Transcriptional co-regulation of microRNAs and protein-coding genes

A thesis submitted to the University of Manchester

for the degree of Doctor of Philosophy

in the Faculty of Life Sciences

2013

By

Aaron Webber

Table of Contents

Abstract 8

Declaration 9

Copyright statement 10

Acknowledgements 11

Preface 12

Chapter 1: Introduction 13

1.1 The protein-coding gene expression pathway 13

1.2 Regulation of transcription 17

1.2.1 Transcription factors 19

1.2.2 Transcriptional regulatory regions 22

1.3 The physical architecture of DNA 23

1.4 Alternative transcription and alternative splicing 25

1.5 RNA processing and transport to the ribosome 27

1.6 Post-transcriptional gene silencing 28

1.6.1 MicroRNAs 29

1.6.2 Mechanisms of microRNA-mediated gene repression 32

1.7 Protein folding and post-translational modification 35

1.8 Crosstalk between the regulators of gene expression 36

1.8.1 Transcriptional regulation of microRNAs 36

1.8.2 Integrated regulatory network of transcription factors and microRNAs 37

1.9 Genome-wide measurement 40

1.9.1 Gene expression 40

1.9.2 Proteins in contact with DNA 41

1.10 Objectives of the research 44

Chapter 2: MicroRNA host genes are highly regulated and biased towards developmental functions 47

Abstract 47

2.1 Introduction 48

2.2 Materials and methods 51

2.3 Results 54

2.3.1 MicroRNA host genes are bound by many transcriptional regulators 54

2.3.2 Chromatin modifications to microRNA host genes favour activation of transcription 54

2.3.3 Expression patterns of intragenic microRNAs and their host genes 57

2.3.4 Characteristics of microRNA host genes 59

2.3.5 MicroRNA host genes are enriched for developmentally important genes 64

2.4 Discussion 67
2.5 Conclusions 69
Chapter 3: Coupled regulation of microRNA genes and their protein-coding neighbours 70
Abstract 70
3.1 Introduction 71
3.2 Methods 74
3.3 Results 77
3.3.1 Classification of microRNAs 77
3.3.2 Intergenic microRNAs are found within regions of high protein-coding gene density 78
3.3.3 MicroRNAs are proximal to protein-coding genes across animal species 80
3.3.4 Favoured orientations of proximal microRNAs 81
3.3.5 Comparison between microRNA flanking and host genes 84
3.3.6 *Cis*-regulatory regions of flanking genes 85
3.3.7 Candidate bidirectional promoters of intergenic microRNA and protein-coding gene pairs 87
3.3.8 Functional association of transcription factors and microRNAs 88
3.4 Discussion 91
3.5 Conclusions 93
Chapter 4: Feedforward regulation by transcription factors and microRNAs 94
Abstract 94
4.1 Introduction 95
4.2 Methods 98
4.3 Results 103
4.3.1 Construction of an integrated regulatory network 103
4.3.2 Relationships between TRF- and microRNA-mediated regulation and target gene expression 106
4.3.3 Bidirectional targeting between a TRF and microRNA pair is linked to TRF regulatory sign 111
4.3.4 Feedforward regulation by TRFs and microRNAs of common gene expression pathways 113
4.3.5 Connectivity between TRFs and microRNAs defines gene expression level 114
4.3.6 Feedforward circuits are associated with stable mRNA expression levels 116
4.3.7 Feedforward circuits are associated with core cellular processes 118
4.4 Discussion 120
4.5 Conclusions 123

Chapter 5: Species-specific gene expression and transcriptional regulation in pathogenic versus natural host SIV infections 124
Abstract 124
5.1 Introduction 126
5.2 Methods 129
5.3 Results 132
5.3.1 Gene expression patterns reflect species-specific T cell dynamics 132
5.3.2 Species-specific patterns of overlap between gene expression clusters 137
5.3.3 Expression changes within peripheral blood CD4+ T cells from individual animals 138
5.3.4 Transcriptional regulators of species-specific gene expression programs 140
5.3.5 Interactions of viral proteins with differentially expressed host genes 151
5.4 Discussion 154
5.5 Conclusions 157
Chapter 6: Discussion 158
6.1 Transcription and microRNA gene birth 158
6.2 MicroRNA gene regulation depends upon the protein-coding context 159
6.3 Regulatory feedback between transcription factors and microRNAs 162
6.4 An integrated regulatory network of transcription factors and microRNAs 162
6.5 MicroRNAs in development 163
6.6 Perturbations of regulatory networks 164
6.7 Final comment 166
Appendix S1: Supplementary material for Chapter 1 167
Appendix S2: Supplementary material for Chapter 2 173
Appendix S3: Supplementary material for Chapter 3 176
Appendix S4: Supplementary material for Chapter 4 178
Appendix S5: Supplementary material for Chapter 5 191
References 200

Word count: 72,488

(**main text**: 55,920; **appendices**: 7,953; **references**: 8,615)

List of Figures and Tables

Figure 1.1. Transcription and translation 15
Figure 1.2. The RNA pol II holoenzyme 18
Figure 1.3. Protein-coding mRNAs expressed from a bidirectional promoter region 19
Figure 1.4. The IFN-α/β activated JAK-STAT cascade 21
Figure 1.5. Transcriptional regulatory regions 23
Figure 1.6. Alternative transcription and alternative splicing of protein-coding genes 26
Figure 1.7. Arrangements of microRNA genes 30
Figure 1.8. The microRNA biogenesis pathway in animals 31
Figure 1.9. MicroRNA and RISC complex binding to the 3'-UTR of a target mRNA 33
Figure 1.10. Varieties of feedback and feedforward circuits containing microRNAs 38
Figure 1.11. ChIP-seq protocol 43
Table 2.1. Acetylation marks within microRNA host gene promoter regions 55
Table 2.2. Methylation marks within microRNA host gene promoter regions 56
Table 2.3. Host gene – microRNA pairs within significant co-expression correlation across human tissues 57
Figure 2.1. Expression and regulation of protein-coding genes and intragenic microRNAs 58
Table 2.4. Characteristics of microRNA host genes 60
Figure 2.2. Transcriptional regulation of microRNA host and non-host genes 62
Table 2.5. Structural and functional enrichments of the microRNA host gene class 65
Table 2.6. Relationship between microRNA host gene function and number of bound transcriptional regulatory factors 66
Figure 3.1. Arrangements of microRNA genes 72
Table 3.1. Numbers and types of microRNA genes in 4 animal species 77
Figure 3.2. Protein-coding and chromatin environment of human intergenic microRNAs 79
Table 3.2. Enrichment of proximal intergenic microRNAs in mammals and invertebrates 81
Figure 3.3. Distance distributions of microRNA clusters around protein-coding TSSs 82
Figure 3.4. Distance distributions of microRNA clusters around protein-coding TSSs and TESs across animal species 83
Figure 3.5. Numbers of transcription factors binding promoter regions of microRNA flanking genes 86
Table 3.3. Regulatory loops between microRNAs and flanking gene promoter regions 89
Figure 3.6. TRF/microRNA feedback loops via bidirectional promoter regions 92
Figure 4.1. Types of TRF – microRNA containing regulatory motifs 96

Figure 4.2. Distributions of TRFs mapping to predicted target genes 104
Figure 4.3. Relationships between gene expression and numbers of regulators..... 107
Table 4.1. Comparison of target/non-target expression ratios for TRF activators, repressors, and those of variable sign 108
Table 4.2. Overlaps between TRF regulatory signs from literature and the Gene Ontology ... 108
Table 4.3. Correlations between number of microRNA target sites and gene expression 109
Figure 4.4. Tissues with more microRNA expression show greater target gene repression…..110
Table 4.4. Bidirectional targeting between transcriptional regulators and microRNAs..... 111
Figure 4.5. Significance of frequencies of TRF – microRNA feedforward loops 113
Figure 4.6. Gene expression levels across subsets of TRF-microRNA feedforward loops..... 115
Figure 4.7. Relationships between feedforward loops and the stability of gene expression... 117
Figure 4.8. Cellular processes significantly enriched for the bidirectional FFL motif 119
Table 5.1. Time series datasets across species and tissue compartments..... 127
Figure 5.1. Gene clusters with common expression profiles upon SIV infection 133
Figure 5.2. Composition of genes in each expression profile by species and T cell source..... 134
Figure 5.3. Expression patterns of cell-cycle and viral-response enriched gene clusters 135
Figure 5.4. Significant gene intersections between expression profiles 138
Figure 5.5. Expression profiles of single and clustered genes from individual monkeys 139
Table 5.2. Collection of co-expressed TRFs from AP CD4+ T cells within cluster 7 141
Figure 5.6. Relationships between transcriptional regulators and co-expression collections of genes 144
Table 5.3. Factors enriched uniquely within *cis*-regulatory regions of genes in cluster 1..... 144
Table 5.4. Predicted target set overlaps of STAT1, STAT2, BATF and IRF4 147
Figure 5.7. Expression perturbations within families of transcription factors following SIV infection 148
Figure 5.8. Expression patterns of selected perturbed genes within AGMs and RMs 150
Figure 5.9. Gene regulatory networks by tissue type and monkey species 152
Table 5.5. Enrichment for virus protein interactions among expression profiles 153
Table S1.1. Transcriptional regulatory factors with ChIP-seq data from YALE and HAIB ENCODE consoria..... 167
Table S2.1. Matched samples from protein-coding and microRNA expression atlases 173
Table S2.2. GO terms with most highly regulated microRNA host genes..... 174
Figure S3.1. Relationship between microRNA cluster density and chromatin state.....176
Table S3.2. 5’-antisense microRNA clusters and their flanking genes..... 177
Figure S4.1. Density of bZIP and ETS family TFs around protein-coding 5’-TSSs 178
Figure S4.2. Degree distributions of transcriptional regulators 179

Figure S4.3. MicroRNA expression level against number of transcriptional regulators 180
Table S4.4. Comparison of expression levels of targets and non-targets of individual TRFs ... 181
Figure S4.5. Fraction of POU2F2 and Suz12 target genes as expression level varies 184
Figure S4.6. Variation in gene expression with numbers of miRanda-predicted microRNA target sites 185
Figure S4.7. Variation in gene expression with numbers of seed sequence predicted microRNA target sites 186
Table S4.8. Hypergeometric p-values for incidence of the bidirectional (TRF, microRNA) feedforward circuit 187
Figure S4.9. Variation in motif significance with minimum TRF peak quality score 188
Figure S4.10. Variation in motif significance with *cis*-regulatory region size 188
Figure S4.11. Gene expression levels across subsets of TRF-microRNA feedforward loops 189
Table S4.12. Correlations between individual node z-scores across target predictors 190
Table S5.1. Gene lists within each of 15 expression profile clusters 191
Table S5.2. ChIP-seq datasets for transcription regulatory factors used in Chapter 5 194
Table S5.3. Predicted target set overlaps of STAT1, STAT2, BATF and IRF4 197
Table S5.4. Significantly perturbed families of transcription factors 199

Abstract

This thesis was presented by Aaron Webber on the 4th December 2013 for the degree of Doctor of Philosophy from the University of Manchester. The title of this thesis is 'Transcriptional co-regulation of microRNAs and protein-coding genes'. The thesis relates to gene expression regulation within humans and closely related primate species. We have investigated the binding site distributions from publically available ChIP-seq data of 117 transcription regulatory factors (TRFs) within the human genome. These were mapped to *cis*-regulatory regions of two major classes of genes, ≈ 20,000 genes encoding proteins and ≈ 1500 genes encoding microRNAs. MicroRNAs are short 20 - 24 nt noncoding RNAs which bind complementary regions within target mRNAs to repress translation. The complete collection of ChIP-seq binding site data is related to genomic associations between protein-coding and microRNA genes, and to the expression patterns and functions of both gene types across human tissues.

We show that microRNA genes are associated with highly regulated protein-coding gene regions, and show rigorously that transcriptional regulation is greater than expected, given properties of these protein-coding genes. We find enrichment in developmental proteins among protein-coding genes hosting microRNA sequences. Novel subclasses of microRNAs are identified that lie outside of protein-coding genes yet may still be expressed from a shared promoter region with their protein-coding neighbours. We show that such microRNAs are more likely to form regulatory feedback loops with the transcriptional regulators lying in the upstream protein-coding promoter region.

We show that when a microRNA and a TRF regulate one another, the TRF is more likely to sometimes function as a repressor. As in many studies, the data show that microRNAs lying downstream of particular TRFs target significantly many genes in common with these TRFs. We then demonstrate that the prevalence of such TRF/microRNA regulatory partnerships relates directly to the variation in mRNA expression across human tissues, with the least variable mRNAs having the most significant enrichment in such partnerships. This result is connected to theory describing the buffering of gene expression variation by microRNAs. Taken together, our study has demonstrated significant novel linkages between the transcriptional TRF and post-transcriptional microRNA-mediated regulatory layers.

We finally consider transcriptional regulators alone, by mapping these to genes clustered on the basis of their expression patterns through time, within the context of CD4+ T cells from African green monkeys and Rhesus macaques infected with Simian immunodeficiency virus (SIV). African green monkeys maintain a functioning immune system despite never clearing the virus, while in rhesus macaques, the immune system becomes chronically stimulated leading to pathogenesis. Gene expression clusters were identified characterizing the natural and pathogenic host systems. We map transcriptional regulators to these expression clusters and demonstrate significant yet unexpected co-binding by two heterodimers (STAT1:STAT2 and BATF:IRF4) over key viral response genes. From 34 structural families of TRFs, we demonstrate that bZIPs, STATs and IRFs are the most frequently perturbed upon SIV infection. Our work therefore contributes to the characterization of both natural and pathogenic SIV infections, with longer term implications for HIV therapeutics.

Declaration

No portion of the work referred to in this thesis inclusive of Chapters 1 – 4, Chapter 6, and appendices has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Chapter 5 is the result of collaboration between two joint first authors, and additional collaborators. A proportion of the material in this chapter has been submitted in support of the award of Doctor of Philosophy for the other joint first author, Dr J. MacPherson. The respective contributions to material in Chapter 5 are laid out in the preface to that chapter. Only those contributions to Chapter 5 by the author of this thesis are submitted in support of the present application for the award of Doctor of Philosophy.

Copyright statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks or other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and / or Reproductions

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's policy on Presentation of Theses.

Acknowledgements

On the academic side, foremost I would like to thank my supervisor, Dr. Sam Griffiths-Jones, for his effort, patience and guidance in helping me to prepare this thesis. I would also like to thank Professor David Robertson, who was the principal investigator for the research in Chapter 5. Many other colleagues were generous with their time, in particular, Dr. Jamie Macpherson, for inviting me to participate in the collaboration reflected in Chapter 5, Dr. Antonio Marco, for critical reading of Chapter 2 and many helpful discussions, Dr. Matthew Ronshaugen, for critical reading of both Chapter 2 and Chapter 3, Professor Michaela Müller-Trütwin and Dr. Beatrice Jacquelin, from the Pasteur Institute in Paris, for their biological insight and detailed scrutiny of the material in Chapter 5, and Dr. Ana Kozomara, for assistance with miRBase. I am also grateful to these colleagues, and others, notably Dr. Hubert Rogers, Kasia Hooks, and Maria Ninova, for their all-round good humour. More distantly now, some of my secondary school teachers, including Iain MacLean, Alan Callum, Dr. Sloper and others, helped to foster an enthusiasm for scientific ideas. My undergraduate Philosophy tutors, especially Dr. Maria Frasca-Spada, did their best to teach me some critical thinking skills. Finally, I am grateful to the British Biochemical and Biological Sciences Research Council for funding the research.

I am indebted to Bee, Paul, Bernard and Wendy, and the rest of the family, for their love and generosity, particularly during the write-up period. They gave me the space I needed to get the job done, but they were there whenever I needed to unwind. I am especially grateful to Ed, Nick, Will and James for being marvellous friends...hopefully the missed social occassions are forgiven. I enjoyed my time in Ashburne Hall, where I worked as a pastoral tutor for a little over 2 years, thanks to the tutorial team and especially the warden Norman Gillson, as well as easy access to a Bechstein grand. In the final months, Sparky my cat spent many a night sleeping on the desk as I tapped away. This cheered me up no end.

The thesis is dedicated to my step-grandfather Graham Cashmore (1942 – 2012). He was a wonderfully kind and supportive man who would have celebrated the completion of this project in his ever cheerful way.

Preface

My main interest as a science researcher is to understand regulatory mechanisms governing the expression of protein-coding genes. In this thesis I have focused upon transcriptional regulation by DNA-binding proteins, including transcription factors, and post-transcriptional regulation by short endogenously produced noncoding RNAs, termed microRNAs. My project supervisor, Dr. Sam Griffiths-Jones, designed and maintains a database of microRNA genes and related data across numerous animal and plant genomes (http://www.mirbase.org) (Griffiths-Jones, Grocock *et al.* 2006). Dr. Griffiths-Jones also has a significant background in publishing work on microRNAs and more generally on noncoding RNAs. Both kinds of regulator, transcription factors and microRNAs, number over 1000 distinct genes in human, which can regulate one another, as well as co-regulating collections of protein-coding genes (Shalgi, Lieber *et al.* 2007). There is significant research scope to characterise the resulting system of interactions.

The thesis consists of an introductory chapter describing the scientific background to the research as a whole, followed by four chapters containing original research in focused areas (Chapters 2 – 5), and finally a discussion, which again relates to the thesis as a whole (Chapter 6). There are also appendices S1 – S5 for data tables and supplementary analysis, which are referred to in the main text. As areas of focused research were developed, it became apparent that the material could be arranged into a number of distinct but related sets of findings suitable for submission to peer-reviewed journals. We therefore chose to prepare the thesis in the alternative format. This means that each research chapter is written in the form of a journal article, with self-contained introduction, methods, results and discussion.

Chapter One

Introduction

This thesis is concerned with regulatory processes controlling the expression patterns of proteins within humans and related species. The chapter structure has therefore been shaped by proceeding along the protein-coding gene expression pathway, from transcriptional regulation to post-translational modification. The research focuses in particular upon two classes of protein-coding gene expression regulator, transcription factors, and microRNAs. The material within this chapter summarizes the scientific background to this research. The chapter concludes with a short discussion of the overall objectives of my research.

1.1 The protein-coding gene expression pathway

Proteins are composed of linear chains of chemical components, termed amino acids, folded into an effectively limitless variety of intricate 3-dimensional structures. These structures give rise to the specialized functions of proteins, including as receptors, signalling molecules, enzymes, regulators, transporters, and scaffolds, so that proteins are active in all aspects of cellular organization and behaviour (Alberts, Johnson *et al.* 2007). Many biological processes, for example the replication of a cell, oxidative phosphorylation and photosynthesis, or the response of the immune system to a pathogen, require the coordinated expression of hundreds of proteins (Niehrs and Pollet 1999). Aberrant expression patterns of proteins provide signatures of genetic diseases including cancer (Hanash 2003). Molecular processes determining the protein contents of a cell are of fundamental importance to the functioning of living things.

As early as 1941, a link was established between specific metabolic proteins and heritable genetic elements in the nucleus of a cell, leading to the 'one gene - one enzyme' hypothesis of George Beadle and Edward Tatum (Beadle and Tatum 1941). While heritable genetic traits had already been mapped to locations along the lengths of chromosomes, particularly through work by Thomas Morgan and Alfred Sturtevant on genetic maps in fruit flies, the molecular basis for this was unknown (Sturtevant, Bridges *et al.* 1919). At that time, the DNA part of chromosomes was generally considered to be little more than a scaffold for information-rich nuclear proteins, which were the true bearers of heredity. Between 1944 and 1953, this picture was overturned, as purified DNA and not protein was shown to carry genetic information between *Streptococcus pneumoniae* bacterial cells, determining their virulence in

mice (Avery, Macleod *et al.* 1944); then in 1952, DNA was confirmed convincingly as the genetic material of the T2 phage (Hershey and Chase 1952); and in 1953, the atomic structure of DNA was solved by James Watson and Francis Crick, using crystallographic data determined by Rosalind Franklin and Maurice Wilson (Watson and Crick 1953). The famous double-helix geometry, with two hydrogen-bonded chains of complementary nucleotides, was consistent with replication, and therefore inheritance, via unwinding of the chains, and polymerisation of a pair of new DNA strands complementary to each of the originals (Meselson and Stahl 1958). The consistent pairing of adenine with thymine, and cytosine with guanine, known as Watson-Crick base pairing, meant that the sequence on one strand would determine the sequence of the other. The fundamental question arose how information in the double-stranded DNA polymer could direct the synthesis of polypeptides.

It was not clear whether genetic traits were encrypted through the order of nucleotides along a DNA strand, or otherwise through its chemical or electrostatic characteristics. The suggestion that linear sequences of nucleotides were sufficient to specify sequences of amino acids comprising protein molecules became known as the *Sequence Hypothesis (Crick 1958)*. Since DNA consists of chains of 4 canonical kinds of nucleotide base, while proteins contain up to 20 common amino acids, it was apparent that a nucleotide sequence could not be simply transcribed into a polypeptide chain. The physicist George Gamow suggested that sequential overlapping nucleotide words of length 3 might be related through a one-to-one mapping with amino acids, a kind of Beadle & Tatum hypothesis in miniature (Gamow 1954). In the mechanism proposed, amino acids were fitted directly within the grooves of the double-helix itself bonding to quartets of bases. However, the site of protein synthesis was pinpointed just one year later, by George Palade and Albert Claude, to particles outside the nucleus, and therefore far away from the genome (Palade 1955). These were termed *microsomes*, and were bound to the surface of a folded membrane structure, the rough endoplasmic reticulum (Porter, Claude *et al.* 1945). When purified, it was clear that microsomes contained no DNA, but rather, the closely related ribonucleic acid, RNA, leading to their current naming as *ribosomes*. The RNA fragments from these were shown to hybridize with DNA from corresponding gene regions. Thus, there appeared to be a flow of information from gene regions stored in nuclear DNA, through short, mobile, complementary RNA messenger elements (mRNAs), to a polypeptide chain constructed by the ribosome (Crick 1958).

Synthesis of an RNA strand from a DNA template, or *transcription*, was understood as the writing out of a DNA message into a complementary RNA message, now termed a pre-mRNA, with the base uracil on the RNA strand replacing the base thymine from the coding DNA strand (Figure 1.1A). The molecular catalyst of mRNA synthesis in *E. coli* was identified by Jerard

Hurwitz, and others, as a ribonucleotide-ligating enzyme, termed RNA polymerase (RNAP) (Hurwitz, Bresler *et al.* 1960; Stevens 1960). A variety of RNA polymerase enzymes is known today, with distinct functions, and species-specific forms. Nevertheless, all known RNA polymerases are related through structure and catalytic mechanism (Eick, Wedel *et al.* 1994; Sekine, Tagami *et al.* 2012).

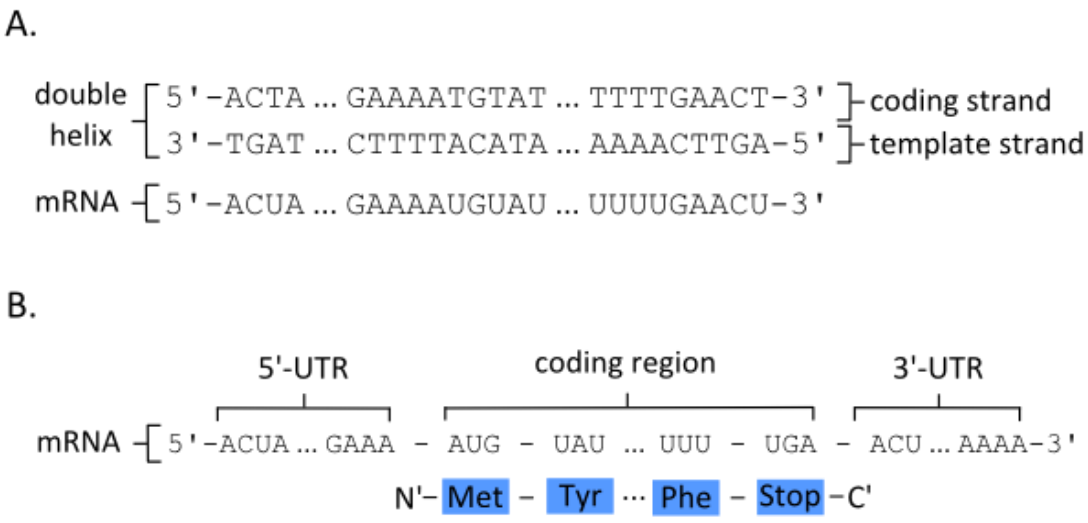


Figure 1.1. Transcription and translation.

A. Transcription: A pre-mRNA polymer is formed matching the DNA sequence of a gene , with the base uracil (U) in RNA substituting for thymine (T) in DNA. Polymerization is achieved through complementary base pairing between the nascent mRNA and the progressively unwound DNA template strand. Complementary base pairing follows the simple rules G with C, and A with U (or T). Strand orientations are defined by the direction from 5' to 3' carbon atoms within the sugar moieties from each nucleotide. Differences between the pre-mRNA and the mature mRNA derived from this are described later (Sections 1.4 and 1.5).

B. Translation: The central part of the mature mRNA sequence, termed the coding region, consists of a sequence of nucleotide triplets termed codons, from start (AUG) to stop (in this case, UGA), specifying the sequence of amino acids that are linked to form a polypeptide chain. Polypeptide chain orientation is defined from the amino (N) to the carboxyl (C) groups of amino acid residues within the chain. Outer parts of the mRNA sequence not encoding the polypeptide sequence are termed the 5' and 3' untranslated regions (UTRs).

Synthesis of a polypeptide strand from an RNA template was termed *translation*, to reflect that the chemical languages of nucleic acids and proteins differ fundamentally. By means of inserting and deleting single nucleotides into RNA templates, Francis Crick and Sydney Brenner, together with Leslie Barnett and R. J. Watts-Tobin, showed that amino acids were coded for by non-overlapping triplets of nucleotides, termed codons, in mRNAs from the bacteriophage T4, and were read as a continuous sequence from the start to the end of a *coding region* (Figure 1.1B) (Crick, Barnett *et al.* 1961). Soon after, by using synthetic

polynucleotide constructs within an *in vitro* ribosomal system, Marshall Nirenberg and Heinrich Matthaei began to enumerate the correspondences between codons and amino acids (Matthaei, Jones *et al.* 1962). Collectively, these correspondences are known as the genetic code. By 1968 the genetic code in *E. coli* cells had been solved completely, emphatically supporting the *Sequence Hypothesis* (Nirenberg, Leder *et al.* 1965). Also in the early 1960s, a physical basis for the mappings in the genetic code was determined with the characterization of an ancient class of RNAs, the transfer RNAs, which each possess a triplet of nucleotides complementary to a particular codon (Holley, Apgar *et al.* 1965). Specialized enzymes link the amino acid corresponding to the recognised codon to each tRNA (O'Donoghue and Luthey-Schulten 2003). Protein and RNA components of the ribosome then choreograph a process of binding the correct tRNA to a template codon, catalysing transfer of the tRNA-linked amino acid to the growing polypeptide chain, and then shifting relative to the messenger sequence to begin reading the next codon in sequence. Four codons were found to have special meanings, the methionine codon signalling also the start of a coding mRNA sequence, and 3 chain-termination or stop codons signalling the end. Francis Crick hypothesised that the genetic code would be constant across all living things, and today, we know that this is very nearly true (Crick 1963; Wong 1976). There are only occasional genome-specific variations, mainly around stop and start signals, for example in certain prokaryotes and *archaea,* and in the shared genetic code of the chloroplast and mitochondrion (Osawa, Jukes *et al.* 1992).

The remarkable achievements of the post-war years significantly advanced the science of molecular biology. Near universal principles had been uncovered, which laid the foundations to explain some of the key features of life, including the mechanism of heredity, and the expression of the protein-coding genome. Investigation into protein-coding gene expression has been continuously active ever since, and we can identify many distinct areas of ongoing research, e.g.:

[1] The regulation of transcription

[2] The physical architecture of DNA

[3] Alternative transcription and splicing

[4] mRNA maturation and transport from nucleus to ribosome

[5] Post-transcriptional regulation, especially repression, of mRNAs

[6] Protein folding and post-translational covalent modifications to proteins

[7] Widespread regulatory crosstalk between any of [1] – [6].

This study is concerned mainly with areas [1], [5] and [7], and to a lesser degree with [2] (Sections 1.2, 1.6 and 1.8). The remaining areas are discussed only briefly (Sections 1.3, 1.4, 1.5 and 1.7). There is then a short account of some of the key genome-wide technologies (Section 1.9), and finally, a short discussion of the key objectives of my research.

1.2 Regulation of transcription

Transcription begins with the engagement of an RNA polymerase on DNA in a region called *the core promoter*, which includes the first transcribed nucleotide. Here, we are only concerned with the eukaryotic enzyme RNA polymerase II, a 10 – 12 subunit enzyme which is largely conserved from yeast to humans (Myer and Young 1998). The process of anchoring RNA pol II at gene starts depends on numerous factors. These assist in locating start sites, creating a favourable environment for anchoring the polymerase to the DNA, unwinding coding and template strands, and transmitting information between a variety of molecular complexes associated with transcription (Figure 1.2). This begins with the stepwise assembly of a minimum of 5 initiation factors (TFIIs B, D, E, F, and H), which interact with DNA and with RNA pol II to form the *transcription preinitiation complex* (PIC) (Liu, Bushnell *et al.* 2013).

Most of the components of the PIC are themselves multimeric. For example, TFIID, which is among the very first factors to recognise and bind specific DNA regions, contains the TATA-box binding protein (TBP) together with from 8 to 16 TBP-associated factors (TAFs) (Ranish, Yudkovsky *et al.* 1999). The factor TFIIH includes the helicases XPB and XPD. These helicases promote local unwinding of double-stranded DNA to create a 12 - 19 nt *transcription bubble* suitable for DNA – RNA binding during RNA polymerisation (Zaychikov, Denissova *et al.* 1995; Kim, Ebright *et al.* 2000; Pal, Ponticelli *et al.* 2005). The PIC is joined in configuring DNA and RNA pol II for transcription by a variety of other transcription factors, chromatin-remodelling agents, and co-regulatory molecules, collectively termed the RNA pol II holoenzyme (Figure 1.2). In particular a region of RNA pol II makes contacts with a complex of at least 25 proteins, the *mediator* (aliases: TRAP, SMCC, or DRIP complex in various species), discovered by Roger Kornberg and collaborators (Kim, Bjorklund *et al.* 1994). The large surface-area of the mediator allows it to integrate regulatory signals from more distal factors and transmit these signals to the PIC (Ito and Roeder 2001). A number of rounds of starting transcription but disengaging within the first 20 nt are required, before RNA pol II is able to escape the transcription bubble and begin elongating a complete RNA transcript of the gene, often 100s of kilo-base pairs (kb) in length (Goldman, Ebright *et al.* 2009; Liu *et al.* 2013).

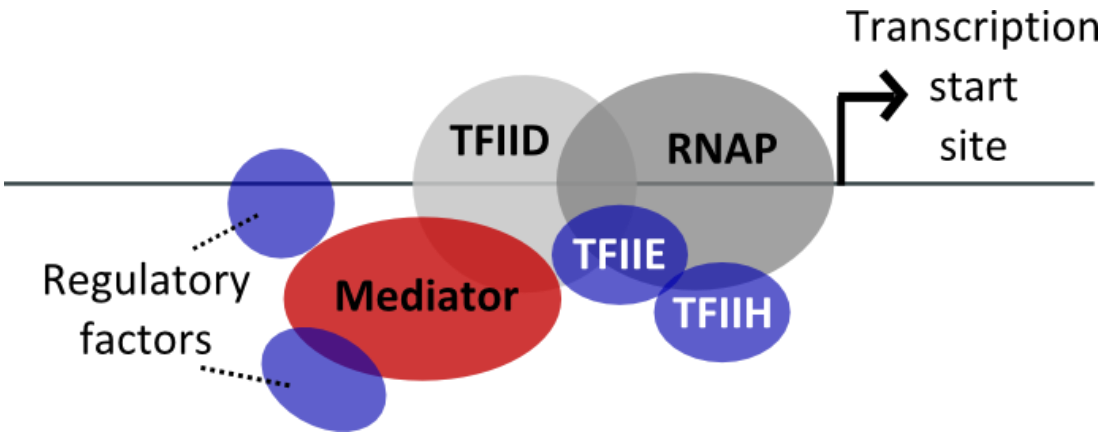


Figure 1.2. The RNA pol II holoenzyme.

Components of the RNA pol II (RNAP) holoenzyme, engaged prior to transcription of protein coding gene regions. The multimeric mediator complex transmits signals from DNA-bound and non-DNA-bound regulatory factors to the C-terminal domain of RNAP.

Recognition of core eukaryotic promoter regions by components of the PIC is a complex process. In the classical view, the TFIIB component TBP binds a canonical sequence 5'-TATAAA-3' or near variant, called the TATA box (Lifton, Goldberg *et al.* 1978). Analysis of the 5' ends of protein-coding genes has since revealed that only around 10 - 25% of eukaryotic genes possess a TATA box (Carninci, Sandelin *et al.* 2006). Other general recognition elements that can interact with TBP or other GTF components include the initiator element (Inr), B-recognition element (BRE), and downstream promoter element (DPE), and of these only Inr is present at more than 25% of promoters (Xi, Yu *et al.* 2007; Juven-Gershon, Hsu *et al.* 2008).

A much more common characteristic of promoter regions in higher eukaryotes is the possession of CG dinucleotide pairs (within each strand) at rates far in excess of the background. Such regions are termed CpG islands (Deaton and Bird 2011). These are associated with around 70% of vertebrate gene promoters, representing the most common vertebrate promoter type (Saxonov, Berg *et al.* 2006). Most CpG island promoters are associated with scattered transcription initiation sites giving a spread of 5' ends of genes. By contrast, most TATA box promoters are associated with a very precise transcription start site giving a consistent 5' gene end (Carninci *et al.* 2006). In addition, CpG island promoters typically result in higher levels of overall gene expression (Carninci *et al.* 2006; Seila, Calabrese *et al.* 2008; Deaton and Bird 2011).

Most promoter regions have a preferred direction, so that the DNA sequence lying downstream (and therefore the coding strand) is clearly defined (Carninci *et al.* 2006). Despite this, most transcription start sites are associated with a variety of sense and antisense transcripts of differing lengths, and it is not yet clear how much of this represents functional

sequence rather than cellular noise (Taft, Pheasant *et al.* 2007; Seila *et al.* 2008; Wei, Pelechano *et al.* 2011). Some promoter regions however are known to produce functional RNAs in both directions, and an example is shown in Figure 1.3 (Trinklein, Aldred *et al.* 2004; Carninci *et al.* 2006). This may occur either from closely spaced but divergent core promoter regions, or from a single and therefore truly bidirectional core promoter. Consistent with a less constrained architecture, bidirectional promoter regions tend to be associated with CpG islands and overwhelmingly lack TATA boxes (Trinklein *et al.* 2004).

We build the case for a novel general class of bidirectional promoter regions in Chapter 3.

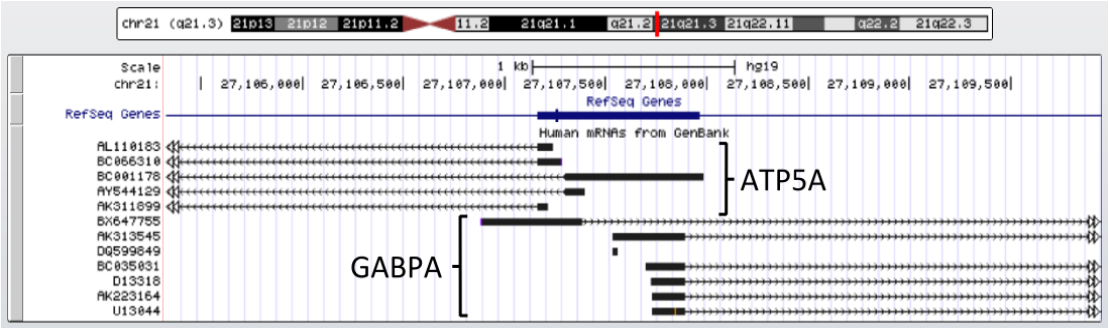


Figure 1.3. Protein-coding mRNAs expressed from a bidirectional promoter region.

The figure shows a screenshot from the UCSC genome browser (Karolchik, Baertsch *et al.* 2003) of overlapped, bidirectionally oriented, transcripts for the genes ATP5A and GABPA, on human chromosome 21. Both genes have a number of alternative transcription start sites, including sites within transcribed regions of the other gene (Carninci *et al.* 2006).

1.2.1 Transcription factors

Transcription factors (TFs) can be defined as DNA binding proteins which affect the process of transcription. Components of the PIC are called general transcription factors (GTFs) because they are engaged prior to most protein-coding gene transcription events, even though some of them do not bind DNA directly (Sawadogo and Sentenac 1990). Most TFs however regulate only a subset of genes, and this regulation often varies according to context, such as cell type. The DNA-binding domains of both types of TF recognise specific DNA sequences, called *recognition elements*, but the term *sequence-specific TF* is usually reserved for the latter (Latchman 1993). The binding of a TF to DNA is determined by stereochemistry and by hydrogen bonds between amino acid side chains and electron acceptors or donors within the grooves of the double helix. This can be achieved through a variety of protein structural conformations, examples of which include the zinc-finger, leucine zipper, winged helix-loop-helix, or homeobox, domains (Latchman 1993). Sometimes RNA pol II is referred to as a GTF, and since it is capable of progressing along arbitrary DNA sequences, it is also a true non-

sequence-specific factor. In the human genome, estimates of the total number of TFs are variable, but more than 2600 proteins are predicted to possess a DNA-binding domain, and of these, no fewer than 1400 are confidently annotated as functional TFs (Venter, Adams *et al.* 2001; Babu, Luscombe *et al.* 2004; Gerstein, Kundaje *et al.* 2012).

Many features of TFs allow them to be arranged into related sets, for example, mechanism of activation, DNA-binding domain, or function. Some TFs are active within nearly every cell, such as Sp1 or RNA pol II, but most are switched on only in particular developmental contexts, or in response to the appropriate signals (Brivanlou and Darnell 2002). Developmental TFs, e.g. GATA proteins and HOX proteins, are activated or repressed within some very specific cell types and time periods, and often remain in this state throughout the rest of the life of the organism, in order to maintain cell type identities (Schughart, Kappen *et al.* 1988; Rothenberg and Pant 2004). Signal-activated TFs have levels which vary according to the stimulus that is present. And indeed, for regulators involved in cell and organism homeostasis, the ability to match their activity to environmental fluctuations is necessary to compensate for these.

Transmission of signals to TFs often falls under one of a small number of general mechanisms (Brivanlou and Darnell 2002). Some TFs respond directly to small chemical messengers via *ligand-binding domains*. These include TFs activated by ligands of extracellular origin, including nuclear receptors, responsive to hormones such as testosterone (Evans 1988). Likewise, there are TFs activated through binding by an intracellular ligand, such as the sterol-sensitive factor SREBP1 (Brown and Goldstein 1997). Other TFs respond only indirectly to signals, through information relayed to them by a series of intermediate reactions termed a signal-transduction cascade. The primary activation signal is detected by a cell-surface receptor, and then often transmitted to intracellular TFs via one or more sequential phosphorylation reactions carried out by kinase enzymes (Darnell, Kerr *et al.* 1994; Pearson, Robinson *et al.* 2001). Very many TF families, e.g. STATs, SMADs, NF-kB, AP-1/ATF superfamily members, Heat shock factors (HSFs), etc., are activated this way, especially in response to stressful conditions, including assault by viruses and other pathogens (Darnell *et al.* 1994; Sinha, Jaggi *et al.* 2011; Lupino, Ramondetti *et al.* 2012). An example of this type of signal transduction cascade is shown in Figure 1.4.

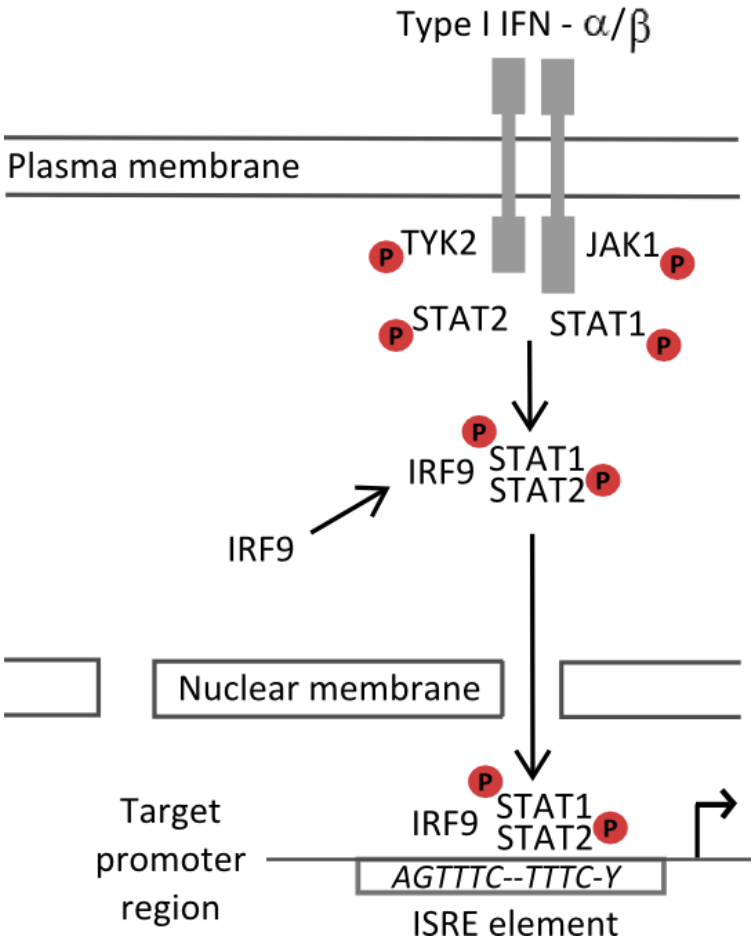


Figure 1.4. The IFN-α/β activated JAK-STAT cascade.

The signalling molecules interferon-α or interferon-β (IFN-α/β) bind to transmembrane receptors on target immune system cell types, leading to phosphorylation by tyrosine kinase 2 (TYK2), and Janus-activated Kinase 1 (JAK1), of the factors STAT1 and STAT2. These factors then heterodimerize and bind interferon regulatory factor 9 (IRF9). The trimeric complex is translocated to the nucleus and binds to copies of the 14mer interferon-stimulated regulatory element (ISRE) within the promoter or enhancer regions of target genes. The characters '-' and 'Y' within the ISRE sequence represent 'any nucleotide', and 'either cytosine (C) or thymine (T)', respectively.

In Chapter 5, these kinds of cascades are explored through dynamic time series data from two species of monkeys with a differential immune response to simian immunodeficiency virus (SIV) infection.

Once engaged at a binding site through the DNA-binding domain, TFs regulate transcription through their *transactivation (*or *transrepression)* domains. These domains interact with other regulatory TFs, with transcriptional co-regulators not bound to DNA, or with components of the RNA pol II holoenzyme (Ito and Roeder 2001; Piskacek, Gregor *et al.* 2007). Some TFs are recognised very generally as activators or repressors of transcription, while others such as the

yin-yang factor YY1 have a regulatory sign that varies depending upon the binding context (Hahn 1992).

Within chapters 2 - 4, and as is common in network analysis, TFs are mainly not distinguished on the basis of activation mechanism, or DNA binding domain. They are treated as units of regulatory information, counted within a region of DNA, or as nodes in a network. Occasionally, sequence-specific and general TFs are distinguished. In Chapter 4, TFs are separated into activators, repressors, and those with variable regulatory sign, and properties of regulatory networks related to these functional classes. In Chapter 5, the functions of specific TFs are examined in relation to a specialized literature from the field of primate immunology. The complete collection of human TFs used within this thesis is provided as Supplementary Table S1.1.

1.2.2 Transcriptional regulatory regions

Regions of the genome significantly enriched in TF binding sites are termed *regulatory modules* and can be distinguished by location relative to target genes (Figure 1.5) (Howard and Davidson 2004). A stretch of nucleotides immediately around the core promoter of a gene is termed its *cis*-regulatory region, and is associated with a particularly high density of regulatory TFs (Xi *et al.* 2007; Juven-Gershon *et al.* 2008). The interval between around 250 to 1000 nt upstream of the core promoter is termed the *proximal promoter region*, and further upstream, the *distal promoter region*. With a looped chromosome conformation, more distant collections of TFs can make efficient contacts with factors in proximal promoter regions (Tolhuis, Palstra *et al.* 2002; Zhao, Tavoosidana *et al.* 2006). Regulatory modules that increase or decrease the rate of transcription are called *enhancers* or *silencers* respectively. Enhancers and silencers are often assumed to lie upstream of the gene regions they regulate. Nevertheless, they can fall within genes, downstream of genes, and on the same or different chromosomes to their targets (Kowalczyk, Hughes *et al.* 2012).

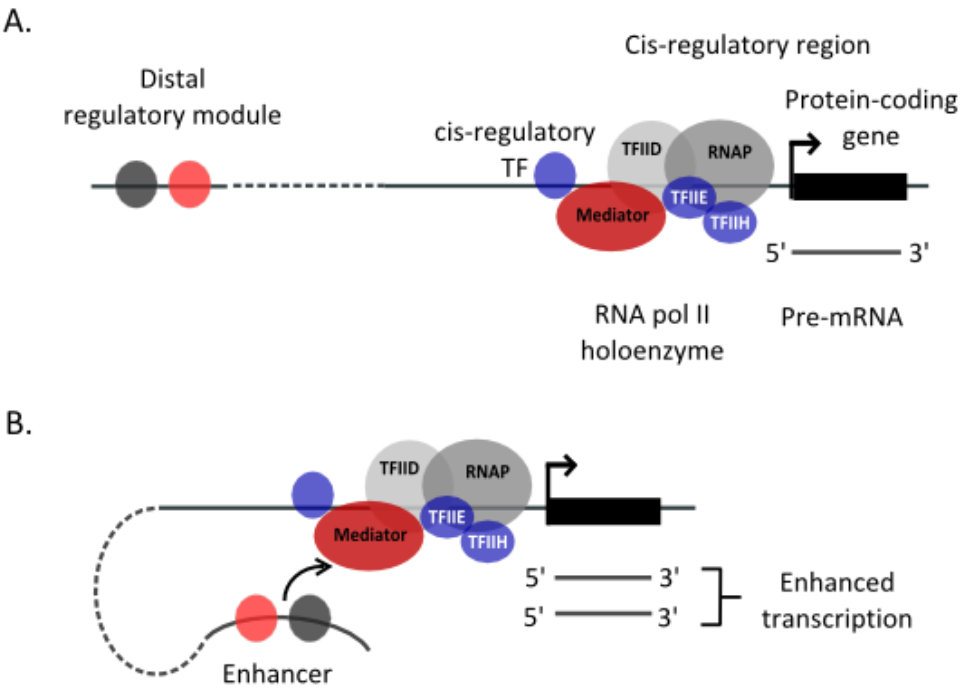


Figure 1.5 Transcriptional regulatory regions.

A. *Cis*-regulatory and distal regulatory regions lying upstream of a protein-coding gene. Within the *cis*-regulatory region, a regulatory TF activates components of the RNA pol II holoenzyme.

B. Chromosome looping brings an enhancer module into proximity with *cis*-regulatory TFs and the RNA pol II holoenzyme, leading to enhanced transcription.

The present work is restricted to data sampled from *cis*-regulatory regions of protein-coding or microRNA genes (Section 1.6.1). Nevertheless, it is highly likely that many regulators from enhancer or silencer regions have been sampled via physical contacts with factors bound to *cis*-regulatory regions (Section 1.9.2).

1.3 The physical architecture of DNA

Chromatin refers to DNA arranged into chromosomes, together with a multitude of proteins bound to these. Within chromatin, the most abundant fraction of proteins are named histones, first isolated by Albrecht Kossel in 1884 (Kossel 1928). Families of histones are conserved throughout eukaryotic organisms, with related molecules in many *archaea*, but they are absent from prokaryotes (Hentschel and Birnstiel 1981; Ouzounis and Kyrpides 1996). Histones are arranged into structures termed nucleosomes, first identified by Roger Kornberg (Kornberg 1974). Nucleosomes repeat along the length of the DNA double helix, with consecutive pairs separated by variable linker regions of around 80 nt of DNA (Kornberg 1974). Each nucleosome

consists of 4 pairs of histones (H2A, H2B, H3 and H4) arranged cubically, around which roughly 150 nt of DNA is looped. This leads to a drop in the extended length of the DNA molecule, to give a more condensed structure. A fifth histone, H1, can also be attached to a pair of adjacent nucleosome core particles, leading to further condensation into what is termed a chromatin *fibre* (Bassett, Cooper *et al.* 2009). During cell division, many more packaging proteins are recruited to chromatin fibres, leading to highly condensed, manoeuvrable chromosomes with the characteristic 'X' shapes that can be seen under a light microscope (Woodcock and Ghosh 2010).

The dynamics of nucleosomes have been shown to provide a rich source of gene regulation. The winding of DNA around the core particle is disruptive to transcription, so that the presence of static nucleosomes in core promoter regions is associated with gene repression (Han and Grunstein 1988; Schones, Cui *et al.* 2008). Nucleosomes can be energetically propelled along DNA in reactions catalysed by a variety of regulatory factors termed chromatin remodelling enzymes, including members of the SMARC family (Whitehouse, Flaus *et al.* 1999; Ramirez and Hagman 2009; Erdel and Rippe 2011). The TF CTCF (CCCTC-binding factor) leads to static nucleosomes which resist clearance and stabilise chromatin (Fu, Sinha *et al.* 2008). Regions with significant CTCF binding are termed insulators, since contacts between enhancers and promoters are blocked, and the spread of a compact chromatin state from neighbouring regions on the DNA strand is resisted (Kim, Abdullaev *et al.* 2007). Thus, regulation of nucleosome dynamics provides a regulatory control point over gene expression.

As well as lateral movements, histone proteins comprising nucleosomes are subject to regulatory covalent modifications, including acetylation, methylation, ubiquitination, and phosphorylation, and removal of any of these (Strahl and Allis 2000). The best studied are acetylation and mono-, di-, and trimethylation of lysine and arginine residues within the tail regions of histones 3 and 4 (Strahl and Allis 2000; Wang, Zang *et al.* 2008). These modifications are catalysed by numerous factors and complexes and can be tissue specific and responsive to environmental signals (Li 2002). Histone modifications are closely tied to regional gene expression, either positively (e.g. trimethylated lysine-4 on histone 3, which is written as H3K4me3) or negatively (e.g. H3K9me3, with the same notational conventions) (Barski, Cuddapah *et al.* 2007; Karlic, Chung *et al.* 2010). At least two general regulatory mechanisms are recognised. First, the modification can alter points of contact between nucleosome and DNA, leading to relaxation or tightening of chromatin (Smith 1991). Second, the modification can anchor protein factors and complexes leading them to interact with the transcriptional machinery (Lachner, O'Carroll *et al.* 2001).

In Chapter 2, we assess the significance of numbers of histone marks deposited over a set of protein-coding genes that have regulatory importance due to their association with microRNA genes (See also Section 2.8.1). In Chapter 3, we show that distributions of microRNA genes are themselves enriched within genomic regions according to the regulatory context defined by chromatin state.

1.4 Alternative transcription and alternative splicing

Protein-coding genes in higher eukaryotes generally express many variant mRNA products, termed isoforms. The RNA pol II holoenzyme can often assemble at more than one location, leading to RNA transcripts with different 5' ends. Alternative transcription start site (TSS) selection can be determined by chromatin state (Maunakea, Nagarajan *et al.* 2010). Termination of transcription is thought to be less precisely regulated than initiation, and significantly differing 3' ends are commonplace (Carninci, Kasukawa *et al.* 2005). As will be described shortly, the 3' end of an mRNA is an extremely important target for regulation during translation (Section 2.5: Post-transcriptional gene silencing). Variation at both ends of gene transcripts is often linked to cell type or cell state, and the variant mRNA and protein isoforms that result can differ functionally (Chiu, Touhalisky *et al.* 2001; Lutz 2008; Kowalczyk *et al.* 2012).

The transcribed product of RNA polymerase II is termed the pre-mRNA, and a collection of pre-mRNAs with variable start and end coordinates from the same coding locus are termed *alternative transcripts* (Figure 1.6A). During or after transcription, the pre-mRNA may undergo further processing before the mature mRNA is produced. This processing includes excision of regions of pre-mRNA termed introns, independently identified in 1977 in the laboratories of Richard Roberts and Phillip Sharp (Berget, Moore *et al.* 1977; Chow, Gelinas *et al.* 1977). Introns are removed by multimeric RNA – protein complexes termed the major or minor spliceosomes, found only in eukaryotes (Hall and Padgett 1996; Saltzman, Pan *et al.* 2011). The non-intronic, expressed regions, termed *exons*, are spliced together during mRNA maturation. Splicing is far more common in multicellular eukaryotes than in unicellular organisms, e.g. with more than 95% of human genes split in this way, but less than 5% of yeast genes (Pleiss, Whitworth *et al.* 2007). Since introns do not code for protein sequences, they serve as a particularly rich source of regulatory elements within genes. In addition, they can harbour functionally autonomous RNA genes including microRNAs and snoRNAs (Fedorova and Fedorov 2003; Rodriguez, Griffiths-Jones *et al.* 2004; Dieci, Preti *et al.* 2009).

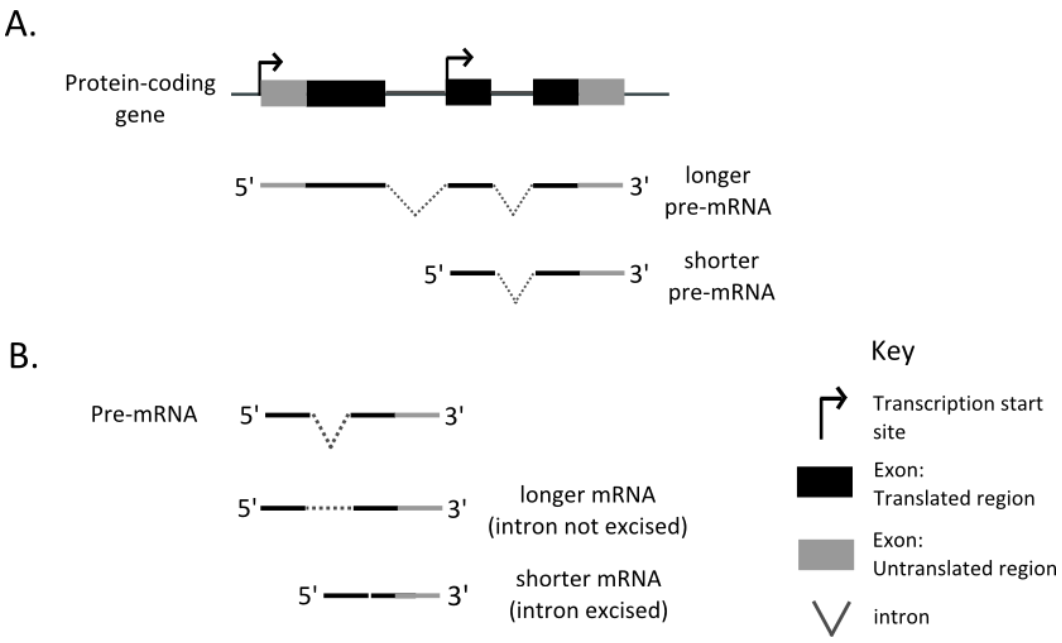


Figure 1.6. Alternative transcription and alternative splicing of protein-coding genes.

A. Alternative transcription. Two distinct pre-mRNA transcripts are transcribed from a common gene region, with transcription start sites at different positions within the gene.

B. Alternative splicing. Two distinct mature mRNAs are expressed from a common pre-mRNA precursor transcript, according to whether or not the pre-mRNA intronic region is excised.

During or after transcription, additional mRNA diversity results from the regulation of splicing. While some introns are removed under all conditions, others are *alternatively* expressed according to regulatory signal (Fagnani, Barash *et al.* 2007). The resulting mRNAs are termed *splice variants* (Figure 1.6B). The selection of alternative exons can be coordinated across functionally related mRNAs, including the collection of mRNAs encoding the splicing factors themselves (Fagnani *et al.* 2007; Saltzman *et al.* 2011). The spliceosome can bind directly to the C-terminal domain of RNA polymerase II, and interact with many other factors involved in transcription (David and Manley 2011). This allows splicing to be linked to the kinetics of transcription. For example, in the expression of the gene CD44, transcriptional elongation, chromatin state modification, and splicing kinetics, are a mutually dependent collection of processes (Ameyar-Zazoua, Rachez *et al.* 2012).

Evidence will be presented in Chapter 2 that the number of transcriptional isoforms of a gene is related to the number of transcriptional regulators of the gene, suggesting (albeit loosely) a connection between regulatory, and organism, complexity.

1.5 RNA processing and transport to the ribosome

Eukaryotic nuclear precursor mRNA transcripts are modified at both the 5' and 3' ends, and are sometimes subject to single nucleotide editing events (Xia, Yang *et al.* 2005). At the 5' end, a methylated guanosine residue is attached and termed the 5'-cap. This protects the transcript from degradation by 5'-exonuclease enzymes, and is a recognition element for enzymes transporting the mRNA from the nucleus to the cytoplasm (Reddy, Singh *et al.* 1992; Sonenberg and Gingras 1998). At the 3' end, excepting pre-mRNAs encoding histone proteins, a run of up to around 250 adenosine residues is added, termed the poly(A) tail (Lewis, Gunderson *et al.* 1995; Davila Lopez and Samuelsson 2008). A number of copies of a regulatory factor, poly(A) binding protein (PABP), attach to the poly(A) tail, and act as chaperones for the mRNA as it is manoeuvred from the nucleus to the cytoplasm (Bernstein and Ross 1989).

The movement of mRNA and associated proteins from nucleus to cytoplasm is biased by ATP-catalysed reactions in favour of export from the nucleus (Vargas, Raj *et al.* 2005). Once the rough endoplasmic reticulum is reached, translation initiation factors recognise the 5'-cap and poly(A) tail, as well as copies of PABP (Borman, Michel *et al.* 2002). At least 12 general translation factors bind the 5' and 3' ends of the mRNA, leading to attachment of the mRNA to the ribosome (Aitken and Lorsch 2012). The ribosome binds the mRNA 5' end, or an internal mRNA binding site, and moves along the mRNA until the message start codon is reached (Gilbert 2010). Polypeptide synthesis continues from the start to the end of the coding region, signalled by a stop codon (Figure 1.1B). Regions of mRNA outside of the polypeptide coding message boundaries are termed the 5' and 3' untranslated regions (5'-UTR and 3'-UTR). Since the 5'-UTR and 3'-UTR do not contain codons, they provide the majority of recognition elements for binding of general and regulatory translation factors (Pichon, Wilson *et al.* 2012).

In prokaryotes, mRNAs often contain several different coding regions, encoding a series of functionally related polypeptide chains. The co-expression of sequential enzymes in a metabolic pathway was anticipated in 1960 by François Jacob and Jacques Monod *et al*. (Jacob, Perrin *et al.* 1960). They defined an operon as a set of genes with a shared expression pattern, under the control of a common operator, or regulator. In an analysis of the genome-wide transcriptome of *Listeria*, under many conditions, more than 60% of genes were found to belong to operons (Toledo-Arana, Dussurget *et al.* 2009). In eukaryotes, however, the expression of many polypeptide chains from a single mRNA is rare. Although eukaryotic genes in common pathways have related expression patterns, this is not usually the result of the arrangement of the genes in a sequence along a chromosome (Niehrs and Pollet 1999).

In Chapters 2 and 3, special classes of eukaryotic transcriptional units (TUs) will be discussed, which may have the character of operons (Section 2.8.1. Transcriptional regulation of microRNAs: see sections on clustered microRNAs, and intronic microRNAs, together with their host genes). In Chapter 2, co-regulation of a well-characterized class of such TUs is discussed. A case is then made for co-regulation of a novel class of linked transcript pairs in Chapter 3.

In chapter 5, we consider how coordinated expression is achieved for collections of genes that are scattered across chromosomes. This is based upon the principle that functionally related genes often have common regulatory elements in their promoters, and this allows them to be co-expressed, both through time and in particular tissues or cell types (Niehrs and Pollet 1999; Borman *et al.* 2002). Dr. Jamie Macpherson's work on clustering of expression patterns is combined with my analysis of transcription factor enrichment and expression patterns, to infer sets of regulators controlling a differential inflammatory immune response in two species of monkeys.

1.6 Post-transcriptional gene silencing

In the early 1980s, studies in the regulation of bacterial plasmid copy number, and then in plants and animals, observed repression of mRNA translation by short RNA transcripts with regions of antisense complementarity to the mRNA (Light and Molin 1982; Light and Molin 1983; Simons and Kleckner 1983; Coleman, Green *et al.* 1984; Ecker and Davis 1986). Often, these were transcribed antisense to a gene region and shown to inhibit the rate of protein synthesis but not transcription. In 1993, Rosalind Lee, Victor Ambros and others showed that this post-transcriptional repression could alter animal body plans, as the null mutant of a repressive RNA in *C. elegans* induced developmental abnormalities (Lee, Feinbaum *et al.* 1993). The repressive RNA was named lin-4, with sequence complementary to short segments within the 3'-UTR of a target mRNA from the LIN14 gene locus, a key regulator of early differentiation in nematodes (Hristova, Birse *et al.* 2005). The LIN4 gene gave rise to two RNA transcripts, of 61nt and 22nt in length, but the repressive action of lin-4 was pinpointed to the shorter transcript. This was the first time a developmentally important microRNA had been observed, with clues to the microRNA biogenesis pathway (Bartel 2004). Repression of translation of a target mRNA by complementary regulatory RNAs was termed post-transcriptional gene silencing (PTGS). It would be another 7 years before the discovery of the second microRNA in *C. elegans*, let-7, and the deletion of let-7 proved fatal to the organism (Rougvie 2001). Today hundreds of microRNA families are recognised, across all life, and

contributing to the post-transcriptional regulation of the majority of protein-coding genes within metazoa (Bartel 2004; Griffiths-Jones *et al.* 2006; Friedman, Farh *et al.* 2009).

1.6.1 MicroRNAs

MicroRNAs are short 20 – 24 nt RNAs which bind stretches of complementary nucleotides of target mRNAs, most often within the 3'-UTR, leading to inhibition of translation (Bartel 2004). The number of microRNA genes in humans is roughly 8% of the total number of protein-coding genes (see e.g. miRBase v.19 and RefSeq v.57 protein-coding genes)(Griffiths-Jones *et al.* 2006; Pruitt, Tatusova *et al.* 2007). MicroRNAs have been identified in significant numbers within numerous animal and plant species, as well as within viral genomes and in some unicellular organisms (Griffiths-Jones *et al.* 2006).

1.6.1.1 MicroRNA gene arrangements

MicroRNA genes are expressed as primary transcripts from a variety of genomic loci. They are commonly classified according to their associations with protein-coding gene transcripts, as (i) *intergenic*, with no pre-existing gene annotation on the strand surrounding the microRNA precursor sequence, (ii) *intronic*, residing within introns of genes of either proteins or noncoding RNAs, or (iii) *exonic* when overlapping exon regions of longer ncRNA transcripts or of protein-coding exons (Figure 1.7) (Rodriguez *et al.* 2004). Occasionally, the microRNA precursor gene sequence may span the entire length of an intron, and is then termed a mirtron (Berezikov, Chung *et al.* 2007; Okamura, Hagen *et al.* 2007). The primary transcript of a microRNA gene region is termed the pri-miRNA, and as indicated in Figure 1.7A, this may be the same RNA molecule as a protein-coding host gene pre-mRNA. Alternatively the pri-miRNA can be autonomously transcribed, from a microRNA-specific promoter region either within a protein-coding gene region, or within intergenic regions (Figures 1.7B and 1.7C). MicroRNA genes can also be clustered as sets of precursor sequences that are expressed within the same primary microRNA transcript (pri-miRNA), requiring separate reactions to excise each microRNA precursor (Figure 1.7B) (Altuvia, Landgraf *et al.* 2005). A set of clustered microRNAs can form an operon, since they are co-regulated and often target mRNAs within the same cellular processes (Kim, Yu *et al.* 2009; Merchan, Boualem *et al.* 2009).

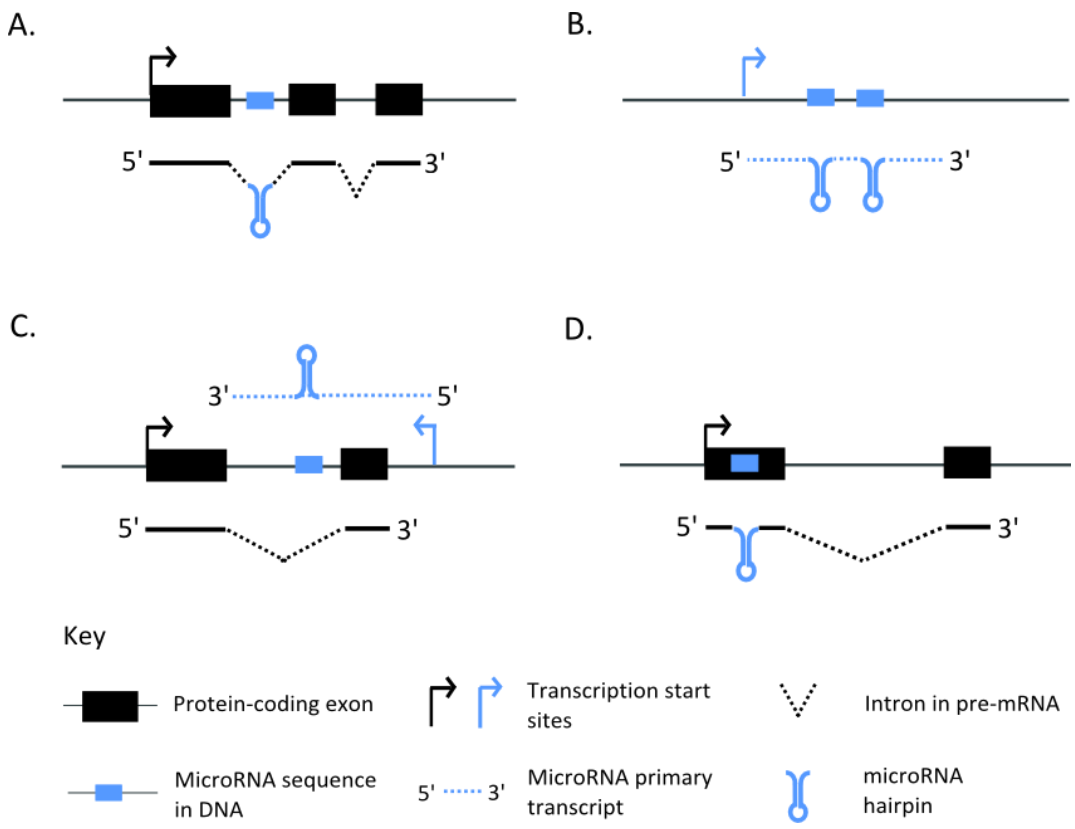


Figure 1.7 Arrangements of microRNA genes.

A. Intronic microRNA
B. Intergenic microRNA (cluster of 2 precursor sequence
C. Opposite intronic microRNA
D. Exonic microRNA (infrequent)

Throughout chapters 2 and 3, we mainly examine the transcriptional regulation of protein-coding host genes of intronic microRNAs, and of protein-coding neighbouring genes of intergenic microRNAs.

1.6.1.2 MicroRNA biogenesis

The microRNA biogenesis pathway is a highly conserved multistep process (Figure 1.8), and its intermediates serve as a key means for distinguishing microRNAs from other types of noncoding RNA (Filipowicz, Bhattacharyya *et al.* 2008; Ding, Weiler *et al.* 2009). There are minor differences between plants and animals, and for the plant biogenesis pathway, see for example a review by Chen *et al*. (Chen 2005). The key steps in animals are expression from the microRNA gene locus of the pri-miRNA, excision from this of a hairpin precursor (pre-miRNA), the export of this hairpin from the nucleus, and then excision of the loop part of the hairpin structure to give 1 or often 2 mature microRNA strands (Bartel 2004).

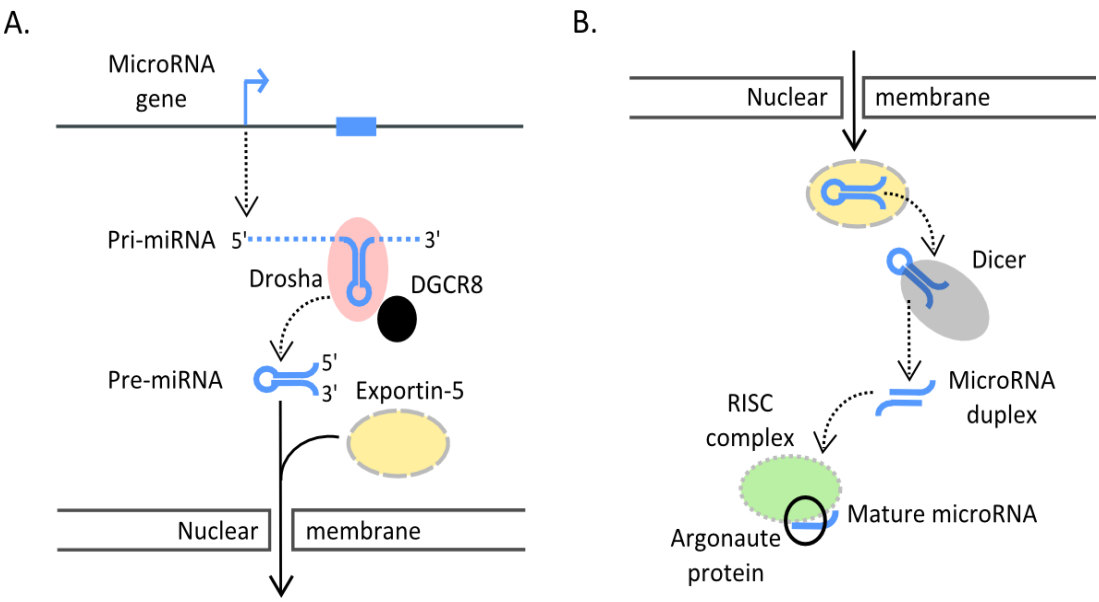


Figure 1.8 The microRNA biogenesis pathway in animals.

A. MicroRNA primary transcription and nuclear export of the pre-miRNA hairpin.
B. MicroRNA maturation and Argonaute binding within the cytoplasm.

