcelus, M. G., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science*, 325(5945):1246–1250.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

Dowell, R. D. (2010). Transcription factor binding variation in the evolution of gene regulation. *Trends in genetics*, 26(11):468–475.

Doyle, B., Fudenberg, G., Imakaev, M., and Mirny, L. A. (2014). Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS computational biology*, 10(10):e1003867.

Dreszer, T. R., Karolchik, D., Zweig, A. S., Hinrichs, A. S., Raney, B. J., Kuhn, R. M., Meyer, L. R., Wong, M., Sloan, C. A., Rosenbloom, K. R., et al. (2012). The ucsc genome browser database: extensions and updates 2011. *Nucleic acids research*, 40(D1):D918–D923.

Dupont, C. A., Dardalhon-Cuménal, D., Kyba, M., Brock, H. W., Randsholt, N. B., and Peronnet, F. (2015). Drosophila cyclin g and epigenetic maintenance of gene expression during development. *Epigenetics & chromatin*, 8(1):1–17.

Duttke, S. H., Lacadie, S. A., Ibrahim, M. M., Glass, C. K., Corcoran, D. L., Benner, C., Heinz, S., Kadonaga, J. T., and Ohler, U. (2015). Human promoters are intrinsically directional. *Molecular cell*, 57(4):674–684.

Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). Discovering motifs in ranked lists of dna sequences. *PLoS computational biology*, 3(3):e39.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10(1):48.

Elango, N. and Soojin, V. Y. (2011). Functional relevance of cpg island length for regulation of gene expression. *Genetics*, 187(4):1077–1083.

Elgar, G. (2009). Pan-vertebrate conserved non-coding sequences associated with developmental regulation. *Briefings in functional genomics & proteomics*, 8(4):256–265.

Elkon, R., Ugalde, A. P., and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics*, 14(7):496–506.

Elnitski, L., Hardison, R. C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M. J., Schwartz, S., Miller, W., and Chiaromonte, F. (2003). Distinguishing regulatory dna from neutral sites. *Genome research*, 13(1):64–72.

Epstein, D. J. (2009). Cis-regulatory mutations in human disease. *Briefings in functional genomics & proteomics*, 8(4):310–316.

Esteller, M. (2007). Epigenetic gene silencing in cancer: the dna hypermethylome. *Human molecular genetics*, 16(R1):R50–R59.

Faghihi, M. A. and Wahlestedt, C. (2009). Regulatory roles of natural antisense transcripts. *Nature reviews Molecular cell biology*, 10(9):637–643.

Fairclough, S. R., Chen, Z., Kramer, E., Zeng, Q., Young, S., Robertson, H. M., Begovic, E., Richter, D. J., Russ, C., Westbrook, M. J., et al. (2013). Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate salpingoeca rosetta. *Genome Biol*, 14(2):R15.

Fatemi, M., Pao, M. M., Jeong, S., Gal-Yam, E. N., Egger, G., Weisenberger, D. J., and Jones, P. A. (2005). Footprinting of mammalian promoters: use of a cpg dna methyltransferase revealing nucleosome positions at a single molecule level. *Nucleic acids research*, 33(20):e176–e176.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2013). Ensembl 2014. *Nucleic acids research*, page gkt1196.

Forbes, S. A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C. Y., Jia, M., Ewing, R., Menzies, A., et al. (2009). Cosmic (the catalogue of somatic mutations in cancer): a resource to investigate acquired mutations in human cancer. *Nucleic acids research*, page gkp995.

Forrest, A., Consortium, T. F., et al. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.

Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C., Aitken, S., et al. (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular systems biology*, 11(12):852.

Frith, M. C. et al. (2014). Explaining the correlations among properties of mammalian promoters. *Nucleic acids research*, 42(8):4823–4832.

Frith, M. C., Ponjavic, J., Fredman, D., Kai, C., Kawai, J., Carninci, P., Hayshizaki, Y., and Sandelin, A. (2006). Evolutionary turnover of mammalian transcription start sites. *Genome research*, 16(6):713–722.

Gaiti, F., Fernandez-Valverde, S. L., Nakanishi, N., Calcino, A. D., Yanai, I., Tanurdzic, M., and Degnan, B. M. (2015). Dynamic and widespread lncrna expression in a sponge and the origin of animal complexity. *Molecular biology and evolution*, 32(9):2367–2382.

Galanello, R. and Origa, R. (2010). Review: beta-thalassemia. *Orphanet J Rare Dis*, 5(11).

