ABSTRACT

Title of Document: MAPPING AND CHARACTERIZATION OF FUNCTIONAL INNOVATIONS IN CIS-ACTING ELEMENTS AND TRANS-ACTING FACTORS

Shrutii Sarda, Doctor of Philosophy, 2017

Directed By: Dr. Sridhar Hannenhalli, Professor
Dept. of Cell Biology and Molecular Genetics

The primary mediators of transcriptional regulation are the cis-regulatory elements (CREs), viz., promoters and enhancers, and the trans-acting factors (TFs) that bind to the CREs. First, the landscape of distinct sequence elements that regulate the spatio-temporal activity profiles of genes is far from complete. For example, several (or alternate) CREs can in a context-specific fashion regulate transcription of one gene. Second, mutations that occur in the coding sequences of TFs, or those occurring in CREs that determine TF binding sites, may change the identity of the cognate TF or alter the affinity with which a site is bound, respectively. This in turn introduces a change in the logic of the transcriptional regulatory circuits harboring these modifications and leads to adaptations in the form of novel gene expression patterns, or robust responses to internal or external signals. CREs and trans-acting factors thus provide an extensive platform for regulatory innovation; the extent of which is only beginning

to be appreciated. In this thesis, we discuss three yet-unexplored avenues of regulatory innovation and provide novel insights into each program.

Cis-regulatory rewiring mediated by CREs: A co-regulated module of genes ("regulon") can have evolutionarily conserved expression and yet have diverged upstream regulators across species, such as the ribosomal regulon which is regulated by the transcription factor (TF) TBF1 in *C. albicans*, instead of RAP1 in *S. cerevisiae*. Only a handful of such rewiring events have been established, and the prevalence or conditions conducive to such events are not well known. Here, we develop a novel probabilistic scoring method to comprehensively screen for rewiring within regulons across 23 yeast species. Our analysis recapitulates known events, and suggests TF candidates for certain processes reported to be under distinct regulatory controls in *S. cerevisiae* and *C. albicans*, for which the implied regulators are not known. Independent functional analyses of rewiring TF pairs revealed greater functional interactions, common upstream regulators and shared biological processes between them. Our study reveals that cis-rewiring is pervasive; and generated a high-confidence resource of specific events.

Interaction-mediated regulatory rewiring in TFs: Similar to evolutionary changes in the sequence of CREs, changes within coding regions of TFs can allow for altered protein-protein interaction capabilities and function, through motif and domain turnover across evolution. For example, FTZ, has switched from a homeotic TF in ancestral insect species, to being involved in segmentation in the *Drosophila* genus by the loss of a YPWM motif, and the gain of a LXXLL motif. Elucidating the occurrence of, and mechanisms underlying these switches in TF

function is critical to our understanding of evolution. To this end, we developed an approach to detect protein interaction regulatory rewiring across 1200 TFs in 12 related arthropod species. Simulation studies show that the accuracy of event detection is approximately ~80-85%. We recapitulate the known FTZ rewiring event; and find several members of “enhancer of split” complex represented amongst top events, consistent with previous knowledge that the latter has undergone lineage specific losses and duplications across arthropod evolution. Overall, this work establishes that interaction-rewiring is quite prevalent in arthropod development, and provides a high-confidence list of such candidates.

Orphan CGI alternative promoter potential: CGIs are regions with a relatively high frequency of CpG sites. CGIs that occur within gene promoters are historically well studied. Yet, about 50% of all CGIs lie outside of promoter regions (called orphan CGIs), and not much is understood about their biological significance. We show through extensive analysis of the methylome and transcriptome in 34 tissues, that in many cases of highly expressed genes with methylated-promoters, transcription is initiated by a distal orphan CGI located several tens of kb away that functions as an alternative promoter. We found strong evidence of transcription initiation at the upstream CGI and a lack thereof at the methylated proximal promoter itself. CGI-initiated transcripts are associated with signals of stable elongation and splicing that extend into the gene body, as evidenced by tissue-specific RNA-seq and other DNA-encoded splice signals. Overall, our study describes an unreported mechanism of transcription of methylated proximal promoter genes in a tissue-specific fashion.

MAPPING AND CHARACTERIZATION OF FUNCTIONAL INNOVATIONS IN CIS-ACTING ELEMENTS AND TRANS-ACTING FACTORS

By

Shrutii Sarda

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Sridhar Hannenhalli, Chair
Associate Professor Hector Corrada Bravo
Associate Professor Michelle Girvan
Professor Carlos Machado
Professor Leslie Pick

DEDICATION

To my mother and grandmother, without whose infinite love,

I wouldn’t be who I am today.

AKNOWLEDGEMENTS

I am extremely grateful to my advisor, Dr. Sridhar Hannenhalli, for the opportunity to join his lab and provide me with an environment of support, encouragement and intellectual freedom to pursue questions of a diverse nature. Throughout this journey, I have constantly been inspired by his dedication, as well as his approach to science. I thank my committee members, Drs. Leslie Pick, Hector Corrada Bravo, Carlos Machado and Michelle Girvan, who have been generous with their time and provided great insights on my research.

I would also like to acknowledge other members of the Hannenhalli Lab, especially Avinash Das, Mahfuza Sharmin and Kun Wang, for the many scientifically stimulating discussions we've had over the years.

Finally, I feel very lucky to have incredible parents, and an unconditionally loving grandmother, who have sacrificed endlessly to afford me every opportunity that came along in my lifetime. I am also blessed to have the most understanding and supportive best friend, Rohit Mathews, who was with me, cheering me on every step of the way. These people have always put my interests first, and I cannot thank them enough for everything that they have done.

PREFACE

Portions of the material presented in this dissertation have either been published at or are being prepared for submission to peer-reviewed journals. Please see below for the list of papers that constitute my dissertation, as well as others that I have contributed to during my time at University of Maryland.

Chapter 2

Sarda S, Hannenhalli S. 2015. High throughput identification of cis-regulatory rewiring events in yeast. *Mol Biol Evol* **32**: 3047–3063.

Chapter 3

Sarda S, Das A, Vinson C, Hannenhalli S. 2017. Distal CpG islands can serve as alternative promoters to transcribe genes with silenced proximal promoters. *Genome Res* **27**: 553–566.

Chapter 4

Sarda S, Ben-Hur A, Pick L, Hannenhalli S. High-throughput detection and analysis of protein interaction-based regulatory rewiring events (in preparation).

Other publications

Sarda S, Hannenhalli S. 2014. Next-generation sequencing and epigenomics research: a hammer in search of nails. *Genomics Inform* **12**: 2–11.

Sarda S, Hannenhalli S. Co-option of CGIs as alternative promoters. Invited peer-reviewed perspective article by the journal Transcription (in preparation).

TABLE OF CONTENTS

ABSTRACT i

DEDICATION ii

AKNOWLEDGEMENTS iii

PREFACE iv

TABLE OF CONTENTS v

LIST OF FIGURES viii

LIST OF TABLES ix

CHAPTER 1: Introduction 1

Regulation of Genome Activity 1

Transcriptional Regulation and its Complexities 3

cis-Acting Regulatory Sequences: Promoters and Enhancers 4

trans-Acting Regulatory Factors: Transcription Factors (TFs) 6

Chromatin Structure & DNA Accessibility 8

Histone Modifications and Nucleosome Remodeling 10

DNA Methylation 12

Regulatory Innovations in the Genome 14

Cis-regulatory Rewiring mediated by CREs 15

Interaction-mediated Regulatory Rewiring in trans-acting Factors 17

Orphan CGI alternative promoter potential 20

Organization of Thesis 22

CHAPTER 2: High throughput identification of cis-regulatory rewiring events in yeast 24

Abstract 24

Introduction 25

Results 30

A probabilistic framework to detect rewiring events 30

High-throughput computation of rewiring scores across all orthogroups, TFs and lineages 31

Regulon-level rewiring of transcription factor usage 34
Our method recapitulates previously established rewiring events in yeast 35
Identifying candidate TFs for known rewiring events..... 41
Rewiring events are strongly supported by co-expression between the regulator and targets 44
Functional connections between rewiring TFs and their properties..... 48
Gene-level assessment of rewiring using rotation test 54
Discussion 58
Materials and Methods..... 63
Gene orthology groups, annotations and sequences..... 63
Probabilistic rewiring score..... 63
PWM based TF binding probability..... 65
Species tree and selected lineages to assess rewiring..... 65
Expression data..... 66
Regulon discovery..... 66
Generation of species-specific TF-TF networks 68
Phylogeny-preserving rotation test to assess significance of rewiring score at gene-level .. 68
Nucleosome occupancy data 69
CHAPTER 3: Distal CpG islands can serve as alternative promoters to transcribe genes with silenced proximal-promoters 70
Abstract..... 70
Introduction 71
Results..... 74
Highly expressed genes with methylated promoters..... 74
Association of distal CGI with methylated-promoter gene expression..... 77
Transcription initiation occurs at distal CGI, and not the promoter of MethExp genes..... 81
Evidence of transcriptional elongation and splicing occurring between distal CGIs and their associated MethExp gene promoters. 88
Aberrant gene expression in cancer linked to hypomethylated distal CGIs. 94
Discussion 106
Materials and Methods..... 111
Datasets 111
Primary processing of genes and pooling into gene groups..... 114
Evidence of gene body alternative promoter usage..... 116

Tissue specificity index (TSI) 116
Evolutionary conservation 117
Cell-type specific regulation of alternative promoter CGIs 117
Sequence-based splicing signals 118
Gene Ontology (GO) enrichment 118

CHAPTER 4: High-throughput detection and analysis of protein interaction-based regulatory rewiring events 119
Abstract 119
Introduction 120
Results 124
Quantifying evolutionary interaction rewiring 124
Estimating interaction probabilities between a pair of proteins 125
Simulation-based power analysis of rewiring event detection 130
Recapitulation of the FTZ rewiring event 132
Functional insights about interaction based rewiring in arthropod species 133
Discussion 135
Materials & Methods 137
TF annotations, sequences and orthology groups 137
Domain and linear motif detection 138
Domain-Domain (DDI) and Domain-Linear Motif Interaction (DLI) detection 138
PPI data source, treatment and negative set generation 138
Alternative PPI prediction methods 139
Probabilistic rewiring score 141
Identification of developmental TFs 141

CHAPTER 5: Perspective and future work 143

APPENDICES 149
Appendix A: Supplemental Material from Chapter 2 149
Appendix B: Supplemental Material from Chapter 3 158
Appendix C: Supplemental Material from Chapter 4 205

BIBLIOGRAPHY 206

LIST OF FIGURES

Figure 1-1. An overview of yeast Ribosomal Protein (RP) promoters. …… 16
Figure 1-2. Diversity in Ftz cofactor interaction motifs across arthropod evolution. …… 19
Figure 2-1. Overview of the cis-rewiring detection approach. 30
Figure 2-2. Distribution of all computed rewiring scores. 33
Figure 2-3. Rewiring scores of the ribosomal regulon for RAP1-TBF1 and IFH1-CBF1 switches across branches. 40
Figure 2-4. Rewiring scores of other known rewiring events across branches. 41
Figure 2-5. Rewiring scores of predicted rewiring events across branches. 44
Figure 2-6. Co-expression analyses for regulon rewiring events. 48
Figure 2-7. Functional analyses of rewired TFs in regulons. 53
Figure 2-8. Rewiring at the individual gene level. …… 56
Figure 2-9. Functional analyses of rewired TFs in individual genes. …… 58
Figure 3-1. Association of distal CGI with the expression of MethExp genes. …… 80
Figure 3-2. Transcription initiation occurs at an upstream alternative CGI promoter, and not at the proximal promoters of MethExp genes. …… 85
Figure 3-3. Transcriptional elongation and splicing signals between CGI and the gene. 93
Figure 3-4. An illustrative example. 94
Figure 3-5. Use of distal CGI as alternative promoter by MethExp genes in cancer. …… 97
Figure 3-6. Functional enrichment of MethExp genes in cancer that potentially utilize an upstream CGI as promoter. …… 105
Figure 4-1. Illustration of approach to assess interaction rewiring. …… 124
Figure 4-2. Phylogenetic tree of select arthropod species. …… 125
Figure 4-3. Off-the-shelf PPI prediction method evaluation. 128
Figure 4-4. Fly to human predicted PPI probability distribution and resulting rewiring event detection accuracy. ……132
Figure 4-5. Distribution of rewiring scores for all triplets with the TF FTZ. ……133

LIST OF TABLES

Table 1-1. Examples of steps in the genome expression pathway at which regulation can be exerted. 2
Table 3-1. Metadata associated with the 34 tissue types used in our analyses. …….. 75
Table 3-2. The number of genes before and after applying a filtration step. …… 76
Table 3-3. List of hypermethylated-promoter genes in breast cancer that use distal CGIs as alternative promoters. …… 99
Table 3-4. A GO functional annotation of breast cancer MethExp genes in breast cancer differentially expressed from their normal counterparts. ….... 104
Table 3-5. Fraction of tested MethExp loci per cell type that show strong evidence of alternative promoter usage based on stringent thresholds. ……..…… 108

CHAPTER 1: Introduction

Regulation of Genome Activity

The central dogma of molecular biology suggests that it is the “realization” of the genome that specifies the content of the proteome, which in turn determines the biochemical signature of a cell. Therefore, early on, efforts were made to obtain quantitative measures of “genetic distances”, as these were considered to hold clues to the vast diversity observed at the organismal level – that is, at the level of anatomy, physiology, and behavior. Although, in the 1970s, King and Wilson (King and Wilson 1975) strongly suggested that even the genetic distance between humans and the chimpanzee was too small to account for their substantial organismal differences. In order to explain how species that have such similar genes (and proteins) can differ so substantially, one must look for differences at the regulatory level, i.e., mechanisms that control the spatio-temporal expression of genes rather than the sequence changes within genes themselves. The importance of the role of regulation in the genome becomes even more apparent when differences within an individual organism are considered (Marbach et al. 2016). Given that all cells in an individual share identical DNA (not withstanding somatic mutations, and some immune cell subtypes), and it is the same genetic information that is translated into morphologically and phenotypically distinct cells that comprise that individual, it becomes clear that the same genome can be “realized” in distinct ways. Thus, studying the complexities of regulation of genome activity becomes crucial.

There are many factors that influence genome expression at various stages of the process, including transcription, mRNA processing as well as protein synthesis and processing (Brown 2002). The following table lists many of the individual steps involved in genome expression and the types of regulatory controls exerted at each stage of the process.

Step	Example of regulation	Cross-reference
Transcription		
Gene accessibility	Locus control regions determine chromatin structure in areas that contain genes	Ch. 8, Sec. 1 (Brown 2002)
	Histone modifications influence chromatin structure and determine which genes are accessible	Ch. 8, Sec. 2 (Brown 2002)
	Nucleosome positioning controls access of RNA polymerase and transcription factors to the promoter region	Ch. 8, Sec. 2 (Brown 2002)
	DNA methylation silences regions of the genome	Ch. 8, Sec. 2 (Brown 2002)
Initiation of transcription	Productive initiation is influenced by activators, repressors and other control systems	Ch. 9, Sec. 3 (Brown 2002)
Synthesis of RNA	Prokaryotes use antitermination and attenuation to control the amount and nature of individual transcripts	Ch. 10, Sec. 1 (Brown 2002)
Eukaryotic mRNA processing		
Capping	Some animals use capping as a means of regulating protein synthesis during egg maturation	-
Polyadenylation	Translation of *bicoid* mRNA in *Drosophila* eggs is activated after fertilization by extension of the poly(A) tail	Ch. 12, Sec. 3 (Brown 2002)
Splicing	Alternative splice site selection controls sex determination in *Drosophila*	Ch. 10, Sec. 1 (Brown 2002)
Chemical modification	RNA editing of apolipoprotein-B mRNA results in liver- and intestine-specific versions of this protein	Ch. 10, Sec. 3 (Brown 2002)
mRNA degradation	Iron controls degradation of transferrin receptor mRNA	Ch. 11, Sec. 2 (Brown 2002)
Protein synthesis and processing		
Initiation of translation	Phosphorylation of eIF-2 results in a general reduction in translation initiation in eukaryotes	Ch. 11, Sec. 2 (Brown 2002)
	Ribosomal proteins in bacteria control their own synthesis by modulating ribosome attachment to their mRNAs	Ch. 11, Sec. 2 (Brown 2002)
	In some eukaryotes, iron controls ribosome scanning on ferritin mRNAs	Ch. 11, Sec. 2 (Brown 2002)
Protein synthesis	Frameshifting enables two DNA polymerase III subunits to be translated from the *Escherichia coli dnaX* gene	Ch. 11, Sec. 2 (Brown 2002)
Cutting events	Alternative cleavage pathways for polyproteins result in tissue-specific protein products	Ch. 11, Sec. 3 (Brown 2002)
Chemical modification	Many proteins involved in signal transduction are activated by phosphorylation	Ch. 12, Sec. 1 (Brown 2002)

Table 1-1. Examples of steps in the genome expression pathway at which regulation can be exerted. Table adapted from “Genomes” by T.A. Brown (Brown 2002).

It is no surprise that we can nominate examples of regulation for every point in the genome expression pathway. But are all these control points of equal importance in regulating the activity of the genome as a whole? Our current perception is that they are not. Our understanding may be imperfect, but it appears that the critical controls over genome expression - the decisions about which genes are switched on and which are switched off - are exerted, to a large extent, at the level of transcription (Jacob and Monod 1961). For most genes, control that is exerted at later steps, i.e. in mRNA processing, translation, and post-translational modifications (PTM), mainly serves to modulate expression and protein levels, but does not act as the primary determinant of whether the gene is on or off.

Transcriptional Regulation and its Complexities

As in bacteria, transcription initiation and expression in eukaryotic cells is controlled by proteins that bind to specific regulatory sequences and modulate the activity of RNA polymerase. The intricate task of regulating gene expression in the many differentiated cell types of multicellular organisms is accomplished primarily by the combined actions of multiple different transcriptional regulatory proteins. In addition, several epigenomic signatures characterized by the packaging of DNA into distinct chromatin functional domains, and its modification by methylation, histone modifications and nucleosome remodeling impart further levels of complexity to the control of eukaryotic transcription initiation and expression. The landscape of these determinants, as well as their exact role and

significance in transcriptional regulation is described at length below.

cis-Acting Regulatory Sequences: Promoters and Enhancers

A promoter is a genomic region of a particular gene where the transcription is initiated. It is located upstream (at the 5' end) of the gene, and also provides a control point for regulated gene transcription. Promoters represent critical elements that can work in concert with other regulatory regions (enhancers, silencers, boundary elements/insulators) to direct the level of transcription of a given gene (Maston et al. 2006). The promoter contains specific DNA sequences that are recognized and bound by proteins known as transcription factors, recruiting RNA polymerase, the enzyme that synthesizes RNA from the DNA template of the gene. It is generally comprised of the two following elements, (i) the core promoter, and (ii) the proximal promoter.

Genes transcribed by RNA polymerase II most commonly have either one of two, mostly mutually exclusive core promoter elements, the TATA box or the Inr sequence, that serve as specific binding sites for general transcription factors (Smale and Kadonaga 2003; Smale and Baltimore 1989). The core promoter is the minimal portion of the promoter required to properly initiate transcription. Other cis-acting sequences within the proximal promoter serve as binding sites for a wide variety of regulatory factors that control the expression of individual genes (Lynch 2006). These cis-acting regulatory sequences are frequently, though not always, located upstream of the core promoter. For example, two regulatory sequences that are found in many eukaryotic genes were identified by studies of the promoter of the herpes simplex virus gene that encodes thymidine

kinase. Both of these sequences are located within 100 base pairs upstream of the TATA box: Their consensus sequences are CCAAT and GGGCGG (called a GC box) (Jonkers and Lis 2015). Specific proteins that bind to these sequences and stimulate transcription have since been identified.

In contrast to the relatively simple organization of CCAAT and GC boxes in the herpes thymidine kinase promoter, many genes in mammalian cells are controlled by regulatory sequences located farther away (sometimes more than 10 kilobases) from the transcription start site. These sequences, called enhancers, were first identified by Walter Schaffner in 1981 during studies of the promoter of another virus, SV40 (Banerji et al. 1981). In addition to a TATA box and a set of six GC boxes, two 72-base-pair repeats located farther upstream are required for efficient transcription from this promoter. These sequences were found to stimulate transcription from other promoters as well as from that of SV40, and, surprisingly, their activity depended on (i) neither their distance, (ii) nor their orientation with respect to the transcription initiation site. They could stimulate transcription when placed either upstream or downstream of the promoter, in either a forward or backward orientation (Müeller-Storm et al. 1989).

Surprisingly, enhancers, like promoters, are also bound by transcription factors (TFs) that then regulate RNA polymerase. This is possible because of DNA looping, which allows a TF bound to a distant enhancer to interact with RNA polymerase or general transcription factors at the promoter. Thus, TFs bound to distant enhancers work by the same mechanisms as those bound adjacent to promoters (Weingarten-Gabbay and Segal 2014). An important aspect of

enhancers is that they usually contain multiple functional sequence elements that bind different transcriptional regulatory proteins. These elements include some *cis*-acting regulatory sequences that bind transcriptional activators, that activate transcription in particular cell types (Schirm et al. 1987), as well as other regulatory sequences that bind repressors that inhibit transcription in inappropriate cell types. Tissue-specific expression patterns result from the combination of the individual sequence elements that make up the complete enhancer (Bulger and Groudine 2010).

trans-Acting Regulatory Factors: Transcription Factors (TFs)

One of the prototypes of eukaryotic transcription factors was initially identified by Robert Tjian and his colleagues during studies of the transcription of SV40 DNA. This factor (called Sp1, for specificity protein 1) was found to stimulate transcription from the SV40 promoter, but not from several other promoters, in cell-free extracts. Then, stimulation of transcription by Sp1 was found to depend on the presence of the GC boxes in the SV40 promoter: if these sequences were deleted, stimulation by Sp1 was abolished (Kadonaga et al. 1986). Taken together, these results indicate that the GC box represents a specific binding site for a specific transcriptional factor - Sp1. Similar experiments have established that many other transcriptional regulatory sequences, including the CCAAT sequence (Jones et al. 1987), also present recognition sites for sequence-specific DNA-binding proteins.

Because transcription factors are central to the regulation of gene expression, understanding the mechanisms of their action (Spitz and Furlong 2012) is a

major area of ongoing research in cell and molecular biology. The most thoroughly studied of these proteins are transcriptional activators, which, like Sp1 (Kadonaga et al. 1987), bind to regulatory DNA sequences and stimulate transcription. In general, these factors have been found to consist of two domains: one region of the protein specifically binds DNA; the other activates transcription by interacting with other proteins or components of the transcriptional machinery. Molecular characterization has revealed that the DNA-binding domains of many of these proteins are related to one another (Latchman 1997). Zinc finger domains contain repeats of cysteine and histidine residues that bind zinc ions and fold into looped structures (“fingers”) that bind DNA. Other families of DNA-binding proteins include leucine zipper, helix-turn-helix, helix-loop-helix proteins, etc. (Frankel and Kim 1991). The activation domains of transcription factors are not as well characterized as their DNA-binding domains. Some, called acidic activation domains, are rich in negatively charged residues (aspartate and glutamate); others are rich in proline or glutamine residues. These activation domains are thought to stimulate transcription by interacting with general transcription factors (Ptashne and Gann 1997), such as TFIIB or TFIID, thereby facilitating the assembly of a transcription complex on the promoter.

Gene expression in eukaryotic cells is regulated by repressors as well as by activators. Like their prokaryotic counterparts, eukaryotic repressors bind to specific DNA sequences and inhibit transcription. In some cases, eukaryotic repressors simply interfere with the binding of other transcription factors to DNA. For example, the binding of a repressor near the transcription start site can block

the interaction of RNA polymerase or general transcription factors with the promoter, which is similar to the action of repressors in bacteria (Gaston and Jayaraman 2003). Other repressors compete with activators for binding to specific regulatory sequences. Some such repressors contain the same DNA-binding domain as the activator but lack its activation domain. As a result, their binding to a promoter or enhancer blocks the binding of the activator, thereby inhibiting transcription. In contrast to repressors that simply interfere with activator binding, many repressors (called active repressors) contain specific functional domains that inhibit transcription via protein-protein interactions (Hanna-Rose and Hansen 1996). The functional targets of repressors are also diverse. Some repressors inhibit transcription by interacting with general transcription factors, such as TFIID; others are thought to interact with specific activator proteins, irrespective of its site of binding to DNA (Ptashne 2014).

Chromatin Structure & DNA Accessibility

The DNA of all eukaryotic cells is tightly packaged into chromatin. The basic structural unit of chromatin is the nucleosome, which consists of 146 base pairs of DNA wrapped around two molecules each of histones H2A, H2B, H3, and H4, with one molecule of histone H1 bound to the DNA as it enters the nucleosome core particle. The chromatin is then further condensed by being coiled into higher-order structures, called 30 nm chromatin fibers (Bartova et al. 2008) organized into large loops of DNA.

Chromatin structure is hierarchic, ranging from the two lowest levels of DNA packaging – the nucleosome and the 30nm chromatin fiber – to the metaphase

chromosomes, which represent the most compact form of chromatin in eukaryotes and occur only during nuclear division (Woodcock and Ghosh 2010). After division, the chromosomes become less compact and cannot be distinguished as individual structures. When non-dividing nuclei are examined by light microscopy all that can be seen is a mixture of lightly and darkly staining areas within the nucleus. The dark areas, which are concentrated around the periphery of the nucleus, are called heterochromatin and contain DNA that in a relatively compact organization (Hendzel et al. 1997), although still less compact than in the metaphase structure. Two types of heterochromatin are recognized:

- Constitutive heterochromatin is a permanent feature of all cells and represents DNA that contains no genes and so can always be retained in a compact organization. This fraction includes centromeric and telomeric DNA as well as certain regions of some other chromosomes (Haaf and Schmid 1991). For example, most of the human Y chromosome is made of constitutive heterochromatin.
- Facultative heterochromatin is not a permanent feature but is seen in some cells some of the time. Facultative heterochromatin is thought to contain genes that are inactive in some cells or at some periods of the cell cycle (Trojer and Reinberg 2007). When these genes are inactive, their DNA regions are compacted into heterochromatin.

The organization of heterochromatin is so compact that proteins involved in gene expression simply cannot access the DNA. In contrast, the remaining regions of chromosomal DNA, the parts that contain active genes, are less compact and

permit entry of the expression proteins (Moazed 2001). These regions are called euchromatin and they are dispersed throughout the nucleus.

This form of packaging of eukaryotic DNA in chromatin has diverse functions well beyond mere compaction, including having important consequences in terms of its availability as a template for transcription which makes chromatin structure a critical aspect of gene expression in eukaryotic cells. First, open and easily accessible regions of DNA within the chromatin are indicative of local territories of transcriptional activity (Dekker et al. 2013). Actively transcribed genes are found in decondensed chromatin that is more accessible to transcription factors than is the rest of the genome. Second, coordinated activity of distal elements is orchestrated by short- and long-range DNA interactions, which is determined by the 3D chromatin structure. For instance, chromatin conformation/looping mediates a promoter's access to its enhancers, thereby determining the transcriptional fate of a gene (Harmston and Lenhard 2013).

Histone Modifications and Nucleosome Remodeling

Decondensation of chromatin, however, is not sufficient to make the DNA an accessible template for transcription. Even in decondensed chromatin, actively transcribed genes remain bound to histones and packaged in nucleosomes, so transcription factors and RNA polymerase are still faced with the problem of interacting with chromatin rather than with naked DNA (Mirny 2010). The tight winding of DNA around the nucleosome core particle is a major obstacle to transcription, affecting both the ability of transcription factors to bind DNA and the ability of RNA polymerase to transcribe through a chromatin template.

Chromosomal DNA is wrapped around histone octamers, essentially composed of 4 kinds of subunits; viz., H2A, H2B, H3, and H4. These proteins are subject to chemical modifications at specific residues of histone tails; some well-studied modifications include phosphorylation, methylation, and ubiquitination (Peterson and Laniel 2004). These mostly reversible modifications are involved in setting the stage for directed transcriptional activation and repression by controlling DNA accessibility or recruitment (Bannister and Kouzarides 2005) of other protein complexes. For example, the inhibitory effect of nucleosomes is relieved by acetylation of histones which reduces their net positive charge and weaken their binding to DNA. This idea has been extended into the “histone code” hypothesis (Strahl and Allis 2000; Jenuwein and Allis 2001) that complex combinations of distinct histone modifications, like H3K27me3 (a mark of repressed regions), H3K4me3 (a mark of gene promoters), H3K27ac (a mark of transcriptionally active regions), etc., underlie specific transcriptional programs (Bannister and Kouzarides 2011). This notion has been further extended in more recent works into an ‘epigenomic code’ to include epigenomic marks other than histone modifications (Ernst and Kellis 2012).

Additional proteins called nucleosome remodeling factors facilitate the binding of transcription factors to chromatin by altering nucleosome structure. The mechanism of action of nucleosome remodeling factors is not yet clear, but they appear to increase the accessibility of nucleosomal DNA to other proteins (such as transcription factors) without removing the histones (Alkhatib and Landry 2011). One possibility is that they catalyze the sliding of histone octamers along

the DNA molecule, thereby repositioning nucleosomes to facilitate transcription factor binding. The mechanisms by which nucleosome remodeling factors are targeted to actively transcribed genes also remain to be established, although some studies suggest that they can be brought to enhancer or promoter sites in association with transcriptional activators or as components of the RNA polymerase II holoenzyme (Kireeva et al. 2002).

DNA Methylation

One of the more stable and heritable, and the most studied, epigenetic marks is DNA methylation, which provides another general mechanism by which control of transcription in vertebrates is linked to chromatin structure. Cytosine residues in vertebrate DNA can be modified by the addition of methyl groups at the 5-carbon position. DNA methylation is found in three different sequence contexts: CG (or CpG), CHG or CHH (where H correspond to A, T or C), although in mammals most methylation occurs at CG dinucleotides (Bird 1992). Two types of methylation activity have been distinguished. The first is maintenance methylation which, following genome replication, is responsible for adding methyl groups to the newly synthesized strand of DNA at positions opposite methylated sites on the parent strand. The maintenance activity therefore ensures that the two daughter DNA molecules retain the methylation pattern of the parent molecule (Eckhardt et al. 2006). The second activity is de novo methylation, which adds methyl groups at totally new positions and so changes the pattern of methylation in a localized region of the genome. Through in vitro experiments, it was originally thought that Dnmt1 was responsible for both types of methylation

in mammalian cells. It was subsequently discovered that knockout mice that have an inactivated gene for Dnmt1 can still carry out de novo methylation. This led to the search for new enzymes and the eventual discovery of Dnmt3a and Dnmt3b, which are now considered to be the main de novo methylases of mammals, with Dnmt1 primarily responsible for the maintenance activity (Bird 1984).

DNA methylation is correlated with reduced transcriptional activity of genes that contain high frequencies of CpG dinucleotides in the vicinity of their promoters. Methylation inhibits transcription of these genes via the action of a protein, MeCP2, that specifically binds to methylated DNA and represses transcription. Interestingly, MeCP2 functions as a complex with histone deacetylase, linking DNA methylation to alterations in histone acetylation and nucleosome structure (Robertson 2000). Although DNA methylation is capable of inhibiting transcription, its general significance in gene regulation is unclear. In many cases, methylation of inactive genes is thought to be a consequence, rather than the primary cause, of their lack of transcriptional activity (Baylin and Bestor 2002). The human genome is highly methylated; approximately 80% of cytosines in CpG dinucleotides are chemically modified at their fifth carbon atom with a methyl group (Tucker 2001). Although historically, DNA methylation was associated with transcriptional silencing (as evidenced by many promoter-based studies) (Newell-Price et al. 2000), genome-scale profiling of this epigenetic mark revealed that in some instances DNA methylation is correlated with transcriptional activation, such as when it is enriched in the gene bodies of active genes (Lister et al. 2009). That is, DNA methylation can inhibit transcription

initiation but not elongation. Thus, the associations between an epigenomic mark and functional output may be location-specific, thereby complicating their functional interpretation.

Regulatory Innovations in the Genome

Variability in transcriptional regulation underlies a large portion of phenotypic variability across tissues, as well as across species. The primary mediators of variability in transcriptional regulation are the cis-regulatory elements (CREs), viz., promoters and enhancers, and the trans-acting factors that bind to the CREs. First, the landscape of distinct sequence elements that regulate the spatio-temporal activity profiles of genes is far from complete. For example, several (or alternate) CREs such as promoters and enhancers can act in a coordinated fashion to regulate transcription of one gene. This ensures that genes are only expressed when they are needed allowing the generation of phenotypic variance, organizational maintenance, and energy conservation. Second, mutations that occur in the coding sequences of trans-acting factors, or those occurring in cis-regulatory elements (CREs) that determine TF binding sites, may change the identity of the cognate TF or alter the affinity with which a site is bound, respectively. This in turn introduces a change in the logic of the transcriptional regulatory circuits harboring these modifications and leads to adaptations in the form of novel gene expression patterns. But regulatory changes representative of conserved expression patterns also occur frequently. Since transcriptional output is sometimes the end point of signal transduction

pathways, these changes have probably allowed the two systems to respond differently to internal or external signals.

CREs and trans-acting factors thus provide an extensive platform for regulatory innovation; the extent of which is only beginning to be appreciated. Studies exploring diverse aspects of regulatory innovation across species have compared sequence conservation of their cis-regulatory modules (Maeso et al. 2013), turnover of binding sites in accessible DNA (Stergachis et al. 2014), as well as the spatial configuration of regulatory DNA (Dixon et al. 2012), thereby contributing to our understanding of transcriptional regulation in different ways. We discuss three yet-unexplored avenues of regulatory innovation, that form the bases of this dissertation, below:

Cis-regulatory Rewiring mediated by CREs

As outlined earlier, there is extensive plasticity in the cis and trans-regulatory circuitry that is representative of both conserved and diverged expression programs across species. TF rewiring is a prominent mechanism of evolutionary changes in cis-regulation. Over evolutionary time, some genes (or sets of co-regulated genes) have undergone a switch in cis-regulation whereby they are regulated by one TF in a few species, but have opted this factor out and are instead regulated by a different TF in other related species. A well-known example of such regulatory rewiring in yeast species occurred in a set of functionally related co-expressed genes, viz., the ribosomal regulon. This regulon in *Candida albicans* (Calb) and related yeast species is under the control of the DNA binding factor TBF1, whereas in the more recently evolved *Saccharomyces*

cerevisiae (Scer), the repressor-activator protein RAP1 regulates the transcription of the same regulon (Tanay et al. 2005). This switch of regulatory factors is due to the loss of the cis-element, i.e., transcription factor binding sites for TBF1, and the appearance of the RAP1 transcription factor binding sites in the promoters of 60+ genes that comprise this regulon.

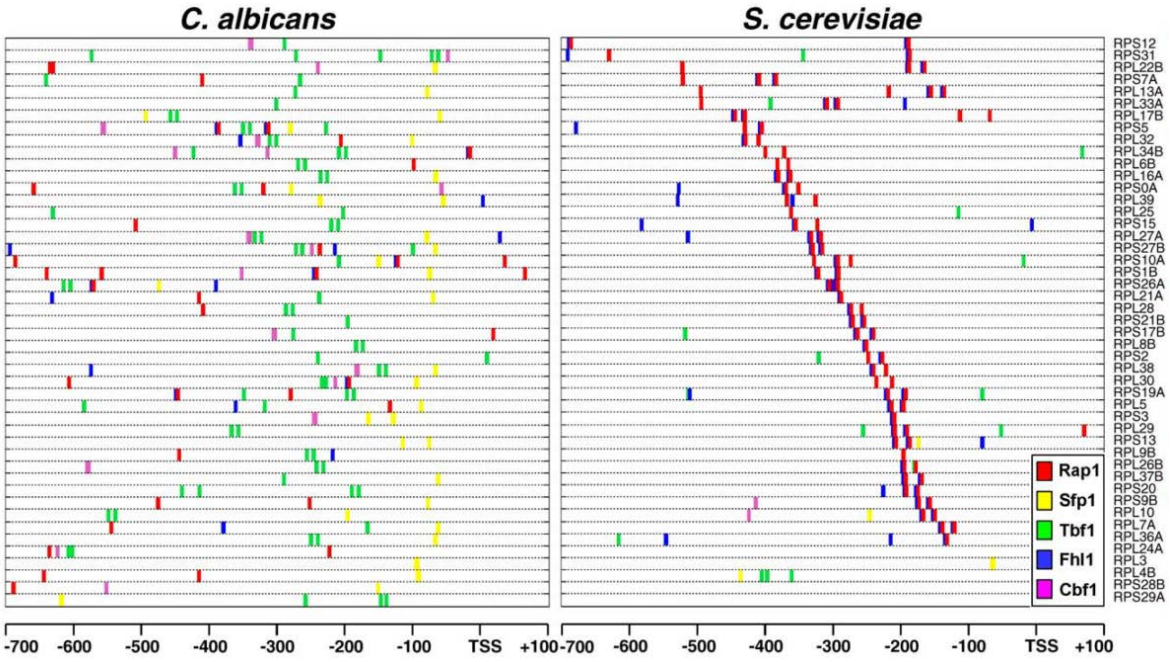


Figure 1-1. An overview of yeast Ribosomal Protein (RP) promoters. Each line displays information for one RP gene (left to right: promoter regions in *C. albicans* and *S. cerevisiae* (from -700 to +100 bp, relative to transcription start site (TSS) and gene name). RPs were chosen based on annotations in *S. cerevisiae*, and restricted to those that have a 1:1 ortholog mapping to *C. albicans*. Colored boxes indicate locations of predicted TF binding sites (see key in bottom right corner). TFs were selected that have at least threefold binding site enrichment in promoter regions of RP genes in either species (relative to randomly selected promoter sets). Genes are sorted (top to bottom) in order of most distal appearance of Rap1 binding sites in *S. cerevisiae* (relative to TSS) among 47 RP genes. Figure adapted from (Weirauch and Hughes 2010)

More recently, a rewiring event involving the regulation of Leloir-pathway genes (enzymes acting to metabolize galactose) was reported in yeast species. In Scer, these genes are positively regulated by Gal4, however this transcription factor is found to regulate genes unrelated to galactose metabolism in Calb. Instead,

activation of expression of Leloir-pathway genes in the latter species requires Cph1, the homolog of the Ste12 transcription factor of *S. cerevisiae* (Martchenko et al. 2007). Prior studies have characterized a few cases of regulatory rewiring of specific genes/gene-sets in great depth (Martchenko et al. 2007; Lavoie et al. 2010; Hogues et al. 2008), and yet these are expected to represent just the tip of the iceberg. These previous studies provide important insights into aspects of rewiring. For example, Mallick and Whiteway (Mallick and Whiteway 2013) showed how regulatory connections local to rewired TFs can change to preserve gene target expression patterns (for example, recruitment of *IFH1-FHL1* to ribosomal gene targets is maintained in both systems). Yet, there are several aspects of regulatory rewiring that are poorly understood. For instance, (i) how widespread is a wholesale shift in transcriptional regulation of a regulon?, (ii) what are the features of target genes that make them amenable to rewiring?, (iii) what characterizes rewired TFs, etc. By gathering more candidate rewiring events and collectively analyzing their trends, we can potentially answer these questions and gain further insights into conditions conducive to rewiring, as well as enable discovery of clade/species-specific instances of regulatory innovation. Yet, no efforts have been made to discover in an unbiased and high-throughput manner, the presence of such regulatory rewiring.

Interaction-mediated Regulatory Rewiring in trans-acting Factors

Similar to evolutionary changes in the sequence of CREs, changes in the coding sequence of genes encoding TF proteins (in terms of their activation or interaction domains) can also lead to regulatory rewiring. Specifically, due to

sequence changes coding for secondary structures like motifs and 3D domains of a protein, a given TF can now interact with different partners, via altered protein-protein interaction capabilities, such that it is might be involved in regulating a different set of genes.

A few instances of such domain related regulatory rewiring have been well established in a wide range of species. For example, the cis-element rewiring between RAP1 and TBF1 in yeast was also accompanied by a change in the protein domains of co-factors that they interact with. Essentially, a dimer containing IFH1 is considered to be the primary regulator of ribosomal protein genes in both Scer and Calb. In Scer, the dimer is recruited to the ribosomal gene promoter by RAP1 to activate expression, whereas in Calb, this dimer is directed to its target by TBF1. Intriguingly, correlated with the transition to the RAP1-regulated circuit in Scer, the Sc-IFH1 now contains a RAP1 interaction domain that is not present in the Calb protein (Mallick and Whiteway 2013).

A similar phenomenon driven by changes in protein linear motifs has evolved during the development of diverse insect species. The TF FTZ, has switched from serving a homeotic role in ancestral insect species, to being involved in segmentation in the *Drosophila* genus. This switch in FTZ's function is accompanied by the loss of YPWM, a protein sequence motif that is responsible for cofactor interactions with homeotic regulators, and the gain of a LXXLL motif that enables interaction with segmentation-related cofactors and targets (Heffer et al. 2010). Specifically, having acquired the LXXLL motif, the FTZ now possesses the capability to dimerize with FTZ-F1 (via the AF-2 domain), a TF

controlling segmentation processes across diverse insect species. These TFs cooperatively bind to their target gene engrailed, to activate its expression and initiate this stage of development in *Drosophila* (Florence et al. 1997).

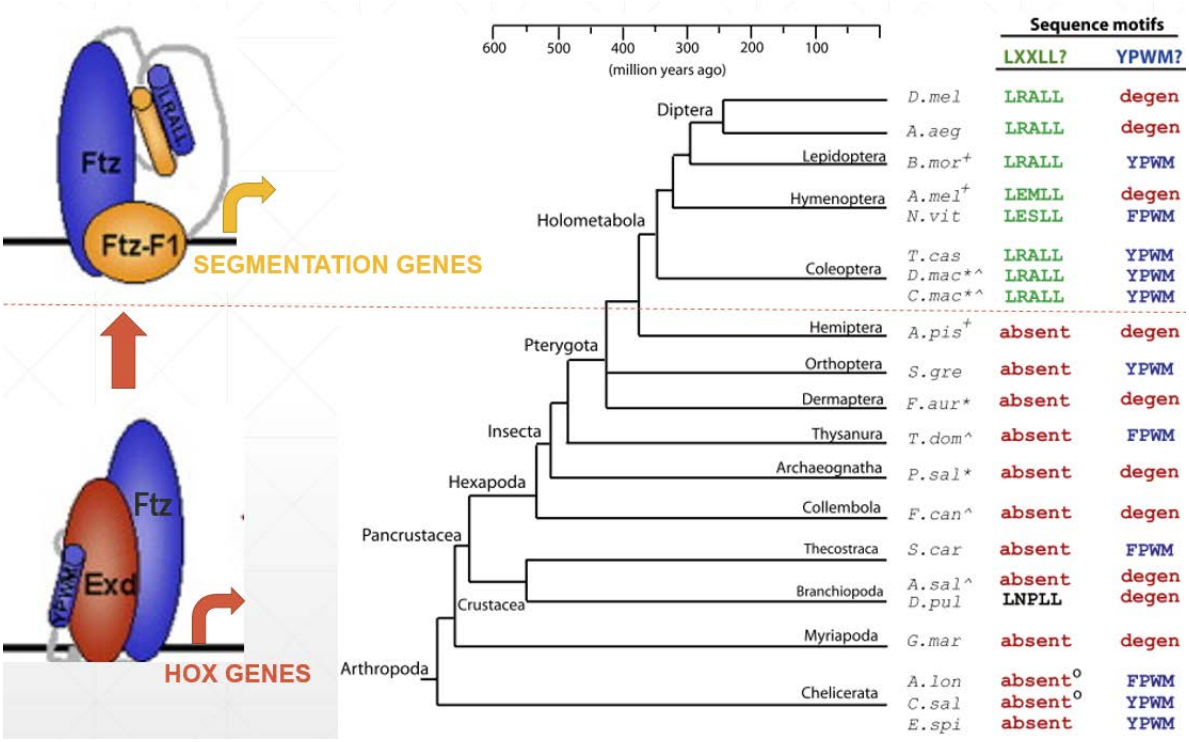


Figure 1-2. Diversity in Ftz cofactor interaction motifs across arthropod evolution. An arthropod phylogeny showing the presence/absence of functional Ftz motifs which was accompanied by a change in function of Ftz from regulating homeotic genes in early arthropods to segmentation genes in holometabolites. The LXXLL motif (green) is required for pair-rule function in *Drosophila* and mediates interaction with the Ftz cofactor Ftz-F1. LXXLL was stably acquired at the base of Endopterygota. The YPWM mediates interaction with the homeotic cofactor Exd. This motif is present in Ftz in some arthropods (blue), but has degenerated in many lineages (red). Figure adapted from (Heffer et al. 2013)

Despite the extent of protein domain changes across clades, very few studies have been reported that map these in a genome-wide fashion, specifically as they pertain to regulatory rewiring. Moore and Bornberg-Bauer (Moore and Bornberg-Bauer 2012) explored the functional implications of protein domain gain, loss and emergence in the proteomes of 20 arthropod species of the pan-

crustacean clade. Although they map evolutionary changes in protein structure to function, this study (1) is mainly focused on contrasting the evolution and function of novel protein domains vs. that of conserved domains across species, (2) does not venture to explain changes in interaction preferences between TFs on account of sequence changes, and (3) is primarily a domain-centric analysis, and does not account for the effect of evolutionary changes in short linear motifs (SLiMs), like the LXXLL/YPWM motifs in insects. SLiMs are interaction modules that have been implicated in greatly diversifying functions of protein isoforms (Weatheritt and Gibson 2012). A recent study suggests that weak domain-linear motif interactions (DLIs) are more likely to connect distinct biological modules than strong domain-domain interactions (DDIs) (Kim et al. 2014), supporting the notion that they contribute significantly to innovation in regulatory networks.

Orphan CGI alternative promoter potential

As only a small percentage of the genome is responsible for coding proteins (around 2%), almost all of the remaining DNA was predicted to have no biological function. Yet, the international Encyclopedia of DNA Elements (ENCODE) project uncovered, by direct biochemical approaches, that at least 80% of human genomic DNA has biochemical activity (not to be misconstrued with biological function). This accompanied by the discovery of many functional noncoding regions and their involvement in epigenetic activity and complex networks of genetic interactions suggests that a rising percentage is being shown to have regulatory functions. The discovery of crucial regulatory functions in novel regions of non-coding DNA continues to occur as the amount of genome-scale

data increases, and is expected to rise even further in the next few decades.

The two major cis-regulatory sequences in non-coding DNA that regulate the composition of genes expression in different cell types are promoters and enhancers. They orchestrate this process by acting as templates that allow binding of specific factors recruited in response to various stimuli. Promoters are located immediately upstream of genes and serve to bind the preinitiation complex (PIC) to initiate transcription, whereas enhancers modulate transcriptional rate in a cell type specific fashion by either recruiting or stabilizing the PIC at the promoter (Maston et al. 2006).

CpG islands (or CGIs) are regions with a high frequency of CpG sites of at least 200 bp in length, and a GC percentage that is greater than 50%, and with an observed-to-expected CpG ratio that is greater than 60% (however, other operational definitions of CGI have been used). CGIs near genes are historically well studied; they are recognized as regulatory elements capable of transcription initiation that lie within promoters. Yet, about 50% of all CGIs lie outside of known promoter regions (also called orphan CGIs), and not much is understood about their biological significance. Although a previous study showed that a CGI in intron 10 of the imprinted *Kcnq1* gene (Mancini-DiNardo et al. 2003) was found to promote the initiation of a noncoding transcript (*Kcnq1ot1*) required for the imprinting of several genes at this locus. Further, tissue-specific alternative promoter activity was recently detected at an orphan CGI that promotes a specific isoform of the Rapgef4 gene (Hoivik et al. 2013). Finally, orphan CGIs have been shown to be co-opted over evolutionary time by nearby promoter-less

genes in humans (viz., retrocopies) to transcribe their gene products (Carelli et al. 2016). Thus, cumulative evidence suggests that most, perhaps all, CGIs (1) have promoter-like characteristics, (2) are sites of transcription initiation, and (3) might be poised as transcriptional initiation sites, such that in a contextually favorable configuration can function as alternative promoters for a proximally located neighboring gene. To our knowledge, aside from the anecdotal pieces of evidence, no study so far has reported unexpected promoter activity at orphan CGIs at a large scale across cell types, or reported the existence of a general regulatory mechanism involving orphan CGIs.

Organization of Thesis

In Chapter 2, we develop a novel probabilistic scoring method to comprehensively screen for cis-regulatory rewiring within regulons across 23 yeast species. Besides recapitulating known events, this work generates a high-confidence resource of previously unknown rewiring events spanning functional gene-sets and individual genes, which is then followed by extensive analysis of their functional properties.

In Chapter 3, by analyzing data from ENCODE, Epigenomics Roadmap, FANTOM amongst other sources, we show for the very first time, the pervasive ability of intergenic orphan CGIs located several kilobases upstream of methylated-promoter genes to serve as their alternative promoters. Such CGI-initiated transcription explains the tissue-specific expression of a large fraction (~ 50%) of genes with methylated promoters, as observed across 34 human tissues and cell types. We found that distal CGI-initiated transcription also explains

aberrant tumor expression of certain genes with methylated promoters, implicating the observed transcriptional mechanism in cancer. Our work adds an important piece to the puzzle that intergenic CGIs present with respect to their overall functional role.

In Chapter 4, we used a pairwise SVM method for species specific PPI prediction between 1200 TFs in 12 related arthropod species, followed by detection of protein interaction based regulatory rewiring. Based on simulation studies, we show that the accuracy of detection of rewiring events using the above PPI prediction method is approximately ~80-85%, which recapitulates the known FTZ-EXD to FTZ-FTZ1 interaction rewiring event amongst the top 5% of all events involving FTZ. We find rewiring events involving several protein members of the “enhancer of split” complex amongst the top 1% detected events, which is known to have undergone lineage specific gene losses and duplications. We expect that a deeper investigation of the rewiring events involving these protein members may reveal crucial information about regulatory network changes in neurogenesis across insect evolution.

Finally, in Chapter 5, we conclude with overall perspective and potential future directions of this work.

CHAPTER 2: High throughput identification of cis-regulatory rewiring events in yeast

Abstract

A co-regulated module of genes ("regulon") can have evolutionarily conserved expression patterns and yet have diverged upstream regulators across species. For instance, the ribosomal genes regulon is regulated by the transcription factor (TF) TBF1 in *C. albicans*, while in *S. cerevisiae* it is regulated by RAP1. Only a handful of such rewiring events have been established, and the prevalence or conditions conducive to such events are not well known. Here, we develop a novel probabilistic scoring method to comprehensively screen for regulatory rewiring within regulons across 23 yeast species. Investigation of 1713 regulons and 176 TFs yielded 5353 significant rewiring events at 5% FDR. Besides successfully recapitulating known rewiring events, our analyses also suggests TF candidates for certain processes reported to be under distinct regulatory controls in *S. cerevisiae* and *C. albicans*, for which the implied regulators are not known: 1) oxidative stress response (Sc-MSN2 to Ca-FKH2), and 2) nutrient modulation (Sc-RTG1 to Ca-GCN4/Ca-UME6). Further, a stringent screen to detect TF rewiring at individual genes identified 1446 events at 10% FDR. Overall, these events are supported by strong co-expression between the predicted regulator and its target gene(s) in a species-specific fashion (>50-fold). Independent functional analyses of rewiring TF pairs revealed greater functional interactions and shared biological processes between them (p=1e-03).

Our study represents the first comprehensive assessment of regulatory rewiring; with a novel approach that has generated a unique high-confidence resource of several specific events, suggesting that evolutionary rewiring is relatively frequent and may be a significant mechanism of regulatory innovation.

Introduction

Gene expression variability (a biomarker of phenotypic diversity) within and across species is largely brought about by the differences in transcriptional control mechanisms (Stranger et al. 2012; King and Wilson 1975) that are partly reflected in the sequences of regulatory elements, such as transcription factor (TF) binding sites (TFBS), and sequences that effect nucleosome positioning (Wittkopp and Kalay 2011; Connelly et al. 2014; Wray 2007). The converse is not necessarily true; it has been observed that genes with highly conserved spatio-temporal transcriptional patterns have highly divergent *cis*-regulatory configurations in different species (for example, *Endo16* in sea urchins (Romano and Wray 2003), *eve* and *runt* in *Drosophila* species, and many more (Weirauch and Hughes 2010)). Further, a recent comparative study of TF footprints between human and mouse showed only a small (20%) fraction of the footprints to be shared between the two species indicating a large turnover of transcription factor binding sites (Stergachis et al. 2014). Collectively, these observations support the idea that there is extensive plasticity in the *cis*-regulatory circuitry that is representative of both conserved and diverged expression programs across species - the extent of which is only beginning to be appreciated (Wray 2007;

Weirauch and Hughes 2010).

TF rewiring is a prominent mechanism of evolutionary changes in *cis*-regulation, and can occur over relatively short evolutionary timescales (Tuch et al. 2008). Essentially, specific genes (or a set of co-regulated genes) have undergone a switch in *cis*-regulation; whereby in the ancestral species the genes were regulated by a particular TF, but at a specific evolutionary lineage (represented by a subset of extant species) the genes are instead regulated by a different TF (Fig. 2-1). Such evolutionary rewiring of TFs may or not result in changes in downstream expression patterns. A well-known example of the latter type of regulatory rewiring in yeast species occurred in a set of functionally related co-expressed genes, viz., the ribosomal regulon. This regulon in *Candida albicans* and related yeast species is under the control of the DNA binding factor *TBF1*, whereas in the more recently evolved *Saccharomyces cerevisiae*, the repressor-activator protein *RAP1* regulates the transcription of the same regulon (Hogues et al. 2008). This switch of regulatory factors (from *TBF1* to *RAP1*) is likely due to the loss of binding sites for *TBF1*, and the simultaneous gain of the *RAP1* TFBS in the promoters of 60+ genes that comprise this regulon (Weirauch and Hughes 2010). As mentioned previously, in this case, the function and expression pattern of the regulon is maintained in the two species, however, since transcriptional output is the end point of signal transduction pathways, this rewiring has probably allowed the two species to respond differently to internal or external signals. Furthermore, such rewiring might even constitute changes essential for maintaining robustness in regulatory connections (Isalan et al. 2008).

Only a few examples of TF rewiring of co-regulated genes (regulons) with conserved expression patterns across species have been reported. In addition to the above mentioned RAP1-TBF1 switch in ribosomal genes, a GAL4/TYE7-GCR1 switch in glucose metabolism genes across yeast species has previously been characterized, and a GAL4-CPH1 switch in regulation was recently observed in the galactose metabolism regulon (Martchenko et al. 2007) – although these are expected to represent just the 'tip of the iceberg'. Identification of additional cases of rewiring will facilitate comparative analysis of regulation, help discover clade/species-specific instances of regulatory innovation, inform the contribution of TF rewiring in genes/processes towards adaptability, and also enable investigations of evolutionary conditions conducive to such regulatory switching. Despite its importance, no genome-wide efforts to detect rewiring events have been reported.

Here we develop a genome-scale approach to identify potential TF rewiring events in 23 related species of yeast. We utilize comprehensive DNA binding motifs for 176 yeast TFs, annotation of gene promoters and established orthology groups across 23 divergent yeast (ascomycetes) species (Matys et al. 2006; Wapinski et al. 2007b, 2007a), to inform a probabilistic function that tests for clade-specific and gene-specific rewiring of TFs. Briefly, for a TF-pair (rewiring candidate), and a select evolutionary branch (that partitions 23 species into two groups), we compute a probabilistic score which assesses the proposition that a gene is regulated by one of the TFs (say, X) in one group of species and by another TF (say, Y) in the other group of species, as illustrated in

Figure 2-1A. We thus compute a *rewiring score* (*RS*) for every gene (more precisely, orthologous gene family) and every TF-pair across six select partitions of the yeast evolutionary tree (only the branches numbered in bold/larger font in Fig. 2-1B).