The key steps in microRNA biogenesis are:

(i) transcription from a microRNA gene locus, of a long 100s - 10,000s nt primary transcript termed the pri-miRNA. This is catalysed by RNA polymerase II and subject to all the forms of transcriptional regulation described for protein-coding genes. In rare instances, microRNAs may be transcribed by RNA polymerase III (Borchert, Lanier *et al.* 2006).

(ii) excision from the pri-miRNA of one or more imperfectly base-paired 60 – 70 nt hairpin structures termed pre-miRNAs. The hairpins themselves are recognised by DGCR8 in mammals, or the homolog Pasha in fruit flies and nematodes. Excision of the hairpin is catalysed by a nuclear RNase III endonuclease, Drosha, which together with DGCR8/Pasha is called the *microprocessor complex* (Denli, Tops *et al.* 2004; Gregory, Yan *et al.* 2004). Processing of the pri-miRNA occurs rapidly, and often concurrently with transcription (Morlando, Ballarino *et al.* 2008). In the case of mirtrons, then the pre-miRNA is generated by splicing, independently of the microprocessor complex (Berezikov *et al.* 2007; Okamura *et al.* 2007).

(iii) transfer of the expressed pre-miRNA hairpins from nucleus to cytoplasm by exportin-5, together with other proteins (Yi, Qin *et al.* 2003). The 61 nt fragment expressed from the lin-4 locus in *C. elegans* is now recognised as a pre-miRNA (Lee *et al.* 1993). It is this stem-loop hairpin precursor, of an approximately uniform length, that best serves to distinguish microRNAs from other species of short noncoding RNA.

(iv) excision of the loop part of the pre-miRNA by Dicer, to leave a ≈ 22nt RNA duplex formed from the stem arms, with 2 nt overhangs at the 3' ends (Bartel 2004). The two strands forming a duplex are termed the 3' and the 5' *arms* according to which end of the pre-miRNA hairpin they derive from.

(v) uptake of one half of this duplex, which forms the mature 20 – 24 nt microRNA sequence into the RNA-induced silencing complex (RISC), including an Argonaute protein family member (Fabian and Sonenberg 2012). The 22 nt fragment encoded by the LIN4 gene in *C. elegans* can be recognised today as the mature lin-4 microRNA (Ambros, Lee *et al.* 2003).

Mature microRNAs are taken up by the RNA-induced silencing complex (RISC), including an Argonaute protein family member (Figure 1.8B) (Fabian and Sonenberg 2012). Selection of the correct microRNA strand for uptake into the RISC, from a nearly symmetrical RNA duplex, provides a puzzle which is still not fully resolved. The RNA duplex may have differential stability along its length, inducing an orientation and a favoured strand: In this case, the microRNA strand that combines with the RISC is referred to as the *guide*. The other strand is termed the *passenger*, and will sometimes be associated with an alternative AGO protein, or be degraded (Czech, Zhou *et al.* 2009; Okamura, Liu *et al.* 2009; Ghildiyal, Xu *et al.* 2010). With deep-sequencing datasets revealing lower abundance RNA molecules, and more extensive sampling of cell types and developmental stages, it is becoming clear that both microRNA strands are often functional (Yang, Phillips *et al.* 2011). The strand that is functional sometimes switches between species (de Wit, Linsen *et al.* 2009; Griffiths-Jones, Hui *et al.* 2011).

1.6.2 Mechanisms of microRNA-mediated gene repression

The association between microRNA and Argonaute protein leads to the RISC complex being guided to target mRNAs. In animals, target recognition between a microRNA and an mRNA generally permits bulges and mismatches, while in contrast, closely related small-interfering RNAs (siRNAs) are perfectly complementary to their RNA targets. This is because siRNAs are often derived directly from copies of the viral genomes to which they subsequently bind. Perfect recognition between an siRNA and viral genomic sequences ensures that these are destroyed by an endonucleolytic AGO (slicer) protein bound to the siRNA (Ghildiyal, Seitz *et al.* 2008; Bartel 2009; Llave 2010). For animal microRNAs, repression of the target RNAs is rarely based upon slicer activity. Once a microRNA has guided the RISC to a target, translation can be inhibited in one of many ways, for example: (i) the RISC may interfere with initiation or post-initiation steps, or (ii) signal the target mRNA for 5’ - 3’ degradation via deadenylation or 5’-cap removal, or (iii) signal the target mRNA for removal from the translation pool for storage, or degradation (Bhattacharyya, Habermacher *et al.* 2006; Nilsen 2007; Filipowicz *et al.* 2008; Flynt and Lai 2008; Friend, Campbell *et al.* 2012). In total, there may be more than 10 different repressive mechanisms, though not all of these are validated (Morozova, Zinovyev *et al.* 2012). Some of the possible mechanisms of microRNA-mediated repression are shown in Figure 1.9. Rarely, a microRNA upregulates a target mRNA (Ghosh, Soni *et al.* 2008).

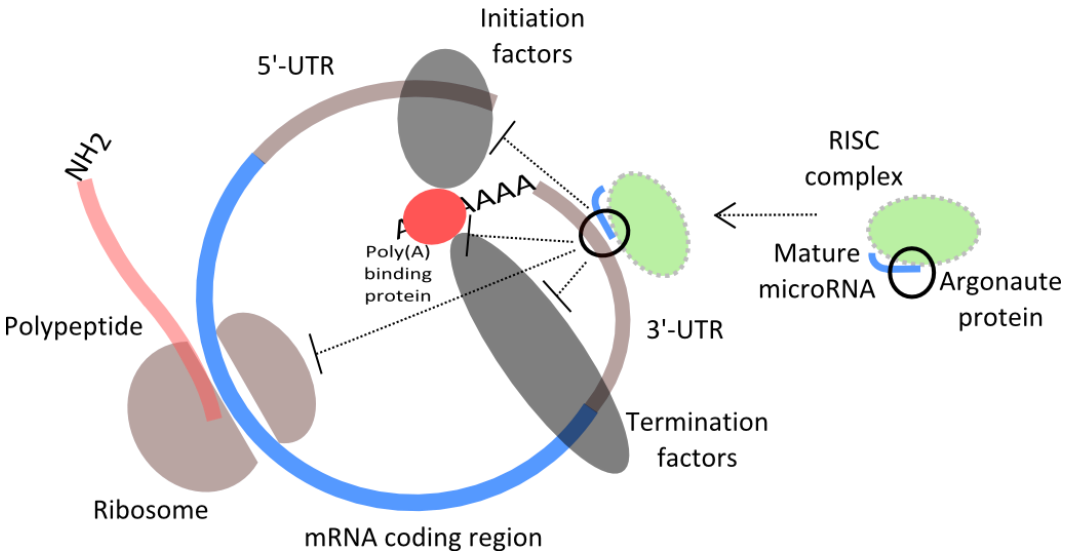


Figure 1.9 MicroRNA and RISC complex binding to the 3'-UTR of a target mRNA.

The figure displays a schematic of eukaryotic mRNA translation, adapted from (Morozova *et al.* 2012). A ribosome is progressing through the mRNA coding region, directing translation of a novel polypeptide chain. The Poly(A) binding protein (PABP) is bound to the 3' end of the mRNA. The mRNA is often circularized through interactions between the initiation and termination factors that bind to PABP. A microRNA has guided a RISC complex to its target site in the 3'-UTR region of the mRNA. This leads to repression of translation, shown by the hammerhead arrows, through a range of mechanisms.

Several lines of evidence indicate that microRNAs recognise their target mRNAs through complementary base-pairing between the mRNA and the terminal 8 nucleotides at 5' end of the microRNA (Doench and Sharp 2004; Grimson, Farh *et al.* 2007; Bartel 2009). In mutagenesis studies, microRNA functions were affected much more significantly by mutations introduced at the 5' end of the microRNA (Doench and Sharp 2004). A microRNA 5' end mediated interaction is also supported by structural analysis of Argonaute proteins, since the 3' ends of microRNAs are wound into the Argonaute PAZ domain (Ma, Ye *et al.* 2004). Thus, only the 5' end of the microRNA is free to bind target mRNAs. Prediction of microRNA binding sites therefore always focuses upon the 5' end of the microRNA (John, Enright *et al.* 2004; Grun, Wang *et al.* 2005; Krek, Grun *et al.* 2005; Friedman *et al.* 2009). A short sequence of 6 - 8 nucleotides at the 5' end of a microRNA (typically from the 2nd to the 8th nt) is termed the *seed* sequence, and target sequences within mRNAs are termed *seed matches*. Subsets of the seed sequence can also induce repression of translation, though generally not as effectively (Grimson *et al.* 2007). The seed region is the most conserved portion of a microRNA between members of the same microRNA family (Lall, Grun *et al.* 2006).

The repression of mRNAs by microRNAs depends upon factors in addition to complementary pairing between microRNA seeds and their target seed matches (Grimson *et al.* 2007; Saetrom,

Heale *et al.* 2007). The secondary structure of the target mRNA affects access of the RISC complex to target sites (see Figure 1.9) (Kertesz, Iovino *et al.* 2007). In principle, thermodynamics of the microRNA – mRNA duplex region affects the stability of the microRNA – mRNA binding interaction (John *et al.* 2004; Grimson *et al.* 2007). Where a single mRNA is targeted by more than one microRNA, then the strength of repression is linked to the distance between the binding sites (Saetrom *et al.* 2007). When their binding sites are spaced far apart, the combined repressive effect due to two microRNAs is roughly additive. However, when the binding sites are closer together, the combined repressive effect is often greater than additive, which might suggest a synergistic interaction between adjacent RISC complexes (Grimson *et al.* 2007). These features have been incorporated into a variety of microRNA target site predictors (Rhoades, Reinhart *et al.* 2002; John *et al.* 2004; Krek *et al.* 2005; Bartel 2009). In addition, microRNA binding is highly promiscuous across many target mRNAs (Friedman *et al.* 2009), and thus the circulating levels of mRNAs with competing target sites can mediate the effectiveness of microRNA-mediated repression. In particular, the repressive effect of microRNAs may be suppressed by species of RNA which are significantly enriched in target sites, and thus act as sponges (or sinks) for microRNAs (Franco-Zorrilla, Valli *et al.* 2007; Hansen, Jensen *et al.* 2013).

Calculations within Chapter 5 show that mean expression level of mRNAs across human tissues is related linearly to the rank of the mRNA when sorted by number of microRNA binding sites. We show also that patterns of connection between microRNAs, TFs and their common target genes vary with microRNA seed type.

Because microRNAs are short, applying simple rules and searching within a target space of millions of nucleotides of 3’-UTRs leads to a significant number of chance matches. A common way to measure the frequency of chance matches is to predict target sites of microRNAs with shuffled sequences (Lewis, Shih *et al.* 2003). Since shuffled microRNAs are artificial, their predicted seed matches are clearly accidental, so the number of such matches provides an estimate of the false positive rate for the predictor. This test has indicated that at least 50% of seed matches may be functionless. High rates of functionless sites have also been estimated from microRNA perturbation experiments, where predicted targets often show no responsiveness to the microRNA (Grimson *et al.* 2007; Baek, Villen *et al.* 2008). Nevertheless, a strong repressive effect might be distributed across many weakly bound microRNA-RISC complexes. Indeed, co-expressed microRNAs are more likely to target common genes (Xu, Li *et al.* 2011). According to this picture, target prediction might reflect better the nature of the RISC-complex milieu within a pool of ribosomes, when applied to collections of microRNAs mapped to shared targets, rather than to individual microRNA – mRNA pairs. Target predictors

often reflect this concept, implicitly, by increasing the score of a predicted binding site when it occurs in conjunction with other sites (Lewis *et al.* 2003; John *et al.* 2004; Krek *et al.* 2005).

In this thesis, the emphasis to some extent is placed more strongly upon the binding sites of transcriptional regulators, rather than of microRNAs. Nevertheless, we have utilised two well known methods of microRNA target prediction. The first and simplest is based on searching for exact matches to different kinds of seed sequence (detailed within Section 4.2). We also used a microRNA target predictor, miRanda, that takes into account, among other factors, relaxations to Watson-Crick base-pairing rules specific to RNA duplexes (wobble base-pairing rules), differential repression effectiveness throughout the length of the seed region, and predicted thermodynamic stability of the microRNA : mRNA duplex (Enright, John *et al.* 2003; John *et al.* 2004).

1.7 Protein folding and post-translational modification

To form a functional protein, one or more polypeptides must be folded into the correct 3-dimensional structures. The process of folding usually occurs spontaneously, but is dependent upon properties of the intracellular solution (pH, temperature, salinity, etc.) (Anfinsen 1973; Das and Baker 2008). In some cases, proteins fold incorrectly or inefficiently unless bound by molecular chaperones. Folding can be regulated co-translationally, and unsuccessfully folded proteins are often targeted for degradation (Hagiwara and Nagata 2012). In principle, interactions between the process of RISC assembly, the ribosome, and the rate of production of a polypeptide, might determine whether a protein is degraded. Examples of this are not yet known (Morozova *et al.* 2012), but relationships have been found between microRNA targeting, mRNA decay rate, and the proportion of a polypeptide that is structurally disordered (Edwards, Lobley *et al.* 2009; Schad, Tompa *et al.* 2011).

Proteins are further subject to co-translational and post-translational modifications (PTMs), increasing the diversity of functional products of a single coding region. Typically, PTMs are covalent attachments to proteins of small molecules, commonly phosphate groups, acetyl and methyl groups, ubiquitin, carbohydrates, or lipid moieties (Prabakaran, Lippens *et al.* 2012). Examples include regulatory acetylation and methylation marks linked to the histone family of proteins, together with phosphate groups attached to TFs under the control of signal transduction cascades (Strahl and Allis 2000; Pearson *et al.* 2001).

1.8 Crosstalk between the regulators of gene expression

Communication between general classes of regulators or control points is termed *crosstalk* (Klaus, Bijsterbosch *et al.* 1987). There is an effectively limitless array of potential feedback circuits linking regulators and targets of various types. Examples include: crosstalk between transcription termination and post-transcriptional gene silencing, through the redefinition of microRNA – mRNA interactions arising from variable length 3'-UTRs (Ghosh *et al.* 2008); alternatively, crosstalk occurs between post-translational modification and transcription, as transcription factors are conditionally activated according to their PTMs (Section 1.2.1); as a final example of crosstalk, signalling cascades that activate TFs affect microRNA biogenesis, via post-translational modifications to Dicer or to partners of Drosha, mediated by Ras/MAPK and TGF-β/SMAD signalling respectively (Saj and Lai 2011; Blahna and Hata 2013). In the present study, we are particularly concerned with crosstalk between the transcriptional and microRNA-mediated post-transcriptional regulatory layers. This requires identification of the microRNA regulators of each TF (as above: Section 1.6), the transcriptional regulators of each microRNA, and their combined actions upon protein-coding gene expression pathways.

1.8.1 Transcriptional regulation of microRNAs

The annotation of the promoter regions of protein-coding genes has a long history, with both experimental and computational methods proving successful (Pedersen, Baldi *et al.* 1999; Megraw, Pereira *et al.* 2009; Gupta, Wikramasinghe *et al.* 2010; Takahashi, Kato *et al.* 2012). The task is more challenging for microRNA genes, since autonomous pri-miRNAs are typically degraded more rapidly than pre-mRNAs, and the DNA sequence between 5' end and microRNA precursor lacks common sequence or structural motifs (Rodriguez *et al.* 2004; Morlando *et al.* 2008). Unlike mRNAs, therefore, pri-miRNA transcripts with readily accessible 5' and 3' ends are rarely detected in nuclear RNA samples. Intronic microRNAs that are excised from host gene pre-mRNA transcripts must clearly be under the control of the protein-coding gene promoter, so that their transcription would be driven by RNA pol II. From a handful of carefully characterized cases, pri-miRNA transcripts expressed from microRNA-specific promoter regions also bear hallmarks of transcription by RNA pol II, including poly(A)-tail and 5'-cap, and lengths are typical of RNA pol II transcripts (Cai, Hagedorn *et al.* 2004; Lee, Kim *et al.* 2004; Cui, Xu *et al.* 2009). The 5' ends of autonomous microRNA genes are therefore inferred from a range of genomic features associated with RNA pol II promoter regions (Megraw, Baev *et al.* 2006; Saini, Griffiths-Jones *et al.* 2007; Zhou, Ruan *et al.* 2007; Marson, Levine *et al.* 2008; Ozsolak, Poling *et al.* 2008; Corcoran, Pandit *et al.* 2009; Megraw *et al.* 2009; Wang, Xuan *et al.*

2009). The consensus from these analyses suggests that the majority of primary transcripts are likely to be of an intermediate length (1 - 20 kb). Reliable annotation of microRNA promoters is one of the most important current problems in microRNA research.

1.8.2 Integrated regulatory network of transcription factors and microRNAs

Once transcriptional regulators have been assigned to microRNA and protein-coding genes, and microRNAs have been assigned to target mRNAs, it is natural to ask what kinds of regulatory circuits result. A number of experimentally validated and well-characterized circuits are shown in Figure 1.10. A simple case is the negative feedback loop, where a regulator represses itself, via a pathway upstream of its own expression. For example, in Figure 1.10A, the microRNA let-7 has a conserved target site in the 3'-UTR of Dicer, which as we have seen is a key protein in the microRNA biogenesis pathway (Figure 1.8) (Tokumaru, Suzuki *et al.* 2008). It is important to maintain cytoplasmic copies of this microRNA within an appropriate range, since let-7 mutants are nearly always lethal (Maller Schulman, Liang *et al.* 2008). In Figure 1.10B, a more complex negative feedback system is shown. The cell-cycle regulating TFs MyC, and E2F family members, activate one another, as well as driving expression of a microRNA cluster mir-17 ~ 92 (which contains 6 pre-miRNA hairpins, mir-17/-18a/-19a/-20a/-19b/-92-1). Members of the cluster target both MyC and each of the E2F family members, leading to negative feedback (Aguda, Kim *et al.* 2008). The circuit dynamics resulting from this set of connections depend on many parameters, including initial concentrations of the TFs and microRNAs, relative strengths of the activators or repressors, response times of their targets, and external perturbations (Aguda *et al.* 2008). Nevertheless, put simply, these negative feedback loops reduce extreme fluctuations.

Autoregulation of a microRNA by itself has not yet been described, although it may indeed target the the mRNA transcript of its own host gene further downstream in the expression pathway. By contrast, many TF repressors do bind their own promoters. This is the case for the neural cell differentiation regulating TF Hes1, in Figure 1.10C. The effect is an oscillating expression pattern for the gene, alternately accumulating, and shutting down (Bonev, Stanley *et al.* 2012). In turn, Hes1 is placed in reciprocal repression with the microRNA, mir-9, so mir-9 is periodically expressed during the parts of the cycle when Hes1 levels are low. Since Hes1 has a shorter half-life in the cell than mir-9, mir-9 accumulates and at a critical point drives Hes1 into a permanently repressed state. This leads a neural progenitor cell to the irreversible commitment to differentiate (Bonev *et al.* 2012).

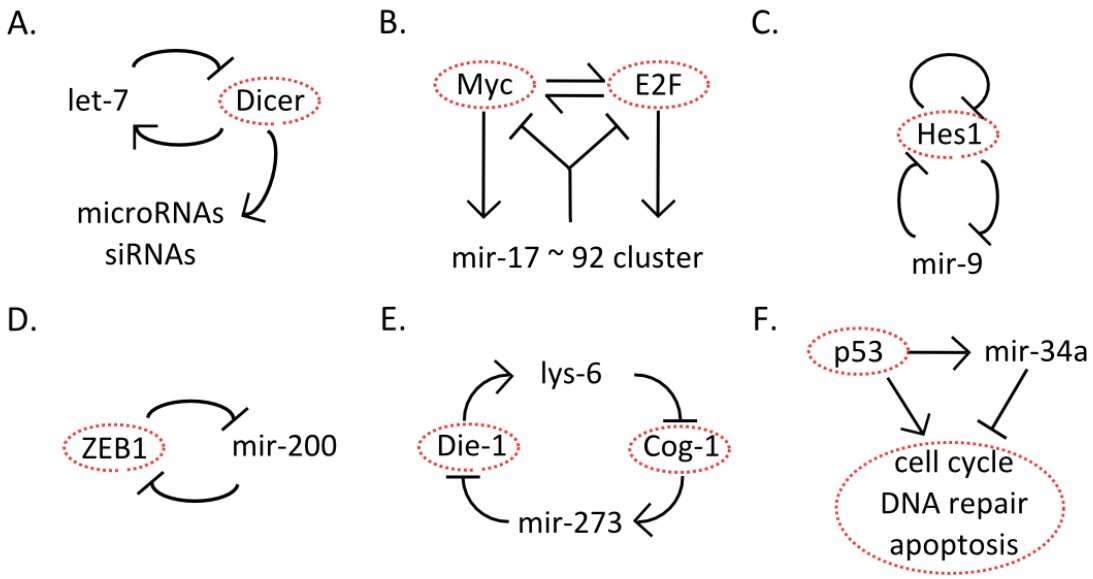


Figure 1.10 Varieties of feedback and feedforward circuits containing microRNAs.

A. Negative feedback by microRNA let-7 on microRNA biogenesis pathway via Dicer
B. Complex negative feedback circuit of c-Myc, E2F family members, and the mir-17~92 cluster
C. – E. Reciprocal repression between microRNAs and TFs as a theme in developmental switching
F. Incoherent feedforward loop over common target genes of a TF upstream of a microRNA.

Red circles indicate transcription factors, together with the Dicer protein, and protein-coding targets of p53 and mir-34a. All other regulators are microRNAs.

Reciprocal repression can lead to one of two clearly distinct final states, and is a common feature of regulatory circuits with a developmental outcome. Figure 1.10D shows the reciprocal repression between ZEB1 and mir-200. The balance between these two regulators contributes to the determination of epithelial-to-mesenchymal transition in the developing embryo (Burk, Schubert *et al.* 2008). Figure 1.10E displays a more complex case, from *C. elegans*, with the same underlying structure: The microRNAs lys-6 and mir-273, respectively, repress the transcription factors Cog1 and Die1, which in turn activate the other microRNA. The arrangement is kinetically balanced, and leads to the permanent silencing of one of the pairs lys-6/Cog1 or mir-273/Die1, chosen randomly, together with commitment of a nematode neural cell to terminal differentiation (Johnston and Hobert 2003). The unpredictability of the final state explains why the corresponding neural cell phenotype in nematodes is not inheritable (Hobert 2006).

MicroRNAs intimately connected with cell fate commitment or proliferation rates are likely to be essential for organism fitness. However, the functions of most microRNAs probably cannot be pinpointed to critical developmental events. In support of this view, it has been shown within nematode worms that only 20% of microRNA null mutants suffer obvious abnormalities (Miska, Alvarez-Saavedra *et al.* 2007). The percentage has since been revised upwards, as the null mutant worms have been exposed to a greater number of environmental conditions, suggesting some microRNAs have protective functions during periods of environmental stress (Brenner, Jasiewicz *et al.* 2010). It remains perplexing however why many microRNAs that are conserved appeared dispensable in these experiments. A theory which might explain this is that many microRNAs act collectively to fine-tune or to stabilise protein-coding gene expression levels (discussed briefly within Section 1.6.2) (Herranz and Cohen 2010). To detect this function might require the parallel knock-down of several microRNA genes.

The apparent dispensability of many individual microRNAs has led to fresh perspectives on the functions of the microRNA-mediated regulatory layer. A popular line of argument is to infer novel microRNA functions from global patterns of connection with transcription factors within an integrated regulatory network (IRN). For example, Figure 1.10F displays a feedforward loop where the TF p53 activates mir-34a, and both regulators target many of the same genes, in processes such as the cell cycle, DNA repair, and apoptosis pathways. This feedforward loop (FFL) is incoherent, since the repressive signal from the microRNA contradicts the activating signal from the TF (Chang, Wentzel *et al.* 2007). Incoherent FFL patterns with TFs lying upstream of microRNAs, and sharing common target gene sets, have been shown to be widespread in a number of animals (Shalgi *et al.* 2007; Yu, Lin *et al.* 2008; Cheng, Yan *et al.* 2011; Gerstein *et al.* 2012). This kind of regulatory motif can lead to dampened oscillations within the mRNA and protein product outputs of the pathway (Herranz and Cohen 2010; Osella, Bosia *et al.* 2011). This effect may be distributed over many expressed microRNAs, so that a significant number would need to be deleted from the genome before an impact upon organism fitness became evident. Thus, the microRNA-mediated regulatory layer might function as a generalized dampener of oscillations in protein-coding gene expression pathways.

Throughout Chapters 2 - 4, we focus particularly upon crosstalk between transcriptional DNA-binding regulators, and post-transcriptional microRNA regulators. As mentioned above, feedback between signalling pathways and microRNA biogenesis is possible. In chapter 4, we identify a pathway that is highly regulated by TF – microRNA circuits, and discover from the literature that this pathway has been singled out as a key regulator of Dicer (see also Figure 1.8B). We examine this and related regulatory circuits in human, and examine links between TF

– microRNA feedforward regulation, and the stability of transcript expression levels across human tissues.

1.9 Genome-wide measurement

Given significant short-range and long-range interactions between genomic regions, it is important to be able to consider the genome as an integrated whole. To meet this challenge, significant progress has been made towards capturing genome-wide data in a single experiment. One of the principal tasks of computational biology is to make sense of this type of data. We will consider (1) how expression levels of genes and RNAs are measured and (2) how genome-wide locations of DNA binding proteins can be assessed.

1.9.1 Gene expression

The measurement of gene expression levels within a tissue or single cell requires simultaneous quantification of thousands of different RNA species, and is thus an ambitious goal (Toledo-Arana *et al.* 2009). Just 4 decades ago, there was no general method to determine the sequence of nucleotides of a DNA or RNA molecule. The publication in 1972 of the 474 nucleotide sequence of the mRNA coding for bacteriophage MS2 coat protein was a ground-breaking accomplishment (Min Jou, Haegeman *et al.* 1972). Today, the fastest sequencing machines allow sequencing of hundreds of millions of base pairs in a day (Margulies, Egholm *et al.* 2005). Two fundamentally different principles underpin most of the history of the modern technologies. The first principle depends upon hybridization of unknown nucleic acid sequences to complementary nucleic acids with known sequence, termed probes. This gave rise to the Southern blot, after Edward Southern (Southern 1975). The second principle is based upon *de novo* nucleic acid polymerisation against a template sequence, with chemical chain terminators specific to each of the four nucleotides in turn. The resulting collection of nucleic acid fragments is size fractionated. By measuring the final incorporated base within each fraction, the complete sequence of nucleotides to be deduced. Such methods were developed in parallel by many researchers, but it was a technique invented by Fred Sanger in 1977, using dideoxynucleotide chain-terminators, that dominated the first significant waves of gene sequencing (Sanger, Nicklen *et al.* 1977).

In their original forms, these approaches allowed researchers to examine one by one genes or RNAs or interest. By 1997, significant progress was made towards genome-wide measurement of gene expression, through adhering 1000s of known sequence probes to a single chip, termed a *microarray chip* (Maskos and Southern 1992; Lashkari, DeRisi *et al.* 1997). A sample

of DNA sequences is fluorescently labelled, hybridised to the microarray, and the fluorescent intensities at each spot in the array indicate roughly the concentration of complementary sequences in the sample. The technology can be applied to any sample of DNA sequences, including whole cell cDNA libraries. Complete genomes were known for a number of simple organisms including *E. coli* and yeast (*S. cerevisiae*). Microarrays were designed with probes tiling thousands of genomic locations, or their entire genomes (Goffeau, Barrell *et al.* 1996; Blattner, Plunkett *et al.* 1997). For more complex genomes, with billions of nucleotide base pairs, either multiple chips are required, or probe locations have to be spaced apart along chromosomes (Russo, Zegar *et al.* 2003).

In the present work, we have made extensive use of microarray-derived protein-coding expression measurements from many human tissues (Chapters 2 and 4), through the Novartis atlas (Su, Wiltshire *et al.* 2004), and from monkey CD4+ T cells (Chapter 5) (Jacquelin, Mayau *et al.* 2009). Although highly reproducible, microarray technology has begun to be superseded by newer approaches based upon direct sequencing of millions of DNA or RNA fragments (Heintzman, Stuart *et al.* 2007). Unlike the microarray, the set of DNA sequences that can be detected is not predefined. In effect, as microarrays parallelize Southern blotting, these simultaneous sequencing technologies, termed *next generation*, *ultra high-throughput*, or *deep sequencing*, parallelize the principle of Sanger sequencing (though the chemistry is distinct). Like Sanger sequencing, the technology rests upon *de novo* synthesis of DNA complementary to fragments in the sample, but successively incorporated nucleotides are determined optically rather than chemically (Sanger *et al.* 1977; Margulies *et al.* 2005). The method is fast, relatively inexpensive, produces reads of sufficient length for *de novo* genome assembly (up to 700 bp) and has high accuracy in general though performs poorly for homopolymeric sequences (Lysholm 2012). While we have not utilised deep sequencing datasets for the measurement of gene expression levels, we have made extensive use of regulatory datasets derived using some of the same underlying technology (see below).

1.9.2 Proteins in contact with DNA

A novel method for identifying DNA regions bound by a protein was devised in 1984 by John Lis and David Gilmour (Gilmour and Lis 1984). Termed *chromatin immunoprecipitation (ChIP)*, the experimental steps consist in adding antibody specific to the protein, cross-linking protein to DNA, shearing DNA into fragments, precipitating the antibody together with its target protein and bound DNA, reversing cross-links, then finally purifying and examining sequences of the DNA fragments within the sample (Figure 1.11). At the time, DNA fragments within a sample were assayed by means of complementary probes, as in a Southern blot, to test which

part of gene region each fragment derived from (Southern 1975). The invention of microarray chips provided a means to identify bound DNA fragments genome-wide. When applied to DNA fragments derived from ChIP experiments, a microarray DNA sequence assay is termed ChIP-chip (Lee, Rinaldi *et al.* 2002; Heintzman *et al.* 2007). When mapped to the genome, DNA sequences from a ChIP experiment tend to cluster on the sense and antisense DNA strands either wide of the real TF binding site, with the interpolated maximum read density closely related to the true binding site location (Johnson, Mortazavi *et al.* 2007). This pattern is called a *peak* and algorithms have been developed to automate the identification and scoring of peaks, and inference of TF binding sites from these, e.g. (Zhang, Liu *et al.* 2008; Rozowsky, Euskirchen *et al.* 2009).

In Chapter 5, we consider the impact of TF peak quality scores on derived network properties. However, when compared with the parallel case of testing different stringency microRNA seed matches, network properties were relatively insensitive to TF peak quality scores.

More recently, the ChIP-chip protocol has been superseded by methods with higher resolution, by combining ChIP with deep-sequencing, in a protocol termed ChIP-seq. The first example of ChIP being combined with high-throughput DNA sequencing dates from 2007 in a survey of the DNA binding sites of the human transcriptional repressor REST (NRSF) in Jurkat T cells (Johnson *et al.* 2007). Many of the factors that have been mentioned in this introduction, including GTFs, NF-kB, STAT proteins, CTCF, HSF and others, have already been subjected to ChIP-seq within a variety of human cell lines. The number of published ChIP-seq studies has grown significantly year on year (Barski *et al.* 2007; Wang *et al.* 2008; Gerstein, Lu *et al.* 2010; Roy, Ernst *et al.* 2010; Gerstein *et al.* 2012). Much of this work has been coordinated by the ENCODE consortium, which is undertaking a long-term project to measure systematically locations of regulatory marks within the human genome (Celniker, Dillon *et al.* 2009). ChIP-seq datasets relating to round 8% of the confidently annotated DNA binding proteins in humans are examined within this thesis. The work presented here assumes that this $\approx$ 8% representation of the total TFs in human and monkey is sufficient to extrapolate properties of the complete but presently unobservable space of transcriptional regulatory interactions within these species.

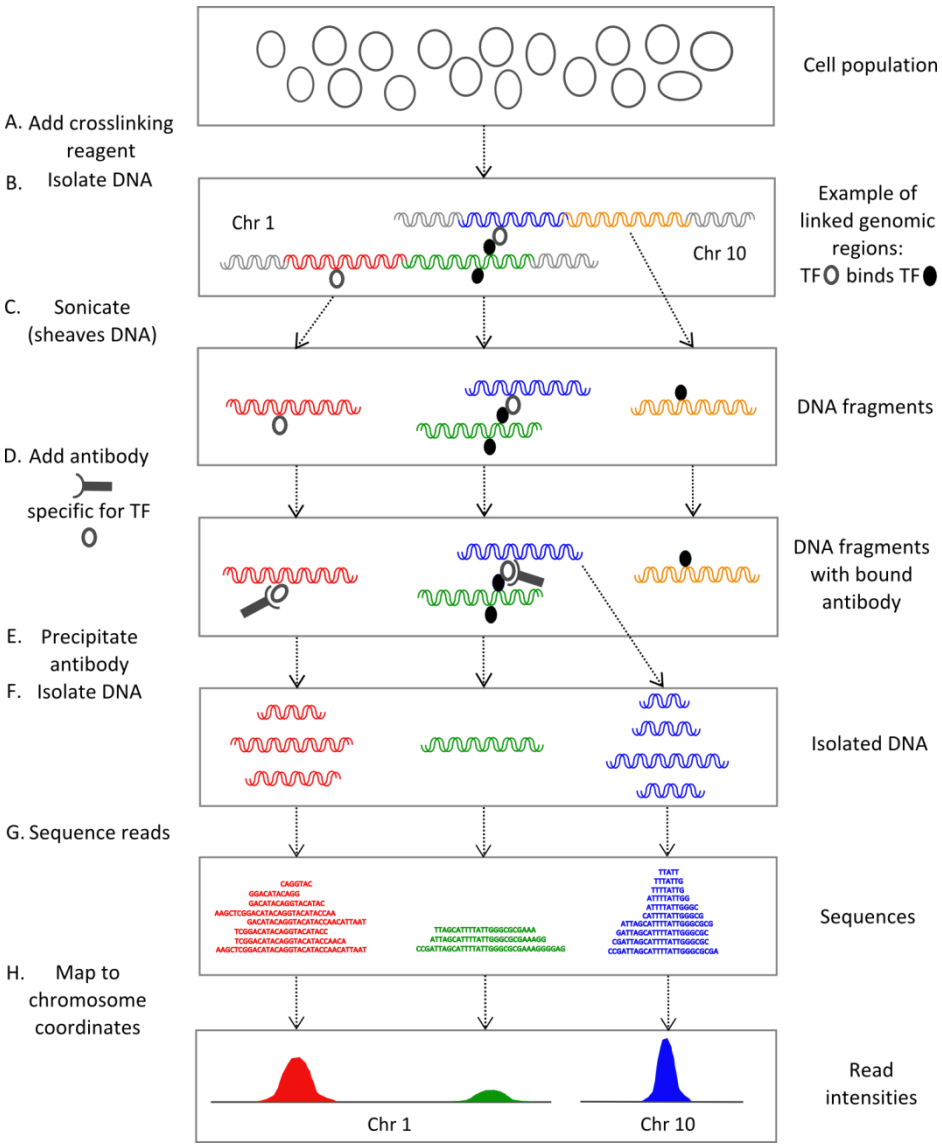


Figure 1.11. ChIP-seq protocol.

As shown in the diagram, since formaldehyde traps protein – protein interactions as well as protein – DNA interactions, then some sequenced DNA fragments may be related to the protein factor of interest through an intermediate protein. Thus, both *cis*-regulatory and *trans*-regulatory interactions may be detected. Despite this, the results of a ChIP-Seq TF study are generally referred to as *binding sites*. Note that (i) the choice of chromosomes 1 and 10 is arbitrary, (ii) ChIP-seq peaks shown are the result of merging regions of read density from each strand in turn, offset by a gap roughly equal to the sequenced fragment length.

1.10 Objectives of the research

The broad objective at the beginning of the thesis was to apply ChIP-seq derived maps of TF binding sites to areas of research involving the transcription of microRNA genes. Up until 2009, the majority of genome-wide research into the regulation of microRNA gene expression relied upon computationally predicted TF binding sites. For example, I had previously examined a number of TF – microRNA integrated regulatory network papers in human resting solely upon computational predictions of TF binding patterns, e.g. (Shalgi *et al.* 2007; Yu *et al.* 2008; Tu, Yu *et al.* 2009). At around the same time, a number of studies were published capitalizing upon new genome-wide maps of modifications to chromatin, and RNA polymerase II binding sites, to annotate transcription-initiation sites of microRNA genes, e.g. (Marson *et al.* 2008; Ozsolak *et al.* 2008; Barski, Jothi *et al.* 2009; Corcoran *et al.* 2009; Xu *et al.* 2011). This led to our plan of building an updated integrated regulatory network of microRNAs and TFs, with ChIP-seq datasets replacing the TF binding motifs from the previous network research. This plan benefitted from a more than 10-fold growth in the number of publically available ChIP-seq datasets in human during the course of the research (ENCODE 2011).

Since the network project was relatively complex, we began with some simpler investigations in order to start to gain a feel for the datasets and some understanding of the statistical methods that might be used. We found the result that human microRNA host gene promoter regions contained a higher density of transcriptional regulators, compared to other genes. We then noted that this regulation might be related to other biases within the microRNA host gene class, for example to their greater lengths or numbers of splice forms, since these properties are themselves related to the density of transcriptional regulators in gene promoter regions e.g. (Golan, Levy *et al.* 2010; Warnefors and Eyre-Walker 2011). This set in motion the research activity for Chapter 2: 'MicroRNA host genes are highly regulated and biased towards developmental functions'. The key objective here was to test as rigorously as possible links between transcriptional regulation and other properties of microRNA host genes. The key figure summarizing our methodology is Figure 2.2, showing that the above-average level of transcriptional regulation of microRNA host genes persists after controlling for a number of other gene properties.

An important matter linked to the material within Chapter 2 is whether, in the general case, transcriptional regulators bound to the host gene promoter region drive expression of the intragenic microRNA within the gene region. This question continues to exercise the research community, with the amount of alternative transcription of primary microRNA transcripts, not

known e.g. Figure 2.1A and 2.1C in this thesis, and (Baskerville and Bartel 2005; Marson *et al.* 2008; Ozsolak *et al.* 2008; He, Li *et al.* 2012). We had hoped to settle the matter more clearly using the ChIP-seq datasets. In the course of the research, we realised that such a line of inference, from distributions of transcriptional regulators to patterns of splicing of microRNA host gene regions, may be extremely hard to argue. Perhaps it could be done with matched expression datasets within enough cell lines and of sufficiently high quality, but for microRNAs such datasets are difficult to obtain. We therefore sought alternative ways to use the regulatory data. This led us to ask whether transcriptional regulators within protein-coding gene promoter regions might also drive the expression of primary transcripts for microRNAs lying *outside* of protein-coding gene regions. In the general case this has not been considered before, and only rare examples of bidirectional promoter regions driving expression of a protein-coding gene and an intergenic microRNA cluster have been noted (Kim, Saetrom *et al.* 2008; Toyota, Suzuki *et al.* 2008; Barski *et al.* 2009). We therefore carried out a step-by-step exploration of spatial and regulatory relationships between intergenic microRNAs and their protein-coding neighbours. This led to the research in Chapter 3: 'Coupled regulation of intergenic microRNAs and their protein-coding neighbours'.

Taken together, the material within Chapters 2 and 3 provide two independent but closely related investigations into the transcriptional regulation of microRNA genes in humans. The material within Chapter 4 builds upon these chapters through integrating post-transcriptional and transcriptional regulatory layers, to form an integrated regulatory network of TFs and microRNAs over common target protein-coding gene expression pathways. A purely transcriptional network inferred from a similar collection of ChIP-seq datasets in human was published late in 2012, together with much more preliminary analyses of various noncoding RNA extensions to this (Gerstein *et al.* 2012). Our work provides a significantly more detailed analysis of relationships between the TF and microRNA mediated regulatory layers.

The material within Chapter 5 was not planned at the start of the research programme. Rather, this collaborative study was discussed only towards the end of the second year of work. The objective of the study was to provide experimental collaborators with novel testable hypotheses, relating to the natural and pathogenic responses of simian immune systems to SIV infection. The collection of ChIP-seq datasets which I had been working with was mapped to upstream regulatory regions of genes within co-expression clusters in monkeys, identified by Dr. Jamie MacPherson. This chapter is therefore relatively independent, in terms of scientific topic, from the other three research chapters, since no connections to microRNA biology or function were made. The common area lies mainly in methods to identify

significance of enrichment of transcriptional regulators upstream of particular classes of protein-coding genes.

A major priority from the start of this project was the aim of providing reasonably rigorous controls for genomic results. The issue continues to be one of my main preoccupations and some important progress might have been made in this respect. For example, through methods to control the density of transcriptional regulators for the lengths of genes (and many other properties) in chapter 2 (Chapter 2); or to control the density of microRNA genes for the numbers of intergenic nucleotides within fixed distances from protein-coding gene boundaries, in various genomes (Chapter 3); or through considering as often as possible metrics relating to characteristics of specific genes, in place of genome-wide averages or correlations, with the aim of less oversimplification of the data. In the context of Chapter 4, this style of analysis provides much of the novelty in the material, compared to earlier studies (Shalgi *et al.* 2007; Cheng *et al.* 2011), or research sometimes conflicting with our own (Lu and Clark 2012). For example, we attempt to shift the focus from the single summative significance scores typically assigned to significance of a pattern within a regulatory network as a whole, to these scores adapted to each node of the network. In principle this may allow discovery of biological properties of regulators and genes most commonly associated with particular network patterns. We provide some examples of this kind of discovery process in the final results sections of Chapter 4.

Chapter 2

MicroRNA host genes are highly regulated and biased towards developmental functions

Abstract

Approximately half of all human microRNA genes reside within protein-coding host genes. Here, we use 386 genome-wide human transcriptional regulatory factor (TRF) and histone modification ChIP-seq datasets to investigate the *cis*-regulatory regions of microRNA host genes. We find that microRNA host gene *cis*-regulatory regions have an active chromatin state, and are bound by significantly more transcriptional regulators than expected for genes of similar length, number of splice forms, age and function. We show that very little of this regulation is due to the greater lengths of host genes. However, we can link some of the regulation to other characteristics of microRNA host genes, such as the greater numbers of isoforms and greater ages of these genes. Taken together, these properties of microRNA host genes suggest that the *de novo* origin of microRNA genes is favoured within actively transcribed regions. Our work provides novel evidence for the involvement of the host gene *cis*-regulatory region in controlling the regulation and the expression of many intragenic microRNAs. Finally, we identify for the first time significant functional enrichments within the microRNA host gene class, including genes involved in the development of nerves, muscle tissue, and blood vessels. This generalizes from known cases of intragenic microRNAs located within developmentally significant gene regions.

Contributions

The research within this chapter was supervised by Dr. Sam Griffiths-Jones.

2.1 Introduction

MicroRNAs are 20 - 24 nt noncoding RNAs, which typically function as post-transcriptional repressors of protein-coding gene expression pathways. The mature microRNA is derived from microRNA gene regions via a multistep biogenesis pathway (Bartel 2004). The key steps in microRNA biogenesis are expression from a microRNA gene region of a long primary transcript (pri-miRNA), from which one or more 65 – 70 nt RNA stem-loop structures (pre-miRNAs) are rapidly excised (Gregory *et al.* 2004; Marco, Ninova *et al.* 2013). After transport to the cytoplasm, the ≈ 15 – 20 nt loop part of a pre-miRNA is cleaved by the endoribonuclease enzyme Dicer (Bernstein, Caudy *et al.* 2001), and the remaining 20 – 24 nt pre-miRNA stems equate with one or two functional microRNAs (Czech *et al.* 2009; Okamura *et al.* 2009). MicroRNAs repress the production of proteins from mRNAs by binding via their 5'-ends to complementary sequences in target mRNAs, to disrupt translation, signal the target for destruction, or sequester the mRNA from the ribosomal translation pool (Bartel 2004; Bhattacharyya *et al.* 2006). Dysregulated expression patterns of microRNAs are associated with many genetic diseases, including lethal developmental abnormalities, together with numerous cancer phenotypes (Carrington and Ambros 2003; Calin and Croce 2006; Woods, Thomson *et al.* 2007; Hagen and Lai 2008; Pencheva and Tavazoie 2013). There is therefore significant interest in studying the mechanisms of transcriptional regulation that critically determine the expression patterns of microRNA genes, e.g. (Shalgi *et al.* 2007; Marson *et al.* 2008; Ozsolak *et al.* 2008; Yu *et al.* 2008; Barski *et al.* 2009; Tu *et al.* 2009; Cheng *et al.* 2011).

Depending upon the species, ≈35 – ≈65 % of vertebrate microRNA genes are located within protein-coding genes, termed microRNA host genes (Rodriguez *et al.* 2004; Saini *et al.* 2007; Berezikov 2011; Meunier, Lemoine *et al.* 2013). In all species examined, this is a much higher proportion than expected by chance, given the total genomic space available within protein-coding genes (Berezikov 2011; Meunier *et al.* 2013). The vast majority of intragenic microRNAs reside within intronic regions, consistent with protein-coding functional constraints excluding them from the host gene exons (Rodriguez *et al.* 2004; Berezikov 2011). Evolutionary associations between microRNAs and their host genes are stable, and at least for the most conserved microRNAs, their expression levels show a tendency to be correlated with those of their host genes (Baskerville and Bartel 2005; Hoeppner, White *et al.* 2009; Biasiolo, Sales *et al.* 2011). Since ≈ 80% of intragenic microRNAs also have the same transcriptional orientation to the host gene, intragenic pre-miRNAs are generally considered to be expressed within a host gene precursor mRNA (pre-mRNA), from a shared promoter region (Rodriguez *et al.* 2004; Baskerville and Bartel 2005; Meunier *et al.* 2013). Even in cases where the microRNA and host

gene are transcribed from distinct promoter regions (Marson *et al.* 2008; Ozsolak *et al.* 2008), the microRNA and host gene primary transcripts must still overlap, so that crosstalk between the biogenesis pathways of the protein-coding and microRNA genes seems inevitable (Shomron and Levy 2009).

Through computational analysis, TFs and microRNAs are predicted to target one another at much higher rates than expected, given a random shuffling of genes (Cui, Yu *et al.* 2006; Shalgi *et al.* 2007 ; Yu *et al.* 2008). Since these studies, significant numbers of datasets have been generated providing genome-wide binding locations of transcriptional regulators, including TFs, transcriptional cofactors, and chromatin modifying enzymes, e.g. (2011; Gerstein *et al.* 2012). It is therefore timely to build upon earlier computational analysis of TF-microRNA networks using experimental data. An important observation which has received little attention within this context is that the microRNA host gene class is biased in favour of certain gene characteristics. For example, host genes have been shown to be much longer than average, and to have longer introns (Golan *et al.* 2010). This would be expected under a neutrally evolving model of *de novo* microRNA birth in transcribed sequences, since longer host genes contain more intronic space. It is clearly possible that the greater numbers of computationally predicted TF motifs within microRNA host gene *cis*-regulatory regions are related to properties of the host gene itself. Indeed, many properties of protein-coding genes are connected with the densities of transcriptional regulators bound to their promoter regions, including gene function, expression, and age (Cui *et al.* 2006; Barski *et al.* 2009; Warnefors and Eyre-Walker 2011; Yan, Enge *et al.* 2013).

We undertook in this study to examine the transcriptional regulation of intragenic microRNAs and their host gene transcripts in human, using experimentally determined genome-wide binding site datasets for 117 transcriptional regulators and maps of 38 histone modifications. For brevity, we term TFs together with other families of transcriptional regulators sampled by ChIP-seq as transcriptional regulatory factors (TRFs). We provide validation from experimental data that microRNA host gene *cis*-regulatory regions are bound by significantly elevated numbers of TRFs. We characterise the microRNA host gene class, uncovering several new properties, including a preferentially open chromatin state for these genes, a shift towards greater evolutionary ages, and significant enrichments in functional categories of genes, particularly relating to development. We refute the hypothesis that gene length alone is sufficient to account for the greater numbers of TRFs bound to host gene cis-regulatory regions. Nevertheless, we show that other properties of the microRNA host gene class, such as the greater numbers of splice variants of these genes, can greatly affect whether the numbers of bound TRFs within the host gene *cis*-regulatory region is considered significant. We also

show that the number of transcriptional regulators within microRNA host gene *cis*-regulatory regions is linked, albeit weakly, with microRNA expression levels. Our study therefore provides support for a model of microRNA gene birth that is favoured within actively transcribed host gene regions.

2.2 Materials and methods

Gene sets

The coordinates of human protein-coding transcript datasets were obtained from BioMart (http://www.biomart.org/; Ensembl transcript collection, v.65) (Kasprzyk 2011). Human microRNA gene locations were downloaded from miRBase (http://www.mirbase.org/; v.18) (Griffiths-Jones *et al.* 2006; Kozomara and Griffiths-Jones 2011). MicroRNA genes with coordinates overlapping a protein-coding gene on the same strand were annotated as intronic. The protein-coding gene *cis*-regulatory region was defined as the proximal region 2000 nt either side of the 5'-most TSS of the gene from all transcriptional isoforms available in the Ensembl v.65 human protein-coding gene collection. Varying the definition of proximal promoter by up to 2000 nt did not significantly affect the results.

Genome-wide mapping of DNA binding proteins

Genome-wide ChIP-seq peak locations of DNA-binding proteins in human cell lines were obtained from collections submitted by the Yale and HAIB Consortia to the UCSC genome browser (http://genome.ucsc.edu/), listed under the table browser's 'regulation' tab (restricted datasets as of April 2013 not included) (Karolchik, Hinrichs *et al.* 2004; 2011). Details of these TRFs are provided in Supplementary Table S1.1. In total, 590 datasets were obtained, including 242 in pairs of replicates from the HAIB consortium, and 106 from the YALE consortium. Replicate pairs were combined by taking the intersection of ChIP-seq peak regions between the two datasets (with the condition for overlap of 1 nt between replicate peaks). This gave 357 separate ChIP-seq datasets relating to 117 TRFs surveyed across a variety of cell lines and experimental conditions. Literature review was used to separate the TRFs into classes, including 95 TFs associated with RNA pol II promoters, 3 RNA pol II components, 4 nucleosome remodelling agents, 7 chromatin modifying enzymes, 1 insulator, 3 DNA repair enzymes, and 5 RNA polymerase III components and partners. We also obtained genome-wide read densities obtained from ChIP-seq performed with antibodies specific to a range of histone modifications in human CD4+ T cells (Barski *et al.* 2007; Wang *et al.* 2008). These comprised 18 covalent acetylation marks and 20 methylation marks on histones H2, H3 and H4.

Protein and microRNA expression atlases

We downloaded the Novartis human protein-coding gene expression atlas across 79 human tissue samples (Su *et al.* 2004). Probe identifiers were mapped (via well-defined gene symbols

where necessary) to 15087 unique Ensembl gene identifiers, using probe annotation files from BioGPS (Wu, Orozco *et al.* 2009) and from http://www.affymetrix.com/, and gene symbols from the HUGO gene names consortium (http://www.genenames.org/). Average gene expression was measured as the mean of the normalised expression values of all probe sets mapped to the gene, across the 79 tissue samples. We downloaded a microRNA gene expression atlas across 172 human tissue samples (Landgraf, Rusu *et al.* 2007). Mature microRNA sequences from the atlas were mapped to all possible precursor sequences from miRBase (v.18). Matched healthy tissue samples between the protein-coding and microRNA expression atlases are provided as Supplementary Table S2.1.

Ages of protein-coding genes

A catalogue of evolutionary ages of 20259 protein-coding genes was obtained from (Domazet-Loso and Tautz 2010) using times from (Hedges, Dudley *et al.* 2006) as in (Warnefors and Eyre-Walker 2011). Gene symbols were converted to 16935 IDs present in the Ensembl v.65 human protein-coding gene collection, using mappings provided by BioMart and downloaded from the HGNC website (http://www.genenames.org/).

Host gene properties

For each numerical property of a gene (length, number of splice forms, age, and average expression), a measure of the difference in the genome-wide distributions of the property between microRNA host genes and non-host genes was calculated, using the two-sample Kolmogorov-Smirnov test (K-S test). Statistical significance of the difference in number of bound TRFs, or of ChIP-seq reads for each examined chromatin modification, between host and non-host gene sets was likewise assessed using the two-sample K-S test.

MicroRNA host and non-host gene collections were divided into classes according to (i) the number of TRFs bound to their *cis*-regulatory regions (Figure 2.1B), and to each of (ii) gene length, number of splice forms, and gene age (Figure 2.2). For gene length and number of bound TRFs, class boundaries were chosen to give almost equal class sizes. For gene age and number of splice forms, classes were defined by the discrete values of these properties, giving smaller class sizes for younger genes and those with more splice variants. To prevent class sizes becoming too small, genes with age <= 910 mya and genes with ≥ 8 splice forms were collected into single classes. Varying class sizes within sensible limits has minimal impact upon findings. Contributions from (i) the number of bound TRFs, to microRNA host gene expression

level, and from (ii) each of microRNA host gene length, number of splice forms, and age, to number of bound TRFs, were then estimated by simulation.

Specifically, within each regulatory, length, splice forms, and age class, the labels of microRNA host and non-host genes were shuffled 10000 times. For example, to calculate the data in Figure 2.1B, the expected expression of microRNA host genes within a given regulation class was estimated as the mean expression level across 10000 shuffled samples of host genes within this class. Variation was assessed using the quartiles of the distribution of simulated expression values within each regulation class. The statistical significance (p-value) of microRNA host gene expression levels within each regulation class was defined as the fraction of simulated host gene sets within this regulation class having mean expression level greater than for the real microRNA host gene set. Global significance of microRNA host gene expression given regulation (across all host genes) is likewise assessed by counting the differences between observed and simulated microRNA host gene expression levels across all regulation classes. A global p-value reflecting the significance of microRNA expression levels given numbers of bound TRFs was calculated in the same way as for individual regulation classes. To calculate Figure 2.2, the same kinds of procedures were used, with random gene samples drawn within gene length, splice form, and age classes, and with mean numbers of TRFs (instead of mean gene expression level) as the dependent variable for each of the tests. Increasing the number of random samples or changing class boundaries leads to no meaningful changes to global p-values.

Functional annotation

Human genes were mapped to terms in the Gene Ontology (GO) hierarchy using the associations provided by the GO consortium [http://www.geneontology.org/GO.downloads.annotations.shtm retrieved on 14/06/2013] (Ashburner, Ball *et al.* 2000). The full GO hierarchy of all links between GO terms (OBO v.1.2) was parsed to collect gene sets with a common ancestral GO term. Mean numbers of TRFs were compared between microRNA host and non-host genes mapping to each GO term. Statistical significant of excess TRFs per host gene was assessed by generating 10000 random samples or genes of equal size to the number of host genes mapping to the GO term. P-values were defined as the fraction of random host gene samples with at least as many TRFs per gene as in the real microRNA host genes matching the corresponding GO term. Due to the large number of GO terms, p-values were adjusted using a multiple testing correction (Benjamini, Drai *et al.* 2001).

2.3 Results

2.3.1 MicroRNA host genes are bound by many transcriptional regulators

We cross-referenced the coordinates of human microRNA genes from miRBase (v.18) and protein-coding transcripts from Ensembl (v.65), identifying 659 genes with one or more intragenic microRNAs. A total of 357 ChIP-seq datasets, corresponding to 117 transcription regulatory factors (TRFs), were mapped to the genome. Of these, around 80% are recognised as sequence-specific DNA-binding TFs, with the remainder reflecting general transcription factors, chromatin modifiers, regulatory co-factors, or DNA repair enzymes. The number of TRFs bound to *cis*-regulatory regions of each protein-coding gene was calculated. C*is*-regulatory regions were defined as 2 kb intervals either side of the 5' end of the gene. The mean number of regulators per gene was then compared between microRNA host and non-host genes. The 659 microRNA host genes are bound on average by 33.2 TRFs compared with an average of 28.1 TRFs for 19316 non-host genes. The data therefore show an 18% increase in the number of TRFs bound to *cis*-regulatory regions of microRNA host genes, and this difference is highly significant ($p = 1.5 \times 10^{-11}$, K-S test, methods). This is consistent with an earlier report in which greater than expected numbers of TF binding motifs were found upstream of intronic and intergenic microRNAs combined (Yu *et al.* 2008). Very similar percentage increases are observed for various subgroups of regulatory factors, including for 95 sequence-specific TFs (17.7%) and for 12 chromatin modifiers (19.1%). The difference was also significant for datasets restricted to specific cell lines (HEPG2: $p = 6.6 \times 10^{-6}$; K562: $p = 1.2 \times 10^{-5}$; HELA: $p = 6.7 \times 10^{-12}$; GM12878: 1.1×10^{-8}). Finally, 110 / 357 ChIP-seq datasets were *individually* significant over the microRNA host gene class ($p < 0.05$ by binomial test, corrected for multiple tests). By contrast, although some ChIP datasets have below-average numbers of binding sites within the *cis*-regulatory regions of microRNA hosts, none of these were significantly depleted. We conclude that intragenic microRNAs, as a class, reside within genes enriched for upstream transcriptional regulators.

2.3.2 Chromatin modifications to microRNA host genes favour activation of transcription

Since ChIP-seq experiments will tend to sample from accessible chromatin regions (Zhang *et al.* 2008; Rozowsky *et al.* 2009), we next asked whether chromatin modifications to microRNA host gene regions reflect an accessible chromatin state. In a number of previous studies, maps of histone protein post-translational modification marks were used to search upstream of

microRNA sequences for elements associated with an RNA polymerase II promoter (Marson *et al.* 2008; Ozsolak *et al.* 2008; Barski *et al.* 2009). In around 2 / 3 of cases, the nearest such promoter signal to the intragenic microRNA was found to coincide with the 5' end of the host gene. Here, we instead ask which histone modifications are enriched within the *cis*-regulatory regions of microRNA host genes, regardless of whether an additional microRNA-specific promoter region may be found within the gene. We first mapped the locations of 18 types of histone acetylation and 20 types of histone methylation to *cis*-regulatory regions of protein-coding genes (Barski *et al.* 2007; Wang *et al.* 2008). We then compared the numbers of ChIP-seq reads for each mark between microRNA host and non-host genes (Tables 2.1 and 2.2).

Table 2.1. Acetylation marks within microRNA host gene promoter regions

		p-value	
Rank	**Acetylation**	**Uncorrected**	**B.-H. corrected**
1	H2BK120ac	1.35E-05	2.70E-04
2	H3K9ac	4.27E-05	4.27E-04
3	H3K18ac	0.000	9.12E-04
4	H3K4ac	0.000	7.81E-04
5	H2AK9ac	2.996E-04	1.20E-03
6	H4K91ac	0.000	1.29E-03
7	H2BK20ac	0.001	1.60E-03
8	H3K27ac	0.001	1.73E-03
9	H2BK5ac	0.001	2.95E-03
10	H3K23ac	0.003	5.88E-03
11	H4K12ac	0.004	6.60E-03
12	H3K36ac	0.005	8.40E-03
13	H4K16ac	0.010	1.51E-02
14	H4K8ac	0.011	1.52E-02
15	H2BK12ac	0.022	2.90E-02
16	H4K5ac	0.025	3.12E-02
17	H3K14ac	0.268	3.15E-01
18	H2AK5ac	0.301	3.34E-01

Significance of enrichment was calculated by comparing the distribution of histone mark read counts between microRNA host and non-host genes, using the Kolmogorov-Smirnov test, with p-values calculated in R. P-values were corrected for multiple testing using the Benjamini-Hochberg correction (Benjamini *et al.* 2001). Acetylation marks shaded green are significant with $p \leq 0.05$.

Table 2.2. Methylation marks within microRNA host gene promoter regions

		p-value	
Rank	**Methylation**	**Uncorrected**	**B.-H. corrected**
1	H3K4me3	1.09E-12	2.18E-11
2	H4K20me1	7.52E-11	7.52E-10
3	H3K9me1	7.14E-08	4.76E-07
4	H3K4me1	8.50E-08	4.25E-07
5	H3K4me2	1.21E-07	4.86E-07
6	H3K79me3	4.41E-05	1.47E-04
7	H3K79me2	7.81E-05	2.23E-04
8	H3K79me1	1.597E-04	3.99E-04
9	H2BK5me1	3.212E-04	7.14E-04
10	H3K36me1	0.005	9.86E-03
11	H3K27me1	0.011	2.07E-02
12	H3K27me3	0.037	6.23E-02
13	H4K20me3	0.084	1.29E-01
14	H3R2me2	0.094	1.34E-01
15	H3K36me3	0.136	1.81E-01
16	H3K9me3	0.147	1.84E-01
17	H3R2me1	0.171	2.01E-01
18	H4R3me2	0.318	3.53E-01
19	H3K27me2	0.374	3.94E-01
20	H3K9me2	0.758	7.58E-01

P-values were calculated in the same manner as for acetylation marks (Table 2.1).

We find that 16 / 18 of the acetylation marks, but only 11 / 20 of the methylation marks, are significantly enriched in *cis*-regulatory regions of microRNA host genes (Tables 2.1 and 2.2). The high level of acetylation suggests an open, active chromatin state (Wang *et al.* 2008). Examining the set of significantly enriched methylation marks confirms this, since modifications which enhance transcription are present at high levels. These include H3K4-me1,-me2,-me3 ($p < 10^{-6}$ by K-S test, corrected for multiple tests; see also methods), H3K9-me1 ($p < 10^{-6}$) and H3K79-me1,-me2,-me3 ($p < 10^{-3}$) (Barski *et al.* 2007). By contrast, three prominent repressive methylation marks are present at marginally lower levels in *cis*-regulatory regions of microRNA host genes (H3K9-me2, H3K27-me2, H3K36-me3) (Lachner and Jenuwein 2002; Barski *et al.* 2007). We conclude that modifications to histone proteins at the 5' end of microRNA host genes favour activation of transcription of these genes.

2.3.3 Expression patterns of intragenic microRNAs and their host genes

It has previously been reported that microRNA and host gene mRNA expression patterns are correlated across tissues, both in human and in other species (Baskerville and Bartel 2005; Meunier *et al.* 2013), although the strength of this finding has also been questioned (He *et al.* 2012). We therefore tested whether the result holds true for the data used in this study. Expression values of mRNAs and microRNAs were obtained from publically available expression atlases, across 79 and 172 human tissue samples respectively (Su *et al.* 2004; Landgraf *et al.* 2007). From the microRNA expression atlas, 237 mature sequences could be matched to intragenic microRNAs, within 150 different host genes from the protein-coding expression atlas. From the protein-coding expression atlas, 18 tissues were found in common with the microRNA expression atlas (Supplementary Table S2.1). Spearman's correlation coefficients (r_s) were then calculated between the expression levels of pairs of intragenic microRNA and host gene mRNA within each tissue. The distribution of r_s values is significantly shifted towards positive values (Figure 2.1A), indicating that intragenic microRNAs generally are expressed in similar tissues to their host genes (t-statistic = 6.02, $p < 10^{-5}$). In total, positively correlated (microRNA / host gene) pairs account for 70.5% of total pairs examined. After correcting for multiple tests, however, only 4 (microRNA mature sequence / host gene) pairs are individually significantly correlated at the 5% level ($r_s \geq 0.611$) (Table 2.3).

Table 2.3. Host gene – microRNA pairs with significant co-expression correlation across human tissues

Gene	Mature microRNA	r_s	p-value	α (B-H corrected)
EGFL7	hsa-mir-126-3p	0.827096	0.000001	0.000446
C1orf61	hsa-mir-9-5p	0.733085	0.000090	0.000893
EVL	hsa-mir-342-3p	0.645976	0.001045	0.001339
EGFL7	hsa-mir-126-5p	0.611387	0.002201	0.001786

Spearman's correlation coefficients between microRNA and host gene mean expression levels are calculated as described in the main text. The significance level $\alpha = 0.05$ was adjusted for multiple tests using the Benjamini-Hochberg correction (Benjamini *et al.* 2001).

We conclude that there is significant class-wide association between the measured expression levels of microRNA and host mRNA transcripts, but that the strength of this association for particular microRNA/mRNA pairs is often quite weak. This could be due to differences in the transcription, degradation, or measurement, of the microRNA and mRNA pair.

We next asked whether microRNA host genes have higher or lower expression than expected given the number of bound TRFs. To test this, protein-coding genes were divided into a

number of equal-sized groups ranked by the number of bound TRFs. We term these groups *regulation classes*. The mean expression of microRNA host and non-host genes was then calculated within each regulation class (Figure 2.1B). Expected mRNA expression level within each regulation class was then estimated by simulation (see methods). As suggested by Figure 2.1B, there is no significant difference between the observed and expected expression level of microRNA host genes in any regulation class. Thus, microRNA host genes are expressed at levels commensurate with their level of regulation. Assuming microRNAs are typically excised from a host gene transcript, this is consistent with the finding that this excision does not lead to a significant reduction in the number of mature host mRNAs within a cell (Kim and Kim 2007).

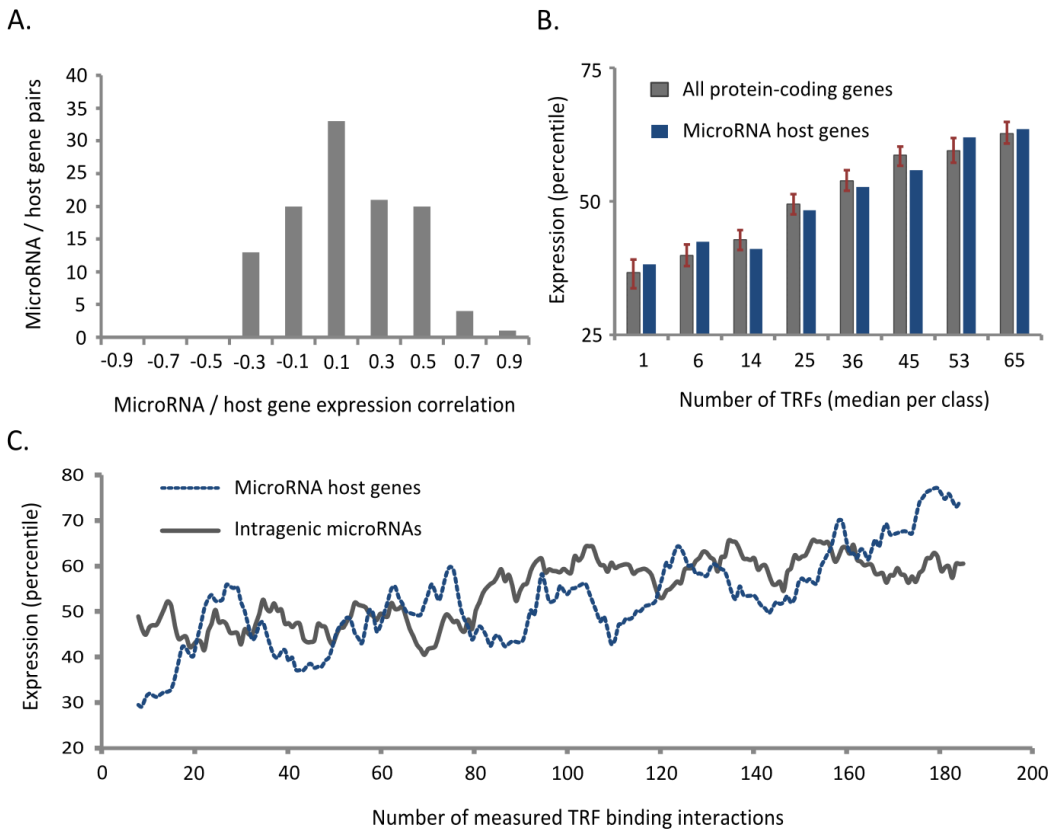


Figure 2.1. Expression and regulation of protein-coding genes and intragenic microRNAs.

A. Distribution of Spearman's correlation coefficients between expression of intragenic microRNA mature sequences and microRNA host gene mRNAs, across human tissues.

B. Comparison of microRNA host and general protein-coding gene expression within 8 regulation classes. Expression was calculated as the mean value across 72 human tissue samples, and converted to percentiles of the mean expression distribution. Errors bars reflect the interquartile ranges of simulated data (methods).

C. Variation in expression of intragenic microRNAs and host gene mRNAs, with numbers of TRF binding interactions detected across all ChIP-seq experiments. The lines display a moving average calculated across groups of 25 genes at a time.

It is also clear from Figure 2.1B and consistent with literature that the expression level of mRNAs increases significantly with increasing numbers of TRFs bound to the gene's *cis*-regulatory region (Yan *et al.* 2013). We tested this more precisely using the Spearman's correlation coefficient (r_s). We find a significant positive correlation genome-wide between the numbers of TRFs bound to the *cis*-regulatory region of a protein-coding gene, and the mean expression level of the gene across tissues ($r_S = 0.278$, $p < 10^{-15}$). This result remains true when restricted to the 659 microRNA host genes ($r_S = 0.257$, $p < 10^{-15}$). Importantly, there is also a significant though weaker positive correlation between numbers of TRFs bound to microRNA host gene *cis*-regulatory regions, and the mean expression rank of the intragenic microRNA ($r_S = 0.204$, $p = 0.0011$). Both mRNA and microRNA expression levels also increase with total number of *cis*-regulatory interactions sampled by ChIP-seq i.e. added up across all experiments performed (Figure 2.1C). The lower correlation between numbers of upstream regulators and expression level for microRNAs, compared to mRNAs, could indicate that some microRNAs are transcribed from alternative promoter regions, compared to the dominant isoforms of host gene transcripts (Marson *et al.* 2008; Ozsolak *et al.* 2008). Alternatively, this might reflect the lower resolution of the technology used to measure microRNA expression (Landgraf *et al.* 2007).

2.3.4 Characteristics of microRNA host genes

We next defined two possible accounts of the high levels of regulation within *cis*-regulatory regions of microRNA host genes. According to the first account, the number of transcriptional regulators relates to underlying properties of the host gene, in particular, to the length of the gene. This is because there are more positions within a longer gene at which a microRNA gene sequence can be located (Golan *et al.* 2010). According to the second account, the above-average number of transcriptional regulators relates to properties of the host gene class favoured by microRNAs, over and above length, and more generally to the function of the intragenic microRNA. This can be because microRNAs are more likely born, and then retained, within highly regulated host genes, or because *cis*-regulatory elements are more likely to be acquired, and then retained, upstream of microRNA-bearing gene regions. To test these accounts, we first identify a number of distinctive characteristics of the microRNA host gene class. We then compare the numbers of TRFs observed in their *cis*-regulatory regions, to the numbers which are expected, given these characteristics of microRNA host genes.

2.3.4.1 MicroRNA host genes are longer, more highly spliced, and older than average

We calculated gene length, number of splice variants, and age, for all protein coding genes, and then compared the distributions of these features between microRNA host and non-host genes (Table 2.4, methods).

Table 2.4. Characteristics of microRNA host genes.