Gardiner-Garden, M. and Frommer, M. (1987). Cpg islands in vertebrate genomes. *Journal of molecular biology*, 196(2):261–282.

Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E. E. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*.

Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P., and White, K. P. (2006). Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature*, 440(7081):242–245.

Gilchrist, D. A. and Adelman, K. (2012). Coupling polymerase pausing and chromatin landscapes for precise regulation of transcription. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(7):700–706.

Gilchrist, D. A., Dos Santos, G., Fargo, D. C., Xie, B., Gao, Y., Li, L., and Adelman, K. (2010). Pausing of rna polymerase ii disrupts dna-specified nucleosome organization to enable precise gene regulation. *Cell*, 143(4):540–551.

Gingeras, T. R. (2009). Implications of chimaeric non-co-linear transcripts. *Nature*, 461(7261):206–211.

Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, 128(4):635–638.

Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A., and Carroll, S. B. (2005). Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in drosophila. *Nature*, 433(7025):481–487.

González, A. J., Setty, M., and Leslie, C. S. (2015). Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nature genetics*, 47(11):1249–1259.

Goodrich, J. A. and Tjian, R. (2010). Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nature Reviews Genetics*, 11(8):549–558.

Gorlova, O., Fedorov, A., Logothetis, C., Amos, C., and Gorlov, I. (2014). Genes with a large intronic burden show greater evolutionary conservation on the protein level. *BMC evolutionary biology*, 14(1):50.

Gregory, T. (2001). Coincidence, coevolution, or causation? dna content, cellsize, and the c-value enigma. *Biological Reviews*, 76(1):65–101.

Gulko, B., Hubisz, M. J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature genetics*, 47(3):276–283.

Haberle, V., Li, N., Hadzhiev, Y., Plessy, C., Previti, C., Nepal, C., Gehrig, J., Dong, X., Akalin, A., Suzuki, A. M., et al. (2014). Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*.

Hackett, J. A., Reddington, J. P., Nestor, C. E., Dunican, D. S., Branco, M. R., Reichmann, J., Reik, W., Surani, M. A., Adams, I. R., and Meehan, R. R. (2012). Promoter dna methylation couples genome-defence mechanisms to epigenetic reprogramming in the mouse germline. *Development*, 139(19):3623–3632.

Hahn, M. A., Wu, X., Li, A. X., Hahn, T., and Pfeifer, G. P. (2011). Relationship between gene body dna methylation and intragenic h3k9me3 and h3k36me3 chromatin marks. *PLoS One*, 6(4):e18844.

Hahn, M. W., Wray, G. A., et al. (2002). The g-value paradox. *Evolution and Development*, 4(2):73–75.

Hao, L., Ge, X., Wan, H., Hu, S., Lercher, M. J., Yu, J., and Chen, W.-H. (2010). Human functional genetic studies are biased against the medically most relevant primate-specific genes. *BMC evolutionary biology*, 10(1):316.

Hardcastle, T. J. and Kelly, K. A. (2010). bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1):422.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., Pahwa, J. S., Moskvina, V., Dowzell, K., Williams, A., et al. (2013). Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease. *Nature genetics*, 45(6):712–712.

He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N., and Kinzler, K. W. (2008). The antisense transcriptomes of human cells. *Science*, 322(5909):1855–1857.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112.

Hill, R. E. and Lettice, L. A. (2013). Alterations to the remote control of shh gene expression cause congenital abnormalities. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 368(1620):20120357.

Horvath, S. (2013). Dna methylation age of human tissues and cell types. *Genome biology*, 14(10):R115.

Hoyle, D. C., Rattray, M., Jupp, R., and Brass, A. (2002). Making sense of microarray data distributions. *Bioinformatics*, 18(4):576–584.

Hu, M. and Polyak, K. (2006). Serial analysis of gene expression. *Nature protocols*, 1(4):1743–1760.

Hughes, A. L. and Rando, O. J. (2014). Mechanisms underlying nucleosome positioning in vivo. *Annual review of biophysics*, 43:41–63.

Hume, D. A. (2012). Plenary perspective: The complexity of constitutive and inducible gene expression in mononuclear phagocytes. *Journal of Leukocyte Biology*, 92(3):433–444.

Hussey, S. G., Mizrachi, E., Groover, A., Berger, D. K., and Myburg, A. A. (2015). Genome-wide mapping of histone h3 lysine 4 trimethylation in eucalyptus grandis developing xylem. *BMC Plant Biology*, 15(1):117.