Next, we apply our novel method to detect rewiring events for groups of genes involved in the same biological process and whose expression are correlated in both *S. cerevisiae* and *C. albicans*. Our broad application to 1713 regulons detected 5353 significant rewiring events (FDR<0.05). While successfully recapitulating the known rewiring events discussed earlier, our results also suggest plausible TF candidates for certain processes reported to be under distinct regulatory controls in *S. cerevisiae* (*Scer*) and *C. albicans* (*Calb*) but for which specific regulators are not known. Specifically, *MSN2*/*4* are known to be major players in controlling the response to oxidative stress in *Scer* (Elfving et al. 2014), although these TFs possess no known roles in regulating the same in *Calb* (Nicholls et al. 2004); we present evidence for the co-option of *FKH2* in regulating this process in *Calb*. Similarly, *RTG1* plays a role in regulating the metabolism of intermediates in *Scer* such that its misregulation leads to amino acid auxotrophies (Homann et al. 2009), while the same does not occur in *Calb*. Our results indicate that the promoters of some of the genes involved in this process seem to have diverged to accommodate binding sites for Ca-*GCN4*/Ca-*UME6*, thereby potentially rewiring their upstream regulator.

Furthermore, independent functional analyses of TF pairs that tend to rewire amongst themselves revealed that they (1) possess greater functional

connections ($p<1e-04$) and shared biological processes ($p<1e-03$), (2) occupy lower levels of the TF hierarchy, and (3) display strong co-expression between the predicted regulator and the target gene(s) in a species-specific fashion (>50-fold enrichment) across rewiring events. Next, to assess the significance of rewiring events at the level of individual genes, we applied a highly stringent control using a phylogeny-preserving permutation technique (called rotation test) to generate a suitable null expectation. At FDR < 0.1, we detected over 1000 significant rewiring events at the individual gene level. Similar to regulon rewiring, gene-level rewiring events are also supported by species-specific co-expression of TFs and targets, as well as greater functional connections between rewiring TFs. Finally, an assessment of TF rewiring within regulons and individual genes across 23 yeast species suggests that evolutionary rewiring is relatively frequent and may be a significant mechanism of regulatory innovation.

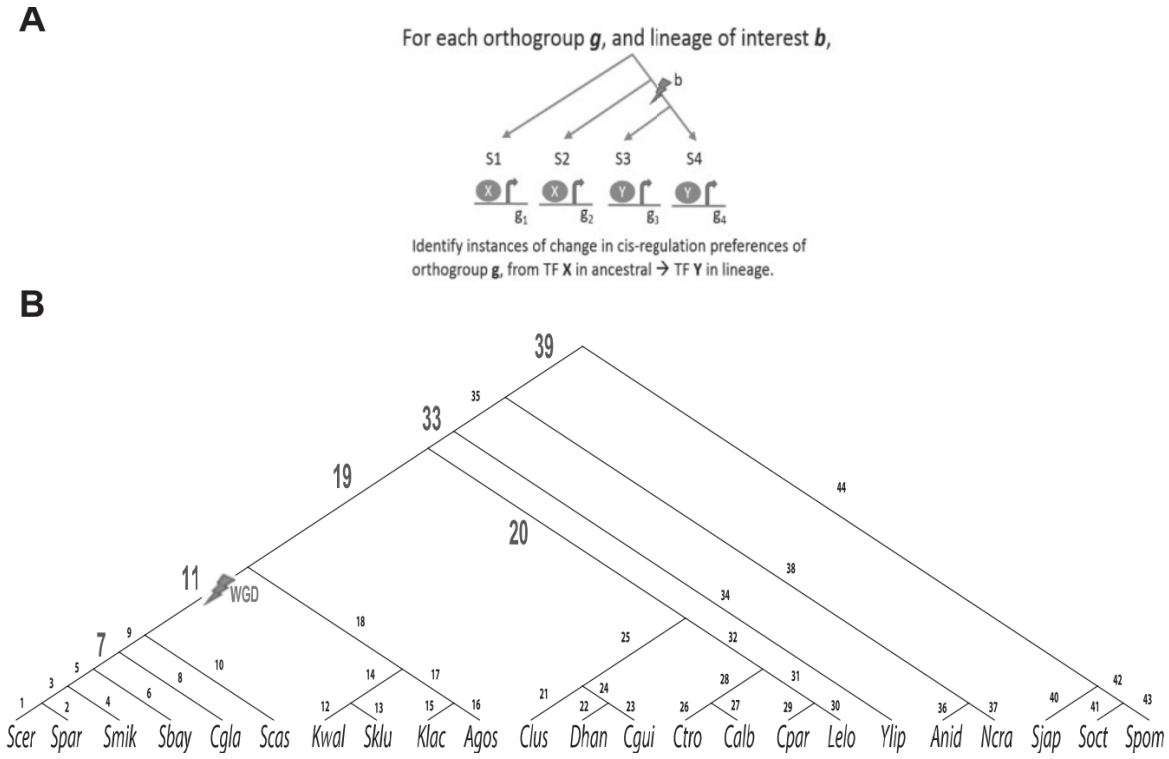


Figure 2-1. Overview of the cis-rewiring detection approach. The figure illustrates the rationale, the probabilistic function, and the search space. (A) *Toy Example:* This sample tree shows 4 species (*s1,s2,s3,s4*) partitioned at a select branch *b* to produce the partition of 2 species in the left clade (*s1, s2 Є S*) and 2 species (*s3, s4 Є T*) in the right clade. Gene locus *g* represents the orthologous group of genes across all the four species (*g1,g2,g3,g4 Є G*) that hypothetically exhibit differential usage of regulating TFs *X* and *Y,* where *X* is used by species in the left clade and *Y* is used by species in the right clade, and not vice-versa. Such an instance of change in *cis*-regulatory preferences of locus *g*, between the TF-pair (X,Y) at branch *b*, can be computed as shown, where P(TF,g,s) represents the probability of a TF binding the promoter of gene *g* in species *s*. Summation of the binding probabilities in each clade yields the rewiring score RS(X,Y,g,b). (B) *Phylogenetic tree of Ascomycetes*: Tree shows relationships between the 23 yeast species surveyed in this analysis. Branches are numbered from 1 to 44. Six branches highlighted in bold and larger font numbering represent the chosen branches across which we partitioned the species to assess lineage-specific *cis*-regulatory rewiring.

Results

A probabilistic framework to detect rewiring events

We define a probabilistic function henceforth referred to as the “rewiring score” (RS) to provide a metric indicative of how likely it is that a given gene locus (including all orthologs across 23 yeast species, or an *orthogroup*) has selectively

switched its regulator in a particular lineage. The RS function is illustrated in Figure 2-1A and described in the Methods section. Very briefly, consider orthogroup *g*, and a phylogenetic tree branch *b* that partitions the 23 species into species set *S* comprising of the species descending from the internal branch *b*, and the complement species set *T*. For TFs X, Y, rewiring score *RS(X,Y,g,b)* calculates the probability that *X* regulates *g* in the species set *S* (and *Y* does not), and *Y* regulates *g* in the species set *T* (and *X* does not). Following previous works (Habib et al. 2012; Levy and Hannenhalli 2002), the probability that a TF regulates a gene in a species is derived from the score of the TF's DNA binding motif against the gene promoter (see Methods).

High-throughput computation of rewiring scores across all orthogroups, TFs and lineages

Our goal was to comprehensively assess rewiring amongst all orthogroups across 23 extant yeast species, for all possible pairs of 176 TFs (annotated for DNA binding motif in *S. cerevisiae*). We chose 6 distinct lineages in the evolutionary tree of 23 ascomycetes to test for rewiring (Fig. 2-1B). The internal branches defining these lineages were selected based on two criteria: (1) each of the two species groups separated by the lineage comprised of at least 3 species, and (2) the partitioning is biologically meaningful, e.g., non *sensu-stricto* & *sensu-stricto* species, pre- & post- whole genome duplication, etc.

We obtained the 3844 orthogroups corresponding to protein coding genes spanning 23 yeast species from the Fungal Orthogroups Repository (Wapinski et al. 2007a). The 600 bps promoter sequences for all genes in all 23 species were

obtained from (Wapinski et al. 2007b). Using the DNA binding motifs for 176 *S. cerevisiae* TFs from TRANSFAC (Matys et al. 2006), we obtained the binding probabilities (a value between 0 and 1) of all TFs in all promoters of 23 species. We thus computed a rewiring score for all 176*175 = 30800 TF pairs for 3844 orthogroups at 6 lineages, resulting in over 118 million rewiring scores per branch. The branch-wise distributions of rewiring scores over all orthogroups and all TF-pairs are shown in Figure 2-2. It is evident that more rewiring has occurred on branch #19 than on other branches. In fact, due to the nature of the rewiring score function, the distribution of rewiring scores is dependent on the species partitioning into distinct clades, and is therefore branch-specific. This is reflected in the variation in rewiring score distributions across branches. (see Methods).

In general, TF binding motifs with high information content (IC) yield a more skewed binding probability distribution relative to TF motifs with low IC. To ensure that this inherent difference in binding properties does not introduce a bias in the rewiring scores, we categorized rewiring scores based on IC values of the two TFs (see Supplemental Fig. A-1). We found that the pooled distributions in different IC bins are not significantly different from each other, suggesting that the rewiring scores are not sensitive to differences in IC of the TF motifs.

Another potential concern is that the TF DNA binding motifs derived from *S. cerevisiae*, are used to estimate binding probability in all yeast species. Divergence in DNA-binding specificity of orthologous transcription regulators across related species is believed to occur infrequently because of pleiotropic consequences of alterations to TF DNA binding specificity (Prud'homme et al.

2007). With some exceptions (e.g. *Matα1* TFBS in yeast species (Baker et al. 2011)), previous studies have observed a strong conservation of the regulatory lexicon (~95% between mouse and human (Stergachis et al. 2014)) as well as the function of several TFs across large evolutionary distances (McGinnis et al. 1990). Our approach cannot identify these exceptions, as we scan promoters for TFBS using known TF motifs, as opposed to de-novo motif detection, which, however, is more error-prone and difficult to interpret. Although in principle, species-specific refinements of the motif can be derived, a recent work based on the same data sets used here showed that such a refinement step did not result in substantial differences in the detection of binding sites (Habib et al. 2012).

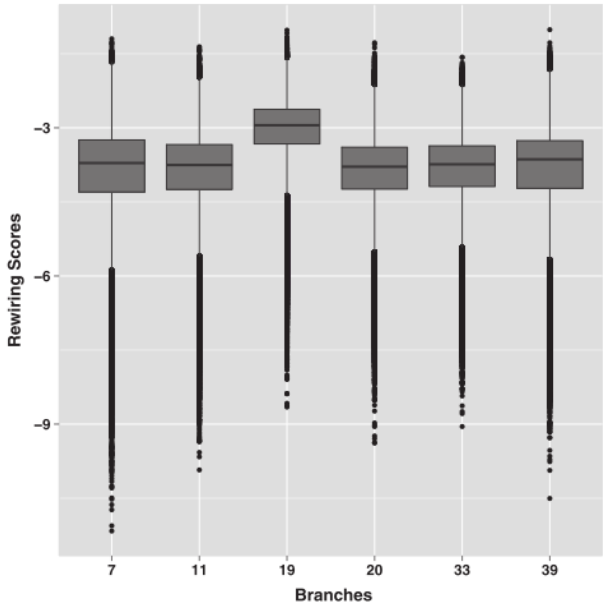


Figure 2-2. Distribution of all computed rewiring scores. Each boxplot here represents the distribution of rewiring scores (log scale terms on the Y axis) across all triplets RS(*X,Y,g*), at a chosen branch *b* ϵ (7,11,19,20,33,39) (shown on the X axis).

Regulon-level rewiring of transcription factor usage

A regulon, as described earlier, is a collection of transcriptionally co-regulated and presumably functionally related genes (Segal et al. 2003). Such coordinated regulation is evidenced by correlated expression patterns of the genes across multiple spatio-temporal conditions. Following our primary motivation of detecting coordinated changes in TF usage that are representative of conserved expression phenotypes across sets of related genes, (similar to ribosomal genes (Weirauch and Hughes 2010)), we specifically assessed those sets of genes that shared a biological function and had strongly correlated expression both in *Scer* and *Calb* (separated by over 300 MY) for rewiring of their putative upstream regulator. Very briefly, starting with gene sets corresponding to 577 distinct biological functions (Gene Ontology (GO) term) or pathways, we identified disjoint subsets of genes that exhibit highly correlated expression across hundreds of spatio-temporal conditions, both in *Scer* and *Calb*. See Methods for details. A total of 1713 gene groups, with an average size of 32 genes were assessed for regulatory rewiring.

To assess regulatory rewiring of a regulon, we computed the rewiring score for each gene in the regulon as described above, yielding a distribution of rewiring scores. To estimate the significance of this distribution, we compared it with the distribution of rewiring scores for all orthogroups (at the same lineage and for the same TF pair) using Wilcoxon test. A significantly higher rewiring score distribution for the regulon genes was interpreted as evidence for rewiring. We thus estimated significance of rewiring for each of the 1713 regulons,

176*175=30800 TF pairs at 4 select lineages (descending from internal branches $b \in 7, 11, 19, 20$) shown in Figure 2-1B. These branches were selected because they partition the two well characterized species with expression data – *Scer* and *Calb*. After correcting for multiple testing (Storey 1995), we identified 5353 significant rewiring events at FDR < 0.05. Given that our method for detecting TF binding is purely sequence-based, it is possible that the apparent "multiplicity" in cases where multiple TFs rewire at the same gene(s) and the same branch, is simply an artifact of motif similarity between detected TFs. We found that while this is true, it explains only a very small fraction of cases (see Supplemental Fig. A-2) to be of any concern.

The detected rewiring events span regulons involved in 577 processes ranging from core processes (ex. sugar and amino acid metabolism, growth, sporulation, etc.) to more specialized ones (ex. response to drug, chemical stimulus etc.), suggesting that regulatory rewiring has occurred extensively across the evolution of divergent yeast species. We discuss the detected rewiring events in the following sections.

Our method recapitulates previously established rewiring events in yeast

Rewiring of ribosomal genes. Ribosomal protein (RP) genes are crucial for cellular growth and viability. As described earlier, this fairly large regulon has a different upstream regulator in *Scer* and its closely related species (*RAP1*) as compared to the ancestral species (*TBF1*). The switch is believed to have specifically occurred along branch 19, as is shown in Fig. 2-1B (Hogues et al. 2008). While this branch represents an evolutionary period of time that precedes

the WGD event, and there might be a possible link between this switch and WGD; there is currently no evidence to support this. This particular regulatory substitution is supported by the presence of binding sites of the rewired factor as well as the explicit loss of binding sites of the replaced factor, in the corresponding species (Weirauch and Hughes 2010). Furthermore, it has been shown that both *RAP1* and *TBF1* in their respective species are involved in recruiting the *IFH1/FHL1* complex to the RP promoters (Mallick and Whiteway 2013), which are the primary regulators of RP genes. Despite the requirement of this dimer in both species, the *cis*-regulatory organization of RP genes in *Calb* is different from those in *Scer; in Calb,* these are mainly dominated by *CBF1* binding sites while lacking discernible *IFH1* sites, while the opposite pattern is observed in *Scer* (Hogues et al. 2008).

These differences in *cis*-element configurations (viz., *RAP1/IFH1* binding sites in *Scer* vs. *TBF1/CBF1* binding sites in *Calb*) of ribosomal genes are immediately apparent in the rewiring scores of the TFs implicated in the above process. Since our method does not consider combinatorial relationships between TF-binding within species, it detects all 4 pairwise combinations of TFs (viz., *RAP1-TBF1*, *RAP1-CBF1*, *IFH1-TBF1* and *IFH1-CBF1*) as having significantly rewired at that lineage. We present in the main text, results for the *RAP1-TBF1* and *IFH1-CBF1* rewiring events only, but the results are similar for all 4 cases (see Supplemental Fig. A-3 for the others).

Figure 2-3 shows the distributions of rewiring scores of the RP regulon versus the background (over all genes) for the implicated TFs. Figure 2-3A compares

the rewiring scores assessing the potential that *RAP1* regulates the genes in species diverging from a given branch *b*, and *TBF1* regulates the ancestral species. We observe that the differential in the rewiring scores for RP genes and the background is indeed the greatest at branch #19 (FDR < 1e-04). Figure 2-3B depicts plots analogous to Figure 2-3A, but for the potential that *TBF1* regulates the genes in species diverging from branch *b*, and *RAP1* regulates the ancestral species. Since this is essentially the complementary configuration, we expected to see a negative shift in rewiring scores of RP genes relative to background. Interestingly, the negative shift at branch #19 in Figure 2-3B is far more extreme than their positive shift counterpart in Figure 2-3A (compared to null; FDR < 1e-16). This is consistent with the fact that this rewiring event was mainly driven by the loss of *TBF1* sites in *Scer* and related species, rather than gain of *RAP1* binding sites (Weirauch and Hughes 2010). Figure 2-3C and 2-3D show qualitatively similar trends for the *IFH1-CBF1* rewiring event and is consistent with the rewiring between the two TFs at branch #19 (Hogues et al. 2008).

Rewiring of galactose metabolism genes. Galactose metabolism is another process that has undergone rewiring of the transcriptional circuitry, such that the upstream regulatory regions of a subset of genes encoding enzymes of this pathway have significantly diverged (viz., *GAL1*, *GAL2*, *GAL3*, *GAL7* and *GAL10*) in related fungi (Rokas and Hittinger 2007). In *S. cerevisiae*, the regulator *GAL4* positively activates transcription of these genes in response to galactose through the recognition sequence CGG(N_{11})CCG (Martchenko et al. 2007). However, in *C. albicans*, *GAL4*-mediated regulation and the same recognition sequence is

found in contexts unrelated to galactose metabolism. Martchenko et al. further suggested that the regulation of these genes in *Calb* are instead mediated by *CPH1*, the homolog of *STE12* in *Scer*; these two factors share 86% sequence similarity in their DNA-binding domain.

Indeed, analogous to RP genes, we detected significant support of rewiring in the galactose regulon genes for the two factors, *GAL4* and *STE12*. Specifically, Figure 2-4A compares the rewiring score distribution of the background (over all genes) against that of galactose metabolism genes for the potential that *GAL4* regulates the genes in species diverging from a given branch *b*, and *STE12* (or *CPH1*) regulates the ancestral species. Here, we see positive shifts in rewiring scores of the galactose regulon across all branches that separate *Scer* and *Calb*, with the highest shift in branch #19 (FDR < 0.02). See Supplemental Figure A-4A for the potential that *STE12* (or *CPH1*) regulates the genes in species diverging from a given branch *b*, and *GAL4* regulates the ancestral species. Similar to the case of RP genes, we observed significantly lower regulon rewiring scores when compared to the null background expectation in branch #19 (FDR < 0.002).

Taken together, these results suggest that this change in *cis*-configuration, and thereby regulation, occurred at branch #19. Martchenko et al. hypothesized that this switch probably occurred as a consequence of WGD (Martchenko et al. 2007), but our analysis suggests that the gain of *GAL4* binding sites, as well as loss of *CPH1* binding sites initiated before the WGD event.

Rewiring of glucose metabolism genes. In *C. albicans,* genes involved in glucose utilization are regulated by *GAL4* and *TYE7*, whereas in *S. cerevisiae* this task

has been taken over by *GCR1* and *GCR2* (Askew et al. 2009; Lavoie et al. 2009). Consistent with this event, we detect significant signal of rewiring in a subset of genes involved in glucose metabolism for the two factors, *GCR1* and *GAL4*. Specifically, Figure 2-4B compares the rewiring score distribution of the background against that of glucose metabolism genes for the potential that *GCR1* regulates the genes in species diverging from a given branch *b*, and *GAL4* regulates the ancestral species. We see positive shifts in rewiring scores of glucose metabolism genes across all branches that separate *Scer* and *Calb*, with the highest shift in branch #19 (FDR < 0.005). Similar to previous cases, we observe significantly lower regulon rewiring scores when compared to the background in the complementary scenario as shown in Supplemental Figure A-4B (FDR < 0.05).

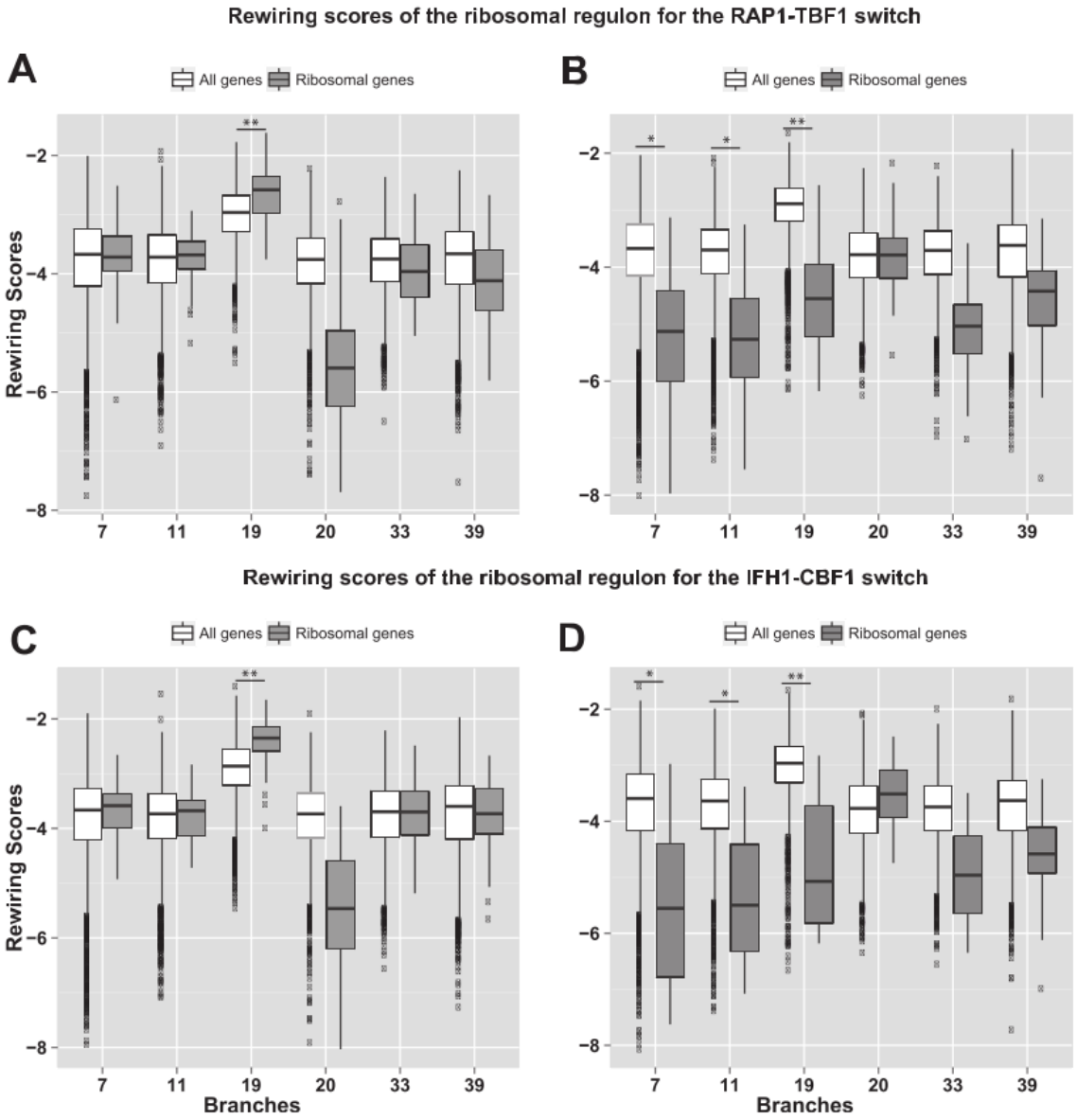


Figure 2-3. Rewiring scores of the ribosomal regulon for *RAP1-TBF1* and *IFH1-CBF1* switches across branches. The rewiring scores are shown on the Y axis, and the selected branches are shown on X axis. **(A)** *RAP1 in lineage & TBF1 in ancestral species*: This plot compares the rewiring score distribution of the background (all genes; in white) and that of ribosomal genes (in grey) for the potential that *RAP1* regulates its member genes in species diverging from a given branch *b*, and *TBF1* regulates the ancestral species. **(B)** *TBF1 in lineage & RAP1 in ancestral species*: This plot compares the rewiring score distribution of the background (in white) and that of ribosomal genes (in grey) for the potential that *TBF1* regulates its member genes in species diverging from a given branch *b*, and *RAP1* regulates the ancestral species. **(C)** & **(D)** are analogous to (A) & (B) respectively for the *IFH1-CBF1* switch in RP genes.

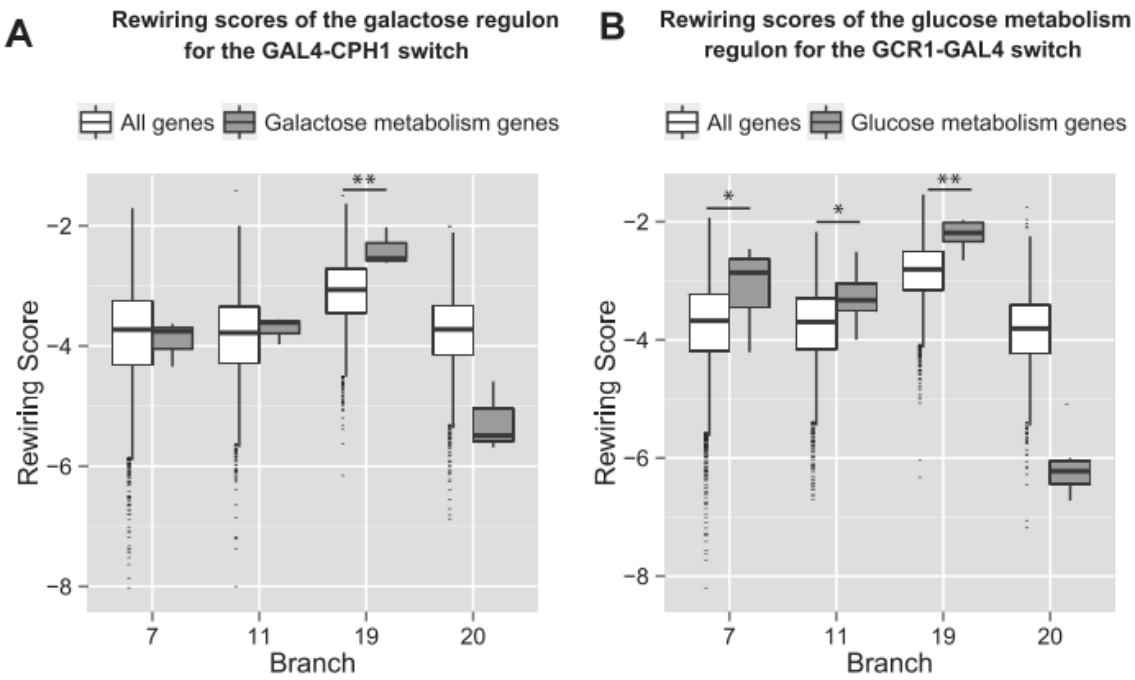


Figure 2-4. Rewiring scores of other known rewiring events across branches. See Figure 2-3 legend for details. **(A)** Galactose regulon rewiring scores for *GAL4 in lineage & CPH1 in ancestral species.* **(B)** Glucose metabolism regulon (subset) rewiring scores for *GCR1 in lineage & GAL4 in ancestral species.*

Identifying candidate TFs for known rewiring events

Next, we searched the literature for processes that are reported to be under distinct regulatory controls in *S. cerevisiae* and *C. albicans*, but for which specific regulators have not been implicated, and assessed whether our probabilistic method can help identify potential regulators in these cases.

Stress response. Zinc finger TFs *MSN2/4* bind to highly similar motifs and are the primary regulators of response to a variety of stresses (nutritional, oxidative, etc.) in *Scer*. Here, MSN2 elicits a complex response to stress, whereby different cohorts of target genes respond differently, resulting in either gene expression activation or repression (Elfving et al. 2014). However, these TFs are not known

to play a role in stress response in *Calb*; the disruption of Ca-*MSN2/4* had no tangible effect on the resistance of the species to heat, oxidative, and osmotic stresses (Nicholls et al. 2004). Consistent with the rewiring of stress response regulators in the two species, we found that the regulon involved in oxidative stress response shows strong signals for regulatory rewiring of *MSN2*/*MSN4* regulating these genes in *Scer* to being regulated by *FKH2* in *Calb*. Specifically, Figure 2-5A compares the rewiring score distribution of the background (over all genes) against that of stress response genes for the potential that *MSN2/4* regulates the genes in species diverging from a given branch *b*, and *FKH2* regulates the ancestral species. Here, we see significant positive shifts in rewiring scores of the regulon in branch #19 (Fig. 2-5A; FDR < 0.01), and significant negative shifts for the complementary scenario akin to previous cases (Supplemental Fig. A-5A; FDR < 0.01). Although short of a direct experimental validation, our finding is supported by a prior study showing that Ca-*FKH2* mutants in *C. albicans* resulted in increased transcript levels of genes involved in stress response (Bensen et al. 2002).

Metabolism. Retrograde (RTG) signaling, triggered by lack of glutamine, modulates carbohydrate and nitrogen metabolism through nuclear accumulation of the heterodimeric transcription factors, *RTG1/3* (Giannattasio et al. 2005). This accumulation and subsequent binding to metabolic gene targets allows cells to maintain synthesis of α-ketoglutarate, which is a precursor to glutamate and glutamine (Crespo and Powers 2002) (the latter is a preferred nitrogen source in yeast (Crespo and Hall 2002); lack of which leads to amino acid starvation). It

has been shown that deletion of TF *RTG1* in *Scer* causes glutamate and aspartate auxotrophies, yet deletion of its ortholog in *Calb* does not result in the same phenotype (Homann et al. 2009).

Our results indicate that the promoters of some of the genes involved in carbohydrate metabolism (glycolysis regulon, as well as a subset of genes involved in glucose metabolism) display an aggregate loss of *RTG1* binding sites in *Calb*, and a concomitant gain of binding sites for Ca-*GCN4* and Ca-*UME6* respectively, thereby potentially rewiring their regulation (Fig. 2-5B,C). Figure 2-5B compares the rewiring score distribution of the background (over all genes) against that of glycolysis genes for the potential that *RTG1* regulates the genes in species diverging from a given branch *b*, and *GCN4* regulates the ancestral species. The plot depicts positive shifts at branches separating *Scer* and *Calb* into distinct clades (FDR < 0.05). Supplemental Figure A-5B shows the complementary scenario with corresponding negative shifts (FDR < 0.05). Figure 2-5C (FDR < 0.05) and Supplemental Figure A-5C (FDR < 0.05) show qualitatively similar trends to Figure 2-5B and Supplemental Figure A-5B respectively for the R*TG1-UME6* rewiring event in glucose metabolism genes, and is consistent with the rewiring between the two TFs. These regulatory changes potentially result in *Calb* evolving alternate responses to the lack of glutamine, or to the lack of intermediates essential for amino acid synthesis to prevent starvation. *GCN4* is known to be involved in amino acid starvation responses that include (i) amino-acid biosynthesis, (ii) increasing expression of autophagy genes, and (iii) repressing genes encoding ribosome proteins

(Hinnebusch 2005). Similarly, *UME6* in *Calb* is part of a signaling cascade that regulates autophagy (Bartholomew et al. 2012) and is also involved in regulating hyphal (filamentous) growth (Banerjee and Thompson 2008), a phenotype better suited for nutrient scavenging.

Note: Graphical illustrations of the binding site profile of rewiring TFs for all significant events described in the above two sections are shown in Supplemental Figure A-6. For conciseness, we only show TFBS profiles for regulons in *Scer* and *Calb* for each rewiring event.

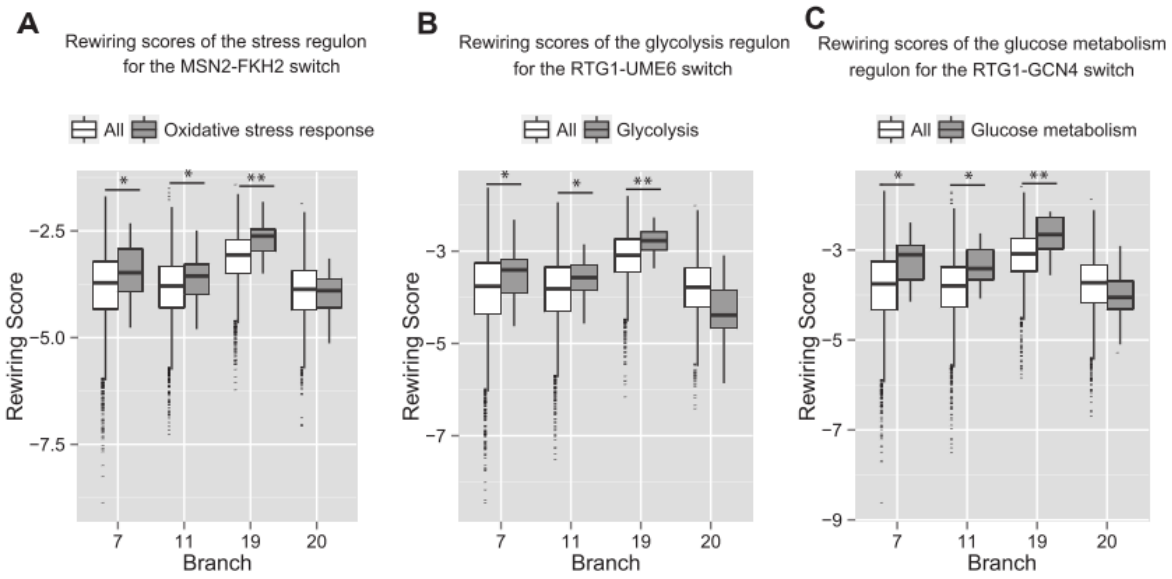


Figure 2-5. Rewiring scores of predicted rewiring events across branches. See Figure 2-3 legend for details. **(A)** Oxidative stress response regulon rewiring scores for *MSN2 in lineage & FKH2 in ancestral species*. **(B)** Glycolysis regulon rewiring scores for *RTG1 in lineage & UME6 in ancestral species*. **(C)** Glucose metabolism regulon (subset) rewiring scores for *RTG1 in lineage & GCN4 in ancestral species*.

Rewiring events are strongly supported by co-expression between the regulator and targets

Even though the causal link between TF gene level and the target gene level is

confounded by (i) low constitutive expression of many TFs and (ii) regulatory mechanisms including post-translational modifications, co-factors etc., in general, expression of TF genes and their target genes are expected to be correlated across different environments to some extent (Basso et al. 2005). We assessed if such correlated expression patterns are apparent among the 5353 detected rewiring events. Specifically, we tested if the expression of the predicted regulator of a regulon correlates with the expression of the regulon's component genes in a species-specific fashion (using expression data in *Scer* and *Calb* from (Ihmels et al. 2005)). For instance, in the case of RP regulon rewiring, we assessed whether *RAP1* is co-expressed with the RP regulon genes in *S. cerevisiae*, but not in *C. albicans*, and whether the converse was true for *TBF1*?

For each significant rewiring event at the regulon level (say between TF *X* and TF *Y* in regulon *R*), we carried out the following analysis. We collected 4 different sets of Spearman correlations between, *viz*. (1) TF *X* and *R* genes in *Scer*, (2) TF *Y* and *R* genes in *Calb*, (3) TF *Y* and *R* genes in *Scer*, and (4) TF *X* and *R* in *Calb*. Out of 5353 events, complete expression correlation data for all 4 sets was available for 3030 cases. Since the detected rewiring events predict which regulator is being used by which species, we simply calculated the number of cases in which the set of correlations between the TF predicted in a given species and the target genes are significantly greater (Wilcoxon p-val ≤ 0.05) than those for the TF not being used in the species. We required this condition to be satisfied in both *Scer* and *Calb*. We found that in 493 of the 3030 cases, the co-expression patterns support the predicted regulatory switch in both species,

which is strong evidence in light of the null expectation (16% vs. 0.25%; 65-fold enrichment). In fact, most of this enrichment is localized to rewiring events specific to the WGD branch (32% vs. 0.25% at WGD branch #11).

In case of RP genes the degree of co-expression in *Scer* of *RAP1* and *TBF1* with the RP genes (Fig. 2-6A) were comparable (Wilcoxon p-val = 0.53). However, in *Calb* the co-expression between *TBF1* and RP genes was significantly higher than that of *RAP1* and RP genes (Wilcoxon p-val < 1e-07). In case of glucose metabolism, co-expression of *GCR1* as well as *GAL4* with the glucose metabolism genes (Fig. 2-6B) is consistent with the direction of TF rewiring in *Scer* (Wilcoxon p-val < 1e-06); this could not be tested in *Calb* due to the absence of an annotated *GCR1* homolog. Finally, in case of the galactose regulon, absence of sufficient co-expression datapoints for *GAL4* and *STE12* (or *CPH1*) (due to small regulon size, comprising of 3 genes) limits the analysis.

In case of oxidative stress response genes, we observe co-expression support (Fig. 2-6C) for the implicated TFs (*MSN2/4* and *FKH2*) only in *Scer* (Wilcoxon p-val = 0.03), but not in *Calb* (Wilcoxon p-val = 0.22; although co-expression median of “True Regulator-Regulon” > “Replaced Regulator-Regulon” in *Calb*). On the other hand, co-expression of regulator and targets in retrograde response is as follows: (1) For genes involved in glycolysis (Fig. 2-6D), we observe strong co-expression support for the implicated factors (*RTG1* and *UME6*) in *Scer* (Wilcoxon p-val < 1e-05) as well as *Calb* (Wilcoxon p-val < 2e-04). (2) For a subset of genes involved in glucose metabolism (Fig. 2-6E), we observe strong co-expression support for the implicated factors (*RTG1* and *GCN4*) only in *Calb*

(Wilcoxon p-val < 0.01), but not in *Scer*. Further it can be seen from Figure 2-6 that even when the co-expression between replaced regulator and regulon genes is relatively low, the co-expression of the replaced regulator with all genes is still high. This suggests that the differences in co-expression levels of true and replaced regulators with their putative gene targets across species, is not simply due to an overall reduced expression of the replaced regulator in a given species, and in most cases the opted-out regulator still functions to regulate genes involved in other processes. For example, although *CPH1* (or its homolog *STE12* in *Scer*) does not regulate galactose metabolism genes in *Scer* anymore, it is still involved in regulating genes involved in mating type determination.

Thus, although co-expression of TFs and targets consistent with the direction of rewiring in the member species is not a necessary condition for rewiring (as illustrated for RP genes), we observe strong overall support for target regulation in a species-specific fashion. Interestingly, as mentioned above, this support is the highest for rewiring events occurring at a branch associated with WGD (branch #11); WGD is linked with higher degrees of expression and protein sequence divergence (Ha et al. 2009; Li et al. 2015).

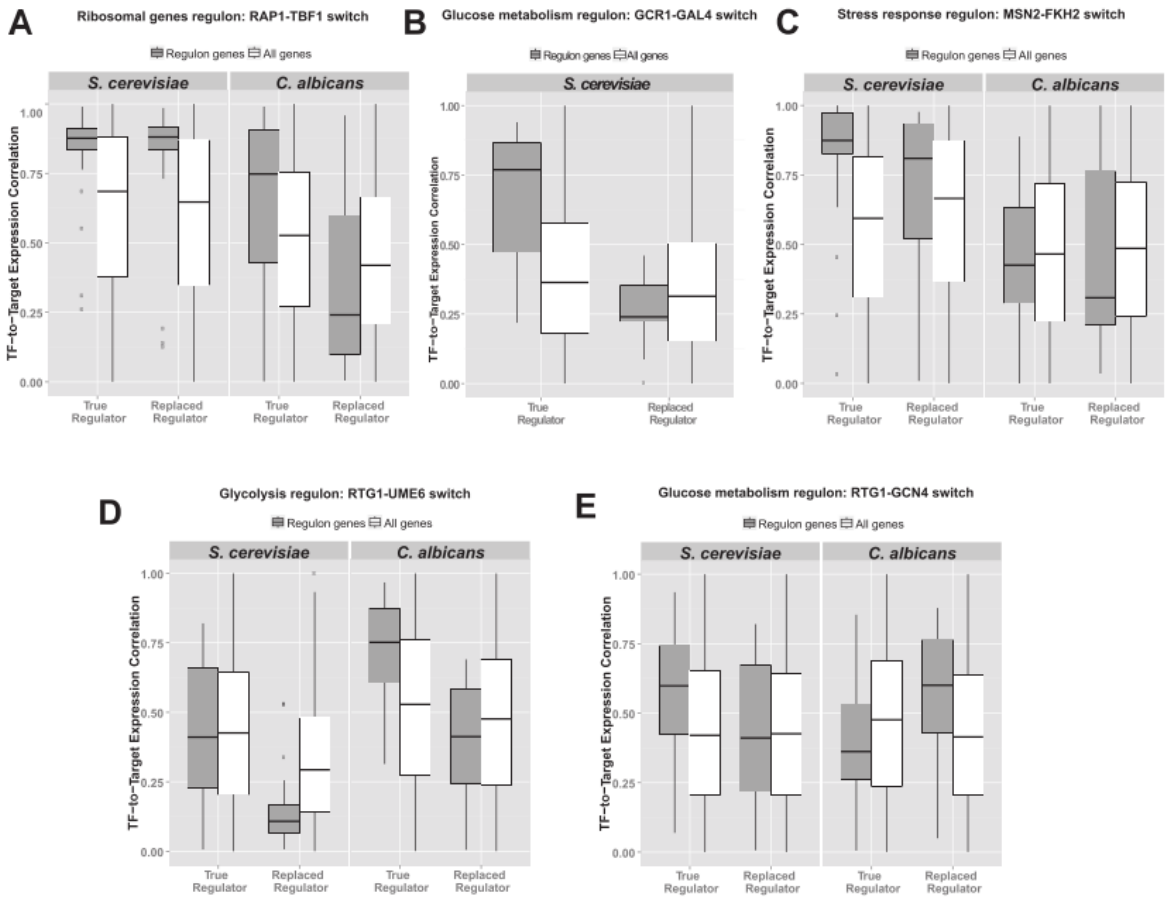


Figure 2-6. Co-expression analyses for regulon rewiring events. In each panel from A-F, we compare the TF-to-Target expression correlations on Y axis for the candidate regulator (e.g., *RAP1* in *Scer*) and the replaced regulator (e.g., *TBF1* in *Scer*) on X axis. The distribution of correlations with regulon gene targets is shown in grey, while that with all genes (comprising the background) is shown in white. The facets in each panel represent individual species (*Scer* and *Calb*).**(A)** *Ribosomal genes regulon (RAP1-TBF1)*. **(B)** *Glucose metabolism regulon (GCR1-GAL4).* **(C)** *Oxidative stress regulon (MSN2-FKH2)*. **(D)** *Glycolysis regulon (RTG1/3-UME6)*. **(E)** *Glucose metabolism regulon (RTG1/3-GCN4).*

Functional connections between rewiring TFs and their properties.

Next, we investigated the functional characteristics of rewired TF pairs that might have enabled or facilitated rewiring. It is plausible that aspects such as protein domain similarities, an increased propensity for physical interaction, coordinated expression across conditions for the two TFs, as well as their mutual involvement

in common biological processes/pathways could individually or synergistically facilitate rewiring between the two TFs. For example, *RAP1*, like *TBF1* is a Myb family protein (Bhattacharya and Warner 2008) and has similar GC-rich binding specificities (Weirauch and Hughes 2010); this could have predisposed *RAP1* to acquire the competency for RP regulation. We assessed the extent to which these different features are enriched among rewiring TFs.

Physical interaction potential: First, we compared the propensity for physical interaction between rewired TFs relative to randomly selected TFs. We used PPI annotations from BioGRID database (Chatr-Aryamontri et al. 2015) for proteins known to physically interact in *S. cerevisiae* and binned TF pairs in a 2x2 contingency table based on whether or not they interact and whether or not they rewire. Based on a Fisher's test, we did not observe a greater propensity for direct interaction between rewiring TF pairs (Fig. 2-7A; Odds-ratio = 1.02; Fisher's pval = 0.93). We obtained qualitatively similar results using PPI annotations from STRING database (Franceschini et al. 2013) (see Supplemental Fig. A-7). While such direct interaction potential between rewiring TFs is absent, it has previously been suggested that if members of rewiring TFs can bind and co-localize with a common co-factor to cooperatively regulate a target(s), then a series of small successive changes in the component interactions comprising such combinatorial control could ultimately result in a regulatory handoff between rewiring TFs across evolutionary time (Tuch et al. 2008). For example, the *cis*-element rewiring between *RAP1* and *TBF1* in RP genes was accompanied by a change in the protein domain of a common co-

factor that they interact with (a heterodimer containing *IFH1 and FHL*). Specifically, correlated with the transition to the *RAP1*-regulated circuit in *Scer*, the Sc-*IFH1* acquired a *RAP1* interaction domain (Mallick and Whiteway 2013) that is not present in the *Ca-IFH1* protein. To assess this possibility, we tested (1) if rewired TF pairs possess a common interacting TF more often than other TF pairs, and (2) if the commonly interacting TF is more likely to bind to the target gene's promoter when compared to other promoters. While the first test only shows mild support (although, not statistically significant) for the expected trend (Fig. 2-7B; Odds-ratio = 1.2; Fisher's p-val = 0.112), the second test showed a highly significant trend (Fig. 2-7C; Odds-ratio = 14.1; Fisher's p-val = 1e-04). Overall this suggests that cooperative binding of rewiring TFs to a common factor is perhaps one of the potential mechanisms facilitating regulatory rewiring.

Structural Similarity: Second, we gathered the structural family annotations of the TFs (Pfam; (Finn et al. 2014)), and tested if rewired TF pairs belong to the same family more or less often than background. We observed that the rewired TFs in fact belong to the same TF family less often than the random non-rewired TF pairs (Fig. 2-7D; Odds-ratio = 0.67; Fisher's p-val = 0.001). Although the reasons for depletion of co-family TFs amongst rewiring TF pairs is not entirely clear, we suspect it may partly be due to functional divergence of paralogous genes, consistent with our other results showing a greater functional similarity between the rewired TFs.

Common Pathways: Third, we hypothesized that the co-option by a group of genes of an alternate regulator may be influenced by whether or not the rewired

TFs are already functioning in the same pathways. Based on KEGG pathway annotations we assessed if TFs implicated in the same pathway are more likely to rewire than those involved in different pathways. TFs annotations in KEGG were limited to cell cycle, signaling pathways and meiosis, which substantially reduced the number of pairs we could test. Nevertheless, we observed greater likelihood for TFs of common pathways to rewire than that expected by random chance (Fig. 2-7E; Odds-ratio = 3.4; Fisher's p-val = 0.001).

Regulatory Hierarchy: Previous studies of the effects of network rewiring events (insertion or deletion of connections) in a broadly constructed regulatory hierarchy of transcriptional factors in yeast suggest that rewiring affecting upper levels of such a hierarchy are much less tolerated and result in cell proliferation and survival defects, when compared to those affecting lower levels of the hierarchy (Bhardwaj et al. 2010). Also, these upper-level regulators were found to exhibit fewer functionally redundant copies across species. In light of these characteristics, we expect that the TFs in the upper level of the hierarchy should be less prone to rewiring. Using data on regulator hierarchy across 90 transcription factors (Bhardwaj et al. 2010), we indeed observe that there is significant depletion of rewiring events involving TFs belonging to the highest level of regulation when compared to lower and middle level TFs (Fig. 2-7F; Odds-ratio = 1.67; Fisher's p-val = 0.004).

Common Upstream Regulator: Next, we assessed the possibility that member TFs of a rewired TF pair are regulated by a common upstream regulator (UpR), and that this differential regulation between the species of a lineage and the

ancestral species has enabled rewiring of the two factors. In general however, this UpR may not be directly regulating either of the rewired TFs, but may instead exist further upstream in the regulatory network, at a point from which two alternative paths leading to the two rewired TFs originate. This possibility can be tested using species-specific TF-TF regulatory networks and checking if the two rewired TFs lead up to a common upstream regulator in their respective species-specific networks. We generated TF-TF regulatory networks for *Scer* and *Calb* independently (see Methods). Next, for every TF pair, we checked if the members of the pair link to a common UpR in their respective species, such that the sum of the shortest path lengths to that UpR is smaller than that expected by random chance (i.e., shortest path length to common UpR for random TF pairs). We use shortest path length from the TF pair to the UpR as a proxy for the presence or absence of the UpR, i.e., the smaller the metric is, the greater the chance that the common UpR exists. Similar to the analyses above, we binned TF pairs into whether or not they are close to a common UpR and whether or not they rewire across species (2X2 contingency table). Using a Fisher's test, we conclude that rewired TF pairs do in fact possess a common UpR, more often than random TF pairs (Fig. 2-7G; Odds-ratio = 1.28; Fisher's p-val = 1e-04). To remove possible confounding effects in the computation of shortest paths due to widely connected master regulators, we removed the top 5% TFs with the greatest degree before computing shortest path between nodes. This however does not affect our conclusion (see Supplemental Fig. A-8). This notion of a nearby common UpR is further supported by a prior study in yeast (Ucar et al.

2009) that recovered *MSN2* and *FKH2* in a TF-TF interaction pathway (in oxidative stress conditions) that they generated by combining ChIP-chip, motif binding sites, nucleosome occupancy and mRNA expression datasets in a probabilistic framework. Similarly, they recovered *RTG1*/3 and *GCN4* from a network generated for amino-acid starvation (AAS) conditions.

Taken together, our results suggest that regulon rewiring under conserved target expression is limited to the lower level TFs in a given pathway, such that they might not necessarily interact with each other, but are be implicated in the same process/pathway in a broader context.

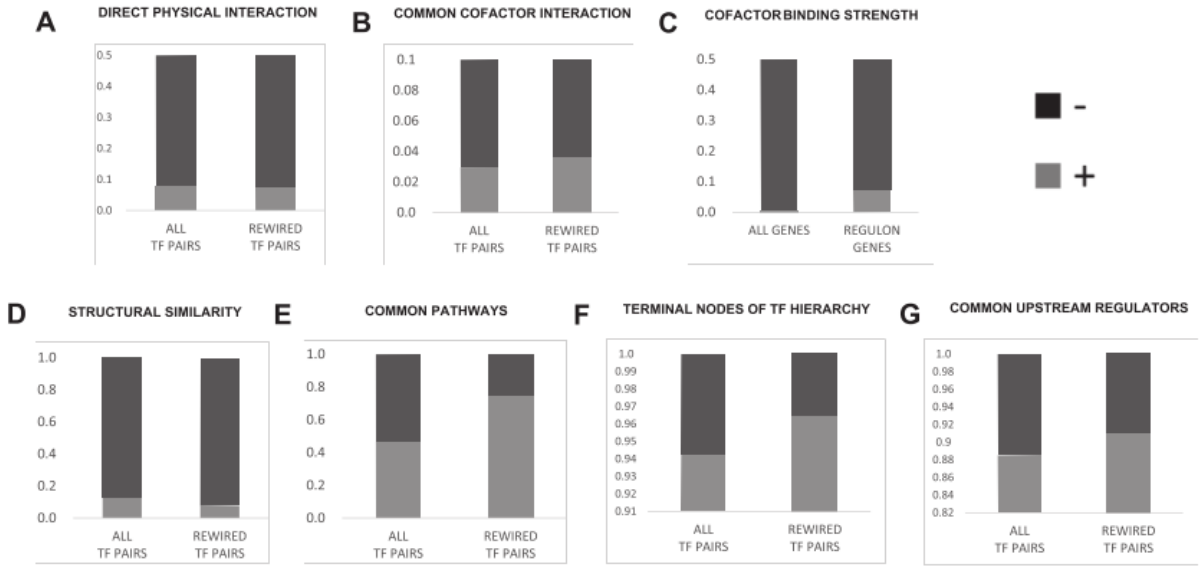


Figure 2-7. Functional analyses of rewired TFs in regulons. Each panel represents Fisher test of a specific hypothesis. In each case rewired TF-pairs and all other TF pairs are binned into two classes based on a given functional criteria (except panel C, where the bins are regulon genes and all other genes), and compared using Fisher's exact test. **(A)** *Direct physical interaction:* Based on BioGRID database, the plot shows the fraction of TF-pairs that do (light-grey) and do not (dark-grey) physically interact. **(B)** *Physical interaction with a common cofactor TF:* Based on BioGRID database, the plot shows the fraction of TF-pairs that do (light-grey) and do not (dark-grey) possess a common cofactor TF. **(C)** *Cofactor binding at target regulons:* This plot shows the fraction of cofactor TFs that do (light-grey) and do not (dark-grey) bind strongly at gene promoters (>= 0.75, vs. < 0.75 binding scores). **(D)** *Structural similarity:* This plot shows the fraction of TF-pairs that do (light-grey) and do not (dark-grey) belong to same structural family. **(E)** *Common KEGG pathways:* This plot shows the fraction of TF-pairs that do (light-grey) and do not (dark-grey) belong to same KEGG pathway. **(F)** *TF Hierarchy:* This plot shows the fraction of TF-pairs

that do (light-grey) and do not (dark-grey) belongs to lowest and middle hierarchies. **(G)** *Common upstream regulator:* This plot shows the fraction of TF-pairs whose distance to a common upstream regulator is ≤ 4 (light-grey) or > 4 (dark-grey).

Gene-level assessment of rewiring using rotation test

The application of the rewiring score function to each orthogroup and TF-pair triplet across 6 different evolutionary lineages resulted in ~650 million individual rewiring scores (~110 million per branch). Thus, a major challenge was to devise a stringent control to assess significance of each individual rewiring score. Therefore, we employed a rotation test (Langsrud 2005) based FDR approach whereby a background distribution of rewiring scores using controlled permutations of TF binding probabilities across species is generated, and compared to the distribution of observed rewiring scores to get a FDR for each datapoint in the set. We expect binding probabilities for a TF at an orthogene in sister species to be very similar due to expected sequence similarity. Traditional permutation of these binding probabilities would sample from each variable (binding probability in a given species) independently, despite the fact that there is an inherent constraint in the range of values each variable can adopt due to their mutual relationship, thus leading to overestimation of significance. The controlled permutation method called rotation test essentially has the effect of permuting the binding probabilities of a given TF across species while preserving the inherent phylogenetic relationships between species to simulate neutral evolution.

This is essentially equivalent to sampling from binding probabilities across species while maintaining a fixed co-variance structure that represents the

constraints of phylogeny; we derive this co-variance matrix from the concatenated binding probabilities for all TFs at all gene loci in each species, which serves as a suitable proxy for phylogeny. Thus, binding probabilities for each TF across 23 species were permuted as above, and the “rotated” binding probability profiles of each TF were subsequently used to compute background rewiring scores. The details of the rotation test are provided in M&M. We estimated the False Discovery Rate (FDR) of every rewiring score. The background generation and FDR calculation was done independently for each of the 6 lineages mentioned above. At 0.1 FDR we identify 1446 significant gene-level rewiring. Although the total number of detected events is much smaller compared to regulon-level rewiring (which is expected due to our use of a highly stringent control), most of these are along branches #19 and #20 (Fig. 2-1B) that best separate *C. albicans* from *S. cerevisiae*, consistent with our results from regulon-level findings.