	Mean value of gene property			
Gene property	Non-host	MicroRNA host	K-S	P-value
Length (kb)	61.3	199.2	0.349	$<1.0 \times 10^{-15}$
Splice forms	6.47	9.63	0.243	$<1.0 \times 10^{-15}$
Age (mya)	1539.1	1678.1	0.128	8.1×10^{-8}

K-S = Kolmogorov-Smirnov 2-sample statistic for comparison of distribution of each gene property between host and non-host genes, and p-value is the significance of the K-S statistic. Length and splice form calculations are based on the complete set of Ensembl v.65 protein-coding genes, with 19316 non-host and 659 host genes. Age calculations are based on a subset of 16323 non-host and 612 host genes with an annotated age retrieved from (Warnefors and Eyre-Walker 2011).

As reported previously, both the average length (199.2 kb) and number of splice forms (9.63) of microRNA host genes are significantly greater than those of non-host genes (Table 2.4) (Golan *et al.* 2010). This is expected, since, as we have already remarked, a longer host gene region has a higher chance of hosting a microRNA sequence, so that the expected length of host genes is far greater than the average protein-coding gene length. Moreover, longer genes are more highly spliced in general (Golan *et al.* 2010). Using ages of all protein-coding genes as in (Warnefors and Eyre-Walker 2011), we find the novel result that microRNA host genes are also older than non-host genes ($p = 8.1 \times 10^{-8}$, K-S test). This is consistent with the structural characteristics of the microRNA host gene class, since older genes are larger and more highly spliced (between gene age and length, $r_S = 0.254$, and between gene age and number of splice forms, $r_s = 0.223$, both significant with $p < 10^{-15}$). In summary, microRNA host genes are significantly longer, more highly spliced, and older than other protein-coding genes. However, these characteristics might be expected under a model where microRNAs arise at random locations within transcribed sequences.

We therefore next asked to what extent these characteristics of microRNA host genes, and especially the length of the gene, contribute to the enrichment in TRFs found in their promoters. Each of the gene properties – length, number of splice forms and age – is related genome-wide to the number of *cis*-regulatory region binding TRFs (r_s = 0.122, 0.293 and 0.174 respectively, $p < 10^{-15}$). The positive relationship between number of splice forms and total number of TRFs could reflect either higher activity, or more complex regulation, of a more highly spliced gene. The positive relationship between regulation and gene age could reflect that TRFs are gained more than they are lost, so that regulatory connections tend to accumulate through time (Warnefors and Eyre-Walker 2011).

2.3.4.2 Contribution of microRNA host gene characteristics to numbers of bound transcriptional regulators

Protein-coding genes were first ranked in order of length, number of alternative splice forms and age, and divided into classes by the values of each of these properties. These classes were termed *length classes*, *splice form classes*, and *age classes*. The mean numbers of TRFs bound to *cis*-regulatory regions were calculated separately for microRNA host and non-host genes within each length, splice form, and age class (Figure 2.2: A, C and E). The significance of differences in numbers of TRFs between microRNA host and non-host genes within each class was assessed by simulation (methods). Simulations were then combined across classes to give a measure of the expected total number of TRFs bound to microRNA host gene *cis*-regulatory regions, given their greater lengths, numbers of splice forms, and greater ages of the host genes (Figure 2.2: B, D and F).

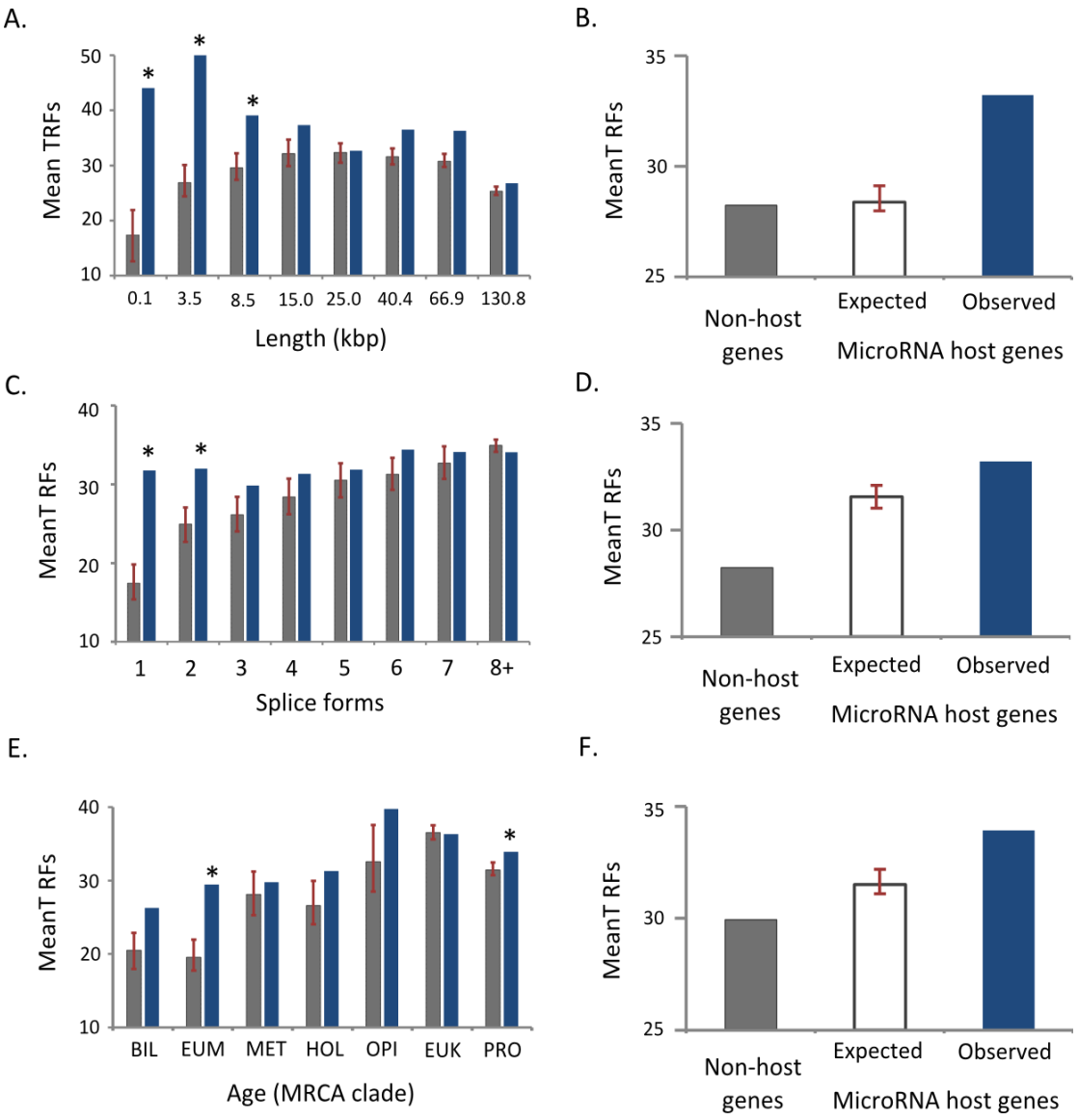


Figure 2.2. Transcriptional regulation of microRNA host and non-host genes.

A, C, E: Genes are divided into classes based on (A) gene length, (C) number of splice forms and (E) gene age (mya), and mean numbers of TRFs for host (blue) and non-host (grey) genes displayed by columns. Gene length was treated as a continuous variable, and cut-offs determined to divide genes into 8 nearly equal-sized cohorts. The number of gene splice forms was treated as categorical data with variable-sized cohorts of genes. Gene age was treated as categorical data, (BIL = bilateria, EUM = eumetazoa, MET = metazoa, HOL = holozoa, OPI = opisthokonta, EUK = eukaryota, PRO = prokaryota), using the most recent common ancestors (MRCAs) of protein-coding genes determined in (Domazet-Loso and Tautz 2010). Due to small sample sizes, data for age ≤ 910mya (subdivisions of bilateria), and for ≥ 8 alternative splice forms, were collected into single cohorts. Error bars reflect uncertainty of estimations from simulation of microRNA host gene regulation within each class (methods). Classes with a significant enrichment in bound TRFs over the host gene class ($p < 0.05$) are indicated by an asterisk.

B, D, F: The observed number of bound TRFs per microRNA host gene compared with the numbers expected, from simulations, given (B) length, (D) splice forms and (F) age.

Host gene Length: We find that *cis*-regulatory regions of microRNA host genes are bound by more TRFs than non-host genes, in all length classes (Figure 2.2A), with the greatest differences for short genes with median length 0.13 – 8.45 kb (number of genes = 79, $p \leq 0.01$, MC simulation). One interpretation for this finding is that shorter host genes include less space for an alternative promoter specific to the microRNA. This would imply that TRFs regulating transcription of the microRNA are more likely to be found in the 5'-most promoter region of the host gene. Overall, the lengths of host genes lead to almost no increase in the expected numbers of TRFs in their *cis*-regulatory regions (from 28.2 to 28.4 TRFs per gene: Figure 2.2B). This is significantly less than the observed rate of 33.2 TRFs per microRNA host gene ($p < 10^{-4}$, MC simulation). We conclude that the model of *de novo* birth of microRNA sequences more often in longer genes cannot account for the high rates of binding of TRFs to host gene *cis*-regulatory regions. Therefore, if microRNAs tend to reside in highly regulated host genes, this must be due to factors over and above the length of the gene.

Number of host gene splice forms: MicroRNA host genes are bound by more TRFs than non-hosts for genes with 1 to 10 annotated splice variants (Figure 2.2C, with results for ≥ 8 splice variants combined). The 38 microRNA hosts with 1 splice form have almost double the background rate of TRFs bound to their promoters (31.8 TRFs per microRNA host compared with 17.4 TRFs per non-host gene, $p < 0.0001$, MC simulation). TRF enrichment within splice form classes remains significant for a further 45 genes with 2 transcript variants ($p = 0.014$). Thereafter, the significance of excess numbers of TRFs bound to microRNA host genes declines, until for the 238 microRNA host genes with ≥ 10 transcript variants, there is no further significant difference in numbers of TRFs compared to non-host genes. (N.B. splice classes for genes with more than 8 transcripts are combined in Figure 2.2C). Combined across all the cohorts of genes, the expected number of TRFs per host gene given their number of splice forms is 31.6 (Figure 2.2D). This is above the background level of 28.2 TRFs for non-host genes, but still significantly less than the 33.4 TRFs observed for microRNA host genes ($p = 0.018$, MC simulation). Thus, the highly spliced character of a subset of microRNA host genes might fully account for the density of bound TRFs for those genes. However, there remains an excess of regulators binding to microRNA hosts, due to exceptional enrichment for the simplest and shortest microRNA host genes (with < 3 splice forms).

Host gene age: Although most protein-coding gene families predate the first microRNA families, protein-coding gene age might nevertheless be related to the subsequent birth of intragenic microRNAs. This is because many of the oldest classes of genes, for example cell cycle regulators, or components of the ribosome, also have the highest transcriptional activity (Ramskold, Wang *et al.* 2009). This pre-existing transcriptional activity may favour *de novo*

birth of functional intragenic microRNAs within the gene region. From the 20574 protein-coding Ensembl genes, 16935 could be linked to gene ages from (Warnefors and Eyre-Walker 2011), with slightly increased mean number of TRFs per gene over this sample. The number of TRFs binding to microRNA host gene *cis*-regulatory regions exceeds non-host TRFs in 6 out of 7 age classes (Figure 2.2E). The expected number of TRFs per host gene, given the combined distribution of gene ages over all 7 age classes, is 31.5. This is significantly below the 33.9 TRFs per microRNA host gene ($p = 0.003$, by simulation. Figure 2.2F). We conclude that the greater ages of microRNA host genes, while perhaps reflecting factors favouring the origin of *de novo* microRNA sequences, do not fully account for the apparent high levels of regulation observed. Taken together, there is a strong case either for microRNAs arising preferentially within more highly regulated genes, or influencing the evolution of their host gene *cis*-regulatory regions so that more regulatory links are acquired.

2.3.5 MicroRNA host genes are enriched for developmentally important genes

Since microRNA host genes are highly regulated, highly spliced, and are relatively old, we might expect this class of genes to sample more often from particular cellular functions. We therefore examined whether microRNA host genes, as a class, have any functional preferences. In previous studies, using less up-to-date database releases, no functional preferences were found (Rodriguez *et al.* 2004; Hoeppner *et al.* 2009). The 659 microRNA host gene Ensembl IDs were submitted to the functional annotation tool at the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Dennis, Sherman *et al.* 2003; Huang da, Sherman *et al.* 2009), with default settings, and using all Ensembl v.65 protein-coding genes as a background. We find that there are indeed enriched terms relating both to structure and function of microRNA host genes (Table 2.5). These terms relate mainly to the regulation of cell shape (e.g. Pleckstrin homology, cell morphogenesis and cytoskeleton organization), and to developmental processes (e.g. neuron development, anterior/posterior pattern formation). In total, more than 40 developmentally significant microRNA host genes are identified, a much larger set than commonly recognised examples such as the HOX gene clusters (Mansfield and McGlinn 2012).

Examples of critical developmental regulators that are also microRNA host genes include NOTCH1, the chromatin modifier EP300, focal adhesion kinase (alias: PTK2), and three members of the forkhead family of transcription factors (FOXP1, FOXP2, and FOXP4) (Goodman and Smolik 2000; Radtke, Wilson *et al.* 2004; Han, Lee *et al.* 2011). To our

knowledge, the possession of an intronic microRNA within each of these three homologous genes has not previously been noted. The DAVID server also provides a functional term clustering tool, allowing detection of enrichment distributed over sets of related terms. This tool detects 14 enriched clusters of functional terms (score ≥ 2), including muscle tissue and blood vessel development, and additional protein domains associated with cytoskeletal proteins (e.g. WD40-repeat, (EGF)-like, FERM domain, and ANKYRIN repeat). In total 202 / 659 microRNA host genes are associated with at least one of these enriched clusters.

Table 2.5 Structural and functional enrichments of the microRNA host gene class.

Category	Term	Number of host genes	P-value
Interpro	Pleckstrin homology-type	31	0.00068
GO_BP	cell morphogenesis	33	0.0074
GO_BP	neuron development	29	0.019
GO_BP	anterior/posterior pattern formation	16	0.041
GO_BP	cytoskeleton organization	32	0.046
SP_PIR	disease mutation	80	0.049
SP_PIR	developmental protein	44	0.072

DAVID draws together structural and functional descriptions of terms from a number of different systems of nomenclature, indicated by the Category column. This includes terms from the Interpro and SwissProt (SP_PIR) databases, and terms from the Gene Ontology consortium, in this case relating to biological processes (GO_BP). Only a subset of the significant terms is included in the table, due to redundancy and / or generality of terms. Multiple testing corrections to modified Fischer's exact test p-values are provided by the DAVID web server (Dennis *et al.* 2003).

As a final test, we compared the regulation of microRNA host genes within each term in the Gene Ontology to all genes matching the same term. To reduce noise, the analysis was restricted to GO terms with a minimum of 50 host genes associated, but the choice of this parameter within sensible limits is not critical to the conclusions. Within the 199 retained GO terms, regulation of microRNA host genes exceeds that of non-hosts in 179 (89.9%) cases, significantly so for 71 GO terms ($p < 0.05$, Benjamini-Hochberg corrected for multiple tests). From these 71 significant terms, for brevity the 15 terms with the largest excess of mean TRFs per host compared to non-host genes are shown in Table 2.6. All significant terms are provided in Supplementary Table S2.2. The largest excess of number of TRFs associated with microRNA host genes was observed for the host GO term 'sequence-specific DNA binding transcription factor activity' (+10.3 TRFs/gene, $p = 6.6 \times 10^{-3}$). We also note the transcriptional regulatory enrichment over microRNA host genes mapping to the term 'regulation of multicellular

organismal development' (+6.7 TRFs/gene, $p = 2.1 \times 10^{-2}$), consistent with the developmental theme identified above. Further, 28 of the 71 (39.4%) significant GO terms include the word 'regulation', in contrast to 4605 of the 39464 (11.7%) terms within the Gene Ontology as a whole. Therefore, intragenic microRNAs are much more frequently associated with host gene TRF enrichment when the host protein has a regulatory role in the cell. Thus, our analysis identifies a further functional theme linking together members of the microRNA host gene class.

Table 2.6. Relationship between microRNA host gene function and number of bound TRFs.

GO term ID	Excess mean TRFs in host promoters (p-value)	GO term description
GO:0003700	10.3 (0.0066)	sequence-specific DNA binding transcription factor activity
GO:0001071	10.3 (0.0060)	nucleic acid binding transcription factor activity
GO:0060089	9.5 (0.0040)	molecular transducer activity
GO:0004871	9.5 (0.0050)	signal transducer activity
GO:0010558	7.5 (0.0286)	**negative regulation of macromolecule biosynthetic process**
GO:0006351	7.1 (0.0100)	transcription, DNA-dependent
GO:0031327	7.1 (0.0301)	**negative regulation of cellular biosynthetic process**
GO:0009890	7.0 (0.0312)	**negative regulation of biosynthetic process**
GO:0006357	6.9 (0.0191)	**regulation of transcription from RNA polymerase II promoter**
GO:0019438	6.7 (0.0104)	aromatic compound biosynthetic process
GO:0006950	6.7 (0.0111)	response to stress
GO:2000026	6.7 (0.0213)	**regulation of multicellular organismal development**
GO:0034654	6.7 (0.0213)	nucleobase-containing compound biosynthetic process
GO:0018130	6.6 (0.0070)	heterocycle biosynthetic process
GO:0051172	6.6 (0.0090)	**negative regulation of nitrogen compound metabolic process**

GO term IDs and term descriptions are taken from the Gene Ontology (methods). Terms containing the word 'regulation' are bold. Mean TRFs bound to host and non-host gene promoters were calculated, for the 199 GO terms with at least 50 microRNA host genes. GO terms were then ranked by the excess of mean TRFs binding microRNA host genes minus mean TRFs binding non-host genes, and statistical significance of the difference for each term determined by MC simulation. P-values were corrected for multiple tests (Benjamini *et al.* 2001). A total of 71 significant terms were identified (Supplementary Table S2.2), with the top 15 of these terms shown in the table.

2.4 Discussion

In this study, we have found a significant enrichment for transcription regulators, both DNA binding and co-regulatory, within the *cis*-regulatory regions of microRNA host genes. The transcriptional regulation of microRNA host genes is therefore higher than expected by chance. Histone modification maps support the picture of a generally open chromatin state, which is favourable to transcription. Our data linking microRNA host gene *cis*-regulators to intragenic microRNA expression level provides novel evidence for the expression of intragenic microRNAs from their host gene *cis*-regulatory regions. This is consistent with a model of microRNA evolution which connects their origin to pre-existing transcribed sequence (Berezikov 2011), and suggests that *de novo* intragenic microRNAs typically arise as a hairpin product excised from a host gene transcript. This model is also supported by the finding from comparative genomics that new microRNAs within clusters preferentially arise within pre-existing microRNA gene regions (Marco *et al.* 2013).

A high level of regulation might be explained by one of a variety of host gene properties. This is important to consider, since there is significant interest in quantifying the levels of regulatory crosstalk between transcriptional regulators and post-transcriptionally regulating microRNAs (Hobert 2006; Shalgi *et al.* 2007; Yu *et al.* 2008; Tu *et al.* 2009; Cheng *et al.* 2011; Gerstein *et al.* 2012). MicroRNA host genes are longer (Golan *et al.* 2010), more highly spliced, and older than non-host genes. Gene age and number of transcriptional isoforms are directly related to the density of transcriptional regulators found within a gene's *cis*-regulatory region (Warnefors and Eyre-Walker 2011). Thus, microRNA host genes may be biased towards a higher level of transcriptional regulation. The key property to consider is gene length, since this is related directly to the probability of a microRNA (or any other nucleotide) sequence being observed within a gene. However, from the variables considered, gene length was least able to account for the number of TRFs bound to microRNA host gene *cis*-regulatory regions ($p < 10^{-4}$). This demonstrates that intragenic microRNAs have a preference for host genes bound by many transcriptional regulators, over and above the consequences of longer intragenic regions having a greater likelihood of hosting microRNA gene sequences.

We next considered whether the *cis*-regulatory enrichment of microRNA host genes might be derivative of microRNAs residing within more highly spliced genes, which represent a more highly regulated class of genes. We do indeed find that for host genes with many splice forms, the number of regulators is commensurate with the number of host gene splice forms. Thus, there might not be any excess of regulators associated with the presence of the microRNA. By contrast, for microRNA host genes with few splice forms, the increased level of regulation

remains highly significant after controlling for splice variant numbers, and is in sharp contrast to other protein-coding genes with few splice forms. In this case, since the probability of a short gene hosting a microRNA is relatively low, we suggest that significant transcriptional activity may generally be required as a pre-condition for microRNA birth. In addition, an intragenic microRNA within a simple host gene may be less liable to be expressed from an alternative downstream promoter region. If so, then the microRNA may be directly linked either through its origin or subsequent evolution with greater fractions of transcriptional regulators within the microRNA host gene cis-regulatory region itself.

The very first microRNAs to be discovered, let-7 and lin-4, have essential functions in the control of developmental timing, including in the regulation of neural tube closure (Lee *et al.* 1993; Johnson, Lin *et al.* 2003). Since then, numerous developmentally important roles have been identified for microRNA genes and gene families e.g.(Lee *et al.* 1993; Johnson *et al.* 2003; Johnston and Hobert 2003; Brabletz, Bajdak *et al.* 2011; Guan, Yang *et al.* 2011). Our examination of microRNA host gene properties uncovered for the first time a significant enrichment in genes relating to neural development, and clusters of terms relating to other developmental pathways, as well as regulation of the cytoskeleton. We also demonstrated a relationship between those microRNA host genes with the most significantly above-average numbers of bound TRFs, and regulatory functions in general. Thus, far from being a random collection of protein-coding genes, the microRNA host gene class possesses a number of enriched functional characteristics. Moreover, evidence for host gene *cis*-regulation linked to the microRNA persists once host gene function is taken into account. Our study therefore serves both to characterise the microRNA host gene class itself, while providing novel evidence for the involvement of the host gene promoter region in the regulation, and expression, of intragenic microRNAs.

2.5 Conclusions

This work has identified many novel characteristics of microRNA host genes. We have shown that protein-coding genes hosting microRNAs within their introns are bound by significantly elevated numbers of transcriptional regulators. Chromatin modifications within the promoter regions of these genes suggest a generally open conformation, favourable to active transcription. We have shown rigorously that the number of bound transcriptional regulators is not a trivial consequence of the greater lengths of the microRNA host gene class. Thus, microRNA genes preferentially arise within highly regulated protein-coding gene regions. A fraction of the excess transcriptional regulators can also be linked to the greater numbers of splice forms of microRNA host genes, though not for the simplest host genes with 1 – 2 splice forms. We identified a significant enrichment in proteins with developmental functions among the microRNA host gene class, including three homologous host genes encoding the transcription factors FOXP1, FOXP2 and FOXP4. Collectively, our results therefore indicate that microRNA gene birth is favoured with highly regulated and developmentally important protein-coding gene regions.

Chapter 3

Coupled regulation of microRNA genes and their protein-coding neighbours

Abstract

MicroRNA genes are found either within protein-coding host genes or within intergenic regions. We show that intergenic microRNAs are found much closer to protein-coding genes than expected by chance, with 30% of human intergenic microRNAs lying within 10 kb of a protein-coding neighbour. Using functional annotations of genomic regions derived from chromatin state, we show that intergenic microRNAs are associated with promoters, enhancers, repetitive sequences, and Polycomb-repressed regions, but are depleted within heterochromatin. In contrast to the prevalence of sense microRNAs inside protein-coding host genes, there is a marked enrichment in antisense intergenic microRNAs at the 5' ends of protein-coding genes, in animal species from human to nematode worm. This corresponds to a divergent protein-coding and microRNA gene pair, which could be expressed from a bidirectional promoter region. A specific case is identified in the human genome, the MIR460b gene lying antisense to the 5' end of the protein-coding gene CUED2, where both expressed products have involvement in a range of human cancers. Using ChIP-seq data corresponding to 106 transcriptional regulatory factors in humans, we show that intergenic microRNAs are more likely to lie downstream of highly regulated protein-coding promoter regions. We then use network pattern frequency analysis to demonstrate that intergenic microRNAs participate in regulatory feedback loops with transcriptional regulators binding to the promoter regions of upstream protein-coding genes. We conclude that a significant fraction of intergenic microRNAs are coupled with their protein-coding gene neighbours.

Contributions

The research within this chapter was supervised by Dr. Sam Griffiths-Jones.

3.1 Introduction

MicroRNAs are short, noncoding RNAs found in all organisms across the plant and animal kingdoms, with critical functions as post-transcriptional regulators of expression of protein-coding genes (Bartel 2004). The mature microRNA sequence is 20-24 nt in length, and binds to target mRNAs in the ribosomal pool, leading to repression of translation (Bhattacharyya *et al.* 2006; Flynt and Lai 2008). The mature microRNA is produced from a stem-loop hairpin precursor, termed the pre-miRNA, which in turn was excised in the nucleus from a longer primary genomic transcript (pri-miRNA) (Saini *et al.* 2007). Some pri-miRNA transcripts contain multiple clustered hairpin precursors (Altuvia *et al.* 2005). In mammalian genomes, such as mouse and human, hundreds of microRNA genes have been identified, mainly by deep sequencing of size-fractionated RNA libraries (Kozomara and Griffiths-Jones 2011). Approximately 30 to 50% of microRNA genes in animal genomes are located within protein-coding genes, termed host genes (Rodriguez *et al.* 2004). The overwhelming majority of these (> 85%) are found within intronic regions, and of these, around 80% reside in the same orientation to the host gene (Rodriguez *et al.* 2004; Kim and Kim 2007; He *et al.* 2012). While the primary transcripts of some intronic microRNAs are expressed from a promoter internal to the host gene, the majority are believed to be co-expressed with their host genes, as a single transcriptional unit, from a shared promoter region (Baskerville and Bartel 2005; O'Donnell, Wentzel *et al.* 2005; Shalgi *et al.* 2007; Ozsolak *et al.* 2008; He *et al.* 2012). Indeed, host promoter regions have been shown to drive the expression of microRNA genes (Johnson *et al.* 2003; Ozsolak *et al.* 2008). The microRNA hairpin is thus excised from the host pre-mRNA (Figure 3.1A), and this processing has been shown to be co-transcriptional (Morlando *et al.* 2008). Interactions between host transcript splicing and microRNA excision have also been demonstrated, and some of the same protein factors are involved in either process (Kim and Kim 2007; Morlando *et al.* 2008; Shomron and Levy 2009). The remaining 50% to 70% of microRNA genes that are located outside of protein-coding genes are termed intergenic microRNAs (Figure 3.1B).

Several studies have applied genome-wide experimental or sequence-based prediction techniques to the problem of locating the start sites of intergenic pri-miRNA transcripts (8,10–13). Features associated with transcriptional units that have been used include RNA polymerase II binding sites, expressed sequence tags, CAGE tags, and histone modifications associated with transcription initiation and elongation (Saini *et al.* 2007; Ozsolak *et al.* 2008; Corcoran *et al.* 2009; Wang, Wang *et al.* 2010; Chien, Sun *et al.* 2011). Without the guide of a protein-coding host gene annotation, characterization of primary microRNAs (pri-miRNAs)

expressed from intergenic microRNA genes has proven challenging, due to their rapid nuclear processing, and lack of recognisable signals of protein-coding gene structure (Saini, Enright *et al.* 2008). Thus, in the majority of cases, there is uncertainty in locating the promoter upstream of an intergenic microRNA, where the majority of *cis*-acting transcriptional regulators of microRNA gene expression are expected. With scarce exceptions (Kim *et al.* 2008; Toyota *et al.* 2008; Barski *et al.* 2009), intergenic microRNA genes have been considered to be autonomous transcriptional units, expressing one or a cluster of microRNAs, independently of other genes (Figure 3.1B) (Zhou *et al.* 2007; Chien *et al.* 2011). However, we propose at least two genomic arrangements that could link the transcriptional regulation of intergenic microRNAs with their neighbouring protein-coding genes: Either microRNAs may be transcribed as a result of run-through transcription of a protein-coding gene (Figure 3.1C), or a microRNA and a protein-coding gene may be expressed under the control of a bidirectional promoter region (Figure 3.1D).

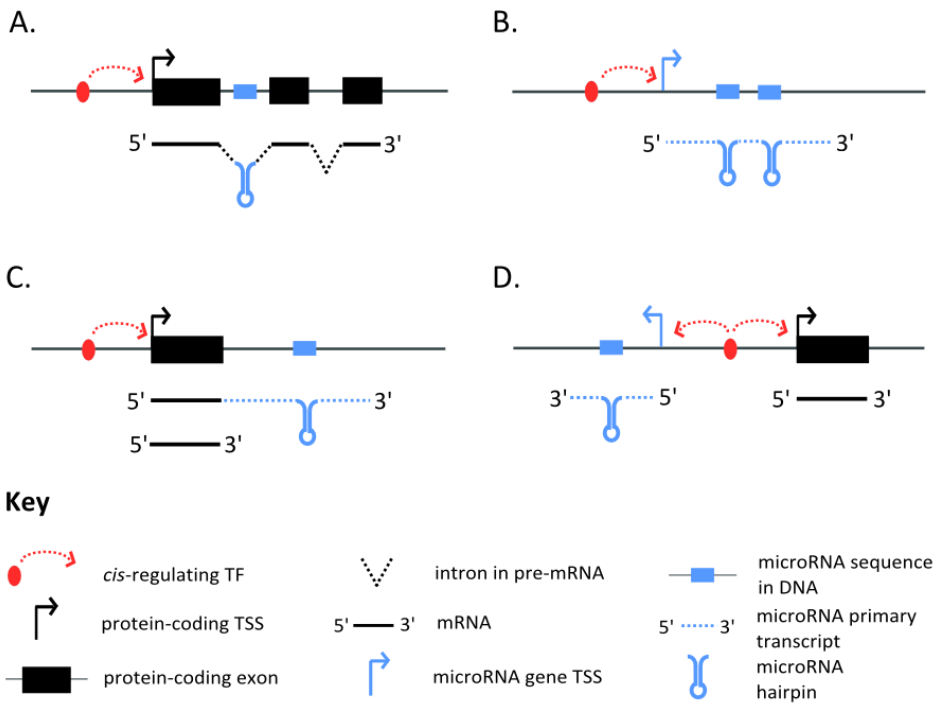


Figure 3.1. Arrangements of microRNA genes.

A. Intronic microRNA gene excised from a protein-coding pre-mRNA sequence.
B. Intergenic microRNA or cluster of microRNAs expressed from an autonomous promoter region.
C. Intergenic microRNA gene with a protein-coding gene neighbour lying upstream, and expressed as a consequence of run-through transcription of this gene.
D. Intergenic microRNA gene with a protein-coding gene neighbour lying upstream but in the antisense orientation to the microRNA. Both genes might be expressed from a shared bidirectional promoter region.

In this study, we have investigated the locations of microRNA gene sequences in intergenic space with respect to neighbouring protein-coding genes across 4 model animal species. We then map experimental datasets representing 106 human transcriptional regulators to *cis*-regulatory regions of protein-coding genes, and consider whether these regulators might also exercise control over the transcription of intergenic microRNAs. We finally ask whether intergenic microRNAs can participate in simple network motifs with the transcriptional regulators that bind to promoter regions of neighbouring protein-coding genes.

3.2 Methods

Datasets

Species and genome builds examined were *H. sapiens* (GRCh37), *M. musculus* (GRCm38), *C. elegans* (WBcel215), and *D. melanogaster* (BDGP5). Protein-coding gene coordinates were obtained from Biomart (Ensembl v.69) (Karolchik *et al.* 2003; Karolchik *et al.* 2004). Gene length was calculated as the full extent of the union of transcripts of a gene. The number of gene splice forms was simply the number of alternative transcripts annotated by Ensembl.

MicroRNA gene coordinates and mature sequences were obtained from miRBase v.19, for the same collection of genome builds (Griffiths-Jones *et al.* 2006; Kozomara and Griffiths-Jones 2011). MicroRNA target sets were predicted in the longest 3'-UTRs of protein-coding genes in Ensembl v.69 using the miRanda algorithm (v.3.3a) (John *et al.* 2004). We imposed a permissive minimum miRanda microRNA target score of 125, but key results were not affected by varying this parameter. We assigned neighbouring pairs of microRNAs to a cluster whenever the inter-microRNA distance was less than or equal to a linkage parameter L. We tested a range of values of L between 5 kb and 20 kb and found very little impact on microRNA clusters, so used the parameter L = 5 kb for all analyses.

Genome-wide human transcription factor ChIP-seq datasets were downloaded from the UCSC genome browser (101 datasets published by the YALE consortium and 171 pairs of replicate datasets from the HAIB consortium, publically available as of 1-July-2012) (Karolchik *et al.* 2003; Karolchik *et al.* 2004 ; 2011) (See Supplementary Table S1.1 for details).

Analysis of microRNA gene frequency density

To compare intergenic regions of variable lengths, we simply expressed microRNA position as a percentage of the distance through an intergenic region. By aligning the protein-coding boundaries of intergenic regions containing microRNA clusters, and rescaling coordinates of microRNA clusters accordingly, we were able to adapt the SeqMiner program to represent the genomic intervals containing intergenic microRNA clusters in human (Ye, Krebs *et al.* 2011). The significance of the difference in numbers of microRNAs between central and end regions of intergenic space was assessed using the binomial distribution. To assess instead the significance of microRNAs within *fixed* numbers of nucleotides from gene ends, we adopted a procedure that normalizes gene counts for the total number of intergenic nucleotides lying within a region of interest. For example, to calculate the probability of a microRNA cluster

falling nearest to and within 10 kb of a gene end, we first count all nucleotides nearest to and within 10 kb of gene ends genome-wide, and divide this by the total number of intergenic nucleotides. The binomial distribution can then be used to estimate how likely it would be to observe at least as many microRNA clusters within 10 kb (or any other interval of interest) around gene ends. Normalizing by total nucleotides within corresponding genomic intervals is particularly important when plotting gene frequencies against distance from gene ends, since the maximum possible distance from a gene end is constrained by the size of the neighbouring intergenic region. This is particularly important when comparing different species, since the size distributions of intergenic regions can differ enormously. Positions of intragenic microRNAs were treated in the same way, normalizing over total numbers of nucleotides found in specified intervals within protein-coding genes.

Chromatin state models

Annotations of genomic regions corresponding to 15 kinds of chromatin state trained using Hidden Markov Models were obtained from a published study, via the UCSC Table Browser (Hinrichs, Karolchik *et al.* 2006; Ernst, Kheradpour *et al.* 2011). Distinct states with identical names were then merged to give 12 states in total. The genomic intervals for each chromatin state were then intersected with intergenic nucleotides, relative to Ensembl v.69 protein-coding genes. The significance of numbers of microRNA clusters within each region type was assessed by the binomial test with n = total intergenic microRNA clusters and p = fraction of total intergenic space of the given type.

Transcription regulatory factor binding sites

Human transcriptional regulator ChIP-seq datasets were converted, where necessary, to the coordinates of genome build GRCh37 using the UCSC LiftOver tool (Hinrichs *et al.* 2006). Coordinates of ChIP-seq binding sites were mapped to proximal promoter regions of protein-coding genes, defined as the region from 1500 nt upstream to 500 nt downstream of the annotated 5’ end of each gene. This definition was varied by increasing the window size up to 2000 nt either side and found to have no effect upon the key results. Because a single TRF can be represented in many datasets, and have multiple ChIP-seq peaks within a single promoter region, we chose to count distinct TRFs rather than distinct ChIP-seq peaks. However, counting peaks instead had no effect on key results.

We define flanking gene - mediated microRNA/TRF feedback loops as cases where a microRNA is predicted to target the 3’-UTR of a TRF gene, and the TRF in turn binds the promoter region

of neighbouring protein-coding gene. We therefore mapped all TRF data to the promoter regions of protein-coding genes neighbouring microRNAs, and mapped all microRNA mature sequences to TRF gene 3'-UTRs. We then constructed matrices representing all TRF/promoter and microRNA/TRF target pairs, and counted (for various gene arrangements) all instances of candidate microRNA/TRF loops. We then shuffled TRF and microRNA target relationships 1000 times, maintaining the degree distributions of both the regulators and the targets in the network, and re-counted microRNA/TRF loops in each of the shuffled networks. P-values were calculated as the proportion of shuffled networks with at least as many microRNA/TRF loops as in the original network. For a particular arrangement of microRNA and protein-coding neighbour, we also counted the percentage of microRNAs with a greater number of microRNA/TRF loops in the original network, compared to the ensemble of shuffled networks. For a particular microRNA we also tested the association between the set of TRFs binding the neighbouring gene's *cis*-regulatory region, and the set of TRFs targeted by the microRNA, by means of Pearson's χ^2-squared test, corrected for multiple tests.

3.3 Results

3.3.1 Classification of microRNAs

We first compared coordinates of human, mouse, fruit fly and nematode microRNA genes from miRBase (v.19) to coordinates of protein-coding genes from the Ensembl (v.69) database (Griffiths-Jones *et al.* 2006; Pruitt *et al.* 2007; Kinsella, Kahari *et al.* 2011). MicroRNA genes were annotated as *intergenic* when lying fully outside of any protein-coding genes, and as *intragenic* when lying fully within a protein-coding gene, termed the *host gene*. Intragenic microRNA genes are further defined as *intronic* if lying on the same DNA strand as the host gene within an intron, and as *antisense* if lying on the opposite strand to the host gene. For the purposes of all analyses, microRNAs separated by less than 5 kb were combined into a single unit. Varying this distance between 5 kb and 20 kb affects only 11.5% of the un-clustered microRNA genes in human, mostly within two regions with high microRNA density (42 microRNA sequences located on chr14 and 46 microRNA sequences located on chr19). Protein-coding genes neighbouring intergenic microRNAs were termed *flanking genes*, and the space between two flanking genes was termed the *intergenic space*.

The numbers of protein-coding host genes, and intra- and intergenic microRNAs based on Ensembl gene builds for each species are shown in Table 3.1. Proportions of intragenic sense, intragenic antisense, and intergenic, microRNAs are consistent with previous reports (Rodriguez *et al.* 2004; Meunier *et al.* 2013), and broadly consistent between species, though with a smaller proportion of intronic microRNAs within *C. elegans*. Within this chapter, we focus mainly upon human data since this provides the largest sample of known microRNAs for analysis. Nevertheless, in section 3.3, a novel property of intergenic microRNAs is outlined, which is shared across each of these species.

Table 3.1. Numbers and types of microRNA genes in 4 animal species.

Species	Protein-coding genes	MicroRNA precursors	Host genes	Intragenic microRNAs	Antisense intragenic microRNAs	Intergenic microRNAs	Intergenic microRNA clusters
H. sapiens	20046	1594	696	751	198	658	486
M. musculus	22571	849	290	385	97	372	261
D. melanogaster	13831	238	112	119	22	97	65
C. elegans	17804	223	46	52	33	138	94

3.3.2 Intergenic microRNAs are found within regions of high protein-coding gene density

We next considered the locations of all human intergenic microRNA gene clusters within intergenic spaces. These 486 intergenic clusters are contained within 395 different intergenic regions, ranging in length from 0.54 kb to 22,800 kb (median = 192.0 kb). The 486 clusters were then sorted by the position of the centre of the cluster, expressed as a percentage of the distance from their upstream to downstream flanking genes. The distribution of cluster locations is shown graphically via 486 horizontally stacked genome tracks in Figure 3.2A. It is clear that human microRNA clusters are not distributed uniformly within intergenic space, with an excess of microRNA clusters nearer to the upstream flanking gene. The total number of clusters within each 10% interval of intergenic space were then counted (Figure 3.2B). We confirm that microRNA clusters are significantly more frequent, compared to the mean intergenic microRNA cluster density, within the 10% of intergenic space nearest to their upstream flanking genes ($p = 3.9x\ 10^{-6}$, by binomial test; see methods.). There is also a smaller microRNA cluster enrichment within the 10% of intergenic space nearest to their downstream flanking genes ($p = 0.021$). We conclude that human intergenic microRNA clusters are more frequently found near to protein-coding gene neighbours, particularly when these lie upstream of the cluster.

In the protein-coding case, it is well known that gene density along chromosomes varies, with regions of enrichment (gene islands) and depletion (gene deserts) (Carninci *et al.* 2005). We next tested whether microRNAs are associated with regions of chromosomes that have greater protein-coding gene density. The numbers of protein-coding transcription start sites (TSSs) were counted within 500 kb either side of (i) 10,000 randomly selected genomic coordinates, and around 5' ends of (ii) intergenic microRNA clusters, (iii) protein-coding genes, and (iv) protein-coding genes ≤ 2.5 kb from intergenic microRNAs. The resulting density distributions of numbers of TSSs are shown in Figure 3.2C. We find a mean value of 13.2 protein-coding TSSs / mb around random coordinates, but a much higher value of 18.8 protein-coding TSSs / mb surrounding microRNA clusters ($p = 2.3 \times 10^{-10}$ by K-S test), and 32.0 protein-coding TSSs / mb surrounding protein-coding TSSs ($p < 10\text{-}16$, by K-S test). Compared to all protein-coding genes, the subset with intergenic microRNAs nearby (<= 2.5 kb) are found within regions of even greater protein-coding TSS density (mean = 37.8 TSSs / mb; $p = 0.03$ by K-S test). These differences in protein-coding TSS densities remain broadly similar regardless of window sizes. Thus, intergenic microRNAs are associated with gene-dense genomic regions.

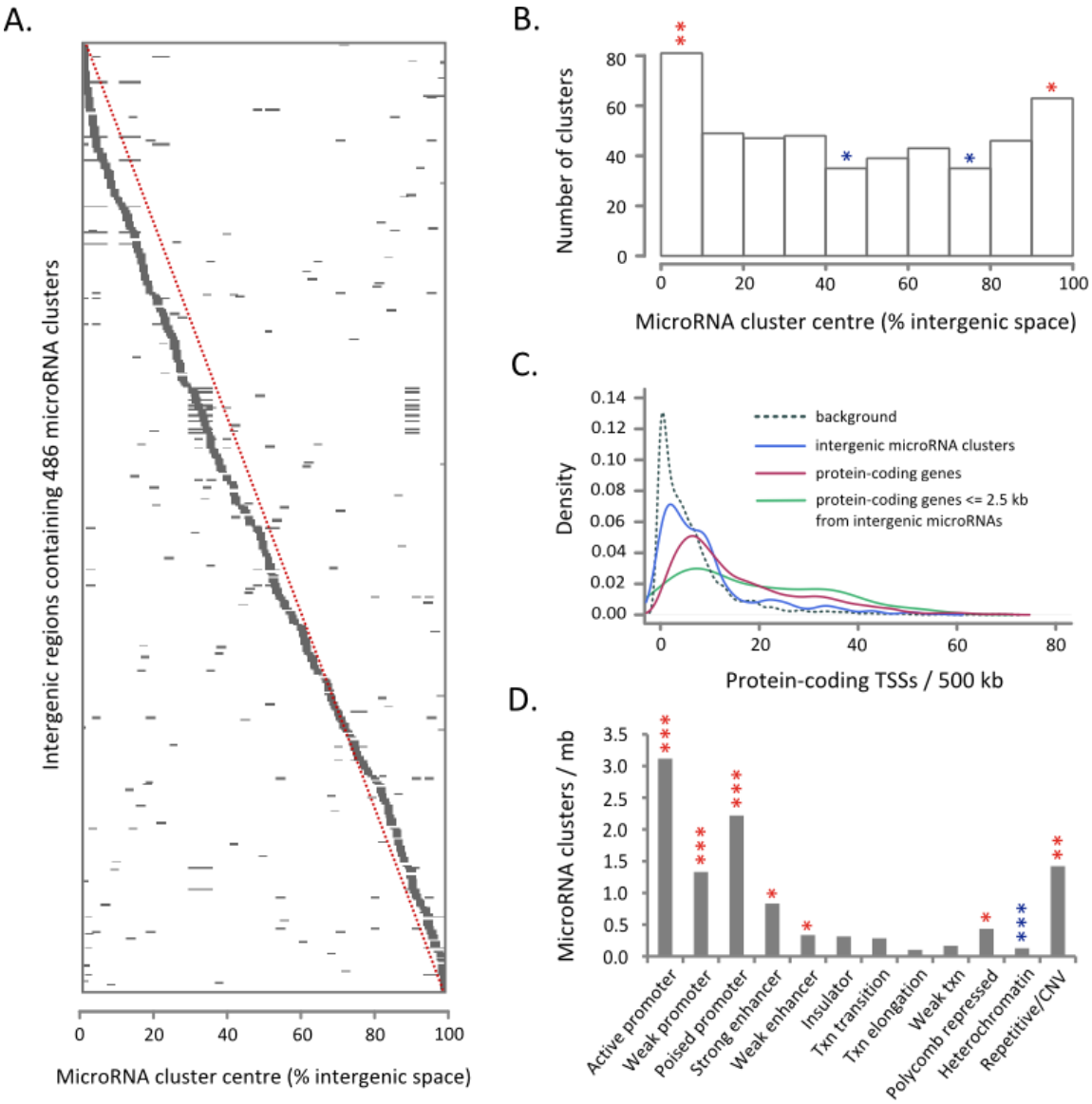


Figure 3.2. Protein-coding and chromatin environment of human intergenic microRNAs.

A. MicroRNA cluster positions within intergenic regions. Intergenic regions containing each of 486 microRNA clusters (grey dashes) are shown as a stacked collection of genome tracks, arranged from top to bottom according to the position of the microRNA cluster in the region. All tracks are oriented in the direction of transcription of the microRNA cluster. The position of the reference cluster within each region is expressed as a percentage of the distance from the upstream to the downstream flanking gene. Regions containing multiple clusters are represented as multiple rows in the diagram, once for each cluster in turn. The red dotted line approximates expected cluster locations given a uniform distribution in intergenic space. The figure was produced using SeqMiner (see methods) (Ye *et al.* 2011).

B. Frequency distribution of microRNA clusters within intergenic space

C. Density distributions of protein-coding TSSs around random genomic coordinates, intergenic microRNAs, and protein-coding genes. Protein-coding TSSs were counted lying within 500 kB of (i) 10,000 random genomic coordinates (background) (ii) every intergenic microRNA cluster 5' end, (iii) every protein-coding gene start (excluding itself), and (iv) 79 protein-coding gene starts within 2.5 kB of an intergenic microRNA.

Distributions of numbers of protein-coding TSSs are represented by density functions calculated in R.

D. MicroRNA cluster density within intergenic regions of different functional types. Functional annotations were obtained from the table 'hg19, regulation, BroadChromHmm, H1HESC' from the UCSC genome browser, based upon systematic studies of chromatin state (Karolchik *et al.* 2004; Ernst and Kellis 2010; Ernst *et al.* 2011). Abbreviations: Txn = transcription, CNV = copy number variation.

In B and D, significance of microRNA cluster enrichment within given genomic regions is indicated by coloured asterisks. Red = enriched, blue = depleted. *: $p < 0.05$, **: $p < 10^{-4}$, *** $p < 10^{-8}$. Significance was calculated using the binomial distribution, with (x, n) = number of clusters in (required region, total); and p = fraction of intergenic space of each functional type.

We next obtained publically available functional annotations of 12 different kinds of genomic regions, inferred from chromatin state, within 9 human cell lines (see methods) (Ernst and Kellis 2010). Chromatin state refers to distributions of covalent modifications especially to histone proteins, associated with the functional variations between regions, such as promoters, enhancers, or heterochromatin (Wang *et al.* 2008; Ernst *et al.* 2011). Numbers of intergenic microRNA clusters, divided by numbers of intergenic nucleotides within each region type in human embryonic stem cells, are shown in Figure 3.2D. The data show clearly that the density of intergenic microRNA clusters per nucleotide is greatest within active, followed by poised and then weak, promoter regions ($p < 10^{-8}$ by binomial test, n = 72 microRNA clusters in total). This is evidently consistent with the enrichment in microRNA clusters near to flanking gene regions. We find a smaller but still significant enrichment within repetitive sequences ($p = 3.9 \times 10^{-6}$, n = 9), enhancer regions (strong: $p = 5.2 \times 10^{-4}$; weak: $p = 7.8 \times 10^{-4}$, n = 42 in total), and within regions repressed by the developmentally critical Polycomb-group factors ($p = 1.7 \times 10^{-3}$, n = 16) (Boyer, Plath *et al.* 2006). Almost 50% of intergenic microRNA clusters are located within heterochromatin (n = 237), but since this accounts for 89.2% of intergenic nucleotides, microRNA clusters are nevertheless significantly depleted within this type of chromatin ($p < 10^{-25}$). Results within the other 8 cell lines are similar, with examples of some minor variations between cell lines shown in Supplementary Figure S3.1. Thus, the density of microRNA genes within intergenic regions varies markedly with chromatin state.

3.3.3 MicroRNAs are proximal to protein-coding genes across animal species

We surmise that intergenic microRNAs have a greater likelihood of interacting with their flanking genes when separated by a smaller distance along chromosomes. In the remainder of this study, we therefore focus upon the microRNAs that lie within 10 kb of a protein-coding neighbour, which we term proximal microRNAs. Percentages of intergenic microRNA clusters proximal to protein-coding genes were calculated within human, mouse, fruit fly, and

nematode worm, and compared to the percentages of total intergenic nucleotides within 10 kb of protein-coding TSSs in each of these genomes (Table 3.2). The fraction of intergenic microRNAs that are proximal is very significant within human and mouse ($p < 10^{-15}$ by binomial test). The fraction of intergenic microRNAs that are proximal is less significant in fruit fly ($p = 4 \times 10^{-6}$) and not significant in nematode ($p = 0.18$). This is due to the compactness of invertebrate genomes, such that a much higher percentage of the intergenic genome is already proximal to a protein-coding gene (Table 3.2). Nevertheless, within smaller distances of flanking genes, we do find a significant enrichment in *C. elegans* intergenic microRNA clusters (e.g. $p < 10^{-4}$ for microRNA cluster centres within 5 kb of the flanking genes' ends, by binomial test). Thus, the favoured association of microRNA clusters with regions relatively near to flanking genes is a shared characteristic of these animal genomes.

Table 3.2. Enrichment of proximal intergenic microRNAs in mammals and invertebrates.

Species	% intergenic genome ≤ 10 kb from protein-coding gene	Clusters ≤ 10 kb of flanking gene (% intergenic clusters, p-value)
H. sapiens	2.3	151 (31.1, $< 10^{-15}$)
M. musculus	4.1	88 (33.6, $< 10^{-15}$)
D. melanogaster	50.6	51 (78.5, 4×10^{-6})
C. elegans	96.7	93 (98.9, 0.18)

Total intergenic space excludes centromeric regions. P-values represent significance of enrichment of microRNA clusters within 10 kb of flanking gene ends, using the binomial test (n = total intergenic microRNA clusters, p = fraction of intergenic nucleotides no greater than 10 kb from a protein-coding TSS).

3.3.4 Favoured orientations of proximal microRNAs

We reasoned that the direction of transcription of a protein-coding neighbour might influence the number of proximal microRNAs upstream or downstream of the protein-coding gene, and in either the sense or antisense orientations. To test this hypothesis, we calculated the densities / kb of microRNA clusters lying within 10 kb either side of protein-coding TSSs and transcription end sites (TESs) in human, mouse, fruit fly and nematode. We considered separately microRNA clusters in the sense and antisense orientations relative to the protein-coding gene. Distributions of human and nematode microRNAs around protein-coding TSSs are shown in Figure 3.3. The distributions of microRNA clusters around mouse and fruit fly TSSs, and around TESs in each of the 4 species, are provided in Figure 3.4.

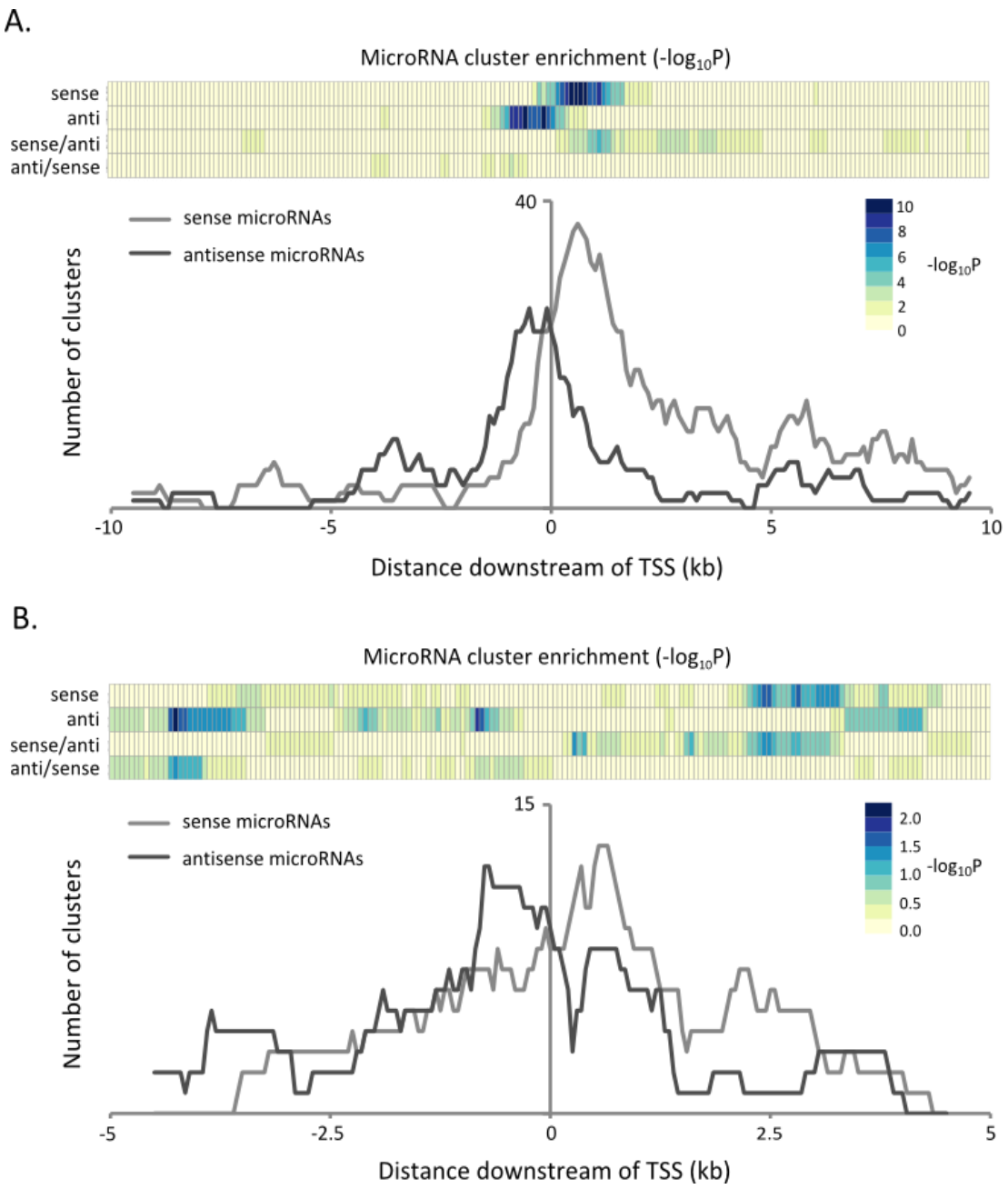


Figure 3.3. Distributions of sense and antisense microRNA clusters around protein-coding TSSs.

A. Human **B.** Nematode

MicroRNA cluster frequencies were calculated within sliding windows of size 1 kb. The heatmap above each microRNA cluster density graph was created using the pheatmap function in R, and displays $-\log_{10}(P)$ values for significance of the numbers of sense and antisense microRNA clusters, and for the excess of sense over antisense clusters (sense/anti), and antisense over sense clusters (anti/sense), within each window. Significance of sense and antisense microRNA cluster frequencies within each window was assessed by binomial test, with parameters n = total clusters in the required orientation, and p = fraction of genomic space within the required window. Significance of the excess of sense over antisense, or of antisense over sense, was calculated by binomial test, with parameters n = total clusters in both orientations within a given window, p = 0.5.

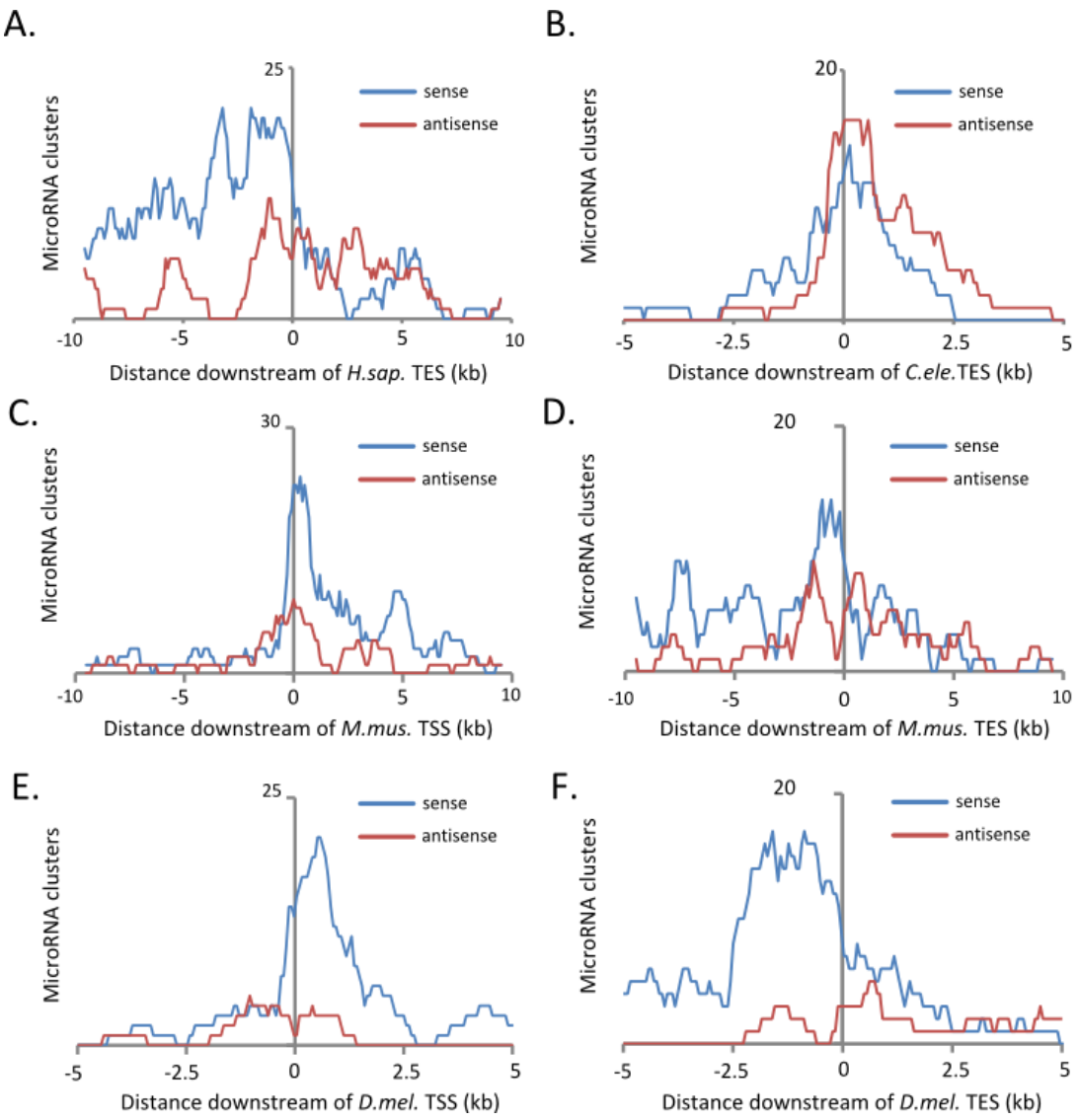


Figure 3.4. Distance distributions of microRNA clusters around protein-coding TSSs and TESs, across animal species.

A. Human sense and antisense microRNA clusters around protein-coding TESs
B. Nematode sense and antisense microRNA clusters around protein-coding TESs
C. Mouse sense and antisense microRNA clusters around protein-coding TSSs
D. Mouse sense and antisense microRNA clusters around protein-coding TESs
E. Fruit fly sense and antisense microRNA clusters around protein-coding TSSs
F. Fruit fly sense and antisense microRNA clusters around protein-coding TESs

MicroRNA gene cluster frequencies are calculated within sliding windows of size 1 kb. Distributions around human and nematode TSSs are shown within Figure 3.3. The representative species together with human are *C. elegans*, *M. musculus*, and *D. melanogaster*.

We find a clearly asymmetrical distribution of microRNA orientations around protein-coding gene start coordinates in both human and nematode. As previously reported, within protein-coding genes, sense microRNAs significantly outnumber antisense microRNAs (see also Table 3.1) (Rodriguez *et al.* 2004; Baskerville and Bartel 2005). In contrast, within all 4 species examined, antisense intergenic microRNAs outnumber sense intergenic microRNAs upstream of the TSS (Figures 3.3 and 3.4:C,E). In human, the enrichment in antisense microRNAs per nt within 1 kb upstream of TSSs is extremely significant ($p \approx 10^{-8}$ - 10^{-6}, as indicated in the antisense track of the heatmap in Figure 3.3A). The number of antisense microRNAs is also significantly greater than the number of sense microRNAs within most intervals up to around to 2 kb from the TSS ($0.0026 \leq p \leq 0.0832$, dependent upon the window). This switch between microRNA cluster orientations is precisely coincident with annotated protein-coding TSSs, in both human and nematode. It is therefore unlikely that these distributions arise from any systematic error in the annotation of protein-coding genes. We conclude that an asymmetrical distribution of microRNA gene orientations around the 5' ends of protein-coding genes is likely to be a shared characteristic of many animal genomes.

Our analysis also highlights interesting contrasts between the vertebrate and invertebrate cases. In the human case, microRNA gene frequencies align with the significance of these frequencies per unit nt, as shown by the heatmap in Figure 3.3A. This is not true, however, for *C. elegans* microRNAs, with the maximum sense and antisense frequencies very near to TSSs not being significant per unit nt (Figure 3.3B). Indeed, *C. elegans* sense and antisense microRNA gene densities are much more significant between around 2.5 kb and 5 kb downstream, and upstream, respectively, from TSSs, even though total numbers of microRNA genes are much lower. The reason for this paradoxical result is again due to the compactness of the *C. elegans* genome, so that the majority of nucleotides lie very near to protein-coding TSSs. This emphasizes the necessity of the statistical tests performed here. Regardless, differences in the densities of genes between vertebrate and invertebrate genomes cannot account for the switch in preferred orientations of microRNAs at gene boundaries. We finally note that sense microRNAs also predominate within genes at the 3' ends, but that a switch in preferred orientation downstream from the gene region is much less evident (Figure 3.4:A,B,D,F). In addition, compared to profiles around gene TSSs, the profiles of sense and antisense microRNAs around gene TESs appear much more variable between these 4 species.

3.3.5 Comparison between microRNA flanking and host genes

We asked whether properties of neighbouring protein-coding genes match those of known host genes of microRNAs. It has previously been shown that microRNA host genes have

distinctive distributions of fundamental properties such as length and number of splice forms (Golan *et al.* 2010). In addition, microRNA host genes are highly regulated, reside within regions of open chromatin, are older than average, and are enriched in developmental functions (Chapter 2). We retrieved all 776 protein-coding genes flanking regions containing intergenic microRNAs. We searched for overrepresented Gene Ontology (GO) terms using the DAVID functional annotation server (Ashburner *et al.* 2000; Dennis *et al.* 2003; Huang da *et al.* 2009)). Although no single terms are significant at $p < 0.05$, surprisingly, the most enriched functional categories are developmental categories (neuron differentiation: $p = 0.073$, and eye morphogenesis: $p = 0.078$. Using the DAVID functional clustering tool, these form significant clusters (Enrichment score ≥ 2) relating to angiogenesis, eye development, and cell morphogenesis. This suggests that both intragenic and intergenic microRNAs are preferentially born within genomic regions having characteristics favourable to the expression of developmental genes.

Despite this thematic link with microRNA host genes, we find that 138 flanking genes within 10 kb of intergenic microRNAs are similar to the genomic average in respect of gene length and number of splice forms, and are significantly different in these respects from microRNA host genes (length: $p < 10^{-8}$; splice forms: $p < 10^{-6}$, K-S tests). We note that the length distributions of the genes flanking proximal microRNAs suggests that they are very unlikely to be, in general, incorrectly annotated microRNA host genes, i.e. genes with missing terminal regions that include the proximal microRNAs. This provides additional evidence that the enrichment of microRNAs near to protein-coding genes, and of 5'-antisense microRNAs in particular, is not generally likely to be explained by errors in gene annotation.

3.3.6 *Cis*-regulatory regions of flanking genes

We have shown that intergenic microRNAs as a whole are preferentially associated with active, poised, and weak promoter regions. We next investigated the transcriptional properties of genes flanking proximal microRNAs, using publically available ChIP-seq datasets corresponding to 106 human transcriptional regulatory factors, mapped to the promoter regions of all protein-coding genes. MicroRNAs upstream of their protein-coding neighbour, and on the same genomic strand, were termed 5'-sense (5'S) microRNAs, with 5'-antisense (5'A), 3'-sense (3'S), and 3'-antisense (3'A) microRNAs defined analogously. We compared the total numbers of TRFs found within promoter regions of protein-coding genes flanking microRNAs within 10 kb, and in each of the four possible arrangements. In total there are 29 5'S, 52 5'A, 32 3'S and 29 3'A microRNA clusters that are proximal to flanking genes. We find that microRNAs in both 5'-antisense and 3'-sense positions are associated with flanking genes that have higher

numbers of TRF binding sites in their promoter regions (Figure 3.5A). The differences between mean numbers of TRFs binding to flanking genes in these orientations, compared with all protein-coding genes as background, are statistically significant ($p = 0.0025$ for genes with 5'-antisense microRNAs, $p = 0.0415$ for 3'-sense microRNAs by K-S test). These cases both correspond to microRNA clusters lying transcriptionally downstream of the protein-coding promoter region (as in Figure 3.5A). In both cases, we therefore suggest that expression of the microRNA may be directly regulated by the promoter region of the flanking gene.

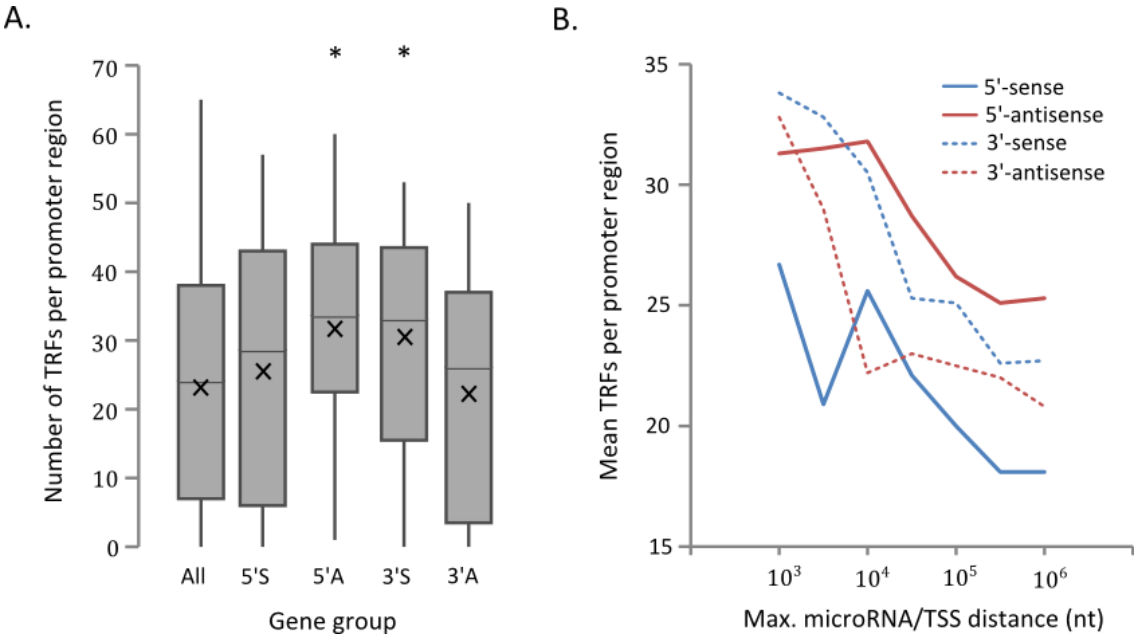


Figure 3.5. Numbers of transcription factors binding promoter regions of microRNA flanking genes.

A. Distributions of numbers of TRFs in promoter regions of genes flanking proximal intergenic microRNA clusters. We consider flanking genes with 5'-antisense (5'A), 5'-sense (5'S), 3'- antisense (3'A), and 3'-sense microRNAs (3'S). The background distribution for all protein-coding genes is shown alongside. The boxes show quartiles (Q1, Q2, Q3); the whiskers show the minimum and maximum values; the crosses represent the mean values; asterisks show gene groups with significantly greater numbers of TRFs ($p < 0.05$ by K-S test).

B. Mean numbers of TRFs across all flanking genes, as maximum distance to nearest neighbouring intergenic microRNA nearest neighbours is varied. Thus, for example, points plotted at 10^4 nt = 10 kb correspond to the mean numbers of TRFs for flanking genes of proximal microRNAs, as in 3.4A.

We next considered whether the number of TRF binding sites in a protein-coding gene varies with distance from intergenic microRNAs. The result is striking: the nearer the microRNAs are to their neighbouring protein-coding genes, the more TRFs are found in the promoter regions of those genes (Figure 3.5B). This confirms that microRNA gene clusters are, on average, found

nearer to more transcriptionally active gene promoters. This relationship holds up to at least a separation of 100 kb. For all arrangements of genes, maximum regulation occurs at the smallest distances examined, i.e. within ≈ 1 kb of a gene end. It is evident that intergenic microRNAs in humans are located closer to protein-coding genes whose promoter regions contain greater numbers of TRF binding sites. Further, over most distances, the total number of binding sites within promoter regions is greater for genes flanking 5'-antisense rather than 5'-sense microRNAs, and for genes flanking 3'-sense rather than 3'-antisense microRNAs (Figure 3.5B). Therefore, as before, we conclude that intergenic microRNAs are preferentially located downstream of transcriptionally active protein-coding promoter regions.

3.3.7 Candidate bidirectional promoters of intergenic microRNA and protein-coding gene pairs

We have shown that intergenic microRNAs are located more often than expected in close proximity to protein-coding genes. In particular, divergent microRNA and protein-coding genes are significantly over-represented ($p < 0.002$ compared with other orientations). We have further argued that these associations are not, in the general case, due to missing annotation of terminal exons of the protein-coding genes. These divergent microRNA/protein-coding genes pairs therefore represent candidates for control by bidirectional promoter regions.