Iwafuchi-Doi, M. and Zaret, K. S. (2014). Pioneer transcription factors in cell reprogramming. *Genes & development*, 28(24):2679–2692.

Jabbari, K. and Bernardi, G. (2004). Cytosine methylation and cpg, tpg (cpa) and tpa frequencies. *Gene*, 333:143–149.

Jacox, E., Gotea, V., Ovcharenko, I., and Elnitski, L. (2010). Tissue-specific and ubiquitous expression patterns from alternative promoters of human genes.

Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33:245–254.

Jay, G. S. (1996). Full house: The spread of excellence from plato to darwin.

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

Jin, S., Tan, R., Jiang, Q., Xu, L., Peng, J., Wang, Y., and Wang, Y. (2014). A generalized topological entropy for analyzing the complexity of dna sequences. *PloS one*, 9(2):e88519.

John, S., Sabo, P. J., Thurman, R. E., Sung, M.-H., Biddie, S. C., Johnson, T. A., Hager, G. L., and Stamatoyannopoulos, J. A. (2011). Chromatin accessibility predetermines glucocorticoid receptor binding patterns. *Nature genetics*, 43(3):264–268.

Jones, B. (2015). Gene expression: Layers of gene regulation. *Nature Reviews Genetics*, 16(3):128–129.

Jonkers, I. and Lis, J. T. (2015). Getting up to speed with transcription elongation by rna polymerase ii. *Nature Reviews Molecular Cell Biology*, 16(3):167–177.

Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., Bertin, N., Kaiho, A., Ninomiya, N., Daub, C. O., et al. (2011). Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome research*, 21(7):1150–1159.

Kechavarzi, B. and Janga, S. C. (2014). Dissecting the expression landscape of rna-binding proteins in human cancers. *Genome Biol*, 15(1):R14.

Keung, A. J., Joung, J. K., Khalil, A. S., and Collins, J. J. (2015). Chromatin regulation at the frontier of synthetic biology. *Nature Reviews Genetics*, 16(3):159–171.

Khaitovich, P., Enard, W., Lachmann, M., and Pääbo, S. (2006). Evolution of primate gene expression. *Nature Reviews Genetics*, 7(9):693–702.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187.

Kindt, A. S., Navarro, P., Semple, C. A., and Haley, C. S. (2013). The genomic signature of trait-associated variants. *BMC genomics*, 14(1):108.

King, M.-C., Wilson, A. C., et al. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116.

Kleinjan, D. A. and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *The American Journal of Human Genetics*, 76(1):8–32.

Koch, C. M., Andrews, R. M., Flicek, P., Dillon, S. C., Karaöz, U., Clelland, G. K., Wilcox, S., Beare, D. M., Fowler, J. C., Couttet, P., et al. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome research*, 17(6):691–707.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., et al. (2006). Cage: cap analysis of gene expression. *Nature methods*, 3(3):211–222.

Koenker, R. (2013). Quantreg: quantile regression. *R package version*, 5.

Köhler, A. et al. (2007). Exporting rna from the nucleus to the cytoplasm. *Nature Reviews Molecular Cell Biology*, 8(10):761–773.

Kolmogorov, A. (1963). Theory of transmission of information. *Amer. Math. Soc. Translation, Ser*, 2(33):291–321.

Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R., Papantonis, A., et al. (2012). Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin*, 5(1):1–1.

Kong, J. and Lasko, P. (2012). Translational control in cellular and developmental processes. *Nature Reviews Genetics*, 13(6):383–394.

Ku, M., Koche, R. P., Rheinbay, E., Mendenhall, E. M., Endoh, M., Mikkelsen, T. S., Presser, A., Nusbaum, C., Xie, X., Chi, A. S., et al. (2008). Genomewide analysis of prc1 and prc2 occupancy identifies two classes of bivalent domains. *PLoS genetics*, 4(10):e1000242.

Ku, W. L., Girvan, M., Yuan, G.-C., Sorrentino, F., and Ott, E. (2013). Modeling the dynamics of bivalent histone modifications. *PloS one*, 8(11):e77944.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Kurdistani, S. K. and Grunstein, M. (2003). Histone acetylation and deacetylation in yeast. *Nature reviews Molecular cell biology*, 4(4):276–284.

Kwak, H., Fuda, N. J., Core, L. J., and Lis, J. T. (2013). Precise maps of rna polymerase reveal how promoters direct initiation and pausing. *Science*, 339(6122):950–953.