Similar to regulon rewiring events, we assessed whether the expression correlation between the TF (say X) and the predicted target in a species is higher than that for replaced TF (say Y) and the same target. To this end, using the entire set of predicted (*X,Y,g*) events, we collected 4 sets of pooled correlations between: (i) *X* and *g* in *Scer*, (ii) *Y* and *g* in *Scer*, (iii) *X* and *g* in *Calb*, (iv) *Y* and *g* in *Calb* across all branches. As shown in Figure 2-8, rewiring events are strongly supported by co-expression as in the case of regulon-level rewiring (Wilcoxon p-val in *Scer* = 1e-05, Wilcoxon p-val in *Calb* = 1e-03). Most of the co-expression support in this case is also driven by branch #11, similar to what we

observe for regulon rewiring (see Supplemental Fig. A-9).

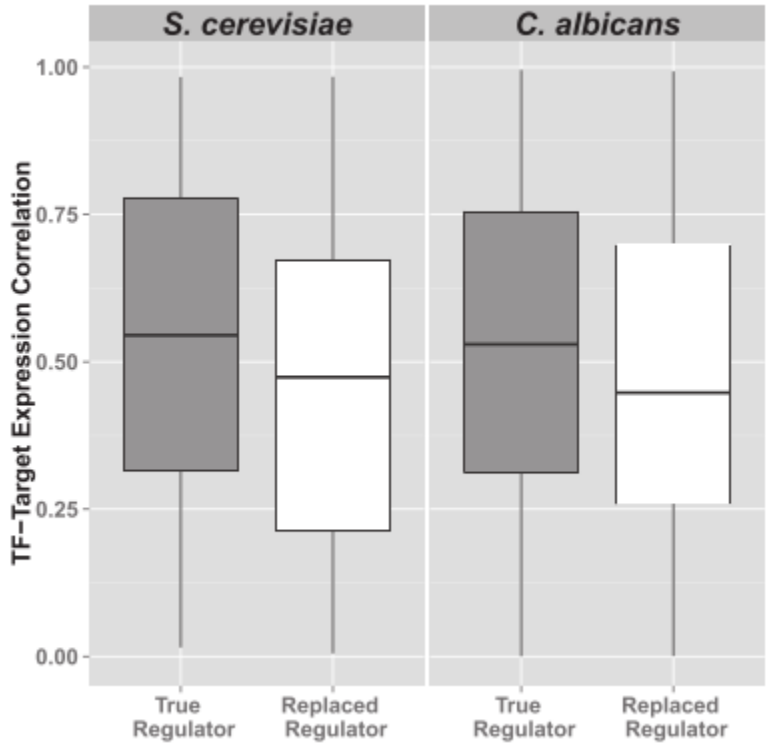


Figure 2-8. Rewiring at the individual gene level. *Species-specific TF-target co-expression analysis:* In each panel, the predicted TF-target expression distribution is shown for the TF predicted to be active in a species (grey) and for the TF predicted not be active in the species (white); the distribution is based on pooled correlations across all significant rewiring events.

Next, we investigated functional characteristics of rewired TF pairs represented in individual gene rewiring events (Fig. 2-9). Akin to our regulon-level analyses, we found that the rewired TF pairs, (1) do not necessarily interact physically with each other (Fig. 2-9A; Odds-ratio = 1.09; Fisher's p-val = 0.9) but yet, show mild potential (although, not statistically significant) for direct interaction with a common co-factor (Fig. 2-9B; Odds-ratio = 1.2, Fisher's p-val = 0.2) that co-operatively regulates its target genes (Fig. 2-9C; Odds-ratio = 16.1, Fisher's p-val= 1e-04), (2) have lower than expected structural similarity (Fig. 2-9D; Odds-

ratio = 0.77; Fisher's p-val = 0.001), (3) are enriched in common pathways, although to a lesser extent than rewired TFs in regulons (Fig. 2-9E; Odds-ratio = 2.15; Fisher's p-val = 0.01), and (4) are enriched in TFs belonging to the lower or middle hierarchies (Fig. 2-9F; Odds-ratio = 2.02; Fisher's p-val = 0.001). However, interestingly, we observed that TF pairs found to rewire in the context of regulating individual genes are 'less' likely to possess a common upstream regulator than random expectation (Fig. 2-9G; Odds-ratio = 0.8; Fisher's p-val < 0.01). This difference between TFs regulating genes with conserved expression patterns (i.e., at the regulon level), and TFs regulating genes with possibly divergent downstream expression (at the gene level) across species, could be an important distinguishing property of the mechanism of rewiring that leads to these alternate scenarios. To test this, we divided the detected gene level targets into groups with conserved and diverged expression patterns, and assessed the potential of rewired TFs in each group to possess a common UpR (see Supplemental Fig. A-10). We found that while TFs regulating genes with conserved expression patterns showed no trend (Fisher's p-val = 0.24), TFs regulating genes with divergent expression patterns were indeed less likely to possess a common UpR than expectation (Fisher's p-val < 0.03). Note that since this analysis was carried out on few high-confidence conserved and diverged expression target genes, their small sizes limit our ability to conclude meaningful functional trends.

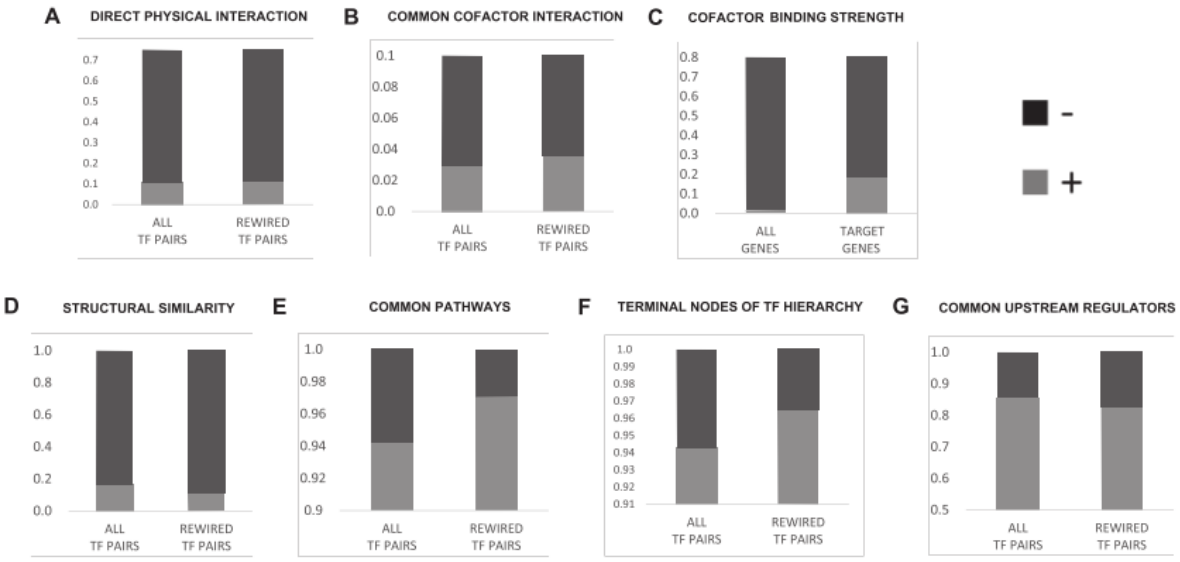


Figure 2-9. Functional analyses of rewired TFs in individual genes. See Figure 2-7 legend for details.

Discussion

Prior studies have characterized a few cases of regulatory rewiring of specific genes/gene-sets in great depth (Martchenko et al. 2007; Lavoie et al. 2010; Hogues et al. 2008). These previous works provide important insights into aspects of rewiring. For example, Mallick and Whiteway (Mallick and Whiteway 2013) showed how regulatory connections local to rewired TFs can change to preserve gene target expression patterns (for example, recruitment of *IFH1-FHL1* to ribosomal gene targets is maintained in both systems). Yet, there are several aspects of regulatory rewiring that are poorly understood. For instance, (i) how widespread is a wholesale shift in transcriptional regulation of a regulon?, (ii) what are the features of target genes that make them amenable to rewiring?, (iii) what characterizes rewired TFs, etc. By gathering more candidate rewiring events and collectively analyzing their trends, we can potentially answer these

questions and gain further insights into conditions conducive to rewiring, as well as enable discovery of clade/species-specific instances of regulatory innovation.

A genome-wide screen for TF rewiring has not been reported thus far. Here, we present the first scalable probabilistic approach to detect rewiring. Its application to 23 yeast species has successfully recapitulated known rewiring events (in ribosomal genes, sugar metabolism genes, etc.), and also has generated specific testable hypotheses of rewiring in many genes, as well as regulons.

Similar to previous related work (Habib et al. 2012; Roy et al. 2013), ours is based on estimated TF binding probabilities and not *in vivo* binding. This is not a limitation of the approach but that of the availability of functional binding information such as ChIP-seq across 100+ TFs or high resolution DNAse footprinting in all 23 species of yeast. Further, reliance on such experimental data is limited due to their condition-specificity, that an *in silico* approach avoids (Roy et al. 2013). On the other hand, there are strong arguments for using experimentally derived data when available, as *in silico* motif-based prediction of *cis*-regulation can be noisy. We attempted to reduce the noise by using experimentally measured nucleosome occupancy data (Tsankov et al. 2010, 2011) from 13 yeast species to gather additional support for functional binding. We have described this analysis in the legend of Supplemental Figure A-11, which suggests that incorporating nucleosome occupancy is not likely to improve the sensitivity of our approach. This is generally expected due to a poor association between nucleosome free regions (NFRs) and TF binding in yeast; Ozonov and van Nimwegen (Ozonov and van Nimwegen 2013) showed that only

10-20 out 158 TFs contribute to inducing NFRs and that nucleosome positioning in mainly determined by intrinsic sequence. Another study by Thompson et al. (Thompson et al. 2013) showed that TF binding sites are depleted from NFRs in most post-WGD species. Thus in our assessment, integrating nucleosome occupancy data in our analyses decreases statistical power without necessarily decreasing noise. Additionally, the presence of several high quality PWM motif matches for a certain TF in a gene promoter would increase the confidence in the corresponding TF-gene regulation. As shown in Supplemental Figure A-12, we found that amongst the detected rewiring events, regulons possess a significantly higher number of motif hits for their true regulator compared to those for the replaced regulator across all species, thus providing support for our approach.

Our rewiring score is based on the partition of the species on a defined lineage, and utilizes the binding probabilities in all extant species. Although our significance assessment does control for the phylogeny, in principle, the inherent phylogenetic relationships between species would be better exploited if ancestral sequences could be inferred at various internal nodes of the tree and rewiring score were computed based on the inferred ancestral sequences. However, ancestral sequence reconstruction (ASR) (for which we used FastML from Ashkenazy et al. 2012) relies critically on the quality of multiple sequence alignment (MSA), which is a major concern because the promoters of extant yeast orthologs are highly diverged with a potentially large amount of binding site turnover. We therefore first assessed the quality of the MSA, generated using two methods: M-Coffee ((Wallace et al. 2006); generates alignment consensus

from multiple progressive and iterative methods) and PRANK ((Löytynoja and Goldman 2008); a phylogeny aware method for sequence alignment). We found that, as suspected, the length of the ancestral sequence produced by both methods were on average twice the length of the longest individual promoter (see Supplemental Fig. A-13), suggesting a very poor alignment with several gaps. Further, an information content calculation based on posterior probabilities of nucleotides at each position of the resulting ancestral sequences revealed that on average the information content is ~0.3 (Min=0, Max=2), which is extremely low and inappropriate for ASR, thus ruling out the suitability of ASR approach to assess rewiring.

We observe many more cases of rewiring at the regulon level, as opposed to the gene level, while in reality one would expect the opposite. There are at least two possible reasons for this outcome. The first has to do with an extremely stringent control imposed by the rotation test in gene-level testing, as discussed further in the Methods section. The second potential reason is increased statistical power in regulon-level testing, i.e., even if the individual gene rewiring events are not strongly evidenced by loss/gain of TF sites as supported by all species (relative to a stringent rotation test), it is easier to detect them when they occur in multiple functionally related genes of a regulon (such as in RP genes). Further, these regulon-level events spanning rewiring at multiple gene loci are likely to have gone through a gradual switch in regulation across species. For instance, some RP genes in *Sklu, Cgla* and *Kwal* contain strong binding sites for both *RAP1* and *TBF1* (Tanay et al. 2005). These RP genes, therefore, are not detectable in our

gene-level analysis, despite retaining a rather strong signal for rewiring at the regulon-level (Fig. 2-3).

In the extreme case, rewiring posits that in any species all genes of a regulon will be regulated by exactly one of the two TFs in question. However, in reality, a gradual evolutionary transition in the regulation of a regulon's member genes is expected. Such transitionary stage is characterized by an ancestral species where the regulon genes are bound by both TFs without a clear winner. Moreover, such transitionary stage may be maintained in some of the extant species. To assess the extent of such transitionary species, for each regulon detected to have undergone rewiring, we estimated the fraction of species (out of 23) that display an intermediate level of rewiring. For a rewiring event involving TFs *X* and *Y*, we defined a species to be transitionary if the fraction of gene promoters (in the particular regulon) more likely to be bound by *X* than by *Y* is between 0.4 and 0.6, i.e., not extreme. We found that, on average across all detected events, ~8 species (of 23) can be considered transitionary, i.e., with regulons potentially regulated by both TFs.

In summary, our probabilistic approach, while recapitulating the well-established cases, implicates specific regulators involved in suspected cases of rewiring, for which the implied regulators are not known. A genome-wide unbiased screen suggests that evolutionary *cis*-regulatory rewiring is relatively frequent and may be a significant mechanism of introducing regulatory innovations and adaptations to changing environments. The detected rewiring events are well-supported by regulator-target species-specific co-expression. Rewiring TF pairs tend to

function in similar biological processes, are generally controlled by a common upstream regulator, and in general occupy lower levels in regulatory hierarchy.

Materials and Methods

Gene orthology groups, annotations and sequences

An orthogroup comprises of orthologs across a set of species. Gene orthogroup assignments for all predicted protein-coding genes across 23 Ascomycete fungal genomes were obtained from the Fungal Orthogroups Repository (Wapinski et al. 2007b) maintained by the Broad Institute (*broadinstitute.org/regev/orthogroups*). For our analysis, we only considered the 3844 orthogroups (Wapinski et al. 2007a) that had mappable orthologs across at least 14 or more species as a compromise between number of genes included and loss of power due to information across fewer species. The genome sequences and gene annotations were obtained from a variety of databases and studies summarized in “Data Sources” at the above link. Gene promoter sequences were defined as 600 bases upstream of ATG and truncated when neighboring ORFs overlapped with this region (also obtained from (Wapinski et al. 2007b)). All promoters of length < 50 were excluded. Mean and standard deviation of lengths of retained promoters were 472.5 bps and 164.2 bps.

Probabilistic rewiring score

We demonstrate here a toy example of the framework used in this analysis. The sample tree in Figure 2-1A shows four species (*s1,s2,s3,s4*) partitioned at a select internal branch *b* to produce the equivalent of 2 species in the left clade

(*s1, s2 ϵ S*) and 2 species (*s3, s4 ϵ T*) in the right clade. Gene locus *g* represents the orthologous group of genes across all the four species (*g1,g2,g3,g4 ϵ G*) that hypothetically exhibits differential usage of regulating transcription factors *X* and *Y,* where *X* is used by species in the left clade and *Y* is used by species in the right clade, and not vice-versa.

The function that tests if transcription factor *Y* is predominantly used by genes in *T*, but was replaced by transcription factor *X* in the *S* in a lineage specific manner is as follows:

$$RS(X,Y,g,b) = \frac{log(P(X,g1,s1)) + log(P(X,g2,s2)) + log(1-P(Y,g1,s1)) + log(1-P(Y,g2,s2))}{2} \\ (+) \\ \frac{log(P(Y,g3,s3)) + log(P(Y,g4,s4)) + log(1-P(X,g3,s3)) + log(1-P(X,g4,s4))}{2}$$

where terms of the form P(TF, gene, species) represent the computed probability of a TF binding to gene's promoter in the species (see below). The denominators in the RHS of both equations represent the size of the left and right clades respectively.

Generalizing the same, we get,

$$RS(X,Y,g,b) = \frac{1}{|L_b|}\sum_{s\in L_b} log(P(X,g,s)) + log(1-P(Y,g,s)) \\ + \frac{1}{|R_b|}\sum_{s\in R_b} log(P(Y,g,s)) + log(1-P(X,g,s))$$

where L_b and R_b denote the sizes of the left and right clades resulting from a partition at branch *b*, respectively.

PWM based TF binding probability

A list of 176 positional weight matrices (PWMs) for *S. cerevisiae* TFs was obtained from TRANSFAC (Matys et al. 2006). A single PWM may map to one or more TFs, and vice-versa. To compute the probability of a TF binding to a gene's promoter in a given species, i.e., P(TF,gene.species), we scan the gene's promoter using PWMSCAN (Levy and Hannenhalli 2002) which provides a p-value for each putative site based on a species specific background of sequence composition. We note the lowest p-value obtained in the promoter and transform that into a promoter-wide probability score based on a previously used approach (Chen et al. 2007) as:

$P(TF, gene, species) = (1 - pval)^{(L-w+1)}$ where L is the length of the promoter, w is the length of the motif.

In (rare) cases where an orthogroup included multiple genes (paralogs) for the same species, we used the average binding probability for all such genes to obtain a species-specific binding probability for the orthogroup. Additionally, for orthogroups missing a gene in a given species, we imputed the value of P(TF,gene.species) by averaging the binding probabilities of all sibling species with detectable orthologs. This essentially has the effect of deriving binding potential from related species, when it cannot be directly estimated by binding scores, thereby providing a suitable proxy.

Species tree and selected lineages to assess rewiring

The species tree showing the relationships between 23 related Ascomycota fungi in Fig 2-1B (obtained from (Wapinski et al. 2007a)) was used to determine the

lineages (partitions) in the phylogeny. The 6 branches were chosen (highlighted in bold in Fig. 2-1B) such that the resulting partitions reflect some clade-specific differences in the biology of these species, viz., sensu-stricto vs non sensu-stricto (branch #7), pre-WGD vs. post-WGD (branch #11), mostly pathogenic vs. mostly non-pathogenic species (branch #20), etc.

Expression data

Expression profiles of *Scer* comprised of data for 6206 genes across 1011 conditions, and *Calb* comprised of data for 6167 genes across 198 conditions (Ihmels et al. 2005). Tab-delimited text files containing the log2 ratios are obtained from *weizmann.ac.il/home/barkai/Rewiring*.

Regulon discovery

We used expression data in *Scer* and *Calb* to identify conserved regulons – a set of genes with similar function that are coordinately expressed both within and across these two species. These two species have diverged sometime between 160 and 800 million years, representing a long evolutionary time. Starting from 1982 manually curated functionally related groups of genes (Field et al. 2009), we generated coexpression networks for each group in *Scer* and for its mappable orthologous genes in *Calb*. The nodes in the network are the component genes and edge weights between them are $|\rho|$, where ρ represents the Spearman correlation between the expression vectors of those genes. The individual networks in the two species were then merged; such that each merged network consist of nodes representing conserved orthologs and the edge weights are the product of the corresponding edge weights in the *Scer* and *Calb* networks. This

unit of edge weight is a proxy for a combined measure of distance based on conserved co-expression (i.e., lesser the distance between nodes, the more likely they are to have conserved co-expression in both species).

Next, each network was subjected to unsupervised clustering to isolate dense subgraphs that are representative of regulons, as per the above definition. We used MCL, a Markov Cluster Algorithm (van Dongen and Abreu-Goodger 2012) to identify these subnetworks using a setting of medium granularity in resolving clusters ('-I 2' option). Since these algorithms are not robust to large graphs with too many edges (despite using edge weights), we removed those edges with low combined measures of co-expression (<= 0.06). This cutoff provides a reasonably good proxy that preserves edges reflecting high correlation, while cutting out the noise significantly (see Supplemental Fig. A-14A). The application of MCL resulted in several subgraphs of functionally related genes with high co-expression (regulons). We excluded regulons that were larger than 100 genes, as well those with 6 or fewer genes (except the galactose regulon), thus identifying 1713 regulons. While some overlap of genes across regulons of different functional processes is expected, it is relatively small (mean Jaccard index = 0.003) to be of any concern (see Supplemental Fig. A-14B).

Note: For cases where a given species possessed multiple genes belonging to the same orthologous group, the expression profiles of the member genes were averaged before computing pairwise correlations with other ortholog groups within that species.

Generation of species-specific TF-TF networks

We used PWMSCAN (Levy and Hannenhalli 2002) to scan the promoter sequences of TF-encoding genes in *Scer* and *Calb*. For all hits detected with a motif-match score of 0.95 (using a species-specific background of nucleotide composition), we assigned a directional edge between the corresponding TF pairs. Using the *igraph* package (Csárdi and Nepusz 2006) in R, we generated a species-specific network using data from the above, which was then used to compute shortest paths between pairs of TFs in a species-specific fashion.

Phylogeny-preserving rotation test to assess significance of rewiring score at gene-level

The overall aim of the rotation test is to enable sampling of related variables from a null distribution such that inherent co-variance structure, i.e., the relationships between variables (TF-binding probability profile across species in our case) is preserved (Langsrud 2005). We first derive the species-by-species (23x23) co-variance matrix $\sum$ based on concatenated binding probabilities for all TFs at all gene loci in each species (estimating co-variance matrix for each TF separately does not influence the overall results). Next, for each TF, we obtain the 23-dimensional vector μ of TF-specific mean binding probabilities of all orthogroups in each of the 23 species. Finally, for a TF, given TF-specific vector μ and the general co-variance matrix $\sum$, we randomly sample from a multivariate normal distribution ($x \sim N(\mu, \sum)$), which is analogous to sampling from the matrix of the TF's binding probabilities in all 3844 orthogroups across all 23 species, while preserving the co-variance structure. We considered this a stringent control as

the co-variance matrix directly captures the relationship of TF binding probabilities across species (as required by the rotation test).

Upon generating these rotated TF binding probabilities for the same number of synthetic orthologous loci (and inverting these distributions back to the probability scale of (0,1)), we applied the rewiring score function to generate a background distribution of rewiring scores. This enabled computation of an FDR value for every observed rewiring score, as summarized for different thresholds in Supplemental Figure A-15.

Although in principle, co-variance matrix derived from the known phylogeny of the 23 species should be similar to the one based on all TF binding probabilities, we found that a phylogeny inferred from TF binding probabilities differs from the known species phylogeny (see Supplemental Fig. A-16). This suggests that evolution of TF binding probability does not strictly follow the neutral expectation. Thus, by directly controlling for overall TF binding probability relationships, our criteria for detecting gene-level rewiring should be considered highly stringent.

Nucleosome occupancy data

Genome-wide nucleosome occupancy data for 12 species, viz., *K. waltii*, *S. bayanus*, *S. cerevisiae*, *Y. lipolytica*, *D. hansenii*, *C. albicans*, *C. glabrata*, *S. castellii*, *S. paradoxus*, *K. lactis*, *S. mikatae* and *S. kluyveri* was obtained from GSE22211 (Tsankov et al. 2010); and for *S. pombe* from GSE28839 (Tsankov et al. 2011).

CHAPTER 3: Distal CpG islands can serve as alternative promoters to transcribe genes with silenced proximal-promoters

Abstract

DNA methylation at the promoter of a gene is presumed to render it silent, yet, a sizable fraction of genes with methylated proximal-promoters exhibit elevated expression. Here, we show, through extensive analysis of the methylome and transcriptome in 34 tissues, that in many such cases, transcription is initiated by a distal upstream CpG island (CGI) located several kilobases away that functions as an alternative promoter. Specifically, such genes are expressed precisely when the neighboring CGI remains unmethylated, but remain silenced otherwise. Based on CAGE and PolII localization data, we found strong evidence of transcription initiation at the upstream CGI and a lack thereof at the methylated proximal promoter itself. Consistent with their alternative promoter activity, CGI-initiated transcripts are associated with signals of stable elongation and splicing that extend into the gene body, as evidenced by tissue-specific RNA-seq and other DNA-encoded splice signals. Furthermore, based on both inter and intra-species analyses, such CGIs were found to be under greater purifying selection relative to CGIs upstream of silenced genes. Overall, our study describes a hitherto unreported conserved mechanism of transcription of genes with methylated proximal promoters in a tissue-specific fashion. Importantly, this phenomenon explains the aberrant expression patterns of some cancer driver genes, potentially due to aberrant hypomethylation of distal CGIs, despite

methylation at proximal promoters.

Introduction

In mammalian DNA, cytosines within CpG dinucleotides are heavily methylated throughout the genome, yet there are several discrete “islands” that contain a high frequency of unmethylated CpG sites. These are called CpG islands (CGI), and their identification has long been considered important in the annotation of functional landmarks within the genome. Historically, CGIs served as landing strips to locate annotated genes (Larsen et al. 1992), and it was for good reason as it was later discovered that 55-60% of all genes contain CGIs at their annotated promoters. While about half of all CGIs in the genome coincide with gene promoters, the remaining half are either intragenic or intergenic and were termed “orphan CGIs” due to their remote location that suggested the uncertainty over their biological significance (Deaton and Bird 2011).

Does there exist evidence to support the idea that orphan CGIs are involved in gene regulation? Indeed, several specific examples, showing promoter activity at orphan CGIs were uncovered in the context of critical functions like imprinting and development (Deaton et al. 2011). For example, a CGI in intron 10 of the imprinted *Kcnq1* gene (Mancini-DiNardo et al. 2003) promotes the initiation of a noncoding transcript (*Kcnq1ot1*) required for the imprinting of several genes at this locus. Tissue-specific alternative promoter activity was detected at another orphan CGI that promotes a specific isoform of the *Rapgef4* gene (Hoivik et al. 2013). Cumulative evidence suggest that most, perhaps all, CGIs have promoter-

like characteristics and are sites of transcription initiation (Illingworth et al. 2010). Additionally, most of the conserved methylation differences between tissues occurred at orphan CGIs (Illingworth and Bird 2009), suggesting that they are tightly regulated. A recent study that derived CGI annotations from experimental methylation data (eCGIs) also showed that promoter-distal eCGIs exhibited the most tissue-specific methylation patterns, and were linked to the tissue-specific production of alternative transcripts (Mendizabal and Yi 2016). In fact, studies profiling CpG methylation patterns have identified differentially methylated regions (DMRs) even in the shores of CpG islands (Pollard et al. 2009). These regions of lower CpG density in close proximity (up to 2 kb) to CGIs, whose differential methylation patterns are strongly related to gene expression, are highly conserved and have distinct tissue-specific methylation patterns (Irizarry et al. 2009a). Thus, over time, CG-dense genomic loci (viz., CGIs and their shores) have been realized to be increasingly important in many functional contexts, whose immense regulatory potential outside of annotated promoters is only beginning to be understood.

Typically, methylation at a gene's promoter renders it silent (Han et al. 2011) by modifying DNA accessibility to the transcriptional machinery (Suzuki and Bird 2008), or by recruiting factors that aid in generating a refractory chromatin conformation unsuitable for transcription (Kouzarides 2007). While several prior studies (Suzuki and Bird 2008; van Eijk et al. 2012; Deaton and Bird 2011; Smith et al. 2012; Sproul et al. 2011) have observed strongly negative correlations between promoter methylation and gene expression, others report more nuanced

relationships between the two (Wagner et al. 2014; Wan et al. 2015; Shilpa et al. 2014; Martino and Saffery 2015), including a lack thereof. Additionally, there are several instances of genes in cancer cells, wherein abnormal expression is persistent despite widespread promoter hypermethylation (Guillaumet-Adkins et al. 2014; Moarii et al. 2015; Van Vlodrop et al. 2011). These collectively indicate that additional factors controlling expression of genes with methylated promoters have not been identified. Furthermore, due to the long standing interest in CGI-promoter genes, most of this knowledge is based on analysis of CGI-promoters (Jones 2012), and the details of the role of methylation in controlling non-CGI transcription start sites (TSSs) have largely been overlooked.

This lack of consensus prompted us to explore the global transcriptional landscape of methylated-promoter genes. We found that substantial numbers of methylated-promoter genes (~1500 in each of 34 tissues) are expressed at high levels; such promoters are predominantly non-CGI, which is consistent with prevailing knowledge on the rarity of methylation at CGI promoters (Illingworth et al. 2010; Brandeis et al. 1994; Lienert et al. 2011). While the expression of many such genes can be attributed to the use of alternate gene body promoters, as has been shown in some normal and cancer cells (Nagarajan et al. 2014; Maunakea et al. 2010), we estimate that the high levels of expression realized by almost 50% of all methylated-promoter genes remain completely unexplained (see Results).

Here we show, through detailed analyses across 34 primary tissues and cell types, that genes with methylated and silenced promoters can in some instances

utilize an upstream, hitherto unknown, CpG island as an alternative promoter to express their gene product. Our results strongly support this previously unreported general regulatory mechanism, which may play a role in promoting aberrant transcription of driver genes in cancer cells.

Results

Highly expressed genes with methylated promoters

We obtained RNA-seq expression and whole genome bisulfite sequencing (WGBS) methylation data for 30 primary tissues and 4 cell lines from the Roadmap Epigenomics Project (Bernstein et al. 2010) and other sources (Lay et al. 2015; Ziller et al. 2013; Djebali et al. 2012; Menafra et al. 2014) (Table 3-1). Henceforth, we will refer to these 34 samples simply as ‘tissue types’. In a given tissue type, there exists, on average about 9000 genes whose primary promoters are maintained in a heavily methylated state (see Methods). Although methylation at a gene’s promoter is expected to render it silent, we observed that ~1500 of such genes exhibited high levels of expression. We then excluded genes whose expression could be explained by alternative gene body promoter activity (see Methods), and this resulted in 700 genes in each tissue whose expression remains unexplained. To specifically assess the involvement of the closest upstream CGI in the expression of these genes, we restricted downstream analysis to only those genes that did not have another gene annotated (including non-coding RNAs) in the genomic region between the gene’s transcription start site and the CGI. This eliminates potential biases owing

to intervening transcriptional activity. We further verified that this subset of genes was not enriched for any specific biological function or expression status compared to the set of all methylated-promoter genes (Supplemental Fig. B-1). These filters resulted in a set of ~3200 methylated-promoter genes (down from ~9000 overall) out of which ~440 (down from ~1500) are highly expressed per tissue. The number of genes at various filtering stages across tissues are provided in Table 3-2. Additionally, methylated-promoter gene names and their methylation status across tissues are listed in Supplemental Table B-1.

Group	Tissue ID	Description	Data Source	Link
ESC	E003	H1 Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
ESC	E016	HUES64 Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
ESC	E024	ES-UCSF4 (4STAR) Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
ES-derived	E004	H1 BMP4 Derived Mesendoderm Cultured Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
ES-derived	E005	H1 BMP4 Derived Trophoblast Cultured Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
ES-derived	E006	H1 Derived Mesenchymal Stem Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
ES-derived	E007	H1 Derived Neuronal Progenitor Cultured Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
ES-derived	E011	hESC Derived CD184+ Endoderm Cultured Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
ES-derived	E012	hESC Derived CD56+ Ectoderm Cultured Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
ES-derived	E013	hESC Derived CD56+ Mesoderm Cultured Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Neurosphere	E053	Cortex derived primary cultured neurospheres	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Neurosphere	E054	Ganglion Eminence derived primary cultured neurospheres	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Thymus	E112	Thymus	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Epithelial	E058	Foreskin Keratinocyte Primary Cells	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Brain	E070	Brain Germinal Matrix	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Brain	E071	Brain Hippocampus Middle	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Muscle	E100	Psoas Muscle	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Heart	E065	Aorta	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Heart	E095	Left Ventricle	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Heart	E104	Right Atrium	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Heart	E105	Right Ventricle	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Digestive	E079	Esophagus	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Digestive	E094	Gastric	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Digestive	E106	Sigmoid Colon	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Digestive	E109	Small Intestine	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Other	E066	Liver	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Other	E096	Lung	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Other	E097	Ovary	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Other	E098	Pancreas	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
Other	E113	Spleen	Roadmap	http://egg2.wustl.edu/roadmap/web_portal/
ENCODE	HepG2	HepG2 Hepatocellular Carcinoma Cell Line	Roadmap, GSE46644	http://egg2.wustl.edu/roadmap/web_portal/ http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46644
ENCODE	IMR90	NHLF Lung Fibroblast Primary Cells	Roadmap, GSE46644	http://egg2.wustl.edu/roadmap/web_portal/ http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46644
ENCODE	MCF7	NHEK-Epidermal Keratinocyte Primary Cells	GSE54693	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54693
ENCODE	K562	K562 Leukemia Cells	Roadmap, GSE64929	http://egg2.wustl.edu/roadmap/web_portal/ http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64929

Table 3-1. Metadata associated with the 34 tissue types used in our analyses.

	PRE-FILTER			POST-FILTER		
Tissue	Promoter-methylated genes	Highly expressed promoter-methylated genes	Highly expressed promoter-methylated genes (unexplained)	Promoter-methylated genes	Highly expressed promoter-methylated genes	Highly expressed promoter-methylated genes (unexplained)
E003	10335	1493	653	3651	419	180
E004	10155	1454	624	3598	413	148
E005	10221	1645	746	3611	474	196
E006	9783	1637	736	3496	454	125
E007	10313	1436	629	3640	421	177
E011	10575	1452	660	3732	424	215
E012	10686	1491	694	3785	445	234
E013	10392	1583	695	3676	436	217
E016	10334	1382	583	3651	403	183
E024	10035	1387	610	3545	414	219
E053	9325	1500	668	3266	420	173
E054	9290	1484	660	3283	422	196
E058	8381	1312	560	2999	361	109
E065	9024	1795	748	3246	554	259
E066	9168	1947	752	3298	557	270
E070	9187	1369	596	3242	382	188
E071	9526	1635	788	3397	455	265
E079	9147	1557	694	3279	433	270
E094	8877	1527	675	3182	425	247
E095	9333	1603	773	3346	445	218
E096	9326	1766	810	3308	483	274
E097	8602	1399	610	3064	388	192
E098	8327	1544	746	2951	434	284
E100	8729	1535	704	3115	407	135
E104	9043	1530	725	3247	428	225
E105	9156	1608	746	3283	440	212
E106	9312	1731	811	3319	486	320
E109	9349	1713	805	3338	492	327
E112	9703	2063	793	3445	559	311
E113	9303	1808	788	3323	493	288
HepG2	4089	1369	586	1330	350	150
K562	2312	1094	558	647	262	109
MCF7	8874	1699	866	3227	493	331
IMR90	6947	1588	704	2415	422	262

Table 3-2. The number of genes before (blue) and after applying a filtration step (green) that discards all loci that contain a neighboring gene spanning the region between them and their associated upstream CGIs in all 34 tissue types. Columns B and F correspond to all promoter-methylated genes. Columns C and G correspond to those that are highly expressed (greater than 50th percentile) in B and F respectively. Columns D and H correspond to those whose expression cannot be explained by alternative gene-body promoter activity in C and G respectively.

In all 34 tissues, we find that a vast majority (~90%) of these genes do not contain CpG islands in their promoters, which is significant enrichment relative to a 30% expectation of non-CGI promoter genes genome-wide (Saxonov et al. 2006). This result agrees with prevailing knowledge on the rarity of methylated CGIs at the promoters of annotated genes (which is only ~ 3% overall (Illingworth et al. 2010)), as well as the lower propensity of CGI promoters to be *de novo* methylated compared to non-CGI promoters (Brandeis et al. 1994; Lienert et al. 2011). Further, they are enriched for cell-type specific functions based on a

quantitative index of tissue-specificity (TSI (Yanai et al. 2005); see Methods) as well as Gene Ontology (GO) enrichment. Supplemental Fig. B-2 shows that the median TSI of expressed methylated-promoter genes is significantly greater compared to that of all genes (10^{-4} < Wilcoxon signed-rank test P < 0.05 in 33/34 tissues showing significant trend). We also present overrepresented functional terms based on GO enrichment in each tissue in Supplemental Fig. B-3. These findings are in line with existing knowledge that a majority of widely expressed genes use CpG island promoters, while most tissue-specific genes have neither CpG islands nor TATA-boxes (Zhu et al. 2008; Larsen et al. 1992) in their promoters.

Association of distal CGI with methylated-promoter gene expression

Across the set of all methylated-promoter genes in a given tissue, we asked if the methylation status of the closest upstream CGI was informative of its expression. Specifically, we categorized these genes into two sets, (i) expressed (MethExp) and, (ii) not expressed (MethNotExp) genes (see Methods), and compared the proportion of methylated distal CGIs in each case. As shown in Fig. 3-1A, we observe a strong negative relationship between CGI methylation and the corresponding gene's expression (1.25 < Odds ratio < 1.75, 10^{-10} < Fisher's exact test P < 0.01 in 26/34 tissue types showing significant trend). We further found that CGIs associated with MethExp genes tend to have significantly lower methylation than those associated with MethNotExp genes (Fig. 3-1B; 10^{-13} < Wilcoxon P < 0.02 in 32/34 tissues showing significant trend). Therefore, we conclude that expression levels of methylated-promoter genes are strongly

associated with the epigenetic status of the distal upstream CGIs.

On average, CGIs upstream of MethExp genes are located at a distance of 10 kb, and interestingly are several fold closer to their associated genes than those upstream of MethNotExp genes (Fig. 3-1C; 10^{-13} < Wilcoxon P < 10^{-4} in 33/34 tissues showing significant trend). While such proximity might not be a pre-requisite for intergenic CpG islands to act as alternative promoters to transcribe genes with silenced primary promoters, it does seem likely that it would be a preferred configuration.

Finally, the CGIs associated with MethExp genes are evolutionarily much more conserved than those associated with MethNotExp genes, both between species (using phastCons scores (Siepel et al. 2005) based on an alignment of 46 vertebrates; Fig. 3-1D, 10^{-4} < Wilcoxon P < 0.05 in 27/34 tissues) and within species (using average derived allele frequencies (DAF) across humans, see Methods; Fig. 3-1E, 10^{-7} < Wilcoxon P < 0.05 in 20/34 tissues). Additionally, from annotations of syntenic blocks between human and 8 related vertebrate species (see Methods), we assessed the extent to which shared synteny between a methylated-promoter gene and its upstream CGI was informed by the expression status of the gene, using a logistic regression framework that controlled for the genomic distance between them. We found that MethExp genes and their upstream CGIs are more often in the same syntenic block than CGIs upstream of all other genes (Fig. 3-1F; 10^{-3} > p-value attached to co-efficient of expression status > 10^{-6} in all 6/8 comparisons). Interestingly, this holds true only in the 6 mammalian species and not in any of the 2 non-mammalian vertebrates, which

suggests that MethExp associated CGIs were only recently co-opted, close to the base of mammalian divergence, to function as alternative promoters. Thus, higher purifying selection acting specifically on MethExp-CGIs that are also in synteny with their associated genes is indicative of their functional role, in this case, as a regulatory element (promoter) facilitating transcription of the downstream gene.

We further hypothesized that the tissue-specific usage of CGIs as alternative promoters may be regulated by cell type-specific transcription factors (TFs). To test this, for every CGI showing evidence of alternative promoter activity in some cell type, we identified the high confidence TF binding sites (see Methods) in those CGIs, and tested if TFs corresponding to these sites show a preference to be expressed in cell types where the CGI was active versus not. Consistent with expectations, a large fraction of these CGIs (~40% vs. a 5% random null expectation; Fisher's $P < 10^{-16}$) do show patterns of cell type-specific regulation.

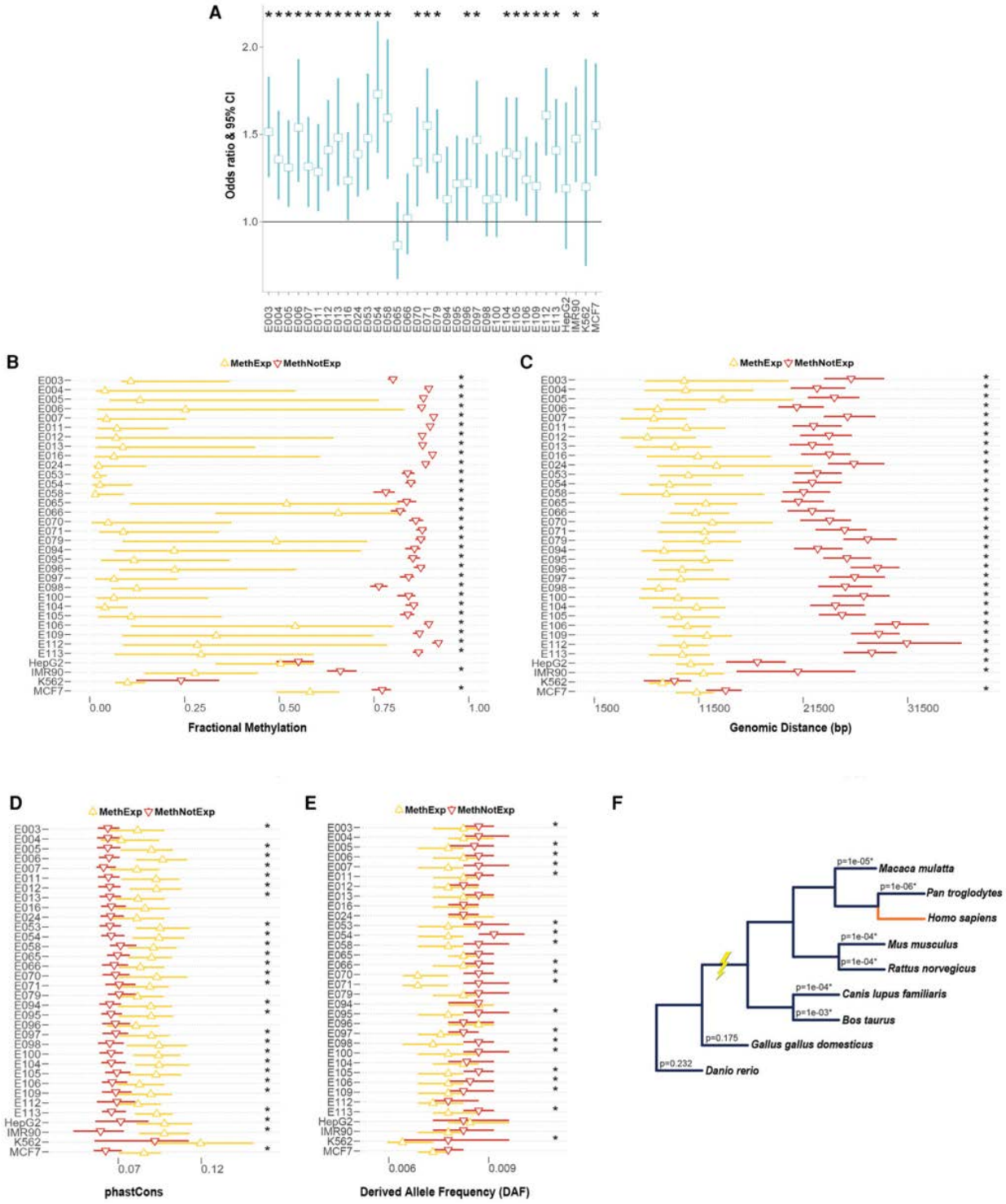


Figure 3-1. ***Association of distal CGI with the expression of MethExp genes*****.** (A) Odds ratio and 95% confidence interval (CI) (Y-axis) of the proportion of unmethylated CGIs upstream to MethExp genes versus MethNotExp genes in 34 tissue types (X-axis). A depletion of methylation at CGIs upstream of MethExp genes corresponds to a higher odds ratio. (B)-(E) compare various properties for the CGIs upstream to MethExp genes (yellow) versus MethNotExp (red) genes,

viz., (B) fractional methylation level, (C) genomic distance to gene, (D) phastCons scores, and (E) derived allele frequencies (DAF). The median and the 95% CI (X-axes) are shown for 34 tissue types (Y-axes). (F) Phylogenetic tree of the 8 vertebrate species used in determining the extent of shared synteny with human amongst methylated-promoter genes. The association between CGI-gene synteny and whether the gene is MethExp or MethNotExp was assessed via regression, while controlling for genomic distance between CGI and the gene. The significance of the association (p-value) is shown on each branch corresponding to the species used to estimate synteny with respect to human. Statistically significant associations ($P < 0.05$) are annotated with an asterisk in all plots.

Transcription initiation occurs at distal CGI, and not the promoter of MethExp genes

Our previous observation of lower methylation and increased conservation at CGIs upstream of MethExp genes is only suggestive of their potential to function as alternative promoters to transcribe them. Here, we explicitly test for transcriptional initiation at these CGIs using two different experimental measures. First, we used single molecule Cap Analysis of Gene Expression (CAGE) data from the FANTOM Consortium (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014), available for 15 tissue types. The CAGE assay produces a snapshot of the 5′ end of the messenger RNA population in a biological sample, which provides a direct quantitative measure of initiation rate at a given locus. Thus, in a tissue-specific fashion, we quantified the transcription initiation signal (number of CAGE tags) at the promoters as well as the associated CGIs of three groups of genes, (i) MethExp, (ii) MethNotExp, as well as (iii) expressed genes with methylation-free promoters (NotMethExp). This third group serves as a baseline for the amount of initiation expected at similarly expressed gene loci. Since expression level is related to the intensity of initiation signal, we ensured by sampling, that the expression level distribution of the selected MethExp and

NotMethExp genes were comparable. Fig. 3-2A and B show the distribution of CAGE levels at the promoters as well as the associated CGIs of MethExp, MethNotExp and NotMethExp genes pooled across all tissues, respectively. In the case of promoters, we observe that the number of CAGE tags is quite low at MethExp genes, and importantly, is several fold lesser than that at similarly expressed NotMethExp genes (Wilcoxon $P = 10^{-6}$). Further, the complete lack of CAGE tags at MethNotExp genes is consistent with the fact that these genes aren't expressed at all. Next, we contrast the transcription initiation signal at upstream CGIs associated with the three gene groups. It is known that most, perhaps all, CGIs are sites of transcription initiation, and it is owing to this property that about 50% of them are adapted for promoter function and coincide with the TSS of annotated genes (Deaton and Bird 2011). Consistent with this expectation, CGIs from all gene groups show substantial CAGE tag levels. Considering that CGIs associated with MethNotExp genes do not contribute to the expression of those genes, the CAGE level at these CGIs may serve as a baseline expectation for orphan CGIs. Then, interestingly, we observe that the CAGE at CGIs associated with MethExp genes is significantly greater than this baseline (Wilcoxon $P = 10^{-4}$). In fact, MethExp-associated CGIs collectively exhibit somewhat greater transcriptional activity (CAGE) than even the NotMethExp-associated CGIs (Wilcoxon $P = 0.04$). Low coverage of CAGE tags at orphan CGIs limits our ability to statistically substantiate comparisons between groups at the per-tissue resolution, but the promoter and CGI CAGE trends we observe across gene groups in the above analyses are consistent in 15/15 and

12/15 tissues respectively, (see Supplemental Fig. B-4) and therefore does not affect our conclusion.

The second measure we used for the quantification of initiation corresponds to a signal associated with serine-5-phosphorylated RNA Polymerase II (PolII-Ser5). Specifically, the initiating form of Pol-II is phosphorylated at Ser5, and as elongation of the mRNA molecule occurs, the enzyme gradually loses Ser5-P, and gains Ser2-P (Phatnani and Greenleaf 2006; Jonkers and Lis 2015). To this end, we used Ser5-P Pol-II chromatin immunoprecipation sequencing (ChIP-seq) data in MCF-7 cell line (based on data availability) to quantify transcriptional initiation at the promoters (Fig. 3-2C) and upstream CGIs (Fig. 3-2D) of the three gene groups. The trends are highly consistent with those obtained from CAGE. Specifically, initiation signal at the promoters of MethExp genes is much lower than NotMethExp genes (Wilcoxon $P < 10^{-5}$) that are expressed at comparable levels. Also, consistently, CGIs at MethExp have higher Ser5-P signals than MethNotExp genes (Wilcoxon $P < 10^{-5}$) as well as NotMethExp (albeit not yielding statistical significance; Wilcoxon $P = 0.15$), suggesting that specifically for MethExp genes, the upstream CGI may serve as an alternative, hitherto undetected, promoter.

In light of the above observations, it is also possible that distal CGIs associated with MethExp genes are in fact their true primary promoters that were misannotated, likely due to the narrow expression breadth (i.e., tissue specificity) of MethExp genes. To distinguish between the alternative scenarios of promoter misannotation and unsuspected context-specific distal promoter usage, we

carried out two specific analyses. First, we performed a locus-specific cross-tissue comparison of CAGE tags at the annotated promoters of MethExp genes when they are categorized as MethExp, MethNotExp and NotMethExp across different tissues (Fig. 3-2E). If our observations were simply due to misannotation, then specifically for these select group of genes whose promoters become methylated in some tissue, we expect to see a ubiquitous lack of transcription initiation at their annotated promoters across all other tissues, regardless of their methylation status. Instead, we find that the CAGE tags at the annotated promoters of these select genes when they are unmethylated is very high (NotMethExp >> MethExp or MethNotExp; Wilcoxon $P < 10^{-3}$ in both cases), supporting the idea that MethExp-associated CGIs serve only as alternative promoters, and not the primary ones. Further, a similar locus-specific cross-tissue comparison of CAGE tags at distal CGIs (Fig. 3-2F) showed that CGIs have higher CAGE tags in tissues where their associated genes are MethExp compared with tissues where they are MethNotExp (Wilcoxon $P = 0.008$). However, MethExp and NotMethExp groups do not show a significant difference in CAGE levels at the distal CGI, suggesting that transcriptional activity at distal CGIs in these instances is generally unlinked with promoter activity. In addition, we also observe that the promoter methylation levels are starkly different when these loci are active versus silent across tissues (Supplemental Fig. B-5).

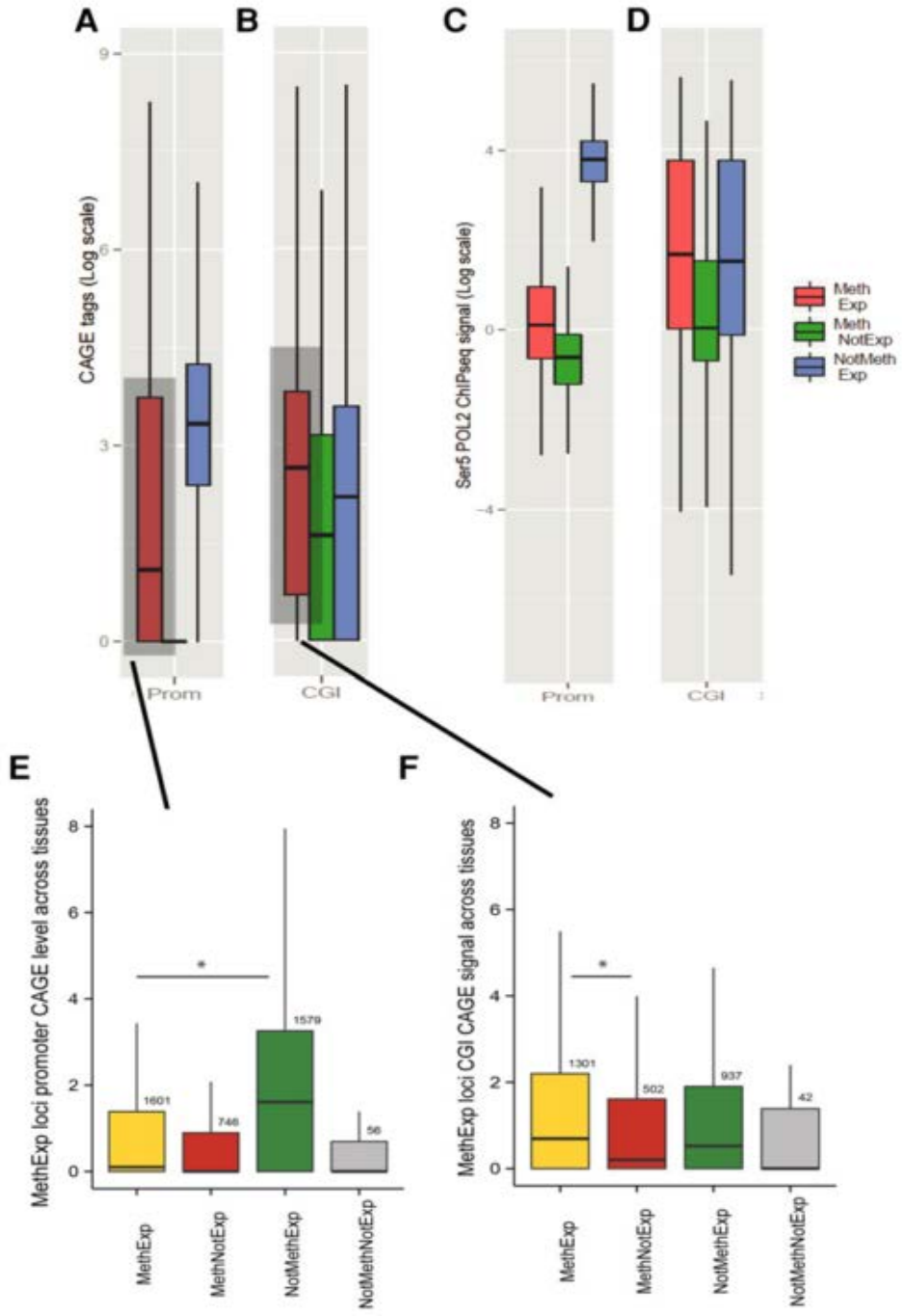


Figure 3-2. ***Transcription initiation occurs at an upstream alternative CGI promoter, and not at the proximal promoters of MethExp genes*****.** The evidence of transcriptional initiation based on CAGE tag intensity (log-transformed Y-axis) is contrasted for three gene groups, MethExp (pink), MethNotExp (green) and NotMethExp (blue) at (A) the proximal-promoter, and (B) the distal CGI (X-axis). (C) and (D) are analogous to (A) and (B), respectively, for observed levels of transcription initiation based on Ser5-PolII ChIP-seq intensity. For the pan-tissue pooled set of MethExp genes, the plot shows the CAGE signal (Y-axis) at the (E) promoter and (F) the associated distal CGIs of these genes when they are MethExp (yellow), MethNotExp (red),

NotMethExp (green) and NotMethNotExp (grey) in other tissues (X-axis).

Next, we assessed the effect of loss of methylation on the relative activity of the annotated promoter of MethExp genes. This analysis is however limited by the availability of data in human. We therefore analysed MethExp genes in mouse embryonic stem cells with (WT; wild type cells) and without DNA methyl transferase activity (DNMT TKO; DNMT triple knockout cells). Using RNA-seq and WGBS methylation data in WT cells, we identified all high-confidence (about 103) MethExp genes using the same protocol as that for tissue types in human. After verifying that they exhibit the same broad features of CGI alternative promoter use as the MethExp genes in human tissue types (Supplemental Fig. B-6), we analyzed their promoter usage patterns in DNMT TKO cells. We hypothesized that if the distal CGI was the only promoter of these genes, then removing methylation at the annotated promoters should not lead to a change in their activity status. Unlike CAGE, RNA-seq does not allow for direct quantification of transcription initiation at these annotated promoters. Therefore, in DNMT TKO cells, we contrasted the mean read density observed upstream of the annotated TSS (TSS-200 bp) to that observed downstream of it (TSS+200 bp), relative to the same in WT, for every gene identified as MethExp in WT. We see that 68 out of 103 MethExp genes show an increase in mean read density (normalized by the corresponding densities in WT) downstream of the annotated TSS (Fisher's P = 0.02), hinting at a potential switch from usage of distal CGIs to the annotated promoters in these cases. Note that in the absence of data in mouse knockouts that directly quantifies initiation rates, we cannot conclusively

ascertain that the increased numbers of reads downstream of the TSS in DNMT TKO cells are from transcripts originating at the annotated TSS; this result therefore must be considered with caution. However, taken together, we conclude that our overall observations are not simply a reflection of erroneous promoter annotation.