In total, 61 5'-antisense microRNA clusters were identified within 10 kb upstream of a protein-coding neighbour (Supplementary Table S3.2. N.B. Nine of these clusters are also nearer to their other flanking protein-coding gene). Using the DAVID functional analysis server, we found that the 61 protein-coding genes are not significantly enriched in any GO category, but rather are spread over diverse functions (Dennis *et al.* 2003; Huang da *et al.* 2009). These include conserved protein-coding genes with pivotal functions in cellular biochemistry, for example the translation initiation factor EI3FL, ribosomal protein S5 (RPS5), the ubiquitous transcription factor JUND, the δ subunit of RNA polymerase III, and genes involved in the control of the cell cycle (e.g. BTG4, CWC35, RAB11).

In human, to our knowledge only two bidirectional promoter regions with both a microRNA and protein-coding product have been experimentally identified: the BTG4 promoter upstream of mir-34b,-34c (Toyota *et al.* 2008), and the POLR3D promoter upstream of mir-320a on chromosome 8 (Kim *et al.* 2008). The mir-34b,-34c case is particularly interesting, since mir-34b is known to regulate CREB (a TF with cell cycle regulatory roles (Rajabi, Baluchamy *et al.* 2005)) in acute myeloid leukaemia (Pigazzi, Manara *et al.* 2009), while the neighbouring protein-coding gene BTG4 has antiproliferative properties and can induce G1 cell cycle arrest

(Winkler 2010). The functions of protein-coding gene and microRNA are thus connected. The second experimentally characterised bidirectional promoter expresses primary transcripts of both mir-320a and of the gene encoding the δ subunit of RNA polymerase on human chromosome 8 (POLR3D). In this case, the microRNA gene has an atypical function, repressing expression of the neighbouring POLR3D gene by transcriptional silencing of POLR3D pre-mRNAs in the nucleus (Kim *et al.* 2008). The paralogous microRNA mir-320b-1 is located in the 5'-antisense orientation to another, apparently unrelated, protein-coding gene IGSF3 on human chromosome 1 (Supplementary Table 3.2). This gene encodes a member of immunoglobulin superfamily 3 but has not been studied in detail.

The candidate bidirectional microRNA/gene pairs include examples of well-studied microRNAs. For example, a member of the developmentally critical oncogenic let-7 family (let-7i) (Zhang, Ma *et al.* 2012) lies 252 nt from the 5' end of a candidate protein-coding gene. The microRNA mir-22, which is associated with several cancers (Huang, Wang *et al.* 2012; Ling, Wang *et al.* 2012; Szczyrba, Nolte *et al.* 2013), is upstream of a neurological disorder-linked gene WDR81 (Gulsuner, Tekinay *et al.* 2011; Sarac, Gulsuner *et al.* 2012). We draw particular attention to the mir-146b/CUEDC2 pair: both genes are individually diagnostic of aggressive grades of specific cancers. CUEDC2 is involved in cell mitotic spindle checkpoint, and regulates the JAK1/STAT3 pathway (Gao, Li *et al.* 2011; Man and Zhang 2011). Over-expression of CUEDC2 leads to chromosomal missegregation, and the protein is highly expressed in many tumour types. The mir-146 family is associated with differential outcomes for tumours of different types, including glioma, breast cancer, non-small-cell lung carcinomas, and the most aggressive forms of papillary thyroid carcinoma (PTC) (Hurst, Edmonds *et al.* 2009; Katakowski, Zheng *et al.* 2010; Yip, Kelly *et al.* 2011; Malleter, Jacquot *et al.* 2012). If the expression of the two products, mir-146b and CUEDC2, is linked through a shared promoter region, then there is a significant possibility of interaction between the components of their downstream pathways in human cancers. Despite a growing literature for both mir-146b and CUEDC2, and their proximity in the genome, there has so far been no attempt to consider them in conjunction.

3.3.8 Functional association of transcription factors and microRNAs

MicroRNAs and TRFs have frequently been shown to be involved in regulatory feedback loops, mutually regulating one another's expression, e.g. (O'Donnell *et al.* 2005; Hobert 2006; Shalgi *et al.* 2007; Yu *et al.* 2008; Brabletz *et al.* 2011). Thus, if 5'-antisense microRNAs are transcribed from a bidirectional promoter region shared with the neighbouring protein-coding gene, we might expect to find that such microRNAs regulate the expression of the TRFs that bind to the promoter region. Conversely, we do not expect a protein-coding gene promoter

region to control transcription of 3'-antisense microRNAs, so no enrichment in targeting of the relevant TRFs by the microRNA is expected. We therefore asked whether numbers of calculated microRNA/TRF loops varied according to the position and orientation of the microRNA, and its distance from a protein-coding neighbour. To this end, we predicted target sites of microRNAs in the 3'UTRs of mRNAs encoding all of the 106 TRFs used in our study, using two independent microRNA target-prediction algorithms (see methods). For each proximal microRNA, the collection of TRFs is divided into four groups, either present or absent from the flanking gene's promoter region, and with or without a microRNA target site in the TRF gene's 3'UTR. We then test the association between the TRF set targeted by the microRNA and the TRF set that binds the neighbouring gene's promoter region. For each microRNA, we can identify the number of TRFs with which it forms a potential microRNA/TRF loop via the flanking gene's promoter region. We then record the percentage of microRNAs with greater than random numbers of microRNA/TRF loops for the four genomic arrangements of microRNAs with their flanking genes, and assess the statistical significance of these percentages by simulation (see methods). The results of these tests are shown in Table 3.3.

Table 3.3. Regulatory loops between microRNA and flanking gene promoter regions.

Position and orientation of microRNA relative to flanking gene	% microRNAs enriched for microRNA/TRF loops (p-value)	Individual microRNAs significantly enriched for microRNA/TRF loops
5'-sense	59.6 (0.19)	mir-194-2
5'-antisense	65.9 (0.07)	mir-22 mir-34b/-34c mir-3677 mir-4470 mir-4795
3'-sense	71.7 (0.01)	mir-4502 mir-197 mir-3197-2
3'-antisense	52.2 (0.39)	None

Results are shown for microRNA target sets calculated using miRanda (methods). Significant microRNAs are those with individually significant chi-squared statistics for association between their target TRFs and TRFs binding their flanking gene's promoter, at the 5% level after Benjamini-Hochberg multiple-testing correction (Benjamini *et al.* 2001).

The data clearly show that microRNAs close to protein-coding genes are more likely to target TRFs binding *cis*-regulatory regions upstream of the microRNA (3'-sense and 5'-antisense cases). The percentage of microRNA/TRF loops is significantly greater than expected for 3'-sense microRNAs ($p = 0.01$) and weakly significant for 5'-antisense microRNAs ($p = 0.07$). We find the most individually significant microRNA/gene pairs in the 5'-antisense orientation (5 microRNAs), and fewest in the 3'-antisense orientation (none). The significant 5'-antisense microRNA/TRF loops include the two-member cluster miR-34b/-34c, which is known to be expressed in the antisense direction from the adjacent promoter of the gene BTG4, as discussed above (Toyota *et al.* 2008). We also find that the percentage of microRNA/TRF loops falls as the distance between microRNA and flanking gene increases. For example, only 43.4% of 5'-antisense microRNAs between 10 and 1000 kb from a protein-coding neighbour are found more often than expected in microRNA/TRF loops, compared with 65.9% for all 5'-antisense microRNAs within 10 kb. We infer that the transcriptional connection between the protein-coding promoter region and the microRNA is broken for more distal microRNAs.

3.4 Discussion

Many lines of evidence suggest that microRNAs may be born *de novo* by the formation of microRNA precursor-like hairpin structures in pre-existing transcripts. For example, *de novo* microRNA birth may be responsible for the widespread presence of microRNA sequences within the introns of protein-coding host genes (Campo-Paysaa, Semon *et al.* 2011), and the expansion of microRNA clusters (Altuvia *et al.* 2005) (Figure 3.1A and 3.1B). Our analyses suggest two further transcriptional substrates in which novel microRNAs may commonly evolve: run-through transcripts of regions downstream of protein-coding genes (Figure 3.1C), and antisense transcripts from bidirectional promoter regions (Figure 3.1D). The work presented here provides empirical support for this model. We show that microRNA frequency is non-uniform within intergenic space, with significantly more microRNA genes close to the starts and ends of flanking protein-coding genes. We note that this is consistent with the controversial suggestion that the majority of all transcripts are associated with known coding regions (Carninci *et al.* 2005; van Bakel, Nislow *et al.* 2010). In addition, it may be consistent with the pervasive expression of short transcripts from gene TSSs and TESs (Carninci *et al.* 2006; Seila *et al.* 2008; Taft, Glazov *et al.* 2009). However, in previous analyses the typical lengths distributions of these transcripts did not match the typical lengths of microRNA precursors, shifting attention away from this functionally important case (Seila *et al.* 2008; Taft *et al.* 2009). Since 30% of human intergenic microRNAs reside within 10 kb of the 5’ or 3’ end of a known protein-coding gene, our findings relate to a very broad class of intergenic microRNAs.

We find that the *cis*-regulatory regions of genes adjacent to 5’-antisense and 3’-sense microRNAs are bound by significantly more transcription factors. Therefore, promoter regions of flanking genes are most likely to have high levels of transcriptional activity when lying upstream of the microRNA. In cases of both 5’-antisense and 3’-sense microRNAs, TRFs that bind the promoter of the protein-coding gene potentially control the expression of the microRNA. These observations suggest a simple model for a feedback loop involving TRFs and microRNAs. Figure 3.6 illustrates the 5’-antisense microRNA case, and an analogous model can be described for 3’-sense microRNAs. Under this model, a TRF regulates the expression of a protein-coding gene. The promoter region of the protein also regulates the expression of an upstream antisense microRNA by bidirectional transcription, or of a downstream sense microRNA by transcriptional run-through. The microRNA is able to post-transcriptionally repress the expression of the TRF in a negative feedback loop. Consistent with this model, we

find that microRNA / TRF pairs of these types are found more frequently than expected by chance when the microRNA lies downstream of the TRF binding sites.

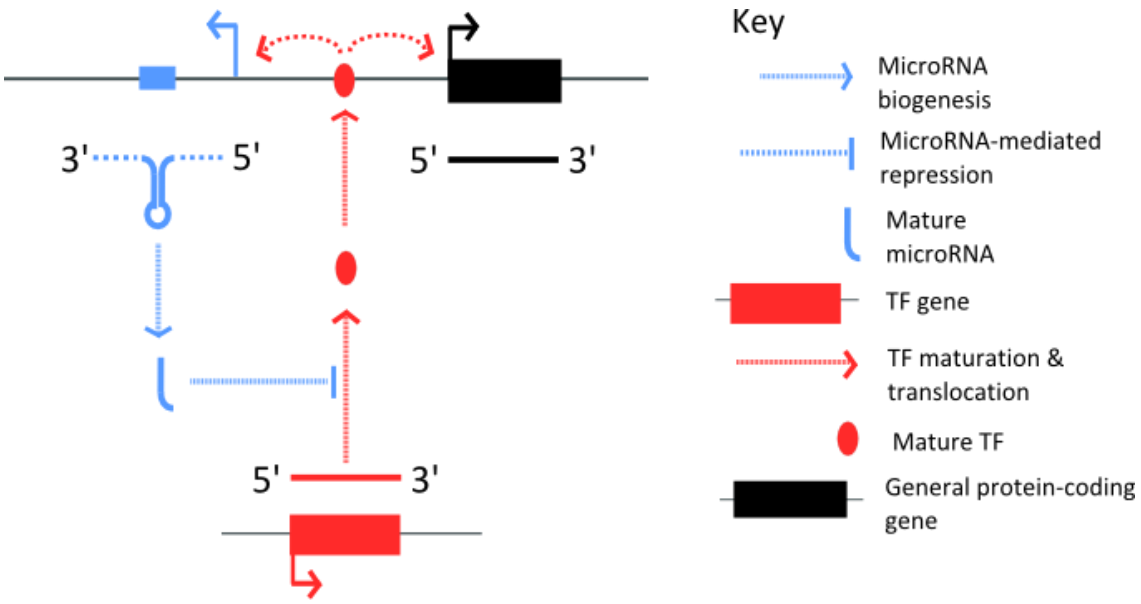


Figure 3.6. TRF/microRNA feedback loops via bidirectional promoter regions.

The figure displays a feedback loop involving a microRNA and protein-coding gene expressed from a bidirectional promoter region, and a TRF that regulates their expression. Symbols and colour scheme not shown within the key are as in Figure 3.1.

We offer a simple interpretation for these results. Intuitively, a microRNA-like hairpin sequence which arises randomly in the genome has a greater likelihood of becoming fixed and functional if it is located downstream of actively transcribed sequences. This could lead to higher rates of birth and retention of intergenic microRNA genes near to, and especially downstream of, active protein-coding neighbours, reflected in greater numbers of microRNAs proximal to protein-coding genes. Functional connections between protein-coding regions and intergenic microRNAs are further indicated by the different regulatory properties observed depending upon the orientation and position of the microRNA gene, as discussed above. Although intergenic microRNAs are conventionally regarded as transcriptional units independent from protein-coding machinery and regulation, this work suggests a shift in perspective may be appropriate, to a model in which the expression of many intergenic microRNAs is coupled with the expression of protein-coding genes. This work therefore has implications for correct description of the evolution, transcription, and function of a significant fraction of microRNAs in animal genomes.

3.5 Conclusions

We have shown within this study that intergenic microRNA genes have a highly non-random distribution within animal genomes. They are significantly more likely to reside near to and downstream of protein-coding neighbours, and in these cases, the neighbouring genes are transcriptionally active and highly regulated loci. Our analysis of microRNA/TRF feedback loops strengthens the case for functional connections between subsets of intergenic microRNAs and the transcription factors that regulate their upstream protein-coding genes. In particular, we have focused upon the subset of microRNAs that are enriched in the antisense orientation upstream of protein-coding TSSs. This identification can lead to a broad generalization from rare known cases of bidirectional transcriptional control of a microRNA gene, and its protein-coding neighbour. We have highlighted one example from the literature where potential functional connections abound between a bidirectionally arranged oncomir and oncogene (mir-146b/CUEDC2). Our study suggests that co-regulation of microRNA genes and their protein-coding neighbours may be a widespread phenomenon within animal genomes.

Chapter 4

Feedforward regulation by transcription factors and microRNAs

Abstract

The expression of messenger RNAs (mRNA) from protein-coding genes is regulated by transcription factors. In turn, the translation of the mRNA into a protein product can be regulated repressively by microRNAs. Here, we examine relationships between the transcriptional and post-transcriptional regulatory layers. We have mapped ChIP-seq data for 117 TFs and other transcriptional regulators in human to microRNA and protein-coding genes, and predicted the targets of 1919 mature microRNA sequences, using two representative microRNA target prediction methods. Our network reflects the largest so far constructed in human. We find that transcriptional regulators with both activator and repressor functions are most likely to participate in a bidirectional feedback loop with a microRNA. As many studies in smaller networks have indicated, feedforward loops consisting of a transcriptional regulator and microRNA pair targeting common genes occur significantly more often than expected by chance. We then score every protein-coding gene according to its frequency of membership in such feedforward loops. We find that, when a transcriptional regulator and microRNA pair regulate common genes, with the microRNA lying downstream of its transcriptional regulatory partner, then their common target genes have less variable expression levels across tissues. This result does not hold when the microRNA lies upstream of its transcriptional regulatory partner. This supports a generalized function for the post-transcriptional microRNA-mediated regulatory layer in the buffering of transcriptional noise. We then identify cellular processes, including several signalling pathways, that have the most significant enrichment in these putatively stabilising TF – microRNA feedforward loops.

Contributions

The research within this chapter was supervised by Dr. Sam Griffiths-Jones.

4.1 Introduction

Noncoding RNAs (ncRNAs) that do not encode a protein product are recognized throughout all life (Griffiths-Jones 2007). Our study will focus upon the microRNA family, an abundant class of small ncRNA post-transcriptional regulators of gene expression (Bartel 2004). Regulation of gene expression at the transcriptional level, by transcription factor (TF) proteins, has led to the concept of gene regulatory networks connecting TF and target gene pairs. In the human genome there are believed to be no fewer than 1400 genes encoding TFs, and many of these can regulate expression of thousands of genes (Vaquerizas, Kummerfeld *et al.* 2009; 2011). Networks of interactions are built up through combinations of activator and repressor TFs, with transcriptional regulation varying between genes, cell types, in the life-cycle of a single cell, through development, and in response to external change (Latchman 1990; Gerstein *et al.* 2012). The aim of this chapter is to study how transcriptional regulatory networks are modified through post-transcriptional regulation by microRNAs. This leads to the analysis of an integrated regulatory network (IRN) of TFs, microRNAs, and target gene expression pathways (Yu *et al.* 2008).

MicroRNAs are endogenous 20-24 nt ncRNAs found in all animals and plants. The functional mature microRNAs are produced from primary transcripts by a multi-step biogenesis pathway (Bartel 2004; Griffiths-Jones 2007). Mature microRNAs bind partially complementary regions of 10s – 1000s of target mRNAs, most often in the 3'-untranslated regions (3'-UTRs). The critical region for microRNA-mRNA binding is the seed sequence, a run of 6 – 8 nucleotides at the 5' end of the microRNA, and in general, the longer the seed match region, the more effective the binding interaction (Grimson *et al.* 2007). Binding can lead to (i) inhibition of translation initiation or ribosome processivity, or (ii) to degradation or sequestration of the target mRNA from the ribosomal pool, both functions catalysed by the RNA Induced Silencing Complex (RISC) (Bhattacharyya *et al.* 2006; Flynt and Lai 2008). Like TFs, microRNAs usually do not function in isolation but as one of a collection of regulators with common targets (Gerstein *et al.* 2012; Xu, Chen *et al.* 2013). In total, more than 1500 microRNA precursors have been identified in the human genome (Griffiths-Jones *et al.* 2006). These are thought to be involved in post-transcriptional silencing, or repressive regulation, of most protein-coding genes (Friedman *et al.* 2009).

Apart from their canonical function as repressors of gene expression pathways, a number of alternative functions of the microRNA regulatory layer have been proposed. These include buffering transcriptional noise in order to stabilise the protein complement of a cell, providing

greater determinism to the process of cell differentiation, and allowing specialization of cells into a greater range of subtypes (Peterson, Dietrich *et al.* 2009; Herranz and Cohen 2010). Many studies identified prevalent network patterns (motifs) based on computationally annotated TF binding sites, which may act as thematic regulatory structures to either stabilise the network, or move it toward some new state (as in cellular differentiation) (Shalgi *et al.* 2007; Yu *et al.* 2008; Tu *et al.* 2009) (Figure 4.1). Mathematical modelling shows that proposed functions of the post-transcriptional regulatory layer, such as noise reduction, can arise from regulators with particular signs and connections between them. For example, if an activating TF drives expression of a repressive microRNA, which preferentially binds mRNAs under regulation by the same TF, then fluctuations in mRNA and protein levels may be reduced (Figure 4.1D) (Osella *et al.* 2011).

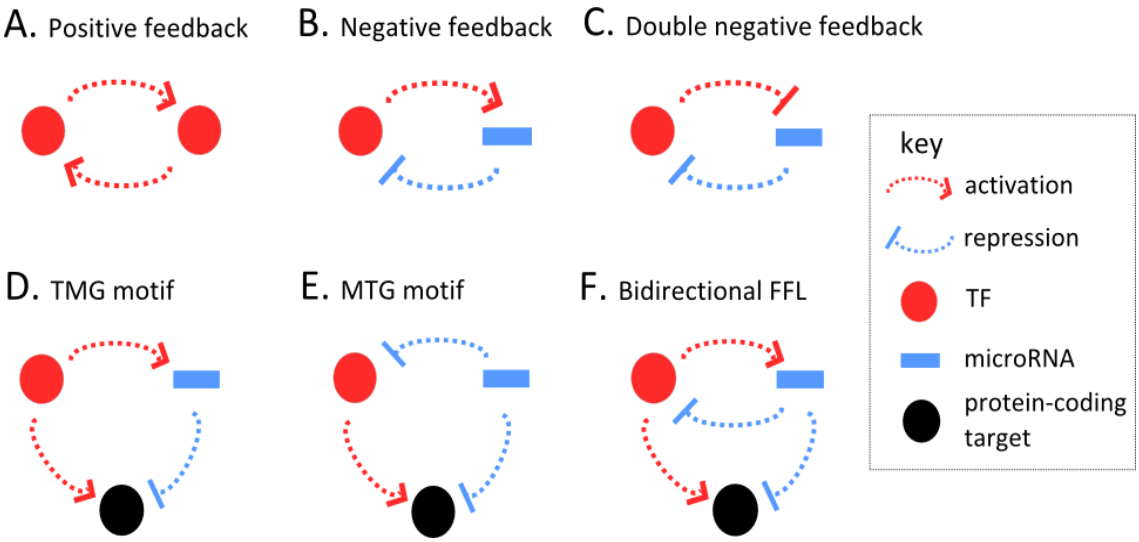


Figure 4.1. Types of TF – microRNA containing regulatory motifs.

A pair of activating TFs can participate in positive feedback (**A**). Since microRNAs are repressive, they can only participate in negative, or double negative, feedback loops (**B** and **C**). Subfigures **D** – **F** display 3-member feedforward circuits consisting of an activating TF (T) and a repressive microRNA (M) with a common protein-coding target (G). The TF could also be a repressor. The protein-coding target represents both the gene and the expressed mRNA.

Currently, investigation of TF – microRNA IRNs relies upon both computational and experimental methods. In a laboratory setting, the connection between regulators and their targets can be examined in various ways. For example, by adjusting the concentration of a microRNA *in vivo*, subsequent changes to mRNA and protein-coding expression levels can be measured (Grimson *et al.* 2007). This technique has been used to improve microRNA target prediction, and to model dynamic features of the TF – microRNA IRN over a small set of experimentally tested microRNAs (Tu *et al.* 2009). An alternative approach is to map the

locations of regulator binding sites, without disturbing the levels of the regulator, but also without measuring the target response. For example, in a chromatin-immunoprecipitation study with DNA sequencing (ChIP-seq), a genome-wide TF binding profile is obtained by trapping copies of the TF *in situ* on the genome, and then using an antibody for the TF to pull out specifically the bound DNA fragments for sequencing and mapping to the reference genome (Johnson *et al.* 2007). It is only very recently that sufficient ChIP-seq datasets have become publically available to make this a viable option for analysis of the TF – microRNA IRN (2011). Many earlier studies have predicted TF binding site locations genome-wide using matches to common features of DNA sequences within known binding sites (Shalgi *et al.* 2007; Yu *et al.* 2008; Tu *et al.* 2009). These features are believed to reflect the binding preferences of the TF (Stormo 2000). While there has been some examination of regulatory networks derived from large collections of ChIP-seq studies (Cheng *et al.* 2011; Gerstein *et al.* 2012), especially of the purely transcriptional network, the majority of the earlier computational work on TF – microRNA IRNs has not yet been reanalysed and developed in light of the new datasets. The use of ChIP-seq data also permits us to consider transcriptional co-factors and chromatin-modifying proteins alongside the sequence-specific TFs. As in previous chapters, we term these collectively as transcriptional regulatory factors (TRFs).

In this study, we estimate the average impact of the microRNA-mediated regulatory layer upon an underlying transcriptional regulatory layer. To this end, we combine protein-coding and microRNA expression atlases with ChIP-seq datasets relating to 117 TRFs (mainly published by the ENCODE consortium (2011)), and with binding sites of microRNAs predicted using a variety of target prediction methods. We then estimate the relative contributions of transcriptional regulatory factors and microRNAs globally, and within an array of human tissue samples, to mRNA expression levels. We examine a wide class of patterns of linkage between the two kinds of regulator and their shared or distinct target sets, and consider how these patterns relate to target gene activity.

4.2 Methods

Datasets

Coordinates of human protein-coding genes were obtained from Biomart (Ensembl v.65) (Kinsella *et al.* 2011). Human microRNA gene coordinates and mature sequences were obtained from miRBase (v.18) (Griffiths-Jones *et al.* 2006). Genome-wide ChIP-seq datasets from the ENCODE consortium (Latchman 1990; 2011) for 117 human transcription regulatory factors (TRFs) in a variety of cell lines were downloaded from the UCSC table browser (Karolchik *et al.* 2003; Karolchik *et al.* 2004). Pairs of replicate datasets under the same experimental conditions were merged by taking the intersections of regions within ChIP-seq peaks from each replicate. Following the classification from (2011), the final compilation of TRFs consists of 13 general TFs, 93 sequence-specific TFs, and 11 that are primarily concerned with covalent regulation of chromatin (Supplementary Table S1.1). Regulatory signs of TRFs were derived from the Gene Ontology (Ashburner *et al.* 2000), and from an extensive literature survey via PubMed.

Mapping transcription factor ChIP-seq datasets to protein-coding gene targets

A gene was annotated as a potential target of a TRF if the midpoint of at least one ChIP-seq peak for the TRF (from any given cell line) was found within the *cis*-regulatory region. We used a range of definitions of *cis*-regulatory regions around the 5'-TSS of each protein-coding gene, from a narrow window 250 nt either side, up to an interval 1 order of magnitude larger, from 3.5 kb upstream to 1.5 kb downstream of the TSS. If a pair of Ensembl gene IDs corresponded to the same or nearby regions, TRF peaks were assigned ambiguously to both. ChIP-seq peak scores calculated using the MACS algorithm and converted to a standard scale from 0 to 1000 were also available for download for the majority of studies (Zhang *et al.* 2008). We explored a range of peak score and peak-TSS distance thresholds. The key properties of the network do not depend upon these parameters.

Mapping transcription factor ChIP-seq datasets to microRNA gene targets

MicroRNA genes were annotated as sharing a common primary transcript whenever the distance between gene coordinates was less than or equal to a parameter L. Varying L between 5 kb and 20 kb affects gene cluster memberships for only 11% of human microRNA genes, so the parameter L = 5 kb was chosen as standard. MicroRNA genes or gene clusters were then divided into two classes: (a) Intragenic and lying in the same sense to the host gene,

assumed for simplicity to be transcribed from a shared primary transcript with the host genes (Baskerville and Bartel 2005); (b) All other microRNAs (i.e. intergenic and intragenic but antisense to a protein-coding gene), assumed to be transcribed from their own primary transcripts. Relationships between microRNAs and other classes of noncoding RNA genes that sometimes host these were not considered (Griffiths-Jones 2007). There was only one instance of a microRNA cluster including a mixture of intronic and non-intronic microRNAs (mir-658 and mir-659), and we defined this cluster using the 5'-member mir-659, which lies outside of a protein-coding region. For intronic microRNAs, we mapped all TRF peaks found in the host promoter region. For autonomously transcribed microRNAs, we mapped all TRF peaks within fixed distances upstream of the 5'-most member of a gene cluster to each microRNA within the cluster. We used a fixed distance of 10 kb as in (Shalgi *et al.* 2007). TRFs mapped to microRNA genes were assigned to microRNA mature sequences, from both arms of the pre-miRNA hairpin, using the mappings between microRNA genes and mature sequences from miRBase v.18. Due to neighbouring or overlapping protein-coding and opposite sense microRNA primary transcripts, a single ChIP-seq peak may be ambiguously assigned to more than one type of gene.

Mapping mature microRNA sequences to target mRNAs

MicroRNA target sets were predicted within the longest 3'-UTRs of protein-coding genes in Ensembl (v.65) using the miRanda algorithm (v.3.3a) (John *et al.* 2004; Kinsella *et al.* 2011). The core miRanda algorithm utilises microRNA sequence matching in target mRNA sequences, with some tolerance for G:U base pairing, together with thermodynamic stability modelling of the predicted microRNA:mRNA duplex. We examined a range of miRanda target site score thresholds (between 130 and 160 in intervals of 5) in order to build a range of microRNA – target gene networks with different edge density. In addition, we predicted target sets by searching in 3'-UTRs of mRNAs for exact matches to microRNA seed sequences of length 7 and 8 nucleotides similar to the TargetScan algorithm (Friedman *et al.* 2009). Two different types of 7mer seed were used, each containing a 6mer seed from the 2nd to the 7th microRNA 5' end nucleotides, but distinguished by either the 1st or the 8th base pair. These are termed 7mer-A1 and 7mer-m8 seeds, and are defined either by an adenine residue at position 1 of the seed sequence, or by Waston-Crick base pairing between the seed match and the 8th nucleotide of the seed sequence (Grimson *et al.* 2007). We did not impose any scoring filters for predicted effectiveness in repressing mRNA levels derived from microRNA perturbation experiments (such as the mirSVR score, often used in conjunction with miRanda targets) (Betel, Wilson *et*

al. 2008). This avoids a circular analysis of the global impact of microRNAs on target expression level.

Gene expression patterns

Protein-coding gene expression measurements were obtained from the Novartis human microarray-generated atlas across 79 human tissue samples (Su *et al.* 2004). Probe IDs for the Affymetrix U133A and GNF1H chips used were mapped to Ensembl gene identifiers, with probe annotation files from BioGPS and from http://www.affymetrix.com/, and gene symbols from the HUGO gene names consortium (http://www.genenames.org/). Where multiple probes map to a single gene, their expression values were averaged. MicroRNA expression measurements from 172 libraries derived from human tissues and cell lines were also downloaded (Landgraf *et al.* 2007). To match these measurements to microRNA mature sequences listed in miRBase (v.18), we used the mature microRNA nucleotide sequences themselves, since these are more stable than microRNA names. We then manually matched samples between the two atlases, identifying 17 unambiguous correspondences between healthy tissues, together with an imprecise match between 'whole brain' from the protein-coding expression atlas, and the union of healthy brain tissues represented in the microRNA expression atlas (Supplementary Table S2.1). TRF expression levels were extracted from the Novartis atlas (Su *et al.* 2004), using the Ensembl IDs for all the TRFs included in our study. For non-parametric permutation tests on mRNA expression and related variables, we fully permuted labels randomly, repeated simulations a minimum of 10,000 times, and counted the proportion of simulated datasets showing a trend at least as strong as was found in the real network. This provides an empirically estimated p-value for significance of the trend.

Integrated regulatory network of TRFs and microRNAs

The complete collection of target relationships between 102 TRFs, 1907 microRNA mature sequences, and 19427 protein-coding genes was represented by matrices. The $(i, j)^{th}$ element of such a matrix is set to 1 if a potential target relationship was identified between regulator i and target j, and to 0 otherwise. We are particularly interested in recurrent patterns within the network, such as feedforward loops (FFL; see Figure 4.1.D-F). The number of FFLs of each type can be counted by checking whether the conditions are satisfied for all possible triples (TRF, microRNA, protein-coding gene). To assess the significance of enrichment of any such pattern, its frequency is compared between the original target relationship matrices, and ensembles of random matrices, shuffled subject to biologically realistic constraints on network structure. In particular, degrees of all the regulators and targets are held constant, using an edge-swapping

algorithm as in (Shalgi *et al.* 2007). The shuffling process is continued until the proportion of edges in common with the original matrix reduces no further, assessed by the coefficient of variation of this proportion between successive sets of 10 batches of 1000 shuffles reaching a stable minimum. Frequencies of network patterns in matrices shuffled in this way are approximately normal across a wide range of biological networks (Shen-Orr, Milo *et al.* 2002; Shalgi *et al.* 2007). Normality for simulated pattern frequencies on this particular network was checked using the Shapiro-Wilk test. We then calculated a z-score using the pattern count in the original matrices to the distribution of counts in shuffled matrices. The z-score reflects whether the original matrix is significantly enriched ($z >> 0$) or depleted ($z << 0$) in the pattern. Z-scores were then transformed to (upper tail) p-values by subtracting from 1 the cumulative density function of the standard normal distribution evaluated at z. Patterns which are significantly enriched ($p < 0.05$) are termed network motifs.

Permutation tests

In figures with a moving average of one quantity plotted against the percentiles of the ranked distribution of another, we often obtained relationships with a linear form. The strength of the linearity was represented by the Pearson Correlation Coefficient (PCC) between the two vectors of plotted values. Because percentiles are equally spaced, and therefore clearly not normal, then standard tables of critical values cannot be used to measure significance of the PCC. Instead, we randomly shuffled the vectors and recalculated simulated PCC values repeating up to 100000 times. The distribution of simulated PCC values was not normal (by Shapiro-Wilk test), so we did not convert them to a standard z-score. Instead, the empirically estimated non-parametric p-value is the proportion of simulations having a PCC value at least as far from zero as the real value.

Variation in gene expression across tissues

We examined two different measures of variation across tissues:

(1) Entropy of gene expression, defined as in (35), where E_g[tissue t] is the expression level of gene g within tissue t averaged across microarray probe sets:

$S = \log_2(\text{number of tissues}) - \Sigma_{\text{tissues}} E_g[\text{tissue t}] * \log_2(E_g[\text{tissue t}])$

This score varies between 0 for a gene that is expressed in just one tissue, and 1 for a gene that is expressed uniformly across all tissues.

(2) Standard deviation of gene expression normalized across tissues.

The two scores give nearly identical results here, so we report results for entropy of gene expression.

Gene Ontology Analysis

Human genes were mapped to terms in the Gene Ontology (GO) hierarchy using the associations provided by the GO consortium, retrieved on 02/09/2011 (Ashburner *et al.* 2000). The full GO hierarchy of all parent-child relationships between terms (OBO v.1.2) was parsed to collect genes sets with a common ancestral GO term.

4.3 Results

4.3.1 Construction of an integrated regulatory network

We began by comparing the positions of ChIP-seq peaks for 117 human TRFs to the 5'-most transcription start sites (5'-TSSs) of 19427 protein-coding genes from Ensembl (Kinsella *et al.* 2011). Summing over TRFs, we find that ChIP-seq peak density reaches a maximum immediately around the TSS, for general and sequence-specific TFs alike, and for regulators that covalently modify histone proteins (Figure 4.2A). The specificity of different TRFs for TSS-proximal regions can vary greatly. For example, chromatin-modifying factors, including 4 members of the SMARCC family, are spread more diffusely than the other classes over 5' ends of genes (Figure 4.2A); the density distribution for general TFs lies further downstream than that of sequence-specific TFs ($p < 10^{-15}$ by Kolmogorov-Smirnov (K-S) test); likewise, bZIP TFs such as FOS are bound significantly further downstream than ETS domain TFs such as GABP (Supplementary Figure S4.1). There is a high density of ChIP-derived regulatory marks lying within protein-coding genes. This could reflect transcription elongation-associated regulators (e.g. RNA polII, or RDBP), factors bound to downstream regulatory regions within a gene, and distal regulators in complexes with proximal regulators (Johnson *et al.* 2007). Thus, it is likely that the set of TRF-target relationships detected is inclusive of a variety of regulatory mechanisms.

The degree distribution of numbers of experimental TRFs per gene is bimodal, with peaks at 0 and at 28-29 transcriptional regulators per gene (Figure 4.2B). The same bimodal pattern is retrieved for TRF datasets drawn from individual cell types, and becomes more pronounced for wider *cis*-regulatory intervals and at more stringent TRF peak quality scores (Supplementary Figure S4.2). When predicted from TF binding motifs in DNA sequences, the TF degree distribution follows either a negative exponential, or negative power law, functional form (Potapov, Voss *et al.* 2005; Balaji, Babu *et al.* 2006). This is in contrast to the bell-shaped Poisson distribution found within a large random network (Albert and Barabasi 2002). The experimentally determined TRF degree distribution shown here resembles a compound of both distribution types. Sources of discrepancy between TRF degree distribution predicted from ChIP-seq data and the TF degree distribution predicted from the binding motifs of sequence specific factors might include: (i) variations in the accessibility of TF motifs *in vivo* (ii) non-sequence specific binding interactions, for example, between ChIP-seq sampled cofactors and DNA-binding TFs, or (iii) experimental noise.

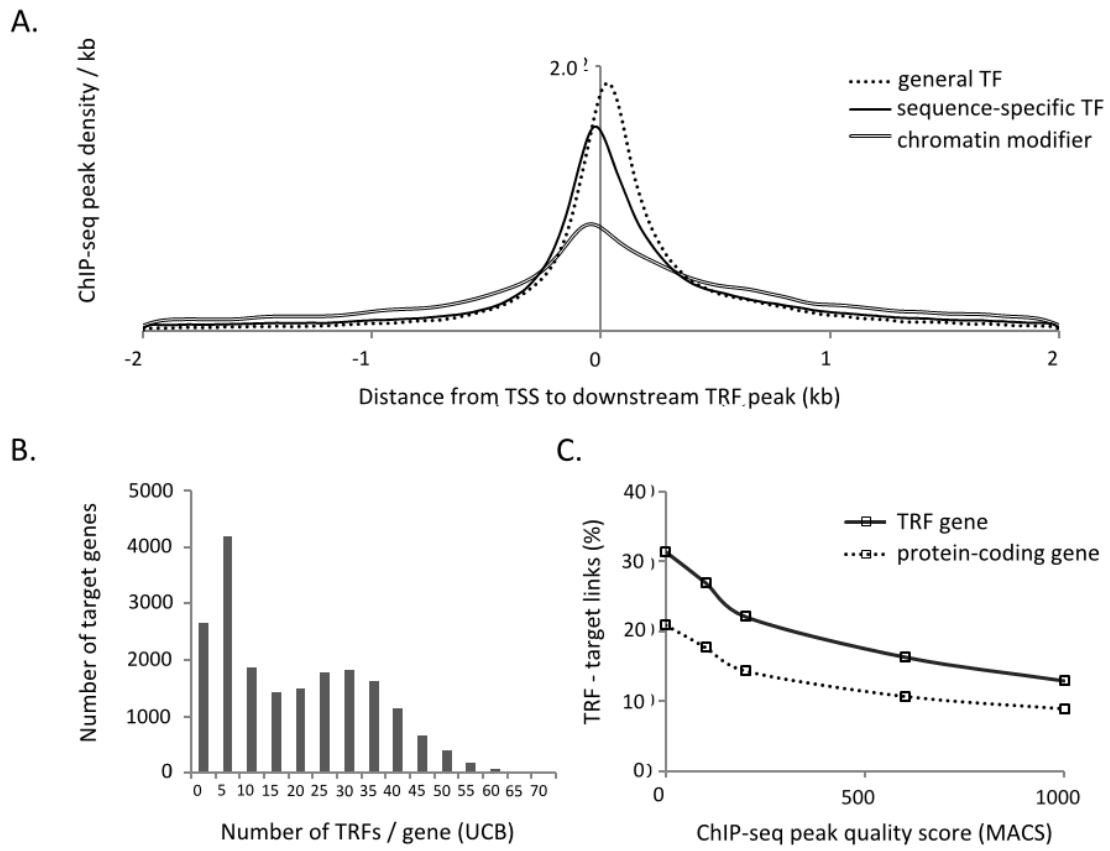


Figure 4.2. Distributions of TRFs mapping to predicted target genes.

A. Distributions of TRF ChIP-seq peaks around 5'-TSSs of protein-coding genes. Distributions for general TFs, sequence-specific (SS) TFs, and chromatin-modifying factors, were calculated in R using the density function, with Gaussian kernels, and optimal bandwidths inferred from data.

B. Distribution of TRF degree per target protein-coding gene. TRF degree was defined as the number of distinct TRFs detected within *cis*-regulatory regions added up across cell types sampled by ChIP. The figure displays results using all ChIP-seq peaks lying within 1.5 kb upstream and 0.5 kb downstream of the set of 5'-TSSs of all protein-coding genes. Note all classes have width 5 except the leftmost class for n(TRFs) = 0.

C. Comparison of TRF and protein-coding target set percentages as MACS peak score is varied

The mean degree is significantly greater than that expected if TRFs are randomly connected to genes ($p < 10^{-15}$ K-S test). Indeed, the target sets of TRFs are highly interdependent: A randomly sampled pair of TRFs overlaps in their protein-coding gene target sets at 1.86 times the expected level, with over 80% of overlaps being significant at $\alpha = 10^{-15}$ (Using the hypergeometric test, with all 19427 genes as background). Using annotations of TRF regulatory signs derived from literature (Supplementary Table S1.1), pairs of activators co-associate (at 1.91 times the expected level) more than pairs of repressors (at 1.71 times the expected level). This might be because repressors, such as CTCF and SIN3A, often induce a local chromatin

state that is silenced, so less favourable to binding by other factors (Cowley, Iritani *et al.* 2005; Kim *et al.* 2007). Unless otherwise stated, results are reported using all TRF peaks (regardless of score) and a window from 1.5 kb upstream to 500 nt downstream of 5'-TSSs. In Figure 4.2C we confirm as elsewhere that TRFs target a substantially higher fraction of TRF genes than protein-coding genes in general; and this result is true independently of the ChIP-seq peak quality score threshold calculated by the MACS algorithm (Zhang *et al.* 2008; Gerstein *et al.* 2012).

We next mapped the same collection of 117 human TRFs to 1523 microRNA genes (methods). Primary transcripts of intragenic microRNAs were assumed to be transcribed from a shared promoter region with their host genes (Baskerville and Bartel 2005). Upstream of intergenic and antisense intragenic microRNA clusters, the majority of TRF density is located within 5 kb of the cluster, with less than 1/3 as many further binding sites detected in the next 5 kb upstream. Extending further upstream, TRF density continues to decline. The error from approximating *cis*-regulatory regions of intergenic or antisense intragenic microRNA clusters as a fixed 10 kb window upstream, as in (Shalgi *et al.* 2007), is therefore likely to be small, and we adopt the same simplification here.

We next predicted target sites of 1919 microRNA mature sequences (including both arms of most pre-miRNA hairpins) in the longest 3'-UTRs of all protein-coding genes, using the miRanda algorithm (v.3.0) (John *et al.* 2004) and an exact microRNA seed matching algorithm similar to TargetScan (Friedman *et al.* 2009). Unless otherwise stated, results are reported for the minimum microRNA target site score of 150 from the miRanda algorithm, and also for each of the possible seed types or their union. Similar to the patterns of regulation by TRFs, we find higher rates of targeting by microRNA sequences of TRF genes than of protein-coding genes in general. This remains true across all settings examined for both the miRanda and seed-matching algorithms. It was previously suggested that this difference could be attributable to differing lengths of 3'-UTRs between mRNAs encoding TRFs and other proteins (Shalgi *et al.* 2007). Indeed, the 3'-UTRs of the 117 TRFs used in our study are longer than average (p = 0058, K-S test). We controlled for this by dividing total predicted microRNA target sites by 3'-UTR length to obtain a measure of target site density per unit nucleotide. We confirm that the density of microRNA binding sites is higher within the 3'-UTRs of TRF genes, regardless of microRNA target prediction method or settings (e.g. for exact 8mer seeds: p = 0.0052 by K-S test). This mirrors the enrichment of TRF regulators in the *cis*-regulatory regions of both TRF genes, and the host genes of intragenic microRNAs.

4.3.2 Relationships between TRF- and microRNA-mediated regulation and target gene expression

We next examined relationships between TRF and microRNA regulators, and the expression of protein-coding genes across 79 human tissues from the Novartis expression atlas (Su *et al.* 2004). Principal component analysis of these 79 protein-coding expression datasets showed a significant overlap in expression patterns between tissues, with the first 4 principal components accounting for 69.2% of the total variance. This indicates that gene expression within individual tissues is similar to the mean expression level across these tissues. We ask whether variation in numbers of TRF or microRNA binding sites is linked to shifts in both mean and tissue-specific target gene expression level.

TRFs: A clear positive relationship is found between total number of TRF regulators and mean protein-coding gene expression (Figure 4.3A). The relationship between number of TRF regulators and microRNA expression level, calculated in the same way, is very similar (Supplementary Figure S4.3). This is consistent with analysis of other collections or TRFs (Yan *et al.* 2013), and with the regulatory signs of TRFs derived from literature (Supplementary Table S1.1), since activating TRFs (n = 44) outnumber repressive TRFs (n = 18). Target gene expression is significantly greater than non-target expression for 110 / 117 TRFs surveyed ($p < 0.01$, with the majority significant at $p < 10^{-15}$, by z-transformed Mann-Whitney tests, Benjamini-Hochberg corrected for multiple tests (Supplementary Table S4.4) (Benjamini *et al.* 2001). The greatest target over non-target expression ratios were found for negative elongation factor RDBP (4.96), which is evidently associated with actively transcribed genes, and for activators ZZZ3 (3.67), TAF7 (2.99) and POL2 (2.81). Only 3 TRFs have target genes with lower expression level than non-target genes, and significantly only for the developmental repressor Suz12 (target : non-target expression ratio = 0.38, $p < 10^{-15}$, z-transformed Mann-Whitney test. See also Supplementary Figure S4.5).

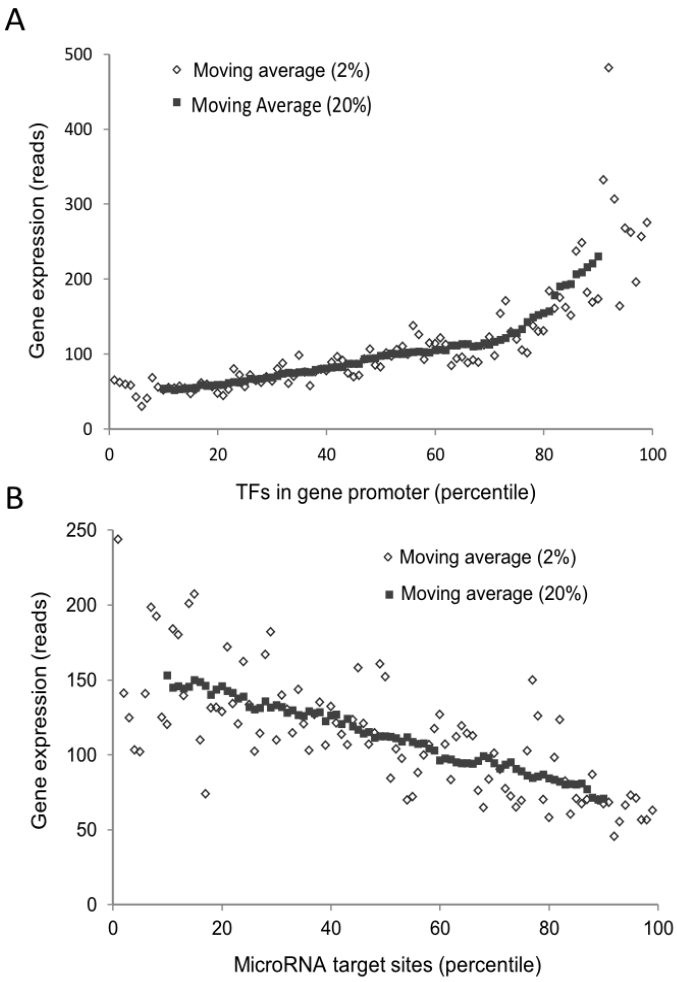


Figure 4.3. Relationships between gene expression and numbers of regulators.

Mean target gene expression averaged over windows representing 2% and 20% of the genes ranked by **A.** measured TRFs per gene **B.** predicted microRNA binding sites per gene. For the TRF distribution, all ChIP-seq peaks were used, with other parameters as standard (methods). For the microRNA binding site distribution, miRanda was used to predict binding sites, with minimum score per site S ≥ 150. Other miRanda score thresholds were checked, and also seeds of each type (Supplementary Figures S4.6 and S4.7).

Co-association of collections of TRFs over active common target sets helps to explain why the majority of repressive TRFs, such as NRSF (neuron-restrictive silencing factor), have targets expressed at relatively high levels, since the repressors may be co-associated with many activating TRFs. It may therefore not be advisable to annotate regulator signs by comparing target to non-target expression levels, as is sometimes attempted (Gerstein *et al.* 2012). Instead, it might be appropriate to ask whether there is a *relative* difference between target and non-target expression ratios for sets of repressors and activators. We tested this by ranking the mean target to non-target expression level, and assessing difference between ranks for activating (n = 44) and repressive (n = 18) TRFs, and those TRFs with alternating regulatory sign (n = 49). The ranks of exclusively repressive TRFs are significantly lower than

the rest (p = 0.0071, Mann-Whitney test); while the ranks of exclusively activating TRFs are significantly greater (p = 0.0042). This provides support for our collection of TRF sign annotations derived from literature. The test was repeated using the annotations 'positive regulation of transcription' (n = 33) and 'negative regulation of transcription' (n = 12) from the Gene Ontology (34), with a further 19 TRFs annotated as both. These GO-derived annotations of regulatory signs failed to separate the sets of TRFs by their target to non-target expression ratios (Table 4.1). Although the two collections of annotations are fairly consistent, the regulatory sign of a significantly higher proportion of TRFs was unclassified by GO, compared to our annotations (Table 4.2).

Table 4.1. Comparison of target/non-target expression ratios for TRF activators, repressors, and those of variable regulatory sign.

	Literature derived annotations		GO derived annotations	
Regulatory sign	**Z**	**p**	**Z**	**p**
Activators	-2.633	0.0044	-0.418	0.3372
Alternating sign	0.863	0.1943	-0.341	0.3666
Repressors	2.342	0.0096	1.17	0.1321

TRFs were ranked by the ratio of their target to non-target expression, and then separated into activators, those of alternating sign, and repressors, using both literature derived and GO-derived annotations. Mann-Whitney U-statistics were then calculated to compare ranks for each of the three classes of TRFs compared to all others. Due to the large sample size, to calculate significance the U statistic was converted to a normal z-score using mean(U) = 0.5 * n(targets) * n(non-targets) and st.dev(U) = sqrt((n(targets) * n(non-targets) * (n(genes) + 1) / 12). P-values were then calculated using the cumulative normal distribution function.

Table 4.2. Overlaps between TF regulatory signs from literature and the Gene Ontology

			Annotations from literature			
			Positive	**Both**	**Negative**	**Unclassified**
		Number of TFs (n = 117)	44	49	18	6
Annotations from GO	**Positive**	33	19	12	0	2
	Both	19	6	11	2	0
	Negative	12	0	3	8	1
	Unclassified	53	19	23	8	3

Annotations from literature were taken from the Genecards database, and from PubMed. See Supplementary Table S1.1 for full details. Annotations from the Gene Ontology (Ashburner *et al.* 2000) were taken from TRF genes that mapped to either 'positive regulator of transcription', 'negative regulator of transcription', both of these, or neither (unclassified). Annotations from literature and from GO never resulted in a reversal in sign; however, comparing the two sets of annotations, many TRFs moved between either the categories 'positive' and 'both', or the categories 'negative' and 'both'.

MicroRNAs: A strong negative linear relationship is found between mean gene expression across tissues and number of predicted microRNA binding sites per gene (using 20% moving averages, Pearson's correlation $r_p = 0.9925$, $p < 10^{-4}$ by permutation test, see methods). With smaller window sizes, the negative linear relationship is still significant but noisier (Pearson's correlation $r_p = 0.7315$, $p < 10^{-4}$). The best fits to linear regression functions are found with the settings $S \geq 150$ for the miRanda algorithm, and using seeds of type 7mer-m8 (Table 4.3). The relationship is perhaps surprising since it is derived from counting potential microRNA binding sites genome-wide without considering expression levels of microRNAs. We suggest that the relationship reflects the repressive impact due to equilibrium levels of microRNAs, averaged through time and across cellular contexts and tissues, merely a fraction of which are engaged at true target sites. Assuming that the number of functional microRNA binding interactions is proportional to total predicted binding sites, then the relationship clearly implies that the total repression of targets is cumulative with number of regulating microRNAs. Analogous relationships have been established experimentally in specific cases (Doensch and Sharp 2004; Grimson *et al.* 2007), and the significance of microRNA target sites is greatly elevated in the context of other target sites, suggesting cooperative functions (Enright *et al.* 2003).

Table 4.3. Correlations between number of microRNA target sites and target gene expression.

Minimum miRanda score S	PCC(microRNA target sites, Expression)	Seed type	PCC(microRNA target sites, Expression)
130	0.666	Eightmer	0.784
140	0.685	Seven-A1	0.755
145	0.679	Seven-m8	0.817
150	0.732	All	0.770
155	0.705	[MiRanda S ≥150] ∪ [Seven-m8]	0.824
160	0.674	[MiRanda S ≥150] ∩ [Seven-m8]	0.678

All genes in the Su *et al*. protein-coding gene expression atlas were ranked in order of number of predicted microRNA binding sites within their longest 3'-UTR. This was repeated for 6 settings of the miRanda algorithm, and for four basic microRNA seed types (8mers, 7-A1, 7-m8, and the union of these = all). Moving average (MA) gene expression was calculated taking 2% of the ranked gene list at a time, from the 1st to the 99th percentile. PCC (Pearson Correlation Coefficient) values were then calculated using all pairs of percentiles and MA expression values. PCC in this context is not suitable for significance testing due to the equidistant spacing of the percentile set (which is therefore clearly not normal). The final two rows on the right-hand side show results when predicted targets from miRanda with score $S \geq 150$ and exact matches to seeds of type 7-m8 are combined, either as a union ($\cup$) or an intersection ($\cap$)

The gradient of the best fit regression line might reflect the mean strength of the global repressive impact on mRNA concentrations due to microRNAs. Across 18 tissues in a human

microRNA expression atlas (Supplementary Table S2.1) (Landgraf *et al.* 2007), the gradient is somewhat steeper in tissues with higher levels of circulating microRNAs, albeit weakly so ($p < 0.1$; Figure 4.4). Considering individual microRNAs, the mean expression of target sets of 1237/1919 (64.5%) mature microRNA sequences is significantly lower than for non-target sets, and significantly higher for none. Averaged across all microRNAs, the target to non-target expression ratio is 0.74 (26% lower expression). This contrasts with a mean ratio of 1.86 for TRFs (86% higher expression).

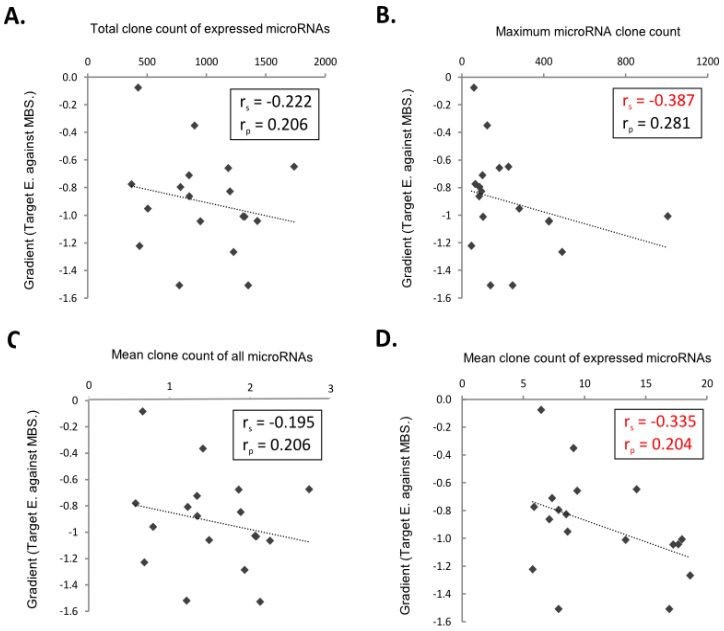


Figure 4.4. Tissues with more microRNa expression show greater target gene repression

We matched tissue samples between protein-coding and microRNA expression atlases, identifying 17 samples shared between each, and in addition matching 'whole brain' from the protein atlas with 3 healthy brain tissues from the microRNA atlas. The change in target gene expression with microRNA indegree (number of microRNAs per gene) was measured as the gradient of best fit lines between target gene expression and percentiles of genes ranked by number of microRNA regulators. This gradient was calculated separately in all 18 tissues identified as shared between the microRNA and mRNA expression atlases. This was then plotted against a variety of measures of microRNA expression level:

A. Total clone count = sum of all individual microRNA clone counts for a tissue
B. Maximum clone count = count for the most highly expressed microRNA in each tissue
C. Mean clone count of all microRNAs = mean averaged across every microRNA (including 85% that are unexpressed, on average).
D. Mean clone count of expressed microRNAs = mean averaged across the 15% of microRNAs that were expressed with clone count ≥ 1 in each tissue.

Correlation coefficients between repression gradient and microRNA expression are shown in the legends: (Pearson's: r_p and Spearman's: r_s are shown in the legends).

4.3.3 Bidirectional targeting between a TRF and microRNA pair is linked to TRF regulatory sign

As in many earlier studies, network motifs were defined as patterns of network connections which occur significantly more frequently than expected given the degree distributions of all the regulators and targets (Shen-Orr *et al.* 2002). When the motif consists of a pair of regulators targeting one another, significance of the frequency of bidirectional links can be calculated using the hypergeometric distribution. This test was to measure the significance of numbers of pairs of TRFs and microRNAs targeting one another, (TRF ↔ microRNA), given the underlying numbers of (TRF → microRNA) and (microRNA → TRF) links (Table 4.4). The null model is that links in the direction (TRF → microRNA) are independent of links in the direction (microRNA → TRF) so that instances of bidirectional targeting (TRF ↔ microRNA) occur randomly.

Table 4.4. Bidirectional targeting between transcriptional regulators and microRNAs.

A. miRanda algorithm

	miRanda minimum score S			
	130	140	150	160
n(TRF, microRNA) pairs	224523	224523	224523	224523
n(TRF → microRNA) links	45191	45191	45191	45191
n(microRNA → TRF) links	43311	23900	10936	3165
n(TRF ↔ microRNA) links	10172	5725	2658	748
p-value	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$	6.5×10^{-7}

B. Exact matches to seed sequences

	Seed type			
	All	7mer-A1	7mer-m8	8mer
n(TRF, microRNA) pairs	224523	224523	224523	224523
n(TRF → microRNA) links	45191	45191	45191	45191
n(microRNA → TRF) links	51686	24952	26767	10224
n(TRF ↔ microRNA) links	10494	4945	5494	2072
p-value	0.129	0.904	0.043	0.364

Hypergeometric p-values were calculated to reflect P(n(TRF ↔ microRNA) links ≥ observed) given the network counts n(TRF, microRNA) pairs , n(TRF → microRNA) links and n(microRNA → TRF) links. Independence was rejected if the hypergeometric p-value was less than 0.05.

We find that the frequency of bidirectional connections (TRF ↔ microRNA) between 117 TRFs and 1919 microRNA mature sequences is much higher than expected regardless of minimum miRanda microRNA target site score (Table 4.4A). The rate of bidirectional links is most significant when restricted to microRNA mature sequences derived from intronic loci, but remains highly significant when restricted to sequences derived from intergenic regions (Supplementary Table S4.8A). This could be because TRF regulators are assigned with greater accuracy to microRNAs derived from a host gene promoter, than to intergenic or antisense intragenic microRNAs. When searching for exact microRNA seed matches of different types, much lower rates of bidirectional linkage are found between TRFs and microRNAs, though still significant at the conventional 5% level for 7mer-m8 seeds ($p = 0.043$).

Bidirectional targeting between regulators can reflect either positive feedback (if both regulators are activators), negative feedback (if one is an activator and the other a repressor), or double negative feedback (if both are repressors) (See Figure 4.1.A-C). In addition, if the regulators switch between acting as activators or repressors depending upon cellular context, then the interaction type might vary. For microRNA repressors with a bidirectionally linked TRF partner, we examined whether any global preference exists between negative feedback, double negative feedback, and variable interaction types. We applied the same classical test, based on the hypergeometric distribution of numbers of bidirectional links found, using miRanda-predicted microRNA target sites, and within subnetworks restricted to TRFs with particular regulatory signs. The rates of bidirectional linkage from TRFs to microRNAs were found to be marginally significant for 44 activating TRFs ($p = 0.072$), but highly significant for 18 repressive TRFs ($p = 6.1 \times 10^{-6}$) and for the remaining 49 TRFs that have variable regulatory signs ($p < 10^{-15}$). We ask whether these contrasts between the classes are sufficient to indicate a preference in bidirectional regulation towards interaction with repressive or variable-sign TRFs. The frequencies of bidirectional links within each class in the real network were compared to those within an ensemble of 10000 simulated networks in which the 117 TRFs were assigned randomly to the three classes, keeping class sizes fixed. We find that bidirectional links, TRF ↔ microRNA, are indeed significantly associated with variable-sign TRFs ($p = 0.0326$), compared with expected rates for repressive TRFs ($p \geq 0.4319$), while they are depleted over the activating TRF set ($p = 0.0239$). Thus, linkage of microRNAs to TRFs with variable regulatory sign, or to repressor TRFs (Figure 4.1C), are more prevalent than simple negative feedback circuits (Figure 4.1B).

4.3.4 Feedforward regulation by TRFs and microRNAs of common gene expression pathways

For motifs with more than 2 edges, classical tests become increasingly intractable, so the significance of enrichment of network patterns is tested by simulation. As in previous studies, the frequency of a given network pattern was compared between the real regulatory network and an ensemble of degree-preserving simulated networks. Frequencies of patterns of connections in degree-preserving simulated networks are approximately normal across a wide variety of biological networks (Shen-Orr *et al.* 2002; Shalgi *et al.* 2007). We calculated the statistical significance of frequencies of FFLs with either a TRF upstream of a microRNA (TMG), or a microRNA upstream of a TRF (Figure 4.1E), or with bidirectional targeting between the two, over common target genes (Figure 4.1F) (see methods). This was repeated across a range of *cis*-regulatory window sizes, score thresholds, and for both microRNA target prediction algorithms.

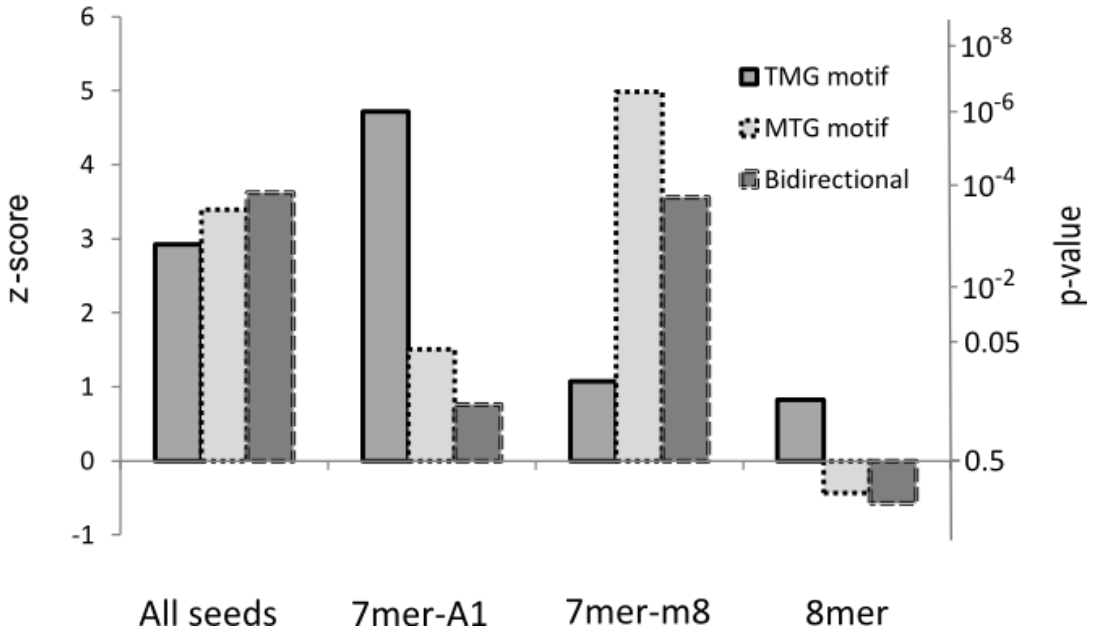


Figure 4.5. Significance of frequencies of TRF-microRNA feedforward loops.

Z-scores calculated from network pattern simulations are shown on the left-hand axis, with the corresponding standard normal p-value on the right-hand axis.

Significance of enrichment of motifs can be read from the right-hand axis (upper tail p-values). Results are shown across a range of miRanda score thresholds and microRNA seed types for 3 types of feedforward loop: TMG refers to the FFL with a TRF upstream of a microRNA, and both regulating a common target gene; MTG refers to the FFL with a microRNA upstream of a TRF, and both regulating a common target gene; Bidirectional refers to the FFL with TRF and microRNA regulating one another, and both regulating common targets.

Considering first the microRNA target predictions from miRanda, all three configurations of FFL are found to be significant regardless of score threshold for microRNA target sites ($p < 0.05$). Using the recommended minimum binding site score of S = 150 (Betel *et al.* 2008), the significance of enrichment ranges from $\approx 10^{-3}$ for the MTG motif, to $\approx 10^{-9}$ for the TMG motif, with the bidirectional motif intermediate. Significance of FFLs generally decreases with decreasing numbers of edges in the network, either through more stringent microRNA binding site predictions or TRF peak quality scores, or through increasing the size of the *cis*-regulatory region (Supplementary Figures S4.9 and S4.10). Similar trends were found using the seed-matching algorithm, though FFL significance was lower than for networks constructed using miRanda scores $S \geq 150$. All three configurations of FFL are detected as significant across a union of the three microRNA seed types (8mer, 7mer-A1 and 7mer-m8), and in some cases, across 7mer-A1 or 7mer-m8 seeds alone (Figure 4.5). When restricted to 8mer seeds, FFL frequencies were insignificantly different from random expectation. Thus, the global significance of a network pattern is sensitive to microRNA seed type.

4.3.5 Connectivity between TRFs and microRNAs defines gene expression level

We ask whether the distributions of FFLs across genes are linked to target gene expression. We first divided all pairs of TRFs and microRNAs into one of four cases depending on the patterns of regulation predicted between them (bidirectional, TRF $\rightarrow$ microRNA, microRNA $\rightarrow$ TRF, or neither). For each of the four patterns, we calculated within each of 18 human tissues (methods) the mean expression level of genes targeted by (A) both the TRF and the microRNA, as well as the mean expression level of genes targeted by (B) just the TRF, (C) just the microRNA, or (D) neither. The expression of genes targeted by neither regulator was taken as the background, and the other three cases (A) – (C) normalized to this. Figure 4.6 shows the cross-tissue average results for miRanda target predictions, with results for exact seed matches to microRNAs provided as Supplementary Figure S4.11. Results are consistent across each of 18 tissues matched between mRNA and microRNA atlases, and across a range of network settings (data not shown).

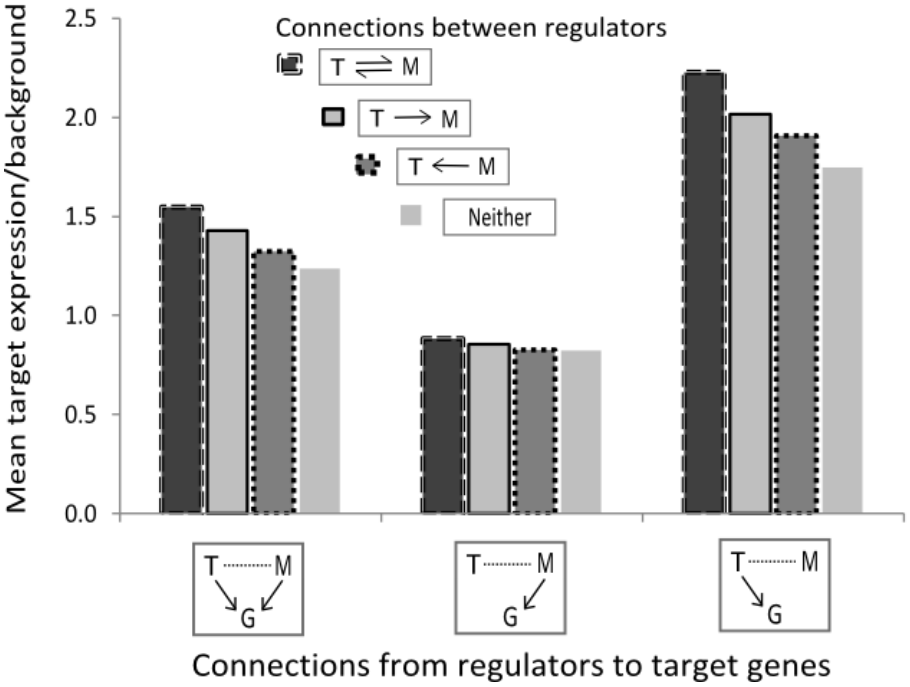


Figure 4.6. Gene expression levels across subsets of TRF-microRNA feedforward loops.

For each TRF-microRNA pair, the collection of all protein-coding genes was separated into four groups according to whether they are predicted as targets of both TRF and microRNA, only the TRF, only the microRNA, or are not targeted. Mean expression was then calculated within the three targeted groups (shown on the horizontal axis), and divided by the mean expression for all untargeted genes (as background). The complete collection of TRF-microRNA pairs was also divided into four classes according to the predicted regulatory connections between TRF and microRNA. Mean expression ratios (targeted/background) were then averaged across all the TRF-microRNA pairs in each class. MicroRNA target predictions were calculated using the mIRanda algorithm.

For typical TRF-microRNA pairs, the expression of genes targeted by both regulators is significantly greater than the expression of those targeted by neither regulator (Ratio E[G: (TRF + microRNA $\rightarrow$ G)]/E[G:(neither TRF nor microRNA $\rightarrow$ G)] $\approx$ 1.3 – 1.5: Mean $p < 10^{-3}$ by Mann-Whitney tests on TRF-microRNA pairs). We previously showed that target expression is related positively to total, and most individual, TRF binding sites, and negatively to microRNA binding sites. Thus the net activation linked with a typical TRF binding site outweighs the net repression linked with a typical microRNA binding site.

Secondly, genes targeted by TRFs are more likely to be highly expressed when the TRF regulates or is regulated by a microRNA. In particular, the greatest expression is found for genes falling within the target sets of bidirectionally regulating TRF-microRNA pairs, followed by those with a microRNA downstream of the TRF, with lowest expression for the targets of an unconnected TRF-microRNA pair. This is true in feedforward circuits, but remains true even when the microRNA is disconnected from the target gene (Figure 4.6). This suggests that the

feedforward circuit *per se* is not the essential feature defining high expression, but rather the connectedness of the TRF to a microRNA, which in turn relates to the TRF degree. The addition of a microRNA → gene link preserves the relationships but with lower average target expression. Very similar results were found using targets predicted by exact seed matches (Supplementary Figure S4.11).

4.3.6 Feedforward circuits are associated with stable mRNA expression levels

It has been argued that microRNAs can stabilise mRNA levels, reducing the impact of transcriptional noise on the proteome (Osella *et al.* 2011). We therefore ask whether the frequencies of TRF / microRNA feedforward circuits incident upon an mRNA are related to the stability of mRNA expression levels across tissues. To this end, we count the number of instances of patterns that contain each node in the real and in the ensemble of simulated networks, for every node and for each kind of FFL pattern. These counts were converted into a standard z-score reflecting enrichment or depletion of the pattern at each node in the network. We confirm that z-scores for individual genes are highly consistent across microRNA target prediction methods (Pearson's $r \geq 0.443$, $p < 10^{-15}$; See Supplementary Table S4.12 for details). Variation in mRNA expression across tissues was then measured in two ways, using either a quantity derived from the thermodynamic concept of entropy as in (Landgraf *et al.* 2007) (see methods), or a measure of spread, the standard deviation, as in (Lu and Clark 2012). The variation in gene expression across tissues was then compared directly with the network pattern z-score for the gene.