Landauer, R. (1988). A simple measure of complexity. *Nature*, 336:306–307.

Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W., and Hartl, D. L. (2007). Genetic properties influencing the evolvability of gene expression. *Science*, 317(5834):118–121.

Lane, N. and Martin, W. (2010). The energetics of genome complexity. *Nature*, 467(7318):929–934.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences*, 94(24):13057–13062.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, 15(2):R29.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput Biol*, 9(8):e1003118.

Lee, J., Ji, Y., Liang, S., Cai, G., and Müller, P. (2011). On differential gene expression using rna-seq data. *Cancer informatics*, 10:205.

Lee, J.-H., Tate, C. M., You, J.-S., and Skalnik, D. G. (2007). Identification and characterization of the human set1b histone h3-lys4 methyltransferase complex. *Journal of Biological Chemistry*, 282(18):13419–13428.

Lee, N., Iyer, S. S., Mu, J., Weissman, J. D., Ohali, A., Howcroft, T. K., Lewis, B. A., and Singer, D. S. (2010). Three novel downstream promoter elements regulate mhc class i promoter activity in mammalian cells. *PloS one*, 5(12):e15278.

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., Isono, K.-i., et al. (2006). Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, 125(2):301–313.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, 13(4):233–245.

Lesch, B. J. and Page, D. C. (2014). Poised chromatin in the mammalian germ line. *Development*, 141(19):3619–3626.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics*, 12(14):1725–1735.

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945):147–151.

Levings, P. P. and Bungert, J. (2002). The human β-globin locus control region. *European Journal of Biochemistry*, 269(6):1589–1599.

Levo, M. and Segal, E. (2014). In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics*, 15(7):453–468.

Li, B., Carey, M., and Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, 128(4):707–719.

Li, E., Bestor, T. H., and Jaenisch, R. (1992). Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926.

Li, J. J. and Biggin, M. D. (2015). Statistics requantitates the central dogma. *Science*, 347(6226):1066–1067.

Li, R., Mav, D., Grimm, S. A., Jothi, R., Shah, R., and Wade, P. A. (2014). Fine-tuning of epigenetic regulation with respect to promoter cpg content in a cell type-specific manner. *epigenetics*, 9(5):747–759.

Liao, B.-Y. and Weng, M.-P. (2012). Natural selection drives rapid evolution of mouse embryonic heart enhancers. *BMC systems biology*, 6(Suppl 2):S1.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293.

Lin, H.-T. and Li, L. (2005). Analysis of sage results with combined learning techniques. *Analysis*, (1/23).

Lis, J. (1998). Promoter-associated pausing in promoter architecture and postinitiation transcriptional regulation. In *Cold Spring Harbor symposia on quantitative biology*, volume 63, pages 347–356. Cold Spring Harbor Laboratory Press.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., et al. (2013). Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146):1237905.

Liu, R., Holik, A. Z., Su, S., Jansz, N., Chen, K., San Leong, H., Blewitt, M. E., Asselin-Labat, M.-L., Smyth, G. K., and Ritchie, M. E. (2015). Why weight? modelling sample and observational level variability improves power in rna-seq analyses. *Nucleic acids research*, 43(15):e97–e97.

Liu, Y., Zhou, J., and White, K. P. (2014). Rna-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–304.

Lopez-Ruiz, R., Mancini, H. L., and Calbet, X. (1995). A statistical measure of complexity. *Physics Letters A*, 209(5):321–326.

Lowe, C. B., Kellis, M., Siepel, A., Raney, B. J., Clamp, M., Salama, S. R., Kingsley, D. M., Lindblad-Toh, K., and Haussler, D. (2011). Three periods of regulatory innovation during vertebrate evolution. *Science*, 333(6045):1019–1024.

Luo, W., Hankenson, K. D., and Woolf, P. J. (2008). Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. *BMC bioinformatics*, 9(1):467.

Luse, D. S. (2013). Promoter clearance by rna polymerase ii. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1829(1):63–68.

Margueron, R. and Reinberg, D. (2011). The polycomb complex prc2 and its mark in life. *Nature*, 469(7330):343–349.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517.

Martínez, O. and Reyes-Valdés, M. H. (2008). Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *Proceedings of the National Academy of Sciences*, 105(28):9709–9714.

Mathis, D. J. and Chambon, P. (1981). The sv40 early region tata box is required for accurate in vitro initiation of transcription. *Nature*, 290(5804):310–315.