As an additional layer of evidence for transcriptional activity, we quantified the repressive histone modifications (H3K9me3 and H3K27me3) at the promoters of MethExp, MethNotExp and NotMethExp genes (Supplemental Fig. B-7). Consistent with other observed features of active transcription, we find that both of these marks are significantly higher at MethExp then NotMethExp promoters (H3K9me3: 10^{-69} < Wicoxon P < 10^{-5} in 22/32 tissues, H3K27me3: 10^{-113} < Wilcoxon P < 10^{-3} in 19/32 tissues). Further, given that distal CGIs can display transcriptional activity similar to promoters, it is likely that they also harbor histone modifications reflective of their activity status. To this end, we contrasted the ChIP-seq signal of two active (H3K4me3, H3K9ac) and two repressive (H3K27me3, H3K9me3) histone modifications at CGIs associated with MethExp, MethNotExp and NotMethExp genes (Supplemental Fig. B-8). In addition, we also compared the DNase hypersensitivity signal to assess the extent of chromatin accessibility (also reflective of transcriptional activity) at these CGIs. The tests for histone marks were performed for different numbers of tissues as per data availability. Broadly, we observe that active marks are significantly greater in MethExp-CGIs compared to both MethNotExp- (DNase: 10^{-14} < Wilcoxon P < 10^{-4} in 12/12 tissues, H3K4me3: 10^{-16} < Wilcoxon P < 10^{-4} in 32/32

tissues, H3K9ac: 10^{-17} < Wilcoxon P < 10^{-5} in 10/10 tissues) and NotMethExp-CGIs (DNase: 10^{-3} < Wilcoxon P < 0.05 in 8/12 tissues, H3K4me3: 10^{-3} < Wilcoxon P < 0.05 in 23/32 tissues, H3K9ac: 10^{-6} < Wilcoxon P < 0.05 in 8/10 tissues). In the case of repressive marks, while we broadly observe that they are significantly lower in MethExp-CGIs than MethNotExp- (H3K27me3: 10^{-9} < Wilcoxon P < 0.05 in 18/32 tissues, H3K9me3: 10^{-8} < Wilcoxon P < 0.05 in 30/32 tissues) and NotMethExp-CGIs (H3K27me3: 10^{-3} < Wilcoxon P < 0.05 in 18/32 tissues, H3K9me3: 10^{-3} < Wilcoxon P < 0.05 in 21/32 tissues), the differences in these levels are not as pronounced or widespread across tissues as for the active marks. This is consistent with the idea that active repression of an orphan CGI when the locus is not acting as an alternative promoter probably occurs less often, and that these in general tend to prevail as open, accessible actively transcribing entities across the genome.

Evidence of transcriptional elongation and splicing occurring between distal CGIs and their associated MethExp gene promoters.

The emerging trend of strong transcription initiation at distal CGIs that are associated with the expression of downstream MethExp genes, further accompanied by a total lack of transcription initiation at proximal promoters motivated us to probe further for evidence of bona fide promoter action at the CGIs, which we describe in four complimentary analyses that follow. It is known that any transcriptional activity ensuing from intergenic regulatory elements (i.e., true orphan CGIs and enhancers) that are not immediately proximal to coding or non-coding RNA genes does not culminate in the production of long RNA

molecules (Andersson et al. 2014; De Santa et al. 2010). But if indeed MethExp-associated CGIs function as promoters, they are expected to exhibit sustained transcriptional activity and elongation along the entire stretch of the intervening genomic region between the CGI and the downstream gene (henceforth referred to as the "segment"). While the presence of coding or non-coding genic elements in the segment region can bias quantitative measures of elongation, as mentioned earlier, this complication was pre-empted by excluding any such genes from our analyses (see Methods).

To this effect, first, we binned the segment region corresponding to MethExp, MethNotExp and NotMethExp genes into 10 equal-sized bins, and quantified in each bin, three parameters that inform the extent of transcriptional activity, as well as elongation: 1) RNA-seq signal strength (RPKM), 2) RNA-seq signal coverage (fraction of nucleotides supported by a read), and 3) H3K36me3 ChIP-seq signal (histone mark associated with the gene bodies of actively transcribed genes (Hon et al. 2009)). As can be seen in Fig. 3-3A, 3-3B and 3-3C, the segment region corresponding to MethExp genes show significantly greater evidence for transcriptional activity and elongation, than those for MethNotExp and NotMethExp genes (10^{-102} < Wilcoxon P < 10^{-5}) in all tissues (with the exception of H3K36me3 wherein 31/33 and 29/33 tissues show significant trends respectively). We present in the main text only the pooled distribution across all segment bins for each of the above, but the trends remain qualitatively similar in each assessed bin (Supplemental Fig. B-9). These results are consistent with our prediction that CGIs associated with MethExp genes have a greater tendency to

produce long RNA molecules extending into the body of the downstream gene.

As PolII-Ser2 is also a marker of transcriptional elongation, we analyzed the PolII-Ser2 signal in the segment region of MethExp, MethNotExp and NotMethExp genes in MCF-7 cells (Supplemental Fig. B-10). While the effect size of the trend (greater elongation in MethExp-segment regions compared with other groups) using PolII-Ser2 is not as strong as when using RNA-seq and H3K36me3 data, MethExp-segments do show significantly greater elongation signals compared to MethNotExp (Wilcoxon $P < 10^{-3}$).

Second, paired-end (PE) RNA-seq reads whose pairs are split across the segment and the downstream annotated gene region would provide a more direct indication for transcription initiating at upstream CGIs and extending into the body of MethExp genes. Such reads are not expected in the case of MethNotExp and NotMethExp genes, because in both cases, transcription initiates at the annotated primary promoter of these genes, and not their associated upstream CGI. As expected, the proposed evidence is much greater for MethExp genes relative to the other two classes (Fig. 3-3D; Wilcoxon $P < 10^{-4}$ in both cases).

Third, it has been shown that transcripts that initiate from intergenic regulatory elements as well as those that remain unspliced terminate prematurely, and are rapidly cleared away from the cell due to their instability in the absence of splice signals (Almada et al. 2013; Ntini et al. 2013). Previous studies have showed that sequence motifs dictate the production of stable vs. unstable transcripts; presence of a splice donor site facilitates binding of splicing factor U1 which can

suppress polyadenylation site (PAS) dependent termination, thereby promoting elongation of mRNAs (recently shown to be true in all transcript classes (Schwalb et al. 2016)). Core et al. (Core et al. 2014) used this in a Hidden Markov Model (HMM) and showed that U1 binding sites strongly tend to precede PAS on stable transcripts, but not on unstable transcripts. Thus we directly probed the order of occurrence of the above motifs in the sequence of the segment region to inform the stability of transcripts originating from the upstream CGIs associated with all genes. Each gene was deemed “stable” or “unstable” based on the order of the two motifs from the 5’ end of the segment. We then compared the fraction of stable transcripts between genes that are MethExp in at least one tissue to the rest of the genes using Fisher’s exact test. We found that the fraction of “stable” transcripts is significantly greater amongst MethExp genes than amongst other genes (Fig. 3-3E; Fisher’s $P = 10^{-5}$), however the effect size is modest (Odds ratio=1.2).

Finally, while it is not unrealistic for the region intervening the distal CGI and the gene TSS to possess long 5’ UTR (**U**n-**T**ranslated **R**egions; which in eukaryotes can be upto several kb long (Lodish 2008)), it is more likely that it is spliced out in the mature transcript. Therefore, we directly assessed splicing activity by assembling transcripts *de novo* from RNA-seq reads using STAR (Dobin et al. 2013) and mapping splice junctions (see Methods). From the mapped junctions in each of 28 tissues (limited by raw read data availability), we quantified the number of MethExp, MethNotExp and NotMethExp genes that showed evidence of a splice junction connecting their associated CGIs to their coding region

(henceforth called a ‘split junction’). We found strong support for enrichment of split junctions in MethExp genes compared to both MethNotExp (8 < Odds Ratio < 340; 0 < Fisher’s P < 10^{-67}) and NotMethExp (2 < Odds Ratio < 23; 10^{-5} < Fisher’s P < 10^{-40}) genes in all 28 tissues (Fig. 3-3F). We illustrate one such example of *GIGYF1*, which is a MethExp gene in the Esophagus tissue (Fig. 3-4). Alternative promoter activity of its associated upstream unmethylated CGI is apparent in this case, where ensuing transcripts that contain a long intron spanning the segment region is spliced out.

Taken together, these results strongly suggest that CGIs associated with MethExp genes are bona fide promoters that produce transcripts that are stably elongated and spliced into the annotated genes.

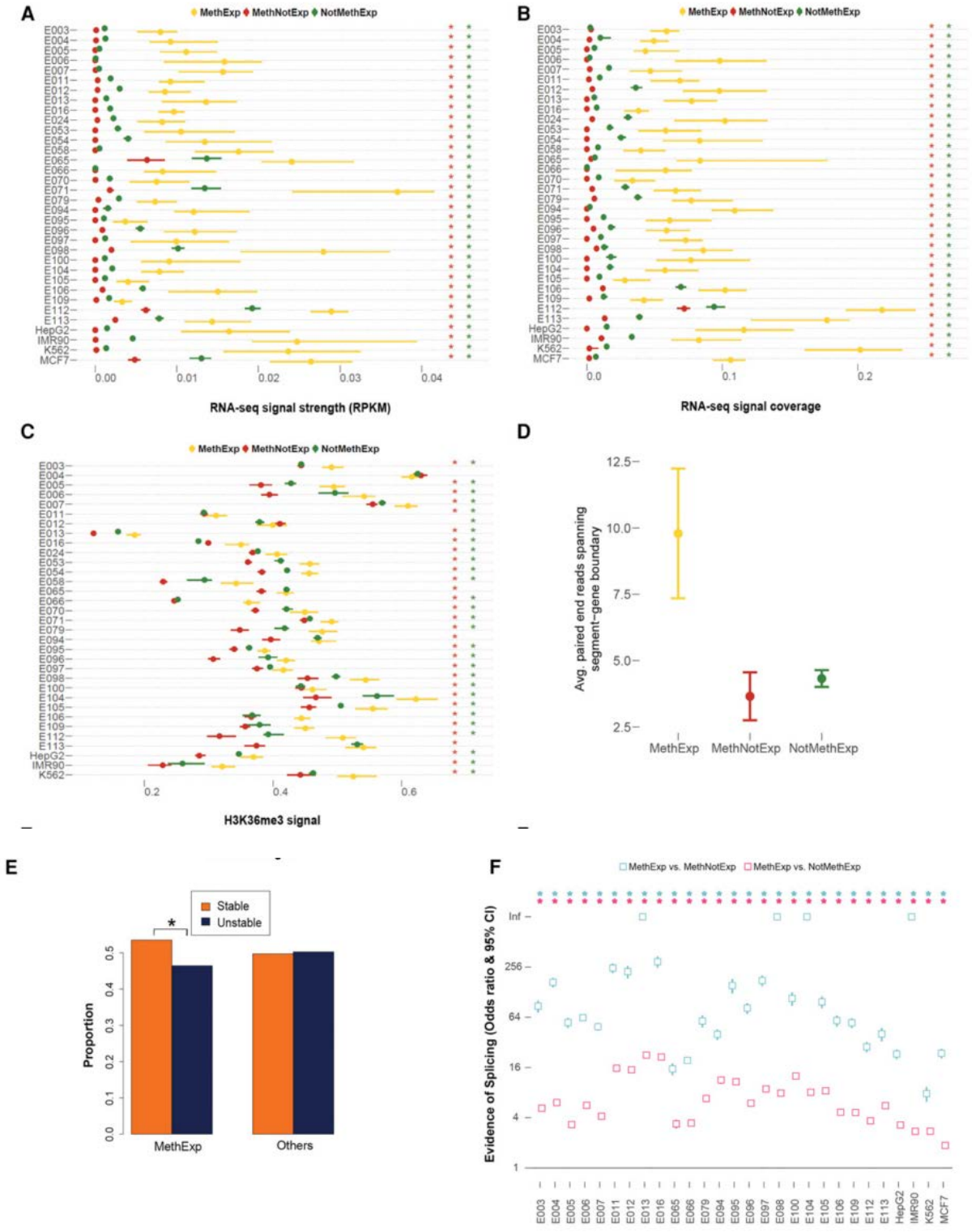


Figure 3-3. ***Transcriptional elongation and splicing signals between CGI and the gene.*** (A) Median RNA-seq RPKM signal, (B) median RNA-seq coverage, and (C) median H3K36me3 ChIP-seq signal and 95% CI (X-axes) associated with the segment region of MethExp (yellow),

MethNotExp (red) and NotMethExp (green) genes across 34 tissue types (Y-axes). (D) The average number of paired-end RNA-seq reads (Y-axis) whose ends lie in both the segment region as well as the annotated gene, as seen across MethExp (yellow), MethNotExp (red) and NotMethExp (green) genes across all tissue types (X-axis). (E) The proportion of "stable" (orange) and "unstable" (dark-blue) transcripts (Y-axis), as determined from a sequence based predictor (U1-PAS motif order in segment region) in those genes that are MethExp in at least one tissue, versus other genes (Fisher's P = 10^{-5}). (F) Odds ratio and 95% CI (Y-axis) of the proportion of (i) MethExp versus MethNotExp genes, (cyan) and (ii) MethExp versus NotMethExp genes (pink) that show evidence of splice junctions between the segment region and annotated gene based on *de novo* transcript assembly across 24 tissue types (X-axis). An enrichment of such splice junctions in the segment region associated with MethExp genes corresponds to a higher odds ratio. Statistically significant associations (P < 0.05) are annotated with an asterisk in all plots, in a matched color scheme, wherever appropriate. In panels (A) through (C), this color corresponds to the color of the background gene group that MethExp genes are contrasted against, i.e., a red asterisk to represent significant difference between MethExp and MethNotExp, and green for that against the NotMethExp group.

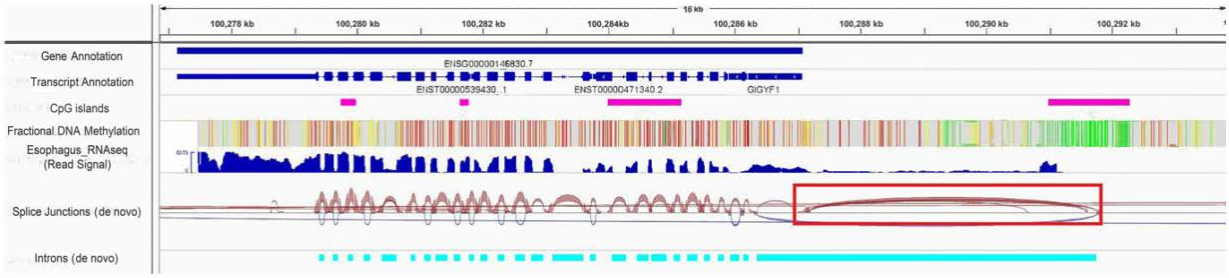


Figure 3-4. An illustrative example. Transcriptomic and epigenetic marks surrounding a gene (GIGFY1) that is expressed despite a hypermethylated promoter in the Esophagus tissue. As shown, the proximal promoter is highly methylated (red corresponds to high methylation, whereas green to low), and yet there is a large signal for expression as can be seen in the RNA-seq signal track. An upstream CGI (in pink) > 6 kb away is free of methylation, and transcription of the gene ensues at this locus extending into the body of the gene. These patterns suggest that (1) the longest transcript starts at the CGI as opposed to its annotated start site in Ensembl/GENCODE, (2) the first intron spans the segment region and extends into the body of the gene (in cyan), and (3) there is a large splice junction (loop in dark red) that is split between a region located inside the segment and an exon inside the annotated gene.

Aberrant gene expression in cancer linked to hypomethylated distal CGIs.

The aberrant DNA methylation landscape associated with cancer cells is considered to be a hallmark of the disease. Cancer is characterized by both global hypomethylation, as well as widespread promoter-associated hypermethylation of important genes like tumor suppressors (Jones and Baylin

2007), that lead to their silencing. We aimed to investigate the extent to which the usage of upstream CGIs as alternative promoters explains the aberrant gene expression patterns observed in cancer phenotypes.

We obtained RNA-seq and Illumina methylation array (450K) data from The Cancer Genomic Atlas (TCGA) for 780 breast cancer (Koboldt et al. 2012) and 315 renal cell carcinoma (Creighton et al. 2013) patients. Due to low coverage of 450K methylation data, only ~3030 genes could be used for which methylation at both upstream CGI and promoter were available. Overall, there exist ~300 (~10%) methylated-promoter genes in each cancer sample that are expressed at high levels (see Methods). Similar to normal cells, these are also mainly protein-coding genes (85%) with mostly non-CGI promoters (90%). Given that hypermethylation in cancer is mainly targeted to CGI-promoters of genes (Sproul et al. 2012), one might expect to see a greater fraction of CGI-promoter genes in the MethExp group, but this was not the case. This strongly supports the idea that methylation at CGI-promoters is almost always accompanied by systematic silencing of that gene locus.

We find that CGIs associated with MethExp genes in cancer cells exhibit very similar properties to those found in normal tissues. Relative to MethNotExp-associated CGIs, MethExp-associated CGIs (i) have significantly lower methylation, (ii) are closer to their associated gene loci, and (iii) show significantly higher transcriptional activity and elongation (based on RNA-seq RPKM and read coverage measures) signals in the segment region. Fig. 3-5 shows these trends for 100 representative breast and kidney cancer samples. As

seen, the two gene groups are significantly different in all the above aspects (10^{-60} < Wilcoxon P < 10^{-8} across all comparisons). Thus, the use of distal CGIs by non-CGI methylated-promoter genes as alternative promoters is a general phenomenon, observed in both normal and cancer cells.

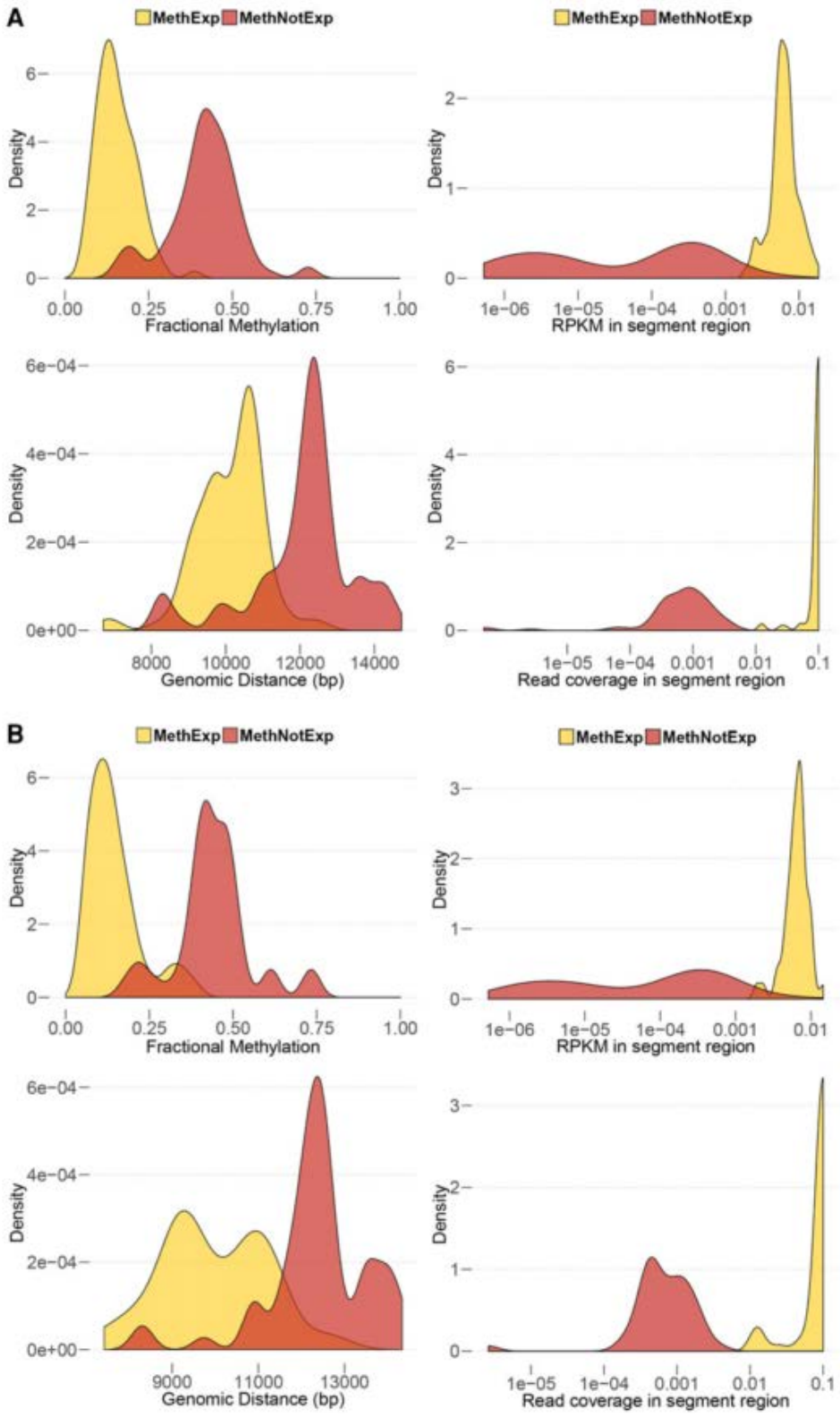


Figure 3-5. ***Use of distal CGI as alternative promoter by MethExp genes in cancer.*** The figure shows four lines of evidence supporting the usage of distal CGI as alternative promoter by

MethExp genes in contrast to MethNotExp genes in (A) breast, and (B) kidney cancer. Each panel shows the distribution of the median (i) fractional methylation at upstream CGIs (top-left), (ii) genomic distance between distal CGI and gene (bottom-left), (iii) RNA-seq RPKM signal (top-right), and (iv) RNA-seq coverage (bottom-right) at the segment region (Y-axes) corresponding to MethExp (yellow) vs. MethNotExp genes (red) across 100 representative samples (X-axes).

Next, we mapped the functional landscape of MethExp genes in cancer. First, we focused specifically on those sets of genes whose promoters are hypermethylated in cancer, and that potentially rely on a distal CGI to express themselves. To this end, in a given cancer type, we identified genes whose promoters are hypermethylated in cancer (in >75% of the samples), and whose expression is associated with CGI methylation (Spearman's ranked correlation $P < 0.05$) but not with proximal promoter's methylation (Spearman's $P > 0.05$). This resulted in 34 genes in breast, and 39 genes in kidney cancer (listed in Table 3-3). GO terms associated with general phenotypes in cancer like cell growth, maintenance or adhesion ($P < 0.05$) are overrepresented in both cases. More interestingly, we found an enrichment of protein sequence features and domains (FDR < 0.05) associated with (i) *EGF*, *EGF*-like and palmitate among genes identified in breast (involved in breast cancer drug resistance (Masuda et al. 2012; Liu et al. 2008); Fig. 3-6A), and (ii) calcium binding, *FOX* transcription factor family and alpha-actinins among genes identified in kidney cancer (key genes/pathways involved in decreased kidney function and cancer (Feng et al. 2015; Linehan et al. 2010); Fig. 3-6B). This suggests that key genes involved in cancer also bypass their inactive promoters and utilize distal CGIs for their expression.

GeneID
ENSG00000075073
ENSG00000081248
ENSG00000101074
ENSG00000102104
ENSG00000111245
ENSG00000121898
ENSG00000128011
ENSG00000130182
ENSG00000131142
ENSG00000132692
ENSG00000143006
ENSG00000143194
ENSG00000147394
ENSG00000148204
ENSG00000157322
ENSG00000159650
ENSG00000162571
ENSG00000162592
ENSG00000167914
ENSG00000168065
ENSG00000170423
ENSG00000178722
ENSG00000180438
ENSG00000181408
ENSG00000183914
ENSG00000184599
ENSG00000185972
ENSG00000187569
ENSG00000188716
ENSG00000188937
ENSG00000205864
ENSG00000206026
ENSG00000215045
ENSG00000215131

Table 3-3. List of hypermethylated-promoter genes in breast cancer that use distal CGIs as alternative promoters.

Next, we focused specifically on genes that are differentially expressed in cancer relative to their normal counterparts, potentially due to differential methylation of their upstream CGI (and not the primary promoter). To this end, we obtained matched normal samples for 80 of the breast cancer samples (normal and cancer tissue from the same individuals), and identified 208 genes that are differentially expressed between cancer and normal samples (Wilcoxon $P < 0.05$) and whose non-zero expression is associated with CGI methylation (Spearman's $P < 0.05$) but not with proximal promoter's methylation (Spearman's $P > 0.05$). These genes are enriched (FDR < 0.05) for GO terms related to cell cycle, cell growth (tyrosine and MAPK kinase signaling) and cell-cell adhesion, which are implicated in cancer progression and metastasis (Fig. 3-6C). Many of these genes exhibit very high negative correlation between upstream CGI methylation status and gene expression across healthy and cancer samples (up to Spearman's rho = -0.45). These include the *YES1* Yamaguchi sarcoma viral oncogene (src tyrosine kinases family) whose paralog *LYN* is involved in mediating treatment resistance in breast cancer (Schwarz et al. 2014) and the *GINS2* gene whose protein product interacts with *CHEK2*, a tumor suppressor gene linked to many cancers including breast (Rantala et al. 2010). An entire list of these genes is provided in Table 3-4.

In summary, this previously unreported phenomenon whereby distal CGIs are utilized as alternative promoters by certain highly expressed genes with methylated proximal promoters is prevalent across several clinically important genes in cancer, and warrants further investigation to chart its full implications.

ID	Gene Name
ENSG00000173157	ADAM metallopeptidase with thrombospondin type 1 motif, 20
ENSG00000144746	ADP-ribosylation-like factor 6 interacting protein 5
ENSG00000133612	ArfGAP with GTPase domain, ankyrin repeat and PH domain 3
ENSG00000172530	BTG3 associated nuclear protein
ENSG00000122507	Bardet-Biedl syndrome 9
ENSG00000157322	C-type lectin domain family 18, member A
ENSG00000221869	CCAAT/enhancer binding protein (C/EBP), delta
ENSG00000141030	COP9 constitutive photomorphogenic homolog subunit 3 (Arabidopsis)
ENSG00000132153	DEAH (Asp-Glu-Ala-His) box polypeptide 30
ENSG00000103423	DnaJ (Hsp40) homolog, subfamily A, member 3
ENSG00000102034	E74-like factor 4 (ets domain transcription factor)
ENSG00000133216	EPH receptor B2
ENSG00000132591	Era G-protein-like 1 (E. coli)
ENSG00000187741	Fanconi anemia, complementation group A
ENSG00000125812	GDNF-inducible zinc finger protein 1
ENSG00000131153	GINS complex subunit 2 (Psf2 homolog)
ENSG00000214367	HAUS augmin-like complex, subunit 3
ENSG00000166189	Hermansky-Pudlak syndrome 6
ENSG00000123485	Holliday junction recognition protein
ENSG00000198589	LPS-responsive vesicle trafficking, beach and anchor containing
ENSG00000196199	M-phase phosphoprotein 8
ENSG00000164077	MON1 homolog A (yeast)
ENSG00000189043	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4, 9kDa
ENSG00000083642	PDS5, regulator of cohesion maintenance, homolog B (S. cerevisiae)
ENSG00000106536	POU class 6 homeobox 2
ENSG00000128563	PRKR interacting protein 1 (IL11 inducible)
ENSG00000174788	Purkinje cell protein 2
ENSG00000101074	R3H domain containing-like
ENSG00000011454	RAB GTPase activating protein 1
ENSG00000157869	RAB28, member RAS oncogene family
ENSG00000167771	REST corepressor 2
ENSG00000163785	RYK receptor-like tyrosine kinase
ENSG00000140750	Rho GTPase activating protein 17
ENSG00000140386	S-phase cyclin A-associated protein in the ER
ENSG00000143499	SET and MYND domain containing 2
ENSG00000163788	SNF related kinase
ENSG00000115107	STEAP family member 3
ENSG00000172845	Sp3 transcription factor
ENSG00000132405	TBC1 domain family, member 14
ENSG00000160606	TLC domain containing 1
ENSG00000162604	TM2 domain containing 1
ENSG00000128564	VGF nerve growth factor inducible
ENSG00000164087	WD repeat domain 51A
ENSG00000175877	Williams-Beuren syndrome chromosome region 28
ENSG00000166896	XRCC6 binding protein 1
ENSG00000100997	abhydrolase domain containing 12
ENSG00000075624	actin, beta
ENSG00000077522	actinin, alpha 2

ENSG00000065000	adaptor-related protein complex 3, delta 1 subunit
ENSG00000122359	annexin A11
ENSG00000198250	anthrax toxin receptor-like
ENSG00000101200	arginine vasopressin
ENSG00000118690	armadillo repeat containing 2
ENSG00000174808	betacellulin
ENSG00000185963	bicaudal D homolog 2 (Drosophila)
ENSG00000105552	branched chain aminotransferase 2, mitochondrial
ENSG00000132692	brevican
ENSG00000149654	cadherin-like 22
ENSG00000157445	calcium channel, voltage-dependent, alpha 2/delta subunit 3
ENSG00000185972	calicin
ENSG00000121898	carboxypeptidase X (M14 family), member 2
ENSG00000141527	caspase recruitment domain family, member 14
ENSG00000164287	cell division cycle 20 homolog B (S. cerevisiae)
ENSG00000140743	cerebellar degeneration-related protein 2, 62kDa
ENSG00000102805	ceroid-lipofuscinosis, neuronal 5
ENSG00000131142	chemokine (C-C motif) ligand 25
ENSG00000168539	cholinergic receptor, muscarinic 1
ENSG00000173728	chromosome 1 open reading frame 100
ENSG00000171067	chromosome 11 open reading frame 24
ENSG00000089916	chromosome 14 open reading frame 118
ENSG00000119669	chromosome 14 open reading frame 4
ENSG00000206026	chromosome 18 open reading frame 62
ENSG00000123144	chromosome 19 open reading frame 43
ENSG00000170279	chromosome 7 open reading frame 33
ENSG00000106603	chromosome 7 open reading frame 44
ENSG00000107020	chromosome 9 open reading frame 46
ENSG00000162592	coiled-coil domain containing 27
ENSG00000175602	coiled-coil domain containing 85B
ENSG00000172346	cold shock domain containing C2, RNA binding
ENSG00000163499	crystallin, beta A2
ENSG00000100243	cytochrome b5 reductase 3
ENSG00000109016	dehydrogenase/reductase (SDR family) member 7B
ENSG00000187569	developmental pluripotency associated 3
ENSG00000149927	double C2-like domains, alpha
ENSG00000133059	dual serine/threonine and tyrosine protein kinase
ENSG00000162999	dual specificity phosphatase 19
ENSG00000119661	dynein, axonemal, light chain 1
ENSG00000197102	dynein, cytoplasmic 1, heavy chain 1
ENSG00000135960	ectodysplasin A receptor
ENSG00000106462	enhancer of zeste homolog 2 (Drosophila)
ENSG00000127884	enoyl Coenzyme A hydratase, short chain, 1, mitochondrial
ENSG00000079819	erythrocyte membrane protein band 4.1-like 2
ENSG00000148730	eukaryotic translation initiation factor 4E binding protein 2
ENSG00000178896	exosome component 4
ENSG00000189057	family with sequence similarity 111, member B
ENSG00000185442	family with sequence similarity 174, member B
ENSG00000184599	family with sequence similarity 19 (chemokine (C-C motif)-like), member A3
ENSG00000128573	forkhead box P2
ENSG00000163251	frizzled homolog 5 (Drosophila)
ENSG00000115042	fumarylacetoacetate hydrolase domain containing 2A

ENSG00000142252	gem (nuclear organelle) associated protein 7
ENSG00000182771	glutamate receptor, ionotropic, delta 1
ENSG00000006007	glycerophosphodiester phosphodiesterase 1
ENSG00000066455	golgi autoantigen, golgin subfamily a, 5
ENSG00000169813	heterogeneous nuclear ribonucleoprotein F
ENSG00000182611	histone cluster 1, H2aj
ENSG00000068024	histone deacetylase 4
ENSG00000122592	homeobox A7
ENSG00000145681	hyaluronan and proteoglycan link protein 1
ENSG00000130956	hyaluronan binding protein 4
ENSG00000169047	insulin receptor substrate 1
ENSG00000027697	interferon gamma receptor 1
ENSG00000102753	karyopherin alpha 3 (importin alpha 4)
ENSG00000205864	keratin associated protein 5-6
ENSG00000050555	laminin, gamma 3
ENSG00000131409	leucine rich repeat containing 4B
ENSG00000143194	maelstrom homolog (Drosophila)
ENSG00000103150	malonyl-CoA decarboxylase
ENSG00000136146	mediator complex subunit 4
ENSG00000076706	melanoma cell adhesion molecule
ENSG00000140406	mesoderm development candidate 1
ENSG00000134046	methyl-CpG binding domain protein 2
ENSG00000120333	mitochondrial ribosomal protein S14
ENSG00000181610	mitochondrial ribosomal protein S23
ENSG00000102738	mitochondrial ribosomal protein S31
ENSG00000116353	mitochondrial trans-2-enoyl-CoA reductase
ENSG00000135341	mitogen-activated protein kinase kinase kinase 7
ENSG00000107968	mitogen-activated protein kinase kinase kinase 8
ENSG00000111245	myosin, light chain 2, regulatory, cardiac, slow
ENSG00000139505	myotubularin related protein 6
ENSG00000171532	neurogenic differentiation 2
ENSG00000088970	non-protein coding RNA 153
ENSG00000125450	nucleoporin 85kDa
ENSG00000166579	nudE nuclear distribution gene E homolog (A. nidulans)-like 1
ENSG00000183828	nudix (nucleoside diphosphate linked moiety X)-type motif 14
ENSG00000188937	nyctalopin
ENSG00000121390	paraspeckle component 1; paraspeckle protein 1 pseudogene
ENSG00000100731	pecanex homolog (Drosophila)
ENSG00000108733	peroxisomal biogenesis factor 12
ENSG00000116120	phenylalanyl-tRNA synthetase, beta subunit
ENSG00000087157	phosphatidylglycerophosphate synthase 1
ENSG00000128655	phosphodiesterase 11A
ENSG00000186642	phosphodiesterase 2A, cGMP-stimulated
ENSG00000197943	phospholipase C, gamma 2 (phosphatidylinositol-specific)
ENSG00000115896	phospholipase C-like 1
ENSG00000068137	pleckstrin homology domain containing, family H (with MyTH4 domain)
ENSG00000156374	polycomb group ring finger 6
ENSG00000173889	polyhomeotic homolog 3 (Drosophila)
ENSG00000137054	polymerase (RNA) I polypeptide E, 53kDa
ENSG00000176407	potassium channel modulatory factor 1
ENSG00000172336	processing of precursor 7, ribonuclease P/MRP subunit (S. cerevisiae)
ENSG00000197170	proteasome (prosome, macropain) 26S subunit, non-ATPase, 12
ENSG00000196605	zinc finger protein 846
ENSG00000153786	zinc finger, DHHC-type containing 7
ENSG00000072121	zinc finger, FYVE domain containing 26
ENSG00000015171	zinc finger, MYND domain containing 11

ENSG00000165912	protein kinase C and casein kinase substrate in neurons 3
ENSG00000171132	protein kinase C, epsilon
ENSG00000105568	protein phosphatase 2 (formerly 2A), regulatory subunit A, alpha isoform
ENSG00000152894	protein tyrosine phosphatase, receptor type, K
ENSG00000204956	protocadherin gamma subfamily A, 1
ENSG00000146676	purine-rich element binding protein B
ENSG00000179889	pyridoxal-dependent decarboxylase domain containing 1
ENSG00000111445	replication factor C (activator 1) 5, 36.5kDa
ENSG00000067533	ribosomal RNA processing 15 homolog (S. cerevisiae)
ENSG00000112306	ribosomal protein S12; ribosomal protein S12 pseudogene 4; ribosomal protein S12 pseudogene 11; ribosomal protein S12 pseudogene 9
ENSG00000070423	ring finger protein 126
ENSG00000189051	ring finger protein 222
ENSG00000184178	sec1 family domain containing 2
ENSG00000112320	sine oculis binding protein homolog (Drosophila)
ENSG00000104969	small glutamine-rich tetratricopeptide repeat (TPR)-containing, alpha
ENSG00000163581	solute carrier family 2 (facilitated glucose transporter), member 2
ENSG00000168065	solute carrier family 22 (organic anion/urate transporter), member 11
ENSG00000013306	solute carrier family 25, member 39
ENSG00000101438	solute carrier family 32 (GABA vesicular transporter), member 1
ENSG00000065054	solute carrier family 9 (sodium/hydrogen exchanger), member 3 regulator 2
ENSG00000175898	sphingosine-1-phosphate receptor 2
ENSG00000136158	sprouty homolog 2 (Drosophila)
ENSG00000108055	structural maintenance of chromosomes 3
ENSG00000136478	testis expressed 2
ENSG00000143337	torsin A interacting protein 1
ENSG00000181585	transmembrane inner ear
ENSG00000204713	tripartite motif-containing 27
ENSG00000104804	tubby like protein 2
ENSG00000184811	tumor suppressor candidate 5
ENSG00000164828	unc-84 homolog A (C. elegans)
ENSG00000181408	urotensin 2 receptor
ENSG00000204103	v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian)
ENSG00000176105	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1
ENSG00000108828	vesicle amine transport protein 1 homolog (T. californica)
ENSG00000119614	visual system homeobox 2
ENSG00000186187	zinc and ring finger 1
ENSG00000100722	zinc finger CCCH-type containing 14
ENSG00000179627	zinc finger and BTB domain containing 42
ENSG00000130182	zinc finger and SCAN domain containing 10
ENSG00000147394	zinc finger protein 185 (LIM domain)
ENSG00000083844	zinc finger protein 264
ENSG00000083817	zinc finger protein 416
ENSG00000124459	zinc finger protein 45
ENSG00000103199	zinc finger protein 500
ENSG00000213015	zinc finger protein 580
ENSG00000142684	zinc finger protein 593
ENSG00000197483	zinc finger protein 628
ENSG00000185730	zinc finger protein 696
ENSG00000142409	zinc finger protein 787
ENSG00000197933	zinc finger protein 823

Table 3-4. A GO functional annotation of all methylated-promoter genes in breast cancer, that are differentially expressed from their normal counterparts across 80 matched samples, and display usage of distal CGI alternative promoters.

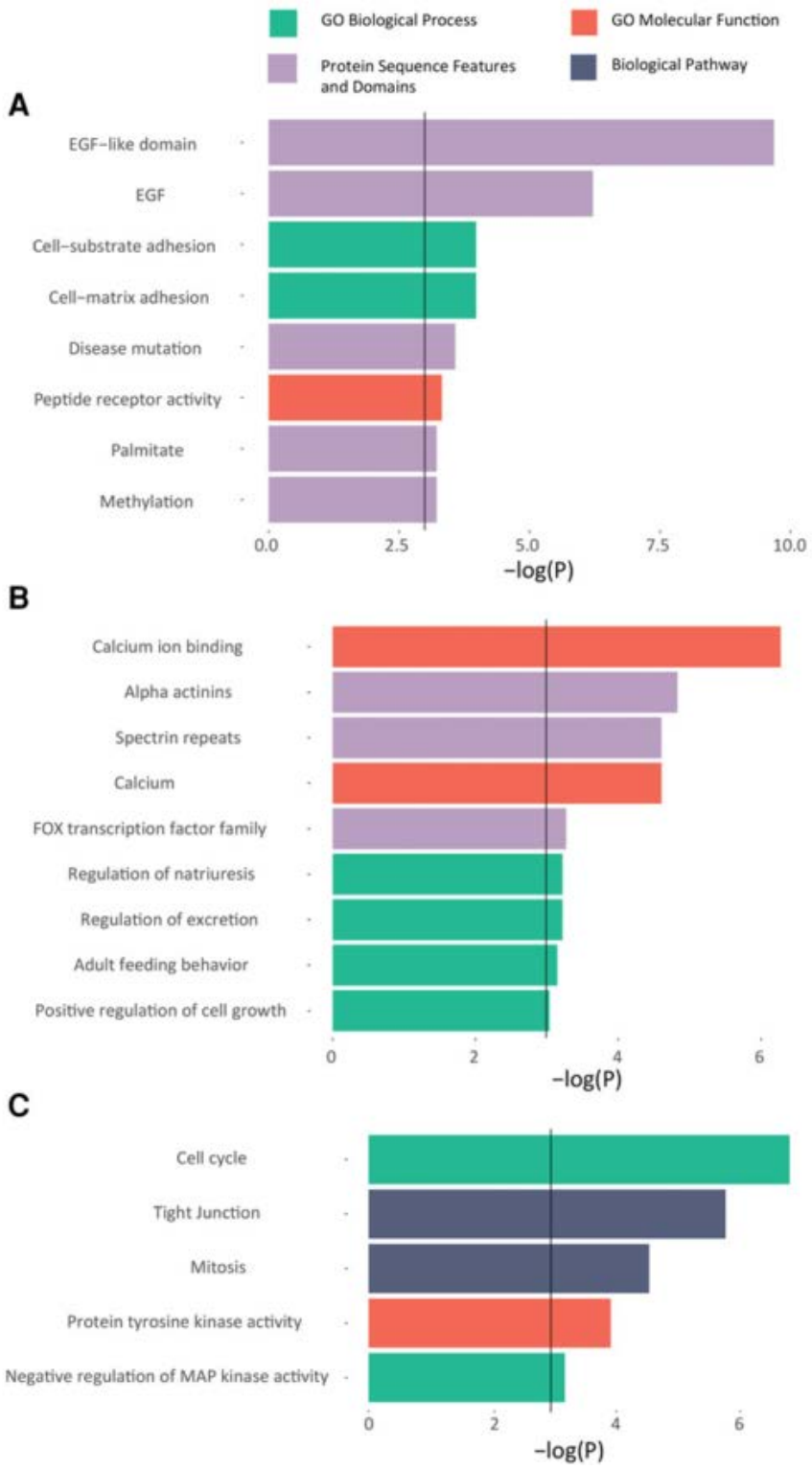


Figure 3-6. ***Functional enrichment of MethExp genes in cancer that potentially utilize an upstream CGI as promoter*****.** GO terms are shown on the Y-axis, along with their corresponding –log (adjusted P) significance measures on the X-axis. Solid black line at P = 0.05 represents the threshold for enrichment. (A) Functional enrichment for genes whose promoters are broadly

hypermethylated across samples, but whose expression across samples is correlated with the upstream CGI's methylation and not that of the proximal promoter, in breast, and (B) kidney cancer. (C) Functional enrichment in genes whose differential expression between normal and breast cancer samples is correlated with the methylation status of the upstream CGIs and not with that of the proximal promoter.

Discussion

CpG islands were first discovered in mouse DNA in the 80s, in seminal work by Adrian Bird and others (Bird et al. 1985). Their unusually high frequency of CG dinucleotides (which are primary targets of DNA methylation in vertebrates), their virtually free-of-methylation disposition (in an otherwise globally methylated genome), as well as the fact that they surround the control regions of most genes led to their quick recognition as important regulatory elements. As a consequence, many of the studies that followed focused mainly on promoter proximal CGIs, which incidentally, also happen to inform much of our understanding of the role of methylation in controlling chromatin structure and gene expression across tissues (Jones 2012). However, it was found that promoter-distal CGIs, despite being remote from annotated TSSs were also capable of transcription initiation (promoter function) (Maunakea et al. 2010), and some of these sites were implicated in transcribing alternative tissue-specific isoforms (Hoivik et al. 2013) as intragenic alternative promoters, or non-coding transcripts involved in imprinting and other functions (Mancini-DiNardo et al. 2003). Furthermore, it is the promoter-distal CGIs (orphan CGIs) that are more often differentially methylated, compared to promoter CGIs (Eckhardt et al. 2006), implicating them in condition-specific regulation. Although, despite these critical observations about orphan CGIs, a global view of their functional

significance is only just beginning to emerge.

Here we report, a previously unknown phenomenon, whereby an intergenic orphan CGI can function as an alternative promoter to express the gene product of a nearby CpG-poor methylated-promoter gene. We found this to occur across hundreds of CpG-poor promoter genes that become methylated in a tissue-specific fashion. In an effort to assess the prevalence of alternative promoter usage of CGI amongst the pool of MethExp genes, we quantified the broad features suggestive of alternative promoter usage, such as CGI methylation, CGI CAGE and RNAPolII-Ser5 (latter only in MCF-7), as well as segment RNA-seq signal and coverage, and computed the percentage of MethExp loci per tissue type that showed strong evidence of these based on stringent thresholds (see Table 3-5 for details). As shown, the fractions of loci with strong support for alterative promoter use are quite high for all of the above features across cell types. This suggests that the usage of upstream CGI as alternative promoters by genes with silenced proximal promoters is widespread.

Tissue	CGI Fractional Methylation	CGI CAGE tag level	PolII-Ser5 signal	Segment RNAseq signal strength	Segment RNAseq read coverage
E003	0.6	1	NA	0.7727	0.7727
E004	0.58	1	NA	0.75	0.77
E005	0.5405	1	NA	0.7094	0.7162
E006	0.5106	1	NA	0.7553	0.734
E007	0.6091	1	NA	0.7727	0.7545
E011	0.7561	1	NA	0.7804	0.7804
E012	0.5426	1	NA	0.7596	0.7596
E013	0.5702	1	NA	0.719	0.719
E016	0.5688	1	NA	0.7614	0.7431
E024	0.6	1	NA	0.8083	0.8
E053	0.6606	1	NA	0.8532	0.8532
E054	0.614	1	NA	0.8157	0.8157
E058	0.6875	1	NA	0.7875	0.7875
E065	0.4674	1	NA	0.6673	0.6728
E066	0.5517	1	NA	0.6812	0.6812
E070	0.5676	1	NA	0.7657	0.7657
E071	0.5714	1	NA	0.6904	0.6904
E079	0.4852	1	NA	0.7396	0.7337
E094	0.5276	1	NA	0.6398	0.6521
E095	0.5714	1	NA	0.6857	0.7
E096	0.538	1	NA	0.7173	0.7228
E097	0.614	1	NA	0.728	0.728
E098	0.4574	1	NA	0.7606	0.7712
E100	0.5941	1	NA	0.7722	0.7722
E104	0.6692	1	NA	0.7368	0.7218
E105	0.5887	1	NA	0.7021	0.6879
E106	0.7265	1	NA	0.6502	0.6547
E109	0.5045	1	NA	0.6818	0.6909
E112	0.5069	1	NA	0.6451	0.6221
E113	0.5102	1	NA	0.6836	0.7091
HepG2	0.5402	1	NA	0.6842	0.6821
K562	0.675	1	NA	0.768	0.765
IMR90	0.5069	1	NA	0.6637	0.662
MCF7	0.541	1	0.6782	0.6382	0.6541

Table 3-5. Fraction of tested MethExp loci per cell type that show strong evidence of alternative promoter usage based on stringent thresholds of: 1) CGI Fractional Methylation - to exclude CGIs that are heavily methylated and potentially silenced, we counted only those with methylation less than the mean methylation level of the intermediately methylated subpopulation identified from a 3-component mixture model fit on the total distribution. 2) CGI CAGE tag level - using the benchmark for CAGE at truly silenced regions, we counted only those with CAGE greater than median CAGE level observed at MethNotExp (silenced) promoters. 3) PolII-Ser5 signal - only in MCF7; using the benchmark for CAGE at truly silenced regions, we counted only those with Ser5 greater than median PolII-Ser5 signal observed at MethNotExp (silenced) promoters. 4) Segment RNAseq signal RPKM - to exclude loci showing elongation activity comparable to other open regions not showing CGI promoter use, we count only those cases with greater than median RNAseq signal in the segment regions of NotMethExp genes. 5) Segment RNAseq read coverage - to exclude loci showing elongation activity comparable to other open regions not showing CGI promoter use, we count only those cases with greater than median RNAseq read coverage in the segment regions of NotMethExp genes.

While the link between CGIs and downstream gene expression can be construed as a mode of distal enhancer mediated regulation, instead of alternative promoter action, we did not find any support to sustain that notion. We found no enrichment of tissue-specific ChromHMM annotated enhancers in MethExp CGIs across 30 tissues (Fisher's P=0.4), and this is consistent with the established

knowledge that enhancers are typically CpG-poor, and are depleted of CGIs (Illingworth et al. 2010; Kim et al. 2010). Further, in addition to observing an enrichment of splice junctions between CGIs and their corresponding MethExp genes, we find some evidence of sequence based predictors that support long, elongating, stable directional transcript production from MethExp associated CGIs. These findings are in conflict with an enhancer model, as it is well known that any transcriptional activity at active enhancers results in short, typically unstable, bidirectional RNA (eRNAs (Kim et al. 2010; De Santa et al. 2010)).

Our findings caution against relying exclusively on proximal TSS platforms in determining the transcriptional outcome of a gene, and implores us to extend focus to alternative distal elements, especially upstream orphan CGIs as they possess a “promoter-like” configuration. Very recently, a study that mapped the processes underlying the evolution of stripped-down retrocopies (intronless and promoterless copies of reverse transcribed RNA inserted into the genome) into new bona fide functional genes discovered that only a marginal fraction (~11%) of these retrocopies piggybacked on existing promoters for their expression, while the majority (~86%) co-opted orphan CGIs and other proto-promoter elements (Carelli et al. 2016). Furthermore, as retrocopies emerged into fully functional genes, most (75%-93%) gained new exons from their upstream flanking sequences; and this overrepresentation of novel 5’ exons suggests that such a gain served to place them under the control of distal promoters, including orphan CGIs.

Nevertheless, the specific molecular mechanisms underlying the context-specific

choice of proximal versus distal promoter in the case of MethExp genes remain unclear. While we cannot exclude the possibility that through some hitherto unknown mechanism, the usage of CGI is actively influenced by the methylation of the proximal promoter, it is also possible that use of the alternative CGI promoter leads to transcriptional silencing of the proximal promoter, consistent with known patterns of high gene body methylation at highly transcribed regions (Laurent et al. 2010). However, our data and the results generally suggest that the usage of CGI occurs independent of the methylation status of the proximal promoter. First, the overall ability of distal CGIs to initiate transcription (evidenced by CAGE tags, for instance) seems largely independent of the methylation status of the proximal-promoter (Fig. 3-2B,3-2D,3-2F). Second, while active histone marks are consistently a lot higher at MethExp-CGIs than NotMethExp-CGIs (i.e. at loci that are actually used as alternative promoters versus those that are not), the difference in the levels of repressive marks between these groups is not as pronounced or consistent across tissue types (Supplemental Fig. B-8), suggesting that active repression of upstream CGIs occurs (if it does) independent of the methylation status of the proximal promoter. Thus, it most likely appears that a MethExp gene utilizes (or co-opts) an already active orphan CGI as an alternative promoter, analogous to the co-option of CGIs as promoters by promoter-less retrocopies of genes discussed above.

It seems likely that MethExp-associated CGIs have been co-opted relatively recently (at a time close to the divergence of mammals from the vertebrates analyzed in this study) for their regulatory role as alternative promoters. First,

gene promoters that are more susceptible to silencing by methylation (namely, CpG-poor promoters) are associated more often with alternative promoter CGIs than CpG-rich promoters, and appear to “co-opt” their usage in specific contexts (as evidenced by locus-specific CAGE analyses). Second, methylated-promoter genes and their upstream CGI elements are more likely to have conserved synteny when they are expressed, and importantly, this tendency increases monotonically as more closely related species are used to ascertain the synteny, suggesting an evolutionary selection to keep the segment region intact. Finally, orphan CGIs have been shown to be co-opted by promoter-less genes in humans (viz., retrocopies) to transcribe their gene products, which together with our findings suggest that this is a general property of orphan CGIs. Thus, a more holistic view of the biological significance of CGIs is beginning to emerge in that they are ubiquitous substrates that are poised as transcriptional initiation sites, such that in a contextually favorable configuration (i.e., unmethylated and upstream of a stable RNA producing transcription elongation-enhancing element) can be selected for alternative promoter activity by a proximally located neighboring gene.

Materials and Methods

Datasets

Expression: RNA-seq expression for 30 primary tissues and 4 ENCODE cell lines analyzed in this study were obtained from Release 9 of the Compendium published by NIH Roadmap Epigenomics Project (Bernstein et al. 2010). This

release comprised of uniformly re-processed data for 111 consolidated epigenomes (Kundaje et al. 2015) (111 primary tissue types), wherein each sample from their original source underwent additional processing in an effort to reduce redundancy, improve quality control and achieve uniformity for integrative analysis. Raw read and processed data are publicly available and were both used in this study.

Methylation: We limited our analyses to tissues with publicly-available WGBS data, that was also sourced from the consolidated epigenomes work. Methylation measures for every CpG dinucleotide was provided in the format of fractional methylation (Reads recording a methylated CpG / Total Reads). BED files with read depth and fractional methylation information are publicly available.

Annotation: The specific version of hg19 genome annotation used in the consolidated epigenomes work cited above was GENCODE v10 (Harrow et al. 2012) (corresponding to Ensembl v65), and has therefore been carried forward in all the analyses performed in this study to maintain consistency. Although, to verify that our results were robust with respect to the latest assembly of the human genome (GRCh38), we repeated few key analyses in one cell type using the GRCh38 gene annotations and the “lifted-over” RNA-seq and methylation data. We observed that the overall trends for differences in CGI methylation levels, genomic distance and transcriptional elongation signals between upstream CGI and gene between the MethExp vs. the MethNotExp categories are consistent between the two versions (Supplemental Fig. B-11).

CpG islands: Annotations of CpG islands were extracted from UCSC Genome

Browser. This track corresponds to a hierarchical HMM model based definition of CpG islands in hg19 (Wu et al. 2010; Irizarry et al. 2009b).

Syntenic blocks: Precomputed syntenic blocks derived from whole genome sequence alignments between human (as the reference) and 6 mammalian species (chimpanzee, rhesus monkey, mouse, rat, dog and cow) as well as 2 non-mammalian vertebrate species (chicken and zebrafish) were downloaded from CINTENY (Sinha and Meller 2007).

CAGE: Single molecule CAGE profiles for 573 human primary cell samples (up to a median depth of 4 million mapped tags per sample) were generated by the FANTOM Consortium (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014). Out of the 34 tissues we analyzed, CAGE was available for 15 of them. This data is pre-processed to report the CAGE peaks associated with TSSs found genome-wide and are available as BED files. To determine the CAGE tag level in a given genomic region (for example, promoters or CGIs), we used the dominant TSS, or the TSS with the highest number of CAGE tags.

Ser5P and Ser2P RNA Pol-II ChIP-seq: ChIP-seq assays targeted to Ser5 and Ser2 phosphorylated molecules of RNA Pol-II is currently limited to only 1 cell type used in our analyses. Fold-change signal data at base pair resolution was obtained from GEO accession GSE54693 (Menafra et al. 2014).