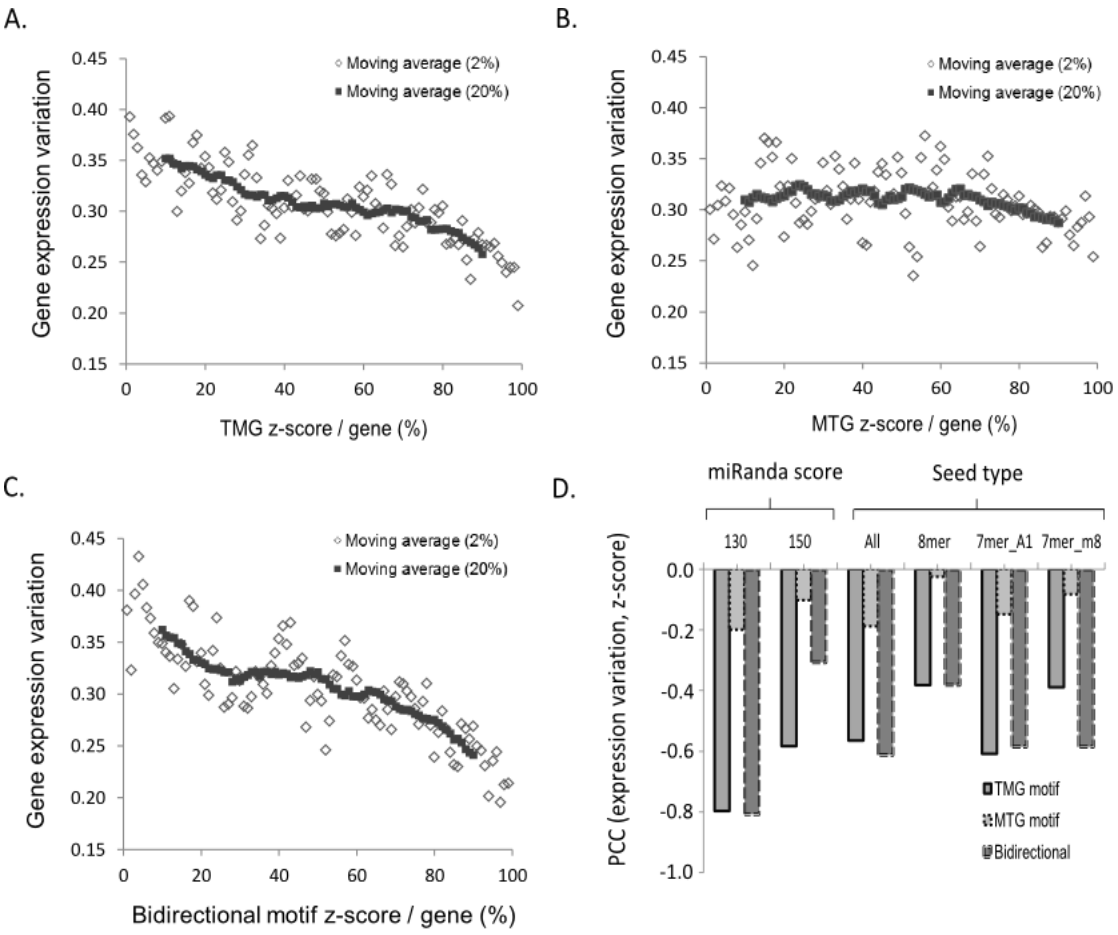


Figure 4.7. Relationships between feedforward loops and the stability of gene expression.

Motif z-scores were ranked for all genes for **A.** TMG, **B.** MTG, and **C.** bidirectional feedforward circuits, and calculated mean entropy of gene expression calculated across 2% and 20% moving averages of the gene ranking (methods). A, B and C show these results with the permissive miRanda score threshold S ≥ 130, which gives the greatest contrast between motifs. After noting the linearity specific to the TMG and bidirectional motifs, we calculated the PCC between the plotted entropy values and z-score rankings, over the 2% moving averages. This was repeated for the three motif types across a range of microRNA target prediction sets (**D**).

Gene expression entropy across tissues declines with the individual gene score for the TMG and bidirectional motif, but is broadly invariant with the distribution of MTG motifs (Figure 4.7). Gene expression entropy is also significantly correlated with gene expression standard deviation here(Spearman's $r_s = 0.947$), and therefore gene expression standard deviation gives almost identical results. Trends for motifs with the microRNA downstream of the TRF are clearly linear for the low specificity collection with minimum score S ≥ 130 ($r_s \approx -0.8$, $p < 10^{-4}$ by permutation test). As this score is raised from 130 to 150, the connection between mRNA stability and the bidirectional TRF ↔ microRNA regulatory patterns becomes weaker (Figure 4.7D). Using exact seed match targets gives similar results to miRanda (data not shown). Thus,

across all network settings considered, genes with the greatest enrichment for TMG and bidirectional TRF ↔ microRNA motifs have the most stable expression levels across tissues, but mRNA stability is significantly less connected with MTG motif significance. Which of the two motifs, TMG or bidirectional, has the tightest link to mRNA stability across tissues is dependent upon network parameters (Figure 4.7D). Furthermore, the variation in TRF expression across tissues is also greatest for the set of activator TRFs (n = 44) that are less likely to be bidirectionally connected to a microRNA. This provides novel evidence that the distribution of feedforward circuits with microRNAs downstream of TRFs and the uniformity of target mRNA levels across tissues are closely related.

4.3.7 Feedforward circuits are associated with core cellular processes

The prevailing tendency to calculate a global network motif significance score leaves open the possibility for a highly imbalanced distribution of the pattern across subnetworks defined by cellular function. We therefore asked whether network patterns are related to specific gene functions and pathways, such as regulatory pathways. To this end, variation in motif frequencies was compared with functional classes of genes, defined by terms in the Gene Ontology (Ashburner *et al.* 2000). After restricting to 1566 GO terms referring to at least 50 genes, the remaining GO terms were then ranked by mean FFL pattern z-scores for all the target genes referred to by the term, for each of the three FFL types (TMG, MTG and bidirectional). The resultant rankings are broadly consistent between target prediction methods and settings (Spearman's coefficient $r \geq 0.488$), so results are reported for miRanda only. The TMG and MTG GO term rankings significantly overlap with the bidirectional motif GO term ranking ($r \leq 0.622$, $p < 10^{-15}$), but less significantly with one another ($r = 0.135$, $p = 4.2 \times 10^{-8}$). For each GO term, we then calculated a p-value measuring how infrequently the mean of a set of normal z-scores lies further from 0 than the observed mean value. These p-values were adjusted using the Benjamini-Hochberg correction, in order to restrict the false discovery rate across 1566 GO terms (Benjamini *et al.* 2001).

Of the three FFL types, the bidirectional motif had the most functional classes with consistently above expectation z-scores (392 significant GO terms). Many of these terms are clearly repetitive, so they were arranged into broadly non-redundant groups, summarised in Figure 4.8. Enrichment for the bidirectional TRF – microRNA regulatory pattern is predicted across numerous aspects of cellular organization, from response to a wide variety of extracellular stimuli (response to cytokines, stress, hormones, virus, and Wnt signalling), through signal transduction cascades, transcription, and epigenetic regulation, to protein post-translational modifications, transport, and degradation. We also identify bidirectional TRF –

microRNA feedforward loops are enriched over a number of critical oncogenic pathways , which control the balance between apoptosis, proliferation and response of a cell to stressful conditions, including p53-signalling, Ras-mediated signalling, and ikB/NF-kB signalling (Cox and Der 2003; He, He *et al.* 2007; Gyrd-Hansen and Meier 2010). Consistent with this, cellular proliferation and apoptosis were also generally enriched, together with differentiation of some specific cell lineages (lymphocytes and neurons). Interestingly, many of the same processes (transcription, Wnt signalling, lymphocyte differentiation, neurogenesis, and cell growth) were identified as significantly enriched common targets of repressor TRF pairs in a regulatory loop with one another (Driskell, Oda *et al.* 2012). We conclude that TRF / microRNA mediated feedforward regulation is indeed prevalent throughout a wide variety of core cellular processes, including key cell signalling pathways.

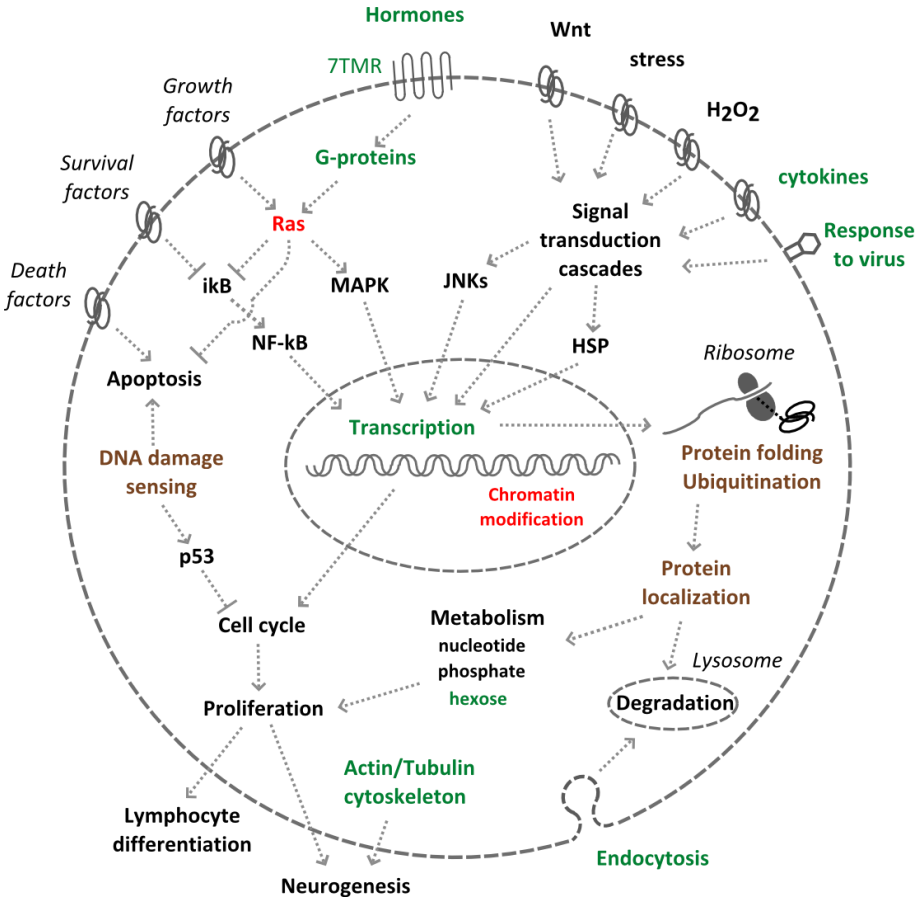


Figure 4.8. Cellular processes significantly enriched for the bidirectional FFL motif.

Cellular processes depicted summarise collections of similar GO terms significantly enriched in the bidirectional FFL of TRF and microRNA regulating one another, and both regulating common target genes. In addition, some processes were found to be significantly enriched in one or both of the unidirectional FFLs as well: TMG and bidirectional (brown), MTG and bidirectional (green); Both TMG and MTG and bidirectional (red). A few additional terms in italics were not significantly enriched in the bidirectional motif, but are included to clarify parts of the figure.

4.4 Discussion

In this study we have analysed integrated regulatory networks comprising links between TRFs, microRNAs, and their common target gene sets in human. Typically in integrated TRF – microRNA network analyses, the regulatory signs of TRFs have not been considered (Potapov *et al.* 2005; Shalgi *et al.* 2007; Yu *et al.* 2008). In other cases, signs have been inferred from target set expression levels (Tu *et al.* 2009; Gerstein *et al.* 2010). We found a non-automated, manually curated, approach to TRF regulatory sign annotation to be preferable to the use of either Gene Ontology terms, or inference from target expression level. In the latter case, many repressors are bound to highly expressed genes, though relatively not as highly expressed as for activators. We also found the novel result that the 2-member bidirectional regulatory loop (TRF $\leftrightarrow$ microRNA) is depleted for activator TRFs (Figure 4.1B) compared to those that are repressors or have variable regulatory sign (Figure 4.1C). Intuitively we can reason that TRFs with more complex regulatory functions, indicated by variable regulatory signs, are more likely to act as regulatory switches, and so are more likely to be interconnected with other regulators, such as microRNAs. In principle, differences in regulatory sign lead to fundamentally different regulator and target dynamics, e.g.(Hobert, 2006; Woods *et al*., 2007; Brabletz *et al*., 2006). The exploration of signed characteristics of regulatory networks is still in its infancy. As more experimental data becomes available, many more sign-specific properties of networks may become clear.

Many studies based upon computationally annotated TF binding sites have predicted global enrichment in thematic network patterns such as the TF/microRNA FFL, e.g. (Shalgi *et al.* 2007; Yu *et al.* 2008; Tu *et al*. 2009; Gerstein *et al*., 2011). We validated this global enrichment for FFLs containing the links (TRF $\leftrightarrow$ microRNA), (TRF $\rightarrow$ microRNA) and (microRNA $\rightarrow$ TRF), over common target gene pathways. We then analysed the distributions of FFL patterns with respect to target gene functions and expression patterns. We identified significant enrichments in many specific core processes within the cell, including developmental processes (lymphocyte differentiation and neurogenesis) (Figure 4.8). This result adds to an already rich literature linking microRNA functions to developmental pathways and regulators, e.g.(Lee *et al.* 1993; Johnson *et al.* 2003; Johnston and Hobert 2003; Burk *et al.* 2008; Peterson *et al.* 2009). We also demonstrated a series of results linking the mean and variation of target gene expression levels to the distributions of transcriptional and post-transcriptional regulators of the gene, and to the connections between these regulators.

It has been shown mathematically, and in specific biological systems, that variation in the output of protein products is lowered as a result of the action of relatively weakly repressive microRNAs on mRNA sequences (Osella *et al.* 2011). Without this post-transcriptional repression, higher levels of noisiness from actively transcribed mRNA populations may be transmitted into the protein population (Baek *et al.* 2008), which can be viewed as a less robust phenotype. It has been speculated that increased robustness drives the expansion of microRNA gene families, which may collectively and often redundantly target the majority of genes in the genome (Peterson *et al.* 2009). Accordingly, this is regarded as one of the most plausible explanations for why microRNAs regulated by specific TRFs often share many of the same target genes, demonstrated biochemically and at the level of genome-wide analysis (Shalgi *et al.* 2007; Yu *et al.* 2008; Tu *et al.* 2009). Our work provides a number of lines of evidence consistent with this model:

Firstly, for a microRNA within a TMG pattern to stabilise but not silence the output of the gene expression pathway, the repression is required to be weak relative to the transcriptional activation (Osella *et al.* 2011). We showed that repression associated with a typical microRNA is weaker than the activation associated with a typical TRF. On average, a TRF was associated with an 86% increase in expression, compared with a 26% drop per microRNA. This suggests that a basic assumption for the noise-buffering role of microRNAs is satisfied for the human TRF – microRNA IRN.

Secondly, we established significant enrichment for three types of TRF – microRNA feedforward circuits (Figure 4.5). The TMG motif was previously identified as significantly enriched using ChIP-seq data together with microRNA target predictions from the TargetScan algorithm (Friedman *et al.* 2009; 2011). Feedforward circuits connecting TRFs and microRNAs have also been predicted from computationally-annotated TRF binding sites (Shalgi *et al.* 2007; Yu *et al.* 2008; Tu *et al.* 2009). Thus, evidence to date suggests frequent coupling of microRNA families to particular TRF-mediated gene expression programmes. We also established that the majority of signal is distributed via promoters of host genes of intronic microRNAs, rather than upstream of microRNAs in other locations. This provides evidence that the model of expression of an intronic microRNA from a common promoter with the host gene performs well (Baskerville and Bartel 2005).

Thirdly, we showed that genes under regulation by TRF – microRNA feedforward circuits tend to be more highly expressed when the TRF lies upstream, so that transcriptional activation outweighs post-transcriptional repression over common target genes (Figure 4.6). Thus, in general, microRNA-mediated regulation is unlikely to result in complete silencing of gene

expression pathways, instead fine-tuning the levels of actively transcribed mRNA and protein products.

Fourthly, we find that the stability of mRNA expression across human tissue samples is directly related to the prevalence of TRF – microRNA feedforward circuits regulating the mRNA (Figure 4.7). Further, this is only true when the TRF lies upstream of the microRNA, consistent with the post-transcriptional layer acting upon an underlying transcriptional regulatory layer. In conclusion, many properties of the ChIP-seq derived IRN support the concept of a significant subset of microRNAs participating in networks that stabilise levels of actively transcribed mRNAs. This is in contrast to a recent article examining expression variation of target mRNAs between individuals and between primate species, suggesting that generally, the targets of microRNAs have higher expression variation (Lu and Clark 2012). The context of that conclusion was not comparable, though, since the placement of the microRNA and gene in relation to common upstream TRFs was not considered. We conclude that our study is consistent with the proposed role of microRNAs as stabilisers of specific transcriptional protein-coding gene expression programmes.

4.5 Conclusions

In this study we have shown that microRNAs are more likely to participate in 2-member regulatory feedback loops with transcriptional regulators that have repressive or variable effects upon transcription. We have shown that TRF – microRNA mediated feedforward regulation is prevalent across common target genes that have low expression variation across human tissues. This property is consistent with the hypothesis that microRNA partners of TRFs can reduce variation in gene expression levels within complex organisms, allowing genetic diversity to accumulate within transcriptional regulatory regions. While many microRNAs have critical roles as silencers of gene expression programs, it is consistent with our results that a large number can act as generalized fine-tuners of gene expression programs. Our analysis also showed that feedforward regulation mediated by TRF and microRNA pairs is enriched over core biological processes including transcription, chromatin modification, signalling pathways, and developmental programs. The cooperativity of transcriptional and post-transcriptional regulatory layers thus reflects an organizing principle within complex cells.

Chapter 5

Species-specific gene expression and transcriptional regulation in pathogenic versus natural host SIV infections

Abstract

Infection of monkeys by simian immunodeficiency virus (SIV) leads to rapid upregulation of interferon-stimulated genes within CD4+ T cells. This upregulation persists in non-naturally infected species, typically rhesus macaques (RMs), and has been linked with chronic proliferation of T cells and eventual immune system exhaustion. By contrast, in infection of natural SIV host species such as African green monkeys (AGMs) or Sooty mangabeys, the initial cytokine response is rapidly suppressed, and long-term infection usually does not lead to pathogenesis. Thus, SIV infection of RMs serves as a model for HIV infection in humans, contrasted with a non-pathogenic outcome in AGMs. In this study we combine (i) expression time series datasets from CD4+ T cells following SIV infection of AGMs and RMs with (ii) genome-wide binding sites for 75 transcriptional regulators surveyed in human lymphatic cell lines, and (iii) a database of HIV – host protein interactions. By clustering gene expression patterns, we describe co-expressed collections of transcripts involved in AGM and RM antiviral and T cell proliferative responses. These species-specific transcriptional dynamics are likely to be regulated by many factors, including statistically significant numbers of STAT, IRF, and bZIP transcription factors, including the pro-inflammatory BATF subfamily recently linked to HIV pathogenesis. In particular, we identify significant co-association of STAT1, STAT2, IRF4 and BATF binding sites over genes involved in the species-specific viral response. Although the initial trigger for these transcriptional cascades remains unclear, analysis of protein-interaction networks suggests that STAT1 and IRF7 activation lies downstream of the lentiviral Tat protein.

Contributions

This study is the result of collaboration between Jamie I MacPherson, Aaron Webber, Beatrice Jacquelin, Arndt Benecke, Michaela C. Muller-Trütwin and David L. Robertson. JIM and AW collaborated as joint first authors for the prospective journal article corresponding to

this chapter. The project was conceived by JIM, BJ, MCMT and DLR. Analysis and preparation of the manuscript was carried out by JIM and AW. Specifically, JIM prepared gene expression clusters and analyzed the gene set overlaps and functional enrichments of these, as in Figures 5.1 to 5.4, as well as the data provided in Supplementary Table S5.1. JIM also analyzed interactions between viral and host proteins, as in Figure 5.9 and Table 5.5, and prepared the first version of the paper. AW prepared the material on transcription factors (Figures 5.5 to 5.8), wrote the accompanying text for these sections, and amended the text for Figures 5.1 to 5.4, following initial reviewer feedback. AB provided data files. All authors commented on and approved the final version of the manuscript.

5.1 Introduction

For the vast majority of infected individuals, human immunodeficiency virus type 1 (HIV-1) causes depletion of CD4+ T cells leading to acquired immunodeficiency syndrome (AIDS). Disease progression is associated with generalized T cell stimulation, and this excessive immune activation is considered the driving force of critical CD4+ T cell depletion (Giorgi, Fahey *et al.* 1987; Picker 2006). Deterioration of the host immune response occurs over a long time period and the onset of acquired immune deficiency syndrome (AIDS) can take years even if the infection is left untreated (Morgan, Mahe *et al.* 2002). By contrast, for many families of the related simian immunodeficiency virus (SIV), infection of natural host species such as African green monkey (AGM), sooty mangabey (SM), or mandrill, is usually not pathogenic. These animals mainly have a normal life span, even though the infection is never cleared (Paiardini, Pandrea *et al.* 2009; Sodora, Allan *et al.* 2009; Souquiere, Onanga *et al.* 2009). When strains of SIV infect non-natural monkey host species, such as rhesus macaque (RM), this type of infection is again pathogenic and leads to generalized T cell activation and eventual immune-system exhaustion similar to human AIDS (Gardner 1996). Infection of RMs by SIV has therefore been used as a model pathogenic system likened to HIV-1 infection of most humans (Gardner 1996; Paiardini *et al.* 2009). Similarly, the ability of natural SIV host monkey species to maintain a functioning immune system, despite never clearing the infection, has the potential to provide insight into protective mechanisms. In turn, this could lead to the identification of new therapeutic targets.

Three recent studies, including one of our own, compared changes in gene expression following infection by SIV of a natural host (AGM or SM) and a non-natural macaque host (RMs or pigtailed macaques) of the respective SIV families (SIVagm, or SIVssm) (Bosinger, Li *et al.* 2009; Jacquelin *et al.* 2009; Lederer, Favre *et al.* 2009). In our study, whole cell RNA was extracted from lymph node and peripheral blood CD4+ cells from both AGMs and RMs, at a number of time points before and after infection with the SIVagm family (Table 5.1). These were then converted to cDNA libraries and hybridized to human protein-coding gene microarrays, and expression changes within specific immune system pathways compared between monkey species. The acute immune response within both species includes production of type 1 interferon (IFN) and strong up-regulation of many IFN-stimulated genes (ISGs) (Diop, Ploquin *et al.* 2008). However, the pattern of gene expression through time linked to a type 1 IFN response differs greatly between primate species. The induction of ISG expression is at least as rapid and as strong in AGMs as in RMs. However, while ISG expression in natural hosts is attenuated by the end of the acute period of SIV infection, ISG expression in

macaques is sustained throughout chronic infection (Bosinger *et al.* 2009; Jacquelin *et al.* 2009; Lederer *et al.* 2009). Based on these results, we and others hypothesised a repressive regulatory mechanism to explain the efficient attenuation of the innate immune response in AGMs (Bosinger *et al.* 2009; Jacquelin *et al.* 2009; Lederer *et al.* 2009). Conversely, the lack of an immune-suppressive control mechanism in RMs might account for a chronic pro-inflammatory cytokine profile within CD4+ T cells in RMs, resulting in generalized T cell proliferation and eventual exhaustion of the RM immune system.

Table 5.1. Time series datasets across species and tissue compartments.

	Time (days post-infection)											
Source	-90	-70	-40	-8	1	6	14	28	41	65	115	Final
MP												
ML												
AP												
AL												

Cells are shaded according to the number of microarray datasets obtained, from a maximum of six AGMs and six RMs at each time point: blue = 6, grey = 5, white = 0. The final time points for collection of CD4+ T cells from lymph nodes were 574 days p.i. and 610 days p.i. in AGMs and RMs respectively. We denote the four combinations of monkey species and CD4+ T cell subpopulation as AL (AGM lymph node), AP (AGM peripheral blood), ML (RM lymph node) and MP (RM peripheral blood).

In this work we set out to identify transcriptional regulatory factors targeting functionally related groups of genes with different expression patterns between natural and non-natural SIV host species. To this end, we combined additional regulatory and protein interaction datasets with the gene expression data from SIV infected AGMs and RMs from our previous study (Jacquelin *et al.* 2009). Consistent with the microarray platform used, the present analysis is conducted with respect to human protein-coding genes. First, we used gene expression clustering to identify unbiased collections of genes most likely to share transcriptional regulators. Next, we obtained publically available ChIP-seq datasets providing genome-wide binding sites for 75 transcriptional regulators in human lymphatic system cell lines. We then identified TFs with the greatest enrichment in binding sites within the *cis*-regulatory regions of clustered collections of genes. These factors include multiple STAT1, STAT2, IRF4 and the pro-inflammatory bZIP factor BATF, recently connected with T cell exhaustion in the context of HIV (Larsson, Shankar *et al.* 2013). When compared to all annotated transcriptional regulatory genes, the same families of factors display the most

significant and species-specific expression perturbations throughout the acute phase of SIV infection in AGMs and RMs. We therefore propose for experimental testing novel regulators of species-specific immune system activation. Finally, we analyse how SIV could contribute to the induction of immune activation, using a database of experimentally-derived viral - host protein interactions (Ptak, Fu *et al.* 2008; Fu, Sanders-Beer *et al.* 2009). Our work contributes to an understanding of the transcriptional events that characterize pathogenic lentiviral infections, and control of CD4+ T cell activation in natural host species.

5.2 Methods

Gene expression in SIV infected primates

Preprocessed gene expression data, from both peripheral blood (PB) and lymph node (LN) CD4+ T cells was obtained from our study of SIV infection among African green monkeys (AGMs) and rhesus macaques (RMs) (Jacquelin *et al.* 2009). Gene expression data included expression levels from individual animals (typically six at each time-point) and also combined data giving mean log2 expression levels for probes compared with baseline expression and statistical significance for differential expression. Mappings between simian and human genes were also present in the expression data. The raw expression data can be downloaded from the MACE database (http://mace.ihes.fr) using accession numbers 3070984318 (AGM) and 2932572286 (RM).

Clustering of gene expression profiles

Probe expression profiles, consisting of mean log2 gene expression values for all mutually available time-points post-infection (1, 14, 28 and 65 days) from AGMs and RMs, for both PB and LN CD4+ cells were selected for differentially expressed probes ($p < 0.1$ at one or more time-point). Expression profiles, regardless of their source, were pooled and clustered using Mfuzz soft clustering with a '*fuzzification*' parameter of 1.25 (Kumar and Futschik 2007), and a stringent within-error (α) of > 0.6 and, hence, the choice of a permissive P value cutoff for differential expression. Probe IDs were assigned to the single cluster that they fit with the largest value for α. A full list of gene name to expression profile ID mappings are given in Supplementary Table S5.1.

Functional enrichment analysis

Functional enrichment analysis of clustered genes was performed using DAVID 6.7 (Dennis *et al.* 2003; Huang da *et al.* 2009), taking the Benjamini and Hochberg (Benjamini *et al.* 2001) corrected P values as a measure of significance. In addition enrichment for interferon stimulated genes (ISGs) was calculated separately for each primate species-cell type combination, for each expression profile. ISGs were retrieved from a database (de Veer, Holko *et al.* 2001). Clusters were tested for enrichment of ISGs by Fisher's exact test if they contained one or more ISG, using the number of genes expressed for the given species-cell type source as a background population. P values were adjusted for multiple tests (Benjamini *et al.* 2001).

Computation of significant intersections between expression profile gene sets

An all-against-all comparison of subsets of clusters from each monkey species and CD4+ source was performed and the number of probe IDs common to two subsets, i.e., from different microarray samples, was identified. Where the intersection was larger than the expected proportion under a null model, a P value was calculated using Fisher's exact test. P values were adjusted for performing multiple tests (Benjamini *et al.* 2001).

Expression of transcription factors in each primate species and cell type

Genome-wide ChiP-seq peaks giving binding site locations for 84 transcriptional regulators, within human lymphatic cell lines (K562, K562B and GM12878), published by the Yale and HAIB consortia, were retrieved from the UCSC genome browser (Supplementary Table S5.2) (Karolchik *et al.* 2003; Karolchik *et al.* 2004). Biological replicates from the HAIB consortium were merged by taking the intersection of peak intervals, with ≥1 nucleotide in common between replicates. To annotate potential *cis*-regulated target genes, the complete collection of Ensembl v.65 protein-coding genes was downloaded from Biomart (Kinsella *et al.* 2011). Transcriptional regulator binding sites between -2 kb and + 2 kb of transcription start sites of these protein-coding genes were then recorded, taking the union across replicates from different lymphatic cell lines. Nine of the regulators were found to target less than 2% of genes genome-wide (mean targets = 110 genes) and were left aside, resulting in the set of 75 regulators analyzed in the remainder of the study.

Relationships between TRFs and gene expression clusters

Expression patterns of TRFs were obtained from microarray datasets, filtered and normalized as described in our previous study (Jacquelin *et al.* 2009). We restricted attention to TRFs with expression time series passing the earlier filtering step in no fewer than 3 of the CD4+ samples from each source. For each transcriptional regulator, mean targets per gene cluster and across the genome as a whole were calculated. Statistical significance was first estimated using the Poisson distribution, taking the expected number of targets per cluster as the genome wide fraction of targets for the TRF, multiplied by the size of the cluster. The number of independent tests carried out is equal to the number of clusters (=15) multiplied by the number of factors (=75), and using this number of tests, a multiple-testing correction was applied to Poisson p-values (Benjamini *et al.* 2001). We then introduced an additional constraint, requiring the total number of TRFs per gene to be held fixed. This provides a more stringent test of specificity of factors for different clusters of genes, independently of the total

number of regulators per cluster. Connections between TRFs and clusters were shuffled by swapping random pairs of connections between TRFs and genes, so that the total number of connections to a given gene never changes. This process was continued within each network simulation, until no further effective randomization was achieved. To determine when further shuffling would be unproductive, we required the coefficient of variation of numbers of edges shuffled, compared to the real network, and calculated using the previous 1000 attempted edge swaps, to fall below 0.05. The entire shuffling process was run 100,000 times, and the number of connections between each TRF and gene cluster counted in each of these 100,000 runs. The significance of links between TRFs and clusters was recorded as a p-value equal to the proportion of simulated runs in which the TRF had more targets within the cluster than in the real network. Computations were performed using code written in Java.

Detection of significantly sized virally activated regulatory subnetworks

Regulatory networks of differentially expressed simian host genes were constructed by mutual information implemented in MRnet (Meyer, Kontos *et al.* 2007), with spearman's correlation as the entropy estimator between their expression profiles between days 1 and 115, for all available time-points. Gene pairs were also filtered by the semantic similarity in their GO annotations. The choice of default parameters in MRnet and threshold semantic similarity was determined as those choices for which the false positive rate was minimized over a curated database of validated protein-protein interactions in yeast (the nearest systematically studied eukaryotic organism) (Stark, Breitkreutz *et al.* 2006). MRnet was then run on the simian expression data to calculate an approximate simian gene regulatory network. Dysregulated genes were obtained from the HIV-1 host protein interaction database (Fu *et al.* 2009) selecting only those cellular genes that are regulated, up-regulated, or downregulated, by one or more HIV-1 proteins (excluding Vpu, which is not conserved in SIV). Subnetworks consisting of cellular genes were then obtained, containing each dysregulated gene from the HIV-1 protein interaction database, and statistical significance for subnetwork sizes calculated using a permutation test. In each permutation, a randomized network was produced by repeatedly swapping one of the two incident nodes between two randomly selected edges, allowing 1000 attempts. Subnetworks for each dysregulated gene were then obtained from the randomized network. The size of the connected component containing the dysregulated gene was calculated for both randomized and original subnetworks. Following 100 permutations, subnetwork sizes for each dysregulated gene were compared for unperturbed and perturbed networks by Mann Whitney U test. P values were adjusted for performing multiple tests (Benjamini *et al.* 2001).

5.3 Results

The expression datasets used in this study comprise 212 AB1700 microarrays, measuring whole cell mRNA from the lymph node and peripheral blood CD4+ T cells of six AGMs and six RMs, before and after SIV infection. We denote the four combinations of monkey species and CD4+ T cell subpopulation as AL (AGM lymph node), AP (AGM peripheral blood), ML (RM lymph node) and MP (RM peripheral blood)., and refer to these as CD4+ *sources*. Samples were obtained from 90 days before to 610 days after SIV infection, but with some variation between tissues and species (Table 5.1). We initially concentrated on five time points with samples from all six animals for each CD4+ source: day 0, defined by the mean expression level pre-infection, and days 1, 14, 28 and 65 post-infection (p.i.).

5.3.1 Gene expression patterns reflect species-specific T cell dynamics

Our first aim was to identify functional trends matching species-specific patterns of gene expression. To do this we first clustered expression profiles of genes through time following SIV infection. We then tested co-expressed collections of genes for functional enrichment. Attention was restricted to genes with the most consistent expression changes between the six animal replicates from each CD4+ source (see methods). The number of genes with sufficiently consistent expression was higher in LN than in PB, and higher in RM than in AGM, corresponding to 484 ML, 338 MP, 268 AL and 210 AP microarray probes. Expression profiles for these probes were then assigned to 15 expression clusters, using the algorithm Mfuzz (Kumar and Futschik 2007). Mfuzz assigned over 90% of probes to clusters (1216 from 1300), with cluster sizes ranging from 13 to 339 probes (mean = 82.8 probes). Expression profiles for the complete set of 15 clusters, together with their sources and complete gene lists, are provided in Figures 5.1 & 5.2 and Supplementary Table S5.1.

We then tested the functional enrichment of each of these clusters by submitting gene lists from each cluster to the Gene Ontology analysis server DAVID (Huang da *et al.* 2009). In total eight clusters, containing 78% of the perturbed probes, are enriched in one or more biological functions ($p < 0.0062$, corrected for multiple tests). The expression patterns of probes in each of these clusters are displayed in Figure 5.3, together with counts of probes from each source, and the genes from the top-scoring functional annotation of the cluster.

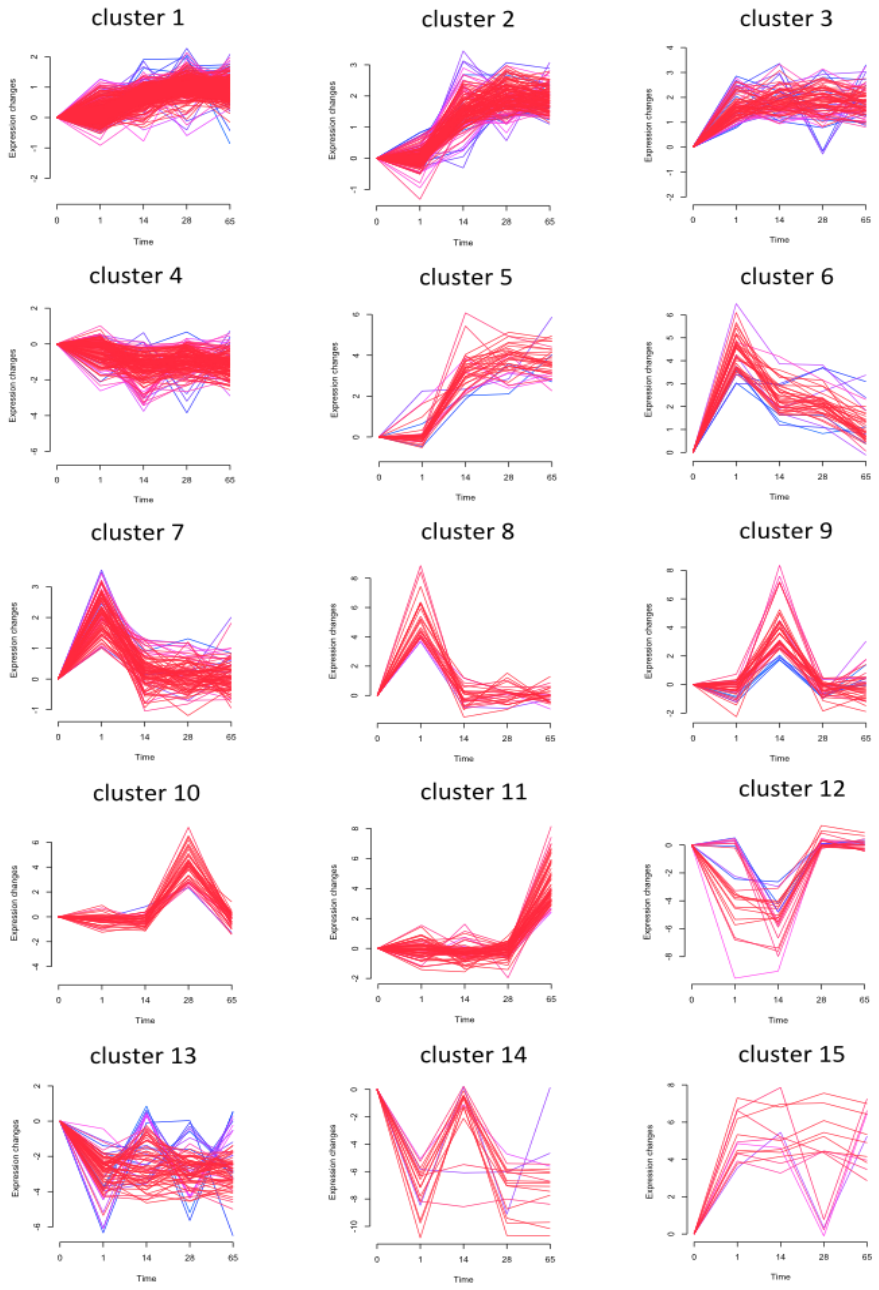


Figure 5.1. Gene clusters with common expression profiles upon SIV infection.

Gene expression clusters were calculated as described in the main text. Each plot represents a set of microarray probes clustered according to their expression profile following SIV infection. Each line represents the expression of a gene from either SIV infected AGMs or RMs, or from either peripheral blood (PB) or lymph node (LN) CD4+ T cells. Line colour indicates the goodness of fit for genes to the given expression profile where red is the best fit, blue the least good fit.

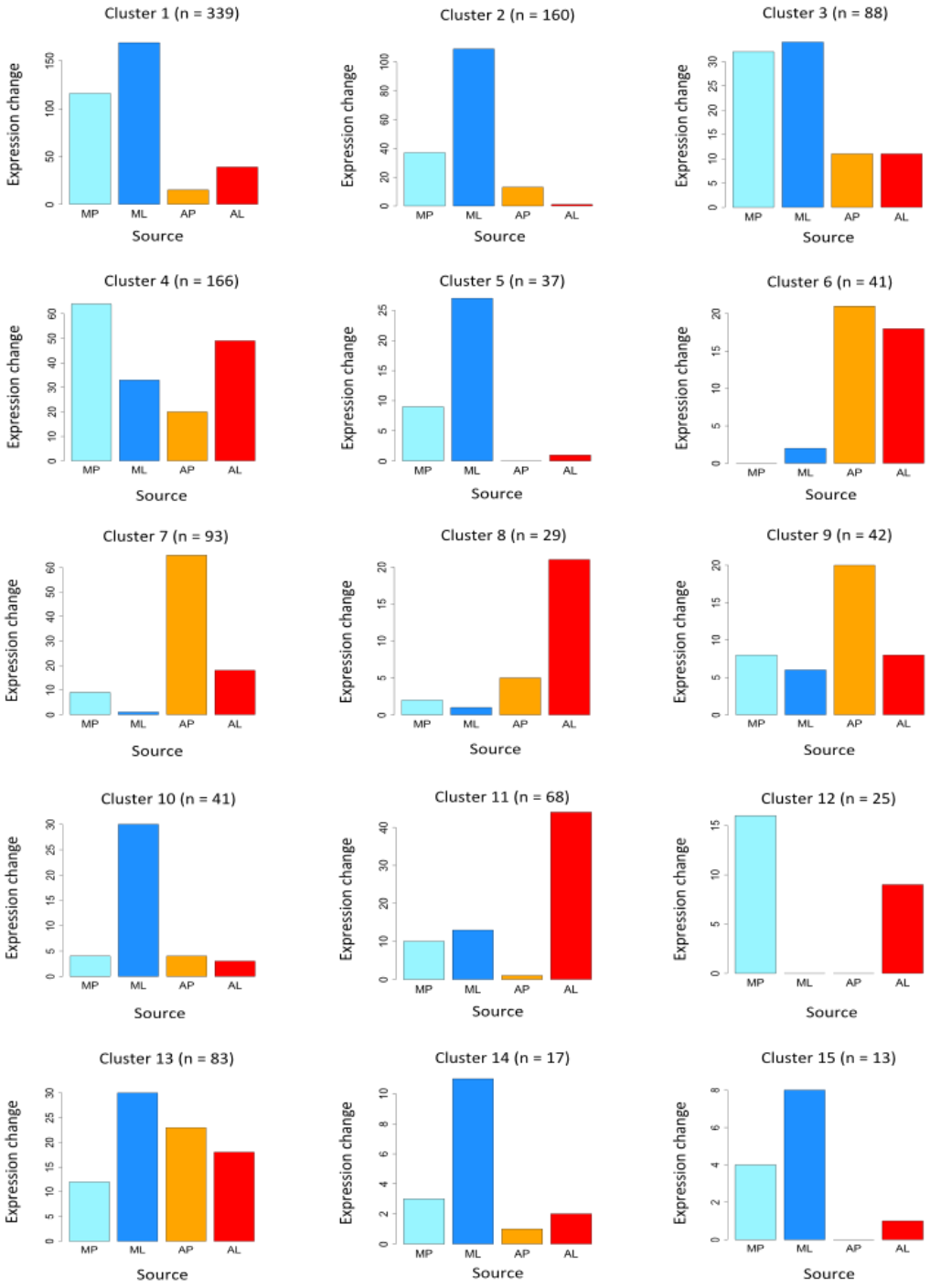


Figure 5.2. Composition of genes in each expression profile by species and T cell source.

Each plot shows the distribution of gene sources (primate species and cell type) from a cluster. The y-axis values denote the number of clustered probes from each source that are present in a cluster.

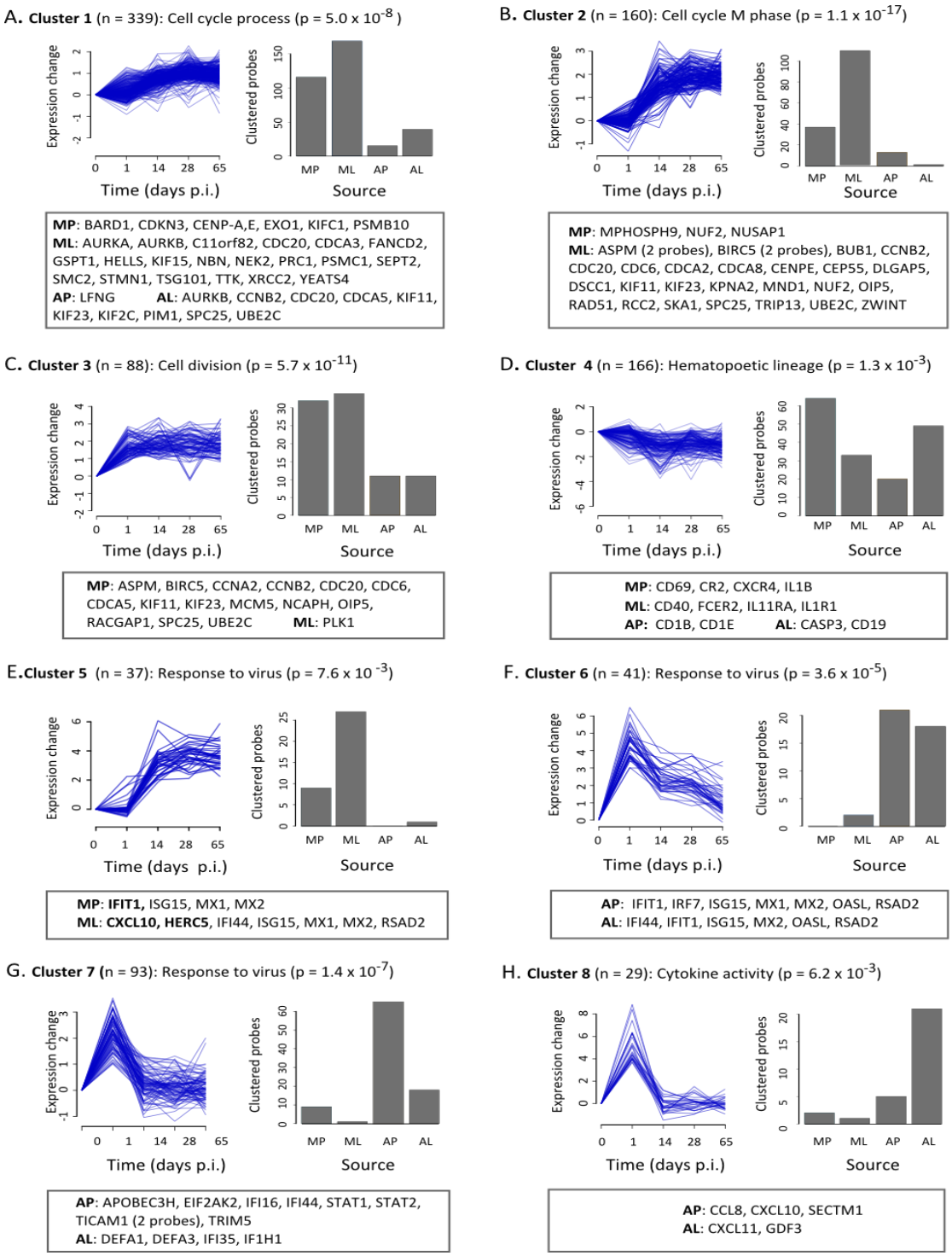


Figure 5.3. Expression patterns of cell-cycle and viral-response enriched gene clusters.

Plots on the left-hand side of each subfigure superimpose $\log_2$(expression) changes between days 0 and 65 p.i. for every microarray probe within the cluster. Plots on the right-hand side display the breakdown of probes according to CD4+ source (MP = macaque peripheral blood; ML = macaque lymph node; AP = AGM peripheral blood; AL = AGM lymph node). Numbers of genes are shown in brackets, together with most enriched GO term and corresponding p-value (see methods). Boxes below each chart show the breakdown by CD4+ source of genes corresponding to the most enriched GO term.

The first three clusters derive mainly from RM and relate to control of the cell cycle, especially mitotic cell division (Figure 5.3.A – C). Mean expression level across these clusters is generally elevated at between 2 – 4 times the pre-infection level at all time points post-infection. Numbers of CD4+ T cells and rates of cell division reflect the outcome of both cell proliferation rates and cell death, due to viral infection and naturally occurring apoptosis (Monceaux, Viollet *et al.* 2007). Significantly more genes involved in cell division were identified in ML than in MP CD4+ T cells ($p = 0.0085$, by binomial test). This is in agreement with measurements showing a significant increase in the fraction of proliferating CD4+ T cells within the tissues of SIV infected compared to healthy macaques (Wang, Xu *et al.* 2013). Peak viral load in RMs occurs around 8 – 12 days (p.i.) (Jacquelin *et al.* 2009; Wang *et al.* 2013), and consistent with this, cell division regulators within host CD4+ T cells have the greatest increase in expression during the time period from days 1 to 14 (p.i.). During and after this period, there are fluctuations around the mean expression level within clusters for some genes within clusters 1 - 3. For example, in cluster 2, the ML cell cycle related genes which increase the most between days 1 and 14 (p.i.) then decrease by day 28 (p.i.), before recovering to a level around 4-fold greater than their expression levels pre-infection. In both macaques and humans, chronically proliferating LN CD4+ T cells are associated with immune system exhaustion and progression to immune deficiency syndrome (Hazenberg, Otto *et al.* 2003; Kornfeld, Ploquin *et al.* 2005; Estes, Gordon *et al.* 2008). These early differences in cell proliferation regulators between AGMs and RMs may therefore be relevant to the long-term tolerance to SIV infection that is achieved in the natural hosts, AGMs.

A number of studies, including one of our own, have demonstrated that Type I interferon-stimulated genes (ISGs), including signal transducers, interferons, anti-viral factors, and other cytokines, are strongly upregulated following SIV infection in both RMs and AGMs, but are then swiftly attenuated in natural hosts (Bosinger *et al.* 2009; Jacquelin *et al.* 2009; Lederer *et al.* 2009). The unbiased gene expression clustering strategy used here has identified groups of genes in clusters 5 – 8 precisely matching these trends (Figure 5.3.E - H). In particular, anti-viral ISGs, including MX1, MX2 and RSAD2, are upregulated in both species, but then rapidly downregulated in AGMs (cluster 5: RMs; clusters 6 – 8: AGMs). This anti-viral response occurs within only a single day following infection in AP and AL CD4+ T cells (clusters 6 - 8), but within 6 days of SIV infection in RM CD4+ T cells (cluster 5). Indeed, the maximum fold changes of ISGs within RMs are significantly later than in AGMs ($p = 6.2 \times 10^{-8}$, by Mann-Whitney U-test). Thus, CD4+ T cells respond significantly faster in the natural than in the pathogenic host system. Significant changes within the first day of infection of AGMs include a roughly 15-fold expression increase in PB of IRF7 and a more than a 100-fold increase in ISG15 expression.

These very early differences in cytokine expression profiles of AGMs and RMs may lead to host species-specific development of the lymph node and peripheral blood CD4+ T cell populations, together with altered relationships both with the virus and with other immune cell subtypes. In particular, the rapid reaction of the AGM CD4+ T cell population to SIVagm infection may lead to timely expression of suppressive factors, which are able to control the rate of CD4+ T cell proliferation within AGMs.

5.3.2 Species-specific patterns of overlap between gene expression clusters

Clusters are collections of genes with similar expression patterns through time (Figures 5.1 and 5.3). Evidently the collections of genes within each cluster may be common to, or different between, species and tissues. We therefore measured all overlaps between subsets of gene clusters from distinct sources and tested whether these subsets have significantly many genes in common with one another, comparing patterns within and between clusters. We found in total thirteen statistically significant overlaps between gene sets from each CD4+ T cell source within each expression profile ($p < 0.05$, by Fisher's exact test, corrected for multiple tests). These overlaps segregated according to the functional trends already identified: cell cycle gene enriched, viral response gene enriched and all other clusters (Figure 5.4). Six overlaps are between genes from different sources present in the same expression profile, shown by edges from a cluster to itself. The majority of these overlaps, within clusters 2, 5 - 6, 9, and 15, reflect a shared set of genes from the two CD4+ T cell sources from the same species, i.e. peripheral blood and lymph node expression patterns of the genes were shared. This is presumably due to movement between the two populations of cells. The remaining self-edge, within cluster 1, corresponds to a collection of genes from both AGM and RM peripheral blood, reflecting a shared rather than a species-specific component of the simian immune response to SIV infection.

We also find seven significant overlaps are between gene lists within CD4+ T cell sources from different clusters. Four of these occur between groups of genes differentially expressed between AGMs and RMs, within a given CD4+ T cell subpopulation. These are therefore candidates for determination of a species-specific immune response. Two of these four overlaps occur between clusters 5 and 6, once for expression patterns in PB and once for expression patterns in LN. We have already mentioned these genes, including IFIT3, ISG15, MX1 and MX2, as giving rise to anti-viral responses specific to each species (Bosinger *et al.* 2009; Jacquelin *et al.* 2009; Lederer *et al.* 2009). The remaining species-specific overlaps are between clusters 9 and 12, and clusters 11 and 13 (Figure 5.3; for the breakdown of these clusters into CD4+ sources, see Figure 5.2). The overlap between clusters 9 and 12 shows a

contrasting response between species, as the expression of shared genes transiently increases in RMs, but decreases in AGMs, at day 14. Interestingly, five genes (DNAJB13, EBF2, FAM124A, PRDM10, SYT6) were common to all four edges within the circuit of overlapping clusters (9 ↔ 12 ↔ 11 ↔ 13 ↔ 9) shown in Figure 5.4C. Despite this, we could not immediately see any shared connection with the response of CD4+ cells to SIV infection. Although EBF2 (early B-cell factor) is a regulator of immune cell differentiation, it is usually associated with immature osteoblastic cells, rather than mature T cells undergoing activation (Kieslinger, Hiechinger *et al.* 2010).

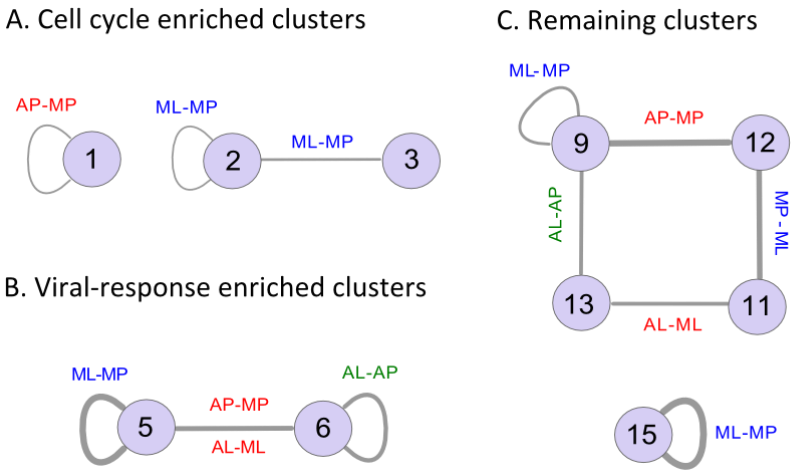


Figure 5.4. Significant gene intersections between expression profiles.

Nodes represent gene expression clusters 1-15 (numbered). Clusters that have statistically significant intersection in their gene sets (Fisher's exact test, $p < 0.05$) stemming from two different sources (AL, AP, ML, MP) are linked, or self-linked. Links are labeled with the sources from which the overlapped set of genes derives. For example, clusters 2 and 3 share a significant number of genes deriving from ML and MP sources, respectively. These labels are colored blue, for a relationship between RM CD4+ subpopulations, green, for a relationship between AGM CD4+ subpopulations, or red, for a relationship between the two species. No significant overlaps were found between clusters from the different groups **A**, **B** and **C**. Edge-width is proportional to the Matthews Correlation Coefficient calculated from the gene set intersections.

5.3.3 Expression changes within peripheral blood CD4+ T cells from individual animals

We next examined the expression of interferon-stimulated genes within each of the six AGM and RM animals individually. We focused on PB since this is sampled at almost twice as many time points within both species (Table 5.1). Expression patterns were examined for a number of ISGs, including the ubiquitin-like interferon-stimulated gene ISG15 (Takeuchi and Yokosawa

2008; Skaug and Chen 2010), from MP in cluster 5 and AP in cluster 6, and the cytokine signal-transducing transcription factor STAT1 (Stark and Darnell 2012), from MP in cluster 2 and AP in cluster 7 (Figure 5.5 A, D and B, E). The expression of these genes is consistent across 5 / 6 RMs, with a roughly 8 – 60 fold upregulation by day 6 in MP samples, followed by a 2 – 16 fold decrease by day 14, and then a return to near-peak levels. This is true for many other individual ISGs (e.g. MX1, MX2, IFIT2, IFIT3 – data not shown) and also for the average expression pattern of all MP probes in cluster 5 (Figure 5.5C). The 6th RM individual, indicated by the red line in Figures 5.5.A – C, has a less intense production of ISGs at day 6 (p.i.), and little variation in ISG expression between days 6 – 28 (p.i.). Interestingly, this individual also expresses cell cycle related genes in cluster 2, from days 28 to 65 (p.i.), at almost twice the mean level of the other five RMs. Although the number of individuals sampled is quite small, this could indicate that the relative intensity of ISG production in RM PB between days 6 - 14 is related to the rate of cell proliferation at later time points.

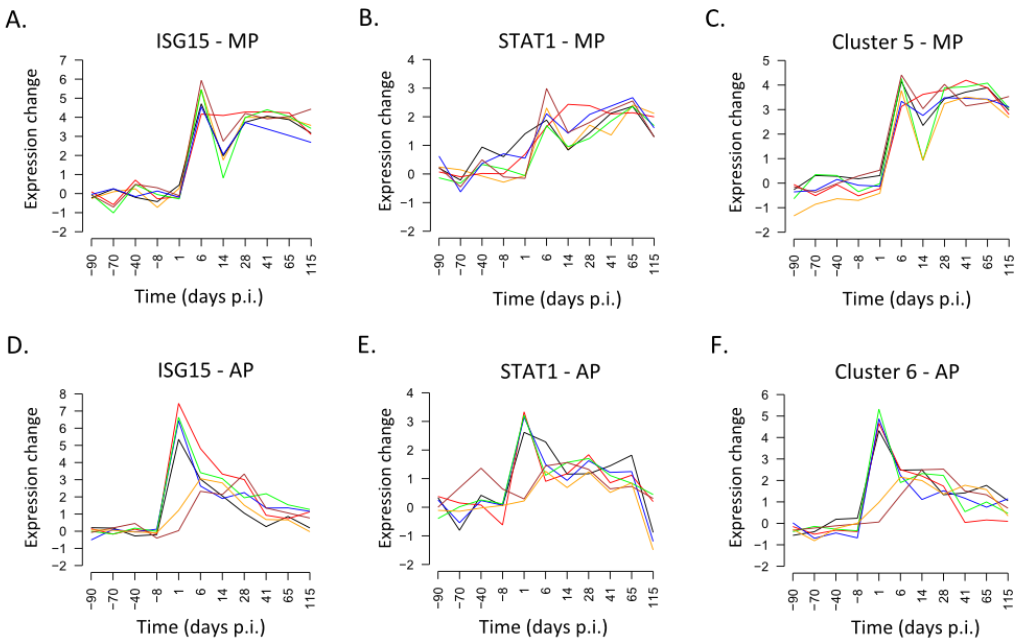


Figure 5.5. Expression profiles of single and clustered genes from individual monkeys.

Log_2(expression) within PB samples from individual RMs (**A, B**) and AGMs (**D, E**) for ISG15 and STAT1. **C, F.** Expression changes within individual RMs and AGMs averaged across all MP probes in cluster 5 (C) and all AGM probes in cluster 6 (G). Lines are colour-coded for individual animals from each of the two species, so that, for example, the red line in figures A – D represents the same RM individual. Missing expression levels for one AGM at day -70 and one RM at day 41 were obtained by linear interpolation.

We next examined the expression levels of genes within individual AGMs. In the majority of AGMs (4 / 6), ISG expression is both rapid and strong but is then attenuated, as shown in clusters 6 – 8, typical of the natural host species (Bosinger *et al.* 2009; Jacquelin *et al.* 2009; Bosinger, Jochems *et al.* 2013). However, for 2 out of 6 AGMs the expression of these anti-viral factors and cytokines is delayed, and from the sampled time points, appears much weaker (indicated by the brown and yellow lines in Figure 5.5.D – H). This is the case for many individual ISGs (e.g. ISG15, STAT1, MX1, MX2, IFIT2), and for the cluster 6 ISG profile as a whole (Figure 5.5G). A possible explanation for two kinds of ISG dynamics in AGMs is that the natural host species displays more than one mechanism of tolerance to and control of SIV infection, which could depend upon different allelic forms of key regulators within the AGM population. Indeed, differences in haplotypes between viral restriction factors including APOBEC3H have been linked to variable degrees of resistance to HIV pathogenesis in human populations (Zhen, Wang *et al.* 2010; Cagliani, Riva *et al.* 2011).

5.3.4 Transcriptional regulators of species-specific gene expression programs

After identifying a number of typical species-specific expression patterns, our next aim was to match these to candidate transcriptional regulators. We first examined expression patterns and target sets of transcriptional regulators conserved within primates. The input datasets were a curated list of 1838 human transcription and transcription regulatory factors (TRFs) (Vaquerizas *et al.* 2009), together with genome-wide ChIP-seq binding site distributions for 85 of these, published by the ENCODE consortium (ENCODE 2011). We restricted attention to 1528 conserved TRFs matching 1791 probes on the AB1700 array platform, and to ChIP-seq datasets for 75 regulators of protein-coding gene expression sampled within lymphatic system derived cell types (K562, K562B and GM12878) (See methods). We consider (i) TRFs identified within gene expression clusters, and where available (ii) the distributions of their binding sites within the *cis*-regulatory regions of protein-coding genes, and (iii) relationships between expression patterns of TRFs and the structural and functional families to which the TRFs belong.

5.3.4.1 Transcriptional regulators in gene expression clusters

The complete collection of 1216 clustered probes was searched for matches to one of the 1528 TRFs present on the microarray platform. In total, 94 probes corresponding to 67 TRFs were identified. We then tested whether the number of TRF probes from each source was greater or less than expected given the total number of probes per cluster. In general, numbers of clustered TRFs from each source and cluster reflect their expected rates ($p > 0.05$

by binomial tests). However, the set of probes from the AP source in cluster 7 is significantly enriched in probes matching TRFs, with 14 observed compared to 5.4 expected ($p = 0.037$, using the binomial distribution, corrected for multiple tests). The TRFs corresponding to these 14 probes are listed in Table 5.2.

Table 5.2. Collection of co-expressed TRFs from AP CD4+ T cells within cluster 7.

Functional class	TRF names (symbol; microarray probe IDs)	Expression change rank from 834 TRFs	
		Up (-8,1)	Down (1, 115)
Signal transducer	Signal-transducer of activated transcription 1 (STAT1; 200004) Signal-transducer of activated transcription 2 (STAT2; 203663)	3* 31	3* 82
Interferon family	Interferon-regulatory factor 2 (IRF2; 167559)	24	56
PML body	Promyelocytic leukemia factor (PML; 129341, 204369, 217558) Speckled nuclear antigen 100 (SP100; 140721)	1* 6*	1* 51
Cytokine	Gamma-interferon-inducible protein Ifi-16 (IFI16; 161333)	10*	21
Leucine zipper (bZIP)	cAMP-responsive element modulator (CREM; 141393) bZIP transcription factor, ATF-like 3 (BATF3; 144215)	2* 5*	404 100
Zinc finger	PR domain zinc finger protein 1 (PRDM1/BLIMP-1; 205036) Ring-finger protein 213 (RNF213; 74799) Zinc-finger protein 267 (ZNF267; 213038)	19 115 96	163 158 359
Other	AF4/FMR2 family member 1 (AFF1; 206144)	12	118

Expression change ranks were calculated based upon a t-statistic, across AGM replicates and including all microarray probe sets for each factor, calculated between time points -8 days before infection and 1 day post-infection, and between 1 day post-infection, and 115 days post-infection. Significant changes in expression level ($p < 0.05$), corrected for multiple tests, are indicated by an asterisk.

We also ranked TRFs by paired t-statistic calculated from their expression level at days -8 and +1 (p.i.), and again between days 1 and 115 (p.i.). In addition, since RNA yields per time point were variable, individual probe sets did not always have a high quality expression time series within all 6 replicate animals (Jacquelin *et al*. 2009). We required an expression time series profile of sufficient quality within at least three of the AGM PB samples, with 834 TRFs satisfying this condition. Ranked TRFs with significant expression change t-statistics ($p < 0.05$, after correction for multiple tests), are indicated by an asterisk. Many of the TRFs in cluster 7, including PML, STAT1, CREM and BATF3, are among the TRFs with most rapidly increasing expression level in AP genome-wide during the first day of infection. STAT1 and PML also rank among the most downregulated between day 1 and the final sampled time point (day 115 p.i.) in PB (by paired t-test, calculated between day 1 and day 115 p.i.). In contrast, cAMP-response element modulator (CREM) shows almost no net reduction in expression level between days 1 and day 115. From the expression time series for this factor, we find that this is due mainly to up-regulation of CREM between days 65 and 115, uniquely within the AGM peripheral blood.

Most of these factors have well-established roles in response to viral infection: as transducers of IFN-α,-β signalling and drivers of ISG expression (STAT1 and STAT2) (Katze, He *et al.* 2002); as double-stranded viral DNA sensors and regulators of cell growth and genomic recombination events (PML, SP100 and IFI16); and as inhibitors of cytokine signalling (IRF2 as a repressor of IRF1-mediated signalling (Honda, Mizutani *et al.* 2004); PML again, as a repressor of IFN-γ signalling (Choi, Bernardi *et al.* 2006); and PRDM1/BLIMP-1 as a generalized repressor of IFN-β signalling and cell growth, and driver of B-cell mediated immunity (Nutt, Fairfax *et al.* 2007). The basic leucine zipper (bZIP) factor BATF3 (alias: SNFT) is more commonly associated with CD8+ T cell-mediated cytokine-signalling pathways, leading to maturation of dendritic cells (DCs) in response to pathogens (Tussiwand, Lee *et al.* 2012). Like its paralogs BATF and BATF2, it is able to dimerize with other bZIP family members, in particular, leading to repression of the cell-growth regulating, immediate early gene, AP-1 complex, consisting of Jun and Fos family dimers (Murphy, Tussiwand *et al.* 2013). Thus, each of the factors PML, SP100, IFI16, PRDM1 and BATF3 has the potential to link the early AP-specific viral response to subsequent modified CD4+ T cell proliferation dynamics within AGMs, compared to RMs. Finally, the bZIP factor CREM is notable as an inhibitor of the chemokine-receptor CCR5 gene promoter, the mature product of which serves as the main recognition element for entry of SIV and HIV virions into host CD4+ cells (Banerjee, Pirrone *et al.* 2011). Taken together, there is a strong case for regarding the dynamics of TRFs within the peripheral blood CD4+ T cells of typical AGMs as part of a fast-acting coordinated reaction to SIV infection.

5.3.4.2 Transcriptional regulators bound to *cis*-regulatory regions of clustered genes

Clusters of genes with similar expression time series might be controlled by shared transcriptional regulators (Herrmann, Van de Sande *et al.* 2012; Zhou, Ma *et al.* 2012). To test for an enrichment in shared transcriptional regulators within each gene cluster, we first obtained ChIP-seq datasets of genome-wide locations of 75 TRFs in human K562, K562B and GM12878 lymphatic system cell lines (Supplementary Table S5.2). These include some of the TRFs already discussed, including STAT1, STAT2 and PML, but also a wide variety of other factors. We counted for each TRF the total ChIP-validated interactions within *cis*-regulatory regions of clustered genes. The *cis*-regulatory interval was set to 2 kb either side of the 5' ends of protein-coding genes from the Ensembl gene collection (v.65) (Kinsella *et al.* 2011). The significance of observed numbers of binding sites for each factor per gene cluster was measured by modelling the number of binding sites as a Poisson variate, with mean defined by the genomic average number of binding sites per gene. After correcting for multiple tests, we find 136 significant interactions between a TRF and a gene expression cluster. Of these, 131 (96%) interactions were between TRFs and clusters 1 - 7 from the eight functionally enriched clusters in Figure 5.3, with cluster 8 having no significant enrichment in binding sites for any of the 75 TRFs. The complete set of significant interactions between TRFs and clusters 1 - 7 is shown in Figure 5.6A, with rows ranked from the most to least connected TRFs. Since clusters 5 and 6 reflect a very similar collection of genes, in RM and AGM respectively (Figure 5.2), these are enriched for exactly the same collection of TRFs (STAT1, STAT2, BATF, IRF4, JUND and GATA2).

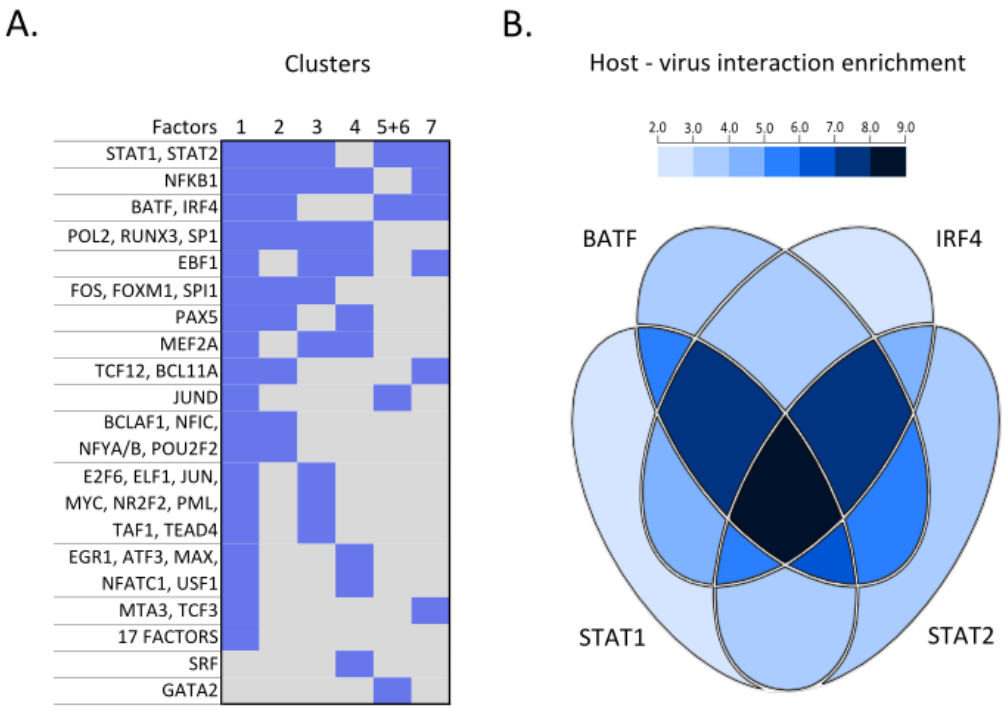


Figure 5.6. Relationships between transcriptional regulators and co-expressed collections of genes.

A. Enrichment of TRF binding sites in *cis*-regulatory regions of clustered genes. The 17 TRFs uniquely targeting cluster 1 are listed in Table 5.3 below.

B. Percentages of host–virus interaction (HVI) related genes within the common target sets of the factors STAT1, STAT2, BATF and IRF4.