McEwen, G. K., Goode, D. K., Parker, H. J., Woolfe, A., Callaway, H., and Elgar, G. (2009a). Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS genetics*, 5(12):e1000762.

McEwen, G. K., Goode, D. K., Parker, H. J., Woolfe, A., Callaway, H., and Elgar, G. (2009b). Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS genetics*, 5(12):e1000762.

McShea, D. W. (1996). Perspective: Metazoan complexity and evolution: is there a trend? *Evolution*, pages 477–492.

Meader, S., Ponting, C. P., and Lunter, G. (2010). Massive turnover of functional sequence in human and other mammalian genomes. *Genome research*, 20(10):1335–1343.

Meselson, M. and Stahl, F. W. (1958). The replication of dna in escherichia coli. *Proceedings of the national academy of sciences*, 44(7):671–682.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560.

Montavon, T. and Duboule, D. (2013). Chromatin organization and global regulation of hox gene clusters. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120367.

Montgomery, S. B. and Dermitzakis, E. T. (2011). From expression qtls to personalized transcriptomics. *Nature Reviews Genetics*, 12(4):277–282.

Moore, M. J. (2005). From birth to death: the complex lives of eukaryotic mrnas. *Science*, 309(5740):1514–1518.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628.

Müller, F. and Tora, L. (2014). Chromatin and dna sequences in defining promoters for transcription initiation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1839(3):118–128.

Nagel, D. H. and Kay, S. A. (2012). Complexity in the Wiring and Regulation of Plant Circadian Networks. *Current Biology*, 22(16):R648–R657.

Nelson, A. C. and Wardle, F. C. (2013). Conserved non-coding elements and cis regulation: actions speak louder than words. *Development*, 140(7):1385–1395.

Nielsen, K. L., Høgh, A. L., and Emmersen, J. (2006). Deepsage—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic acids research*, 34(19):e133–e133.

Nock, A., Ascano, J. M., Barrero, M. J., and Malik, S. (2012). Mediator-regulated transcription through the+ 1 nucleosome. *Molecular cell*, 48(6):837–848.

Nojima, T., Gomes, T., Grosso, A. R. F., Kimura, H., Dye, M. J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N. J. (2015). Mammalian net-seq reveals genome-wide nascent transcription coupled to rna processing. *Cell*, 161(3):526–540.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cdnas. *Nature*, 420(6915):563–573.

Ongen, H., Andersen, C. L., Bramsen, J. B., Oster, B., Rasmussen, M. H., Ferreira, P. G., Sandoval, J., Vidal, E., Whiffin, N., Planchon, A., et al. (2014). Putative cis-regulatory drivers in colorectal cancer. *Nature*, 512(7512):87–90.

Orlando, D. A., Guenther, M. G., Frampton, G. M., and Young, R. A. (2012). Cpg island structure and trithorax/polycomb chromatin domains in human cells. *Genomics*, 100(5):320–326.

Pai, A. A., Pritchard, J. K., and Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. *PLoS genetics*, 11(1):e1004857.

Park, S. G. and Choi, S. S. (2010). Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC evolutionary biology*, 10(1):241.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary dna. *Nucleic acids research*, 37(18):e123–e123.

Phillips, D. (1963). The presence of acetyl groups in histones. *Biochemical Journal*, 87(2):258.

Pombo, A. and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology*.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., Schierup, M. H., and Jensen, T. H. (2008). Rna exosome depletion reveals transcription upstream of active human promoters. *Science*, 322(5909):1851–1854.

Prince, V. E. (2002). The hox paradox: More complex (es) than imagined. *Developmental biology*, 249(1):1–15.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–283.

Raff, R. A., Kaufman, T. C., et al. (1991). *Embryos, genes, and evolution: the developmental-genetic basis of evolutionary change.* Indiana University Press.

Rando, O. J. (2012). Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Current opinion in genetics & development*, 22(2):148–155.

Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Stein, T. I., Bahir, I., Belinky, F., Morrey, C. P., Safran, M., et al. (2013). Malacards: an integrated compendium for diseases and their annotation. *Database*, 2013:bat018.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1998). Genecards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14(8):656–664.

Reynolds, N., O'Shaughnessy, A., and Hendrich, B. (2013). Transcriptional repressors: multifaceted regulators of gene expression. *Development*, 140(3):505–512.

Rhee, H. S. and Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483(7389):295–301.