Histone marks and DNase-seq: Processed data at base-pair resolution for several histone marks and DNase-seq (cleaving DNase hypersensitivity sites) is available from the consolidated epigenomes work cited above.

Datasets used in mouse DNMT knockout analysis: RNA-seq and WGBS

methylation data for mouse wild-type and DNMT knockout embryonic stem cells were obtained from GEO accession number GSE67867 (Domcke et al. 2015). They mapped their data to the mm9 genome assembly version of mouse (NCBIM37), and made available read density and fractional methylation at base pair resolution. Whole genome sequence, CGI and gene annotation files corresponding to mm9 were downloaded from Ensembl and UCSC Genome Browser.

Datasets used in cancer-related analyses: Data for 780 breast cancer samples, 80 matched breast normal samples (matched to their corresponding cancer samples from the same individuals), and 315 renal (kidney) cell cancer samples from TCGA (Koboldt et al. 2012; Creighton et al. 2013) were downloaded using the CGHub Repository (Wilks et al. 2014). Data for each sample comprised of 450K methylation arrays (reporting fractional methylation at select CpG probes) and RNA-seq expression (raw read file FASTQ and processed gene expression in RPKM). To obtain measures of transcriptional activity in the “segment” region (RPKM and read coverage), raw reads from each sample were aligned using STAR (Dobin et al. 2013) and further processed using the BEDTools suite of tools (Quinlan and Hall 2010).

Primary processing of genes and pooling into gene groups

The promoter of a gene was marked as methylated when the average fractional methylation level of all CpG dinucleotides lying within TSS±500 bp was greater than 0.55, and unmethylated when that value was less than 0.45. As vertebrate promoters exhibit a clear bimodal pattern of lowly and heavily methylated

promoters (Elango and Yi 2008), we consider the above thresholds to be fairly stringent. Yet, to be certain that we indeed captured only the highly methylated class of promoters in our MethExp category of genes per cell type using the above threshold, we conduct the following sanity check. We fit a Bcomponent Gaussian mixture model to the overall distribution of promoter methylation levels per tissue type to distinguish three subpopulations corresponding to lowly, intermediate and highly methylated promoters (LMP, IMP, HMP), and then checked the fraction of MethExp promoters, selected based on the aforementioned threshold, belonging to HMP separately in each tissue type. We found that on average, ~97.6% of them belong to HMP (Supplemental Fig. B-12). Further, a gene was considered 'expressed' if its expression was in the top 50^{th} percentile among all genes. The threshold adopted for expression is highly stringent and conservative since we wanted to focus on explaining the mechanisms adopted by highly expressed genes with methylated primary promoters. A gene was considered as not expressed when it had zero expression or its expression value was in the bottom 5^{th} percentile among all genes. The above criteria were used to pool genes into three gene groups, MethExp, MethNotExp and NotMethExp, in each sample.

The distal CGI associated with a given gene was defined as the closest upstream CGI annotated at a minimum distance from TSS-1500 bp. Most annotated CGIs are less than 1 kb long (~83%). Those longer than 1kb were truncated to centrepoint±500 bp for the computation of methylation levels. This did not affect the estimation of methylation levels, as these distributions are almost identical

before and after CGI truncation (Supplemental Fig. B-13). Additionally, we discarded from all three groups, every gene that contained another annotated gene between its TSS and upstream CGI element. This annotated gene could be an ambiguous ORF, or any non-coding RNA including lincRNAs, overlapping sense or antisense RNAs/genes, snRNA, tRNAs etc. annotated by GENCODE. This was done to ensure that there existed no biases from neighboring genes on our observations of intergenic transcriptional activity or neighboring epigenetic and chromatin signatures.

Evidence of gene body alternative promoter usage

To identify the fraction of MethExp genes that initiate transcription from a locus within the gene body distinct from its proximal promoter, we quantified the expression level of all exons within each gene. Then for each MethExp gene, if the expression level (RPKM) of the first exon was zero or in the bottom 5th percentile among all exons of all genes, then that gene was concluded to possess a silenced primary promoter with an active gene body alternative promoter.

Tissue specificity index (TSI)

The quantitative measure of TSI, is defined as:

$$\mathrm{TSI} = \sum_{n=1}^{N}(1 - x_i)/(N - 1)$$

where N is the number of tissues and x_i is the expression profile component normalized by the maximal component value (Yanai et al. 2005).

Evolutionary conservation

Conservation of distal CpG islands was calculated at two distinct levels.

Inter-species conservation: We used genome-wide base-pair resolution phastCons scores that were precomputed from the multiple sequence alignment of 45 vertebrate genomes to the human genome (Siepel et al. 2005).

Intra-species conservation: We used genome-wide human polymorphism data from the 1000 Genomes project (The 1000 Genomes Project Consortium 2012) to infer the extent of intraspecific selection pressure acting on distal CGI elements. A derived allele is one that arises in a population due to a mutation in the original allele in the population (ascertained by comparing with multiple closely related species to human). By definition, the derived allele starts out "rare" and its frequency can increase in a population over time due to genetic drift or in rare occasions, positive selection. If the mutation, or the derived allele is deleterious, its spread will be curtailed due to selection pressures acting on it, thereby resulting in a low derived allele frequency (Vishnoi et al. 2011). Therefore, a low DAF in given region may suggest negative selection in that region. For each CGI, we generated the derived allele frequency (DAF) spectrum by pooling DAFs at all nucleotides within that region. Thus, for each gene loci, there existed a DAF profile corresponding to its upstream CGI.

Cell-type specific regulation of alternative promoter CGIs

Motif information for 642 TFs (those with available Positional Weight Matrix (PWMs) in TRANSFAC (Matys et al. 2006) and expression data across cell types) and the sequences of all CGIs showing evidence of alternative promoter

activity in some cell type were input to PWMSCAN (Levy and Hannenhalli 2002), a tool that scans sequences to identify significant motif matches. Matches with PWM scores in the top 5% were retained, and expression profiles of the corresponding TF genes were obtained. Then for each locus, the distribution of the expression profiles of these TFs in cell-types where the CGI was active was compared to a similar distribution arising from cell types where the CGI was inactive using Wilcoxon test.

Sequence-based splicing signals

Sequences spanning the intergenic region between the TSS of MethExp, MethNotExp and NotMethExp genes and their associated upstream distal CGIs ("segment region") were extracted using the hg19/GRCh37 reference genome from UCSC Genome Browser. Motif information and frequency matrices for the U1 binding site and PAS recognition sequence was obtained from Almada et al. (Almada et al. 2013). The motif frequency data was transformed to position weight matrices and was input to PWMSCAN (Levy and Hannenhalli 2002), a tool that scans sequences to identify significant motif matches. Matches with PWM scores in the top 5% were retained, and the order of motifs on a given sequence was inferred. If the first 1500 bp of the segment region contained a match for U1 before PAS, the corresponding gene loci was assigned the label "stable", and "unstable" in case the motif order was switched.

Gene Ontology (GO) enrichment

DAVID Bioinformatics Resource 6.7 (Dennis et al. 2003) was used for all GO enrichment and functional annotation performed in this study.

CHAPTER 4: High-throughput detection and analysis of protein interaction-based regulatory rewiring events

Abstract

Similar to evolutionary changes in the sequence of CREs, changes within coding regions of TFs can allow for altered protein-protein interaction capabilities and function, through motif and 3D domain turnover across evolution. For example, the TF FTZ, has switched from serving a homeotic role in ancestral insect species, to being involved in segmentation in the *Drosophila* genus. This switch in FTZ's function is accompanied by the loss of YPWM, a protein sequence motif that is responsible for cofactor interactions with homeotic regulators, and the gain of a LXXLL motif that enables interaction with segmentation-related cofactors and targets. Elucidating the occurrence of, and mechanisms underlying these switches in TF function is critical to our understanding of evolutionary plasticity of gene function, especially the highly conserved developmental genes. To this end, we used a pairwise SVM method for species specific PPI prediction between 1200 TFs in 12 related arthropod species, followed by detection of protein interaction-mediated regulatory rewiring. Based on simulation studies, we show that the accuracy of detection of rewiring events using the above PPI prediction method is approximately ~80-85%, which recapitulates the known FTZ-EXD to FTZ-FTZ1 interaction rewiring event amongst the top 5% of all events involving FTZ. Amongst all rewiring events involving all TF protein triplets, we find an overrepresentation of a protein member involved in the "enhancer of split"

complex, which is known to have undergone lineage specific gene losses and duplications. We expect that a deeper investigation of the rewiring events involving this protein member may reveal crucial information about regulatory network changes in neurogenesis across insect evolution. Overall, this work establishes that regulatory rewiring mediated by interaction changes is likely to be prevalent in arthropod development, and provides a high-confidence list of such candidates for future follow up.

Introduction

Embryonic development is a highly conserved process regulated by a set of core regulatory genes and proteins across diverse species. For example, mammalian *Hox* genes (involved in regulating homeotic processes during development) when mis-expressed, could recapitulate some of the same phenotypes caused by their fly counterparts (Lutz et al. 1996). Yet, there are documented instances whereby novel genes are recruited into pre-existing regulatory networks and others are lost (Zhong and Holland 2011), and this appears to be a fundamental feature of a stepwise process underlying metazoan evolution. Understanding mechanisms controlling the balance between constraint and variation in regulatory networks is a primary focus of the Evo-Devo field.

Rewiring of connections within regulatory networks may result through changes in cis-regulatory elements (CREs) controlling gene expression such that it is brought under the control of new regulators. This flexibility is afforded by the modular nature of CREs, which allows retention of ancestral functions even as

new functions are gained, as well as loss of existing functions without affecting the newly acquired functionality. Similarly, regulatory rewiring can also be brought about by the modularity of protein domains, which provides the flexibility for functional changes during evolution for which protein sequence changes would be expected to be highly deleterious. In fact, it has been argued that the evolution of proteins may play a more substantial, but far underestimated, role in developmental evolution. (Lynch and Wagner 2008).

One such example of regulatory rewiring mediated by the alteration of protein sequence and thereby function occurred during the development of diverse arthropod species. Specifically, the TF FTZ, has switched from serving a homeotic role in ancestral arthropod species, to being involved in segmentation in recent holometabolite insects including the *Drosophila* genus. In the blastoderm stage of embryogenesis in *Drosophila*, the major body axes and segment boundaries are determined by “segmentation” genes whereas subsequent development of these segments into morphologically distinct structures like legs, wings, and antennae require the action of “homeotic” genes. The switch in FTZ’s function is accompanied by the loss of YPWM, a protein sequence motif that is responsible for cofactor interactions with homeotic regulators like EXD, and the gain of a LXXLL motif that enables interaction with segmentation-related cofactors and targets (Heffer et al. 2013). Specifically, having acquired the LXXLL motif, the FTZ possesses the capability to dimerize with FTZ-F1 (via the AF-2 domain), a TF controlling segmentation processes across diverse insect species. These TFs cooperatively bind to their target gene

engrailed, to activate its expression and initiate this stage of development in *Drosophila* (Florence et al. 1997; Yu et al. 1997; Guichet et al. 1997).

Despite the extent of protein sequence changes across clades, very few studies have been reported that map these in a genome-wide fashion, specifically as they pertain to regulatory rewiring. Elucidating the occurrence of, and mechanisms underlying these evolutionary switches, are critical to our understanding of the evolution, and for rational design of targeted experiments to probe these processes. One of the challenges in identifying clade-specific protein interaction rewiring is to be able to determine species-specific protein-protein interactions (PPI). However, such data has not been experimentally determined for a majority of species and one must rely on computational estimations of species-specific interactions. To this end, we have used a pairwise SVM (Ben-Hur and Noble 2005) method; with protein motif and domain presence/absence, as well as known interaction potential information as features, to comprehensively screen for evolutionary changes in protein motifs/domains affecting interaction preferences of 1200 TFs across 12 related arthropod species. The method itself achieves a cross-species accuracy (AUROC) of ~70% when predicting protein-protein interactions from sequence. We then applied a probabilistic method (based on a previously developed cis-regulatory rewiring detection method (Sarda and Hannenhalli 2015)) to identify all instances of protein interaction based regulatory rewiring across different evolutionary lineages spanning 12 related arthropod species. Based on simulation studies, we show that the accuracy (AUROC) of detection of rewiring events using the above

PPI prediction method is approximately ~80-85%; a 10-15% increase in power achieved as a consequence of pooling information across 12 species for the rewiring detection step.

Our method recapitulates the known FTZ-EXD to FTZ-FTZ1 interaction rewiring event amongst the top 5% of all events involving FTZ and the other TF pairs, suggesting that the method performs well. Interestingly, upon consideration of just the TF partners highly expressed during some developmental stage (total of 242), the regulatory rewiring landscape of FTZ appears to be quite different. The known FTZ-EXD to FTZ-FTZ1 interaction rewiring event no longer appears in the top 5% of events involving FTZ and other developmental TF pairs, but rather only within the top 25%. This suggests that a lot of restructuring of the developmental TF network occurred across insect/crustacean evolution. Amongst the top 1% rewiring events involving all TF protein triplets, we find several protein members involved in the “enhancer of split” complex. It is known that throughout insect and crustacean evolution, genes encoding this complex (which is critically important for neurogenesis) have undergone lineage specific gene losses and duplications (Dearden 2015). Therefore, we expect that a deeper investigation of the rewiring events involving these proteins may reveal crucial information about regulatory network changes in neurogenesis across insect evolution.

Overall, this work establishes that regulatory rewiring mediated by interaction changes is quite prevalent in arthropod development, and provides a high-confidence list of such candidates for future follow up.

Results

Quantifying evolutionary interaction rewiring

Our overall strategy for detecting interaction-mediated regulatory rewiring is based upon our success in cis-regulatory rewiring detection in yeast (Sarda and Hannenhalli 2015). As shown in Fig. 4-1, for a TF X (rewiring candidate), and a select lineage 'b' (partitioning the species into two groups), we will compute a probabilistic score which assesses the proposition that TF X interacts with TF Y in one subgroup of species and instead with TF Z in the other subgroup of species. We thus compute a rewiring score (RS) for changes in the interaction potential for every TF a branch that separates the holometabolite insects from the other arthropods as shown in the species tree (Fig. 4-2). The analyses can, in principle, be extended to other partitions of species in the tree, but here we report results for just the one indicated in the figure. Following the logic for cis-regulatory rewiring above, the rewiring score RS(X,Y,Z,b) can be estimated as:

$$RS(X,Y,Z,b) = \frac{log(P(X,Y,s1)) + log(P(X,Y,s2)) + log(1-P(X,Z,s1)) + log(1-P(X,Z,s2))}{2} + \frac{log(P(X,Z,s3)) + log(P(X,Z,s4)) + log(1-P(X,Y,s3)) + log(1-P(X,Y,s4))}{2}$$

where P(X,Y,s) quantifies the probability that TF X and Y interact in species s.

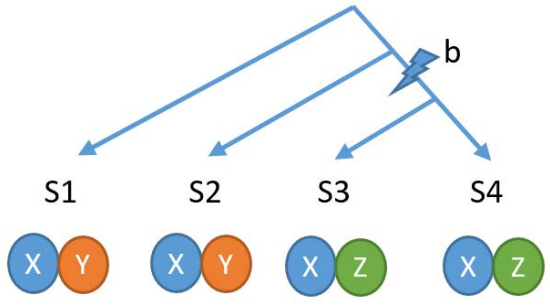


Figure 4-1. Illustration of approach to assess interaction rewiring. This sample tree shows four species (s1, s2, s3, s4) partitioned at a select branch b to produce the partition of two

species in the left clade (s1, s2 ϵ S) and two species (s3, s4 ϵ T) in the right clade. Interaction patterns of transcription factors X, Y and Z are such that, X interacts with Y in species in the left clade S, and Z in species in the right clade T, and not vice versa.

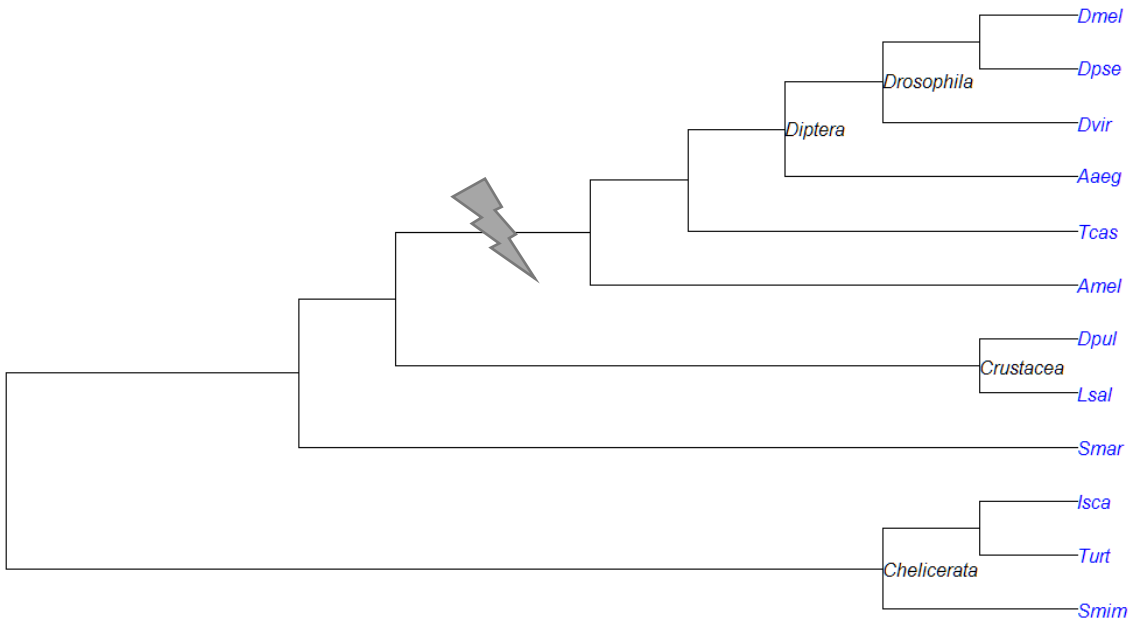


Figure 4-2. Phylogenetic tree of select arthropod species. Tree shows relationships between the 12 arthropod species surveyed in this analysis. The branch across which we partitioned the species to assess lineage-specific trans-regulatory rewiring is indicated by a bolt.

Estimating interaction probabilities between a pair of proteins

The available protein interaction databases such as STRING (Franceschini et al. 2013), BioGrid (Chatr-Aryamontri et al. 2015), HPRD (Keshava Prasad et al. 2009), etc. cannot be used directly to get protein-protein interaction probabilities of pairs of proteins in a given species (i.e., P(X,Y,s) from the previous section), as they are generally not based on species-specific experiments, but instead interaction information is transferred from closely related species based on orthology. In short, species specific PPI data is virtually non-existent for most arthropod species (except in *Drosophila melanogaster*). Therefore, to ensure that we uncover cases of true innovation affecting interaction preferences in a

species-specific fashion, we will computationally estimate the likelihood of binding between a protein pair based only on their sequence. An advantage of methods that predict interactions from just sequences is that, since they learn the interaction rules from the primary amino acid (AA) code which are the fundamental building blocks of proteins, they will easily extend to predicting interactions between proteins from sequences in any species. Thus, a high degree of generalizability in cross-species PPI prediction is expected, which is especially useful for our use case as we shall train on PPI data in *D. melanogaster*, and predict protein interactions using this model in the other 11 species.

Previous methods: Although PPI prediction from primary sequence is recognized to be a challenging problem, it has been addressed by several groups in the past with reasonable success. Following some of the early seminal works that used amino acid subsequence-based features in a graph or SVM based learning model (Bock and Gough 2001; Pitre et al. 2006; Shen et al. 2007; Guo et al. 2008), many groups with slightly different feature extraction protocols and ensemble learning models improved upon the overall prediction accuracy (Zhang et al. 2014; Huang et al. 2015; You et al. 2014). A preliminary evaluation of two such off-the-shelf PPI prediction methods (Yu et al. 2010; You et al. 2015) that use simple sequence features such as count frequency of AA triads, distribution and composition of AA properties like hydrophobicity, polarity within subsequences, etc. revealed that while the accuracies of PPI prediction within species is relatively high (in similar ranges to those reported in the papers; 90-

97%), the methods fail to generalize across species. Both methods were trained on *D. melanogaster* PPIs and tested on *H. sapiens* PPIs (see Methods for PPI data treatment, source and negative set generation), and vice versa, as a proxy for expected arthropod cross-species accuracy estimates. This resulted in an accuracy of less than 53-55%, and we got similar accuracies using their own datasets as well (Figure 4-3). A relatively recent paper (Park and Marcotte 2012) surveyed over 50 similar simple sequence based methods including the above and reported that most of these reach accuracies of ~55% when proteins in the test set were never "seen" during training. As neither of the members of a protein pair are seen in cross-species PPI prediction, we cannot expect any of them to perform well. Therefore, based on our preliminary assessment, we believe that most of the current methods suffer in two crucial aspects: a) simple-sequence based features are inadequate for tools to learn 3D structure rules that are far more relevant for interaction, and b) the encoding of features ignores the pairwise nature of the input; i.e., a simple concatenation of features belonging to each protein member of the pair for which interaction potential is learnt/predicted does not correspond to a space where features/attributes about their "specific paired nature" is encoded.

	Off-the-shelf Method		Published in	Actual accuracy
1)	Features: Model: Reported accuracy:	Amino acid composition and distribution Random Forest Within species: 95% Across species: 87-90%	PLoS One	
2)	Features: Model: Reported accuracy:	Amino acid triad frequencies and significance RVKDE (Variation of SVMs) Within species: 82% Across species: NA	BMC Bioinformatics	

Train ➡ Test ⬇	Pylori	Yeast	Human
Pylori	85*	55	50
Yeast	51	93*	50
Human	55	46	97*

Train ➡ Test ⬇	Fly	Human
Fly	80*	53
Human	51	84*

Figure 4-3. Off-the-shelf PPI prediction method evaluation. A summary of features, models, reported and actual accuracies of two popular off-the-shelf PPI prediction methods when used for within and cross species interaction prediction.

Alternative methods: In seeking to remedy this, we first developed a prototype method that uses features that are representative of (i) the 3D structures (viz., domains and linear motifs) adopted by individual protein members of a given pair, and (ii) interactions that occur within the specific protein pair itself. Specifically, we used as features the frequency of known domain-domain interactions (DDIs), domain-motif interactions (DLIs) and > 200 linear motifs found in a protein pair in a Random Forest framework (see Methods for domain, motif, DDI and DLI detection and quantification strategies). We found that this prototype method achieves ~80% AUROC (5 fold cross-validation) within Human and Fly, with a cross-species AUROC of 71%. This shows that using structural information (even in the naïve manner as the above) recovers some of the generalizability of PPI prediction across species.

Note: Ideally, one would like to also encode features in the learning method

appropriately. Specifically, features like domain and motif frequency are protein-member specific, whereas DDI and DLI are protein-pair specific. Kernel based methods allow encoding protein-member specific features in a paired manner, as shown by Ben-Hur et al. (Ben-Hur and Noble 2005). Briefly, kernels are similarity functions over a pair of datapoints, and are computed in a higher-dimensional space representation of the original features. They can be used in learning methods (like Support Vector Machines) whose optimization function possesses a dual form, which requires only the dot product of input vectors. This allows computation of optimality in any high-dimensional space without explicitly knowing the transformation function that takes features into that space (the kernel trick). Specifically, in the above method, pairwise encoding was achieved by expressing the similarity between pairs of proteins in terms of similarities between individual proteins (see Methods). In fact, the tool published by this group (PyML) allows combining unpaired and paired features in distinct kernels (where kernels can be weighted disproportionately) in a single framework; we use this tool to strengthen our PPI prediction task. Using the same human and fly PPI data, frequencies of over 5000 domains and 200 linear motifs were computed for each member of the protein pair and encoded using a pairwise kernel. We added individual unpaired kernels for DDIs and DLIs as well. With suitable differential weighting of the kernels (see Supplemental Fig. C-1), we achieve upto a maximum of 80% within species accuracy, and 72% cross-species accuracy (based on AUROC estimates). Although, this appears to be upto par with our naïve method, we have not yet been able to assess if our

results and overall conclusions hold by using this method. Therefore, in the main text, we only present results and conclusions based on the application of the Random Forest method on our datasets.

Simulation-based power analysis of rewiring event detection

It is still unclear how effective the prototype PPI detection method, combined with the rewiring score function is in recovering true rewiring events from the overall population. Further, it would be interesting to know how well a sequence based PPI prediction method with 70% accuracy across species performs with respect to rewiring event detection in 12 related species, such as it is in our specific use case. To get a sense of how the performance of rewiring detection is affected or influenced by the accuracy of the underlying PPI prediction method, we carried out a simulation that uses the predicted probabilities of the above Random Forest based method to model rewiring events. Specifically, at a particular branch *b* of the 12 species phylogenetic tree that partitions the tree into two lineages of 6 species each, we simulate 1000 true rewiring events, such that the ground truth is that X interacts with Y on the left clade *l*, and with Z on the right clade *r*, and not vice-versa. We sample probabilities for true interactions {P(X,Y) in *l*, P(X,Z) in *r*} from the overall distribution of predicted probabilities for true PPIs in Human (using the *Drosophila* trained model). Similarly, we sample probabilities for false interactions {P(X,Y) in *r*, P(X,Z) in *l*} from the overall distribution of predicted probabilities for false PPIs in Human using the same model (see Fig. 4-4A for true and false PPIs interaction prediction probabilities). The fly to human cross-species model was chosen for this simulation so as to replicate the expected

output from our use case, which is where the fly model will be used to predict PPIs in the other 11 arthropod species. We then compute rewiring scores of these true rewiring events based on assigned PPI probabilities. To generate a suitable background, for each true event, we shuffle 100 times the assigned PPI probabilities between species using a phylogenetic permutation model, described in Lapointe and Garland 2001 (Lapointe and Garland 2001). A phylogenetically restricted permutation does not shuffle in a completely random manner, but rather shuffles accounting for the fact that the PPI profiles of closer species are more similar. We then compute the “false” rewiring scores based on these phylogenetically permuted PPI probabilities. Using the rewiring scores of the true and false events, for different score cutoffs, we obtain the sensitivity, specificity and accuracy of rewiring detection, as well as the area under the ROC curves. We repeat the above using a few different values for the following additional parameters, viz., 1) loss ϵ {0,1,2} – randomly pick 0,1,2 species from each clade and simulate a loss of PPI interaction, to replicate clade specific loss of interactions. 2) fluidity ϵ {1,2,3,5} – determines how strict the resulting PPI probability permutations are, i.e., lower the value of fluidity, the closer the resulting permuted values are to the original. Here, we show that for loss=1; fluidity=2, the AUROC of rewiring event detection method is 0.85 (Figure 4-4B). We would like to emphasize that the underlying PPI model only achieves a cross-species AUROC of 0.71, and even after using a stringent background and modeling species-specific losses, there is a 15% boost in AUROC for rewiring detection.

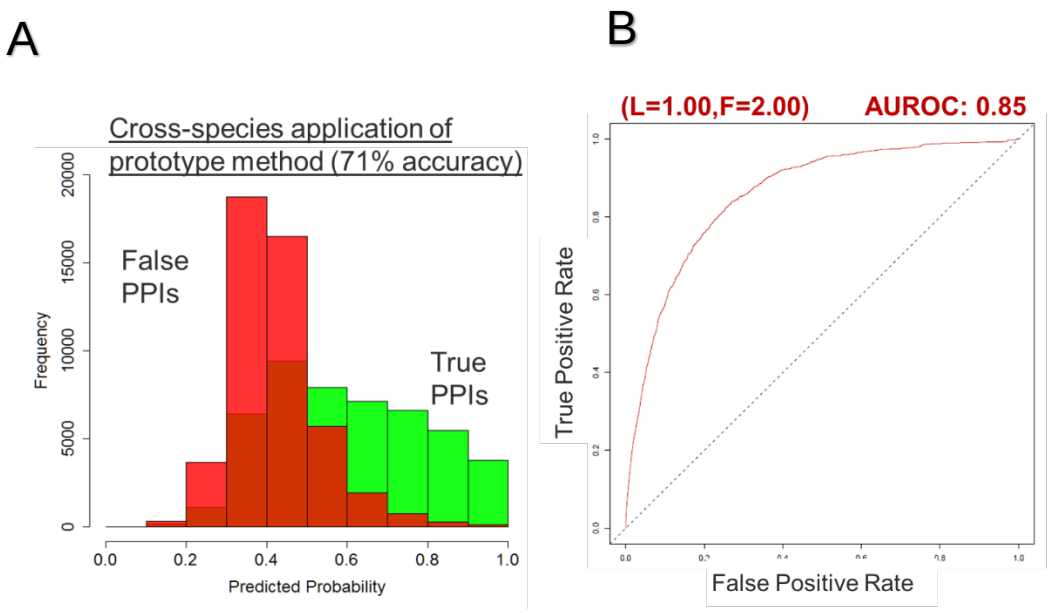


Figure 4-4. Fly to human predicted PPI probability distribution and resulting rewiring event detection accuracy. (A) A histogram depicting the probability distribution resulting from cross-species application of the prototype random forest based method for true and false PPIs in human. (B) A UROC of rewiring event detection method achieved by (i) randomly modeling lineage specific loss of interaction in one species, and (ii) a restricted permutation using a fluidity value of 2 for generation of background events.

Recapitulation of the FTZ rewiring event

We used our prototype random forest method (with AUROC of 0.71 for cross-species prediction) to predict interactions amongst TF pairs in diverse arthropods and asked if our rewiring detection method was capable of successfully recapitulating the known FTZ event. Thus, to this end, for about 800 TFs (out of a total of 1168 TFs in fly) that had mappable orthologs in at least 10 out of 12 arthropod species, we obtained the corresponding protein sequences and ran our PPI prediction method to predict interaction scores between each pair of TFs in each species. This resulted in a total of $^{800}C_2$ * 12 predicted interaction probabilities between every TF pair in every species. Then for each possible TF

protein triplet (X,Y,Z), we computed the rewiring scores at the partition (indicated in Fig. 4-2) using the actual predicted probabilities in each species (see Methods). Fig. 4-5 shows the background distribution of all rewiring scores of protein triplets that involved FTZ with the 5% and 95% quantiles of the distribution. Fascinatingly, we observe that the known event involving FTZ, i.e. (FTZ,FTZ-F1) in the lineage species and (FTZ,EXD) in the basally branching species has a rewiring score of -2.62, which is amongst the top 5% of all events involving FTZ. The background distribution is a TF-specific one, and is thus hence more stringent than looking across the board of all rewiring scores.

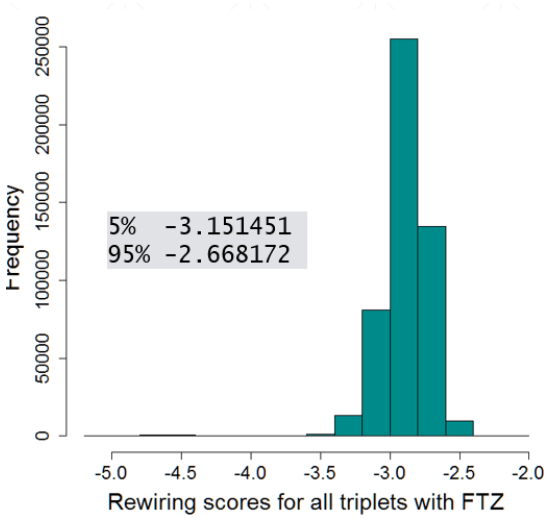


Figure 4-5. Distribution of rewiring scores for all triplets with the TF FTZ. The histogram also depicts the 5th and 95th percentile values of the distribution.

Functional insights about interaction based rewiring in arthropod species

In recent years, several examples have emerged to demonstrate that embryonic regulatory genes can be co-opted into new pathways through changes in cis-elements controlling their own expression. For example, in addition to its typical

Hox-like expression, Antp is expressed in and controls eyespots on butterfly wings (Saenko et al. 2011). Other *Drosophila* developmental genes have also been shown to change function in insect lineages: e.g., eve is not expressed in stripes in grasshoppers (Patel et al. 1992) and functions as a gap gene in Oncopeltus (Liu 2005). In Oncopeltus, a nuclear receptor, E75A, which is not a pair-rule gene in *Drosophila*, has pair-rule function. These variations in gene expression and function during development suggest a prevalence of loss- or gain-of-function across insect evolution.

To get a sense of the how much regulatory interactions vary as they pertain to TFs predominantly involved in the development process, we repeated our rewiring event detection analysis in just 242 of the highly expressed TFs during some stage of embryonic development (see Methods). Interestingly, the regulatory rewiring landscape of FTZ does not seem to be exceptional (the FTZ-EXD-FTZF1 appears among the top 25% of rewiring scores), suggesting that a lot of restructuring of the developmental TF network has occurred across insect/crustacean evolution. We believe that a deeper investigation of these candidates is warranted, and will report our findings in a future version.

Further, amongst the top 1% rewiring events involving developmental TF triplets, we find several protein members involved in the “enhancer of split” complex. It is known that genes encoding this complex arose before the common ancestor of insects and Crustacea, and before the formation of the complex itself. Throughout insect and crustacean evolution, these complex-forming genes have undergone lineage specific gene losses and duplications. The enhancer-of-split

complex is involved in neurogenesis in the genomes that they are found in, but appear to be missing from the genomes of chalcid wasps, raising questions as to how these species carry out neurogenesis in the absence of these genes (Dearden 2015). Therefore, we expect that a deeper investigation of the rewiring events involving this protein member may reveal crucial information about regulatory network changes in neurogenesis across insect evolution.

Discussion

The modularity of protein domains provides the flexibility for functional changes during evolution, demonstrated by changes in the function of embryonic transcription factors such as the case of FTZ across arthropod evolution. Protein sequence changes such as the above would be expected to be highly deleterious and are highly surprising since loss- or gain-of-function changes in segmentation genes in lab animals usually result in lethality. Yet, these are occurring and persistent in much more challenging natural environments. To gain broader insights into change of function, resulting from changes in interaction preferences of regulatory TFs, we have developed a method; with protein motif and domain presence/absence, as well as known interaction potential information as features, to comprehensively screen for evolutionary changes in protein motifs/domains affecting interaction preferences of 1200 TFs across 12 related arthropod species, followed by identification of all instances of protein interaction based regulatory rewiring. Based on simulation studies, we show that the accuracy of detection of rewiring events is approximately ~80-85%.

Predicting interaction potential from protein sequence is an inherently challenging problem, and while a high number of false positives in predicting interaction in an individual species is expected, in computing rewiring scores, our reliance on evidence of interactions in multiple species should reduce false positive inferences of rewiring events. As we mention previously, a 10-15% increase in accuracy was achieved as a consequence of pooling information across 12 species for rewiring event detection.

Most of the binding affinity of a linear motif comes from 4-8 residues, which introduces a significant amount of noise in the predictions, at two levels, (1) in the power of detection within proteins due to its low complexity (unreliable confidence scores or p-values), and (2) in the actual biological stability of detected interactions (such as DLIs, as low affinity often results in transient or reversible interactions). We attempted to alleviate the former by only considering exact matches to the regex definitions provided by ELM for linear motifs. Although noise originating from the latter is a natural outcome of the biology of linear motif mediated interactions (whether it is functional or not), and cannot be approximated by *in silico* methods.

We note that our method here can only identify all those TFs that have globally rewired their interaction preferences, but not those that exhibit differential co-factor binding preferences in a target-specific manner. Such combinatorial behavior generally has cis-element underpinnings that ensure differential localization to target genes. Integrating the cis-framework to allow detection of local interaction rewiring would be challenging because the regulatory regions

can be far away from transcription start sites and are largely unknown, although using approximations with evidence of regulatory potential across multiple species (using conservation scores) might be a viable option.

In summary, our method successfully recapitulates the known FTZ event, provides a high-confidence resource of previously unknown interaction-based rewiring events, and also suggests that the developmental TF network has undergone extensive restructuring across evolution. The results from a deeper investigation of these predicted rewiring events will be presented in future version.

Materials & Methods

TF annotations, sequences and orthology groups

Annotations of TFs were obtained from FlyTF – The *Drosophila* Transcription Factor Database (v1.0) (Pfreundt et al. 2009), based on a loose definition of "TF-like" terms in their description. 11 other arthropod species were probed for orthologues of these selected *D. melanogaster* TFs, viz., *Drosophila pseudoobscura* (fruitfly), *Drosophila virilis* (fruitfly), *Aedes aegypti* (mosquito), *Tribolium casteneum* (red flour beetle), *Apis mellifera* (honeybee), *Daphnia pulex* (common water flea), *Lepeophtheirus salmonis* (salmon louse), *Strigamia maritima* (centipede), *Ixodes scapularis* (black-legged tick), *Tetranychus urticae* (two-spotted spider mite) and *Stegodyphus mimosarum* (African social velvet spider). The choice of species was made based on the availability of whole proteomes, as well as diversity amongst arthropod species. Orthologs of *D.*

melanogaster TFs were determined using two tools, viz., Ensembl Compara (Perl API) and metaPhOrs (Pryszcz et al. 2011), both of which are phylogeny based orthology predictions and merged to increase coverage. The protein sequences of all orthologous TFs across species were downloaded from Ensembl database.

Domain and linear motif detection

Protein domains and linear motifs were detected from each individual TF protein sequence using INTERPROSCAN and ELM motif definitions respectively. Only Pfam and SMART domains detected per sequence were retained resulting in a total of 5948 unique domains across all sequences. A custom script searching for ELM regex definitions returned about 240 unique linear motifs across sequences. The frequency vectors of domain and linear motif counts were generated per TF per species and recorded for further analysis.

Domain-Domain (DDI) and Domain-Linear Motif Interaction (DLI) detection

The annotations of approximately 8200 known DDIs and 270 known DLIs were obtained from DOMINE (Yellaboina et al. 2011) and ELM (Dinkel et al. 2014) respectively. In addition, 2682 DDIs that are experimentally known not to occur were derived from Negatome 2.0 (Blohm et al. 2014), which is a collection of protein domain pairs which are unlikely engaged in direct physical interactions. The frequency vectors of positive DDIs, negative DDIs and positive DLIs were generated per TF-pair per species and recorded for further analysis.

PPI data source, treatment and negative set generation

Approximately 44000 PPI pairs were retained in the positive set after filtering of

D. melanogaster PPIs in STRING (Franceschini et al. 2013) database that had (a) support from experimental studies, and (b) did not involve pairs where either of the protein members had the amino-acid selenocysteine (U) in their sequence. We generated a suitable negative PPI set of equal size by random shuffling of true PPI pairs, and ensuring there was no overlap from the positive set.

Note: All of the above steps, including proteome download, domain and linear motif detection, DDI and DLI detection, PPI positive and negative set generation were repeated in human, so as to have an additional dataset to assess the performance the PPI prediction methods across species. The methods were identical, except that the PPI annotations of over 41000 pairs themselves were downloaded from HIPPIE (Schaefer et al. 2012) instead of STRING, for better data quality.

Alternative PPI prediction methods

Prototype random forest based method: A random forest is a supervised learning algorithm that uses a voting scheme to classify an observation based on the consensus of a collection of 1000s of decision trees that are each trained on a random subset of datapoints and features (Breiman 2001). For the PPI prediction method, we train the method on known PPIs and random PPIs (positive and negative set response variable set to either 0 or 1 respectively) in fly and human. For each protein pair, the features include frequencies of counts of positive DDIs, negative DDIs and positive DLIs found in that pair, concatenated with the unique linear motif counts per protein member totaling upto 481 features. We used the

caret package in R for implementation and report 5-fold cross validation accuracies within and across species in the main text.

Pairwise SVM method: An SVM is also a supervised learning algorithm for classification. An important property of SVMs, viz, the dual form of the optimization function allows the use of kernel functions to SVMs which apply a transformation to the input vectors. This allows the transformation of unpaired features, such as domain and linear motif counts in individual protein members of a pair, to a space of paired features representative of the specific protein pair. This was done by Ben-Hur et al. in 2005 in the context of protein-protein interaction methods. The kernel, i.e., similarity function between input vectors can be transformed to a pairwise kernel by considering the similarity between pairs of input vectors (pairs of features corresponding to a protein pair). The most straightforward way to construct this pairwise kernel is to express the similarity between pairs of proteins in terms of similarities between individual proteins. In the approach, they consider two pairs to be similar to one another when each protein of one pair is similar to one protein of the other pair. For example, if protein X1 is similar to protein X'1, and X2 is similar to X'2, then it can be said that the pairs (X1,X2) and (X'1,X'2) are similar.

These intuitions are translated into the following pairwise kernel:

$K((X1,X2),(X'1,X'2)) = K(X1,X'1)K(X2,X'2) + K(X1,X'2)K(X2,X'1)$,

This kernel takes into account the fact that X1 can be similar to either X'1 or X'2. It is implemented in the PyML package developed by Dr. Asa Ben-Hur, and we used this package to develop the PPI prediction method. We train the method on

known PPIs and random PPIs (positive and negative set response variable set to either -1 or 1 respectively) in fly and human. For each protein pair, the features include (i) unpaired kernels for frequencies of counts of positive DDIs, negative DDIs and positive DLIs found in that pair, and (ii) pairwise kernels of unique domain and linear motif counts per protein member. We report 5 fold cross validation accuracies within and across species in Supplementary Figure C-1. This work is still in progress, although based upon a preliminary assessment it appears as if the method achieves similar accuracies to the prototype method and might replace the results from the latter in the main text in a future version.

Probabilistic rewiring score

Generalizing the rewiring score function presented in the main text, we get

$$RS(X,Y,Z,b) = \sum_{s \in L_b} \frac{log(P(X,Y,s)) + log(1 - P(X,Z,s))}{L_b} + \sum_{s \in R_b} \frac{log(P(X,Z,s)) + log(1 - P(X,Y,s))}{R_b}$$

where X, Y and Z are transcription factors, *s* is a given species, and L_b and R_b denote the sizes of the left and right clades resulting from a partition at branch *b*, respectively. The P(TF1, TF2, species) terms are computed by plugging in the probabilities of interaction between TF1, TF2 in each of the arthropod species, as predicted by the model trained on *D. melanogaster* PPIs.

Identification of developmental TFs

We used a published RNAseq dataset on the developmental transcriptome in *D. melanogaster* (Graveley et al. 2011), where gene expression was measured during every consecutive two-hour interval during embryogenesis. In each of the 12 stages (i.e., within 24 hrs of embryogenesis), we identified the top 10% highly

expressed TFs, and combined them to produce a unique set of 242 developmental TFs.

CHAPTER 5: Perspective and future work

Association studies that map genotype to phenotype or disease unanimously report that a majority of the signal lies in non-coding regions. Yet these signals are notoriously hard to interpret given that the regulatory genome and its range of functional diversity has largely remained an enigma. The recent advances in the systematic annotation of functional noncoding elements, using information about sequence motifs, chromatin state, epigenomic marks, evolutionary conservation etc. have helped develop richer regulatory models seeking to explain the underlying mechanisms that generate phenotypic variation.

Yet, there still remains much to be understood about genome regulation, and this thesis seeks to improve our understanding about some of the lesser explored observations pertaining to the process of regulation. We report that some regulatory innovations, such as cis-regulatory rewiring as well as regulatory TF functional changes (mediated by alteration of interaction partners) are highly prevalent across evolution. We also report, for the first time, a previously unknown general regulatory mechanism by which distal elements (viz., CpG islands) can be poised to serve as alternative promoters to nearby silenced genes, which also has large implications for specific genes expressed in cancer phenotypes.

Specifically, we answer some important questions pertaining to the process of cis-regulatory rewiring. For instance, (i) how widespread is a comprehensive shift in transcriptional regulation of a regulon, and (ii) what are the features of target genes that make them amenable to rewiring? By gathering more candidate

rewiring events and collectively analyzing their trends, we were able to answer these questions and gain further insights into conditions conducive to rewiring, as well as enable discovery of clade/specific instances of regulatory innovation.

Our results indicate that cis-rewiring is pervasive; it is further likely that if our analysis of regulons is not restricted to those with conserved expression across species, it will reveal many more rewiring events that have the potential to shed light on the divergence of functionally related genes' expression mediated by rewiring. Although some of the detected events have functional consequences, it is very likely that a lot of these are manifestations of high *cis*-regulatory plasticity and represent a neutral shift in regulation.

Our observation that the rewired TF pairs tend to function in similar biological processes compliments a previous observation that evolutionarily diverged targets of a TF nevertheless possess common functions (Habib et al. 2012). Taken together, these two observations suggest high plasticity in regulatory networks. We also found that rewired TFs are generally controlled by a common upstream regulator, and occupy lower levels of regulatory hierarchy, which are both consistent with expectations.

It is likely that regulatory rewiring at the individual gene level are frequent and without strong selection (as also noted in (Habib et al. 2012)), while repeated rewiring at multiple loci consisting of regulons (functionally related genes) may partly be due to directional selection. In the future, knowledge from deeper phylogeny could be used to infer the temporal ordering of specific events in *cis*-element evolution (such as the events within a regulon) which may help

distinguish potential seeding events from the ones that follow, likely under selection.

Analogous to cis-rewiring, we show that there have been several instances whereby changes within coding regions of developmentally important TFs have allowed for altered protein-protein interaction capabilities and function, through motif and 3D domain turnover across arthropod evolution. We do so by developing a method; with protein motif and domain presence/absence, as well as known interaction potential information as features, to comprehensively screen for evolutionary changes in protein motifs/domains affecting interaction preferences of 1200 TFs across 12 related arthropod species, and identify all instances of protein interaction-based regulatory rewiring.

While our analyses so far provide some functional insights, we have not yet been able to achieve all of the primary goals we set out to at the onset of this work. We are interested in:

(i) Curating the list and selecting 20 candidates and testing, using Y2H assays and co-immunoprecipitation, whether the specific sequence difference in a TF across species results in the predicted differences in protein interaction. A smaller set of in vitro validated cases will be further examined in *Drosophila* to analyze gene expression and function, by harnessing the array of available experimental tools.

(ii) Performing downstream functional studies to determine conditions conducive to interaction partner rewiring. From the large number of rewiring events we identify, specifically, we will assess whether the

rewired interaction partners (i) share structural similarities (ii) functional similarities (iii) interact with each other. We expect that, similar to cis-regulatory rewiring, these features could individually or synergistically facilitate the opting out of one interaction partner for another. We will test these hypotheses in a manner similar to what we did in the case of cis-regulatory rewiring.

(iii) Providing an outlook on impact of partner switching on downstream gene targets. While it is possible that the two sets of gene targets are involved in similar function consistent with a neutral system drift, the targets may also be involved in very different processes, suggesting a potential switch in the regulator's function. To this end, we will determine potential gene targets of interacting TF dimers in a species-specific fashion, based on presence of DNA binding motifs of the pair of TFs in the gene regulatory regions (such as evolutionary conserved regions within 20kb of the gene start positions). We will then assess, based on Gene Ontology (GO) analysis, whether the two sets of target genes share common functions. This analysis will provide a baseline for a global trend of functional consequences of interaction rewiring, which is currently non-existent.

And finally, we report a previously unknown phenomenon that appears to have evolved in recent vertebrates, whereby an intergenic orphan CGI can function as an alternative promoter to express the gene product of a nearby CpG-poor methylated-promoter gene. We found this to occur across hundreds of CpG-poor promoter genes that become methylated in a tissue-specific fashion across 34

tissues. Importantly, this phenomenon explains the aberrant expression patterns of some cancer driver genes, potentially due to aberrant hypomethylation of distal CGIs, despite hypermethylation at proximal promoters.

From the perspective of all orphan CGIs upstream of a CpG-poor promoter gene (within 50kb), we find that almost 15% of them exhibit significant correlation (Spearman's $P < 0.05$) between CGI methylation and gene expression, and lack such a correlation (Spearman's $P > 0.05$) between the gene's proximal-promoter methylation and expression. Additionally, among all orphan CGIs exhibiting the above property, the corresponding downstream genes are significantly enriched for CpG-poor promoters (Fisher's P < 10-3) compared to CGI promoters. Even more interestingly, we find that the predominantly CpG-poor (i.e., non-CGI) promoters of MethExp genes tend to be more CG-rich than the average non-CGI promoter gene (Wilcoxon $P = 10^{-9}$). It is possible that CpG-poor MethExp promoters are remnants of once CpG-rich promoters that have lost CG dinucleotides (due to mutagenic property of methyl-cytosines) over evolutionary time; the overall impact of this phenomenon, however, remains unclear.

Further, it is not apparent what the causes or implications are, of tissue-specific genes co-opting alternative promoters for their expression.

(i) Does this co-option happen more easily for genes that possess "intermediate" CpG-dense promoters?

(ii) What abilities/advantages does the distal promoter co-option confer to the cell, or to tissue specificity itself? Does alternative CGI promoter activity confer higher than basal levels of expression to tissue-specific genes?

(iii) What are the mechanisms underlying the choice of promoter used by a given gene across cell types?

We believe the results produced as an outcome of this work warrant further investigation so as to chart out the big picture. Therefore, we plan to follow up with analyses that will help gain insights into some of these broad questions in our upcoming perspective paper in the journal, Transcription.

APPENDICES

Appendix A: Supplemental Material from Chapter 2

BRANCH 7	*LOW IC MOTIF*	*HIGH IC MOTIF*
LOW IC MOTIF (to)	-3.75	-3.70
HIGH IC MOTIF (to)	-3.74	-3.73

BRANCH 11	*LOW IC MOTIF*	*HIGH IC MOTIF*
LOW IC MOTIF (to)	-3.77	-3.74
HIGH IC MOTIF (to)	-3.77	-3.73

BRANCH 19	*LOW IC MOTIF*	*HIGH IC MOTIF*
LOW IC MOTIF (to)	-2.96	-3.01
HIGH IC MOTIF (to)	-2.88	-2.96

BRANCH 20	*LOW IC MOTIF*	*HIGH IC MOTIF*
LOW IC MOTIF (to)	-3.80	-3.78
HIGH IC MOTIF (to)	-3.80	-3.81

BRANCH 33	*LOW IC MOTIF*	*HIGH IC MOTIF*
LOW IC MOTIF (to)	-3.74	-3.72
HIGH IC MOTIF (to)	-3.74	-3.76

BRANCH 39	*LOW IC MOTIF*	*HIGH IC MOTIF*
LOW IC MOTIF (to)	-3.65	-3.67
HIGH IC MOTIF (to)	-3.61	-3.65

Figure A-1. Information content (IC) based distribution of rewiring scores per branch. In each branch, we pooled rewiring scores based on whether the rewiring is between (1) a low IC motif-to-low IC motif, (2) a low IC motif-to-high IC motif, (3) a high IC motif-to-low IC motif, or (4) a high IC motif-to-high IC motif. These tables summarize the medians of each binned category per branch.

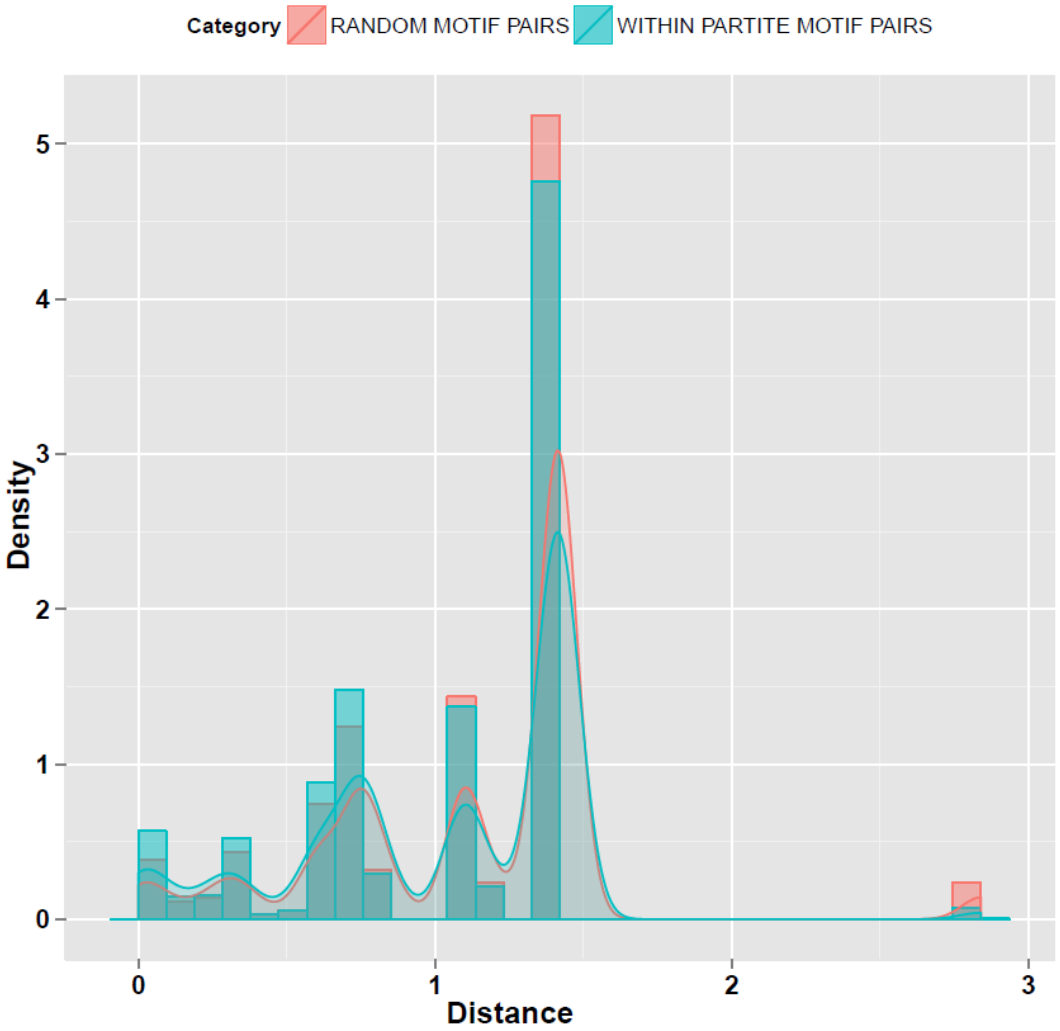


Figure A-2. Effect of motif similarity on detection of "multiplicity" of rewiring at regulons. Overlapping histograms and density profiles of distance between consensus sequences of motifs (a proxy for pairwise motif similarity) of 1) TFs rewiring at the same regulon and the same branch (in blue), and 2) random TF pairs representing the background distribution of motif similarity across TFs (in pink).