Table 5.3. Factors enriched uniquely within *cis*-regulatory regions of genes in cluster 1

Factor	Poisson p-value	Factor	Poisson p-value
ATF2	$2.0E^{-15}$	SIN3A	$8.0E^{-7}$
CBX3	0.003	SIRT6	0.0002
CEBPB	$2.0E^{-6}$	SIX5	$1.0E^{-7}$
CREB1	$1.0E^{-13}$	SMARCA4	0.002
E2F4	$1.0E^{-5}$	SP2	0.0006
ETS1	$5.0E^{-7}$	STAT5A	$<1.0E^{-16}$
GABPB1	$9.0E^{-16}$	YY1	$<1.0E^{-16}$
GTF2B	0.0017	ZBTB33	0.0027
RXRA	$7.0E^{-5}$		

P-values are calculated using the Poisson distribution, as described in the main text, using the genomic mean number of binding sites per protein-coding gene for each factor, and the observed number of binding sites per gene within cluster 1.

From Figure 5.6A, it is clear that some clusters are significantly enriched with binding sites from many different TRFs (e.g. cluster 1; see also Table 5.3) while others are selectively enriched for only a few TRFs (e.g. clusters 5 and 6). This is likely to reflect differences in the activity of the clustered genes across experimental conditions in general, with highly active cell-cycle related genes in cluster 1 being sampled much more frequently by ChIP-seq experiment. If the total number of links between factors and clusters is controlled for, by shuffling TRF – gene connections using a degree-preserving randomization algorithm (see methods), we find that most of the non-specific links to cluster 1 are no longer significant, while links between specific factors and clusters 5 – 7 are retained ($p < 0.05$, corrected for multiple tests). Thus, clusters 5 – 7 have particularly robust connections to their respective enriched TRFs. The TRFs with significant interactions to the most clusters are STAT1 and STAT2, with significant numbers of binding sites over 6 from the 7 functionally enriched and highly targeted gene clusters ($p < 10^{-15}$). These factors heterodimerize in response to IFN-α or IFN-β in order to drive expression of genes harbouring a specific regulatory sequence, termed the interferon-response element (ISRE) (consensus: AGTTTCNNTTTCNY), within their promoter regions (Ghislain, Wong *et al.* 2001). We downloaded genome-wide locations for the ISRE motif conserved across human, mouse and rat, using the 'tfbsConsSites' table from the hg19 TFBS Conserved track in the UCSC Table Browser (Karolchik *et al.* 2003; Karolchik *et al.* 2004). Consistent with ChIP-seq data, copies of the ISRE motif are significantly enriched within the *cis*-regulatory regions of genes in clusters 5 – 7 ($p < 10^{-15}$, by Poisson test). This suggests that the dimerization of STAT1 specifically with STAT2 is important in mediating cytokine signalling within CD4+ T cells in response to SIV infection.

Three further TRFs, NF-κB1, BATF and IRF4, are enriched over a majority (5 / 8) of functionally enriched clusters, and from these, BATF and IRF4 are also enriched over the species-specific clusters 5 and 6. From the microarray data, expression levels of NF-κB1 are relatively constant within CD4+ cells from the two species (data not shown). By contrast, the expression of BATF in AP increases more than 4-fold during the first day of SIV infection, but by day 115 (p.i.) declines to less than half its expression level pre-infection. In MP, BATF level rises slowly throughout the first 115 days (p.i.), to roughly three times its pre-infection level. The microarray probe for IRF4 showed significant (> 32-fold) variations in expression, but these were as large and as regular before infection as after infection, suggesting a technical fault with the IRF4 probe. For BATF and IRF4, the conjunction of these factors over shared gene sets is consistent with the recently discovered function of BATF as a facilitator of IRF4 binding to target DNA sequences in T cells of both mice and humans (Li, Spolski *et al.* 2012; Tussiwand *et al.* 2012). Finally, we note that the promyelocytic leukemia protein (PML) is enriched only over

more generically enriched clusters (1 and 3), alongside many other factors (E2F6, ELF1, JUN, MYC, TAF1, etc). Thus, this factor shows a stronger preference for clusters of genes enriched in cell cycle regulators, rather than those involved in mediating the anti-viral response.

We next considered evidence for cooperative regulation of the host–virus interaction by sets of factors enriched over viral–response related clusters. The most significant contrast between gene expression patterns in RMs and AGMs was detected for viral response genes within clusters 5 and 6. From Figure 5.6A, there are 6 TFs enriched over these gene sets (STAT1, STAT2, BATF, IRF4, JUND and GATA2). We downloaded from UniProt a collection of genes corresponding to the SP_PIR keyword 'host–virus interaction', which we will refer to as HVI genes, with 332 of these matching gene IDs from the Ensembl database (v.65) (Kinsella *et al.* 2011; Magrane and Consortium 2011). We then calculated all possible overlaps between collections of genes bound by each of the 6 TFs enriched over clusters 5 and 6. The percentage of HVI targets was found within each TF target set overlap, with statistical significance assessed using the hypergeometric distribution. Full results are provided in Table 5.4.

Almost all intersections between STAT1, STAT2, BATF and IRF4 are significantly enriched for HVI genes, compared with very few for JUND and none for GATA2 ($p < 0.05$, corrected for multiple tests; see Table 5.4). The percentage of HVI genes within target set overlaps increases to a maximum when all four of the factors STAT1, STAT2, BATF and IRF4 are bound (Figure 5.6B and Table 5.4). There are 69 genes bound by all four factors, of which 7 (8.9%) are annotated as a part of the host–virus interaction (CFLAR, SP100, MAPK1, PSMB3, SYNCRIP, STAT3, and TAP2) (Full Gene list: Supplementary Table S5.3). This reflects a more than 5-fold enrichment compared to the frequency of HVIs in the genome (1.7%). We also note that the HVI annotation set is incomplete, since unannotated anti-viral factors (e.g. HERC6, IFITM1, MX2) were also found within the set of 4-way shared targets. Interestingly, heteromeric relationships between STAT3, IRF4, and the bZIP factor BATF have recently been implicated in differentiation of the pro-inflammatory T cell subset Th17 in mice (Ciofani, Madar *et al.* 2012; Yosef, Shalek *et al.* 2013). The present data may therefore lead to a generalization of CD4+ T cell specific STAT – IRF4 – BATF regulatory modules, in which alternative STAT family members interact with BATF and IRF4 to modulate gene dynamics within CD4+ T cells.

Table 5.4. Predicted target set overlaps of STAT1, STAT2, BATF and IRF4

Factors in overlap	Bound genes	Bound HVIs	% HVI targets	P – value (hyp.geom.)	Rank	P – value (corrected)
STAT2	2950	100	3.39	5.31E-13	1	2.98E-11
STAT1,STAT2	1698	62	3.65	2.72E-09	2	7.62E-08
STAT1	3235	95	2.94	7.04E-09	3	1.31E-07
STAT1,STAT2,IRF4	543	29	5.34	3.68E-08	4	5.15E-07
STAT2,IRF4	783	36	4.60	3.70E-08	5	4.15E-07
STAT1,IRF4	796	36	4.52	5.59E-08	6	5.22E-07
IRF4	2091	59	2.82	3.42E-05	7	2.74E-04
STAT2,BATF	244	14	5.74	6.27E-05	8	4.39E-04
STAT1,BATF	236	13	5.51	1.71E-04	9	1.06E-03
STAT1,STAT2,BATF	149	10	6.71	2.00E-04	10	1.12E-03
STAT2,BATF,IRF4	125	9	7.20	2.45E-04	11	1.25E-03
STAT1,STAT2,BATF,IRF4	79	7	8.86	3.42E-04	12	1.60E-03
STAT1,BATF,IRF4	113	8	7.08	6.04E-04	13	2.60E-03
BATF	1066	32	3.00	9.28E-04	14	3.71E-03
BATF,IRF4	565	20	3.54	1.31E-03	15	4.90E-03
BATF,IRF4,JUND	191	8	4.19	1.48E-02	16	5.18E-02

Binding sites for STAT1, STAT2, BATF, IRF4, JUND and GATA2 were identified in intervals from -2 kb to + 2 kb around transcription start sites for 19,975 protein-coding genes in Ensembl v.65. The input binding site collections were calculated from ChIP-seq experiments by the ENCODE consortium, and obtained from the UCSC genome browser (HAIB and YALE TFBS)(Karolchik *et al.* 2003; Karolchik *et al.* 2004; 2011). Attention was restricted to cell lines most similar to the T cell system examined (K562, K563 and GM12878 lines). Host-virus interaction (HVI) genes were downloaded from Interpro and 332 HVIs matched to Ensembl protein coding IDs, using gene identifier conversions from Biomart (Kinsella *et al.* 2011). For all possible overlaps of the 6 transcription factors examined we then compared the total number of bound genes to the number of bound HVIs. The percentage of bound genes that are HVIs is shown in the third column. A hypergeometric p-value was calculated to measure the significance of the numbers of bound HVIs. This was then corrected using the Benjamini-Hochberg multiple testing correction (Benjamini *et al.* 2001). Since factor overlaps are inclusive, p-values for rows containing factor groups that are subsets of other rows are positively correlated. This means that the multiple-testing correction used here is stringent.

5.3.4.3 Expression patterns of families of transcription factors

We considered whether particular families of functionally related TRFs are more likely to be perturbed following SIV infection of either AGMs and RMs. The complete collection of TRFs was divided into 34 overlapping categories according to their structural domains, which define the DNA-binding and dimerization preferences of TRFs (Supplementary File S5.8). For every gene on the microarray we then calculated a t-statistic to measure its expression change between all available pairs of time points in each of the four CD4+ T cell sources. We then counted the numbers of TRFs within each structural family lying within the extreme 2.5%, 5%

and 10% tails of the t-distribution at each time point within each source. The percentage cut-off used does not ranks of the key structural families (Supplementary Table S5.4), so we use the 5% cut-off as standard in the text. Statistical significance of the observed number of expression changes per TRF family was then modelled using the Poisson distribution, with an excellent fit to results obtained by other methods, such as randomization test. Results are shown for a selection of the most and least variable TRF families in Figure 5.7A (full results of simulations are provided in Supplementary Table S5.4).

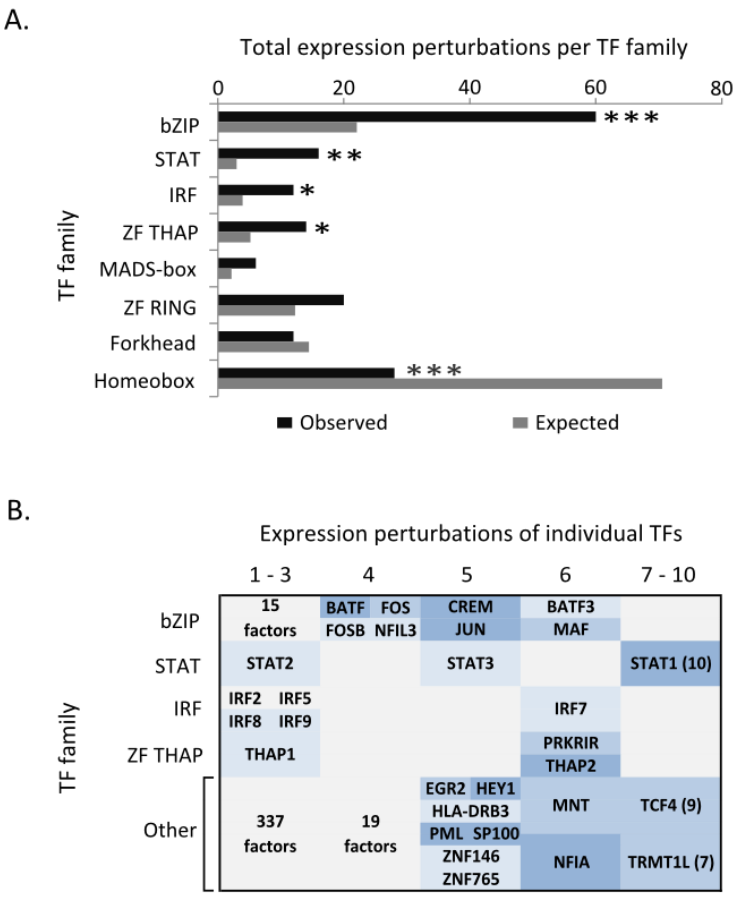


Figure 5.7 Expression perturbations within families of transcription factors following SIV infection.

A. Numbers of perturbations in expression of members of a subset of TRF families between days -8 and 65 (p.i.) across all four CD4+ sources. Perturbations were defined using the t-statistic distribution at each time point and within each source (main text), and the expected number calculated by the mean across simulations (methods). Statistical significance (two tailed) is indicated by asterisks: *** - $p < 10^{-6}$; ** - $p < 10^{-4}$; * - $p < 10^{-2}$. The 6 most frequently perturbed TRF families are shown, of which 4 are significant, together with the least frequently perturbed family (Homeobox).

B. Expression perturbations of individual TRFs. Expression perturbations of TRFs were calculated as in (A), and then counted for each factor. Statistical significance was assessed using a Poisson model. The factors were then grouped according to family and number of perturbations.

We find that the bZIP, STAT, IRF, and Zinc finger THAP families are significantly more likely to be found within the tails of the expression change t-distributions, compared to randomly selected TRFs ($p = 7.6 \times 10^{-10}$; 2.0×10^{-6}; 9.3×10^{-3}; 8.6×10^{-3}, respectively, corrected for multiple tests). This is also true within each of the four CD4+ T cell sources individually (data not shown). Thus, each of these families is perturbed within CD4+ T cells significantly more often than expected following SIV infection. By contrast, for example, the body-pattern regulating family of HOX genes are significantly depleted within the tails of the expression change t-distributions. Thus, changes in expression of transcriptional regulators in response to SIV infection are enriched for specific structural families.

We then identified individual TRFs most responsible for these family-wide enrichments. The total numbers of expression changes were counted for each TRF in turn, and compared to expected numbers of changes per TRF, once again using the Poisson distribution. After correcting for multiple tests, 43 TRFs were found to be perturbed significantly more often than expected by chance ($p < 0.05$). Each of these TRFs was detected four or more times within the extreme 5% of the t-distribution for particular combinations of time point and CD4+ T cell source (Figure 5.7B). Many of the factors correspond to those identified within gene expression clusters, or as enriched in clustered target sets. These include: the most frequently perturbed factor, STAT1, together with STAT2; components PML and SP100 of a multifunctional, homologous recombination regulating, complex termed the PML (or nuclear) body; the interferon regulatory factor IRF7 detected in a number of the expression clusters; and the bZIP factors BATF, BATF3 and CREM. In addition to these, we identified: (i) A further STAT protein (STAT3), with several relatively weak perturbations in both AGMs and RMs, (ii) bZIP components of the cell growth regulating AP-1 complex, Jun, Fos, and FosB, where Jun may alternatively be repressed by its dimerization partners BATF and BATF3, or facilitate BATF – IRF4 interactions (Li *et al.* 2012) (iii) 5 of the 9 interferon regulatory factors in primates (IRF1, 2, 7, 8, 9), with most frequent expression changes for the ISG-promoting factor IRF7, and (iv) THAP1, THAP2 and PRKRIR from the zinc finger THAP-domain family. THAP1 is sometimes found within PML bodies , providing additional evidence for a role for this complex in mediating the CD4+ T cell response of simian species to SIV infection (Roussigne, Cayrol *et al.* 2003).

Expression patterns for a number of these regulators (IRF7, PML, THAP2, BATF, BATF3, CREM) are provided in Figure 5.8.A-G. Each figure shows expression changes in AGMs and RMs side by side, with the interquartile ranges in expression within each species indicated by the error bars. For IRF7, expression patterns are shown for both CD4+ T cell subpopulations (PB and LN) (Figure 5.8.A-B). In other cases, we selected PB or LN expression patterns according to which

show the greatest contrasts between species (Figure 5.8.C-F). CREM expression was perturbed significantly in different CD4+ subpopulations according to species (Figure 5.8G). All available time points for the selected CD4+ T cell sources are shown. In several of these examples (e.g. IRF7 and PML in LN; THAP2 and BATF in PB), expression patterns reflect expression patterns for STAT1 and ISGs (Figure 5.4), with a transient increase in expression within AGMs, but a much more sustained increase within RMs. For BATF3 in LN, expression is likewise much higher in RMs than in AGMs, though without any initial increase in expression of this factor within AGMs (Figure 5.8F). Thus, RM CD4+ T cell populations are characterized by upregulation of mRNAs in the chronic phase of infection, relative to AGMs.

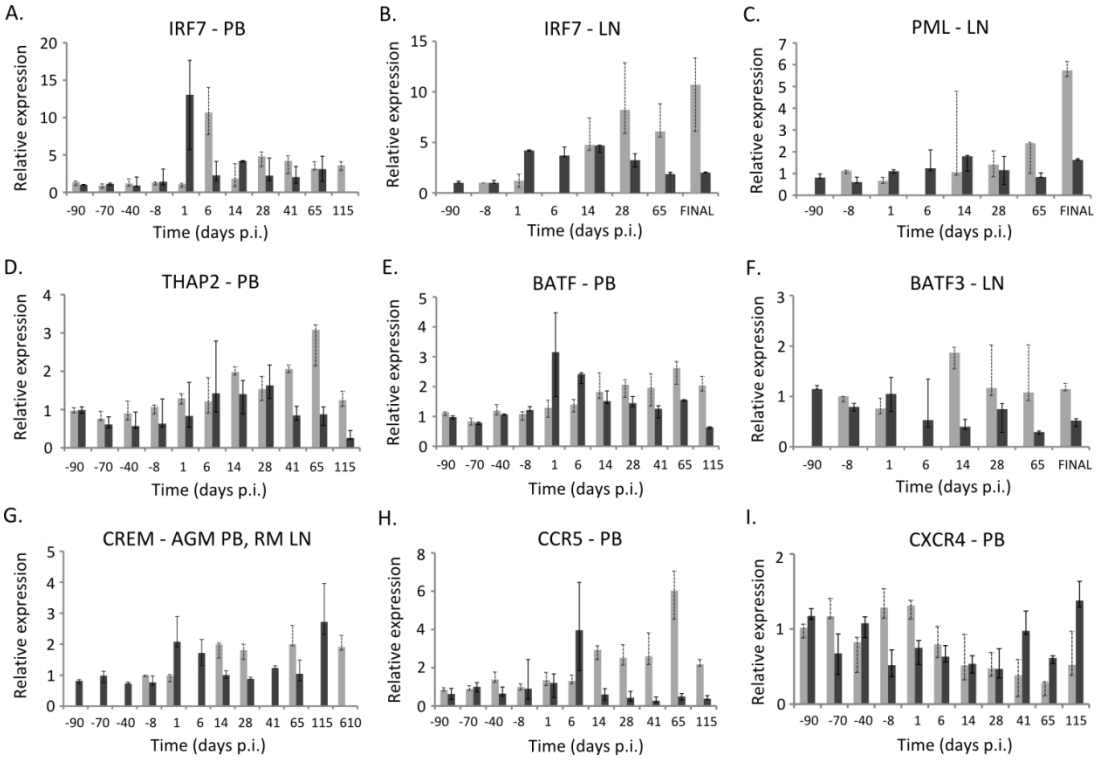


Figure 5.8. Expression patterns of selected perturbed genes within AGMs and RMs.

Expression levels of selected genes within AGMs (black) and RMs (light grey) at all available time points. Expression was normalized within each animal to the level at the first sampled time point (day -90 in AGMs and in RM peripheral blood, and day -8 within RM lymph node). Error bars reflect the interquartile ranges of relative expression across all available animals (≥5) and probe sets for the gene (≥1) at the given time point.

The bZIP factor CREM displays a strong increase in expression between days 65 and 115 (p.i.) in AGM PB (Figure 5.8G). Since the CCR5 and CXCR4 receptors are critical to the accessibility of CD4+ cells to viral incursion, the expression dynamics of these receptors in PB are shown

alongside those of CREM (Figure 5.8.H-I). As noted earlier CREM repressively interacts with the promoter region of the CCR5 cytokine receptor gene (Banerjee *et al.* 2011). Thus, CREM is a candidate to explain the suppression of the CCR5 receptor mRNA observed in AGMs relative to RMs by day 115 (p.i.). The dynamics of the CCR5 receptor mRNA are also consistent with this gene being recognised as an ISG (Hariharan, Douglas *et al.* 1999). Interestingly, there is a noticeable reversal of these dynamics for the CXCR4 receptor mRNA in PB CD4+ cells, with a transient drop in CXCR4 expression throughout the acute phase of infection in both AGMs and RMS, and an eventual recovery to a higher than pre-infection level in AGMs. Importantly, the CXCR4 receptor is almost entirely impervious to SIVagm entry (Marx and Chen 1998). This might therefore reflect an adaptive measure by the AGM CD4+ T cell population towards a patterning of cell surface receptors that is less receptive to the entry by the SIVagm family. Regardless, the species-specific changes in CREM, CCR5 and CXCR4 expression point towards links between the bZIP family of TFs and species-specific regulation of the SIV lifecycle.

5.3.5 Interactions of viral proteins with differentially expressed host genes

We finally investigated connections between viral proteins and host genes, in order to suggest initial steps contributing to differential immune response between AGMs and RMs. We examined viral protein–host gene regulatory interactions in humans (described as 'upregulates', 'downregulates', or without indication of a regulatory sign, as 'regulates'), taken from an HIV-1 human protein interaction database (Ptak *et al.* 2008). We counted the number of known interactions with a viral protein across networks of interacting partners with the host gene, and calculated the statistical significance of overrepresentation by simulation (See methods). Statistically significant network interactions are illustrated in Figure 5.9. The majority of regulatory interactions with these host proteins involved the Tat protein, in both species, and in both tissue compartments. A number of species and tissue-specific interactions were observed, for example, Tat's potential interactions with MX1, IRF7, and STAT1 in RMs but not in AGMs, and Tat's interaction with HSPA1A in AGM LN only. In humans, HIV Tat protein has been shown to have a synergistic effect, together with JAK, STAT and MAPK proteins, to activate expression of the toxic cytokine CXCL10 in astrocyte cells, contributing to HIV-associated dementia (Williams, Yao *et al.* 2009). In addition, Tat has recently been shown to engage the IRF7 gene promoter in CD4+ macrophages, leading to increased STAT1 phosphorylation and activation of an ISG cascade (Kim, Kukkonen *et al.* 2013). A functional relationship between SIVagm Tat and AGM or RM STAT1 within CD4+ T cells therefore seems highly likely, given the abundance of inferred connections from Tat and STAT1 to the same sets of co-expressed genes.

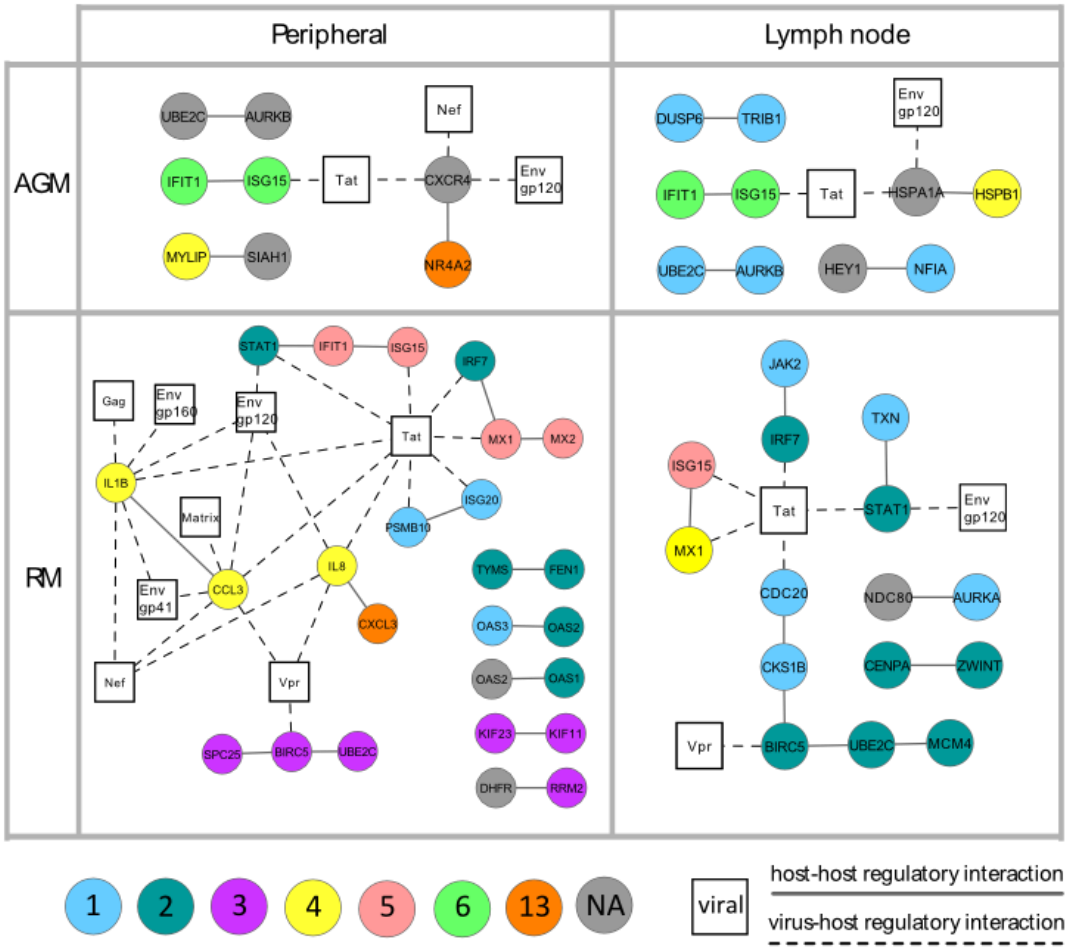


Figure 5.9. Gene regulatory networks by tissue type and monkey species.

Genes are represented by nodes. Round nodes represent cellular genes and square nodes viral genes. Cellular genes are colour coded according to the expression profile from which they are derived (see key). Nodes coloured grey (NA) signify cellular genes with a significant interaction with HIV or cellular proteins, but which did not cluster with any other genes. Edges represent gene regulatory relationships between pairs of host genes (solid line), or HIV proteins and host genes (dashed line).

Tat's potential network interactions in humans were significant in all samples for only one differentially expressed simian gene, ISG15. This chemokine was previously examined, within the species-specific expression patterns illustrated in clusters 5 and 6 in Figures 5.3 and 5.5. Indeed, across the microarray platform as a whole, and within all four CD4+ T cell sources, ISG15 ranks within the top three most consistently and significantly perturbed genes (See Methods). ISG15 is an ubiquitin-like protein modifier, which is believed to have an anti-viral function in response to HIV infection (Takeuchi and Yokosawa 2008). It is possible that Tat activates ISG15 expression in RMs, and subsequently evades an antiviral mechanism of ISG15. Alternatively, ISG15 may contribute to a generalized and sustained immune system activation

within RMs, leading to a larger pool of CD4+ cells which furthers the proliferation of the virus itself. Two additional RM-specific network interactions were those between Env and STAT1, and between Vpr and survivin (BIRC5). The second case serves as validation of our network-based statistical methods, since up-regulation of survivin by HIV-1 Vpr has been experimentally demonstrated (Zhu, Roshal *et al.* 2003). Of particular interest, given its unusual expression dynamics, is the interaction between the host CXCR4 receptor within AP, and the three HIV proteins Tat, Nef and Env. This indicates there are many genes perturbed within AP that are related to both CXCR4 and to Tat, Nef and Env. Finally, cross-referencing all significant interactions between host and viral proteins against our 15 gene expression clusters, we found that seven expression profiles were significantly enriched ($p < 0.05$) (Table 5.5). As for individual genes, the most significantly enriched expression profiles were categorised as 'up-regulated by Tat', in profiles 2, 4, 5 and 6 (Table 5.5). The greatest significance of enrichment in Tat interactions was detected for the RM and AGM- biased clusters 5 and 6, respectively, reflecting once again the importance of this group of genes in mediating the responses of both AGMs and RMs to SIV infection. We conclude that the available evidence from literature suggests that CD4+ T cells are most responsive to factors that interact with Tat, and that these interactions may pinpoint initial events in the species-specific AGM and RM anti-viral responses.

Table 5.5. Enrichment for virus protein interactions among expression profiles.

Cluster	Virus interaction	Number of genes	Rate of enrichment	Corrected P-value
2	Tat up-regulates	12	3.2	2.9×10^{-3}
4	Tat up-regulates	9	4.0	2.9×10^{-3}
5	Tat up-regulates	10	7.0	2.2×10^{-6}
6	Tat up-regulates	7	7.3	9.9×10^{-5}
7	Tat up-regulates	8	3.4	0.022
1	Gag binds	3	10.0	0.040
9	Nef up-regulates	2	23.5	0.044
2	Vpr inactivates	2	18.5	0.044
7	Matrix is stimulated by	2	17.9	0.044
7	Vpr competes with	2	17.9	0.044

Profiles are listed in order of significance of interaction with a viral protein, from most to least significant. Rate of enrichment is the ratio of observed to expected numbers of genes intersecting between a profile and the network of interactions with the viral protein (see methods).

5.4 Discussion

We have utilised an unsupervised gene expression clustering algorithm to extract collections of genes with closely associated expression pattern in a given CD4+ cell sample. Our results demonstrate that unsupervised expression profile clustering is sufficient to detect species-specific patterns of gene expression that are consistent with CD4+ T cell dynamics in response to SIV infection. This allows us to precisely compare and contrast natural and non-natural host species' gene expression programs following SIV infection. These expression patterns cover a variety of processes, including activation of ISGs in both species, early viral response in both AGMs and RMs, and sustained, although fluctuating, activation of cell cycle genes and ISG expression in RMs.

The initial activation of the innate immune response involves interactions between a pathogen and immune cell sensors, including dendritic cells and macrophages (Kushwah and Hu 2011). In HIV or SIV infection, CD4+ T cell subtypes are directly infected (Marx and Chen 1998; Wang *et al.* 2013) and thus perhaps also can act as early immune system anti-viral sensors. Viral proteins with statistically significant network interactions with host cell proteins include the viral transcription factor Tat, and envelope glycoproteins, Nef and Vpr (Ayyavoo, Mahboubi *et al.* 1997; Conti, Fantuzzi *et al.* 2004; Qiao, He *et al.* 2006). The viral Tat gene in particular has been suggested to play a central role in perturbing the host immune response, particularly through dysregulation of host cytokine signalling pathways, including type I interferons and regulators of inflammation, such as interleukins and TNF-α (Li, Yim *et al.* 2010). Furthermore, expression of HIV-1 Tat in dendritic cells can induce expression of many interferon-inducible genes (Izmailova, Bertley *et al.* 2003). Here, the expression clusters most enriched for the cellular factors altered by HIV-1 Tat protein are clusters 5 and 6, which we identified as enriched in viral-response genes. This further emphasizes that immune dysregulation in both pathogenic and non-pathogenic infections is likely to involve Tat activity. In addition, a key transcription factor that may be activated by Tat is STAT1, which together with STAT2, showed the greatest relative rate of targeting of viral response genes. Thus, multiple lines of evidence converge upon viral response genes within clusters 5 – 6, and also cluster 7 within AP specifically, as critical effectors of the acute response to SIV infection.

We found significantly more differentially expressed TFs in AGM peripheral blood than expected by chance, much more so than in AGM lymph node. The AP CD4+ cells respond remarkably rapidly to SIV infection, much faster than in MP. Although it is likely to be in lymph node that the chronic state of the system is defined, since 40% of T cells are found there,

compared to only 2% in PB (Di Mascio, Paik *et al.* 2009), nevertheless the initial infection of T cells occurs peripherally. This can be either in peripheral blood or genital mucosa, according to route of transmission, and then some of these peripheral CD4+ cells will migrate to lymph nodes (Ding, Xu *et al.* 2012). It is plausible that a rapid response to SIV infection within AGM peripheral blood may lead to altered T cell dynamics with lymph nodes at later time points. Indeed, considering RMs, we found evidence that the patterns of ISG expression during the first 14 days might be related to expression dynamics of cell-cycle related genes at later time points. However, this finding is tentative, since a much larger sample of individuals would be needed to establish that there are subtypes of monkeys within a population showing altered early ISG and later correlated cell cycle perturbations. Regardless, it is very notable that peripheral blood CD4+ cells in RMs react much more slowly than in AGMs yet eventually fail to control the infection, leading to a chronically proliferating T cell population and eventual exhaustion of the RM immune system. We speculate that CD4+ T cells within AGM peripheral blood act either as sentinel or signal amplifying cells, to ensure an efficient recovery from the initial assault of the virus upon the host immune system.

Our results relating to transcriptional regulators indicate that significant differences exist between AGMs and RMs, and we surmise that these differences are related to the contrasting responses between the natural and pathogenic host species. We find evidence for a role for several components of PML bodies (PML, SP100, THAP1), which could mediate dsDNA sensing, and recombination of host and viral genome, as well as regulating the interferon-stimulated response (Roussigne *et al.* 2003; Choi *et al.* 2006; Bernardi and Pandolfi 2007; Geoffroy and Chelbi-Alix 2011). We have clarified that among the possible dimerization partners of STAT1, STAT2 and perhaps STAT3 are most likely to play an important role during SIV infection. We also identified a significant enrichment over the targets of STAT1 and STAT2 in binding sites for BATF and IRF4, lying upstream of the viral-response genes that are co-expressed in a species specific way. The overlap between bound genes from all four of these factors is small but includes the PML body component SP100, together with the critical cell-proliferation regulator MAPK1 and T cell differentiation regulator STAT3 (Bernardi and Pandolfi 2007; Ciofani *et al.* 2012; Yosef *et al.* 2013). The factor BATF appears to be part of a wider network of bZIP factors and this is notable since bZIP factors bring about transcriptional regulation through first dimerizing with one another. Significant regular changes in expression were detected for several basic leucine-zipper (bZIP) factors (e.g. MAF, BATF, BATF3, CREM, JUN, FOS), and interferon regulatory factors (IRFs). We mention in particular upregulation of both BATF and BATF3 within RMs, compared to a transient up-regulation within different CD4+ subpopulations in AGMs. Due to the exchangeability of the BATF, BATF2, and BATF3 homologs

in binding to AP-1 factors including Jun, and to either IRF4 and IRF8 (Tussiwand *et al.* 2012), there is evidently scope for many of the bZIP family members to heterodimerize and thereby to regulate the DNA binding profile of IRFs in response to cytokines (Li *et al.* 2012). Since no ChIP-seq data was yet available for BATF3, our study is limited regarding the relative contributions of BATF and BATF3 to gene expression dynamics in different CD4+ T cell contexts.

Interestingly, BATF is increasingly associated with inflammatory T cell kinetics in a range of contexts, including within exhausted T cells in HIV (Quigley, Pereyra *et al.* 2010) and in chronic hepatitis infection in humans (Moal, Textoris *et al.* 2013), within CD4+ T cells differentiating towards the pro-inflammatory Th17 lineage (Ciofani *et al.* 2012; Yosef *et al.* 2013), and also within a severe lymphoproliferative disorder within mice (Logan, Jordan-Williams *et al.* 2012). Our identification that BATF homologs and binding partners are significantly perturbed and in a species-specific manner within natural and pathogenic SIV host species is thus intriguing. In particular, we have shown that BATF and BATF3 are both up-regulated within RMs compared to AGMs, well after the end of the acute period of SIV infection. Clarifying relationships of cause and effect between BATF homologs and other bZIP factors within these model animal systems may not feasible within the context of a genome-wide transcriptional study. Nevertheless, the most coherent features that we have identified within an extensive sample of genome-wide datasets point towards these regulators as important in modulation of the viral response within AGMs and RMs.

5.5 Conclusions

Our work extends earlier studies which have investigated differential gene expression programs between natural and pathogenic monkey hosts of SIV infection, and between rapid progressor and viremic non-progressor human hosts in HIV infection (Bosinger *et al.* 2009; Jacquelin *et al.* 2009; Lederer *et al.* 2009; Rotger, Dalmau *et al.* 2011). As in these studies we have found high initial upregulation of ISGs in both natural and pathogenic cases, followed by marked attenuation in the natural host (AGMs here). We have also identified modules of cell cycle and T cell proliferation genes which are specifically upregulated in the pathogenic host (RMs here). Here we aimed for the first time to identify regulatory networks and factors that could regulate them. Network analysis shows that viral response genes, especially Tat, display significant interactions with STAT1. A significant overlap between binding sites of STAT1 and STAT2, and BATF and IRF4, was detected over collections of viral-response genes with species-specific expression patterns. More broadly, our data suggest that regulation by bZIP domain containing factors is a central component in the response of simian CD4+ T cells to SIV infection. Our study provides novel biochemical and genetic reference points for understanding the progression to disease following SIV and HIV infection.

Chapter 6

Discussion

The material presented within this thesis relates to the regulation of gene expression within eukaryotic organisms, focusing mainly on primates. In various settings, I have examined distributions of transcription factor binding sites mapped to the human genome, and microRNA target sites mapped to the human protein-coding transcriptome. Transcription factor binding site patterns were compared to many other genomic features, including the locations of microRNA and protein-coding genes. Expression patterns of the transcribed products of these gene families were considered across human tissues, and related to the numbers of bound regulators, and distributions of regulatory chromatin modifications, within protein-coding promoter regions. What began as a trickle of genome-wide data became a torrent during the time period of the study, mostly generated by member institutions of the ENCODE consortium (2011). The total amount of human raw data from ENCODE alone is now in excess of 30 TB, exceeding the storage capacity of a typical university work station (Edgar, Domrachev *et al.* 2002; see also http://genome.ucsc.edu/ENCODE/FAQ/). The study here was carried out in time to capture around 8% of the more than 1500 known transcription factors in human (Vaquerizas *et al.* 2009). Now, more than 10% have been sampled in at least one human cell line (2011). Deep-sequencing of whole cell small RNA libraries has led to detection of over 1000 hitherto unknown microRNA genes in human during the past 4 years (Griffiths-Jones *et al.* 2006; Kozomara and Griffiths-Jones 2011). This thesis therefore reflects the momentum in biological sciences towards genome-wide integrative analysis, and this type of research may reflect the future of cellular biology in general.

Each of the research chapters can be taken as a standalone piece of research. The aims of this chapter are to consider themes in common to much of this research, and to reflect upon the strengths and limitations of the research as a whole.

6.1 Transcription and microRNA gene birth

We have proposed that the numbers of regulators of protein-coding genes is a significant selective factor driving the origin of microRNA sequences within and near to these genes. We have shown that microRNA sequences are often associated with highly regulated protein-coding genes, which can be either their hosts or their neighbours along chromosomes. This is consistent with a higher rate of *de novo* microRNA gene birth or retention within highly regulated gene regions. MicroRNAs downstream from protein-coding *cis*-regulatory regions

are more likely to target the transcriptional regulators bound to these regions. This is true for intragenic microRNAs, supporting the model of co-transcription of these microRNAs and their host transcripts, from a shared promoter region (Baskerville and Bartel 2005). It is also true for intergenic microRNA genes in locations consistent with transcription from a protein-coding gene promoter region, either upstream in an antisense orientation, or downstream in a sense orientation. Transcriptional regulators bound to protein-coding promoter regions in these cases could co-regulate the expression of both the pre-mRNA and pri-miRNA transcriptional products.

This enrichment of ChIP-seq peaks for transcriptional regulators upstream of microRNA genes was consistent with our expectation, since, based upon much smaller samples of genes and regulators, computationally predicted TF binding motifs were shown to be enriched within upstream regions of microRNA genes (Yu *et al.* 2008). Nevertheless, this is the first time that experimental data has been used to support this conclusion. We also demonstrated that the property of enrichment within *cis*-regulatory regions applies individually to the majority of general DNA binding, sequence-specific, and chromatin modifying factors. Previously, transcriptional activity has been considered as a pre-requisite for the birth of new microRNA genes, by providing a ready source of transcription for randomly evolved hairpin sequences (Baskerville and Bartel 2005; Berezikov 2011). Here, we have provided evidence for this idea, by demonstrating that the sum of transcriptional activators is significantly positively related to expression levels of both mRNA and microRNA transcripts lying downstream, and to the frequencies of microRNA genes within these regions.

6.2 MicroRNA gene regulation depends upon the protein-coding context

Previously, regulatory effects of flanking genes on intergenic microRNAs, and the host genes on intragenic microRNAs, have been largely overlooked (Shalgi *et al.* 2007; Tu *et al.* 2009; Cheng *et al.* 2011). Here, we took into account the protein-coding gene context of upstream regulation of intragenic and intergenic microRNAs. The analyses in chapters 2 and 3 indicate many layers of interaction between protein-coding gene regulation by transcription factors and chromatin-modifying enzymes, and the distributions and expression levels of associated downstream microRNA genes.

6.2.1 Characterization of microRNA host genes

Our study adds to a literature which has attempted to characterise the microRNA host gene family as a whole e.g. (Rodriguez *et al.* 2004; Baskerville and Bartel 2005; Kim and Kim 2007;

Saini *et al.* 2007; Morlando *et al.* 2008; Saini *et al.* 2008; Hoeppner *et al.* 2009; Golan *et al.* 2010; He *et al.* 2012; Meunier *et al.* 2013). It is intuitive that the longer a gene is, the greater the odds of hosting the precursor sequence of a microRNA hairpin. And in fact, microRNA host genes were previously confirmed to have three times the length of average protein-coding genes (Golan *et al.* 2010). Further, we observe that the number of bound transcriptional regulators is a variable function of gene length (Figure 2.2B). However, in chapter 2, we reject the hypothesis that microRNA host gene regulation can be accounted for in terms of the increased mean length of the host gene set.

We also considered whether numbers of splice variants and the ages of microRNA host genes might explain microRNA host gene regulation (Figure 2.2.C - F). Both of these properties are positively linked with the numbers of bound *cis*-acting regulators of the gene (Warnefors and Eyre-Walker 2011). Consistent with this, we showed that host genes are older than average, and have greater numbers of transcript variants. Even for host genes with few splice forms, an excess of *cis*-acting regulatory factors might also reflect a condition favouring microRNA gene birth, rather than a consequence of selection due to microRNA function. At minimum, the data support the view that *de novo* birth of microRNA gene sequences is favoured within gene regions bound by many transcriptional regulators, presumably due to a supply of host transcripts to serve as a vessel for an emergent functional microRNA precursor. The proportion of intragenic microRNA transcripts expressed from the 5'-most host gene promoter region is the subject of debate (Baskerville and Bartel 2005; Marson *et al.* 2008; He *et al.* 2012; Meunier *et al.* 2013). Our study favours the conventional view that the majority of intragenic microRNAs are co-transcribed with their host gene transcripts.

6.2.2 Intergenic microRNA clusters proximal to protein-coding genes

The material in Chapter 3 extends many of the ideas developed in Chapter 2, to additional subsets of microRNAs that are located outside but near to protein-coding genes. The most important findings are that over 30% of intergenic microRNA gene clusters in human are found within 10 kb of protein-coding genes, and that many of these are likely linked to the expression of the protein-coding gene via common transcriptional regulators. Although it is well known that both microRNAs and protein-coding genes are often clustered, the clustering of intergenic microRNAs with protein-coding neighbours has not been explicitly addressed before. We found that intergenic microRNA clusters were significantly more likely to lie downstream of protein-coding transcription start sites, on either strand. Distributions of microRNAs were considered separately at both 5' and 3' ends of genes, and on both the protein-coding sense and antisense DNA strands. There is a clear reversal in the dominant orientation of microRNAs

at the 5'-boundaries of genes, from the sense strand within a protein-coding gene, to the antisense strand just outside of the protein-coding gene. This result is conserved across species from human to invertebrates, and is reminiscent of deep-sequencing experiments showing an abundance of short antisense transcripts at gene ends (Carninci *et al.* 2005; Carninci *et al.* 2006; Taft *et al.* 2009). In general, the length distributions for short promoter- and gene end-associated transcripts do not precisely match the typical lengths of microRNA mature sequences (Taft *et al.* 2009). Thus, although the expression of short transcripts extending outwards from genes is well-known, the occurrence of microRNAs in these genomic positions has generally been overlooked. Indeed, the intergenic microRNA class has often, though not always (Barski *et al.* 2009), been explicitly separated from the microRNAs found inside protein-coding genes (Shalgi *et al.* 2007; Chien *et al.* 2011).

Here, we propose that the biology of intergenic microRNAs can often be connected with that of the neighbouring protein-coding genes. MicroRNAs that are proximal to protein-coding genes might be thought of as a distinct subclass, the biology of which is probably quite different to that of intergenic microRNAs located much more distally. The identification of links between intergenic microRNAs and neighbouring protein-coding gene promoter regions is not entirely without a precedent, since in a study of promoter-associated histone marks, for 11 intergenic microRNA clusters the strongest upstream promoter signals were found to lie within promoter regions of protein-coding genes (Barski *et al.* 2009). However, we have identified significantly more (>80) microRNA clusters in human, in locations consistent with a direct transcriptional link to the protein-coding region lying upstream.

We found that intergenic microRNA clusters are preferentially located within active promoters compared to weak or poised promoters, and depleted within heterochromatin (Figure 3.2D). Significantly elevated numbers of transcriptional regulators were identified, bound to promoter regions of protein-coding genes near to intergenic microRNAs, especially when these lie in the 5'-antisense, and 3'-sense, orientations (with the language as described in Chapter 3) (Figure 3.5A). This serves to generalize the argument that microRNAs are more likely to be found within transcriptionally active, or highly regulated, regions of the genome. As for intragenic microRNAs, we also find a positive relationship between the expression level of the microRNA and the number of regulators associated to the nearby gene. Systematic analysis of microRNA expression patterns across tissues is still quite difficult, since more than half of the known microRNAs in humans have been detected within the past 4 years (Griffiths-Jones *et al.* 2006). Indeed, there is sometimes just one experiment in which the microRNA has been detected (Kozomara and Griffiths-Jones 2011). This also places some limitation upon the analysis of co-expression patterns of microRNAs, together with their host genes or neighbours.

Recent datasets based upon deep-sequencing of small and long RNA libraries in a range of human cell lines (2011) might provide an advance over the clonal microRNA frequencies and microarray-based expression values used here (Su *et al.* 2004; Landgraf *et al.* 2007).

6.3 Regulatory feedback between transcription factors and microRNAs

Given that microRNAs can target mRNAs expressed from TF genes, reciprocal regulation between TFs and microRNAs can occur. Indeed, this type of regulatory feedback is known to be of critical importance in many specific cellular contexts e.g. (Johnston and Hobert 2003; Hobert 2006; Chang *et al.* 2007; Burk *et al.* 2008). In Chapters 3 and 4, we considered the prevalence of regulatory feedback motifs between TFs and microRNAs, genome-wide. Previous studies have consistently detected significant rates of mutual TF/microRNA regulation, though based upon different and generally much smaller datasets (Shalgi *et al.* 2007; Yu *et al.* 2008; Gerstein *et al.* 2012). Our study extends earlier results in a number of other of ways. First, we find that feedback loops between TFs and microRNAs are predicted most often for TFs capable of repressing gene expression, giving rise to a double-negative feedback loop. Second, we identified a significant number of predicted feedback loops between the transcription factor gene targets of proximal intergenic microRNAs, and the microRNA targets of proximally bound transcription factors. This provides a strong indication that transcriptional regulators bound to the promoter regions of protein-coding genes can be actively involved in regulating transcription of primary intergenic microRNAs. Finally, a significant fraction of feedback loops involved TFs linked to intragenic microRNAs via their host genes' *cis*-regulatory regions. This provides a further line of evidence for the importance of the host promoter region in regulating expression of intragenic microRNAs, as argued in Chapter 2.

6.4 An integrated regulatory network of transcription factors and microRNAs

Transcription factors and microRNAs are also often found as regulatory partners of one another, with a significant enrichment in common target gene sets (Shalgi *et al.* 2007; Yu *et al.* 2008; Tu *et al.* 2009; Cheng *et al.* 2011; Gerstein *et al.* 2012). The transcriptional and post-transcriptional regulatory layers are often incoherent, in the sense that the TF is more often an activator, and the microRNA is almost exclusively a repressor (Herranz and Cohen 2010; Osella *et al.* 2011). We found that lower variation in mRNA transcript levels across tissues is associated with a greater incidence of feedforward regulation from TFs through microRNAs to common target genes. This property is consistent with theoretical arguments suggesting that the microRNA regulatory partners of TFs can act to reduce the levels of gene expression noise

within multicellular organisms (Herranz and Cohen 2010; Osella *et al.* 2011; Singh 2011). This theory has implications for the role of microRNAs in evolution, since the phenotype of a system in which transcriptional noise is buffered is evidently more robust to genetic mutation, which could allow greater genetic diversity to accumulate within species expressing microRNAs from their genomes (Peterson *et al.* 2009). However, the broader evolutionary interpretation is contentious, since it involves genetic selection acting at the level of a population, rather than an individual (Leigh 2010). It seems necessary that additional factors must underpin retention of new microRNA genes within the genomes of individual animals, including to repress co-expressed target mRNA transcripts. The two views might be made consistent, since the microRNA-mediated regulatory layer could evidently buffer transcriptional noise, in the context of a system that already possesses sufficiently many microRNAs to bring this about.

Although our study is consistent with this theory, there are more direct methods available to test the evolutionary argument. We might have asked, for example, whether the frequencies of TF - microRNA regulatory partnerships are directly related with the levels of protein-coding sequence diversity, between different species. The required analysis would not be straightforward, due to many other factors affecting nucleotide sequence diversity, including the recent demographic history of the species (Tajima 1989). Nevertheless, a model of protein-coding sequence diversity, in relation to the microRNA-mediated regulatory layer, and acting downstream of transcription factors, appears to be a natural starting point to test the evolutionary consequences of microRNA-mediated noise-buffering.

6.5 MicroRNAs in development

In our study of microRNA host genes, we found that microRNA host genes are enriched for developmental functions. This was unexpected, since analysis of earlier microRNA database releases did not find any statistically significant enrichment (Rodriguez *et al.* 2004). It is tempting to connect this finding with the well-established roles of many microRNAs within key developmental pathways (Lee *et al.* 1993; Johnson *et al.* 2003; Johnston and Hobert 2003; Brabletz *et al.* 2011; Guan *et al.* 2011). Indeed, this observation can be connected with the preceding remarks on roles for microRNAs in stabilizing mRNA expression levels. These properties may be linked, since the ability of microRNAs to buffer noise and to allow greater protein-coding genetic diversity to accumulate, could allow a greater range of organism morphologies to be reached through evolution. It has therefore been speculated that a period of rapid expansion of animal body-plans during the Cambrian era might be linked with the emergence of microRNA families at around the same time (Peterson *et al.* 2009).

We also found in Chapter 3 that intergenic microRNA clusters are associated with regions of the genome having higher densities of protein-coding genes (Figure 3.2C), and such regions are enriched in genes encoding developmental regulators. In Chapter 4, we then showed that feedforward regulation by microRNAs and TRFs is enriched over protein-coding gene targets within key biological processes, including cell growth, response to cytokines, transcription, and developmental pathways, including neurogenesis (Figure 4.8). Thus, we have identified a number of new ways in which microRNAs are connected genome-wide with developmental proteins. Many of the processes enriched in co-regulation by TF/microRNA pairs themselves regulate the expression and functions of microRNAs and TFs, for example, through the impact of cellular signalling pathways upon components of the microRNA biogenesis pathway (Saj and Lai 2011; Blahna and Hata 2013). This is analogous to the simpler two element TF/microRNA regulatory loop discussed previously, and indeed such regulatory feedback appears to be a pervasive characteristic of cellular systems, at all levels of their organization, see e.g. (Shalgi *et al.* 2007; Kosti, Radivojac *et al.* 2012). This observation clearly strengthens the case for integrative analysis, as cellular elements may not really exist in isolation in any meaningful sense.

6.6 Perturbations of regulatory networks

The property that signalling molecules return information to their upstream stimuli was further examined in the final chapter. After stimulation by interferons, signalling via the Jak-STAT cascade leads to expression of many genes with both repressive and activating effects upon upstream cytokine signalling pathways (Darnell *et al.* 1994; Stark and Darnell 2012). We asked how, in the case of the SIVagm natural host, the system is able to return to an equilibrium, which largely reflects the pre-infection state, while for the SIVagm pathogenic host, the immune system continues to be stimulated, to the point of eventual exhaustion. In earlier chapters, expression levels of microRNAs and mRNAs were considered to be at equilibrium. Here, we were interested in the effects of transcriptional regulators on changing patterns of gene expression through time, as a host immune system responds to viral infection. Gene expression clusters together with their functional enrichments and gene set overlaps were calculated by a collaborator, Dr. Jamie MacPherson (Figures 5.1 to 5.4). Species-specific expression patterns were then characterised, including a rapid but then attenuated response to viral infection within the peripheral blood CD4+ T cells of African Green Monkeys, together with chronic up-regulation of interferon-stimulated genes and cell division regulators within Rhesus Macaques. My contribution was to relate these findings to the expression patterns and binding site locations of a large collection of transcriptional regulators.

I found that the regulators STAT1, STAT2, BATF and IRF4 are highly enriched over co-expressed collections of viral response genes (Figure 5.6), and that each of these regulators belongs to a family of structurally similar TFs, with expression patterns significantly disrupted by SIV infection (Figure 5.7). The STAT and IRF families are key effectors of cytokine-mediated signalling pathways (Ghislain *et al.* 2001; Honda, Yanai *et al.* 2005; Gao, Wang *et al.* 2012; Stark and Darnell 2012), and individual basic leucine zippers (bZIPs), such as BATF, also have well-established immune system functions (Tussiwand *et al.* 2012). Since bZIPs can homo- and heterodimerize, before being able to bind to DNA (Glover and Harrison 1995; Li *et al.* 2012), it is perhaps not surprising that expression changes were distributed over many individual bZIPs. This lead to the novel idea of a network of interacting bZIPs as mediators of gene expression dynamics in CD4+ T cells, during the acute and chronic phases of SIV infection. The discovery that AP1-repressive bZIP factors BATF, and BATF3, are master regulators controlling the differentiation of T cells towards pro-inflammatory lineages, has been the subject of recent high-profile research (Ciofani *et al.* 2012; Yosef *et al.* 2013). An involvement of BATF in a range of inflammatory disorders is also recognised (Quigley *et al.* 2010; Logan *et al.* 2012), and this factor has also been associated with T cell exhaustion, and poor prognosis in HIV infection (Quigley *et al.* 2010; Larsson *et al.* 2013). We have here shown that this factor and its molecular partners are dysregulated not only within the chronic phase of infection but also within peripheral blood and lymph node CD4+ T cells, within the very first day following SIV infection. In conjunction with IRF4, BATF may serve to activate expression of cytokine genes (Yosef *et al.* 2013). We therefore argue that this factor, and its binding partners, may contribute to the chronic activation and eventual exhaustion of CD4+ T cells, in the context of pathogenic infection by SIV.

The study design for the material in Chapter 5 was intended to draw upon as little prior knowledge as possible, in order to favour unbiased discovery. Therefore, the maximum possible number of curated TF gene annotations, and TF ChIP-seq datasets from appropriate cell lineages, were used. There is therefore a good chance for the factors identified in more than one (and sometimes many) tests to genuinely play a role within the context of SIV infection, as argued. Nevertheless, the study was not an empirical one, and it will be essential to test each of its more specific biological hypotheses in detail, before accepting a proof of concept for this approach. We are therefore working with our experimental collaborators to design the follow-up experiments to test this research within a laboratory setting. It would be of particular interest to obtain ChIP-seq datasets for key factors at a number of time points post-infection within these model species or within a cell line system. Combining ChIP-seq with expression pattern datasets through time has recently been successfully used to define gene

regulatory networks within related contexts, such as the differentiation of T cells (Yosef *et al.* 2013).

6.7 Final comment

Regulation of the protein-coding gene expression pathway is as interesting as it is complex. With powerful technologies to mine the distributions of molecules within cells, we begin to have some chance to understand this complexity. Genome-wide and integrative programmes of research therefore seem assured to dominate within the biological sciences in coming decades. It is hoped that others will build upon the content of this thesis, leading by turns to further scientific progress, as our collective knowledge of the genome gathers pace.

Appendix S1

Supplementary material for Chapter 1

Table S1.1. Transcription factors with YALE/HAIB ChIP-seq used in the thesis.

TF	As activator	As repressor	Type	Family/Domain
AR	Yes.		SS	
ATF2	Yes. Binds CRE with Jun. Also a HAT.		SS	bZIP
ATF3	Binds CRE	At ATF sites.	SS	bZIP
BATF		Binds JUN proteins at TRE and more weakly at CRE but inert.	SS	
BCL11A			SS	Zinc finger
BCL3	Yes. Of NFKB1 target genes		SS	
BCLAF1		Repressor interacting with BCL2	SS	
BDP1	Activator at POL3 promoters		G	Homeodomain
BHLHE40			SS	Helix-loop-helix
BRF1	Activator at POL3 promoters		G	
BRF2	Activator at POL3 promoters		G	
CBX3		Silencer. Component of heterochromatin.	S	Chromobox
CEBPB	YES. regulates cytokines		SS	bZIP
CEBPD	YES. regulates cytokines		SS	bZIP
CREB1	Yes. Binds CRE		SS	
CTCF	YES. e.g. of APP. Binds a HAT	YES. e.g. of MYC. Binds an HDAC	SS	Zinc finger
CTCFL		YES. Recruits methylation marks	SS	Zinc finger
E2F1	YES. Binds E2 site.		SS	wHLH
E2F4	YES. Believed.	YES, primarily? Binds E2 site.	SS	wHLH
E2F6		YES. Lacks activation domain	SS	wHLH

EBF1	Activator.		SS	IPT/TIG
EGR1	Yes.		SS	Zinc finger
ELF1	Yes. Enhancer	Yes. Repressor	SS	Zinc finger
EP300	Yes. HAT		G	
ESR1	Yes. Through ERE or ERE independently with other TFs	Yes. Of NFKB1, mutually.	SS	Nuclear receptor
ESRRA			SS	Nuclear receptor
ETS1	Yes	Yes	SS	ETS
FOS	Yes. With Jun family binds AP-1 sites.	Perhaps YES. FOS has an inhibitory domain.	SS	bZIP
FOSL1	Yes. With Jun family binds AP-1 sites.	Perhaps YES. FOS has an inhibitory domain.	SS	bZIP
FOSL2	Yes. With Jun family binds AP-1 sites.	Perhaps YES. FOS has an inhibitory domain.	SS	bZIP
FOXA1	Yes		SS	Forkhead
FOXM1	Yes		SS	Forkhead
FOXP2		Yes	SS	Forkhead
GABPB1	Yes. In tetramer with GABPA		SS	ETS
GATA1	Yes		SS	Zinc finger
GATA2	Yes		SS	Zinc finger
GATA3	Yes		SS	Zinc finger
GTF2B	Yes. At POL2 promoters		G	wHLH
HDAC2		Yes. HDAC	C	
HNF4A	Not sure.	Yes.	SS	Nuclear receptor
HNF4G	Not sure.	Not sure.	SS	Nuclear receptor
HSF1	Yes. Binds HSE		SS	wHLH
IRF4	Yes.	Yes. Negatively regulates Toll-like receptor pathway.	SS	
JUN	Yes. With c-FOS in AP-1		SS	bZIP
JUND	Yes. AP-1 component	Yes.	SS	bZIP
MAX	Yes. MYC-MAX binds E-box	Yes. MAD-MAX. Binds E-box.	SS	Helix-loop-helix
MBD4			SS	

MEF2A	Yes. Binds MEF element	Yes. Represses NUR77 leading to synaptic differentation	SS	MADs-box
MEF2C	Yes. Binds MEF element	Not sure.		MADs-box
MTA3		Yes, through HDACs	SS	
MYBL2	Yes. Activates cyclin D	Yes.	SS	
MYC	Yes. MYC-MAX binds E-box.		SS	Helix-loop-helix
NANOG	Yes	Yes	SS	Homeodomain
NFATC1	Yes. Induces IL-2/IL-4		SS	
NFE2			SS	bZIP
NFIC	Yes		SS	
NFKB1	Yes. Homo or heterodimer binds kappa-B sites	Yes. As p50-p50 homodimer.	SS	p53
NFYA	Yes. CCAAT binding		SS	CBF-NFY
NFYB	Yes. CCAAT binding		SS	CBF-NFY
NR2C2	Yes. Binds hormone response elements (HRE)	Yes.	SS	Nuclear receptor
NR2F2	Yes. As transactivator	Yes. Represses GR & GATA2	SS	Nuclear receptor
NR3C1	Yes	Yes	SS	Nuclear receptor
NRF1	Yes		SS	
PAX5	Yes. Activates B-cell specific genes	Yes. Switches off lineage-inappropriate genes.	SS	Homeodomain
PBX3	Yes		SS	Homeodomain
PML	Yes e.g. DNA repair	Yes e.g. inhibits ELF4	SS	TRIM
POL2	Yes		G	
POLR3A	Yes - active chromatin gates Pol III accessibility to the genome.		G	
POLR3G	Yes - active chromatin gates Pol III accessibility to the genome.		G	
POU2F2	Yes.		SS	Homeodomain
POU5F1	Yes	Yes?	SS	Homeodomain
PPARGC1A	Yes		SS	

RAD21		Yes. Epigenetic silencing	C	
RDBP		Yes. Negative elongation factor.	SS	
REST		Yes.	SS	Zinc finger
RUNX3	Yes	Yes	SS	
RXRA	Yes. Ligand-bound RXR/RAR	Yes. non-ligand-bound RXR/RAR	SS	Nuclear receptor
SETDB1		Yes. Histone methyltransferase	C	
SIN3A		Yes. Co-repressor with REST	G	
SIX5	Yes. Binds ARE element		SS	Homeodomain
SMARCA4	Yes. Relieves repressive chromatin structures SWI/SNF	Yes. Co-repressor of ZEB1	C	
SMARCB1	Yes. relieves repressive chromatin structures SWI/SNF	Yes. Co-repressor of ZEB1	C	
SMARCC1	Yes. relieves repressive chromatin structures SWI/SNF	Yes. Also represses some genes	C	Homeodomain
SMARCC2	Yes. relieves repressive chromatin structures SWI/SNF	Yes. Also represses some genes	C	Homeodomain
SOX2	Yes	Yes?	SS	Homeodomain
SP1	Yes.	Yes.	SS	Zinc finger
SP2	Yes.	Yes.	SS	Zinc finger
SP4	Probable		SS	Zinc finger
SPI1	Yes. Binds PU box		SS	ETS
SREBF1	Yes. Binds SRE		SS	Helix-loop-helix
SREBF2	Yes. Binds SRE		SS	Helix-loop-helix
SRF	Yes. Activates many IEGs	Yes. Binds SRE. Repressor of SMAD-mediated TGF pathways	SS	MADs-box
STAT1	Yes. Binds IFN-stimulated response element (ISRE)		SS	STAT
STAT2	Yes. Binds IFN-stimulated response element (ISRE)		SS	STAT
STAT5A	Yes. Binds IFN-stimulated response element (ISRE)		SS	STAT
SUZ12	Yes. Polycomb repressor.		C	

TAF1	Yes. Binds core promoter		G	
TAF7	Yes. Component of TFIID		G	
TCF12	Yes. Binds E-box		SS	Helix-loop-helix
TCF3	Yes. Binds E-box		SS	
TCF7L2	Yes. With CTNNB1	Yes. In absence of CTNNB1	SS	HMG
TEAD4	Yes. Binds M-CAT motif		SS	
TFAP2A	Yes	Yes. Represses MYC	SS	AP2
TFAP2C	Yes	Yes. Represses MYC	SS	AP2
THAP1			SS	Zinc-finger
TP63	Yes	Yes	SS	P53
TRIM28		Yes. Co-repressor for KRAB-domain zinc finger proteins	C	TRIM
USF1	Yes. Binds E-box & Inr elements		SS	
USF2	Yes. Binds E-box & Inr elements		SS	
XRCC4			G	
YY1	Yes. Via HATs	Yes. Via HDACs	SS	Zinc finger
ZBTB33	Yes. For subset, with CTNND2	Yes. Mainly repressive, binds methylated CpG dinucleotides and recruits N-CoR complex	SS	Zinc finger
ZBTB7A		Yes. Repressive Notch signals	SS	Zinc finger
ZEB1	Yes (mildly). Enhances some genes.	Yes. Binds E-box	SS	Zinc finger
ZNF263		Putative repressor	SS	Zinc finger
ZNF274		Probable. Recruits SETDB1 and TRIM28	SS	Zinc finger
ZZZ3	Yes. Component of HAT ATAC complex		SS	homeodomain

Key: **SS**: Sequence-specific transcription factor **C**: Chromatin-remodelling agent/modifying enzyme G**:** General transcription factor

Regulatory sign annotations are derived from an extensive sample of papers deposited in PubMed. An already abundant literature relating to these 117 factors is expanding all the time,

so these annotations are meant only as an approximate guide to current knowledge. Knowledge of regulatory signs for most TRFs is evidently incomplete, since the effects of most regulators have been examined in detail only in a subset of biological contexts. Nevertheless, many regulators can be confidently assigned as either activators or repressors in most of the biological contexts observed to date. Note that a small number of these factors (11 in total) were not included within the current analysis in Chapter 3 (i.e. Figure 3.5 and Table 3.3. are based on a slightly reduced collection of TRFs).

Appendix S2

Supplementary material for Chapter 2

Table S2.1. Matched samples from protein-coding and microRNA expression atlases

Sample from protein-coding atlas	Sample from microRNA atlas
CD14._Monocytes	hsa_Monocytes-CD14
CD19._BCells.neg._sel..	hsa_B-cell-CD19-pool
CD34.	hsa_HSC-CD34
CD4._Tcells	hsa_T-cell-CD4
CD8._Tcells	hsa_T-cell-CD8
Cerebellum	hsa_Cerebellum-adult
Heart	hsa_Heart
Liver	hsa_Liver
Lymphoma.burkitt.s.Raji.	hsa_Burkitt-raji
Ovary	hsa_Ovary
Pancreas	hsa_Pancreatic-islets
Pituitary	hsa_Pituitary
Placenta	hsa_Placenta
Prostate	hsa_Prostata
Testis	hsa_Testis
Thyroid	hsa_Thyroid
Uterus	hsa_Uterus
Wholebrain	hsa_Cerebellum-adult hsa_Frontal-cortex-adult hsa_Midbrain-adult

Human tissue samples were manually compared between the mRNA and microRNA expression atlases used in this study (Su *et al.* 2004; Landgraf *et al.* 2007). Only healthy tissues were considered**.** Mean microRNA expression was taken across the 3 brain region samples matched to 'wholebrain'.

Table S2.2. GO terms with most highly regulated microRNA host genes

Total genes	Host genes	Mean TRFs (all genes)	Mean excess TRFs (host genes)	p-value (corrected)	GO term
13510	532	30.7	4.0	0.0E+00	cell part
15515	584	29.1	4.6	0.0E+00	cellular_component
14053	544	29.7	4.5	0.0E+00	molecular_function
4180	207	33.9	5.2	0.0E+00	**regulation of cellular metabolic process**
3953	191	34.4	5.6	2.8E-03	**regulation of macromolecule metabolic process**
1511	52	15.6	9.5	3.3E-03	molecular transducer activity
13640	544	30.4	3.7	4.0E-03	biological_process
5555	212	23.5	4.9	4.4E-03	membrane part
1511	52	15.6	9.5	5.0E-03	signal transducer activity
4112	206	33.7	5.1	5.3E-03	**regulation of primary metabolic process**
8135	364	30.3	3.8	5.4E-03	**biological regulation**
7334	335	30.7	3.9	5.7E-03	**regulation of cellular process**
941	51	30.6	10.3	6.0E-03	nucleic acid binding transcription factor activity
4582	223	33.5	4.9	6.1E-03	**regulation of metabolic process**
3357	139	36.4	6.1	6.2E-03	cellular biosynthetic process
2354	97	36.8	6.7	6.3E-03	nucleobase-containing compound biosynthetic process
4617	163	21.8	5.3	6.6E-03	intrinsic to membrane
7704	346	30.5	3.8	6.6E-03	**regulation of biological process**
939	51	30.6	10.3	6.6E-03	sequence-specific DNA binding transcription factor activity
3087	120	37.4	6.3	7.0E-03	nucleic acid binding
3520	145	36.1	5.8	7.0E-03	biosynthetic process
2419	101	36.8	6.6	9.0E-03	heterocycle biosynthetic process
3456	143	36.2	5.6	9.2E-03	organic substance biosynthetic process
2945	149	34.7	5.4	9.4E-03	**regulation of cellular macromolecule biosynthetic process**
3236	166	33.6	5.3	9.6E-03	**regulation of nucleobase-containing compound metabolic process**
3317	168	33.6	5.1	9.6E-03	**regulation of nitrogen compound metabolic process**
2534	103	36.4	6.5	9.6E-03	organic cyclic compound biosynthetic process
2457	97	37.8	6.4	9.6E-03	cellular macromolecule biosynthetic process
3181	160	33.8	5.2	9.7E-03	**regulation of biosynthetic process**
11454	472	31.5	3.1	1.0E-02	cellular process
4506	159	22.0	5.2	1.0E-02	integral to membrane
1994	81	35.9	7.1	1.0E-02	transcription, DNA-dependent
3020	152	34.5	5.5	1.0E-02	**regulation of macromolecule biosynthetic process**
2420	100	36.6	6.7	1.0E-02	aromatic compound biosynthetic process
10969	450	31.8	3.3	1.1E-02	binding
2261	90	31.2	6.7	1.1E-02	**response to stress**
2470	102	36.7	6.2	1.1E-02	cellular nitrogen compound biosynthetic process
3155	159	33.9	5.1	1.1E-02	**regulation of cellular biosynthetic process**
2486	124	28.7	5.6	1.2E-02	**response to chemical stimulus**
3150	161	34.4	4.9	1.2E-02	**regulation of gene expression**
5538	254	28.4	3.9	1.3E-02	response to stimulus
2778	108	38.7	5.7	1.4E-02	RNA metabolic process

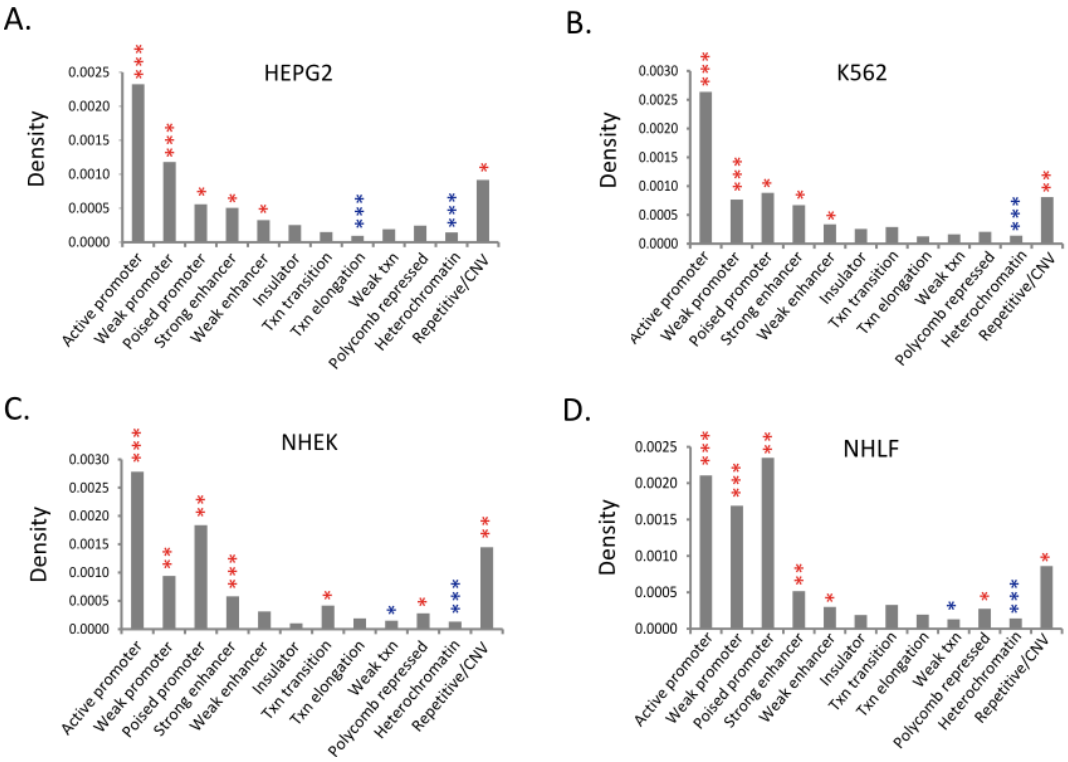
4579	206	36.5	4.0	1.4E-02	nucleus
2632	108	37.2	5.7	1.4E-02	macromolecule biosynthetic process
2699	135	34.0	5.1	1.5E-02	**regulation of transcription, DNA-dependent**
3254	127	20.6	4.9	1.6E-02	plasma membrane
3215	124	39.3	5.0	1.7E-02	nucleic acid metabolic process
2782	138	34.3	4.9	1.9E-02	**regulation of RNA metabolic process**
2718	136	34.1	4.9	1.9E-02	**regulation of RNA biosynthetic process**
3178	162	34.7	4.5	1.9E-02	protein complex
9334	405	29.4	2.9	1.9E-02	single-organism process
1186	70	33.6	6.9	1.9E-02	**regulation of transcription from RNA polymerase II promoter**
1732	93	25.9	5.8	1.9E-02	**regulation of multicellular organismal process**
7014	283	35.2	3.4	1.9E-02	primary metabolic process
2163	85	36.7	6.0	1.9E-02	RNA biosynthetic process
2185	82	35.3	6.1	1.9E-02	DNA binding
5728	240	36.1	3.7	1.9E-02	macromolecule metabolic process
1058	64	26.2	6.7	2.1E-02	**regulation of multicellular organismal development**
5387	229	26.0	3.6	2.2E-02	membrane
3754	176	36.5	4.0	2.5E-02	macromolecular complex
7567	307	34.7	3.1	2.6E-02	metabolic process
908	51	34.9	7.5	2.9E-02	**negative regulation of macromolecule biosynthetic process**
1789	99	33.7	5.1	2.9E-02	**positive regulation of macromolecule metabolic process**
950	54	34.0	7.1	3.0E-02	**negative regulation of cellular biosynthetic process**
964	54	34.0	7.0	3.1E-02	**negative regulation of biosynthetic process**
7240	296	34.9	3.0	3.2E-02	organic substance metabolic process
8176	361	29.8	2.5	4.4E-02	single-organism cellular process
3027	144	31.5	3.9	4.5E-02	**positive regulation of cellular process**
902	51	34.2	6.6	4.6E-02	**negative regulation of nitrogen compound metabolic process**
1840	98	33.4	4.9	4.7E-02	**positive regulation of cellular metabolic process**
2710	122	23.3	4.0	5.0E-02	multicellular organismal process

Significance of excess mean TRFs (transcriptional regulatory factors) was calculated by simulation as described in the main chapter methods (Section 2.2). P-values were corrected for multiple tests using the Benjamini-Hochberg correction (Benjamini *et al.* 2001). Note that since many GO terms within the list have part-whole relationships with one another, the test is more stringent here than for an independent collection of tests.