Rhie, S. K., Hazelett, D. J., Coetzee, S. G., Yan, C., Noushmehr, H., and Coetzee, G. A. (2014). Nucleosome positioning and histone modifications define relationships between regulatory elements and nearby gene expression in breast epithelial cells. *BMC genomics*, 15(1):331.

Riising, E. M., Comet, I., Leblanc, B., Wu, X., Johansen, J. V., and Helin, K. (2014). Gene silencing triggers polycomb repressive complex 2 recruitment to cpg islands genome wide. *Molecular cell*, 55(3):347–360.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., Pachter, L., et al. (2011). Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biol*, 12(3):R22.

Robertson, A. G., Bilenky, M., Tam, A., Zhao, Y., Zeng, T., Thiessen, N., Cezard, T., Fejes, A. P., Wederell, E. D., Cullum, R., et al. (2008). Genome-wide relationship between histone h3 lysine 4 mono-and tri-methylation and transcription factor binding. *Genome research*, 18(12):1906–1917.

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

Roider, H. G., Lenhard, B., Kanhere, A., Haas, S. A., and Vingron, M. (2009). Cpg-depleted promoters harbor tissue-specific transcription factor binding signals—implications for motif overrepresentation analyses. *Nucleic acids research*, 37(19):6305–6315.

Roy, A. L. and Singer, D. S. (2015). Core promoters in transcription: old problem, new insights. *Trends in biochemical sciences*, 40(3):165–171.

Roy, B., Haupt, L. M., and Griffiths, L. R. (2013). Review: Alternative splicing (as) of genes as an approach for generating protein complexity. *Current genomics*, 14(3):182.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Stein, T. I., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., et al. (2010). Genecards version 3: the human gene integrator. *Database*, 2010:baq020.

Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2002). Using the transcriptome to annotate the genome. *Nature biotechnology*, 20(5):508–512.

Schad, E., Tompa, P., and Hegyi, H. (2011). The relationship between proteome size, structural disorder and organism complexity. *Genome Biol*, 12(12):R120.

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

Schotta, G., Lachner, M., Sarma, K., Ebert, A., Sengupta, R., Reuter, G., Reinberg, D., and Jenuwein, T. (2004). A silencing pathway to induce h3-k9 and h4-k20 trimethylation at constitutive heterochromatin. *Genes & development*, 18(11):1251–1262.

Schroder, K., Irvine, K. M., Taylor, M. S., Bokil, N. J., Le Cao, K.-A., Masterman, K.-A., Labzin, L. I., Semple, C. a., Kapetanovic, R., Fairbairn, L., Akalin, A., Faulkner, G. J., Baillie, J. K., Gongora, M., Daub, C. O., Kawaji, H., McLachlan, G. J., Goldman, N., Grimmond, S. M., Carninci, P., Suzuki, H., Hayashizaki, Y., Lenhard, B., Hume, D. a., and Sweet, M. J. (2012). Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. *Proceedings of the National Academy of Sciences of the United States of America*.

Schug, J., Schuller, W.-P., Kappen, C., Salbaum, J. M., Bucan, M., and Stoeckert, C. J. (2005). Promoter features related to tissue specificity as measured by shannon entropy. *Genome biology*, 6(4):R33.

Scruggs, B. S., Gilchrist, D. A., Nechaev, S., Muse, G. W., Burkholder, A., Fargo, D. C., and Adelman, K. (2015). Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Molecular cell*, 58(6):1101–1112.

Sebé-Pedrós, A., de Mendoza, A., Lang, B. F., Degnan, B. M., and Ruiz-Trillo, I. (2011). Unexpected repertoire of metazoan transcription factors in the unicellular holozoan capsaspora owczarzaki. *Molecular biology and evolution*, 28(3):1241–1254.

Sebé-Pedrós, A., Zheng, Y., Ruiz-Trillo, I., and Pan, D. (2012). Premetazoan origin of the hippo signaling pathway. *Cell reports*, 1(1):13–20.

Sebestyén, E., Zawisza, M., and Eyras, E. (2015). Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic acids research*, page gku1392.

Seki, Y., Hayashi, K., Itoh, K., Mizugaki, M., Saitou, M., and Matsui, Y. (2005). Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice. *Developmental biology*, 278(2):440–458.

Severin, J., Lizio, M., Harshbarger, J., Kawaji, H., Daub, C. O., Hayashizaki, Y., Bertin, N., Forrest, A., Consortium, F., et al. (2014). Interactive visualization and analysis of large-scale sequencing datasets using zenbu. *Nature biotechnology*, 32(3):217.