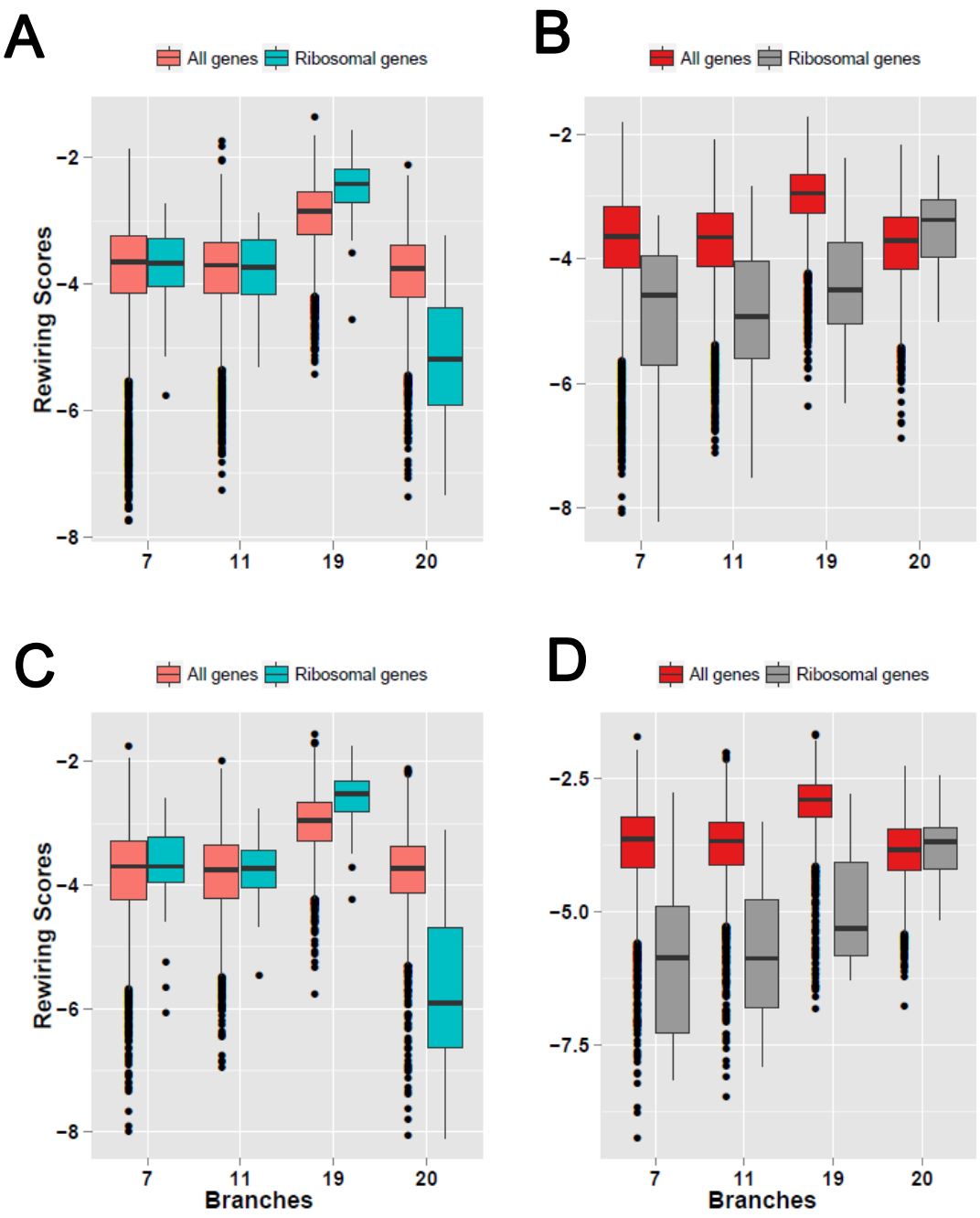
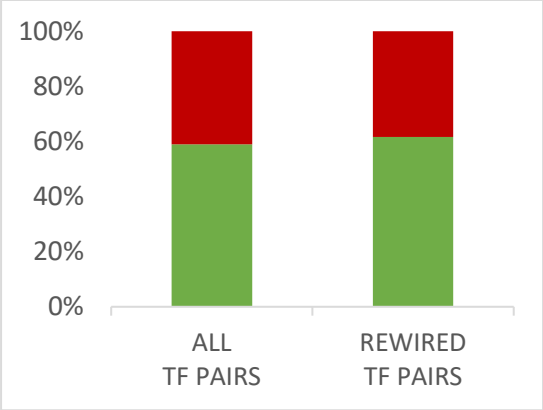


Figure A-3: Rewiring scores of the ribosomal regulon for *RAP1-CBF1* and *IFH1-TBF1* switches across branches. The rewiring scores are shown on the Y axis, and the selected branches are shown on X axis. **(A)** *RAP1 in lineage & CBF1 in ancestral species*: This plot compares the rewiring score distribution of the background (all genes; in pink) and that of ribosomal genes (in blue) for the potential that *RAP1* regulates its member genes in species diverging from a given branch *b*, and *CBF1* regulates the ancestral species. **(B)** *CBF1 in lineage & RAP1 in ancestral species*: This plot compares the rewiring score distribution of the background (in red) and that of ribosomal genes (in grey) for the potential that *CBF1* regulates its member genes in species diverging from a given branch *b*, and *RAP1* regulates the ancestral species. **(C)** & **(D)** are analogous to (A) & (B) respectively for the *IFH1-TBF1* switch in RP genes.



PHYSICAL INTERACTION POTENTIAL	
Odds Ratio	1.11
Fisher's p-value	0.14

Figure A-4. Physical interaction between rewired TFs in regulons. Rewired TF-pairs and all other TF pairs are binned into two classes based on propensity for physical interaction based on annotations in STRING database. The plot shows the fraction of TF-pairs that do (green) and do not (red) physically interact. Odds ratio (rewired versus other TF pairs) and Fisher's test p-value (2x2 table) are also shown.

COMMON UPSTREAM REGULATOR	
Odds Ratio	1.32
Fisher's p-value	0.002

Figure A-5. Proximal shared upstream regulator of rewired TFs in regulons. Rewired TF-pairs and all other TF pairs are binned into two classes based on their distance to a common upstream regulator. This plot shows the fraction of TF-pairs whose distance to a common UpR is ≤ 4 (green) or > 4 (red). Odds ratio (rewired versus other TF pairs) and Fisher's test p-value (2x2 table) are also shown.

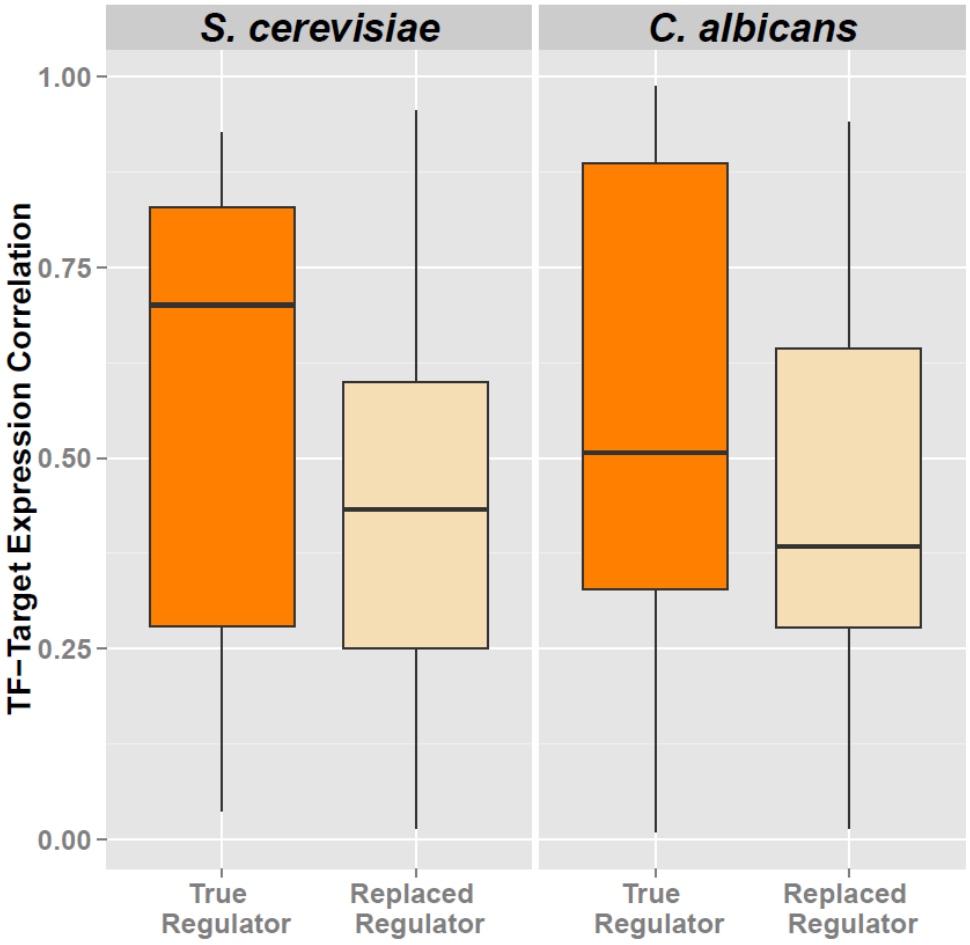


Figure A-6. Species-specific TF-target co-expression for rewiring events detected at branch 11. The predicted TF-target expression distribution is shown for the TF predicted to be active in a species (dark) and for the TF predicted not be active in the species (light); the distribution is based on pooled correlations across all significant rewiring events.

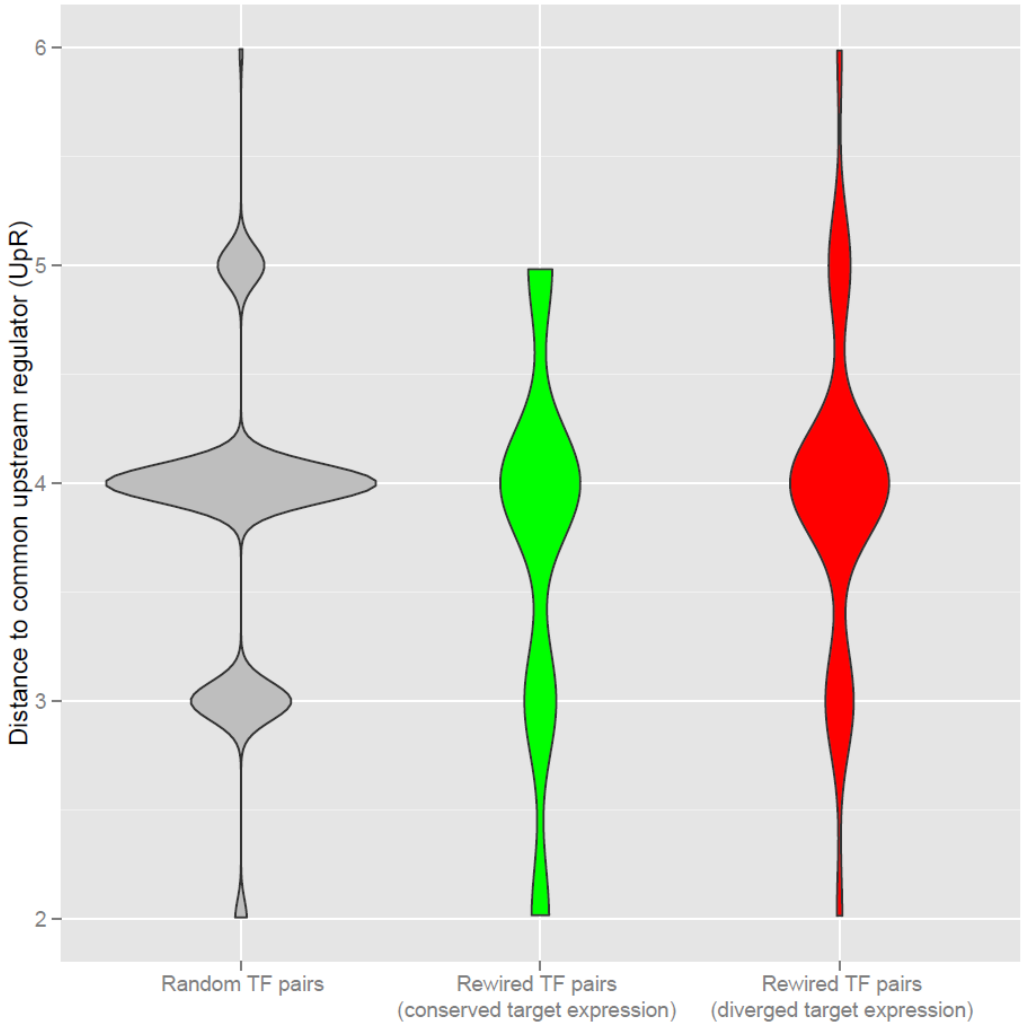


Figure A-7. Distributions of distance (path length) of rewiring TF pairs to their common UpR across species, in different bins of rewiring events segregated based on conservation levels of gene-targets expression. The target genes of rewiring were binned into conserved (in green) and diverged (in red) expression groups, and their respective distributions of distance to a common upstream regulator is shown in violin plots. As background expectation, the distribution of distance to common UpR for random TF-pairs (in grey) is also shown.

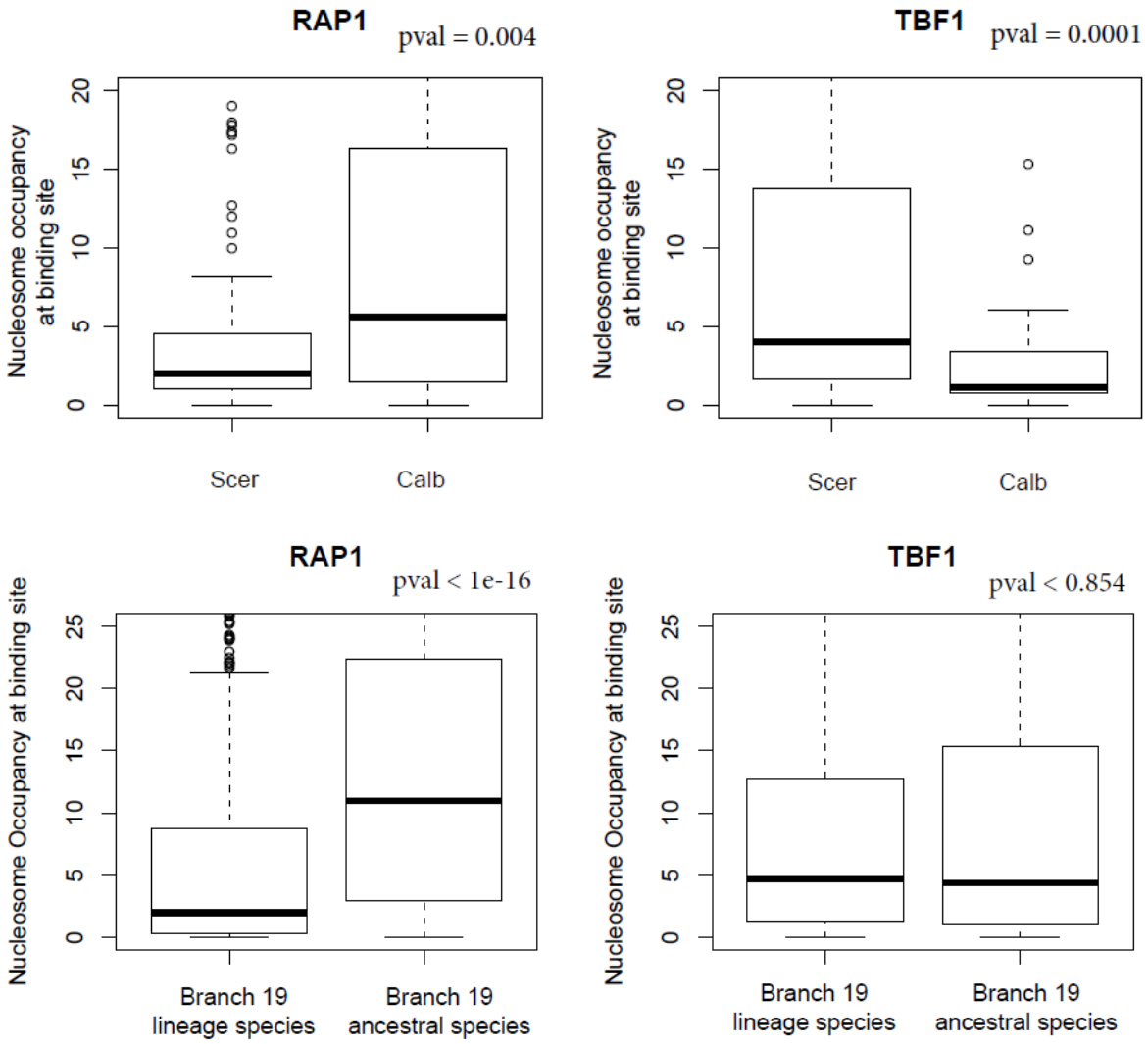


Figure A-8. Nucleosome occupancy of RAP1 and TBF1 at ribosomal genes. Boxplots showing the distributions of nucleosome occupancy scores (Y axis) at TF binding sites in different yeast species (X axis). **(A)** Nuc. Occupancy at RAP1 sites in *Scer* vs. *Calb*. **(B)** Nuc. Occupancy at TBF1 sites in *Scer* vs. *Calb*. **(C)** Nuc. Occupancy at RAP1 sites in lineage vs. ancestral species. **(D)** Nuc. Occupancy at TBF1 sites in lineage vs. ancestral species. If the known rewiring between *RAP1* and *TBF1* in RP genes is supported by nucleosome occupancy, we expect to see lower nucleosome occupancy for *RAP1* sites in *Scer* and lineage-specific species, compared with that in *Calb* and ancestral species; and vice-versa for TBF1 sites. We find that although these trends are individually consistent in *Scer* and *Calb* (Fig. S8-A, B), we did not observe these patterns in the nucleosome occupancy profiles of *RAP1* vs. *TBF1* binding sites in other species (Fig. S8-C, D) (i.e., there is absence of support across clades despite the fact that this rewiring event spans promoters across all species).

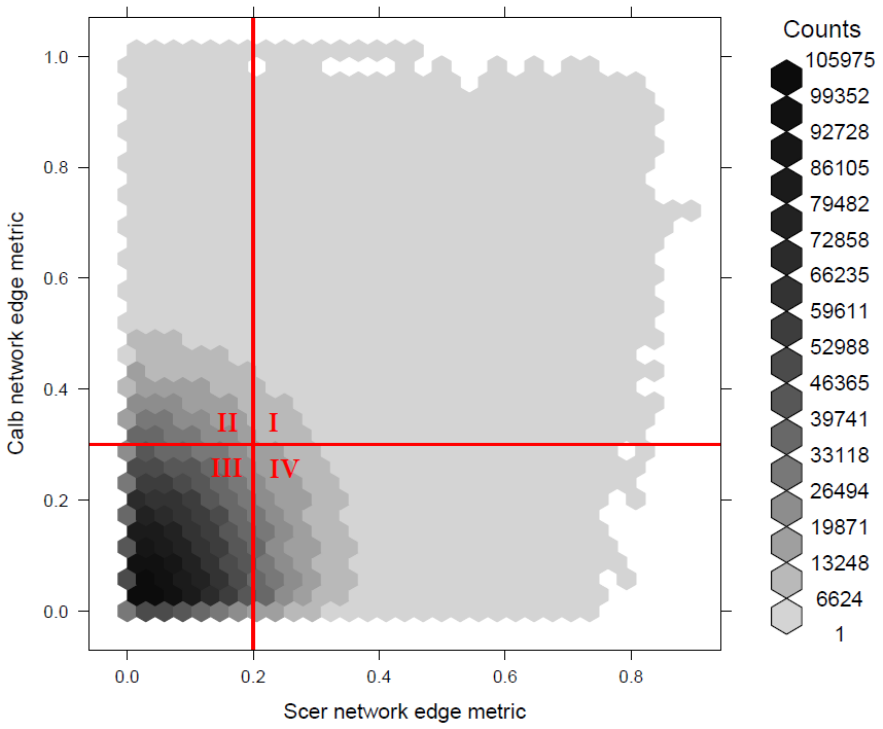


Figure A-9. Density scatter plot of edges (representing expression correlation) between pairs of genes in *Scer* and *Calb* co-expression networks. Each point represents the level of co-expression of a pair of orthologous genes in the *Scer* coexpression network (X axis) and its corresponding level in the *Calb* coexpression network (Y axis) respectively. Datapoints are clustered into hexagons and colored based on point density. All points in quadrant III were discarded from further analyses.

#Significant rewiring events at the gene level for various FDR thresholds						
Branch → / FDR ↓	**7**	**11**	**19**	**20**	**33**	**39**
0.01	1	22	38	186	4	2
0.02	1	38	46	552	4	2
0.05	1	81	120	1001	4	2
0.1	1	164	228	1043	8	2

Figure A-10. Significant rewiring events at the gene level for various FDR thresholds. The number of rewiring events occurring across internal branch partitions at various FDR thresholds.

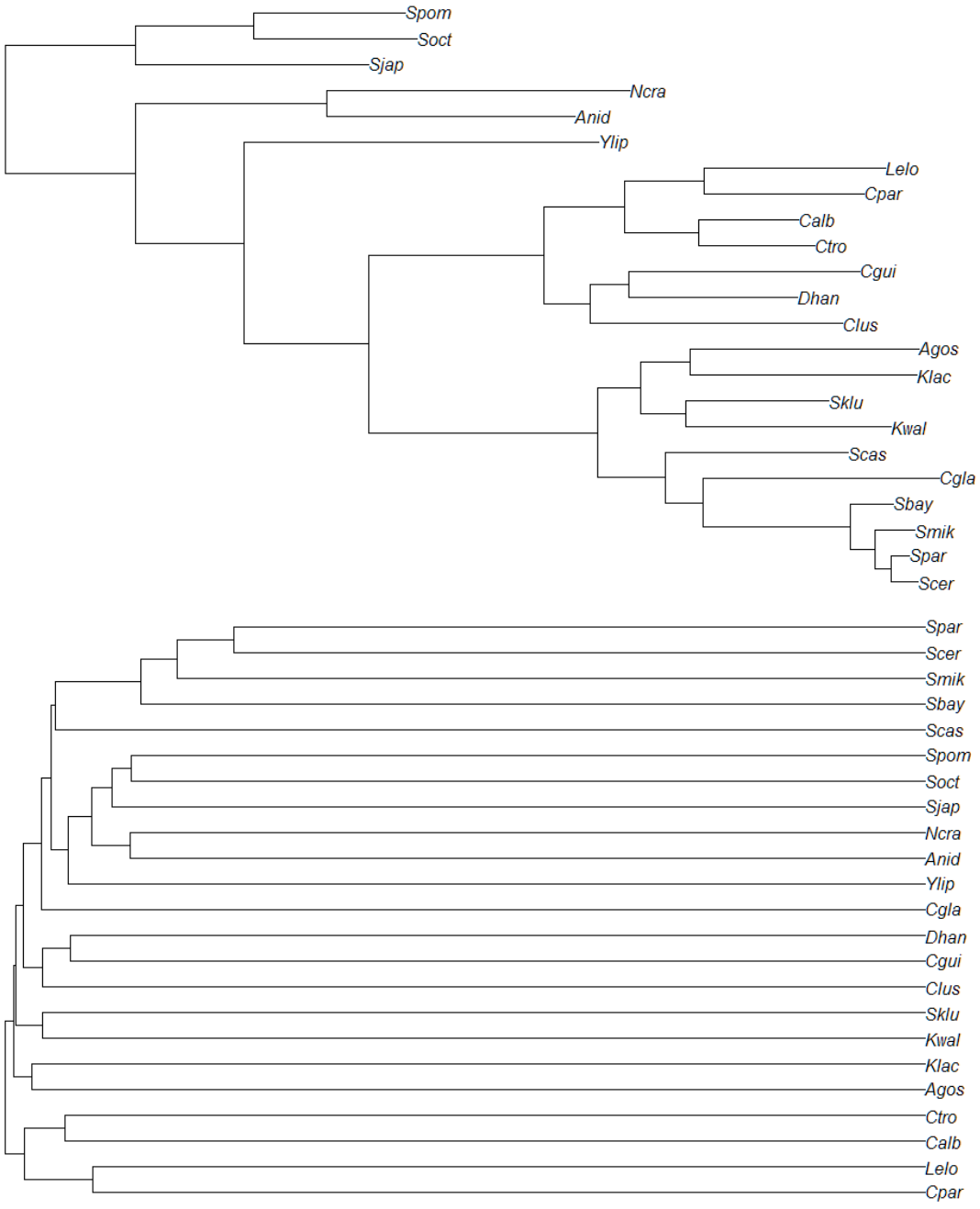
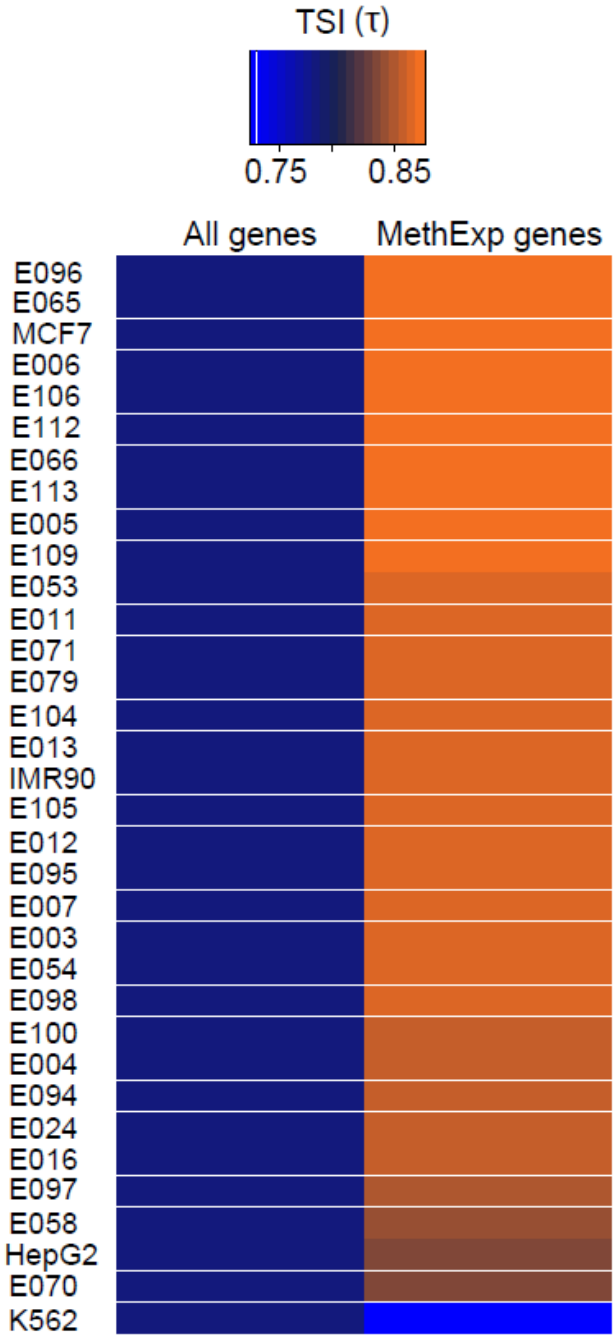


Figure A-11. Yeast species phylogeny trees inferred from aligned genomic blocks and TF binding profiles. Phylogenetic relationships between 23 yeast species with relative branch lengths. **(A)** Known species phylogeny from genomic sequence, **(B)** Phylogeny derived from clustering binding probability profiles of 176 TFs across orthologous promoters in 23 species.

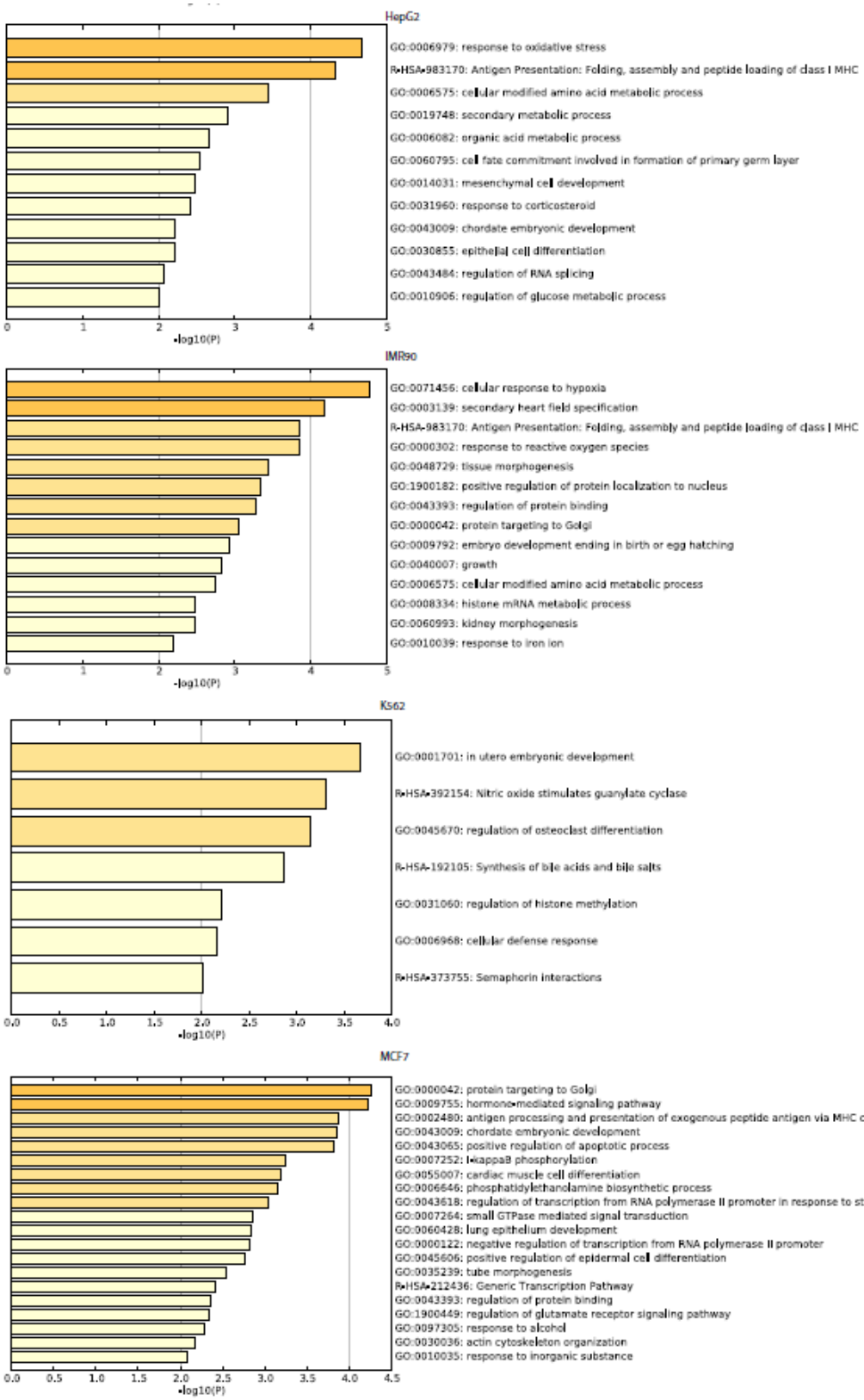
Appendix B: Supplemental Material from Chapter 3



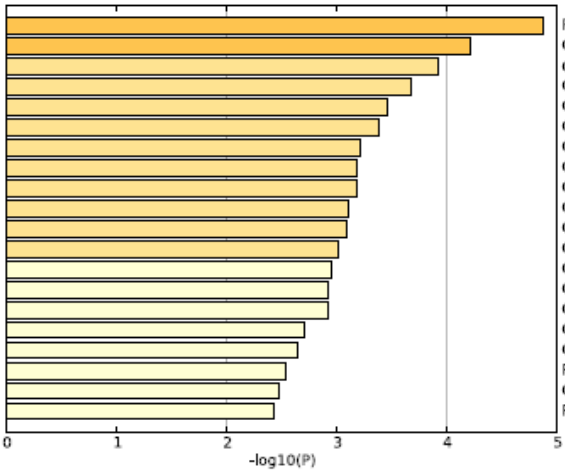
Supplemental Figure B-1. Heatmap representation of the median tissue specificity index (TSI) of MethExp genes versus all genes (columns) across 34 tissues (rows). A color scale for TSI values is provided above the heatmap

Supplemental Figure B-2a. Stacked barplots displaying the proportion (X-axes) of three gene subtypes, viz., protein-coding (green), long-noncoding (pink), and short-noncoding (blue) before and after applying a filtration step (Y-axes) to MethExp genes, that discards all loci that contain a neighboring gene spanning the region between them and their associated upstream CGIs. Application of the filter resulted in no differences in the proportions of these gene subtypes across all 34 tissue types.

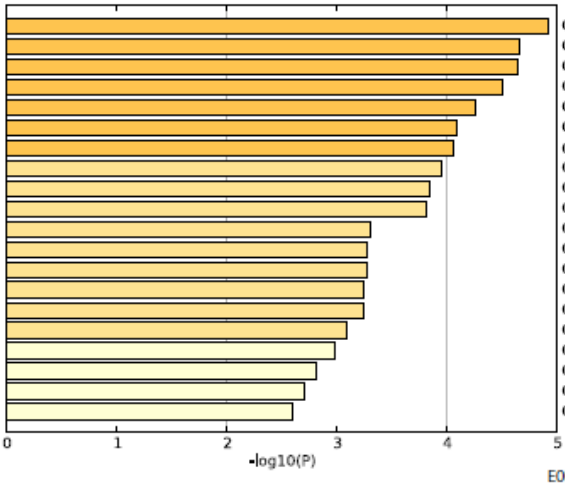
Supplemental Figure B-2b. Boxplots showing the distribution of expression levels (Y-axes) of MethExp genes before and after a filtration step (X-axes) that discards all loci that contain a neighboring gene spanning the region between them and their associated upstream CGIs. Application of the filter resulted in no differences in the overall expression levels of MethExp genes across 34 tissue types.


HepG2
GO:0006979: response to oxidative stress
R-HSA-983170: Antigen Presentation: Folding, assembly and peptide loading of class I MHC
GO:0006575: cellular modified amino acid metabolic process
GO:0019748: secondary metabolic process
GO:0006082: organic acid metabolic process
GO:0060795: cell fate commitment involved in formation of primary germ layer
GO:0014031: mesenchymal cell development
GO:0031960: response to corticosteroid
GO:0043009: chordate embryonic development
GO:0030855: epithelial cell differentiation
GO:0043484: regulation of RNA splicing
GO:0010906: regulation of glucose metabolic process
0 1 2 3 4 5
-log10(P)
IMR90
GO:0071456: cellular response to hypoxia
GO:0003139: secondary heart field specification
R-HSA-983170: Antigen Presentation: Folding, assembly and peptide loading of class I MHC
GO:0000302: response to reactive oxygen species
GO:0048729: tissue morphogenesis
GO:1900182: positive regulation of protein localization to nucleus
GO:0043393: regulation of protein binding
GO:0000042: protein targeting to Golgi
GO:0009792: embryo development ending in birth or egg hatching
GO:0040007: growth
GO:0006575: cellular modified amino acid metabolic process
GO:0008334: histone mRNA metabolic process
GO:0060993: kidney morphogenesis
GO:0010039: response to iron ion
0 1 2 3 4 5
-log10(P)
K562
GO:0001701: in utero embryonic development
R-HSA-392154: Nitric oxide stimulates guanylate cyclase
GO:0045670: regulation of osteoclast differentiation
R-HSA-192105: Synthesis of bile acids and bile salts
GO:0031060: regulation of histone methylation
GO:0006968: cellular defense response
R-HSA-373755: Semaphorin interactions
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
-log10(P)
MCF7
GO:0000042: protein targeting to Golgi
GO:0009755: hormone-mediated signaling pathway
GO:0002480: antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent
GO:0043009: chordate embryonic development
GO:0043065: positive regulation of apoptotic process
GO:0007252: I-kappaB phosphorylation
GO:0055007: cardiac muscle cell differentiation
GO:0006646: phosphatidylethanolamine biosynthetic process
GO:0043618: regulation of transcription from RNA polymerase II promoter in response to stress
GO:0007264: small GTPase mediated signal transduction
GO:0060428: lung epithelium development
GO:0000122: negative regulation of transcription from RNA polymerase II promoter
GO:0045606: positive regulation of epidermal cell differentiation
GO:0035239: tube morphogenesis
R-HSA-212436: Generic Transcription Pathway
GO:0043393: regulation of protein binding
GO:1900449: regulation of glutamate receptor signaling pathway
GO:0097305: response to alcohol
GO:0030036: actin cytoskeleton organization
GO:0010035: response to inorganic substance
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5
-log10(P)

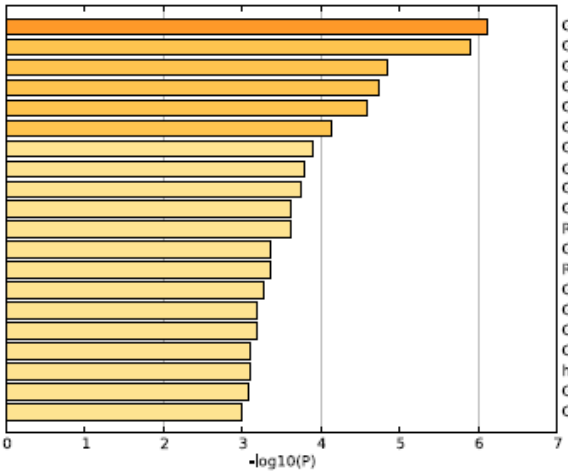
E003



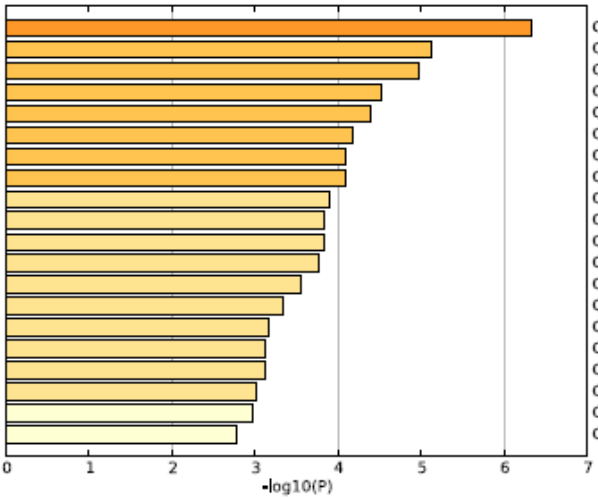
E004



E005

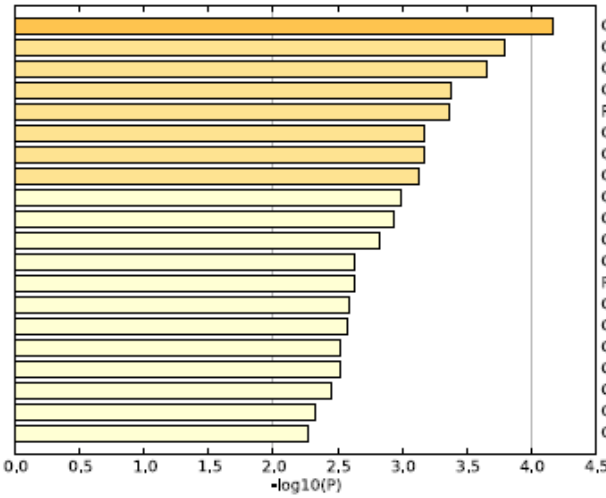


E006



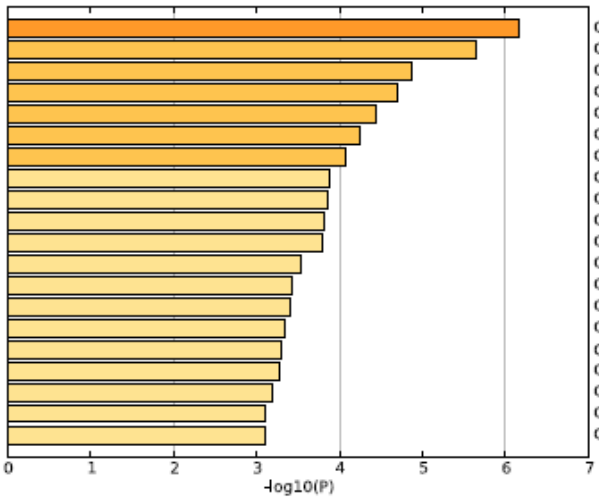
GO:0043393: regulation of protein binding
GO:1900182: positive regulation of protein localization to nucleus
GO:0006575: cellular modified amino acid metabolic process
GO:0043620: regulation of DNA-templated transcription in response to stress
GO:0000302: response to reactive oxygen species
GO:0009792: embryo development ending in birth or egg hatching
GO:0009725: response to hormone
GO:0061061: muscle structure development
GO:0019370: leukotriene biosynthetic process
GO:0006954: inflammatory response
GO:0061138: morphogenesis of a branching epithelium
GO:0040007: growth
GO:0048193: Golgi vesicle transport
GO:0032411: positive regulation of transporter activity
GO:0051897: positive regulation of protein kinase B signaling
GO:0071243: cellular response to arsenic-containing substance
GO:0007252: I-kappaB phosphorylation
GO:0032075: positive regulation of nuclease activity
GO:1902532: negative regulation of intracellular signal transduction
GO:0000042: protein targeting to Golgi

E007



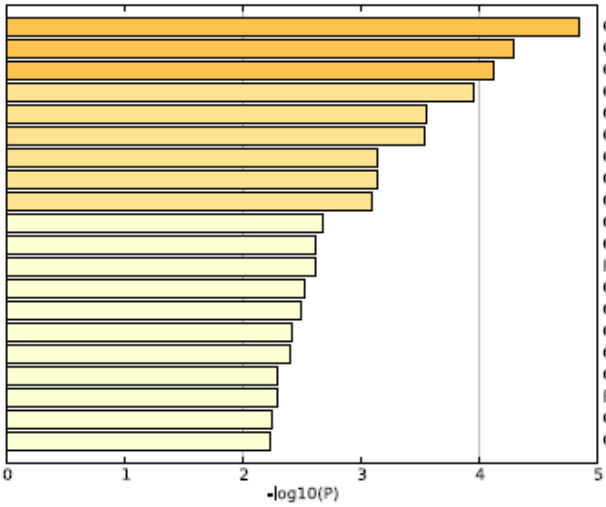
GO:0043393: regulation of protein binding
GO:0043457: regulation of cellular respiration
GO:0043620: regulation of DNA-templated transcription in response to stress
GO:0061061: muscle structure development
R-HSA-8849471: PTK6 Regulates RHO GTPases, RAS GTPase and MAP kinases
GO:0007252: I-kappaB phosphorylation
GO:0071243: cellular response to arsenic-containing substance
GO:0008625: extrinsic apoptotic signaling pathway via death domain receptors
GO:0060135: maternal process involved in female pregnancy
GO:0042542: response to hydrogen peroxide
GO:0000042: protein targeting to Golgi
GO:0048193: Golgi vesicle transport
R-HSA-6794362: Protein-protein interactions at synapses
GO:0040007: growth
GO:0032411: positive regulation of transporter activity
GO:0046942: carboxylic acid transport
GO:0070076: histone lysine demethylation
GO:0001837: epithelial to mesenchymal transition
GO:0006575: cellular modified amino acid metabolic process
GO:0008334: histone mRNA metabolic process

E011



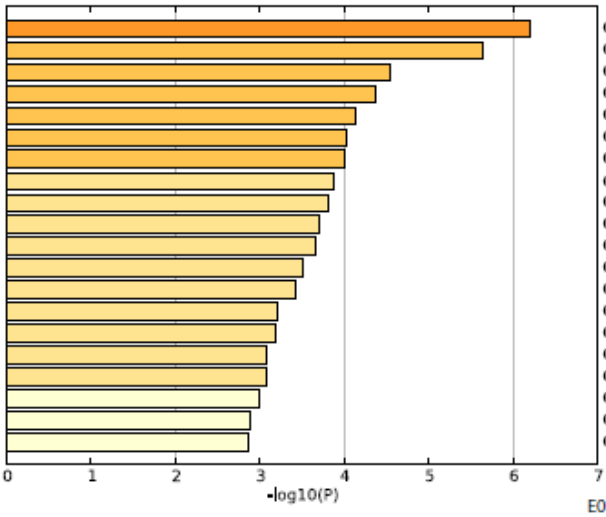
GO:0032411: positive regulation of transporter activity
GO:0003139: secondary heart field specification
GO:0014706: striated muscle tissue development
GO:0043393: regulation of protein binding
GO:0040007: growth
GO:1900182: positive regulation of protein localization to nucleus
GO:0009792: embryo development ending in birth or egg hatching
GO:1901339: regulation of store-operated calcium channel activity
GO:0036294: cellular response to decreased oxygen levels
GO:0043401: steroid hormone mediated signaling pathway
GO:0060135: maternal process involved in female pregnancy
GO:0006979: response to oxidative stress
GO:0090009: primitive streak formation
GO:0072378: blood coagulation, fibrin clot formation
GO:0000226: microtubule cytoskeleton organization
GO:0007610: behavior
GO:1901698: response to nitrogen compound
GO:0043065: positive regulation of apoptotic process
GO:0071243: cellular response to arsenic-containing substance
GO:0007252: I-kappaB phosphorylation

E012



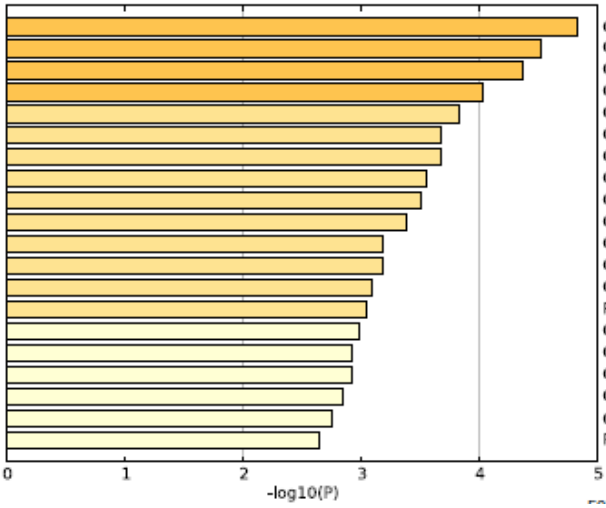
GO:0043393: regulation of protein binding
GO:0032411: positive regulation of transporter activity
GO:0000042: protein targeting to Golgi
GO:0003206: cardiac chamber morphogenesis
GO:0046942: carboxylic acid transport
GO:0040007: growth
GO:0007252: I-kappaB phosphorylation
GO:0071243: cellular response to arsenic-containing substance
GO:0008625: extrinsic apoptotic signaling pathway via death domain receptors
GO:0009755: hormone-mediated signaling pathway
GO:0071731: response to nitric oxide
R-HSA-8849932: SALM protein interactions at the synapses
GO:0007626: locomotory behavior
GO:0070076: histone lysine demethylation
GO:0043623: cellular protein complex assembly
GO:0010469: regulation of receptor activity
GO:0006575: cellular modified amino acid metabolic process
R-HSA-2132295: MHC class II antigen presentation
GO:0008334: histone mRNA metabolic process
GO:0042177: negative regulation of protein catabolic process

E013



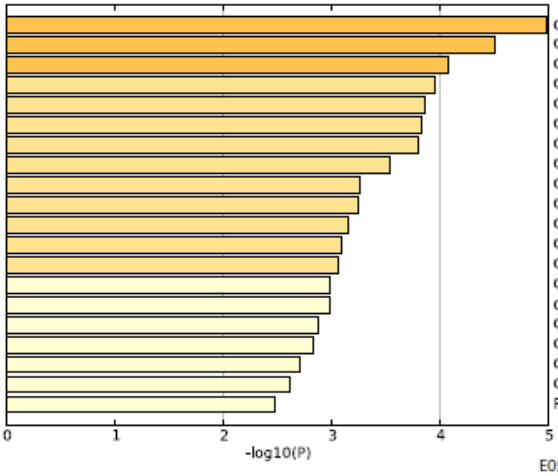
GO:0043393: regulation of protein binding
GO:0003139: secondary heart field specification
GO:0061383: trabecula morphogenesis
GO:0009755: hormone-mediated signaling pathway
GO:0032409: regulation of transporter activity
GO:0043065: positive regulation of apoptotic process
GO:0009792: embryo development ending in birth or egg hatching
GO:0048820: hair follicle maturation
GO:0008625: extrinsic apoptotic signaling pathway via death domain receptors
GO:0051601: exocyst localization
GO:0060389: pathway-restricted SMAD protein phosphorylation
GO:0043620: regulation of DNA-templated transcription in response to stress
GO:0006575: cellular modified amino acid metabolic process
GO:0040007: growth
GO:0097067: cellular response to thyroid hormone stimulus
GO:0071243: cellular response to arsenic-containing substance
GO:0007252: I-kappaB phosphorylation
GO:0032075: positive regulation of nuclease activity
GO:0051129: negative regulation of cellular component organization
GO:0060135: maternal process involved in female pregnancy

E016



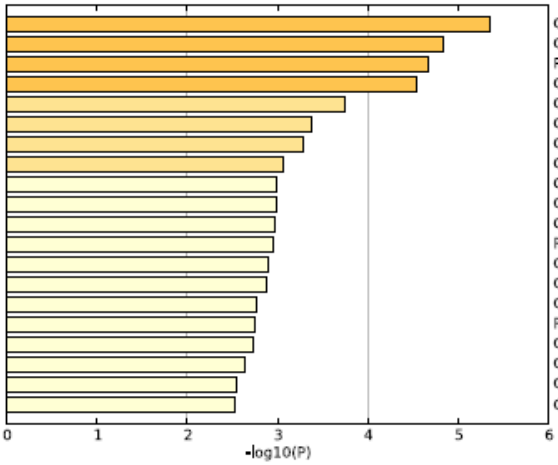
GO:0055007: cardiac muscle cell differentiation
GO:0040007: growth
GO:0032411: positive regulation of transporter activity
GO:0043401: steroid hormone mediated signaling pathway
GO:0006979: response to oxidative stress
GO:0043620: regulation of DNA-templated transcription in response to stress
GO:1900182: positive regulation of protein localization to nucleus
GO:0032868: response to insulin
GO:0009792: embryo development ending in birth or egg hatching
GO:0060391: positive regulation of SMAD protein import into nucleus
GO:0071243: cellular response to arsenic-containing substance
GO:0007252: I-kappaB phosphorylation
GO:0032075: positive regulation of nuclease activity
R-HSA-2132295: MHC class II antigen presentation
GO:0006575: cellular modified amino acid metabolic process
GO:0035282: segmentation
GO:0008356: asymmetric cell division
GO:0000042: protein targeting to Golgi
GO:0007610: behavior
R-HSA-6794362: Protein-protein interactions at synapses

E024



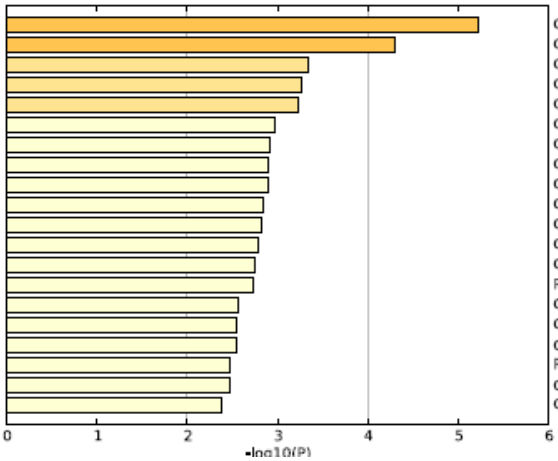
GO:0055007: cardiac muscle cell differentiation
GO:0032411: positive regulation of transporter activity
GO:0006979: response to oxidative stress
GO:0007565: female pregnancy
GO:0046942: carboxylic acid transport
GO:0006575: cellular modified amino acid metabolic process
GO:0043620: regulation of DNA-templated transcription in response to stress
GO:0009755: hormone-mediated signaling pathway
GO:0008625: extrinsic apoptotic signaling pathway via death domain receptors
GO:0007252: I-kappaB phosphorylation
GO:0032075: positive regulation of nuclease activity
GO:0060795: cell fate commitment involved in formation of primary germ layer
GO:1900182: positive regulation of protein localization to nucleus
GO:0008356: asymmetric cell division
GO:0007610: behavior
GO:0009792: embryo development ending in birth or egg hatching
GO:0031344: regulation of cell projection organization
GO:0071731: response to nitric oxide
GO:0006665: sphingolipid metabolic process
R-HSA-983170: Antigen Presentation: Folding, assembly and peptide loading of class I MHC

E053



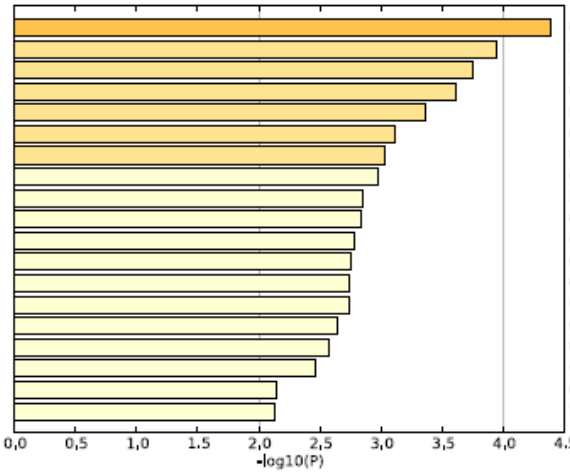
GO:0071456: cellular response to hypoxia
GO:0060537: muscle tissue development
R-HSA-6794362: Protein-protein interactions at synapses
GO:0010035: response to inorganic substance
GO:0009725: response to hormone
GO:0007565: female pregnancy
GO:0071243: cellular response to arsenic-containing substance
GO:1901698: response to nitrogen compound
GO:0043484: regulation of RNA splicing
GO:0007613: memory
GO:0046889: positive regulation of lipid biosynthetic process
R-HSA-556833: Metabolism of lipids and lipoproteins
GO:0001890: placenta development
GO:0006103: 2-oxoglutarate metabolic process
GO:0040007: growth
R-HSA-8849932: SALM protein interactions at the synapses
GO:0010942: positive regulation of cell death
GO:0010469: regulation of receptor activity
GO:0006575: cellular modified amino acid metabolic process
GO:0060343: trabecula formation

E054



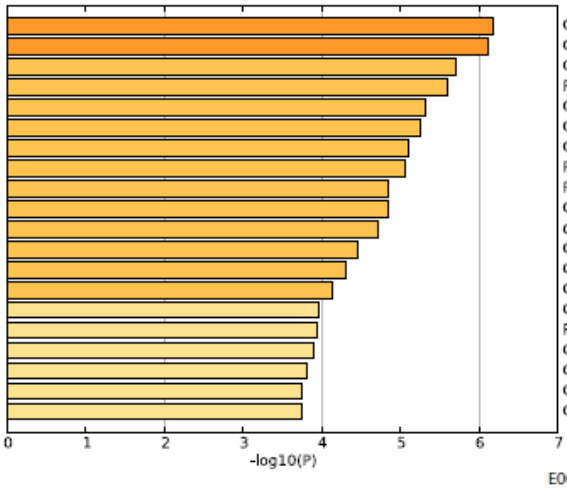
GO:0071456: cellular response to hypoxia
GO:0048738: cardiac muscle tissue development
GO:0010035: response to inorganic substance
GO:0007565: female pregnancy
GO:0007252: I-kappaB phosphorylation
GO:0006766: vitamin metabolic process
GO:0046889: positive regulation of lipid biosynthetic process
GO:0040007: growth
GO:0043484: regulation of RNA splicing
GO:1901698: response to nitrogen compound
GO:0000302: response to reactive oxygen species
GO:0008202: steroid metabolic process
GO:0072358: cardiovascular system development
R-HSA-6794362: Protein-protein interactions at synapses
GO:0051098: regulation of binding
GO:0010469: regulation of receptor activity
GO:0007610: behavior
R-HSA-983170: Antigen Presentation: Folding, assembly and peptide loading of class I MHC
GO:0060343: trabecula formation
GO:0071407: cellular response to organic cyclic compound