Appendix S3

Supplementary material for Chapter 3

Figure S3.1. MicroRNA cluster density within intergenic regions of different functional types.



Functional annotations were obtained from the track 'hg19, regulation, BroadChromHmm' across 9 cell lines, from the UCSC genome browser, based upon systematic studies of chromatin state (Karolchik *et al.* 2004; Ernst and Kellis 2010; Ernst *et al.* 2011). Four of these cell lines are shown, with minor variations in microRNA density distributions compared to the example shown in the main text (Figure 3.2C. Densities relative to chromatin state in human embryonic stem cells):

A. HEPG2 (immortalised hepatocarcinoma)
B. K562 (immortalized myelogenous leukemia)
C. NHEK (renal epithelial cell line)
D. NHLF (immortalized lung fibroblast)

Significance of microRNA cluster enrichment within given genomic regions is indicated by coloured asterisks. Red = enriched, blue = depleted. *: $p < 0.05$, **: $p < 10^{-4}$, *** $p < 10^{-8}$. Significance was calculated using the binomial distribution, with (x, n) = number of clusters in (required region, total); and p = fraction of intergenic space of each functional type.

The relative enrichments of microRNAs differ between higher in weak, and higher in poised, promoter regions (in terms of the underlying chromatin state model). In 1 / 9 cell lines (NHLF), there is actually a marginally higher density of microRNAs within poised promoters, compared to active promoters. Nevertheless, due to the different total fractions of intergenic space contained within these two region types, microRNA density remains most significant within active promoter regions in all lines examined.

Table S3.2. 5’-antisense microRNA clusters and their flanking genes.

Intergenic microRNA	Flanking gene
let-7i	C12orf61
mir-1247	DIO3.1
mir-1285-2	PCYOX1
mir-1302-11	ACOO8993.1
mir-146b	CUEDC2
mir-1539	C18orf32
mir-210,-132	PHRF1
mir-212	HIC1
mir-22	WDR81
mir-3166	RAB38
mir-3178	PDPK1
mir-3188	JUND
mir-3199-2	PITPNB
mir-320a	POLR3D
mir-320b-1	IGSF3
mir-345	SLC23A49
mir-34b,-34c	BTG4
mir-3529	AEN
mir-3677,-	RNPS1
mir-3678	GRB2

Intergenic microRNA	Flanking gene
mir-3908	RILPL1
mir-3912	NPM1
mir-3913-1	CCT2
mir-3928	RNF185
mir-3936	SLC22A5
mir-3939	FGFR1OP
mir-4256	WNT2B
mir-4314	ALOX12B
mir-4325	SPO11
mir-4443	CDC25A
mir-4448	PARL
mir-4453	FBXW7
mir-4470	ASPH
mir-4482-1	GSTO2
mir-4496	SELPLG
mir-4508	MKRN3
mir-4514	MESCDC1
mir-4530	PLEKHG2
mir-4536-1	MAGEH1
mir-4686	TH

Intergenic microRNA	Flanking gene
mir-4710	RP11
mir-4712	GABPB1
mir-4727	CWC25
mir-4733	NF1
mir-4734	MLLT6
mir-4739	AC105337.
mir-4752	LILRB2
mir-4754	RPS5
mir-4795	CHMP2B
mir-4801	KIAA1239
mir-497	BCL6B
mir-4999	RAB11B
mir-5091	BOD1L
mir-548al	PGM2L1
mir-5680	NCALD
mir-5695	SYEC2
mir-5696	RNF149
mir-5707	PTPRN2
mir-607	LCOR
mir-659	EIF3L
mir-92b	MUC1

All clusters within 10 kb of the 5’ end of an Ensembl protein-coding gene are listed, including 9 that are also near to their other flanking gene (including for example the mir-34b,-34c cluster).

Appendix S4

Supplementary material for Chapter 4

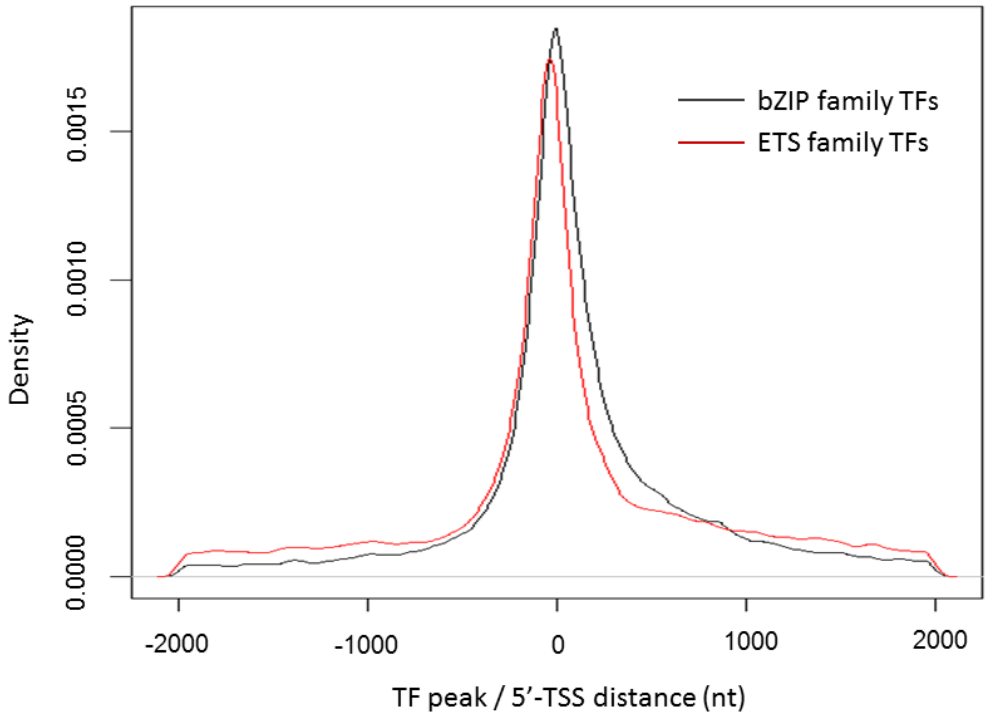


Figure S4.1. Density of bZIP and ETS family TFs around protein-coding 5'-TSSs

Densities were plotted using kernal density estimators in R using the collection of all the distances of ChIP-peaks from bZIP and ETS TF families to 5'-TSSs of human protein-coding genes in Ensembl v.65 (See Supplementary Table S1.1 for TF annotations). The bZIP density plot lies further downstream than the ETS family density plot, and is also spread more diffusely over the 4000 nt shown ($p < 10^{-15}$ by K-S tests).

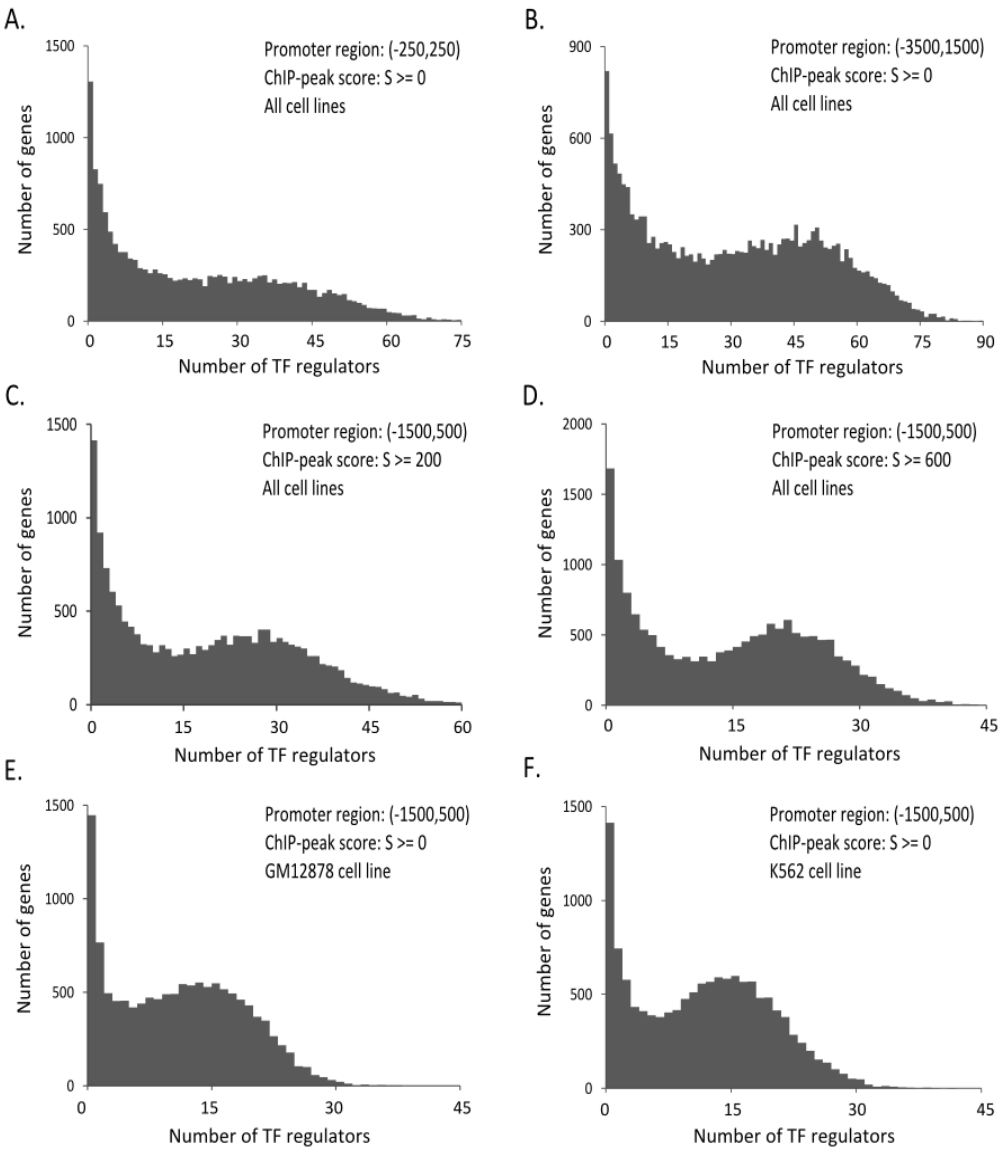


Figure S4.2. Degree distributions of transcriptional regulators

TRF degree distribution was calculated straightforwardly as described in the main text.

A & **B.** Varying *cis*-regulatory window size

C & **D.** Varying minimum TRF peak quality score (from the MACS algorithm)

E & **F.** Restricting to TRF datasets derived from examples of specific cell types (GM12878 and K562 cell lines). Distributions for other cell types, e.g. HELA, are similar.

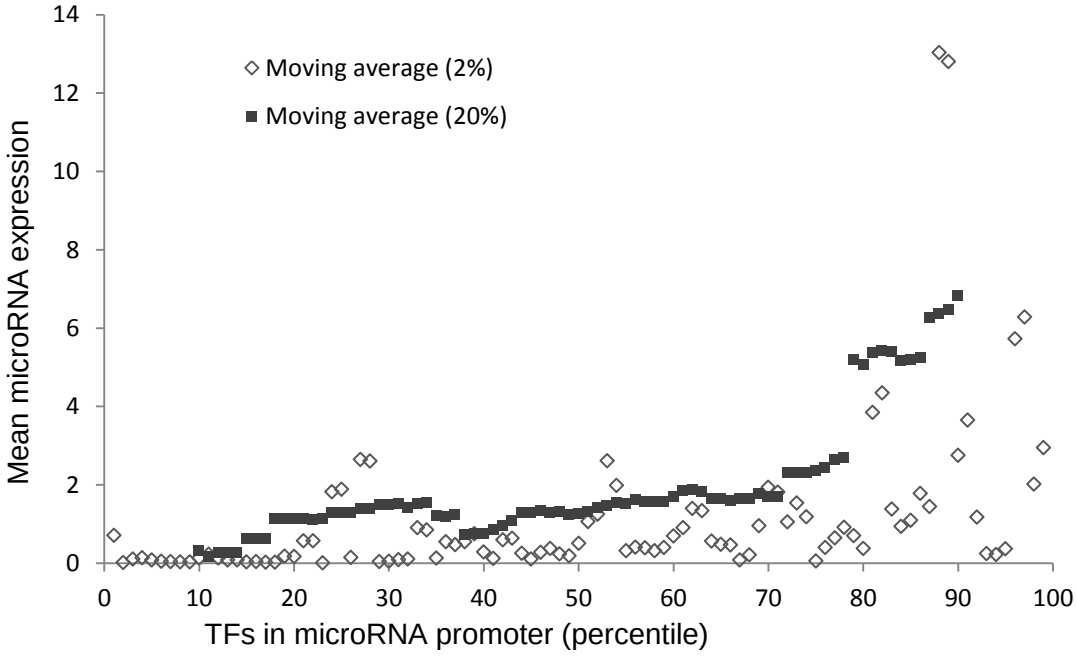


Figure S4.3. MicroRNA expression level against number of transcriptional regulators

MicroRNA expression (in units of clone counts from microRNA libraries) for 636 microRNA mature sequences was averaged across all 172 human samples from (Landgraf *et al.* 2007). MicroRNAs from this atlas were ranked by number of predicted TF regulators using either the host promoter region of intronic microRNAs, or a fixed interval of 10 kb upstream of all other microRNAs. Mean microRNA expression was then averaged across sets of microRNAs representing moving averages of either 2% or 20% of the complete collection, at intervals of 1%. Although lymphatic system samples are significantly overrepresented, a similar distribution is obtained using samples from other tissue systems. We note that the microRNA atlas is relatively sparse, with 87.5% of clone counts (per microRNA per tissue) equal to zero.

Table S4.4. Comparison of expression levels of targets and non-targets of individual TRFs

TF	Mean target expression	Mean non-target expression	Expression ratio (Target/Non)	Normal z-score	Regulatory Sign(lit)	Regulatory Sign(GO)	PCC (Expression , % targets)
RDBP	453.16	91.4	4.96	15.35	NEG	UNC	0.876
ZZZ3	366.08	99.8	3.67	7.47	POS	UNC	0.703
TAF7	234.34	78.3	2.99	20.39	POS	BOTH	0.986
POL2	128.47	45.8	2.81	26.46	POS	UNC	0.988
POU2F2	185.57	66.8	2.78	29.31	POS	UNC	0.996
STAT5A	192.09	69.6	2.76	26.38	POS	POS	0.994
NFATC1	234.08	85.5	2.74	19.18	POS	UNC	0.979
PML	168.91	62.0	2.72	31.93	BOTH	NEG	0.992
TAF1	141.24	52.2	2.71	29.02	POS	POS	0.989
ZBTB33	231.38	85.9	2.69	17.60	BOTH	UNC	0.978
TCF3	201.55	75.8	2.66	24.28	POS	POS	0.990
MYC	151.18	57.0	2.65	29.77	POS	POS	0.989
MTA3	189.80	71.8	2.64	25.44	NEG	UNC	0.990
FOXM1	183.68	69.7	2.64	25.88	POS	UNC	0.989
ATF2	189.86	72.7	2.61	22.66	POS	UNC	0.991
BCL3	210.80	81.7	2.58	20.31	POS	BOTH	0.985
ELF1	149.20	58.1	2.57	29.69	BOTH	POS	0.992
RUNX3	147.81	57.8	2.56	30.69	BOTH	UNC	0.992
YY1	144.88	57.1	2.54	25.96	BOTH	UNC	0.992
REST	163.06	64.9	2.51	23.97	NEG	NEG	0.991
ETS1	182.01	72.6	2.51	25.01	BOTH	POS	0.987
SIX5	210.08	84.2	2.50	18.31	POS	UNC	0.975
NFKB1	135.54	54.7	2.48	27.41	BOTH	BOTH	0.992
CEBPD	208.46	84.4	2.47	21.14	POS	UNC	0.989
NFIC	185.14	75.1	2.46	23.00	POS	BOTH	0.982
E2F4	151.66	62.6	2.42	27.90	BOTH	UNC	0.986
TEAD4	185.69	77.0	2.41	21.90	POS	POS	0.984
SP1	150.72	63.3	2.38	26.54	BOTH	POS	0.991
MYBL2	169.43	71.3	2.38	25.42	BOTH	UNC	0.986
MBD4	213.33	90.4	2.36	17.44	BOTH	UNC	0.982
SIN3A	164.43	69.8	2.36	23.34	NEG	NEG	0.985
FOSL2	207.46	92.0	2.25	14.65	BOTH	UNC	0.954
ATF3	197.15	87.5	2.25	18.98	BOTH	UNC	0.984
TCF7L2	142.96	63.6	2.25	27.04	BOTH	POS	0.981
PAX5	172.68	76.9	2.25	23.05	BOTH	NEG	0.992
TCF12	169.95	75.8	2.24	20.00	POS	UNC	0.985
MAX	130.09	59.0	2.21	21.63	BOTH	UNC	0.978
CREB1	173.36	78.8	2.20	24.71	POS	POS	0.986
BCLAF1	184.36	84.7	2.18	20.53	NEG	NEG	0.979
NR2F2	184.56	86.6	2.13	20.19	BOTH	BOTH	0.990
SP4	185.26	87.5	2.12	17.57	POS	UNC	0.993
NR3C1	201.97	98.5	2.05	11.67	BOTH	UNC	0.943
IRF4	205.69	100.9	2.04	10.65	BOTH	POS	0.952
SMARCB1	134.70	67.2	2.00	26.89	BOTH	POS	0.982
PPARGC1A	215.53	109.1	1.97	8.21	POS	POS	0.910
SMARCC1	140.25	71.8	1.95	26.65	BOTH	POS	0.982
JUN	156.96	83.3	1.88	20.85	POS	POS	0.989
E2F1	137.42	73.0	1.88	23.55	POS	BOTH	0.978
SMARCA4	154.11	82.5	1.87	21.21	BOTH	BOTH	0.984
STAT2	182.16	98.5	1.85	14.37	POS	UNC	0.958
USF2	201.11	108.9	1.85	7.00	POS	POS	0.874
GABPB1	154.04	84.8	1.82	20.00	POS	UNC	0.978
EP300	158.15	88.2	1.79	15.73	POS	POS	0.972
E2F6	131.85	74.2	1.78	20.88	NEG	NEG	0.975
MEF2A	190.50	107.9	1.77	7.58	BOTH	UNC	0.840
USF1	151.89	86.2	1.76	17.58	POS	POS	0.982
SPI1	166.69	94.7	1.76	14.68	POS	BOTH	0.969
JUND	155.92	89.6	1.74	16.93	BOTH	UNC	0.991
STAT1	154.21	88.7	1.74	18.01	POS	UNC	0.985

SREBF2	183.61	106.8	1.72	11.12	POS	POS	0.974
BDP1	186.32	109.7	1.70	2.66	BOTH	UNC	0.552
BCL11A	183.35	109.0	1.68	6.49	BOTH	UNC	0.890
SMARCC2	143.12	85.7	1.67	22.74	BOTH	BOTH	0.973
CEBPB	147.99	88.7	1.67	20.53	POS	POS	0.982
CBX3	164.77	98.8	1.67	13.48	NEG	NEG	0.936
NR2C2	167.55	100.9	1.66	12.85	BOTH	UNC	0.961
SRF	170.90	103.0	1.66	12.93	BOTH	POS	0.963
PBX3	173.82	106.3	1.63	10.85	POS	UNC	0.920
NRF1	162.54	100.5	1.62	15.02	POS	UNC	0.970
EGR1	136.94	84.7	1.62	22.76	BOTH	BOTH	0.991
SP2	167.47	104.3	1.61	12.63	BOTH	UNC	0.943
THAP1	170.08	106.1	1.60	11.73	BOTH	UNC	0.967
NFE2	172.83	110.1	1.57	6.02	BOTH	POS	0.816
FOSL1	169.58	109.9	1.54	6.01	BOTH	POS	0.661
EBF1	154.96	100.4	1.54	14.70	POS	POS	0.983
HSF1	163.46	107.2	1.53	10.36	POS	UNC	0.939
SREBF1	155.67	102.8	1.51	16.23	POS	POS	0.991
MEF2C	169.89	112.3	1.51	2.91	BOTH	BOTH	0.663
XRCC4	169.91	112.6	1.51	2.07	BOTH	UNC	0.175
GTF2B	150.27	101.9	1.47	12.57	POS	UNC	0.956
POLR3G	153.31	108.3	1.42	6.98	BOTH	UNC	0.832
ESRRA	152.00	108.3	1.40	10.55	BOTH	POS	0.973
FOXP2	153.20	109.4	1.40	10.00	NEG	NEG	0.950
ZEB1	149.31	107.6	1.39	7.87	BOTH	BOTH	0.859
FOS	141.90	102.4	1.39	14.84	BOTH	POS	0.969
GATA3	145.82	107.4	1.36	5.64	POS	UNC	0.896
NFYA	141.57	104.8	1.35	13.71	POS	POS	0.965
HNF4A	135.27	100.3	1.35	14.71	BOTH	POS	0.984
GATA1	148.48	110.7	1.34	6.54	POS	UNC	0.809
ZNF263	131.52	98.8	1.33	15.96	NEG	UNC	0.968
NANOG	146.42	110.2	1.33	2.52	BOTH	UNC	0.613
RXRA	145.65	111.2	1.31	5.75	BOTH	BOTH	0.826
SOX2	144.72	111.0	1.30	1.32	BOTH	BOTH	0.330
TFAP2C	134.21	103.5	1.30	13.47	BOTH	UNC	0.974
TFAP2A	135.17	105.6	1.28	11.61	BOTH	UNC	0.957
ESR1	141.85	111.3	1.27	1.21	BOTH	UNC	-0.092
CTCFL	137.51	108.3	1.27	8.59	NEG	UNC	0.914
NFYB	132.75	105.9	1.25	13.17	POS	POS	0.962
HDAC2	134.62	108.5	1.24	7.10	NEG	BOTH	0.870
BHLHE40	136.25	112.3	1.21	5.28	UNC.	NEG	0.899
BRF2	135.55	112.5	1.21	2.51	BOTH	UNC	0.648
FOXA1	127.95	106.5	1.20	9.58	POS	BOTH	0.948
TRIM28	132.37	111.5	1.19	-0.32	NEG	BOTH	0.023
ZBTB7A	125.65	106.3	1.18	15.69	NEG	NEG	0.968
TP63	132.02	112.4	1.17	0.47	BOTH	BOTH	0.104
BATF	130.90	111.5	1.17	6.09	NEG	UNC	0.842
BRF1	127.15	112.3	1.13	1.42	BOTH	POS	0.519
AR	124.81	111.8	1.12	3.15	POS	POS	0.616
SETDB1	119.52	110.3	1.08	10.27	NEG	UNC	0.898
RAD21	115.96	110.1	1.05	7.01	NEG	UNC	0.764
CTCF	116.22	111.2	1.04	3.70	BOTH	BOTH	0.693
GATA2	104.70	112.9	0.93	2.63	POS	POS	0.589
HNF4G	98.43	113.4	0.87	3.77	BOTH	UNC	0.808
POU5F1	78.60	113.5	0.69	-2.64	BOTH	UNC	-0.441
ZNF274	64.95	112.7	0.58	2.50	NEG	UNC	0.074
SUZ12	45.41	120.8	0.38	-14.02	NEG	NEG	-0.979

Key for Table S4.4:

Mean target and non-target expression: Calculated for each TF across 14889 all genes with mRNA expression measurements averaged across the 79 tissue samples in the Novartis expression atlas (Su *et al.* 2004).

Expression ratio (Target / non): Mean target expression divided by mean non-target expression

Normal z-score: All genes were ranked by expression level, and the Mann-Whitney U statistic calculated comparing target to non-target ranks for each TF. Due to the large sample size, to calculate significance the U statistic was converted to a normal z-score using mean(U) = 0.5 * n(targets) * n(non-targets) and st.dev(U) = sqrt((n(targets) * n(non-targets) * (n(genes) + 1) / 12). P-values were then calculated using the cumulative normal distribution function. P-values were then corrected for the 117 tests carried out using the Benjamini-Hochberg method (Benjamini *et al.* 2001). Cells are highlighted darker pink (Target expression > Non-target expression: $p < 10^{-15}$), lighter pink (Target expression > Non-target expression: $p < 0.01$), or blue (Target expression < Non-target expression: $p < 10^{-15}$). No TFs were found with Target expression < non-target expression at only the weaker threshold $\alpha = 0.01$.

Regulatory sign (lit): TF regulatory signs derived from manual examination of literature: POS (exclusive activator), NEG (exclusive repressor), BOTH (alternating sign depending on context), and UNC (unclassified TFs).

Regulatory sign (GO): TF regulatory signs derived from the Gene Ontology (Ashburner *et al.* 2000), specifically from the terms 'positive regulatory of transcription' (POS), 'negative regulatory of transcription' (NEG), with some TFs falling under both terms (BOTH), and many that are unclassified (UNC).

PCC (Expression, % targets): Linear regression correlation coefficients derived from the changing proportion of genes targeted across genes ranked by mean expression level. Because the regression lines are based on equidistantly spaced gene expression percentiles, it is not possible to consult statistical tables to find the significance of the PCC value. These can be calculated instead by permutation tests and are significant at $a = 10^{-5}$ for the majority of the TFs in the sample.

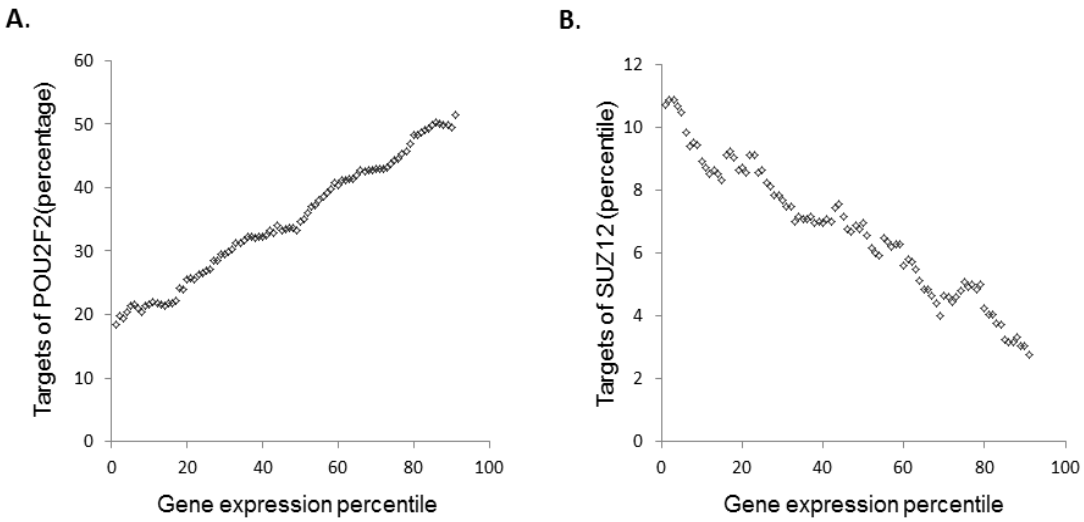


Figure S4.5. Fraction of POU2F2 and Suz12 target genes as expression level varies.

We show examples for two TFs:

A. POU2F2, a strong activator
B. Suz12, a strong repressor

Genes in the Novartis mRNA expression atlas were first ranked by mean expression level across tissues. Then the number of targets for each TRF were found within windows of 2% of the data at a time, moving in steps of 1% through the ranked list, and converted to a percentage of total genes within the window. Plots for the majority of TRFs show similar, though weaker, positive linear trends (Supplementary Table S4.4).

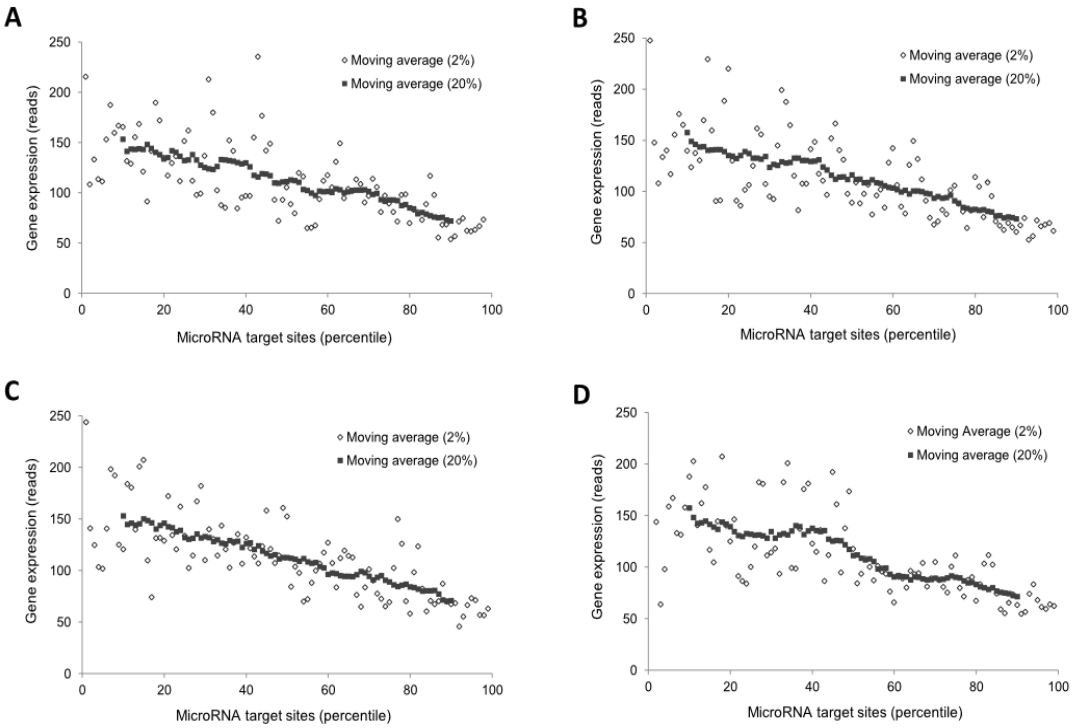


Figure S4.6. Variation in gene expression with numbers of miRanda-predicted microRNA target sites

Plots display moving averages of mean gene expression across tissues in the Su *et. al.* atlas, measured within windows from the list of all protein-coding genes ranked by number of miRanda-predicted microRNA target sites. Four different minimum miRanda scores S were examined:

A. S ≥ 130 **B.** S ≥ 140 **C.** S ≥ 150 **D.** S ≥ 160

Within each figure, two different window sizes were used for the calculation of moving averages, representing intervals of either 2% or 20% of the complete gene collection.

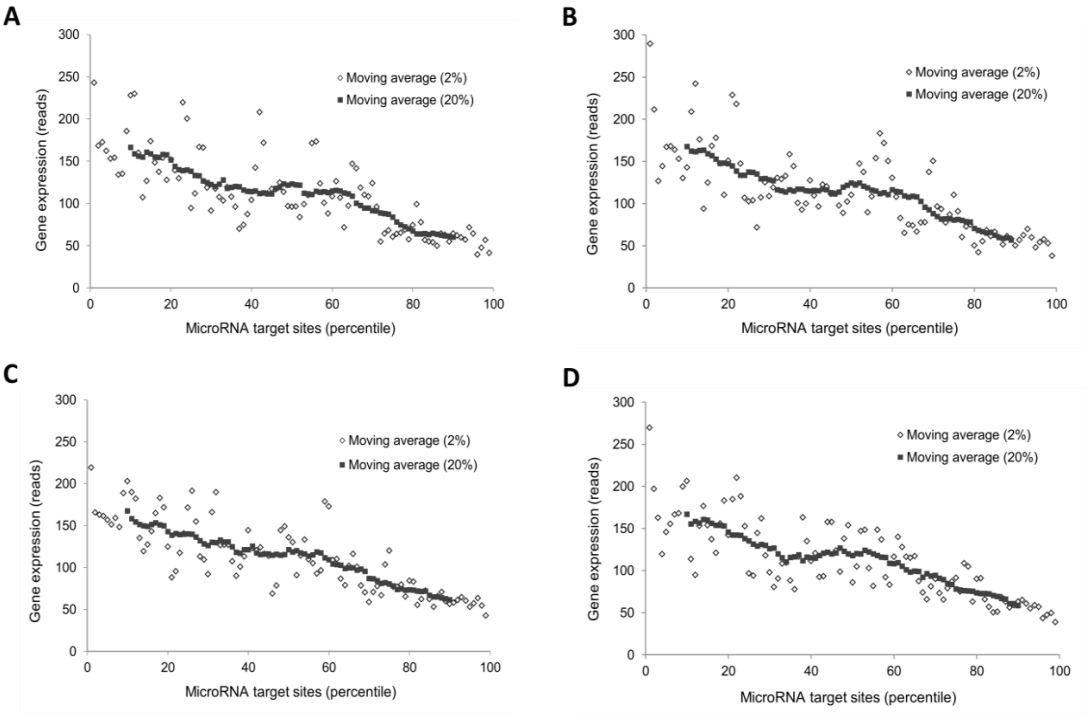


Figure S4.7. Variation in gene expression with numbers of seed sequence predicted microRNA target sites

Plots display moving averages of mean gene expression across tissues in (Su *et al.* 2004), measured within windows from the list of all protein-coding genes ranked by number of seed-match-predicted microRNA target sites. Four different collections of seeds were examined:

A. All seeds **B**. 7mer-A1 seeds **C.** 7mer-m8 seeds **D**. 8mer seeds

Within each figure, two different window sizes were used for the calculation of moving averages, representing intervals of either 2% or 20% of the complete gene collection.

Table S4.8. Hypergeometric p-values for incidence of the bidirectional (TRF ↔microRNA) feedforward circuit.

A. MicroRNA target sites from miRanda v.3.0

		(TF ↔ microRNA) FFL p-values			
		Minimum miRanda microRNA site score			
MicroRNAs	**TFs**	**130**	**140**	**150**	**160**
All	**All**	$< 10^{-15}$	$< 10^{-15}$	8.2×10^{-13}	9.5×10^{-5}
Intronic	**All**	$< 10^{-15}$	3.8×10^{-12}	2.1×10^{-7}	0.45
Intergenic	**All**	7.5×10^{-8}	1.0×10^{-4}	2.1×10^{-3}	0.078
Opposite	**All**	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$	6.5×10^{-7}
All	**+ regulator**	3.1×10^{-3}	0.018	0.072	0.24
All	**- regulator**	$< 10^{-15}$	9.5×10^{-10}	6.1×10^{-6}	0.20
All	**alternating**	$< 10^{-15}$	$< 10^{-15}$	10^{-15}	3.4×10^{-9}

B. MicroRNA target sites that are exact matches to seed sequences

		(TF ↔ microRNA) FFL p-values			
		Seed type			
MicroRNAs	**TFs**	**All**	**7mer-A1**	**7mer-m8**	**8mer**
All	**All**	7.7×10^{-3}	0.17	4.5×10^{-3}	0.28
Intronic	**All**	0.99	1.00	0.87	0.95
Intergenic	**All**	0.89	0.98	0.57	0.036
Opposite	**All**	0.13	0.90	0.043	0.36
All	**+ regulator**	1.00	1.00	1.00	1.00
All	**- regulator**	8.1×10^{-4}	0.062	0.17	0.022
All	**alternating**	1.1×10^{-7}	5.8×10^{-3}	5.2×10^{-8}	8.0×10^{-4}

We calculated the probability of occurrence of the real number of bidirectional links (TRF ↔ microRNA) given the numbers of (TRF → microRNA) links and (TRF ← microRNA) links, using the hypergeometric distribution. We calculated this p –value for the complete network (with standard settings as described in the main text), and for subnetworks based on classes of (i) MicroRNAs: intronic (inside a host gene, in the same sense to it), intergenic (lying outside of protein-coding genes), and opposite(inside a host gene, but in the opposite sense to it); and (ii) TRFs: positive regulators, negative regulators, and those with alternating functions depending on context. (See TRF annotations in Supplementary Table S1.1). We repeated this using four choices of minimum microRNA target site score in miRanda; and for target sites based upon 3 types of microRNA seed, as well as the union of these (all seeds). Colour-coding: Non-significant = blue. Significant = black. Significant < 10^{-15} = red.

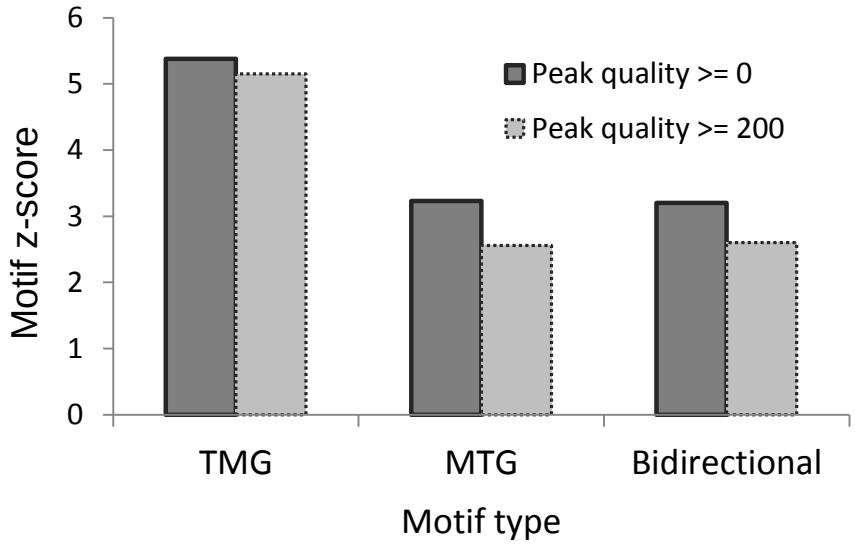


Figure S4.9. Variation in motif significance with minimum TRF peak quality score

Peak quality scores are pre-calculated, using the MACS algorithm, and available for download from the UCSC genome browser (Karolchik *et al.* 2003; Hinrichs *et al.* 2006; Zhang *et al.* 2008). Motif significance for the three basic types of feed-forward circuit are calculated as described in Chapter 4.

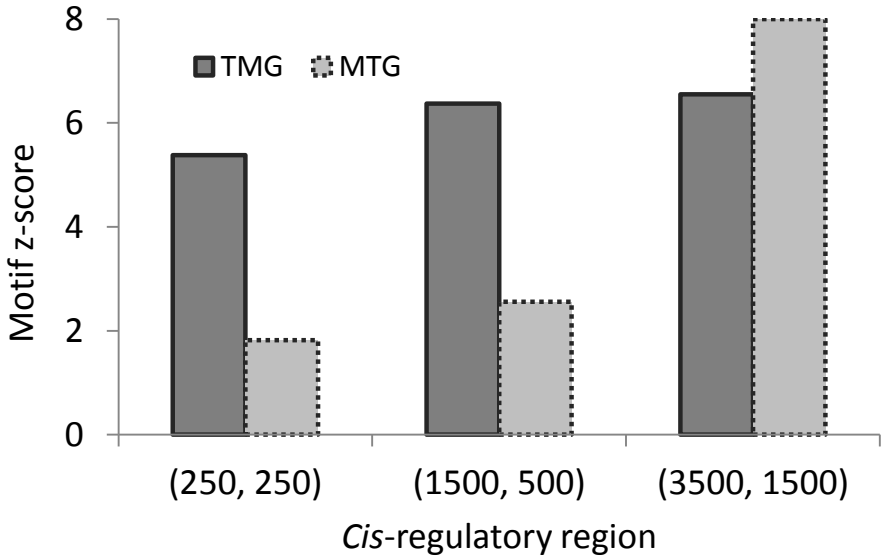


Figure S4.10. Variation in motif significance with *cis*-regulatory region size

We ran motif calculations on three subsets of the network for a narrow, medium, and wide-interval *cis*-regulatory region. *Cis*-regulatory regions around 5'-TSSs of protein-coding genes were defined by the upstream and downstream intervals, shown in brackets in the horizontal axis labels. E.g. the standard interval from 1500 nt to 500 nt downstream of 5'-TSSs is shown as (1500, 500). All TRF peaks were used with MACS score ≥ 500, but the trends are independent of this parameter.

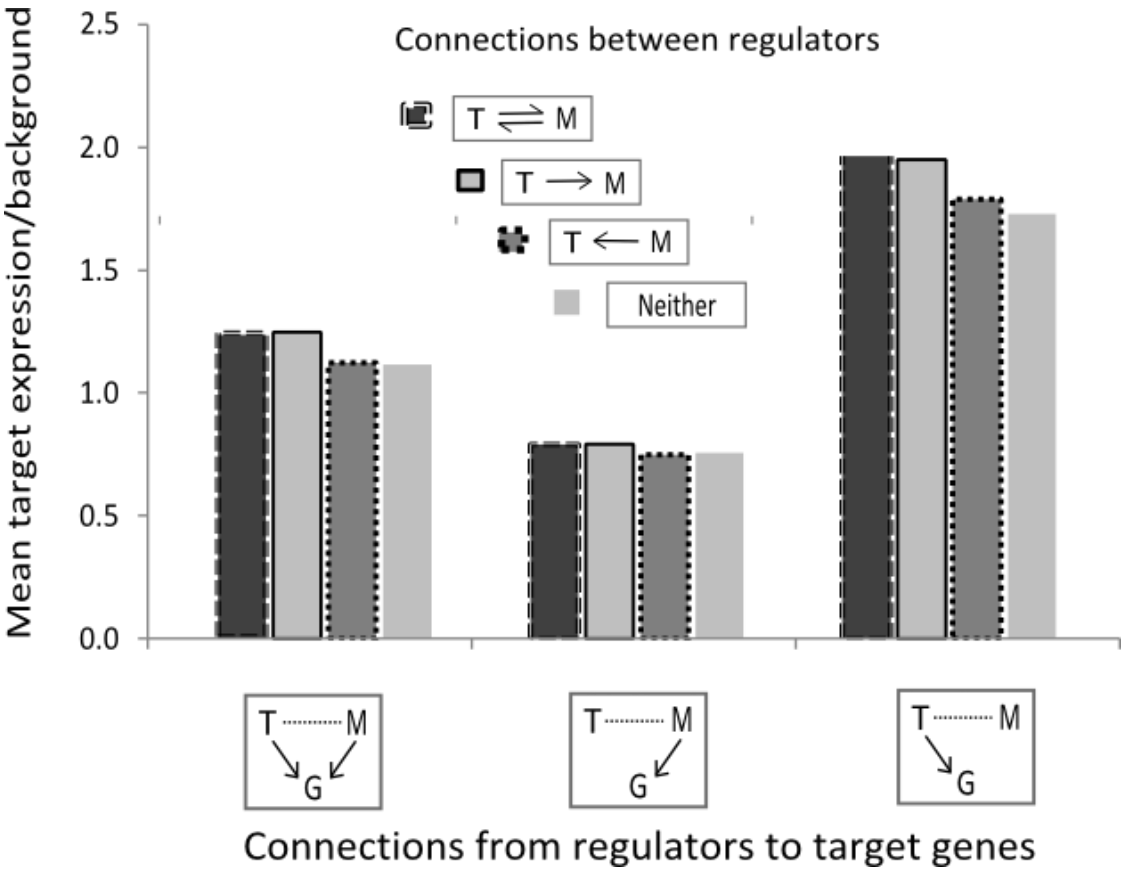


Figure S4.11. Gene expression levels across subsets of TRF-microRNA feedforward loops

MicroRNA target predictions were calculated using all matches to 7mer-A1, 7mer-m8 and 8mer seeds. For each TRF-microRNA pair, the collection of all protein-coding genes was separated into four groups according to whether predicted as targets by both TRF and microRNA, by only the TRF, by only the microRNA, or untargeted. Mean expression was then calculated within the three targeted groups (shown on the horizontal axis), and divided by the mean expression for all untargeted genes (as background). The complete collection of TRF-microRNA pairs was also divided into four classes according to the predicted regulatory connections between TRF and microRNA (shown as the legend 'connections between regulators). Mean expression ratios (targeted/background) were then averaged across all the TRF-microRNA pairs in each class, and displayed as the bars in the figure.

Table S4.12. Correlations between individual node z-scores across target predictors

		Correlation between individual z-scores between pairs of target sets		
		Motif		
	MicroRNA target set pair	TMG	MTG	Bidirectional
Protein-coding genes	(M130, M150)	0.667	0.786	0.632
	(M130, all seeds)	0.691	0.481	0.627
	(M150, all seeds)	0.587	0.443	0.606
TRFs	(M130, M150)	0.945	0.841	0.516
	(M130, all seeds)	0.104	0.117	0.318
	(M150, all seeds)	0.252	0.138	0.436
microRNAs	(M130, M150)	0.950	0.416	0.332
	(M130, all seeds)	0.953	0.222	0.283
	(M150, all seeds)	0.919	0.297	0.341

Individual node z-scores were calculated for all 19472 protein-coding genes, 117 TRFs, and 1919 microRNA mature sequences (methods). Since these scores are normally distributed, Pearson's correlation was used to test agreement between microRNA target predictors and settings, and these PCC values are shown in the table. We show examples for three settings: miRanda with minimum site score S ≥ 130, miRanda with minimum site score S ≥ 150, and exact matches to seed sequences (the union of 8mer, 7mer-A1 and 7mer-m8 seeds). Significant PCC values ($\alpha < 0.01$) are highlighted.

Appendix S5

Supplementary material for Chapter 5

Table S5.1. Gene lists within each of 15 expression profile clusters.

CD4+ source: A = African green monkey, M = Rhesus macaque, L = lymph node, P = peripheral blood

Cluster	Source	Genes
1	AL	ACTG2 AURKB BFSP2 C9orf142 CCDC19 CCNB2 CDC20 CDCA5 CHGB CHN1 CKAP2 CTTN CXCR3 DUSP6 EGLN3 FBLN5 GSDMD HRSP12 IL10 KCNK5 KIF11 KIF23 KIF2C KLK2 LILRA4 MT1CP MYLK NFIA P2RX5 PIM1 POU2AF1 SMPDL3A SPC25 ST6GALNAC5 SYNPO TIGIT TRIB1 UBE2C ZSWIM3
	AP	C1orf56 DNAJB6 DNAJC19 FAM118B FLAD1 IGKC IRF9 LFNG LOC440871 LOC441198 MAP3K4 RSPH10B SNX19 TMEM165
	ML	DEK DHFR DYNC1LI1 ENC1 ENTPD1 EPHB1 ERAP1 EYA2 FAM54A FANCD2 FASN GCLC GIMAP4 GINS4 GLA GPR155 GSPT1 HCCS HDAC2 HELLS HMGB1 HMGB1L1 HSPD1 IFI16 IFITM3 IFNGR1 INTS7 ISG20 ITGA4 JAK2 KIF15 KIF4A LAMP2 LGALS8 LILRB5 LMNB1 LOC389873 MOBKL1B MRPL42 MRPS18B MYL6B MYO6 NADK NAGK NBN NEK2 NOX3 NUFIP2 NUP88 OPN3 P2RY14 PACSIN1 PANX1 PARP2 PARP9 PCIF1 PDHX PDIA6 PDK1 PDP1 PDZD11 PECI PGAM4 PITPNA PITPNB PITPNB PNO1 PPM1K PPP2CB PRC1 PREPL PSMC1 RARRES3 RFC2 RGS2 RPS6KC1 RTP4 SAMHD1 SAP30 SCLT1 SCP2 SEPT2 SERBP1 SFXN1 SIGLEC7 SLFN12L SMC2 SNX1 SP100 SPAG1 SPTBN1 STAT2 STMN1 STMN2 STOML1 SUMO3 SUSD1 THBD TMEM128 TRIM38 TROAP TSFM TSG101 TSPAN12 TTC39A TTK TWF1 TXN USP18 USP25 USP41 WDHD1 XRCC2 YEATS4 ZBTB32 ZNF267
	MP	ACTL6A AHSA1 AK2 ALDH1A1 AP2B1 APAF1 APOBEC3H ARHGAP11A ARV1 ATP6V1C1 BARD1 BET1 BST2 C1orf103 C1orf31 C2orf29 C4orf46 CARD16 CARD17 CCDC76 CD38 CDKN3 CENPA CENPE COX11 DAB1 DDX60L DHFR DNAJA1 DNAJB6 EMG1 ERGIC2 ERMP1 EXO1 FAM118B FAM54A FBXO22 FBXO6 FLAD1 GBP2 GDAP2 GIMAP1 GIMAP7 GLMN GMPR2 GPR180 GPR68 HDHD2 HIF1AN HMGB3 HSBP1 HSP90AA1 HSP90AA2 HSPD1 IFI35 IFITM3 IRF9 ISG20 JMJD6 KIFC1 KRIT1 LMO2 METTL4 MOBKL1B MRPL42 MRPS14 NT5C3 NUP88 OAS3 OMA1 PANK2 PARP9 PARVG PFDN6 PLEKHO2 PPID PPP2CB PSMB10 QRSL1 RHOT1 RNF213 SACM1L SAMD9 SAMD9L SARS SFRS6 SLC35A5 SLC35B3 SLFN12L SNAPC3 SNX19 SP140L SRCAP STARD3NL SYNCRIP SYNJ2BP TARS TMEM126B TMEM140 TMEM33 TNFAIP8L2 TRANK1 TREML2 TSR1 TTC9C USP25 WDR61 XRN1 YIPF4 ZC3H10 ZMPSTE24 ZNF613 ZNF672
2	AL	GALNT8
	AP	IGHD IGHD IGKC IGKC IGKV1-5 IGL@ IGLC1 IGLC1 IGLC2 IGLC2 IGLJ3 IGLV2-14 IGLV3-25 IGLV4-3 LOC439957 LOC440786
	ML	BIRC5 BIRC5 BUB1 C1orf54 C6orf173 CCNB2 CD274 CD38 CDC2 CDC20 CDC6 CDCA2 CDCA8 CDKN3 CENPA CENPE CEP55 CKAP2 CLEC4GP1 CSGALNACT1 DDX58 DEPDC1B DHX58 DLGAP5 DSCC1 DTL DTX3L E2F8 EIF2AK2 ELOVL7 EPB41L3 EPSTI1 FAM26F FAM49A FAM49A FBXO6 FGFBP2 GBP1 GCNT1 GPC1 GPSM2 HAVCR2 HERC6 HIP1 HMGB3 HMMR ICAM1 IDO1 IL21 IL6 IL8RB IRF7 IRF8 KCNMA1 KIAA1598 KIF11 KIF14 KIF23 KLHL6 KPNA2 KYNU LEPREL1 LGALS3BP LILRA4 LITAF MAR1 MCM4 MCM6 MELK MND1 MYBL2 NT5C3 NUF2 OAS1 OAS2 OAS2 ODF3B OIP5 PLA2G4C POLE2 PPM1J PRR11 RACGAP1 RAD51 RASA4 RASA4P RCC2 RNF213 SAMD9L SECTM1 SERPINE2 SGTA SHCBP1 SKA1 SPATS2L SPC25 STAT1 STAT1 SULF2 TMEM140 TOP2A TRIM21

		TRIP13 UBE2C VLDLR WARS WDR63 ZWINT
	MP	APOBEC3A ARHGAP11A ASPM ASPM APOL2 BCL2L14 C14orf119 CCR5 CMPK2 CXorf21 DDX60 DEPDC1B DHX58 DTL ENPP2 EPB41L3 EPSTI1 FEN1 GIMAP4 GPC1 GPSM2 HERC5 HERC6 IFI44 IFIT3 IRF7 MPHOSPH9 NUF2 NUSAP1 OAS1 OAS2 PLA2G4C PSAT1 SPATS2L STAT1 TSFM TYMS UBE2T UHRF1 USP18 XAF1
3	AL	APOBEC3H APOBEC3H HSH2D IRF1 IRF7 IRF9 ISG20 MTFR1 NEAT1 SPATS2L TRIM21 TRIM22
	AP	APOBEC3H CEBPB CISH CLDN5 HSPA1A IER2 IFI27 SDPR SPR SPSB2 TMEM140
	ML	C10orf11 C19orf6 CCRL2 CSTA CTSC DBNL EFNA2 EXOC5 FAM108B1 FHAD1 GNPDA1 LOC284379 LRP10 MAN1B1 MYL4NAIP NCRNA00094 NONO PAPSS1 PDK4 PLIN1 PLK1 RAB13 RGS18RHOT2 SPPL2A SPTBN5 ST6GALNAC6 TMEM177 TNFSF12-TNFSF13 TNFSF13 WNK1 ZNF618
	MP	ASPM BIRC5 CCDC33 CCNA2 CCNB2 CDC20 CDC6 CDCA5 CDT1 CENPP DAB2IP DLGAP5 GBP1 GPR132 HSPB1 IDO1 KIAA0101 KIF11 KIF23 MCM5 MELK MTHFS MYBL2 NCAPH OIP5 PCBP3 RACGAP1 RAD51 RRM2 SECTM1 SPC25 UBE2C
4	AL	AKAP5 ATAD2B BATF3 C11orf41 C15orf28 C21orf96 CACNA1A CCDC80 CCL17 CCL20 CCNT1 CD1B CD1E CDC42BPB CEP170 CLIC2 CNN3 DAB2IP DAPK1 DEPDC7 G0S2 GALNT10 GPR157 HSPB1 IDH1 IRAK2 IRX3 JAG1 KRT18 KRTAP4-1 LAMB1 LAMP3 LPPR5 MARCKS MS4A7 MT1H NRP2 PLIN1 RASSF4 RPS27A SH3PXD2A SORL1 TMC2 TMEM150C TMEM177 TSC1 TTC39B USP34 ZBTB20
	AP	BYSL CASP3 CD19 DAB2IP DDIT3 EXOC8 FEM1B INA KRTAP4-2 LCMT2 MYLIP NSUN3 PFN4 RNF139 RWDD3 TCEANC TMEM167AVPREB3 ZNF134 ZNF211
	ML	CA8 CD40 CLIC5 CRYAA FAM164A FCER2 FOXD1 GNAZ IGF2 IL11RA IL1R1 KLRB1 LOC151009 LOC284454 LOC400987 LOC440894 NPAS2 PDE4B PROC PTK7 RPS23 SLC7A10 ST8SIA1 WASH2P WASH3P YPEL5 ZBTB7A
	MP	AGPAT9 ASB12 C17orf71 C20orf111 CACNA1A CCDC64 CCDC73 CCL3 CD69 CDC42SE2 CDO1 CIRBP CR2 CSPG5 CXCR4 DAAM1 DENND4A DENND4A DUSP5 FAM116AFAM117B FAM65B FBXL12 FEM1B FOXN3 GAS2L1 GPR183 hCG_1820801 HERPUD1 IER3 IL1B IL8 KIF26B KIF2B KLHL24 KLHL28 LNX2 LOC284454 LOC440459 MAPT MED9 MXI1 MYL5 MYLIP PDE4D PIP4K2A PLAUR PLXNC1 RGS1 RLIM RNF139 SBDS SIAH1 SLC16A6 SMAD5 SNIP1 TCP11L2 TMC2 TMC4 WDR43 YPEL5 ZNF256 ZNF548 ZNF614 ZNF771
5	AL	CRNN
	AP	
	ML	BCL2L14 BOP1 C3 CLEC7A CMPK2 COL1A2 CXCL10 CXCL11 DDX60 HERC5 IFI27 IFI44 IFI44L IFI6 IFIT1L IFIT3 ISG15 LVRN MNDA MX1 MX2 OASL RSAD2 SUCNR1 USP18 XAF1
	MP	HSPA1A IFI27 IFI44L IFI6 IFIT1 ISG15 MX1 MX2 OASL
6	AL	BCL2L14 CMPK2 DDX60 HERC6 IFI44 IFI44L IFI6 IFIT1 IFIT3 ISG15 MX2 OASL RSAD2 TMEM140 TTC39A USP18 XAF1
	AP	BCL2L14 DDX60 DHX58 EPSTI1 GBP1 HERC6 IFI44L IFI6 IFIT1 IFIT3 IRF7 ISG15 MX1 MX2 OAS2 OASL RSAD2 SCAMP1 SPATS2L USP18 XAF1
	ML	KIF2B TTC39B
	MP	
7	AL	CCL2 CDKN1A COG2 CSF2RA DEFA1 DEFA3 DRD3 EPSTI1 IFI35 IFIH1 KLK7 LUM OAS2 PARP9 PLA2G4C SAMD9L SECTM1 SMCHD1
	AP	AFF1 ANKRD22 APOBEC3H BATF3 C19orf59 C6orf150 CABP5 CAV2 CD274 CEACAM6 CNP CREM CRISPLD2 DSCR8 EFNA1 EIF2AK2 ENTPD3 FBXO6 FTSJD2 GCA GOLGA6L4 GTPBP1 HSH2D HSPA1A HSPA1B ICAM1 IFI16 IFI44 IFIT5 IGSF6 IL8RB IRF2 LINS1 LMO2 MBOAT1 NBN NCOA7 NLRC5 P2RX7 PARP9 PCDH18

		PLA2G4C	PLSCR2	PML	PML	PML	PRDM1	RAPGEF6	RNF213	RPL10L
		SBNO2	SCIN	SLFN5	SNTB1	SOBP	SP100	STAT1	STAT2	TICAM1
		TICAM1	TMEM171		TRANK1	TRANK1	TRIM5	USP18	USP41	VCPIP1
		ZNF267								
	ML	DIS3								
	MP	C10orf71	GDA	IL1F8	PCDHGA3		RASA4	RASA4P	RNF26	SFRP4
8	AL	ACOT12	BLMH	C15orf43	CXCL11	FAM66C	FLJ34503		GALNT13	GALNTL2
		GDF3	GNMT	GTF2IP1	HAND1	HERC5	KRT14	KRTAP23-1		
		PACSIN3	RNASE11							
	AP	CCL8	CXCL10	HERC5	SECTM1					
	ML	ATXN7L3								
	MP	C12orf40		ZNF619						
9	AL	C2orf85	GRIK3	LOC159110		LYPD6	PARD3B	RXFP3	STON2	VWDE
	AP	ALDH3B2		ANO6	DNAJB13	DST	EBF2	EML5	ESYT3	FAM124A
		FAM181B		HOXB6	KCNA5	MED31	METT5D1		POM121L8P	
		PRDM10	STARD13		SYT6	tcag7.929		ZMYND12		
	ML	FATE1	FCN2	GPC6	GZMA	S100P	TTLL6			
	MP	DPP10	FCN2	GZMA	HAVCR2	KIAA1324L		PTPRD	TCEAL6	UBQLNL
10	AL	MYH15	TTC9	TTLL6						
	AP	EGF	FAM30A		LOC13886					
	ML	AADACL4		ANGPT2	C15orf5	C1orf213		CDKL4	CETN1	CLVS2
		FAM13A		FGF20	FLJ22536		FLJ37035		hCG_1980447	
		KBTBD12	KCNMB3	KRT80	LOC158376		LOC286135		LRAT	MYO1C
		MYOD1	PDCD1LG2		PLOD2	PLS3	SLC26A4	SLC6A7	SRGAP3	TCF21
		TM4SF18		TPO						
	MP	FOLH1	FOLH1B	OR52J3						
11	AL	ACRBP	AKR1D1	ALDH1L2	C17orf47		C1orf53	CBLB	CCDC89	DENND5B
		DGKB	DNAJB13	DUSP16	ENPP3	FADS2	FAM123B		FAM90A1	FOXP2
		FRY	GRIA4	GRM7	KATNAL2	KLF1	LOC146336		LOC153910	
		LOC91316		MAP3K15		MED14	MPPED1	NLN	OR1N2	PLA2G4A
		PTN	RGN	RSPO3	SLC17A3	SLC1A3	STARD13	THSD4	TSPO2	VIL1
		ZNF558								
	AP	DKFZp547G183								
	ML	CHST10	DNAJB13	DST	EBF2	EML5	FAM124A		HOXD1	LOC285141
		PRDM10	PXMP2	SPHK1	SYT6	tcag7.929				
	MP	ATPGD1	C7orf55	FANK1	GAB2	IRGM	KCNG1	LOC286467		LYZL4
		POF1B								
12	AL	CCDC33	CEP192	HERC2P4	IL17RB	KIF2B	LOC151475		LOC404266	
		PION	QRFPR							
	AP									
	ML									
	MP	DNAJB13	DST	EBF2	EML5	FAM124A		FAM181B		GRIK3
		LOC91948		LVRN	METT5D1		PRDM10	RAB6C	STARD13	SYT6
		tcag7.929								
13	AL	ABCA4	CEL	CPNE6	DNAJB13	EBF2	FAM124A		FAM181B	
		FAM71A	HBE1	KLRD1	KRTAP4-2		LOC285389		PRDM10	PRO2852
		SYT6	tcag7.929		TRIM16	UBOX5				
	AP	ATP8B3	C12orf42		C15orf43		C22orf34		C4BPB	CAPS2
		CCNYL2	DAPL1	ELF5	FAM110	B	hCG_38984		HDAC5	IRGM
		KCNJ5	LYZL4	MAN1B1	NR4A2	SLC10A1	TFEB	USP46	ZSCAN20	
	ML	AHNAK2	ARHGEF17		BMP8A	C3orf24	CCDC62	CPA3	CXorf27	FOXE1
		IL1RN	ITGA7	KCNIP3	LOC400573		MCTP2	MMP26	MTFR1	OSBPL5
		PLA2G4E	PLEKHH2	PMS2	RNASET2	SLC28A2	TBC1D17	WNT5B		
	MP	ADD2	ANO4	CXCL3	CYP2C18	EYA2	FAM131B		HBE1	HESX1
		MMP1	RGS16	SERPINB2		SERPINC1				
14	AL	DST	PART1							
	AP	C11orf84								
	ML	hCG_33730		HPR	IRF4	OXCT2	RSPO4	SCAMP1	SLC2A3P2	
		SOHLH2	ST6GAL1	WNK3	ZNF662					
	MP	ACY3	HPX	USP46						

15	AL	MX1
	AP	
	ML	CEP192 HERC2P4 IL17RB LOC151475 MGC24103 PION QRFPR
	MP	CEP192 HERC2P4 IL17RB QRFPR
Unclustered	AL	APOL2 BRSK2 C11orf64 CCDC25 CXCL13 EML5 ESYT3 FAM124A HEY1 HRASLS2 HSPA1A HSPA1B IFI27 IFITM3 LAG3 LOC91948 LVRN METT5D1 MGAM MND1 POPDC3 PTPRD S1PR3 STAT1 TARP ZBP1 ZFP64 ZNF432
	AP	APOL2 FAM124A GZMA HBE1 HPX HRASLS2 IGHG1 IL8 MUC19 MYC S1PR3 SLC14A1 TAP2 TRIM16 UGT2B4 ZAK
	ML	ADAMDEC1 C2orf69 C5orf25 CCDC33 CDCA7 CHD7 DAB2IP GJA1 LAG3 LOC202181 LOC404266 NDC80 PCBP3 PLA2G7 RESP18 SCFD1 TMC2 TREM2 VPS16 ZBED2 ZNF598
	MP	CHAF1A DHFR GPR158 LOC151475 LOC404266 MRPL17 OAS2 OSBPL5 PION PPP1R12B RSAD2 SPAG5 TRIP13 TTC39B UGT2B4

Unclustered probes from particular sources did not have an expression pattern sufficiently close to any of the 15 expression profiles detected by Mfuzz (Kumar and Futschik 2007). Gene names listed more than once for the same source and cluster correspond to cases where a single gene was sampled by more than one probe.

Table S5.2. ChIP-seq datasets for transcription regulatory factors used in Chapter 5

TRF	Type	Family/Domain	Lymphatic system cell lines	Consortia
ATF2	SS	bZIP	GM12878	HAIB
ATF3	SS	bZIP	GM12878; K562	HAIB; HAIB and YALE
BATF	SS	bZIP	GM12878	HAIB
BCL11A	SS	Zinc finger	GM12878	HAIB
BCL3	SS	BCL	GM12878	HAIB
BCLAF1	SS	BCL	GM12878; K562	HAIB; HAIB
CBX3	S	Chromobox	K562	HAIB
CEBPB	SS	bZIP	K562	HAIB
CREB1	SS	bZIP	K562	HAIB
CTCFL	SS	Zinc finger	K562	HAIB
E2F4	SS	Winged helix-loop-helix	K562B	YALE
E2F6	SS	Winged helix-loop-helix	K562, K562B	HAIB; YALE
EBF1	SS	IPT/TIG	GM12878	HAIB
EGR1	SS	Zinc finger	K562	HAIB
ELF1	SS	Zinc finger	GM12878, K562	HAIB
EP300	C	P300, KIX, TAZ, IBiD	GM12878	HAIB

ETS1	SS	ETS	G12878, K562	HAIB
FOS	SS	bZIP	GM12878, K562	YALE
FOSL1	SS	bZIP	K562	HAIB
FOXM1	SS	Forkhead	GM12878	HAIB
GABPB1	SS	ETS	GM12878, K562	HAIB
GATA1	SS	Zinc finger	K562B	YALE
GATA2	SS	Zinc finger	K562	HAIB
GTF2B	G	Winged helix-loop-helix	K562	YALE
HDAC2	C	Histone deacetylase	K562	HAIB
IRF4	SS	Inteferon regulatory factor	GM12878	HAIB
JUN	SS	bZIP	K562	YALE
JUND	SS	bZIP	K562B	YALE
MAX	SS	Helix-loop-helix	GM12878; K562	YALE; HAIB and YALE
MEF2A	SS	MADs-box	GM12878	HAIB
MTA3	SS	BAH, zinc finger	GM12878	HAIB
MYC	SS	Helix-loop-helix	K562	YALE
NFATC1	SS	NFAT	GM12878	HAIB
NFE2	SS	bZIP	K562	YALE
NFIC	SS	CTF/NF-I	GM12878	HAIB
NFKB1	SS	p53	GM12878	YALE
NFYA	SS	CBF-NFY	K562	YALE
NFYB	SS	CBF-NFY	K562	YALE
NR2C2	SS	Nuclear receptor	GM12878; K562B	YALE
NR2F2	SS	Nuclear receptor	K562	HAIB
PAX5	SS	Homeodomain	G,12878	HAIB
PBX3	SS	Homeodomain	GM12878	HAIB
PML	SS	TRIM	GM12878; K562	HAIB
POL2	G	Multimeric GTF	GM12878; K562	HAIB and YALE
POLR3A	G	Multimeric GTF	K562	YALE
POU2F2	SS	Homeodomain	GM12878	HAIB

RAD21	C	Unknown	GM12878; K562	HAIB and YALE
RDBP	SS	RRM	K562	YALE
REST	SS	Zinc finger	GM12878; K562	HAIB
RUNX3	SS	Runt	GM12878	HAIB
SETDB1	C	SET	K562B	YALE
SIN3A	G	PAH	K562	YALE
SIRT6	C	Sirtuin	K562	YALE
SIX5	SS	Homeodomain	GM12878; K562	HAIB
SMARCA4	C	SWI/SNF	K562	YALE
SMARCB1	C	SWI/SNF	K562	YALE
SP1	SS	Zinc finger	GM12878; K562	HAIB
SP2	SS	Zinc finger	K562	K562
SPI1	SS	ETS	GM12878	HAIB
SRF	SS	MADs-box	K562	HAIB
STAT1	SS	STAT	K562	YALE
STAT2	SS	STAT	K562	YALE
STAT5A	SS	STAT	GM12878; K562	HAIB
TAF1	G	TAF	GM12878; K562	HAIB
TAF7	G	TAF	K562	HAIB
TCF12	SS	Helix-loop-helix	GM12878	HAIB
TCF3	SS	Helix-loop-helix	GM12878	HAIB
TEAD4	SS	TEA	K562	HAIB
THAP1	SS	Zinc-finger	K562	HAIB
USF1	SS	bHLH-leucine zipper	GM12878; K562	HAIB
YY1	SS	Zinc finger	GM12878; K562	HAIB and YALE
ZBTB33	SS	Zinc finger	GM12878; K562	HAIB
ZBTB7A	SS	Zinc finger	K562	HAIB
ZEB1	SS	Zinc finger	GM12878	HAIB
ZZZ3	SS	homeodomain	GM12878	YALE

The column 'type' defines each transcriptional regulator broadly as either 'SS' (sequence-specific, corresponding to regulatory TFs), 'G' (general, corresponding to core components of RNA pol2/pol3 basal transcription-initiation complexes), or 'C' (chromatin-modifying, which bring about covalent modifications to histone proteins). This annotation system has been previously described for these datasets in publication by ENCODE members (Gerstein *et al.* 2012). The final column identifies the research institutions (HudsonAlpha Institute for Biotechnology (HAIB) and Yale University) at which the data were generated. All data are available through the UCSC genome browser: hg19 regulation tracks HAIB TFBS and YALE TFBS (Karolchik *et al.* 2003; Karolchik *et al.* 2004).