Shen, Y., Ji, G., Haas, B. J., Wu, X., Zheng, J., Reese, G. J., and Li, Q. Q. (2008). Genome level analysis of rice mrna 3'-end processing signals and alternative polyadenylation. *Nucleic acids research*, 36(9):3150–3161.

Sherwin, W. B. (2010). Entropy and information approaches to genetic diversity and its expression: genomic geography. *Entropy*, 12(7):1765–1798.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286.

Sims, R. J., Belotserkovskaya, R., and Reinberg, D. (2004). Elongation by rna polymerase ii: the short and long of it. *Genes & development*, 18(20):2437–2468.

Smith, Z. D. and Meissner, A. (2013). Dna methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25.

Southern, E. M. (1975). Detection of specific sequences among dna fragments separated by gel electrophoresis. *Journal of molecular biology*, 98(3):503–517.

Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R., and Flavell, R. A. (2005). Interchromosomal associations between alternatively expressed loci. *Nature*, 435(7042):637–645.

Stearns, F. W. (2010). One hundred years of pleiotropy: a retrospective. *Genetics*, 186(3):767–773.

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. *Human mutation*, 21(6):577–581.

Stergachis, A. B., Neph, S., Reynolds, A., Humbert, R., Miller, B., Paige, S. L., Vernot, B., Cheng, J. B., Thurman, R. E., Sandstrom, R., et al. (2013). Developmental fate and cellular maturity encoded in human regulatory dna landscapes. *Cell*, 154(4):888–903.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl 2):S231–S240.

Stewart, J. E. (2014). The direction of evolution: The rise of cooperative organization. *Biosystems*, 123:27–36.

Stock, J. K., Giadrossi, S., Casanova, M., Brookes, E., Vidal, M., Koseki, H., Brockdorff, N., Fisher, A. G., and Pombo, A. (2007). Ring1-mediated ubiquitination of h2a restrains poised rna polymerase ii at bivalent genes in mouse es cells. *Nature cell biology*, 9(12):1428–1435.

Struhl, K. and Segal, E. (2013). Determinants of nucleosome positioning. *Nature structural & molecular biology*, 20(3):267–273.

Sun, X., Zou, Y., Nikiforova, V., Kurths, J., and Walther, D. (2010). The complexity of gene expression dynamics revealed by permutation entropy. *BMC bioinformatics*, 11(1):607.

Takahashi, H., Kato, S., Murata, M., and Carninci, P. (2012). Cage (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. *Gene Regulatory Networks: Methods and Protocols*, pages 181–200.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J. S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., et al. (2011). Identification of 67 histone marks and

histone lysine crotonylation as a new type of histone modification. *Cell*, 146(6):1016–1028.

Taylor, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Semple, C. A. (2006). Heterotachy in mammalian promoter evolution.

Tenreiro Machado, J. (2012). Shannon entropy analysis of the genome code. *Mathematical Problems in Engineering*, 2012.

Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A. R., Deaton, A., Andrews, R., James, K. D., et al. (2010). Cpg islands influence chromatin structure via the cpg-binding protein cfp1. *Nature*, 464(7291):1082–1086.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.

Tian, B., Hu, J., Zhang, H., and Lutz, C. S. (2005). A large-scale analysis of mrna polyadenylation of human and mouse genes. *Nucleic acids research*, 33(1):201–212.

Tirosh, I., Weinberger, A., Carmi, M., and Barkai, N. (2006). A genetic signature of interspecies variations in gene expression. *Nature genetics*, 38(7):830–834.

Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active β-globin locus. *Molecular cell*, 10(6):1453–1465.

Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., and Myers, R. M. (2004). An abundance of bidirectional promoters in the human genome. *Genome research*, 14(1):62–66.

Uccelli, A., Moretta, L., and Pistoia, V. (2008). Mesenchymal stem cells in health and disease. *Nature Reviews Immunology*, 8(9):726–736.

VanderMeer, J. E. and Ahituv, N. (2011). cis-regulatory mutations are a genetic cause of human limb malformations. *Developmental Dynamics*, 240(5):920–930.

Vavoulis, D. V., Francescatto, M., Heutink, P., and Gough, J. (2015). Dgeclust: differential expression analysis of clustered count data. *Genome biology*, 16(1):39.

Vavouri, T. and Lehner, B. (2012). Human genes with cpg island promoters have a distinct transcription-associated chromatin organization. *Genome Biol*, 13(11):R110.