E058



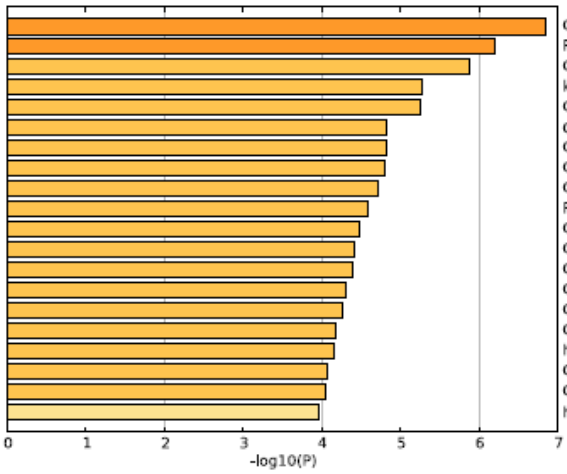
GO:0006575: cellular modified amino acid metabolic process
R-HSA-983170: Antigen Presentation: Folding, assembly and peptide loading of class I MHC
R-HSA-8849471: PTK6 Regulates RHO GTPases, RAS GTPase and MAP kinases
GO:0060537: muscle tissue development
hsa00350: Tyrosine metabolism
GO:0007565: female pregnancy
GO:0006665: sphingolipid metabolic process
GO:0032331: negative regulation of chondrocyte differentiation
R-HSA-556833: Metabolism of lipids and lipoproteins
GO:0045596: negative regulation of cell differentiation
GO:1900180: regulation of protein localization to nucleus
GO:2001235: positive regulation of apoptotic signaling pathway
GO:0048736: appendage development
GO:0043393: regulation of protein binding
GO:0051181: cofactor transport
GO:1901654: response to ketone
GO:0010259: multicellular organism aging
GO:0015711: organic anion transport
R-HSA-8864260: Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors

E065



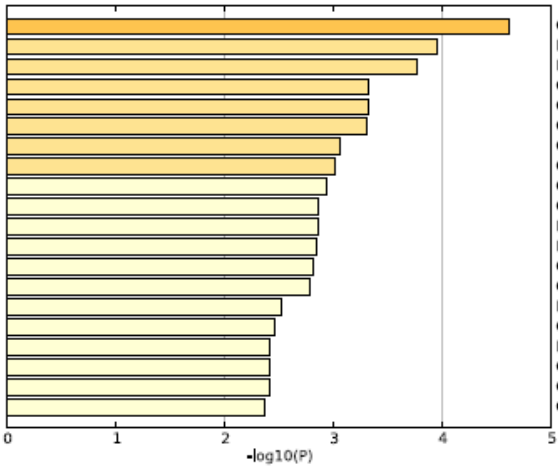
GO:0033559: unsaturated fatty acid metabolic process
GO:0006575: cellular modified amino acid metabolic process
GO:0045087: innate immune response
R-HSA-198933: Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
GO:0042742: defense response to bacterium
GO:0070372: regulation of ERK1 and ERK2 cascade
GO:0010035: response to inorganic substance
R-HSA-556833: Metabolism of lipids and lipoproteins
R-HSA-5686938: Regulation of TLR by endogenous ligand
GO:0043065: positive regulation of apoptotic process
GO:0055007: cardiac muscle cell differentiation
GO:0009725: response to hormone
GO:0042307: positive regulation of protein import into nucleus
GO:0006820: anion transport
GO:0051345: positive regulation of hydrolase activity
R-HSA-6798695: Neutrophil degranulation
GO:0008285: negative regulation of cell proliferation
GO:0043620: regulation of DNA-templated transcription in response to stress
GO:0043393: regulation of protein binding
GO:0036294: cellular response to decreased oxygen levels

E066



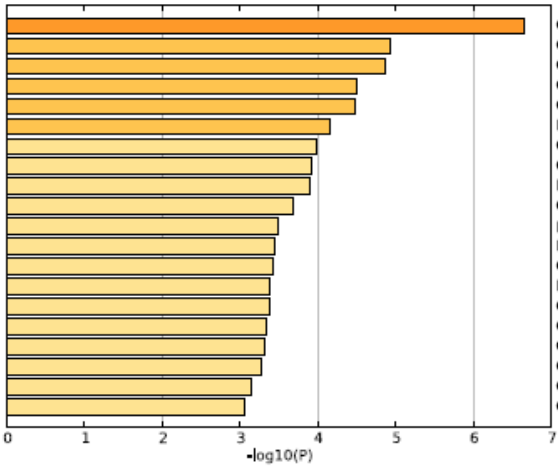
GO:0006954: inflammatory response
R-HSA-556833: Metabolism of lipids and lipoproteins
GO:0050776: regulation of immune response
ko00140: Steroid hormone biosynthesis
GO:0045321: leukocyte activation
GO:0055007: cardiac muscle cell differentiation
GO:0010035: response to inorganic substance
GO:0001816: cytokine production
GO:0009725: response to hormone
R-HSA-6798695: Neutrophil degranulation
GO:0051707: response to other organism
GO:0050900: leukocyte migration
GO:0042307: positive regulation of protein import into nucleus
GO:0097067: cellular response to thyroid hormone stimulus
GO:0015849: organic acid transport
GO:0002274: myeloid leukocyte activation
hsa05204: Chemical carcinogenesis
GO:0006575: cellular modified amino acid metabolic process
GO:0000302: response to reactive oxygen species
hsa00591: Linoleic acid metabolism

E070



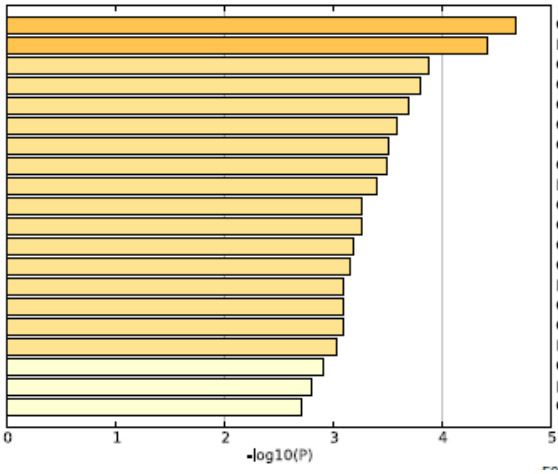
GO:0071456: cellular response to hypoxia
R-HSA-556833: Metabolism of lipids and lipoproteins
R-HSA-2142691: Synthesis of Leukotrienes (LT) and Eoxins (EX)
GO:0071243: cellular response to arsenic-containing substance
GO:0007252: I-kappaB phosphorylation
GO:0006575: cellular modified amino acid metabolic process
GO:0043484: regulation of RNA splicing
GO:0070206: protein trimerization
GO:0034101: erythrocyte homeostasis
GO:0031345: negative regulation of cell projection organization
R-HSA-6794362: Protein-protein interactions at synapses
R-HSA-163560: Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis
GO:0014706: striated muscle tissue development
GO:0071731: response to nitric oxide
R-HSA-2029480: Fcgamma receptor (FCGR) dependent phagocytosis
GO:0010942: positive regulation of cell death
R-HSA-8849471: PTK6 Regulates RHO GTPases, RAS GTPase and MAP kinases
GO:0008334: histone mRNA metabolic process
GO:0010092: specification of organ identity
GO:0070528: protein kinase C signaling

E071



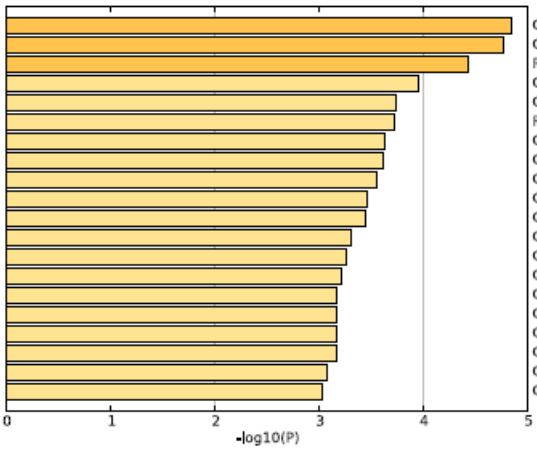
GO:0010035: response to inorganic substance
GO:1902041: regulation of extrinsic apoptotic signaling pathway via death domain receptors
GO:0045087: innate immune response
GO:0006082: organic acid metabolic process
GO:0036294: cellular response to decreased oxygen levels
R-HSA-6794362: Protein-protein interactions at synapses
GO:0051100: negative regulation of binding
GO:0051223: regulation of protein transport
ko04610: Complement and coagulation cascades
GO:0033559: unsaturated fatty acid metabolic process
R-HSA-194068: Bile acid and bile salt metabolism
R-HSA-168898: Toll-Like Receptors Cascades
GO:0002274: myeloid leukocyte activation
R-HSA-1236977: Endosomal/Vacuolar pathway
GO:0015849: organic acid transport
GO:0009725: response to hormone
GO:0007613: memory
GO:0050663: cytokine secretion
GO:0097067: cellular response to thyroid hormone stimulus
GO:0007252: I-kappaB phosphorylation

E079



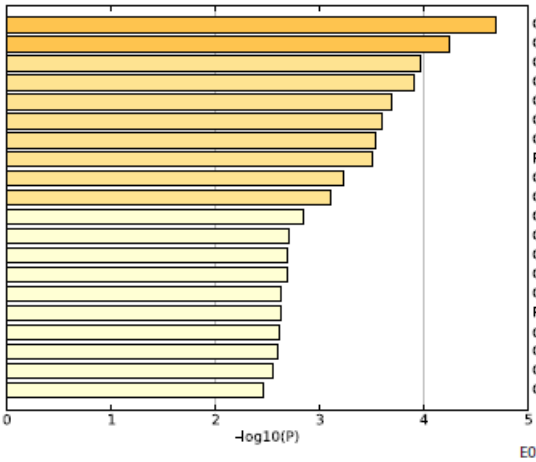
GO:0048608: reproductive structure development
R-HSA-556833: Metabolism of lipids and lipoproteins
GO:0003139: secondary heart field specification
GO:0014031: mesenchymal cell development
GO:0010882: regulation of cardiac muscle contraction by calcium ion signaling
GO:0022612: gland morphogenesis
GO:0060341: regulation of cellular localization
GO:0061383: trabecula morphogenesis
R-HSA-6798695: Neutrophil degranulation
GO:0043393: regulation of protein binding
GO:0071456: cellular response to hypoxia
GO:0097067: cellular response to thyroid hormone stimulus
GO:0009725: response to hormone
hsa04514: Cell adhesion molecules (CAMs)
GO:0071243: cellular response to arsenic-containing substance
GO:0007252: I-kappaB phosphorylation
R-HSA-3906995: Diseases associated with O-glycosylation of proteins
GO:0071496: cellular response to external stimulus
hsa00590: Arachidonic acid metabolism
GO:0070206: protein trimerization

E094



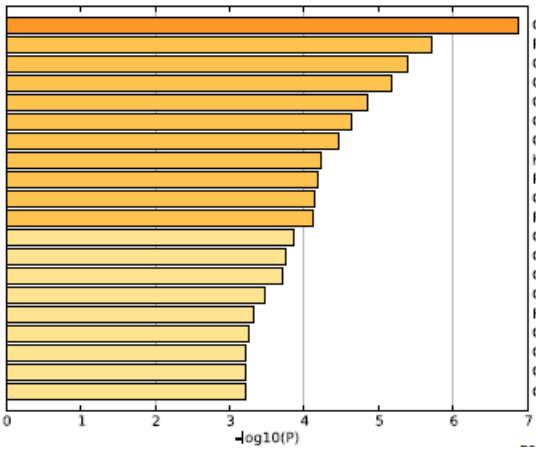
GO:0048738: cardiac muscle tissue development
GO:0042306: regulation of protein import into nucleus
R-HSA-556833: Metabolism of lipids and lipoproteins
GO:0003139: secondary heart field specification
GO:1990266: neutrophil migration
R-HSA-6798695: Neutrophil degranulation
GO:0006575: cellular modified amino acid metabolic process
GO:0006954: inflammatory response
GO:0006766: vitamin metabolic process
GO:0043393: regulation of protein binding
GO:0045087: innate immune response
GO:0043547: positive regulation of GTPase activity
GO:0097067: cellular response to thyroid hormone stimulus
GO:0008285: negative regulation of cell proliferation
GO:0009725: response to hormone
GO:0071243: cellular response to arsenic-containing substance
GO:0007252: I-kappaB phosphorylation
GO:0042448: progesterone metabolic process
GO:0045603: positive regulation of endothelial cell differentiation
GO:0043065: positive regulation of apoptotic process

E095



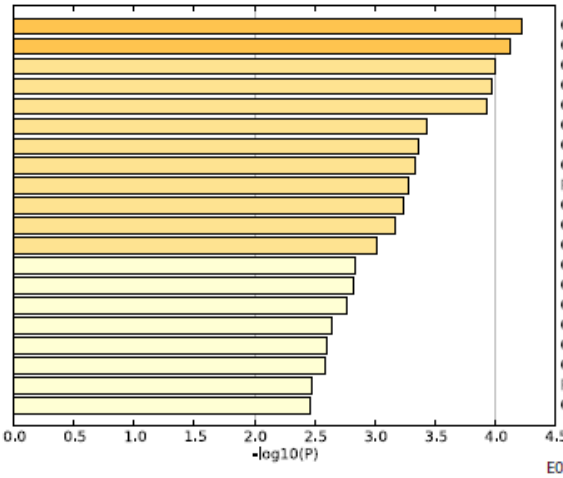
GO:0097067: cellular response to thyroid hormone stimulus
GO:0032409: regulation of transporter activity
GO:0048738: cardiac muscle tissue development
GO:0003139: secondary heart field specification
GO:0009725: response to hormone
GO:0002274: myeloid leukocyte activation
GO:1900182: positive regulation of protein localization to nucleus
R-HSA-983170: Antigen Presentation: Folding, assembly and peptide loading of class I MHC
GO:0032370: positive regulation of lipid transport
GO:0071243: cellular response to arsenic-containing substance
GO:0043618: regulation of transcription from RNA polymerase II promoter in response to stress
GO:0003215: cardiac right ventricle morphogenesis
GO:0043393: regulation of protein binding
GO:0043484: regulation of RNA splicing
GO:0006979: response to oxidative stress
R-HSA-6798695: Neutrophil degranulation
GO:0055072: iron ion homeostasis
GO:0006954: inflammatory response
GO:0060711: labyrinthine layer development
GO:0070076: histone lysine demethylation

E096

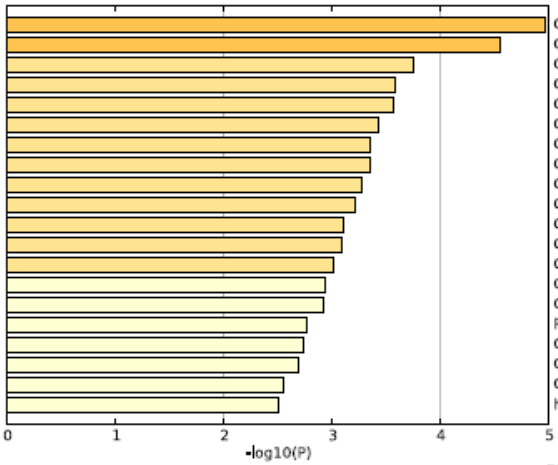


GO:0006575: cellular modified amino acid metabolic process
R-HSA-5083625: Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)
GO:0051222: positive regulation of protein transport
GO:0050663: cytokine secretion
GO:0045087: innate immune response
GO:0002274: myeloid leukocyte activation
GO:0009725: response to hormone
hsa04514: Cell adhesion molecules (CAMs)
R-HSA-556833: Metabolism of lipids and lipoproteins
GO:0042742: defense response to bacterium
R-HSA-391160: Signal regulatory protein (SIRP) family interactions
GO:0007565: female pregnancy
GO:0003139: secondary heart field specification
GO:1903010: regulation of bone development
GO:1904037: positive regulation of epithelial cell apoptotic process
R-HSA-983170: Antigen Presentation: Folding, assembly and peptide loading of class I MHC
GO:0072378: blood coagulation, fibrin clot formation
GO:0010035: response to inorganic substance
GO:0008285: negative regulation of cell proliferation
GO:0055072: iron ion homeostasis

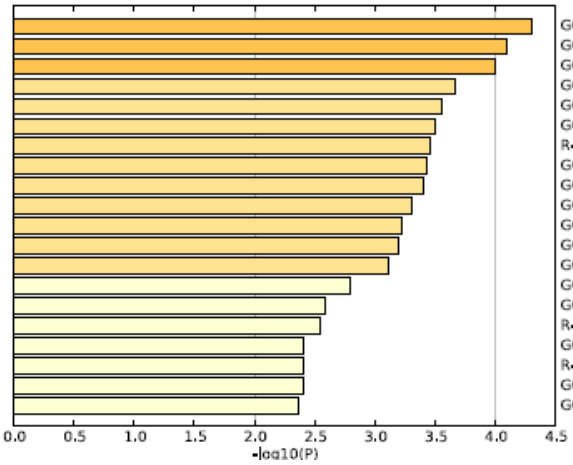
E097



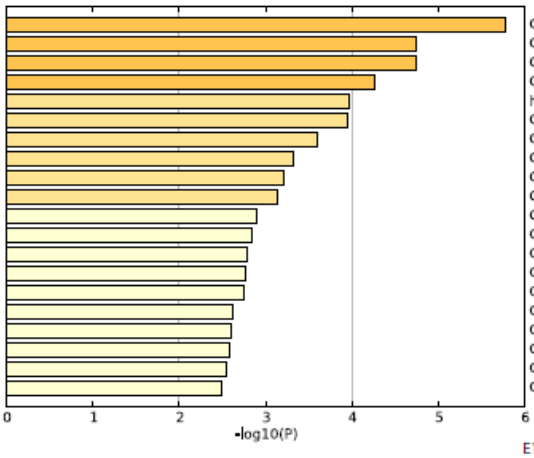
E098



E100

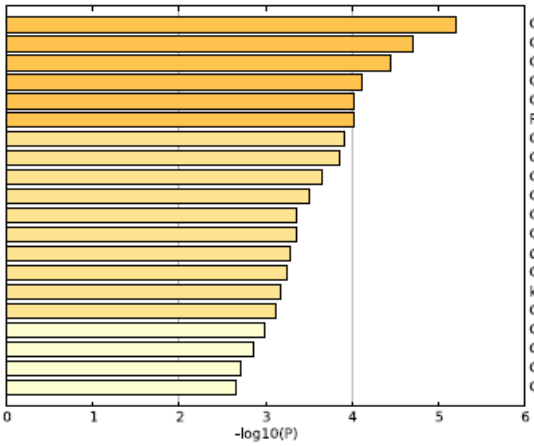


E104



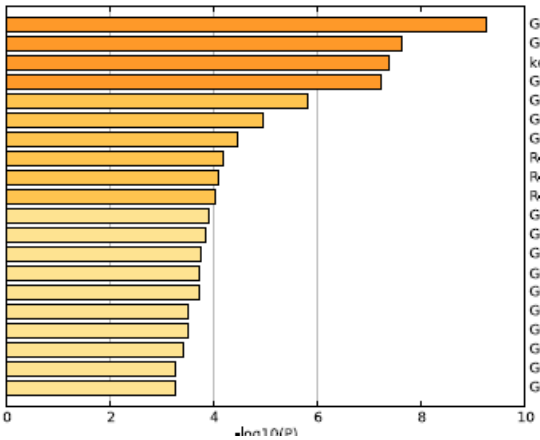
GO:0061061: muscle structure development
GO:0097067: cellular response to thyroid hormone stimulus
GO:0045721: negative regulation of gluconeogenesis
GO:0009725: response to hormone
hsa04514: Cell adhesion molecules (CAMs)
GO:2001257: regulation of cation channel activity
GO:1900182: positive regulation of protein localization to nucleus
GO:0000060: protein import into nucleus, translocation
GO:0043500: muscle adaptation
GO:0071243: cellular response to arsenic-containing substance
GO:0043618: regulation of transcription from RNA polymerase II promoter in response to stress
GO:0040008: regulation of growth
GO:0002437: inflammatory response to antigenic stimulus
GO:0043393: regulation of protein binding
GO:0043484: regulation of RNA splicing
GO:0000302: response to reactive oxygen species
GO:0060711: labyrinthine layer development
GO:0048002: antigen processing and presentation of peptide antigen
GO:0060445: branching involved in salivary gland morphogenesis
GO:0070076: histone lysine demethylation

E105

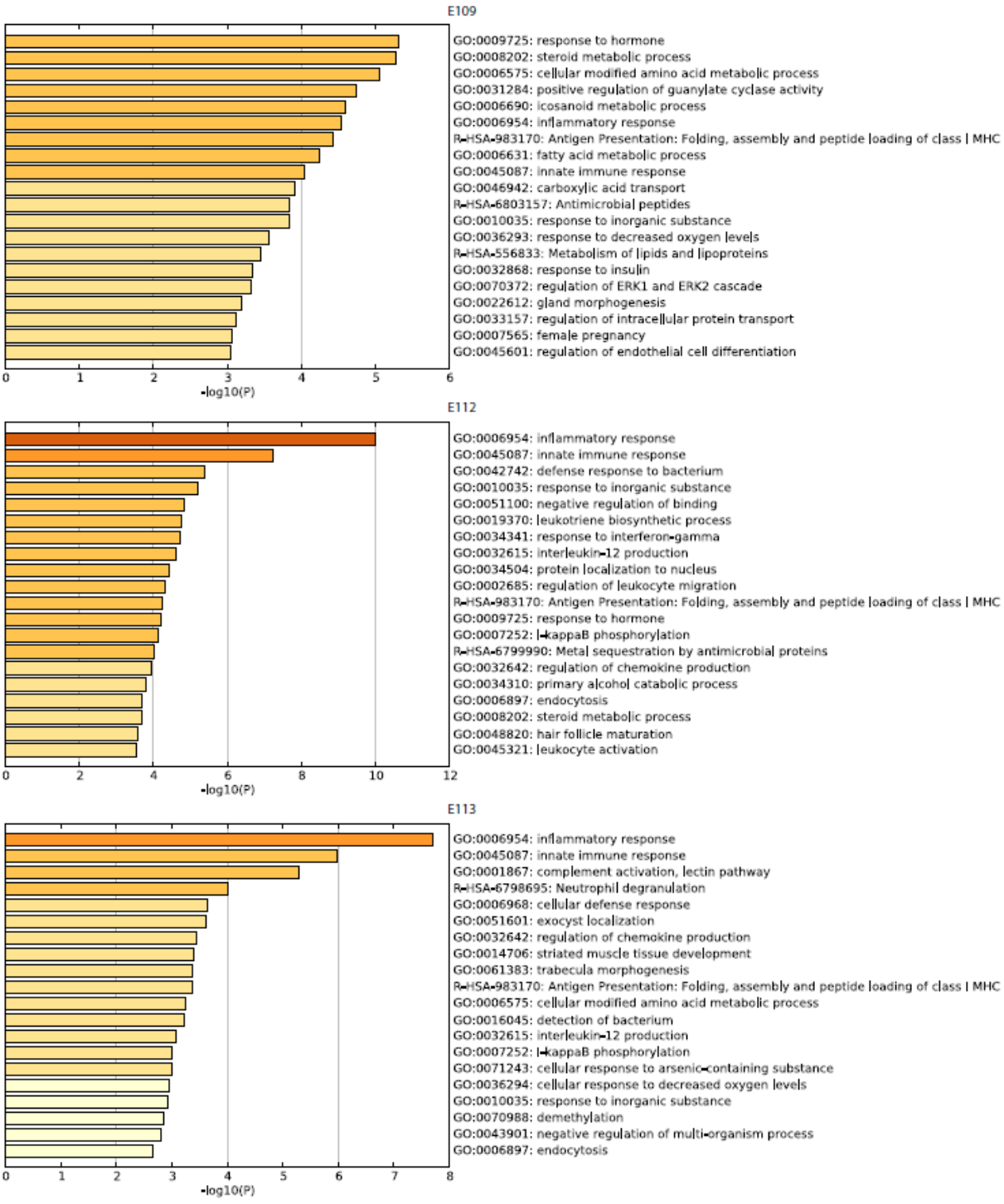


GO:0042307: positive regulation of protein import into nucleus
GO:0097067: cellular response to thyroid hormone stimulus
GO:1901019: regulation of calcium ion transmembrane transporter activity
GO:0061061: muscle structure development
GO:0045087: innate immune response
R-HSA-6798695: Neutrophil degranulation
GO:0003139: secondary heart field specification
GO:0042088: T-helper 1 type immune response
GO:0019882: antigen processing and presentation
GO:0030029: actin filament-based process
GO:0010035: response to inorganic substance
GO:0043393: regulation of protein binding
GO:0045454: cell redox homeostasis
GO:0032615: interleukin-12 production
ko04514: Cell adhesion molecules (CAMs)
GO:0071243: cellular response to arsenic-containing substance
GO:0034694: response to prostaglandin
GO:0043618: regulation of transcription from RNA polymerase II promoter in response to stress
GO:0003215: cardiac right ventricle morphogenesis
GO:1903010: regulation of bone development

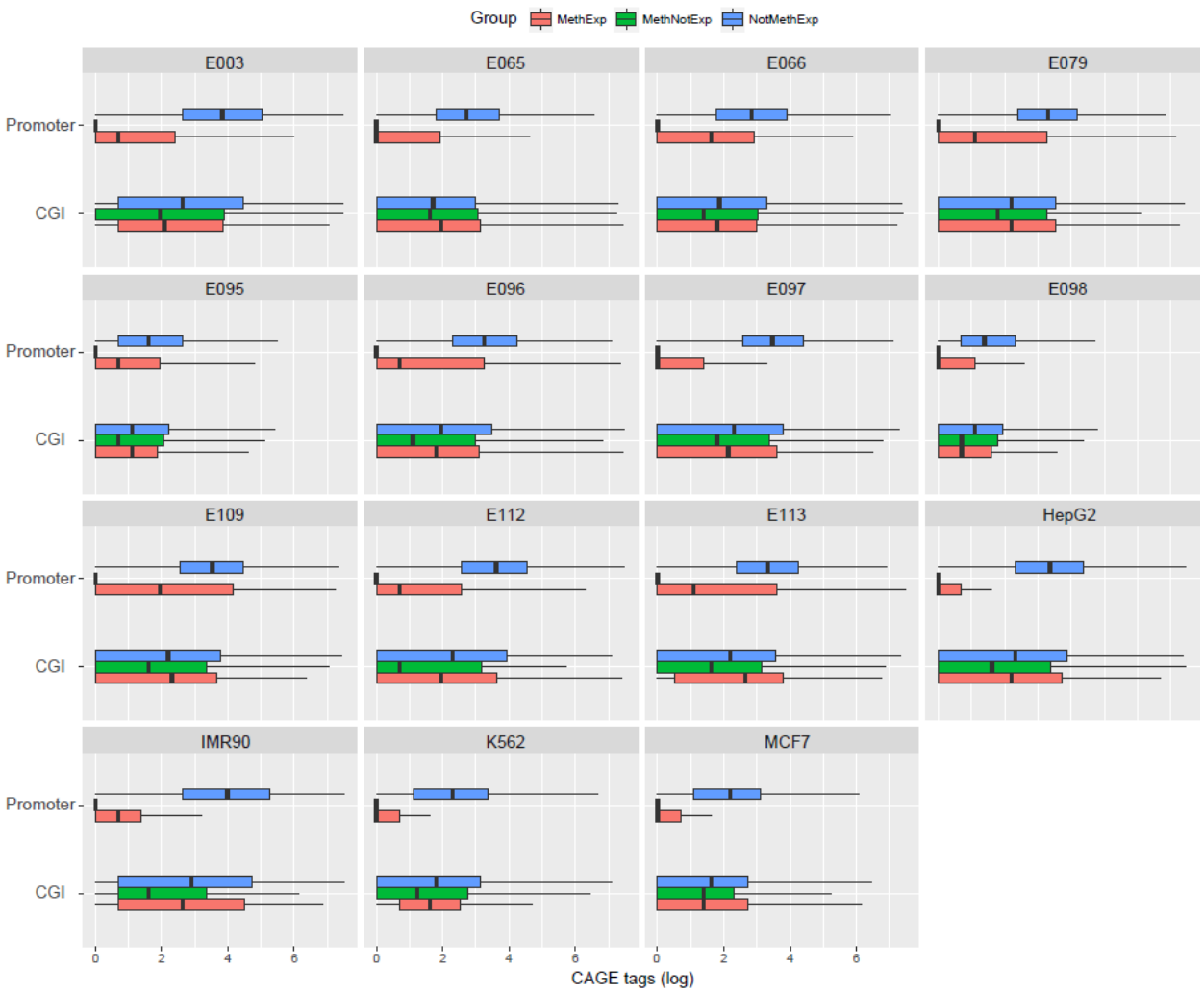
E106



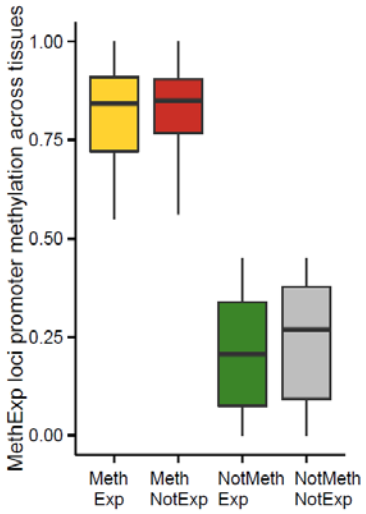
GO:0002684: positive regulation of immune system process
GO:0045321: leukocyte activation
ko04514: Cell adhesion molecules (CAMs)
GO:0006954: inflammatory response
GO:0050777: negative regulation of immune response
GO:0050900: leukocyte migration
GO:0050672: negative regulation of lymphocyte proliferation
R-HSA-193807: Synthesis of bile acids and bile salts via 27-hydroxycholesterol
R-HSA-391160: Signal regulatory protein (SIRP) family interactions
R-HSA-556833: Metabolism of lipids and lipoproteins
GO:0009725: response to hormone
GO:0002274: myeloid leukocyte activation
GO:0034341: response to interferon-gamma
GO:0048820: hair follicle maturation
GO:0043124: negative regulation of I-kappaB kinase/NF-kappaB signaling
GO:0002437: inflammatory response to antigenic stimulus
GO:0006968: cellular defense response
GO:0043065: positive regulation of apoptotic process
GO:0002921: negative regulation of humoral immune response
GO:0051241: negative regulation of multicellular organismal process



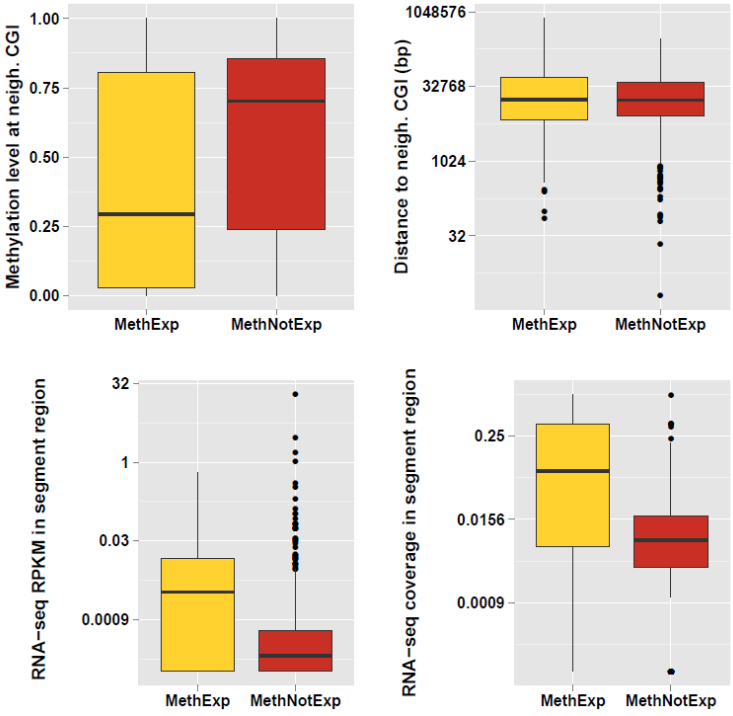
Supplemental Figure B-3. Bar plots showing GO categories (Y-axes) and their enrichment levels in -log(P) (X-axes) for each of the 34 tissue types assessed. A description of tissue IDs is provided in Table 3-1.



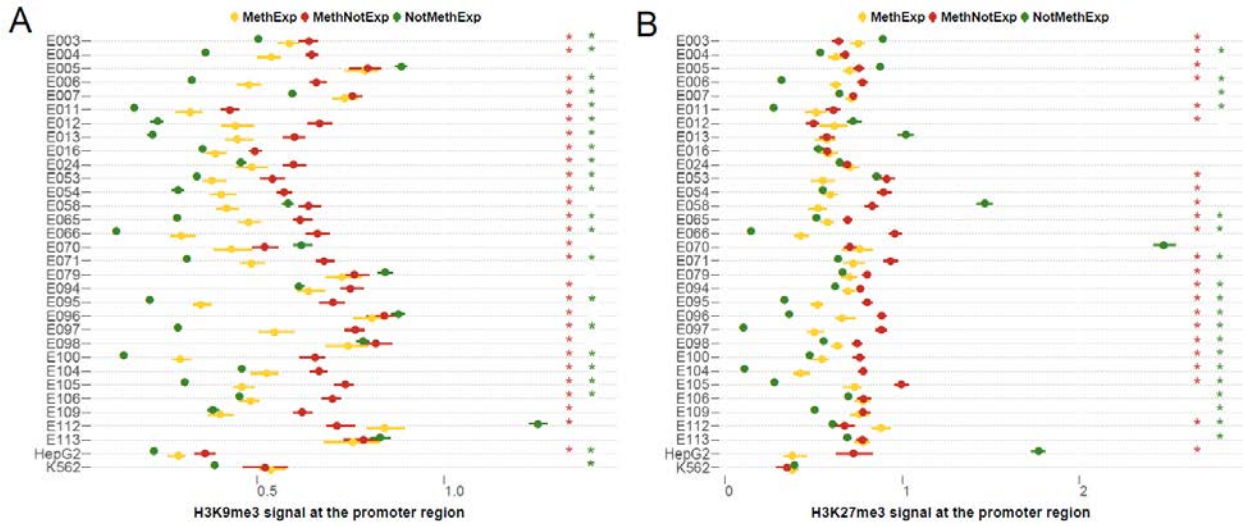
Supplemental Figure B-4. Boxplots showing the evidence of transcriptional initiation based on CAGE tag intensity (Y-axes; log-transformed) contrasted for three gene groups, MethExp (pink), MethNotExp (green) and NotMethExp (blue) at (i) the upstream CGI, and (ii) the proximal-promoter (X-axes) across 15 tissue types.



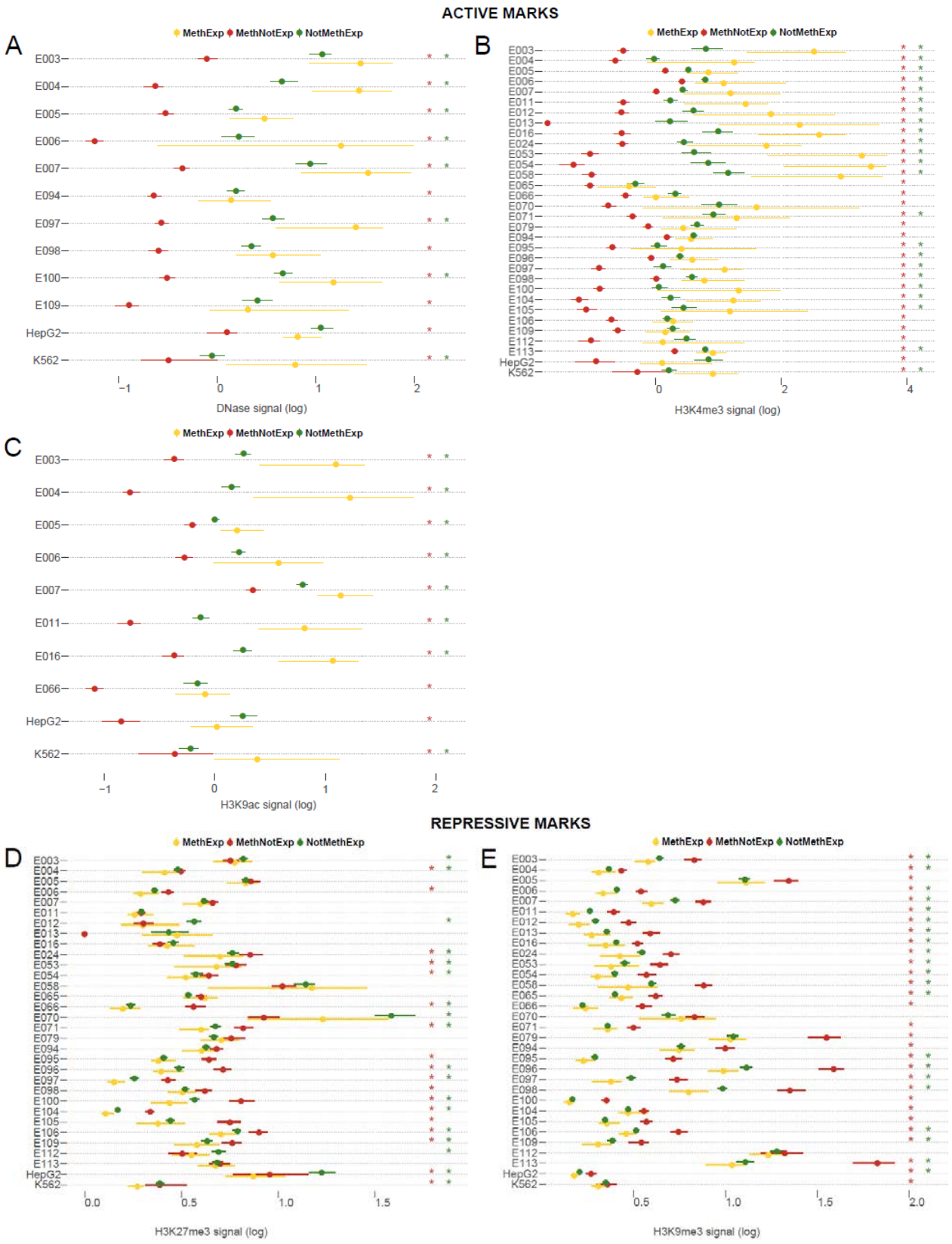
Supplemental Figure B-5. For the pan-tissue pooled set of MethExp genes, the plot shows the level of fractional methylation (Y-axis) at the promoters of these genes when they are MethExp (yellow), MethNotExp (red), NotMethExp (green) and NotMethNotExp (grey) in other tissues (X-axis).



Supplemental Figure B-6. The figure shows four lines of evidence supporting the usage of distal CGI as alternative promoter by MethExp genes in contrast to MethNotExp genes in mouse embryonic stem cells. Each panel shows the distribution of the median (i) fractional methylation at upstream CGIs (top-left), (ii) genomic distance between distal CGI and gene (top-right), (iii) RNA-seq RPKM signal (bottom-left), and (iv) RNA-seq coverage (bottom-right) at the segment region (Y-axes) corresponding to MethExp (yellow) vs. MethNotExp genes (red) (X-axes).

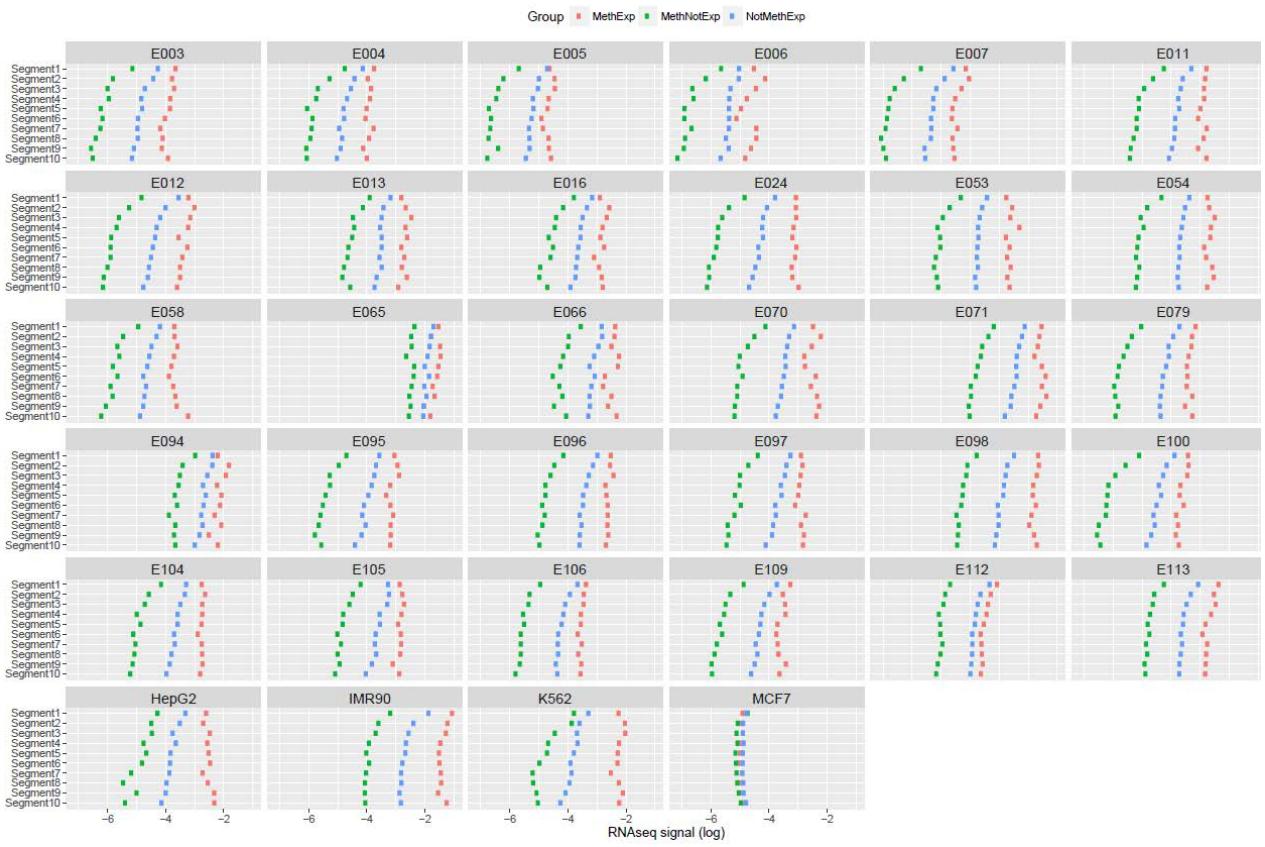


Supplemental Figure B-7. Repressive marks at the promoter region. (A) Median H3K9me3 signal, (B) median H3K27me3 signal, and 95% CI (X-axes) associated with the promoter regions of MethExp (yellow), MethNotExp (red) and NotMethExp (green) genes across 32 tissue types (Y-axes).

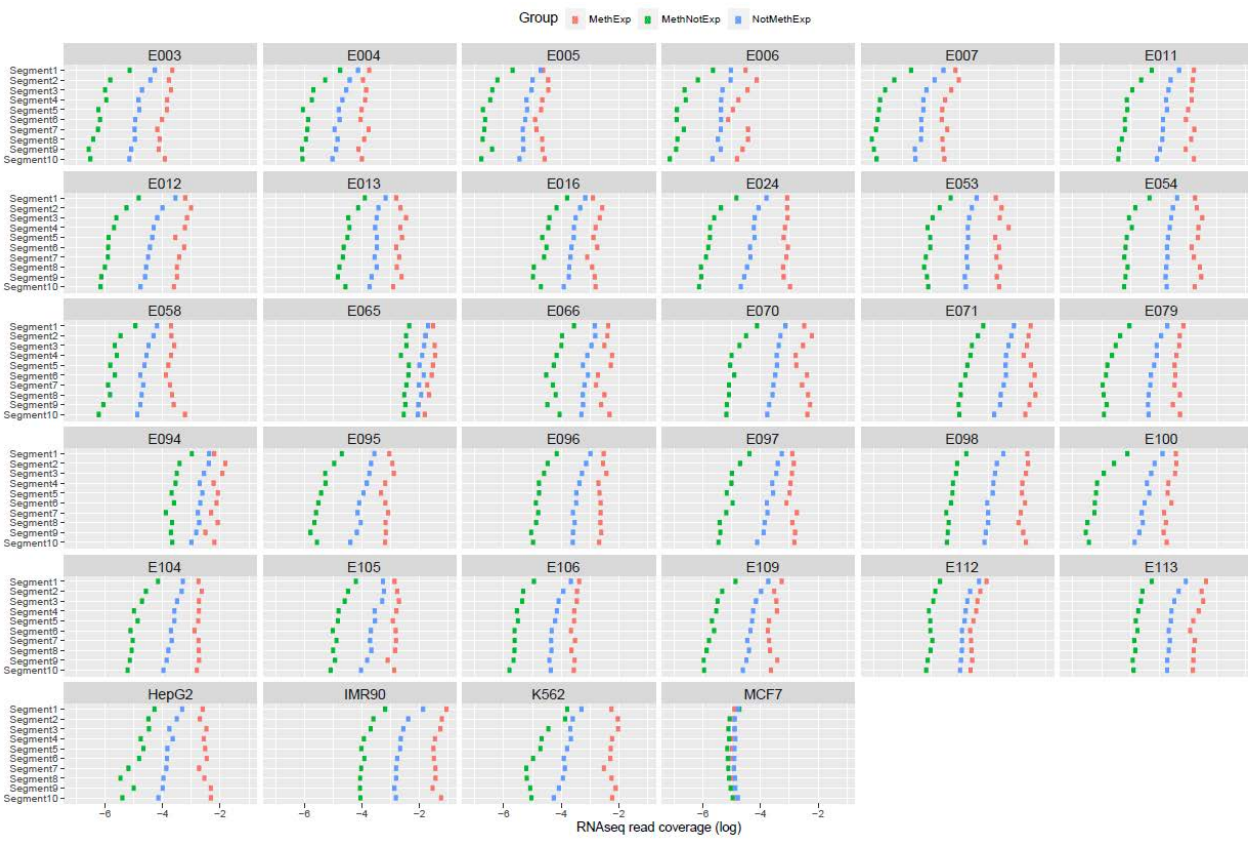


Supplemental Figure B-8. Active and repressive marks at the distal CGIs. (A) Median

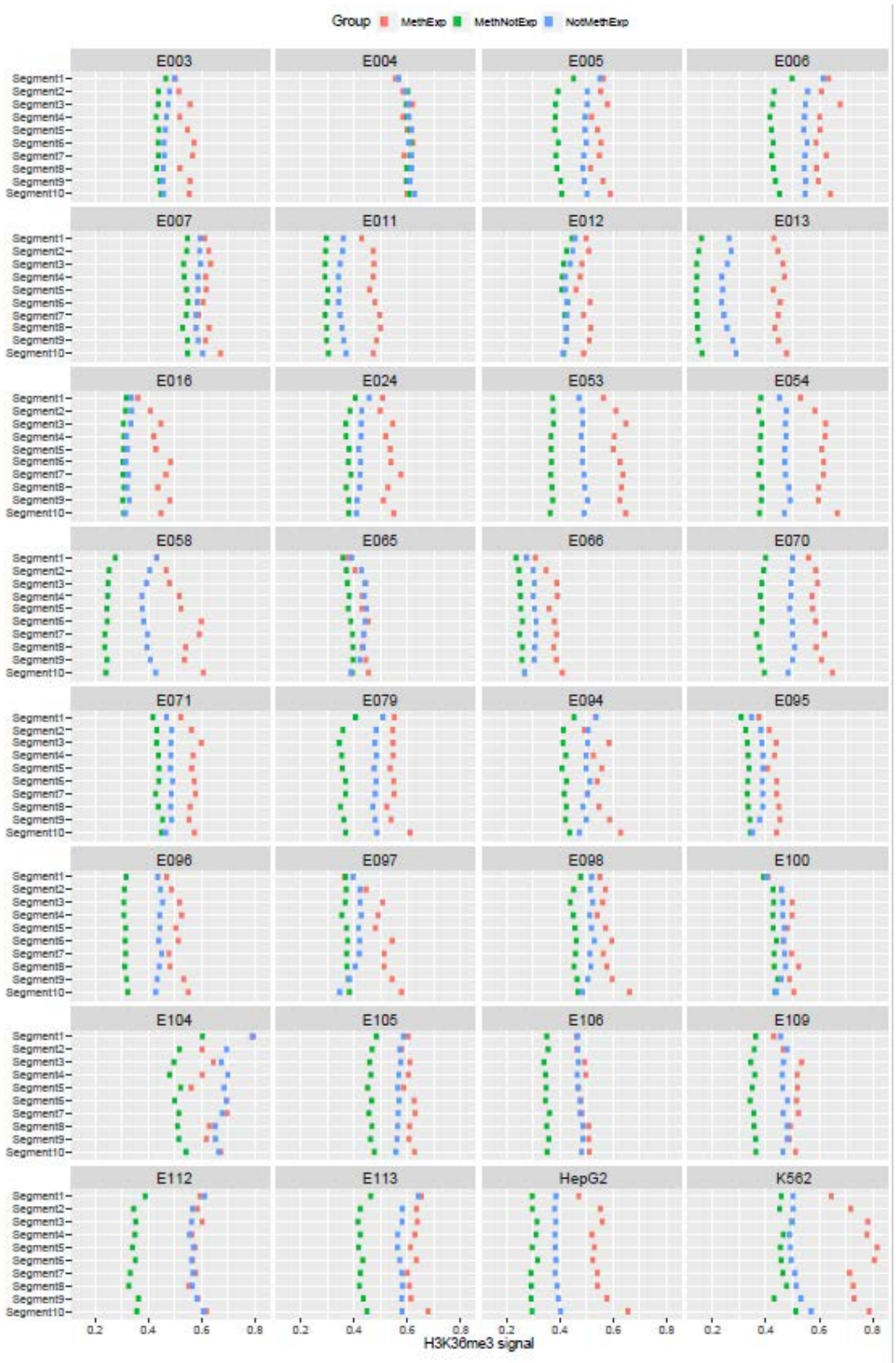
DNAse signal, (B) median H3K4me3 signal, (C) median H3K9ac signal, (D) median H3K27me3 signal,(E) median H3K9me3 signal, and 95% CI (X-axes) at the distal CGIs associated with MethExp (yellow), MethNotExp (red) and NotMethExp (green) genes across several tissue types (Y-axes).



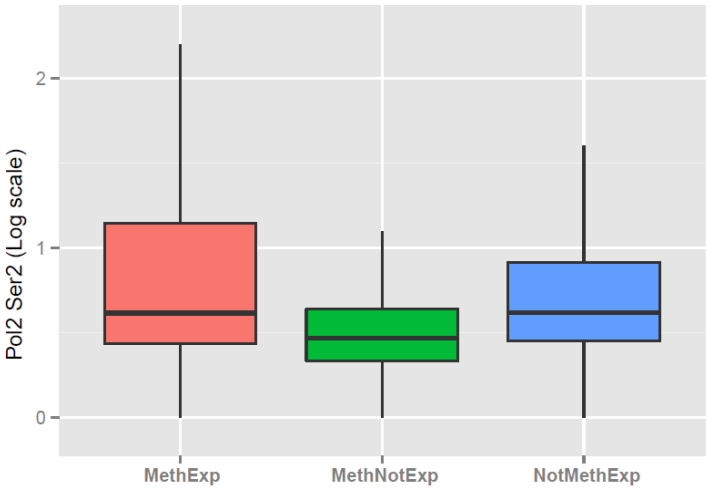
Supplemental Figure B-9a. Median RNA-seq signal (RPKM) (X-axes; log transformed) in each of 10 equal bins of the segment region associated with MethExp (pink), MethNotExp (green) and NotMethExp (blue) genes across 34 tissue types.



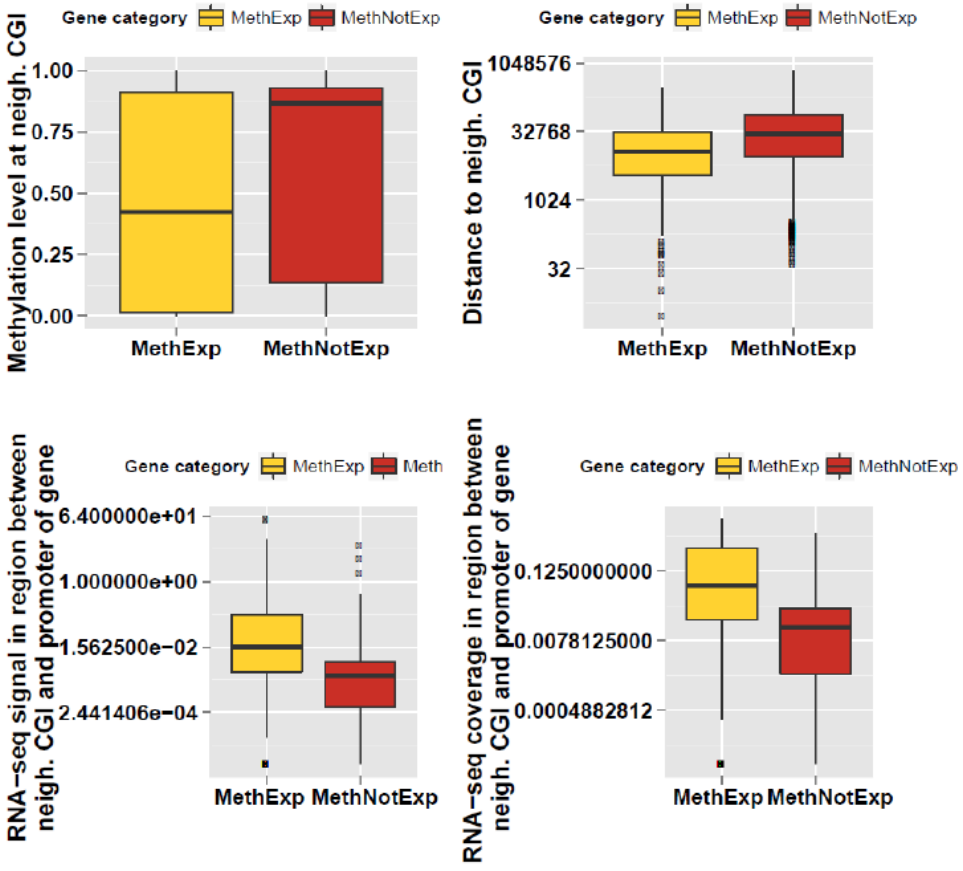
Supplemental Figure B-9b. Median RNA-seq read coverage (X-axes; log transformed) in each of 10 equal bins of the segment region associated with MethExp (pink), MethNotExp (green) and NotMethExp (blue) genes across 34 tissue types.



Supplemental Figure B-9c. Median H3K36me3 ChIP-seq signal intensities (X-axes; log transformed) in each of 10 equal bins of the segment region associated with MethExp (pink), MethNotExp (green) and NotMethExp (blue) genes across 34 tissue types.