Table S5.3. Predicted common *cis*-regulatory targets of STAT1, STAT2, BATF and IRF4

Gene	Name
ABCA3	ATP-binding cassette, sub-family A, member 3
AC083862.1	Uncharacterized protein
AC084082.3.1	Uncharacterized protein
AL583828.1	Uncharacterized mRNA
ANP32E	Acidic nuclear phosphoprotein 32 family, member E
ASXL1	Additional sex combs like 1
B3GAT3	Beta-1,1-glucuronyltransferase 3
BHLHE40	Basic helix-loop-helix family, member E40
BIRC2	Baculoviral IAP repeat containing 2
C12orf57	Chromosome 12 open reading frame 57
C4orf43	Chromosome 4 open reading frame 43
C7orf55	Chromosome 7 open reading frame 55
C8orf37	Chromosome 8 open reading frame 37
CASC5	Cancer susceptibility candidate 5
CBLL1	Casitas B-lineage lymphoma-like 1
CFLAR	CASP8 and FADD-like apoptosis regulator
CPPED1	Calcineurin-like phosphoesterase domain containing 1
CSTF2T	Cleavage stimulation factor 3' pre-RNA subunit 2, tau variant
CUTA	cutA divalent cation tolerance homolog
DENND4A	DENN/MADD domain containing 4A
DPP9	Dipeptidyl-peptidase 9
EHBP1L1	EH domain binding protein 1-like 1
EIF2S3	Eukaryotic translation initiation factor 2, subunit 3 gamma
ENSG00000205534	Uncharacterized protein or pseudogene
ERLIN2	ER lipid raft associated 2
FBXO31	F-box protein 31
FBXO5	F-box protein 5
FRG1	FSHD region gene 1
GPR137	G-protein coupled receptor 137
HERC6	HECT and RLD domain containing E3 ubiquitin ligase family member 6
HIST1H4H	Histone cluster 1, H4h
IFITM1	Interferon induced transmembrane protein 1
INTS5	Integrator complex subunit 5
KDM1A	Lysine (K)-specific demethylase 1A
LASP1	LIM and SH3 protein 1
LUC7L2	

MAK	LUC7-like 2
MAN1A1	Male germ cell-associated kinase
MAP1LC3B2	Mannosidase alpha class 1A member 1
MAPK1	Microtubule-associated protein 1 light chain 3 beta 2
MLL5	Mitogen-activated protein kinase 1
MRPS15	Myeloid/lymphoid leukemia 5
MSRB1	Mitochondrial ribosomal protein S15
MX2	Methionine Sulfoxide reductase B1 (Alias: SEPX1)
NAMPT	Myxovirus resistance 2
NCOA7	Nicotinamide phosphoribosyltransferase
NDUFS7	Nuclear receptor coactivator 7
NOTCH2NL	NADH dehydrogenase ubiquinone Fe-S protein 7
PDE12	Notch 2 N-terminal like
PHF11	Phosphodiesterase 12
PMAIP1	PHD finger protein 11
POLDIP3	Phorbol-12-myristate-12-acetate-induced protein 1
POP7	Polymerase (DNA-directed) delta interacting protein 3
PRDX1	Processing of precursor 7, ribonuclease P/MRP subunit
PSMB3	Peroxiredoxin 1
PTP4A2	Proteasome subunit, beta type, 3
RHBDD2	Protein tyrosine phosphatase type IVA, member 2
RP11-863K10.7.1	Rhomboid domain containing 2
RP11-872D17.8.1	Uncharacterized protein (ENSG00000183154)
RPS12	Uncharacterized protein (ENSG00000254979)
S100PBP	Ribosomal protein S12
SACM1L	S100 family member P-binding protein
SLC39A13	SAC1 suppressor of actin mutations 1-like
SMG1	Solute carrier family 39 member 13
SP100	SMG homolog, phosphatidyl 3-kinase-related kinase
STAT3	SP100 nuclear antigen
SYNCRIP	Signal transducer of activated transcription 3
TAP2	Synaptotagmin binding, cytoplasmic RNA interacting protein
THG1L	Transporter 2, ATP-binding cassette, sub-family B
TMEM140	tRNA-histidine guanylyltransferase 1-like
TROVE2	Transmembrane protein 140
UCHL5	TROVE (TEP1 and Ro60 protein) domain family member 2
USP9X	Ubiquitin carboxyl-terminal hydrolase L5
VAMP5	Ubiquitin specific peptidase 9, X-linked
WDR74	Vesicle-associated membrane protein 5
XXbac-BPG246D15.9.1	WD repeat domain 74 Uncharacterized protein
YARS	Tyrosyl-tRNA synthetase
ZNF131	Zinc finger protein 131
ZNFX1	Zinc-finger, NFX1-type containing 1

Gene symbols are HGNC identifiers, where available.

Table S5.4. Significantly perturbed families of transcription factors

	Numbers of TF expression changes for each TF domain								
	2.5% tails			5% tails			10% tails		
TF domain	OBS	EXP	p(Poisson)	OBS	EXP	p(Poisson)	OBS	EXP	p(Poisson)
bZIP	33	11.3	1.2E-07	60	22.1	2.0E-11	109	44.0	1.1E-16
STAT	12	1.5	4.9E-08	16	3.0	1.0E-07	26	6.1	1.8E-09
IRF	5	2.0	5.2E-02	12	3.9	7.3E-04	20	7.7	1.5E-04
ZF THAP	5	2.7	1.4E-01	14	5.1	9.1E-04	24	10.4	2.1E-04
MADs-box	2	1.1	3.1E-01	6	2.1	2.2E-02	8	4.4	7.8E-02
ZF RING	10	6.3	1.1E-01	20	12.2	2.6E-02	45	24.3	1.1E-04
TRIM	5	1.2	6.8E-03	6	2.3	2.8E-02	10	4.4	1.5E-02
ZF B-box	5	1.2	6.8E-03	6	2.3	2.8E-02	10	4.4	1.5E-02
bHLH	27	19.8	7.2E-02	50	38.7	4.6E-02	85	77.3	2.0E-01
ZF ZZ	2	0.5	8.0E-02	3	0.9	5.9E-02	4	1.7	9.4E-02
ZF C2H2	13	11.7	3.9E-01	31	22.9	6.3E-02	59	46.0	3.7E-02
ZF MYND	0	0.2	1.0E+00	2	0.4	6.6E-02	2	0.9	2.2E-01
Zinc finger	93	91.2	4.4E-01	198	178.9	8.4E-02	392	356.1	3.2E-02
wHTH	36	32.9	3.2E-01	75	64.9	1.2E-01	170	128.8	3.0E-04
ZF PHD	4	3.8	5.2E-01	11	7.5	1.4E-01	19	14.8	1.6E-01
ZF DHHC	1	0.7	5.0E-01	3	1.3	1.4E-01	5	2.6	1.2E-01
ZF C3H1	9	6.1	1.6E-01	16	11.9	1.5E-01	34	23.3	2.2E-02
Ets	7	5.1	2.5E-01	12	9.9	2.9E-01	25	19.8	1.5E-01
BCL2	0	0.2	1.0E+00	1	0.4	3.4E-01	2	0.9	2.3E-01
ZF RanBP2	1	0.7	4.9E-01	2	1.3	3.6E-01	6	2.6	5.1E-02
NR	7	8.0	6.9E-01	17	15.9	4.2E-01	33	31.1	3.9E-01
Forkhead	8	7.1	4.2E-01	12	14.4	7.8E-01	34	28.7	1.8E-01
ZF FYVE	1	0.9	6.0E-01	1	1.8	8.3E-01	3	3.5	6.8E-01
ZF KRAB	46	58.8	9.6E-01	97	115.7	9.7E-01	191	230.2	1.0E+00
Homeobox	15	35.5	1.0E+00	28	70.5	1.0E+00	60	139.4	1.0E+00
Chromobox	0	0.2	1.0E+00	0	0.4	1.0E+00	0	0.9	1.0E+00
Paired box	0	2.1	1.0E+00	0	4.0	1.0E+00	1	7.8	1.0E+00
ZF CCHC	0	0.2	1.0E+00	0	0.4	1.0E+00	2	0.8	2.0E-01
ZF GATA	0	1.8	1.0E+00	0	3.5	1.0E+00	2	6.9	9.9E-01
ZF IBR	0	0.0	1.0E+00	0	0.0	1.0E+00	0	0.0	1.0E+00
ZF Phorbol	0	0.0	1.0E+00	0	0.0	1.0E+00	0	0.0	1.0E+00
ZF RING-CH	0	0.0	1.0E+00	0	0.0	1.0E+00	0	0.0	1.0E+00
ZF TRAF	0	0.2	1.0E+00	0	0.4	1.0E+00	1	0.9	5.8E-01
ZF UBP	0	0.2	1.0E+00	0	0.5	1.0E+00	0	0.8	1.0E+00

At each time point (days -8, 1, 14, and 28) within each of the four CD4+ T cell sources (AL, AP, ML, MP), t-statistics were calculated to reflect the change in expression from the time point to the next common time point (days 1, 14, 28 and 65). The extreme 2.5, 5 and 10% tails of the t-distributions for every time point and source were then examined. TF genes within these tails were added up giving total observed perturbations per TF family (OBS). Significance of TF family perturbations was estimated using the Poisson distribution, with mean equal to the expected number of perturbations given the number of TFs in the family (EXP). In the main text, these p-values were then corrected for multiple tests using the Benjamini-Hochberg correction (Benjamini *et al.* 2001).

References

Aguda, B. D., Y. Kim, et al. (2008). "MicroRNA regulation of a cancer network: consequences of the feedback loops involving miR-17-92, E2F, and Myc." Proc Natl Acad Sci U S A **105**(50): 19678-19683.

Aitken, C. E. and J. R. Lorsch (2012). "A mechanistic overview of translation initiation in eukaryotes." Nat Struct Mol Biol **19**(6): 568-576.

Albert, R. and A.-L. Barabasi (2002). "Statistical mechanics of complex networks." Reviews of Modern Physics **74**: 47-97.

Alberts, B., A. Johnson, et al. (2007). "Molecular Biology of the Cell."

Altuvia, Y., P. Landgraf, et al. (2005). "Clustering and conservation patterns of human microRNAs." Nucleic Acids Res **33**(8): 2697-2706.

Ambros, V., R. C. Lee, et al. (2003). "MicroRNAs and other tiny endogenous RNAs in C. elegans." Curr Biol **13**(10): 807-818.

Ameyar-Zazoua, M., C. Rachez, et al. (2012). "Argonaute proteins couple chromatin silencing to alternative splicing." Nat Struct Mol Biol **19**(10): 998-1004.

Anfinsen, C. B. (1973). "Principles that govern the folding of protein chains." Science **181**(4096): 223-230.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Avery, O. T., C. M. Macleod, et al. (1944). "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii." J Exp Med **79**(2): 137-158.

Ayyavoo, V., A. Mahboubi, et al. (1997). "HIV-1 Vpr suppresses immune activation and apoptosis through regulation of nuclear factor kappa B." Nat Med **3**(10): 1117-1123.

Babu, M. M., N. M. Luscombe, et al. (2004). "Structure and evolution of transcriptional regulatory networks." Curr Opin Struct Biol **14**(3): 283-291.

Baek, D., J. Villen, et al. (2008). "The impact of microRNAs on protein output." Nature **455**(7209): 64-71.

Balaji, S., M. M. Babu, et al. (2006). "Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast." J Mol Biol **360**(1): 213-227.

Banerjee, A., V. Pirrone, et al. (2011). "Transcriptional regulation of the chemokine co-receptor CCR5 by the cAMP/PKA/CREB pathway." Biomed Pharmacother **65**(4): 293-297.

Barski, A., S. Cuddapah, et al. (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-837.

Barski, A., R. Jothi, et al. (2009). "Chromatin poises miRNA- and protein-coding genes for expression." Genome Res **19**(10): 1742-1751.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-297.

Bartel, D. P. (2009). "MicroRNAs: target recognition and regulatory functions." Cell **136**(2): 215-233.

Baskerville, S. and D. P. Bartel (2005). "Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes." RNA **11**(3): 241-247.

Bassett, A., S. Cooper, et al. (2009). "The folding and unfolding of eukaryotic chromatin." Curr Opin Genet Dev **19**(2): 159-165.

Beadle, G. W. and E. L. Tatum (1941). "Genetic Control of Biochemical Reactions in Neurospora." Proc Natl Acad Sci U S A **27**(11): 499-506.

Benjamini, Y., D. Drai, et al. (2001). "Controlling the false discovery rate in behavior genetics research." Behav Brain Res **125**(1-2): 279-284.

Berezikov, E. (2011). "Evolution of microRNA diversity and regulation in animals." Nat Rev Genet **12**(12): 846-860.

Berezikov, E., W. J. Chung, et al. (2007). "Mammalian mirtron genes." Mol Cell **28**(2): 328-336.

Berget, S. M., C. Moore, et al. (1977). "Spliced segments at the 5' terminus of adenovirus 2 late mRNA." Proc Natl Acad Sci U S A **74**(8): 3171-3175.
Bernardi, R. and P. P. Pandolfi (2007). "Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies." Nat Rev Mol Cell Biol **8**(12): 1006-1016.
Bernstein, E., A. A. Caudy, et al. (2001). "Role for a bidentate ribonuclease in the initiation step of RNA interference." Nature **409**(6818): 363-366.
Bernstein, P. and J. Ross (1989). "Poly(A), poly(A) binding protein and the regulation of mRNA stability." Trends Biochem Sci **14**(9): 373-377.
Betel, D., M. Wilson, et al. (2008). "The microRNA.org resource: targets and expression." Nucleic Acids Res **36**(Database issue): D149-153.
Bhattacharyya, S. N., R. Habermacher, et al. (2006). "Stress-induced reversal of microRNA repression and mRNA P-body localization in human cells." Cold Spring Harb Symp Quant Biol **71**: 513-521.
Biasiolo, M., G. Sales, et al. (2011). "Impact of host genes and strand selection on miRNA and miRNA* expression." PLoS One **6**(8): e23854.
Blahna, M. T. and A. Hata (2013). "Regulation of miRNA biogenesis as an integrated component of growth factor signaling." Curr Opin Cell Biol.
Blattner, F. R., G. Plunkett, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12." Science **277**(5331): 1453-1462.
Bonev, B., P. Stanley, et al. (2012). "MicroRNA-9 Modulates Hes1 ultradian oscillations by forming a double-negative feedback loop." Cell Rep **2**(1): 10-18.
Borchert, G. M., W. Lanier, et al. (2006). "RNA polymerase III transcribes human microRNAs." Nat Struct Mol Biol **13**(12): 1097-1101.
Borman, A. M., Y. M. Michel, et al. (2002). "Free poly(A) stimulates capped mRNA translation in vitro through the eIF4G-poly(A)-binding protein interaction." J Biol Chem **277**(39): 36818-36824.
Bosinger, S. E., S. P. Jochems, et al. (2013). "Transcriptional profiling of experimental CD8(+) lymphocyte depletion in rhesus macaques infected with simian immunodeficiency virus SIVmac239." J Virol **87**(1): 433-443.
Bosinger, S. E., Q. Li, et al. (2009). "Global genomic analysis reveals rapid control of a robust innate response in SIV-infected sooty mangabeys." J Clin Invest **119**(12): 3556-3572.
Boyer, L. A., K. Plath, et al. (2006). "Polycomb complexes repress developmental regulators in murine embryonic stem cells." Nature **441**(7091): 349-353.
Brabletz, S., K. Bajdak, et al. (2011). "The ZEB1/miR-200 feedback loop controls Notch signalling in cancer cells." EMBO J **30**(4): 770-782.
Brenner, J. L., K. L. Jasiewicz, et al. (2010). "Loss of individual microRNAs causes mutant phenotypes in sensitized genetic backgrounds in C. elegans." Curr Biol **20**(14): 1321-1325.
Brivanlou, A. H. and J. E. Darnell, Jr. (2002). "Signal transduction and the control of gene expression." Science **295**(5556): 813-818.
Brown, M. S. and J. L. Goldstein (1997). "The SREBP pathway: regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor." Cell **89**(3): 331-340.
Burk, U., J. Schubert, et al. (2008). "A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells." EMBO Rep **9**(6): 582-589.
Cagliani, R., S. Riva, et al. (2011). "A positively selected APOBEC3H haplotype is associated with natural resistance to HIV-1 infection." Evolution **65**(11): 3311-3322.
Cai, X., C. H. Hagedorn, et al. (2004). "Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs." RNA **10**(12): 1957-1966.
Calin, G. A. and C. M. Croce (2006). "MicroRNA signatures in human cancers." Nat Rev Cancer **6**(11): 857-866.
Campo-Paysaa, F., M. Semon, et al. (2011). "microRNA complements in deuterostomes: origin and evolution of microRNAs." Evol Dev **13**(1): 15-27.

Carninci, P., T. Kasukawa, et al. (2005). "The transcriptional landscape of the mammalian genome." Science **309**(5740): 1559-1563.
Carninci, P., A. Sandelin, et al. (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-635.
Carrington, J. C. and V. Ambros (2003). "Role of microRNAs in plant and animal development." Science **301**(5631): 336-338.
Celniker, S. E., L. A. Dillon, et al. (2009). "Unlocking the secrets of the genome." Nature **459**(7249): 927-930.
Chang, T. C., E. A. Wentzel, et al. (2007). "Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis." Mol Cell **26**(5): 745-752.
Chen, X. (2005). "MicroRNA biogenesis and function in plants." FEBS Lett **579**(26): 5923-5931.
Cheng, C., K. K. Yan, et al. (2011). "Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data." PLoS Comput Biol **7**(11): e1002190.
Chien, C. H., Y. M. Sun, et al. (2011). "Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data." Nucleic Acids Res **39**(21): 9345-9356.
Chiu, I. M., K. Touhalisky, et al. (2001). "Multiple controlling mechanisms of FGF1 gene expression through multiple tissue-specific promoters." Prog Nucleic Acid Res Mol Biol **70**: 155-174.
Choi, Y. H., R. Bernardi, et al. (2006). "The promyelocytic leukemia protein functions as a negative regulator of IFN-gamma signaling." Proc Natl Acad Sci U S A **103**(49): 18715-18720.
Chow, L. T., R. E. Gelinas, et al. (1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA." Cell **12**(1): 1-8.
Ciofani, M., A. Madar, et al. (2012). "A validated regulatory network for Th17 cell specification." Cell **151**(2): 289-303.
Claus, H., Van Crugden, A. (1983). "Het verdriet van België." ch.42
Coleman, J., P. J. Green, et al. (1984). "The use of RNAs complementary to specific mRNAs to regulate the expression of individual bacterial genes." Cell **37**(2): 429-436.
Conti, L., L. Fantuzzi, et al. (2004). "Immunomodulatory effects of the HIV-1 gp120 protein on antigen presenting cells: implications for AIDS pathogenesis." Immunobiology **209**(1-2): 99-115.
Corcoran, D. L., K. V. Pandit, et al. (2009). "Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data." PLoS One **4**(4): e5279.
Cowley, S. M., B. M. Iritani, et al. (2005). "The mSin3A chromatin-modifying complex is essential for embryogenesis and T-cell development." Mol Cell Biol **25**(16): 6990-7004.
Cox, A. D. and C. J. Der (2003). "The dark side of Ras: regulation of apoptosis." Oncogene **22**(56): 8999-9006.
Crick, F. (1958). On Protein Synthesis. Symp Soc Exp Biol.
Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-163.
Crick, F. H. (1963). "On the genetic code." Science **139**(3554): 461-464.
Crick, F. H., L. Barnett, et al. (1961). "General nature of the genetic code for proteins." Nature **192**: 1227-1232.
Cui, Q., Z. Yu, et al. (2006). "Principles of microRNA regulation of a human cellular signaling network." Mol Syst Biol **2**: 46.
Cui, X., S. M. Xu, et al. (2009). "Genomic analysis of rice microRNA promoters and clusters." Gene **431**(1-2): 61-66.
Czech, B., R. Zhou, et al. (2009). "Hierarchical rules for Argonaute loading in Drosophila." Mol Cell **36**(3): 445-456.

Darnell, J. E., Jr., I. M. Kerr, et al. (1994). "Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins." Science **264**(5164): 1415-1421.
Das, R. and D. Baker (2008). "Macromolecular modeling with rosetta." Annu Rev Biochem **77**: 363-382.
David, C. J. and J. L. Manley (2011). "The RNA polymerase C-terminal domain: a new role in spliceosome assembly." Transcription **2**(5): 221-225.
Davila Lopez, M. and T. Samuelsson (2008). "Early evolution of histone mRNA 3' end processing." RNA **14**(1): 1-10.
de Veer, M. J., M. Holko, et al. (2001). "Functional classification of interferon-stimulated genes identified using microarrays." J Leukoc Biol **69**(6): 912-920.
de Wit, E., S. E. Linsen, et al. (2009). "Repertoire and evolution of miRNA genes in four divergent nematode species." Genome Res **19**(11): 2064-2074.
Deaton, A. M. and A. Bird (2011). "CpG islands and the regulation of transcription." Genes Dev **25**(10): 1010-1022.
Denli, A. M., B. B. Tops, et al. (2004). "Processing of primary microRNAs by the Microprocessor complex." Nature **432**(7014): 231-235.
Dennis, G., Jr., B. T. Sherman, et al. (2003). "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol **4**(5): P3.
Di Mascio, M., C. H. Paik, et al. (2009). "Noninvasive in vivo imaging of CD4 cells in simian-human immunodeficiency virus (SHIV)-infected nonhuman primates." Blood **114**(2): 328-337.
Dieci, G., M. Preti, et al. (2009). "Eukaryotic snoRNAs: a paradigm for gene expression flexibility." Genomics **94**(2): 83-88.
Ding, X. C., J. Weiler, et al. (2009). "Regulating the regulators: mechanisms controlling the maturation of microRNAs." Trends Biotechnol **27**(1): 27-36.
Ding, Y., J. Xu, et al. (2012). "Regulatory T cell migration during an immune response." Trends Immunol **33**(4): 174-180.
Diop, O. M., M. J. Ploquin, et al. (2008). "Plasmacytoid dendritic cell dynamics and alpha interferon production during Simian immunodeficiency virus infection with a nonpathogenic outcome." J Virol **82**(11): 5145-5152.
Doench, J. G. and P. A. Sharp (2004). "Specificity of microRNA target selection in translational repression." Genes Dev **18**(5): 504-511.
Domazet-Loso, T. and D. Tautz (2010). "Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa." BMC Biol **8**: 66.
Driskell, I., H. Oda, et al. (2012). "The histone methyltransferase Setd8 acts in concert with c-Myc and is required to maintain skin." EMBO J **31**(3): 616-629.
Ecker, J. R. and R. W. Davis (1986). "Inhibition of gene expression in plant cells by expression of antisense RNA." Proc Natl Acad Sci U S A **83**(15): 5372-5376.
Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.
Edwards, Y. J., A. E. Lobley, et al. (2009). "Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data." Genome Biol **10**(5): R50.
Eick, D., A. Wedel, et al. (1994). "From initiation to elongation: comparison of transcription by prokaryotic and eukaryotic RNA polymerases." Trends Genet **10**(8): 292-296.
ENCODE (2011). "A user's guide to the encyclopedia of DNA elements (ENCODE)." PLoS Biol **9**(4): e1001046.
Enright, A. J., B. John, et al. (2003). "MicroRNA targets in Drosophila." Genome Biol **5**(1): R1.
Erdel, F. and K. Rippe (2011). "Chromatin remodelling in mammalian cells by ISWI-type complexes--where, when and why?" FEBS J **278**(19): 3608-3618.
Ernst, J. and M. Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nat Biotechnol **28**(8): 817-825.

Ernst, J., P. Kheradpour, et al. (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types." Nature **473**(7345): 43-49.
Estes, J. D., S. N. Gordon, et al. (2008). "Early resolution of acute immune activation and induction of PD-1 in SIV-infected sooty mangabeys distinguishes nonpathogenic from pathogenic infection in rhesus macaques." J Immunol **180**(10): 6798-6807.
Evans, R. M. (1988). "The steroid and thyroid hormone receptor superfamily." Science **240**(4854): 889-895.
Fabian, M. R. and N. Sonenberg (2012). "The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC." Nat Struct Mol Biol **19**(6): 586-593.
Fagnani, M., Y. Barash, et al. (2007). "Functional coordination of alternative splicing in the mammalian central nervous system." Genome Biol **8**(6): R108.
Fedorova, L. and A. Fedorov (2003). "Introns in gene evolution." Genetica **118**(2-3): 123-131.
Filipowicz, W., S. N. Bhattacharyya, et al. (2008). "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?" Nat Rev Genet **9**(2): 102-114.
Flynt, A. S. and E. C. Lai (2008). "Biological principles of microRNA-mediated regulation: shared themes amid diversity." Nat Rev Genet **9**(11): 831-842.
Franco-Zorrilla, J. M., A. Valli, et al. (2007). "Target mimicry provides a new mechanism for regulation of microRNA activity." Nat Genet **39**(8): 1033-1037.
Friedman, R. C., K. K. Farh, et al. (2009). "Most mammalian mRNAs are conserved targets of microRNAs." Genome Res **19**(1): 92-105.
Friend, K., Z. T. Campbell, et al. (2012). "A conserved PUF-Ago-eEF1A complex attenuates translation elongation." Nat Struct Mol Biol **19**(2): 176-183.
Fu, W., B. E. Sanders-Beer, et al. (2009). "Human immunodeficiency virus type 1, human protein interaction database at NCBI." Nucleic Acids Res **37**(Database issue): D417-422.
Fu, Y., M. Sinha, et al. (2008). "The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome." PLoS Genet **4**(7): e1000138.
Gamow, G. (1954). "Possible relation between deoxyribonucleic acid and protein structure." Nature **173**(318).
Gao, B., H. Wang, et al. (2012). "STAT proteins - key regulators of anti-viral responses, inflammation, and tumorigenesis in the liver." J Hepatol **57**(2): 430-441.
Gao, Y. F., T. Li, et al. (2011). "Cdk1-phosphorylated CUEDC2 promotes spindle checkpoint inactivation and chromosomal instability." Nat Cell Biol **13**(8): 924-933.
Gardner, M. B. (1996). "The history of simian AIDS." J Med Primatol **25**(3): 148-157.
Geoffroy, M. C. and M. K. Chelbi-Alix (2011). "Role of promyelocytic leukemia protein in host antiviral defense." J Interferon Cytokine Res **31**(1): 145-158.
Gerstein, M. B., A. Kundaje, et al. (2012). "Architecture of the human regulatory network derived from ENCODE data." Nature **489**(7414): 91-100.
Gerstein, M. B., Z. J. Lu, et al. (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science **330**(6012): 1775-1787.
Ghildiyal, M., H. Seitz, et al. (2008). "Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells." Science **320**(5879): 1077-1081.
Ghildiyal, M., J. Xu, et al. (2010). "Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway." RNA **16**(1): 43-56.
Ghislain, J. J., T. Wong, et al. (2001). "The interferon-inducible Stat2:Stat1 heterodimer preferentially binds in vitro to a consensus element found in the promoters of a subset of interferon-stimulated genes." J Interferon Cytokine Res **21**(6): 379-388.
Ghosh, T., K. Soni, et al. (2008). "MicroRNA-mediated up-regulation of an alternatively polyadenylated variant of the mouse cytoplasmic {beta}-actin gene." Nucleic Acids Res **36**(19): 6318-6332.
Gilbert, W. V. (2010). "Alternative ways to think about cellular internal ribosome entry." J Biol Chem **285**(38): 29033-29038.

Gilmour, D. S. and J. T. Lis (1984). "Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes." Proc Natl Acad Sci U S A **81**(14): 4275-4279.
Giorgi, J. V., J. L. Fahey, et al. (1987). "Early effects of HIV on CD4 lymphocytes in vivo." J Immunol **138**(11): 3725-3730.
Glover, J. N. and S. C. Harrison (1995). "Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA." Nature **373**(6511): 257-261.
Goffeau, A., B. G. Barrell, et al. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-547.
Golan, D., C. Levy, et al. (2010). "Biased hosting of intronic microRNA genes." Bioinformatics **26**(8): 992-995.
Goldman, S. R., R. H. Ebright, et al. (2009). "Direct detection of abortive RNA transcripts in vivo." Science **324**(5929): 927-928.
Goodman, R. H. and S. Smolik (2000). "CBP/p300 in cell growth, transformation, and development." Genes Dev **14**(13): 1553-1577.
Gregory, R. I., K. P. Yan, et al. (2004). "The Microprocessor complex mediates the genesis of microRNAs." Nature **432**(7014): 235-240.
Griffiths-Jones, S. (2007). "Annotating noncoding RNA genes." Annu Rev Genomics Hum Genet **8**: 279-298.
Griffiths-Jones, S., R. J. Grocock, et al. (2006). "miRBase: microRNA sequences, targets and gene nomenclature." Nucleic Acids Res **34**(Database issue): D140-144.
Griffiths-Jones, S., J. H. Hui, et al. (2011). "MicroRNA evolution by arm switching." EMBO Rep **12**(2): 172-177.
Grimson, A., K. K. Farh, et al. (2007). "MicroRNA targeting specificity in mammals: determinants beyond seed pairing." Mol Cell **27**(1): 91-105.
Grun, D., Y. L. Wang, et al. (2005). "microRNA target predictions across seven Drosophila species and comparison to mammalian targets." PLoS Comput Biol **1**(1): e13.
Guan, Y. J., X. Yang, et al. (2011). "MiR-365: a mechanosensitive microRNA stimulates chondrocyte differentiation through targeting histone deacetylase 4." FASEB J **25**(12): 4457-4466.
Gulsuner, S., A. B. Tekinay, et al. (2011). "Homozygosity mapping and targeted genomic sequencing reveal the gene responsible for cerebellar hypoplasia and quadrupedal locomotion in a consanguineous kindred." Genome Res **21**(12): 1995-2003.
Gupta, R., P. Wikramasinghe, et al. (2010). "Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data." BMC Bioinformatics **11 Suppl 1**: S65.
Gyrd-Hansen, M. and P. Meier (2010). "IAPs: from caspase inhibitors to modulators of NF-kappaB, inflammation and cancer." Nat Rev Cancer **10**(8): 561-574.
Hagen, J. W. and E. C. Lai (2008). "microRNA control of cell-cell signaling during development and disease." Cell Cycle **7**(15): 2327-2332.
Hagiwara, M. and K. Nagata (2012). "Redox-dependent protein quality control in the endoplasmic reticulum: folding to degradation." Antioxid Redox Signal **16**(10): 1119-1128.
Hahn, S. (1992). "The Yin and the Yang of mammalian transcription." Curr Biol **2**(3): 152-154.
Hall, S. L. and R. A. Padgett (1996). "Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns." Science **271**(5256): 1716-1718.
Han, J. W., H. J. Lee, et al. (2011). "Promyogenic function of Integrin/FAK signaling is mediated by Cdo, Cdc42 and MyoD." Cell Signal **23**(7): 1162-1169.
Han, M. and M. Grunstein (1988). "Nucleosome loss activates yeast downstream promoters in vivo." Cell **55**(6): 1137-1145.
Hanash, S. (2003). "Disease proteomics." Nature **422**(6928): 226-232.
Hansen, T. B., T. I. Jensen, et al. (2013). "Natural RNA circles function as efficient microRNA sponges." Nature **495**(7441): 384-388.

Hariharan, D., S. D. Douglas, et al. (1999). "Interferon-gamma upregulates CCR5 expression in cord and adult blood mononuclear phagocytes." Blood **93**(4): 1137-1144.
Hazenberg, M. D., S. A. Otto, et al. (2003). "Persistent immune activation in HIV-1 infection is associated with progression to AIDS." AIDS **17**(13): 1881-1888.
He, C., Z. Li, et al. (2012). "Young intragenic miRNAs are less coexpressed with host genes than old ones: implications of miRNA-host gene coevolution." Nucleic Acids Res **40**(9): 4002-4012.
He, L., X. He, et al. (2007). "microRNAs join the p53 network--another piece in the tumour-suppression puzzle." Nat Rev Cancer **7**(11): 819-822.
Hedges, S. B., J. Dudley, et al. (2006). "TimeTree: a public knowledge-base of divergence times among organisms." Bioinformatics **22**(23): 2971-2972.
Heintzman, N. D., R. K. Stuart, et al. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nat Genet **39**(3): 311-318.
Hentschel, C. C. and M. L. Birnstiel (1981). "The organization and expression of histone gene families." Cell **25**(2): 301-313.
Herranz, H. and S. M. Cohen (2010). "MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems." Genes Dev **24**(13): 1339-1344.
Herrmann, C., B. Van de Sande, et al. (2012). "i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules." Nucleic Acids Res **40**(15): e114.
Hershey, A. D. and M. Chase (1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage." J Gen Physiol **36**(1): 39-56.
Hinrichs, A. S., D. Karolchik, et al. (2006). "The UCSC Genome Browser Database: update 2006." Nucleic Acids Res **34**(Database issue): D590-598.
Hobert, O. (2006). "Architecture of a microRNA-controlled gene regulatory network that diversifies neuronal cell fates." Cold Spring Harb Symp Quant Biol **71**: 181-188.
Hoeppner, M. P., S. White, et al. (2009). "Evolutionarily stable association of intronic snoRNAs and microRNAs with their host genes." Genome Biol Evol **1**: 420-428.
Holley, R. W., J. Apgar, et al. (1965). "Structure of a Ribonucleic Acid." Science **147**(3664): 1462-1465.
Honda, K., T. Mizutani, et al. (2004). "Negative regulation of IFN-alpha/beta signaling by IFN regulatory factor 2 for homeostatic development of dendritic cells." Proc Natl Acad Sci U S A **101**(8): 2416-2421.
Honda, K., H. Yanai, et al. (2005). "IRF-7 is the master regulator of type-I interferon-dependent immune responses." Nature **434**(7034): 772-777.
Howard, M. L. and E. H. Davidson (2004). "cis-Regulatory control circuits in development." Dev Biol **271**(1): 109-118.
Hristova, M., D. Birse, et al. (2005). "The Caenorhabditis elegans heterochronic regulator LIN-14 is a novel transcription factor that controls the developmental timing of transcription from the insulin/insulin-like growth factor gene ins-33 by direct DNA binding." Mol Cell Biol **25**(24): 11059-11072.
Huang da, W., B. T. Sherman, et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.
Huang, S., S. Wang, et al. (2012). "Upregulation of miR-22 promotes osteogenic differentiation and inhibits adipogenic differentiation of human adipose tissue-derived mesenchymal stem cells by repressing HDAC6 protein expression." Stem Cells Dev **21**(13): 2531-2540.
Hurst, D. R., M. D. Edmonds, et al. (2009). "Breast cancer metastasis suppressor 1 up-regulates miR-146, which suppresses breast cancer metastasis." Cancer Res **69**(4): 1279-1283.
Hurwitz, J., A. Bresler, et al. (1960). The Enzymatic Incorporation of Ribonucleotides into RNA and the Role of DNA Biochem Biophys Res Commun.

Ito, M. and R. G. Roeder (2001). "The TRAP/SMCC/Mediator complex and thyroid hormone receptor function." Trends Endocrinol Metab **12**(3): 127-134.
Izmailova, E., F. M. Bertley, et al. (2003). "HIV-1 Tat reprograms immature dendritic cells to express chemoattractants for activated T cells and macrophages." Nat Med **9**(2): 191-197.
Jacob, F., D. Perrin, et al. (1960). "[Operon: a group of genes with the expression coordinated by an operator]." C R Hebd Seances Acad Sci **250**: 1727-1729.
Jacquelin, B., V. Mayau, et al. (2009). "Nonpathogenic SIV infection of African green monkeys induces a strong but rapidly controlled type I IFN response." J Clin Invest **119**(12): 3544-3555.
John, B., A. J. Enright, et al. (2004). "Human MicroRNA targets." PLoS Biol **2**(11): e363.
Johnson, D. S., A. Mortazavi, et al. (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science **316**(5830): 1497-1502.
Johnson, S. M., S. Y. Lin, et al. (2003). "The time of appearance of the C. elegans let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter." Dev Biol **259**(2): 364-379.
Johnston, R. J. and O. Hobert (2003). "A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans." Nature **426**(6968): 845-849.
Juven-Gershon, T., J. Y. Hsu, et al. (2008). "The RNA polymerase II core promoter - the gateway to transcription." Curr Opin Cell Biol **20**(3): 253-259.
Karlic, R., H. R. Chung, et al. (2010). "Histone modification levels are predictive for gene expression." Proc Natl Acad Sci U S A **107**(7): 2926-2931.
Karolchik, D., R. Baertsch, et al. (2003). "The UCSC Genome Browser Database." Nucleic Acids Res **31**(1): 51-54.
Karolchik, D., A. S. Hinrichs, et al. (2004). "The UCSC Table Browser data retrieval tool." Nucleic Acids Res **32**(Database issue): D493-496.
Kasprzyk, A. (2011). "BioMart: driving a paradigm change in biological data management." Database (Oxford) **2011**: bar049.
Katakowski, M., X. Zheng, et al. (2010). "MiR-146b-5p suppresses EGFR expression and reduces in vitro migration and invasion of glioma." Cancer Invest **28**(10): 1024-1030.
Katze, M. G., Y. He, et al. (2002). "Viruses and interferon: a fight for supremacy." Nat Rev Immunol **2**(9): 675-687.
Kertesz, M., N. Iovino, et al. (2007). "The role of site accessibility in microRNA target recognition." Nat Genet **39**(10): 1278-1284.
Kieslinger, M., S. Hiechinger, et al. (2010). "Early B cell factor 2 regulates hematopoietic stem cell homeostasis in a cell-nonautonomous manner." Cell Stem Cell **7**(4): 496-507.
Kim, D. H., P. Saetrom, et al. (2008). "MicroRNA-directed transcriptional gene silencing in mammalian cells." Proc Natl Acad Sci U S A **105**(42): 16230-16235.
Kim, N., S. Kukkonen, et al. (2013). "Tat engagement of p38 MAP kinase and IRF7 pathways leads to activation of interferon-stimulated genes in antigen-presenting cells." Blood **121**(20): 4090-4100.
Kim, T. H., Z. K. Abdullaev, et al. (2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." Cell **128**(6): 1231-1245.
Kim, T. K., R. H. Ebright, et al. (2000). "Mechanism of ATP-dependent promoter melting by transcription factor IIH." Science **288**(5470): 1418-1422.
Kim, Y. J., S. Bjorklund, et al. (1994). "A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II." Cell **77**(4): 599-608.
Kim, Y. K. and V. N. Kim (2007). "Processing of intronic microRNAs." EMBO J **26**(3): 775-783.
Kim, Y. K., J. Yu, et al. (2009). "Functional links between clustered microRNAs: suppression of cell-cycle inhibitors by microRNA clusters in gastric cancer." Nucleic Acids Res **37**(5): 1672-1681.

Kinsella, R. J., A. Kahari, et al. (2011). "Ensembl BioMarts: a hub for data retrieval across taxonomic space." Database (Oxford) **2011**: bar030.
Klaus, G. G., M. K. Bijsterbosch, et al. (1987). "Receptor signalling and crosstalk in B lymphocytes." Immunol Rev **99**: 19-38.
Kornberg, R. D. (1974). "Chromatin structure: a repeating unit of histones and DNA." Science **184**(4139): 868-871.
Kornfeld, C., M. J. Ploquin, et al. (2005). "Antiinflammatory profiles during primary SIV infection in African green monkeys are associated with protection against AIDS." J Clin Invest **115**(4): 1082-1091.
Kossel, A. (1928). The protamines and histones. London, New York etc., Longmans, Green and co.
Kosti, I., P. Radivojac, et al. (2012). "An integrated regulatory network reveals pervasive cross-regulation among transcription and splicing factors." PLoS Comput Biol **8**(7): e1002603.
Kowalczyk, M. S., J. R. Hughes, et al. (2012). "Intragenic enhancers act as alternative promoters." Mol Cell **45**(4): 447-458.
Kozomara, A. and S. Griffiths-Jones (2011). "miRBase: integrating microRNA annotation and deep-sequencing data." Nucleic Acids Res **39**(Database issue): D152-157.
Krek, A., D. Grun, et al. (2005). "Combinatorial microRNA target predictions." Nat Genet **37**(5): 495-500.
Kumar, L. and M. E. Futschik (2007). "Mfuzz: a software package for soft clustering of microarray data." Bioinformation **2**(1): 5-7.
Kushwah, R. and J. Hu (2011). "Complexity of dendritic cell subsets and their function in the host immune system." Immunology **133**(4): 409-419.
Lachner, M. and T. Jenuwein (2002). "The many faces of histone lysine methylation." Curr Opin Cell Biol **14**(3): 286-298.
Lachner, M., D. O'Carroll, et al. (2001). "Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins." Nature **410**(6824): 116-120.
Lall, S., D. Grun, et al. (2006). "A genome-wide map of conserved microRNA targets in C. elegans." Curr Biol **16**(5): 460-471.
Landgraf, P., M. Rusu, et al. (2007). "A mammalian microRNA expression atlas based on small RNA library sequencing." Cell **129**(7): 1401-1414.
Larsson, M., E. M. Shankar, et al. (2013). "Molecular signatures of T-cell inhibition in HIV-1 infection." Retrovirology **10**(1): 31.
Lashkari, D. A., J. L. DeRisi, et al. (1997). "Yeast microarrays for genome wide parallel genetic and gene expression analysis." Proc Natl Acad Sci U S A **94**(24): 13057-13062.
Latchman, D. S. (1990). "Eukaryotic transcription factors." Biochem J **270**(2): 281-289.
Latchman, D. S. (1993). "Transcription factors: an overview." Int J Exp Pathol **74**(5): 417-422.
Lederer, S., D. Favre, et al. (2009). "Transcriptional profiling in pathogenic and non-pathogenic SIV infections reveals significant distinctions in kinetics and tissue compartmentalization." PLoS Pathog **5**(2): e1000296.
Lee, R. C., R. L. Feinbaum, et al. (1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." Cell **75**(5): 843-854.
Lee, T. I., N. J. Rinaldi, et al. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science **298**(5594): 799-804.
Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." EMBO J **23**(20): 4051-4060.
Leigh, E. G., Jr. (2010). "The group selection controversy." J Evol Biol **23**(1): 6-19.
Lewis, B. P., I. H. Shih, et al. (2003). "Prediction of mammalian microRNA targets." Cell **115**(7): 787-798.
Lewis, J. D., S. I. Gunderson, et al. (1995). "The influence of 5' and 3' end structures on pre-mRNA metabolism." J Cell Sci Suppl **19**: 13-19.
Li, E. (2002). "Chromatin modification and epigenetic reprogramming in mammalian development." Nat Rev Genet **3**(9): 662-673.

Li, J. C., H. C. Yim, et al. (2010). "Role of HIV-1 Tat in AIDS pathogenesis: its effects on cytokine dysregulation and contributions to the pathogenesis of opportunistic infection." AIDS **24**(11): 1609-1623.
Li, P., R. Spolski, et al. (2012). "BATF-JUN is critical for IRF4-mediated transcription in T cells." Nature **490**(7421): 543-546.
Lifton, R. P., M. L. Goldberg, et al. (1978). "The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications." Cold Spring Harb Symp Quant Biol **42 Pt 2**: 1047-1051.
Light, J. and S. Molin (1982). "The sites of action of the two copy number control functions of plasmid R1." Mol Gen Genet **187**(3): 486-493.
Light, J. and S. Molin (1983). "Post-transcriptional control of expression of the repA gene of plasmid R1 mediated by a small RNA molecule." EMBO J **2**(1): 93-98.
Ling, B., G. X. Wang, et al. (2012). "Tumor suppressor miR-22 suppresses lung cancer cell progression through post-transcriptional regulation of ErbB3." J Cancer Res Clin Oncol **138**(8): 1355-1361.
Liu, X., D. A. Bushnell, et al. (2013). "RNA polymerase II transcription: Structure and mechanism." Biochim Biophys Acta **1829**(1): 2-8.
Llave, C. (2010). "Virus-derived small interfering RNAs at the core of plant-virus interactions." Trends Plant Sci **15**(12): 701-707.
Logan, M. R., K. L. Jordan-Williams, et al. (2012). "Overexpression of Batf induces an apoptotic defect and an associated lymphoproliferative disorder in mice." Cell Death Dis **3**: e310.
Lu, J. and A. G. Clark (2012). "Impact of microRNA regulation on variation in human gene expression." Genome Res **22**(7): 1243-1254.
Lupino, E., C. Ramondetti, et al. (2012). "IkappaB kinase beta is required for activation of NF-kappaB and AP-1 in CD3/CD28-stimulated primary CD4(+) T cells." J Immunol **188**(6): 2545-2555.
Lutz, C. S. (2008). "Alternative polyadenylation: a twist on mRNA 3' end formation." ACS Chem Biol **3**(10): 609-617.
Lysholm, F. (2012). "Highly improved homopolymer aware nucleotide-protein alignments with 454 data." BMC Bioinformatics **13**: 230.
Ma, J. B., K. Ye, et al. (2004). "Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain." Nature **429**(6989): 318-322.
Magrane, M. and U. Consortium (2011). "UniProt Knowledgebase: a hub of integrated protein data." Database (Oxford) **2011**: bar009.
Maller Schulman, B. R., X. Liang, et al. (2008). "The let-7 microRNA target gene, Mlin41/Trim71 is required for mouse embryonic survival and neural tube closure." Cell Cycle **7**(24): 3935-3942.
Malleter, M., C. Jacquot, et al. (2012). "miRNAs, a potential target in the treatment of Non-Small-Cell Lung Carcinomas." Gene **506**(2): 355-359.
Man, J. and X. Zhang (2011). "CUEDC2: an emerging key player in inflammation and tumorigenesis." Protein Cell **2**(9): 699-703.
Mansfield, J. H. and E. McGlinn (2012). "Evolution, expression, and developmental function of Hox-embedded miRNAs." Curr Top Dev Biol **99**: 31-57.
Marco, A., M. Ninova, et al. (2013). "Clusters of microRNAs emerge by new hairpins in existing transcripts." Nucleic Acids Res.
Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.
Marson, A., S. S. Levine, et al. (2008). "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells." Cell **134**(3): 521-533.
Marx, P. A. and Z. Chen (1998). "The function of simian chemokine receptors in the replication of SIV." Semin Immunol **10**(3): 215-223.

Maskos, U. and E. M. Southern (1992). "Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ." Nucleic Acids Res **20**(7): 1679-1684.
Matthaei, J. H., O. W. Jones, et al. (1962). "Characteristics and composition of RNA coding units." Proc Natl Acad Sci U S A **48**: 666-677.
Maunakea, A. K., R. P. Nagarajan, et al. (2010). "Conserved role of intragenic DNA methylation in regulating alternative promoters." Nature **466**(7303): 253-257.
Megraw, M., V. Baev, et al. (2006). "MicroRNA promoter element discovery in Arabidopsis." RNA **12**(9): 1612-1619.
Megraw, M., F. Pereira, et al. (2009). "A transcription factor affinity-based code for mammalian transcription initiation." Genome Res **19**(4): 644-656.
Merchan, F., A. Boualem, et al. (2009). "Plant polycistronic precursors containing non-homologous microRNAs target transcripts encoding functionally related proteins." Genome Biol **10**(12): R136.
Meselson, M. and F. W. Stahl (1958). "The Replication of DNA in Escherichia Coli." Proc Natl Acad Sci U S A **44**(7): 671-682.
Meunier, J., F. Lemoine, et al. (2013). "Birth and expression evolution of mammalian microRNA genes." Genome Res **23**(1): 34-45.
Meyer, P. E., K. Kontos, et al. (2007). "Information-theoretic inference of large transcriptional regulatory networks." EURASIP J Bioinform Syst Biol: 79879.
Min Jou, W., G. Haegeman, et al. (1972). "Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein." Nature **237**(5350): 82-88.
Miska, E. A., E. Alvarez-Saavedra, et al. (2007). "Most Caenorhabditis elegans microRNAs are individually not essential for development or viability." PLoS Genet **3**(12): e215.
Moal, V., J. Textoris, et al. (2013). "Chronic hepatitis E virus infection is specifically associated with an interferon-related transcriptional program." J Infect Dis **207**(1): 125-132.
Monceaux, V., L. Viollet, et al. (2007). "CD4+ CCR5+ T-cell dynamics during simian immunodeficiency virus infection of Chinese rhesus macaques." J Virol **81**(24): 13865-13875.
Morgan, D., C. Mahe, et al. (2002). "Progression to symptomatic disease in people infected with HIV-1 in rural Uganda: prospective cohort study." BMJ **324**(7331): 193-196.
Morlando, M., M. Ballarino, et al. (2008). "Primary microRNA transcripts are processed co-transcriptionally." Nat Struct Mol Biol **15**(9): 902-909.
Morozova, N., A. Zinovyev, et al. (2012). "Kinetic signatures of microRNA modes of action." RNA **18**(9): 1635-1655.
Murphy, T. L., R. Tussiwand, et al. (2013). "Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks." Nat Rev Immunol **13**(7): 499-509.
Myer, V. E. and R. A. Young (1998). "RNA polymerase II holoenzymes and subcomplexes." J Biol Chem **273**(43): 27757-27760.
Niehrs, C. and N. Pollet (1999). "Synexpression groups in eukaryotes." Nature **402**(6761): 483-487.
Nilsen, T. W. (2007). "Mechanisms of microRNA-mediated gene regulation in animal cells." Trends Genet **23**(5): 243-249.
Nirenberg, M., P. Leder, et al. (1965). "RNA codewords and protein synthesis, VII. On the general nature of the RNA code." Proc Natl Acad Sci U S A **53**(5): 1161-1168.
Nutt, S. L., K. A. Fairfax, et al. (2007). "BLIMP1 guides the fate of effector B and T cells." Nat Rev Immunol **7**(12): 923-927.
O'Donnell, K. A., E. A. Wentzel, et al. (2005). "c-Myc-regulated microRNAs modulate E2F1 expression." Nature **435**(7043): 839-843.
O'Donoghue, P. and Z. Luthey-Schulten (2003). "On the evolution of structure in aminoacyl-tRNA synthetases." Microbiol Mol Biol Rev **67**(4): 550-573.
Okamura, K., J. W. Hagen, et al. (2007). "The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila." Cell **130**(1): 89-100.

Okamura, K., N. Liu, et al. (2009). "Distinct mechanisms for microRNA strand selection by Drosophila Argonautes." Mol Cell **36**(3): 431-444.
Osawa, S., T. H. Jukes, et al. (1992). "Recent evidence for evolution of the genetic code." Microbiol Rev **56**(1): 229-264.
Osella, M., C. Bosia, et al. (2011). "The role of incoherent microRNA-mediated feedforward loops in noise buffering." PLoS Comput Biol **7**(3): e1001101.
Ouzounis, C. A. and N. C. Kyrpides (1996). "Parallel origins of the nucleosome core and eukaryotic transcription from Archaea." J Mol Evol **42**(2): 234-239.
Ozsolak, F., L. L. Poling, et al. (2008). "Chromatin structure analyses identify miRNA promoters." Genes Dev **22**(22): 3172-3183.
Paiardini, M., I. Pandrea, et al. (2009). "Lessons learned from the natural hosts of HIV-related viruses." Annu Rev Med **60**: 485-495.
Pal, M., A. S. Ponticelli, et al. (2005). "The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II." Mol Cell **19**(1): 101-110.
Palade, G. E. (1955). "A small particulate component of the cytoplasm." J Biophys Biochem Cytol **1**(1): 59-68.
Pearson, G., F. Robinson, et al. (2001). "Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions." Endocr Rev **22**(2): 153-183.
Pedersen, A. G., P. Baldi, et al. (1999). "The biology of eukaryotic promoter prediction--a review." Comput Chem **23**(3-4): 191-207.
Pencheva, N. and S. F. Tavazoie (2013). "Control of metastatic progression by microRNA regulatory networks." Nat Cell Biol **15**(6): 546-554.
Peterson, K. J., M. R. Dietrich, et al. (2009). "MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion." Bioessays **31**(7): 736-747.
Pichon, X., L. A. Wilson, et al. (2012). "RNA binding protein/RNA element interactions and the control of translation." Curr Protein Pept Sci **13**(4): 294-304.
Picker, L. J. (2006). "Immunopathogenesis of acute AIDS virus infection." Curr Opin Immunol **18**(4): 399-405.
Pigazzi, M., E. Manara, et al. (2009). "miR-34b targets cyclic AMP-responsive element binding protein in acute myeloid leukemia." Cancer Res **69**(6): 2471-2478.
Piskacek, S., M. Gregor, et al. (2007). "Nine-amino-acid transactivation domain: establishment and prediction utilities." Genomics **89**(6): 756-768.
Pleiss, J. A., G. B. Whitworth, et al. (2007). "Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components." PLoS Biol **5**(4): e90.
Porter, K. R., A. Claude, et al. (1945). "A Study of Tissue Culture Cells by Electron Microscopy : Methods and Preliminary Observations." J Exp Med **81**(3): 233-246.
Potapov, A. P., N. Voss, et al. (2005). "Topology of mammalian transcription networks." Genome Inform **16**(2): 270-278.
Prabakaran, S., G. Lippens, et al. (2012). "Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding." Wiley Interdiscip Rev Syst Biol Med **4**(6): 565-583.
Pruitt, K. D., T. Tatusova, et al. (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **35**(Database issue): D61-65.
Ptak, R. G., W. Fu, et al. (2008). "Cataloguing the HIV type 1 human protein interaction network." AIDS Res Hum Retroviruses **24**(12): 1497-1502.
Qiao, X., B. He, et al. (2006). "Human immunodeficiency virus 1 Nef suppresses CD40-dependent immunoglobulin class switching in bystander B cells." Nat Immunol **7**(3): 302-310.
Quigley, M., F. Pereyra, et al. (2010). "Transcriptional analysis of HIV-specific CD8+ T cells shows that PD-1 inhibits T cell function by upregulating BATF." Nat Med **16**(10): 1147-1151.

Radtke, F., A. Wilson, et al. (2004). "Notch signaling in T- and B-cell development." Curr Opin Immunol **16**(2): 174-179.
Rajabi, H. N., S. Baluchamy, et al. (2005). "Effects of depletion of CREB-binding protein on c-Myc regulation and cell cycle G1-S transition." J Biol Chem **280**(1): 361-374.
Ramirez, J. and J. Hagman (2009). "The Mi-2/NuRD complex: a critical epigenetic regulator of hematopoietic development, differentiation and cancer." Epigenetics **4**(8): 532-536.
Ramskold, D., E. T. Wang, et al. (2009). "An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data." PLoS Comput Biol **5**(12): e1000598.
Ranish, J. A., N. Yudkovsky, et al. (1999). "Intermediates in formation and activity of the RNA polymerase II preinitiation complex: holoenzyme recruitment and a postrecruitment role for the TATA box and TFIIB." Genes Dev **13**(1): 49-63.
Reddy, R., R. Singh, et al. (1992). "Methylated cap structures in eukaryotic RNAs: structure, synthesis and functions." Pharmacol Ther **54**(3): 249-267.
Rhoades, M. W., B. J. Reinhart, et al. (2002). "Prediction of plant microRNA targets." Cell **110**(4): 513-520.
Rodriguez, A., S. Griffiths-Jones, et al. (2004). "Identification of mammalian microRNA host genes and transcription units." Genome Res **14**(10A): 1902-1910.
Rotger, M., J. Dalmau, et al. (2011). "Comparative transcriptomics of extreme phenotypes of human HIV-1 infection and SIV infection in sooty mangabey and rhesus macaque." J Clin Invest **121**(6): 2391-2400.
Rothenberg, E. V. and R. Pant (2004). "Origins of lymphocyte developmental programs: transcription factor evidence." Semin Immunol **16**(4): 227-238.
Rougvie, A. E. (2001). "Control of developmental timing in animals." Nat Rev Genet **2**(9): 690-701.
Roussigne, M., C. Cayrol, et al. (2003). "THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies." Oncogene **22**(16): 2432-2442.
Roy, S., J. Ernst, et al. (2010). "Identification of functional elements and regulatory circuits by Drosophila modENCODE." Science **330**(6012): 1787-1797.
Rozowsky, J., G. Euskirchen, et al. (2009). "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls." Nat Biotechnol **27**(1): 66-75.
Russo, G., C. Zegar, et al. (2003). "Advantages and limitations of microarray technology in human cancer." Oncogene **22**(42): 6497-6507.
Saetrom, P., B. S. Heale, et al. (2007). "Distance constraints between microRNA target sites dictate efficacy and cooperativity." Nucleic Acids Res **35**(7): 2333-2342.
Saini, H. K., A. J. Enright, et al. (2008). "Annotation of mammalian primary microRNAs." BMC Genomics **9**: 564.
Saini, H. K., S. Griffiths-Jones, et al. (2007). "Genomic analysis of human microRNA transcripts." Proc Natl Acad Sci U S A **104**(45): 17719-17724.
Saj, A. and E. C. Lai (2011). "Control of microRNA biogenesis and transcription by cell signaling pathways." Curr Opin Genet Dev **21**(4): 504-510.
Saltzman, A. L., Q. Pan, et al. (2011). "Regulation of alternative splicing by the core spliceosomal machinery." Genes Dev **25**(4): 373-384.
Sanger, F., S. Nicklen, et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.
Sarac, O., S. Gulsuner, et al. (2012). "Neuro-ophthalmologic findings in humans with quadrupedal locomotion." Ophthalmic Genet **33**(4): 249-252.
Sawadogo, M. and A. Sentenac (1990). "RNA polymerase B (II) and general transcription factors." Annu Rev Biochem **59**: 711-754.
Saxonov, S., P. Berg, et al. (2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters." Proc Natl Acad Sci U S A **103**(5): 1412-1417.

Schad, E., P. Tompa, et al. (2011). "The relationship between proteome size, structural disorder and organism complexity." Genome Biol **12**(12): R120.
Schones, D. E., K. Cui, et al. (2008). "Dynamic regulation of nucleosome positioning in the human genome." Cell **132**(5): 887-898.
Schughart, K., C. Kappen, et al. (1988). "Mammalian homeobox-containing genes: genome organization, structure, expression and evolution." Br J Cancer Suppl **9**: 9-13.
Seila, A. C., J. M. Calabrese, et al. (2008). "Divergent transcription from active promoters." Science **322**(5909): 1849-1851.
Sekine, S., S. Tagami, et al. (2012). "Structural basis of transcription by bacterial and eukaryotic RNA polymerases." Curr Opin Struct Biol **22**(1): 110-118.
Shalgi, R., D. Lieber, et al. (2007). "Global and local architecture of the mammalian microRNA-transcription factor regulatory network." PLoS Comput Biol **3**(7): e131.
Shen-Orr, S. S., R. Milo, et al. (2002). "Network motifs in the transcriptional regulation network of Escherichia coli." Nat Genet **31**(1): 64-68.
Shomron, N. and C. Levy (2009). "MicroRNA-biogenesis and Pre-mRNA splicing crosstalk." J Biomed Biotechnol **2009**: 594678.
Simons, R. W. and N. Kleckner (1983). "Translational control of IS10 transposition." Cell **34**(2): 683-691.
Singh, A. (2011). "Negative feedback through mRNA provides the best control of gene-expression noise." IEEE Trans Nanobioscience **10**(3): 194-200.
Sinha, A. K., M. Jaggi, et al. (2011). "Mitogen-activated protein kinase signaling in plants under abiotic stress." Plant Signal Behav **6**(2): 196-203.
Skaug, B. and Z. J. Chen (2010). "Emerging role of ISG15 in antiviral immunity." Cell **143**(2): 187-190.
Smith, M. M. (1991). "Histone structure and function." Curr Opin Cell Biol **3**(3): 429-437.
Sodora, D. L., J. S. Allan, et al. (2009). "Toward an AIDS vaccine: lessons from natural simian immunodeficiency virus infections of African nonhuman primate hosts." Nat Med **15**(8): 861-865.
Sonenberg, N. and A. C. Gingras (1998). "The mRNA 5' cap-binding protein eIF4E and control of cell growth." Curr Opin Cell Biol **10**(2): 268-275.
Souquiere, S., R. Onanga, et al. (2009). "Simian immunodeficiency virus types 1 and 2 (SIV mnd 1 and 2) have different pathogenic potentials in rhesus macaques upon experimental cross-species transmission." J Gen Virol **90**(Pt 2): 488-499.
Southern, E. M. (1975). "Detection of specific sequences among DNA fragments separated by gel electrophoresis." J Mol Biol **98**(3): 503-517.
Stark, C., B. J. Breitkreutz, et al. (2006). "BioGRID: a general repository for interaction datasets." Nucleic Acids Res **34**(Database issue): D535-539.
Stark, G. R. and J. E. Darnell, Jr. (2012). "The JAK-STAT pathway at twenty." Immunity **36**(4): 503-514.
Stevens, A. (1960). Incorporation of the adenine ribonucleotide into RNA by cell fractions from E. coli. Biochem Biophys Res Commun.
Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.
Strahl, B. D. and C. D. Allis (2000). "The language of covalent histone modifications." Nature **403**(6765): 41-45.
Sturtevant, A. H., C. B. Bridges, et al. (1919). "The Spatial Relations of Genes." Proc Natl Acad Sci U S A **5**(5): 168-173.
Su, A. I., T. Wiltshire, et al. (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes." Proc Natl Acad Sci U S A **101**(16): 6062-6067.
Szczyrba, J., E. Nolte, et al. (2013). "Identification of ZNF217, hnRNP-K, VEGF-A and IPO7 as targets for microRNAs that are downregulated in prostate carcinoma." Int J Cancer **132**(4): 775-784.

Taft, R. J., E. A. Glazov, et al. (2009). "Tiny RNAs associated with transcription start sites in animals." Nat Genet **41**(5): 572-578.
Taft, R. J., M. Pheasant, et al. (2007). "The relationship between non-protein-coding DNA and eukaryotic complexity." Bioessays **29**(3): 288-299.
Tajima, F. (1989). "The effect of change in population size on DNA polymorphism." Genetics **123**(3): 597-601.
Takahashi, H., S. Kato, et al. (2012). "CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks." Methods Mol Biol **786**: 181-200.
Takeuchi, T. and H. Yokosawa (2008). "Detection and analysis of protein ISGylation." Methods Mol Biol **446**: 139-149.
Tokumaru, S., M. Suzuki, et al. (2008). "let-7 regulates Dicer expression and constitutes a negative feedback loop." Carcinogenesis **29**(11): 2073-2077.
Toledo-Arana, A., O. Dussurget, et al. (2009). "The Listeria transcriptional landscape from saprophytism to virulence." Nature **459**(7249): 950-956.
Tolhuis, B., R. J. Palstra, et al. (2002). "Looping and interaction between hypersensitive sites in the active beta-globin locus." Mol Cell **10**(6): 1453-1465.
Toyota, M., H. Suzuki, et al. (2008). "Epigenetic silencing of microRNA-34b/c and B-cell translocation gene 4 is associated with CpG island methylation in colorectal cancer." Cancer Res **68**(11): 4123-4132.
Trinklein, N. D., S. F. Aldred, et al. (2004). "An abundance of bidirectional promoters in the human genome." Genome Res **14**(1): 62-66.
Tu, K., H. Yu, et al. (2009). "Combinatorial network of primary and secondary microRNA-driven regulatory mechanisms." Nucleic Acids Res **37**(18): 5969-5980.
Tussiwand, R., W. L. Lee, et al. (2012). "Compensatory dendritic cell development mediated by BATF-IRF interactions." Nature **490**(7421): 502-507.
van Bakel, H., C. Nislow, et al. (2010). "Most "dark matter" transcripts are associated with known genes." PLoS Biol **8**(5): e1000371.
Vaquerizas, J. M., S. K. Kummerfeld, et al. (2009). "A census of human transcription factors: function, expression and evolution." Nat Rev Genet **10**(4): 252-263.
Vargas, D. Y., A. Raj, et al. (2005). "Mechanism of mRNA transport in the nucleus." Proc Natl Acad Sci U S A **102**(47): 17008-17013.
Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.
Wang, G., Y. Wang, et al. (2010). "RNA polymerase II binding patterns reveal genomic regions involved in microRNA gene regulation." PLoS One **5**(11): e13798.
Wang, X., H. Xu, et al. (2013). "Divergent Kinetics of Proliferating T Cell Subsets in Simian Immunodeficiency Virus (SIV) Infection: SIV Eliminates the "First Responder" CD4+ T Cells in Primary Infection." J Virol **87**(12): 7032-7038.
Wang, X., Z. Xuan, et al. (2009). "High-resolution human core-promoter prediction with CoreBoost_HM." Genome Res **19**(2): 266-275.
Wang, Z., C. Zang, et al. (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." Nat Genet **40**(7): 897-903.
Warnefors, M. and A. Eyre-Walker (2011). "The accumulation of gene regulation through time." Genome Biol Evol **3**: 667-673.
Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." Nature **171**(4356): 737-738.
Wei, W., V. Pelechano, et al. (2011). "Functional consequences of bidirectional promoters." Trends Genet **27**(7): 267-276.
Whitehouse, I., A. Flaus, et al. (1999). "Nucleosome mobilization catalysed by the yeast SWI/SNF complex." Nature **400**(6746): 784-787.
Williams, R., H. Yao, et al. (2009). "HIV-1 Tat co-operates with IFN-gamma and TNF-alpha to increase CXCL10 in human astrocytes." PLoS One **4**(5): e5709.

Winkler, G. S. (2010). "The mammalian anti-proliferative BTG/Tob protein family." J Cell Physiol **222**(1): 66-72.
Wong, J. T. (1976). "The evolution of a universal genetic code." Proc Natl Acad Sci U S A **73**(7): 2336-2340.
Woodcock, C. L. and R. P. Ghosh (2010). "Chromatin higher-order structure and dynamics." Cold Spring Harb Perspect Biol **2**(5): a000596.
Woods, K., J. M. Thomson, et al. (2007). "Direct regulation of an oncogenic micro-RNA cluster by E2F transcription factors." J Biol Chem **282**(4): 2130-2134.
Wu, C., C. Orozco, et al. (2009). "BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources." Genome Biol **10**(11): R130.
Xi, H., Y. Yu, et al. (2007). "Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1." Genome Res **17**(6): 798-806.
Xia, S., J. Yang, et al. (2005). "Identification of new targets of Drosophila pre-mRNA adenosine deaminase." Physiol Genomics **20**(2): 195-202.
Xu, J., C. X. Li, et al. (2011). "MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features." Nucleic Acids Res **39**(3): 825-836.
Xu, M., X. Chen, et al. (2013). "Synergistic silencing by promoter methylation and reduced AP-2alpha transactivation of the proapoptotic HRK gene confers apoptosis resistance and enhanced tumor growth." Am J Pathol **182**(1): 84-95.
Yan, J., M. Enge, et al. (2013). "Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites." Cell **154**(4): 801-813.
Yang, J. S., M. D. Phillips, et al. (2011). "Widespread regulatory activity of vertebrate microRNA* species." RNA **17**(2): 312-326.
Ye, T., A. R. Krebs, et al. (2011). "seqMINER: an integrated ChIP-seq data interpretation platform." Nucleic Acids Res **39**(6): e35.
Yi, R., Y. Qin, et al. (2003). "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs." Genes Dev **17**(24): 3011-3016.
Yip, L., L. Kelly, et al. (2011). "MicroRNA signature distinguishes the degree of aggressiveness of papillary thyroid carcinoma." Ann Surg Oncol **18**(7): 2035-2041.
Yosef, N., A. K. Shalek, et al. (2013). "Dynamic regulatory network controlling T17 cell differentiation." Nature.
Yu, X., J. Lin, et al. (2008). "Analysis of regulatory network topology reveals functionally distinct classes of microRNAs." Nucleic Acids Res **36**(20): 6494-6503.
Zaychikov, E., L. Denissova, et al. (1995). "Translocation of the Escherichia coli transcription complex observed in the registers 11 to 20: "jumping" of RNA polymerase and asymmetric expansion and contraction of the "transcription bubble"." Proc Natl Acad Sci U S A **92**(5): 1739-1743.
Zhang, P., Y. Ma, et al. (2012). "Comprehensive gene and microRNA expression profiling reveals the crucial role of hsa-let-7i and its target genes in colorectal cancer metastasis." Mol Biol Rep **39**(2): 1471-1478.
Zhang, Y., T. Liu, et al. (2008). "Model-based analysis of ChIP-Seq (MACS)." Genome Biol **9**(9): R137.
Zhao, Z., G. Tavoosidana, et al. (2006). "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions." Nat Genet **38**(11): 1341-1347.
Zhen, A., T. Wang, et al. (2010). "A single amino acid difference in human APOBEC3H variants determines HIV-1 Vif sensitivity." J Virol **84**(4): 1902-1911.
Zhou, F., Q. Ma, et al. (2012). "QServer: a biclustering server for prediction and assessment of co-expressed gene clusters." PLoS One **7**(3): e32660.
Zhou, X., J. Ruan, et al. (2007). "Characterization and identification of microRNA core promoters in four model species." PLoS Comput Biol **3**(3): e37.
Zhu, Y., M. Roshal, et al. (2003). "Upregulation of survivin by HIV-1 Vpr." Apoptosis **8**(1): 71-79.