Venkatesh, S. and Workman, J. L. (2013). Set2 mediated h3 lysine 36 methylation: regulation of transcription elongation and implications in organismal development. *Wiley Interdisciplinary Reviews: Developmental Biology*, 2(5):685–700.

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (2008). High-resolution mapping of expression-qtls yields insight into human gene regulation. *PLoS genetics*, 4(10):e1000214.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., et al. (2015). Enhancer evolution across 20 mammalian species. *Cell*, 160(3):554–566.

Vinogradov, A. E. (2006). 'genome design'model and multicellular complexity: golden middle. *Nucleic acids research*, 34(20):5906–5914.

Visel, A., Akiyama, J. A., Shoukry, M., Afzal, V., Rubin, E. M., and Pennacchio, L. A. (2009). Functional autonomy of distant-acting human enhancers. *Genomics*, 93(6):509–513.

Voigt, P., Tee, W.-W., and Reinberg, D. (2013). A double take on bivalent promoters. *Genes & development*, 27(12):1318–1338.

Wagner, G. P. and Zhang, J. (2011). The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics*, 12(3):204–213.

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

Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., Li, B., Cui, F., Wei, J., Ma, C., et al. (2014a). The locust genome provides insight into swarm formation and long-distance flight. *Nature communications*, 5.

Wang, Y.-L., Duttke, S. H., Chen, K., Johnston, J., Kassavetis, G. A., Zeitlinger, J., and Kadonaga, J. T. (2014b). Trf2, but not tbp, mediates the transcription of ribosomal protein genes. *Genes & development*.

Wang, Y.-M., Zhou, P., Wang, L.-Y., Li, Z.-H., Zhang, Y.-N., and Zhang, Y.-X. (2012). Correlation between dnase i hypersensitive site distribution and gene expression in hela s3 cells. *PloS one*, 7(8):e42414.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M. Q., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, 40(7):897–903.

Warnefors, M. and Eyre-Walker, A. (2011a). The accumulation of gene regulation through time. *Genome biology and evolution*, 3:667–673.

Warnefors, M. and Eyre-Walker, A. (2011b). The accumulation of gene regulation through time. *Genome biology and evolution*, 3:667–73.

Watson, J. D., Crick, F. H., et al. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738.

Wei, C. and Moss, B. (1977). 5'-terminal capping of rna by guanylyltransferase from hela cell nuclei. *Proceedings of the National Academy of Sciences*, 74(9):3758–3761.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006.

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., et al. (2013). Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature genetics*, 45(10):1238–1243.

Whalen, S., Truty, R. M., and Pollard, K. S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics*.

Williams, S. R., Aldred, M. A., Der Kaloustian, V. M., Halal, F., Gowans, G., McLeod, D. R., Zondag, S., Toriello, H. V., Magenis, R. E., and Elsea, S. H. (2010). Haploinsufficiency of hdac4 causes brachydactyly mental retardation syndrome, with brachydactyly type e, developmental delays, and behavioral problems. *The American Journal of Human Genetics*, 87(2):219–228.

Wittkopp, P. J. and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59–69.

Wolf, Y. I. and Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. *Bioessays*, 35(9):829–837.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., et al. (2004). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS biology*, 3(1):e7.

Wortmann, M. (2012). Dementia: a global health priority-highlights from an adi and world health organization report. *Alzheimers Res Ther*, 4(5):40.

Yaeger, L., Griffith, V., and Sporns, O. (2011). Passive and driven trends in the evolution of complexity. *arXiv preprint arXiv:1112.4906*.

Young, R. S., Hayashizaki, Y., Andersson, R., Sandelin, A., Kawaji, H., Itoh, M., Lassmann, T., Carninci, P., the FANTOM consortium, Bickmore, W. A., Forrest, A. R., and Taylor, M. S. (submitted). The frequent evolutionary birth and death of functional promoters in mouse and human. *Unpublished*.

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2014). Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*.

Zentner, G. E. and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology*, 20(3):259–266.

Zentner, G. E., Tesar, P. J., and Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research*, 21(8):1273–1283.

Zhang, X., Zhao, X.-M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J.-K., Liu, Z.-P., and Chen, L. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, 28(1):98–104.

Zhao, X., Valen, E., Parker, B. J., and Sandelin, A. (2011). Systematic clustering of transcription start site landscapes. *PLoS One*, 6(8):e23409.

Zhou, V. W., Goren, A., and Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, 12(1):7–18.