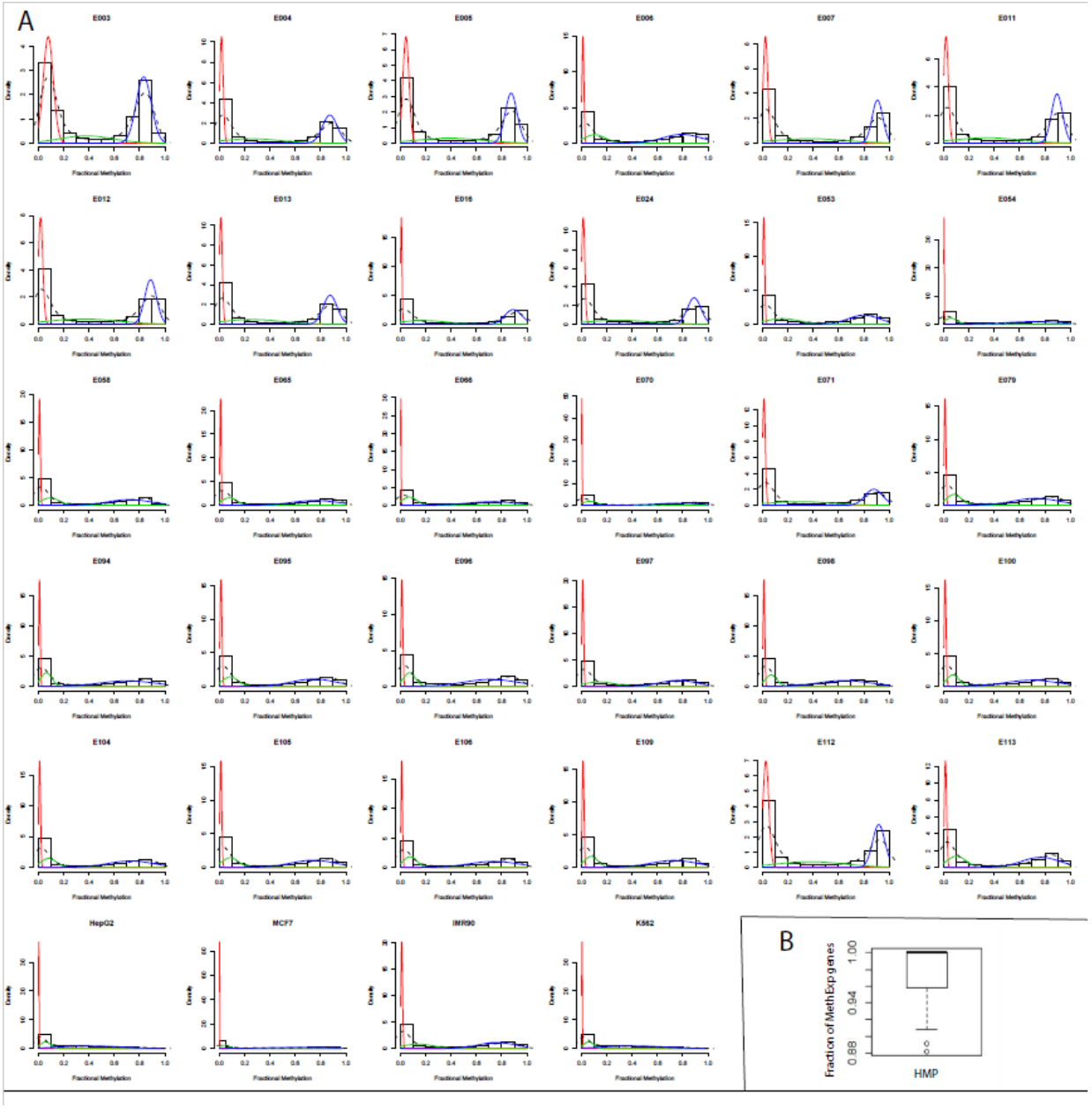


Supplemental Figure B-10. The evidence of transcriptional elongation based on PolII-Ser2 signals (Y-axis) in the segment region associated with three gene groups, MethExp (pink), MethNotExp (green) and NotMethExp (blue) (X-axis).

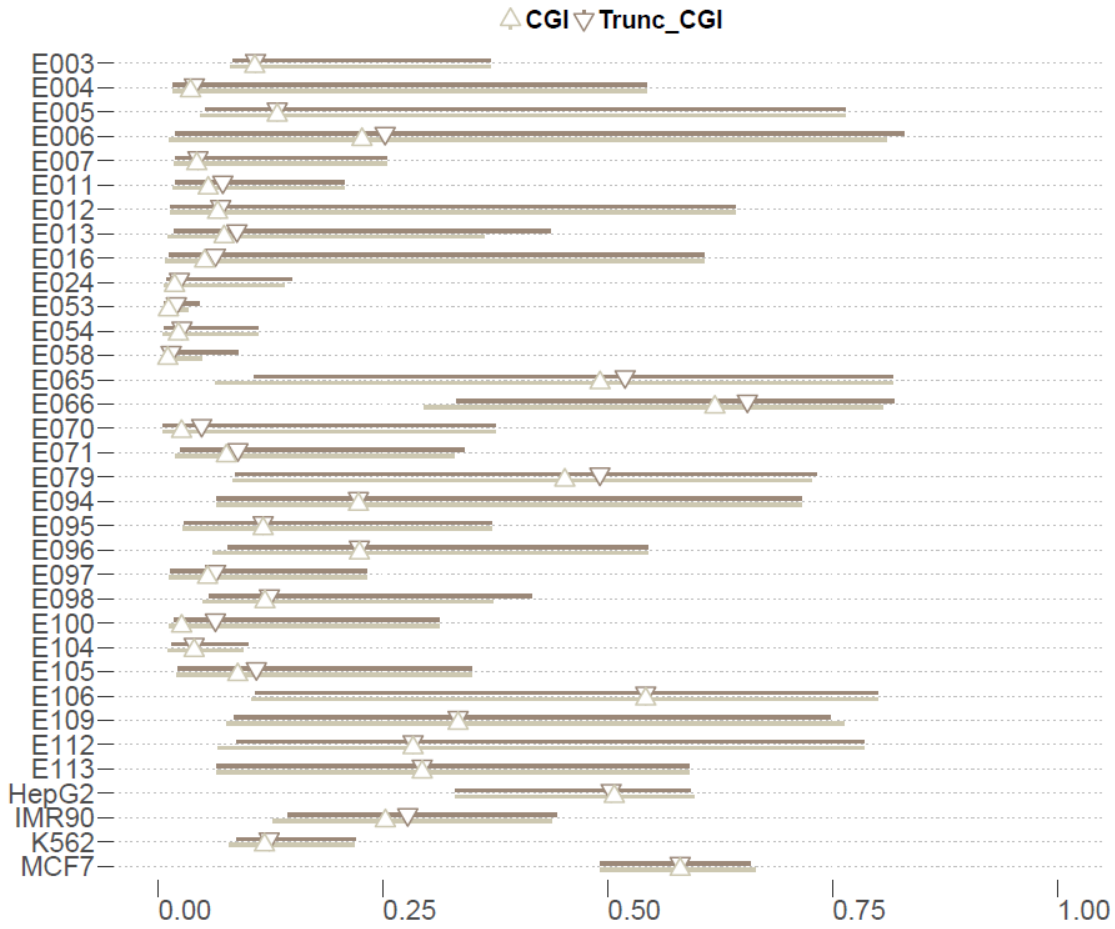


Supplemental Figure B-11. Four lines of evidence supporting the usage of distal CGI as alternative promoter by MethExp genes in contrast to MethNotExp genes in the human spleen (E113). The boxplots show the distribution of (i) fractional methylation at upstream CGIs (top-left),

(ii) genomic distance between distal CGI and gene (top-right), (iii) RNA-seq RPKM signal (bottom-left), and (iv) RNA-seq coverage (bottom-right) at the segment region corresponding to MethExp (yellow) vs. MethNotExp genes (red) computed using the GRCh38/hg38 genome annotation and "lifted-over" RNAseq and methylation data.



Supplemental Figure B-12. (A) Histograms depicting the distribution of fractional methylation across all promoters in each of the 34 tissue types assessed. Each distribution is fit by a 3-component Gaussian mixture model to distinguish subpopulations of lowly methylated (LMP; red), intermediate methylated (IMP; green) and highly methylated (HMP; blue) promoters. (B) Boxplot of the fraction of MethExp promoters that belong to the HMP category across 34 tissue types.



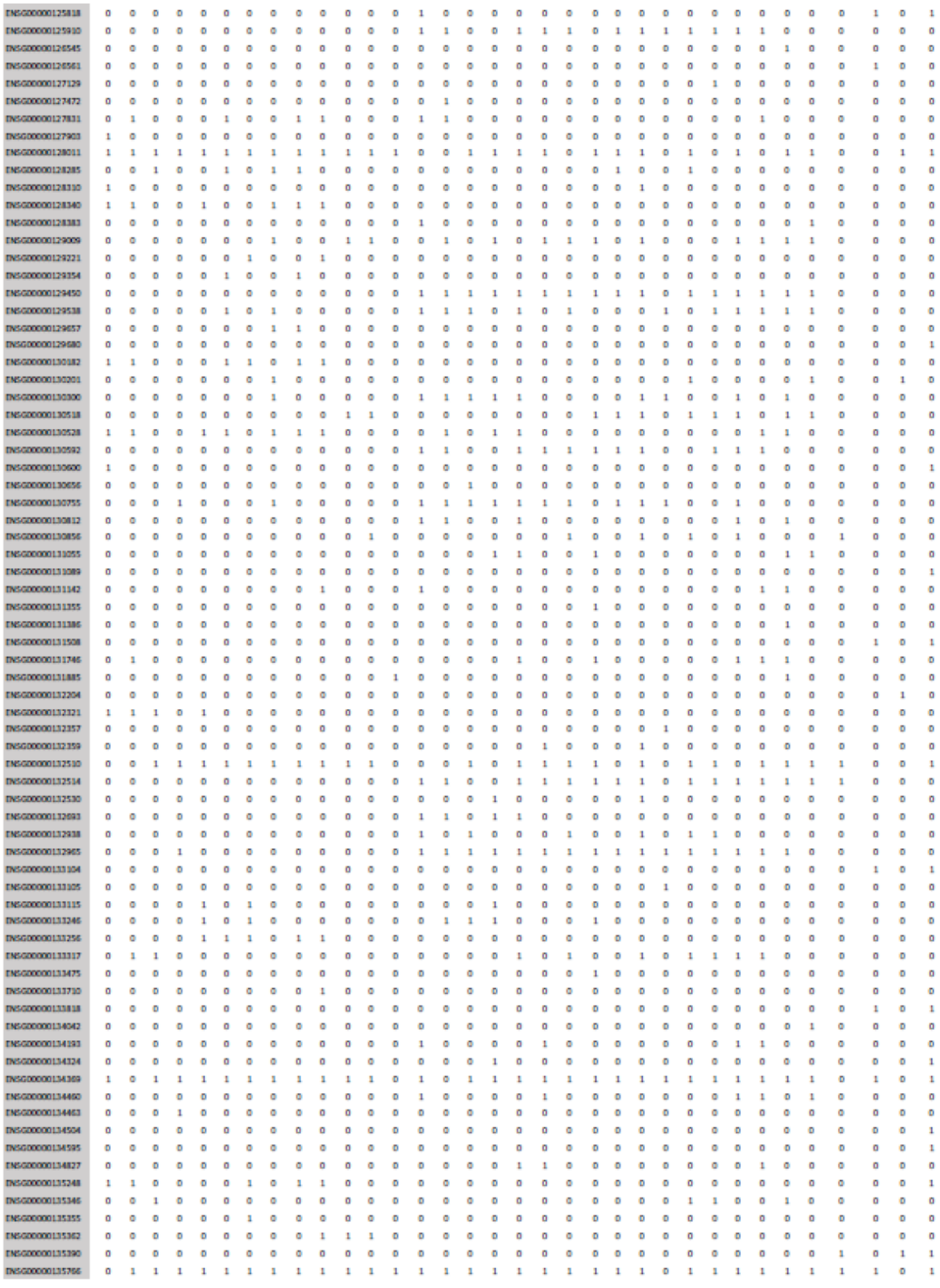
Supplemental Figure B-13. Median fractional methylation levels (X-axis) and 95% CI at the distal CGIs associated with MethExp genes before (light gray) and after (dark gray) truncation of these loci to 1 kb across 34 tissue types (Y-axis). Significant differences are marked with an asterisk. Truncation resulted in negligible differences in these distributions in all tissues.

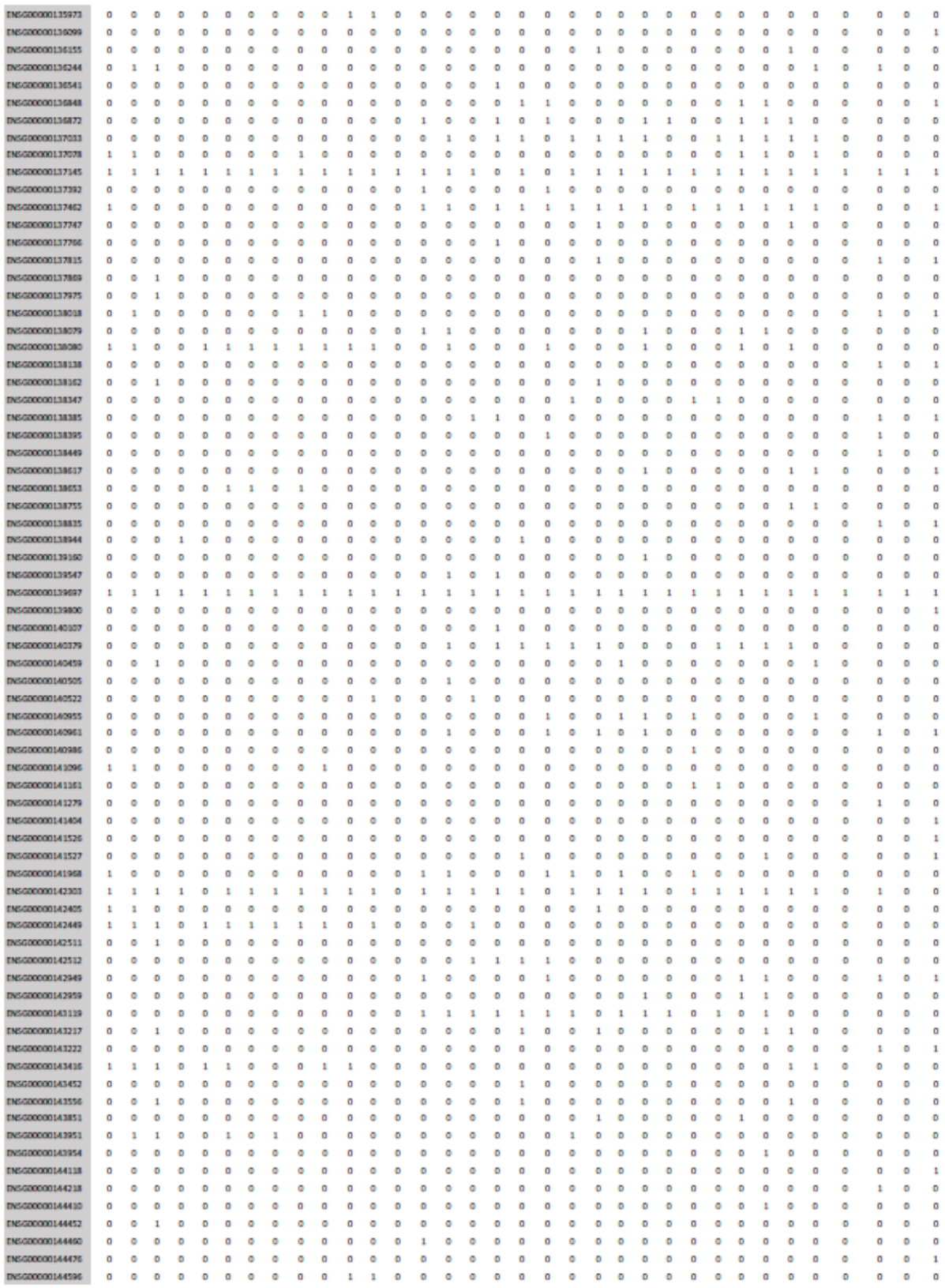
GeneID	E003	E004	E005	E006	E007	E011	E012	E013	E016	E024	E053	E054	E058	E065	E066	E070	E071	E079	E094	E095	E096	E097	E098	E100	E104	E105	E106	E109	E112	E113	HepG2	IMR90	K562	MCF7
ENSG00000000938	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000000971	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000001436	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000005102	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0
ENSG00000006831	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000007038	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000007129	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
ENSG00000007171	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000007216	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ENSG00000007314	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
ENSG00000008130	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000009694	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000009765	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
ENSG00000009790	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ENSG00000010671	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	0	1	0	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	0
ENSG00000011600	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	0
ENSG00000015592	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ENSG00000016602	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000017260	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000017483	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ENSG00000017621	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
ENSG00000018280	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0	0
ENSG00000018625	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000019169	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	1	0	1	1	1	0	0	0	0
ENSG00000020256	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000021300	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000021762	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000028116	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000029993	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000034053	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000034971	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1	0	1	1	0	0	0	0	0
ENSG00000035720	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000039139	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000044012	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ENSG00000049247	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000049249	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
ENSG00000049540	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000049768	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
ENSG00000050730	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
ENSG00000051596	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000054392	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000055957	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000060558	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0
ENSG00000060566	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000062598	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000065150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000065618	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
ENSG00000066405	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000067066	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000067208	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000068394	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000068784	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000069702	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000069998	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000070985	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0
ENSG00000071991	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	0	1	1	0	1	0	0	0	0	0
ENSG00000072163	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000072818	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	1	1	0	0	0	0	0	0
ENSG00000073067	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000074416	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000075073	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
ENSG00000075223	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000077498	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000078687	1	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0
ENSG00000078814	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000079308	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	0	0	0	0
ENSG00000080031	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000082014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000082175	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1

ENSG00000082196	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	1	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	0	1
ENSG00000082482	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000083750	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000084734	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0
ENSG00000085265	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	0	0	1	1	0	0	1	1	0	0	0	0
ENSG00000085998	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	0	1	1	1	0	1	0	0	1	1	0	0	0	1	0	1	0	1
ENSG00000086015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000086289	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000086696	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0
ENSG00000086991	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
ENSG00000087128	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000087365	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000087494	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000088002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	
ENSG00000088053	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000088827	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	
ENSG00000089012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
ENSG00000089057	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1		
ENSG00000089351	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0		
ENSG00000089558	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
ENSG00000090020	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1		
ENSG00000090382	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	0	
ENSG00000090661	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000090920	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	
ENSG00000091106	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0		
ENSG00000091138	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0		
ENSG00000091482	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
ENSG00000091592	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	
ENSG00000091986	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
ENSG00000092621	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
ENSG00000095370	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
ENSG00000095932	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
ENSG00000096395	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	
ENSG00000096996	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	
ENSG00000099331	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000099769	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
ENSG00000099998	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	
ENSG00000100078	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
ENSG00000100121	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	
ENSG00000100197	0	0	0	0	0	1	1	1	0	0	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1	0	0	0
ENSG00000100292	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000100346	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000100934	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
ENSG00000100985	1	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000101162	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0		
ENSG00000101166	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		
ENSG00000101197	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0		
ENSG00000101213	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0		
ENSG00000101276	1	1	1	0	0	1	1	0	1	1	0	0	0	0	0	0	1	1	0	1	0	1	0	0	1	1	0	1	0	0	0	0	0	
ENSG00000101342	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
ENSG00000101435	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000101460	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0	0	1	0	0	
ENSG00000101605	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		
ENSG00000101901	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
ENSG00000101981	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
ENSG00000102053	0	0	0	0	1	1	1	1	1	1	1	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0		
ENSG00000102385	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
ENSG00000102805	0	0	1	1	1	0	1	1	0	0	1	1	1	1	0	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	1
ENSG00000102837	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	1	0	0	0	1	1	1	1	0	0	0	
ENSG00000102891	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0		
ENSG00000102901	0	1	1	1	0	0	0	1	0	0	1	0	1	0	0	0	1	1	1	0	1	1	1	0	0	0	1	0	1	1	0	1	1	1
ENSG00000102935	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
ENSG00000102962	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0		
ENSG00000103355	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
ENSG00000103472	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
ENSG00000103485	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	0	0	0	1	0	1	1	0	0	0	0		
ENSG00000104371	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0		
ENSG00000104375	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000104408	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1		
ENSG00000104419	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		
ENSG00000104537	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0		
ENSG00000104691	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1		
ENSG00000104783	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		

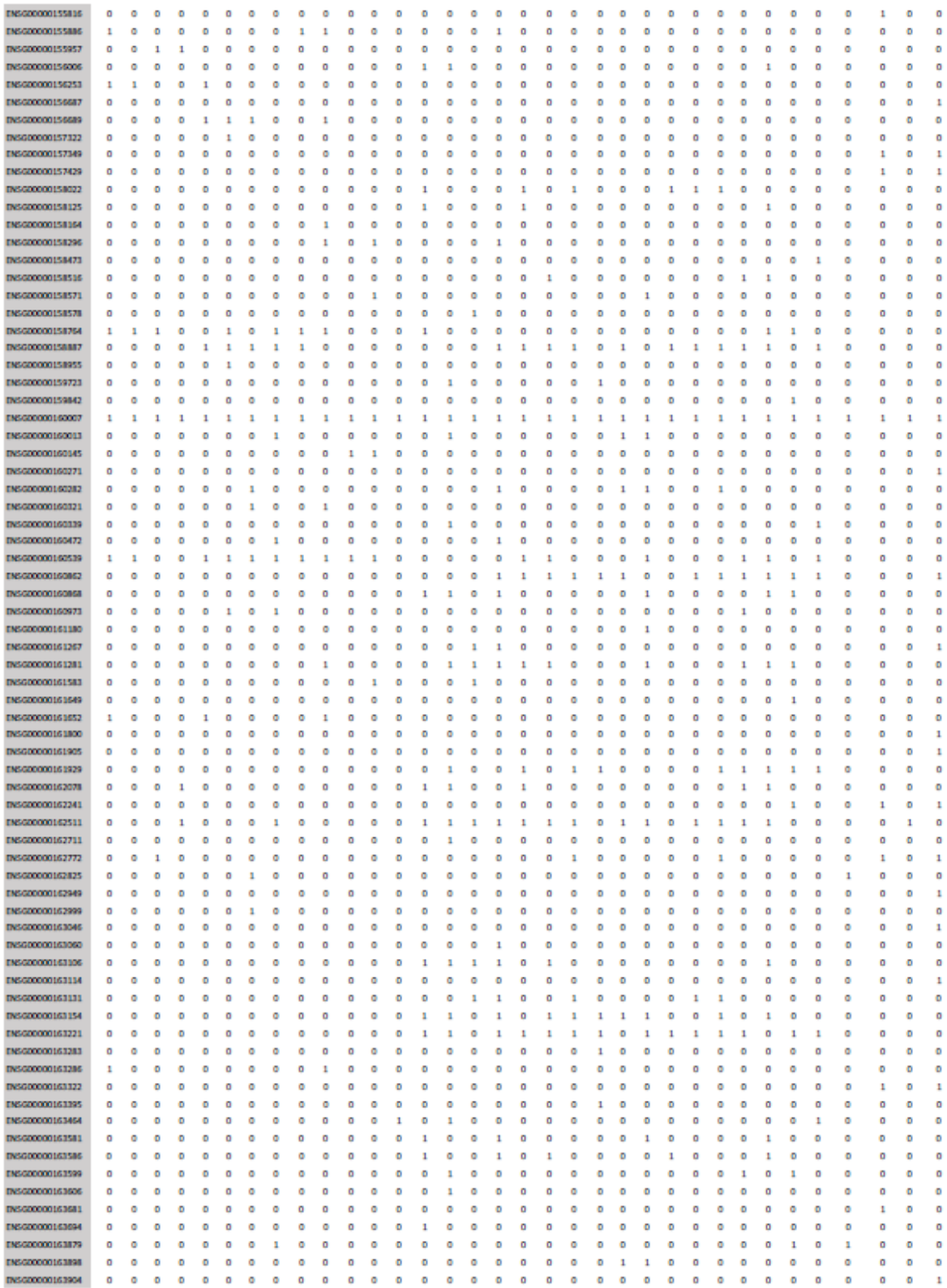
ENSG00000104826	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
ENSG00000104827	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000104879	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000104918	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	0	1	1	1	1	1	0	0	0	0
ENSG00000104972	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000105176	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000105198	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000105289	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000105327	1	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1
ENSG00000105341	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1
ENSG00000105366	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
ENSG00000105370	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000105383	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0
ENSG00000105427	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000105472	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000105492	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000105549	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000105559	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	1	0	0	0
ENSG00000105668	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000105928	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000105967	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000105993	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000106003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
ENSG00000106178	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	
ENSG00000106477	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000107165	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000107951	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
ENSG00000108176	1	1	0	0	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
ENSG00000108231	0	0	1	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000108342	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
ENSG00000108679	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	0	0	0	1	0	1	1	1	1	0	0	1	1	0	0	0	1
ENSG00000108700	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
ENSG00000108786	0	0	1	0	1	0	1	0	0	0	1	1	0	0	0	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	0	1	0	0
ENSG00000108849	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000109063	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000109180	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000109193	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000109576	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000109610	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	
ENSG00000109686	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	
ENSG00000109991	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000110077	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0
ENSG00000110203	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000110427	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000110852	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000110944	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0
ENSG00000111144	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000111181	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000111245	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	
ENSG00000111424	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000111445	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000111700	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000111801	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000111816	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0
ENSG00000111846	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000111859	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1
ENSG00000111907	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000112175	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000112245	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000112339	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000112486	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
ENSG00000112494	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000112619	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0
ENSG00000112742	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000112761	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000112782	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
ENSG00000112812	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	
ENSG00000112818	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
ENSG00000112874	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000113073	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
ENSG00000113248	1	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000113303	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	
ENSG00000113387	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	1

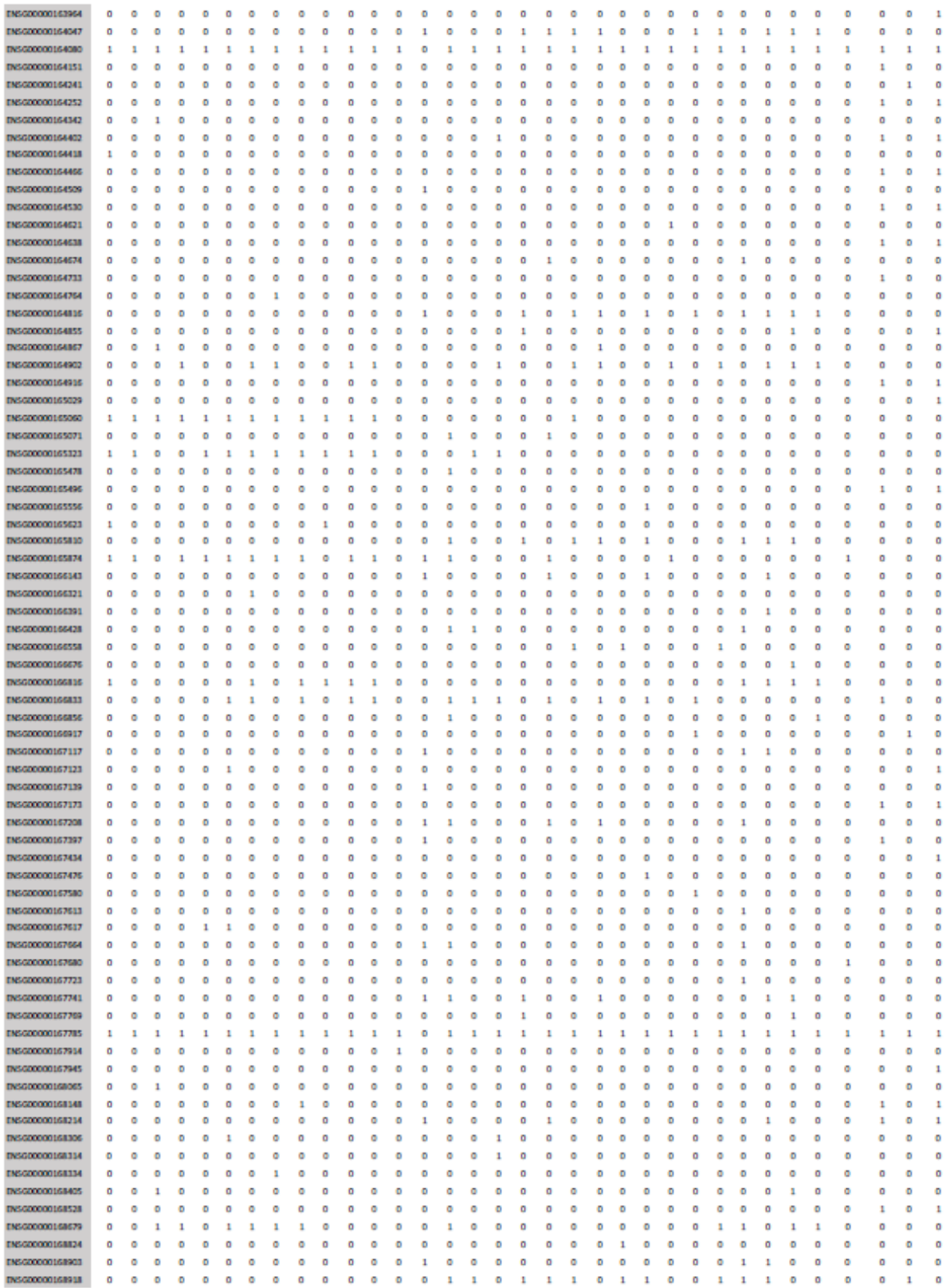
ENSG00000113578	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
ENSG00000113648	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000113721	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000113805	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0
ENSG00000114013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0
ENSG00000114124	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ENSG00000114200	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000114455	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ENSG00000114554	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000115009	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1	1	1	0	1	0	0	0	1	1	1	0	0	0	0
ENSG00000115165	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000115252	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000115363	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
ENSG00000115386	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0
ENSG00000115593	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000115604	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000115648	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000115694	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000115718	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000115806	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	0	0
ENSG00000115919	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	1	0	0	0	0
ENSG00000115963	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000116031	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000116157	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000116176	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ENSG00000116824	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1	0	0	0	0	0	1	1	0	1	0	0	0
ENSG00000116991	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000117090	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000117139	1	0	1	1	1	1	1	1	1	0	0	0	1	1	0	0	0	1	1	1	1	1	1	0	0	0	1	1	0	1	0	0	1
ENSG00000117228	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000117983	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000118004	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000118094	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000118473	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000118520	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0
ENSG00000118526	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000118702	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
ENSG00000118729	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	1	1	1	1	0	0	0	0	0
ENSG00000118785	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000119535	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000119938	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
ENSG00000120055	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0	0	1	0
ENSG00000120211	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000120280	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0
ENSG00000120332	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000120675	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000121691	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000121895	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
ENSG00000122121	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000122133	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000122145	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000122224	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
ENSG00000122497	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000122543	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000122718	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000122852	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000123106	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000123131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
ENSG00000123146	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000123454	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000123610	1	0	0	1	0	0	0	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0	1	1	0	0	0	0
ENSG00000124019	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000124107	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	0	1	1	0	0	1	0	0	0	0
ENSG00000124196	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000124203	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
ENSG00000124225	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1
ENSG00000124333	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000124429	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0
ENSG00000124610	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000124749	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000125089	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0
ENSG00000125378	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1
ENSG00000125746	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1



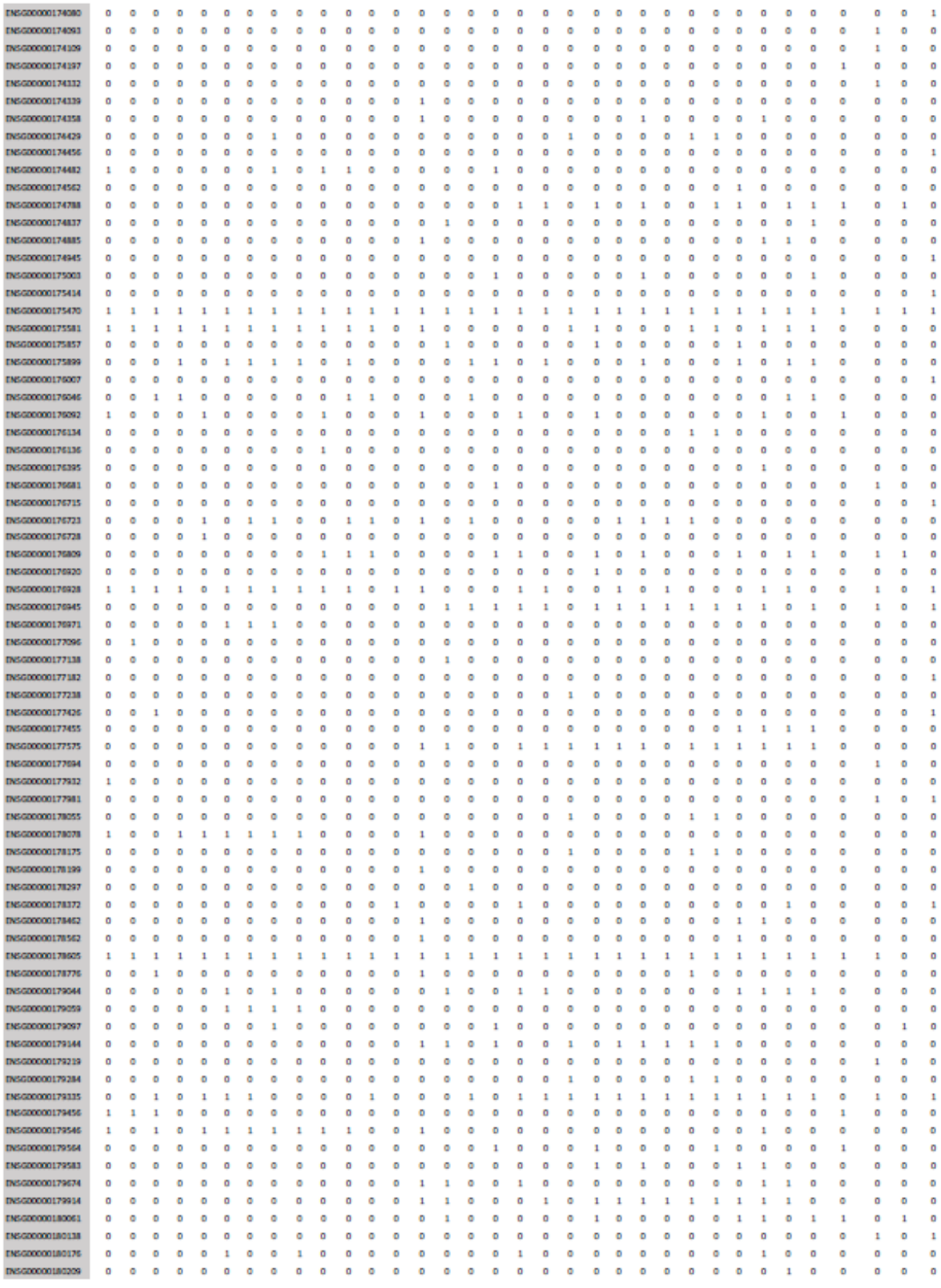


ENSG00000144655	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
ENSG00000144711	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000144736	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000145113	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000145287	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0
ENSG00000145730	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000145850	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
ENSG00000146039	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
ENSG00000146143	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000146166	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000146192	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0
ENSG00000146521	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000146587	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000146830	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000147003	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000147059	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000147117	1	0	0	0	1	1	1	0	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000147144	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0
ENSG00000147206	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000147262	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000147394	1	1	1	0	1	1	1	1	1	1	0	0	1	0	0	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1
ENSG00000147443	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0
ENSG00000147526	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1
ENSG00000147689	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000147697	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
ENSG00000147872	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ENSG00000147996	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000148335	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0
ENSG00000148344	1	0	1	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000148357	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000148396	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000148584	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000148834	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000149043	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000149131	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	1	1	1	1	1	0	0	0
ENSG00000150201	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000150636	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000150687	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1
ENSG00000150961	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000151320	0	0	0	0	1	1	1	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	1	1	0	1	0	0	0	0	1	0
ENSG00000151327	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000151413	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
ENSG00000151474	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000151632	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
ENSG00000151655	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000151835	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000151846	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000151883	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000152213	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ENSG00000152467	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000152580	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0
ENSG00000152601	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ENSG00000152804	0	0	0	1	0	1	0	1	0	0	0	0	0	0	1	0	0	1	1	0	0	1	1	0	0	0	1	1	0	1	1	1	1
ENSG00000153029	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000153071	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000153165	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ENSG00000153303	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000153446	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	0	1	1	0	0	0	0	0
ENSG00000153495	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0
ENSG00000153531	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1
ENSG00000153666	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000153789	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000153922	1	1	1	0	1	1	1	1	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
ENSG00000154493	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000154518	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000154556	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000154760	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000154783	1	1	1	0	0	1	0	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
ENSG00000154874	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000155066	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	1	1	0	1	1	1	0	1	1	0	0	0	0	0
ENSG00000155269	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000155506	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
ENSG00000155719	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



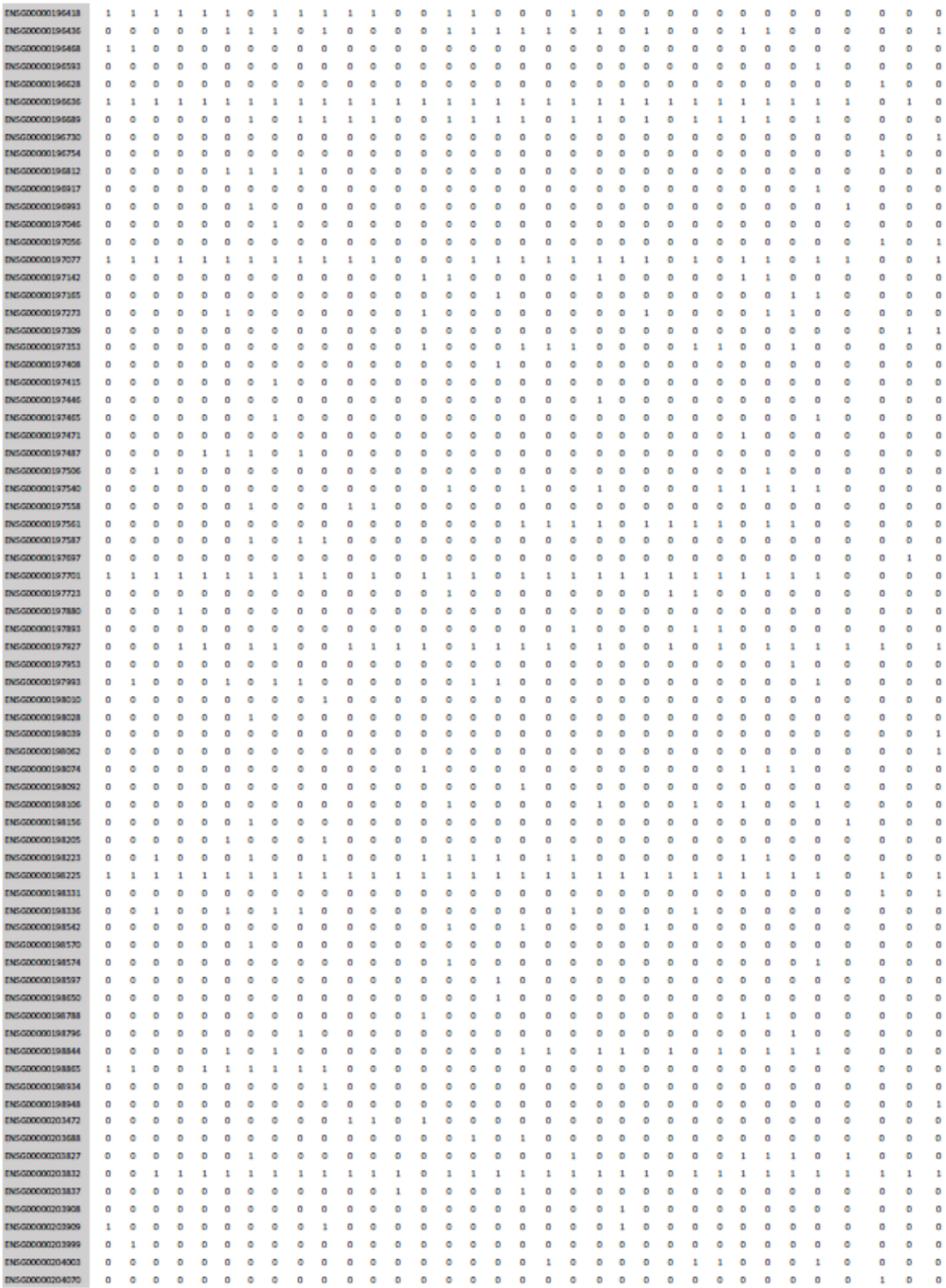


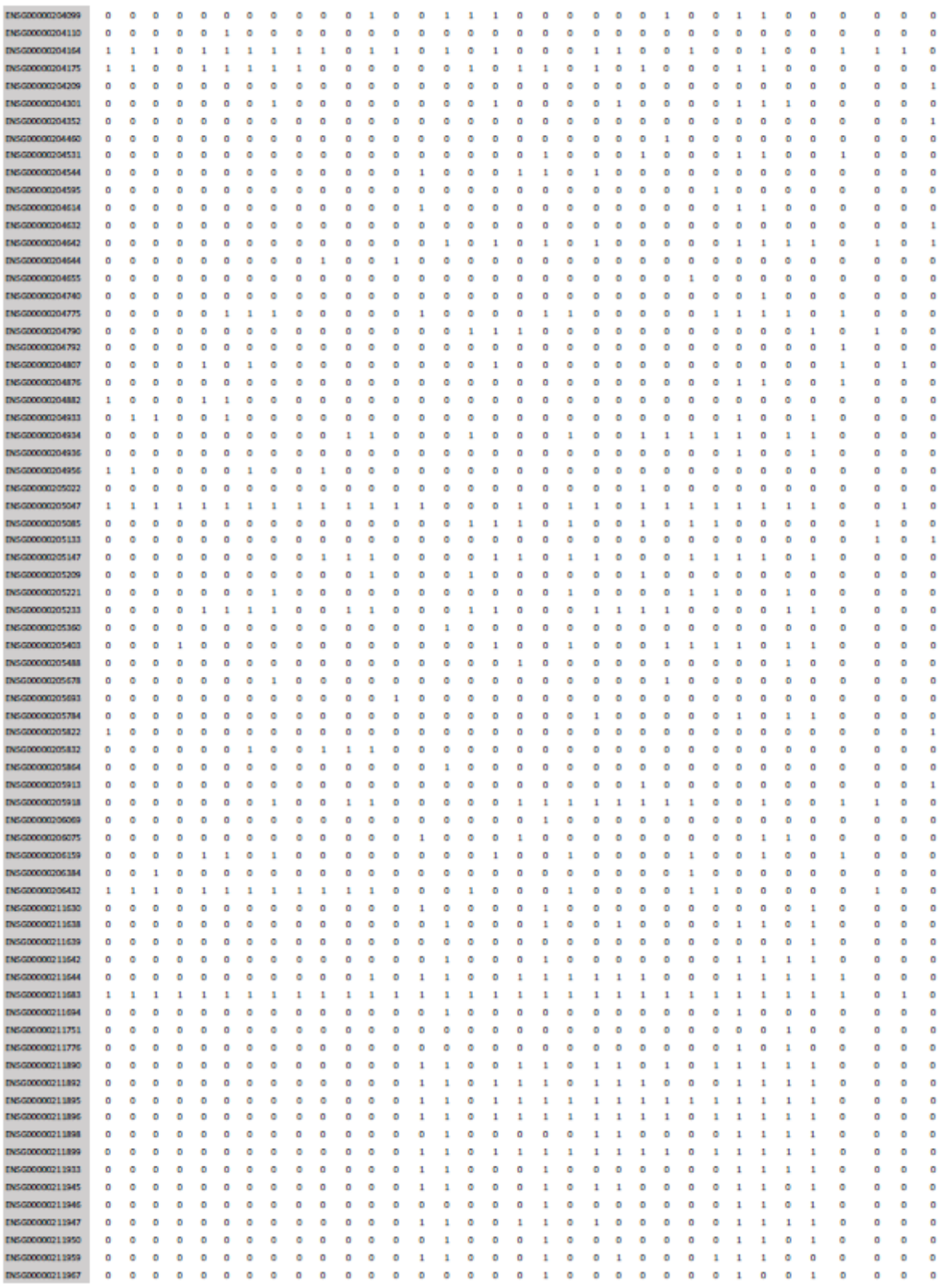
ENSG00000168925	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000168928	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000169220	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000169239	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000169246	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000169271	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	1	0	0	1	1	0	0	0	0	1	1	0	
ENSG00000169575	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000169629	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000169704	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000169877	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000169885	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000169896	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0
ENSG00000169903	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000169962	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0
ENSG00000169992	0	0	0	1	1	1	1	1	1	0	1	1	0	0	0	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	0	1	1	
ENSG00000169994	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000170035	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000170113	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	1	1	0	0	0	0	1	0	1	
ENSG00000170153	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
ENSG00000170160	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000170231	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000170255	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000170298	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
ENSG00000170324	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
ENSG00000170423	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000170523	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000170549	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000170579	1	1	0	0	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000170590	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
ENSG00000170871	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000170891	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000170921	0	0	1	1	0	1	1	1	1	1	1	0	1	0	0	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	1	1	1	
ENSG00000170948	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
ENSG00000171055	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000171102	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000171115	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	
ENSG00000171119	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	1	0	1	0	1	
ENSG00000171121	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	
ENSG00000171124	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000171433	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	
ENSG00000171466	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000171476	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
ENSG00000171557	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000171561	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
ENSG00000171564	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
ENSG00000171604	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000171777	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000171847	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
ENSG00000171951	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000172023	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ENSG00000172179	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000172232	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
ENSG00000172236	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0
ENSG00000172243	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	0	0
ENSG00000172346	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000172482	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000172543	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0
ENSG00000172578	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0
ENSG00000172724	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	1	0	0	1	1	1	1	0	0	0	0	0
ENSG00000172828	0	0	0	0	1	1	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0
ENSG00000172927	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	1	0	0	1	0	0
ENSG00000173040	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000173041	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000173080	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
ENSG00000173213	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
ENSG00000173227	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
ENSG00000173237	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
ENSG00000173302	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000173320	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000173638	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1	0	1
ENSG00000173876	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000173898	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000174030	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	

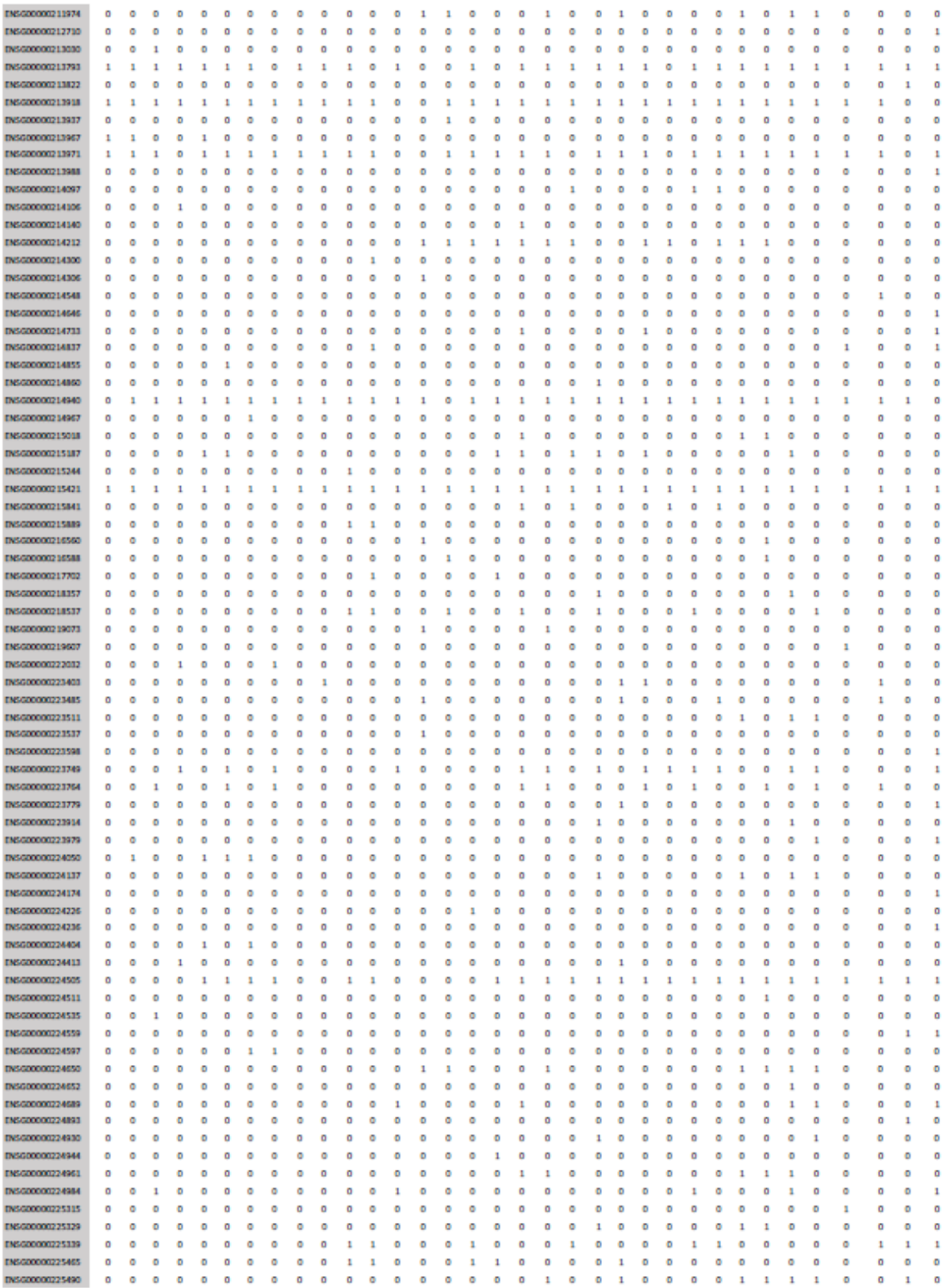


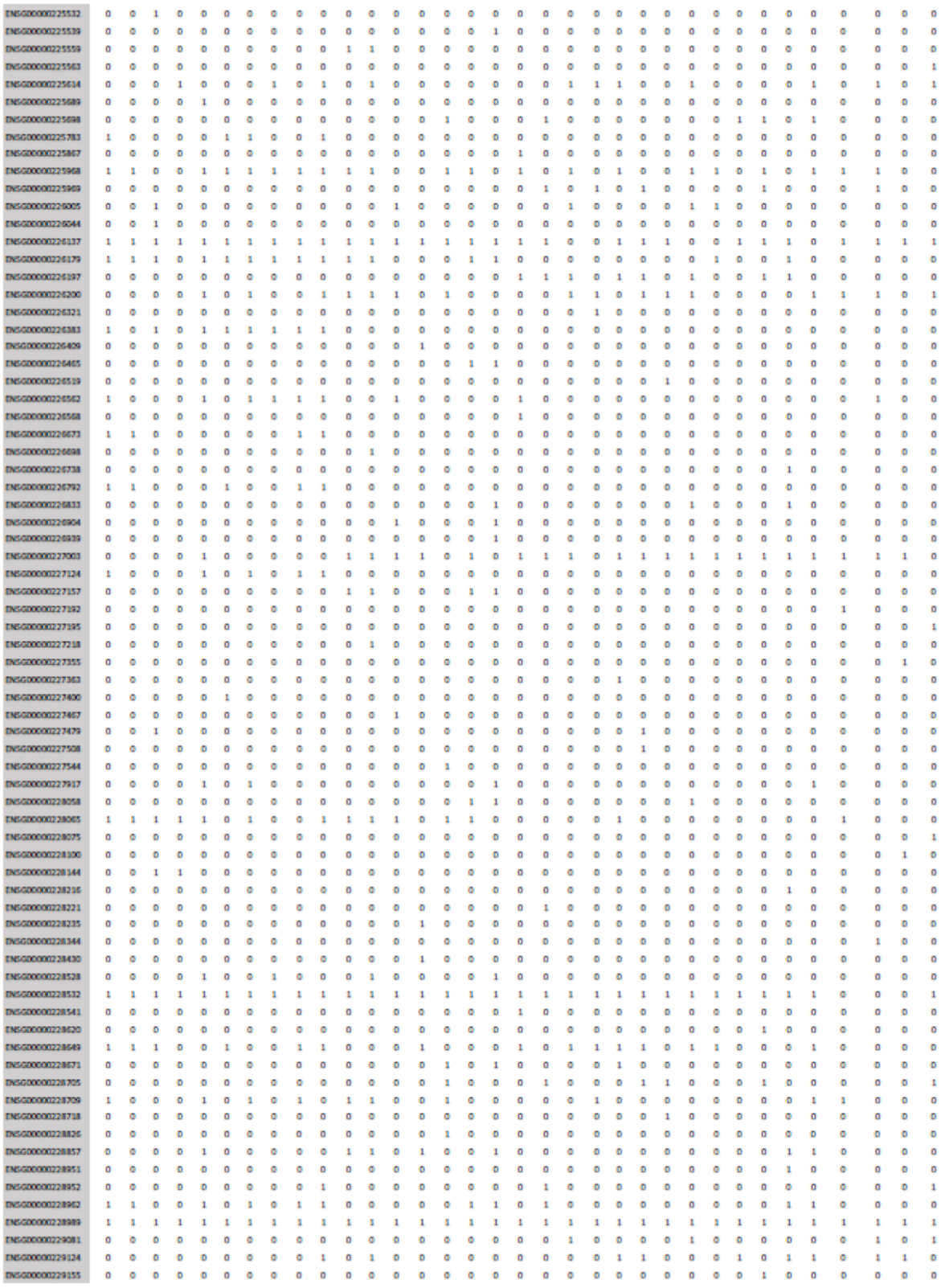
ENSG00000180458	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0
ENSG00000180543	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000180720	1	0	0	0	0	0	1	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	0
ENSG00000180745	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000180772	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000181031	1	0	1	0	1	1	0	1	1	1	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000181092	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	1	1	0	0	0	0	0	0
ENSG00000181143	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000181408	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ENSG00000181626	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000181652	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000181781	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000181804	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000182021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0
ENSG00000182109	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0
ENSG00000182156	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000182162	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0
ENSG00000182208	0	0	1	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000182372	0	1	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000182459	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000182472	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
ENSG00000182504	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000182557	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000182611	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
ENSG00000182636	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000182676	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
ENSG00000182749	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000182851	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0
ENSG00000182902	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
ENSG00000182916	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
ENSG00000182986	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000183019	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000183086	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	0	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0
ENSG00000183145	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000183230	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000183260	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
ENSG00000183281	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000183323	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000183395	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000183431	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	1	1	0	1	0	1	1	1	1	1	1	0	0	0	0
ENSG00000183520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
ENSG00000183615	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	0	0
ENSG00000183688	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000183793	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1	1	0	0	0	0	0	1	0	0	1	1	0	0	1	1	1	0	0
ENSG00000183813	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000183876	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000183878	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	0	0	0	0
ENSG00000183888	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0
ENSG00000183889	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000183908	0	0	0	0	1	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000183914	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000184007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000184113	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000184206	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000184209	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	0	0	1	0	0	0	1	1	0	0	1	1	1	0	0	0
ENSG00000184293	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0
ENSG00000184344	1	1	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000184363	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
ENSG00000184381	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000184584	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000184811	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0
ENSG00000184956	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185101	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185128	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185189	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000185201	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0
ENSG00000185303	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185304	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185332	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185467	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185499	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185559	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185630	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1

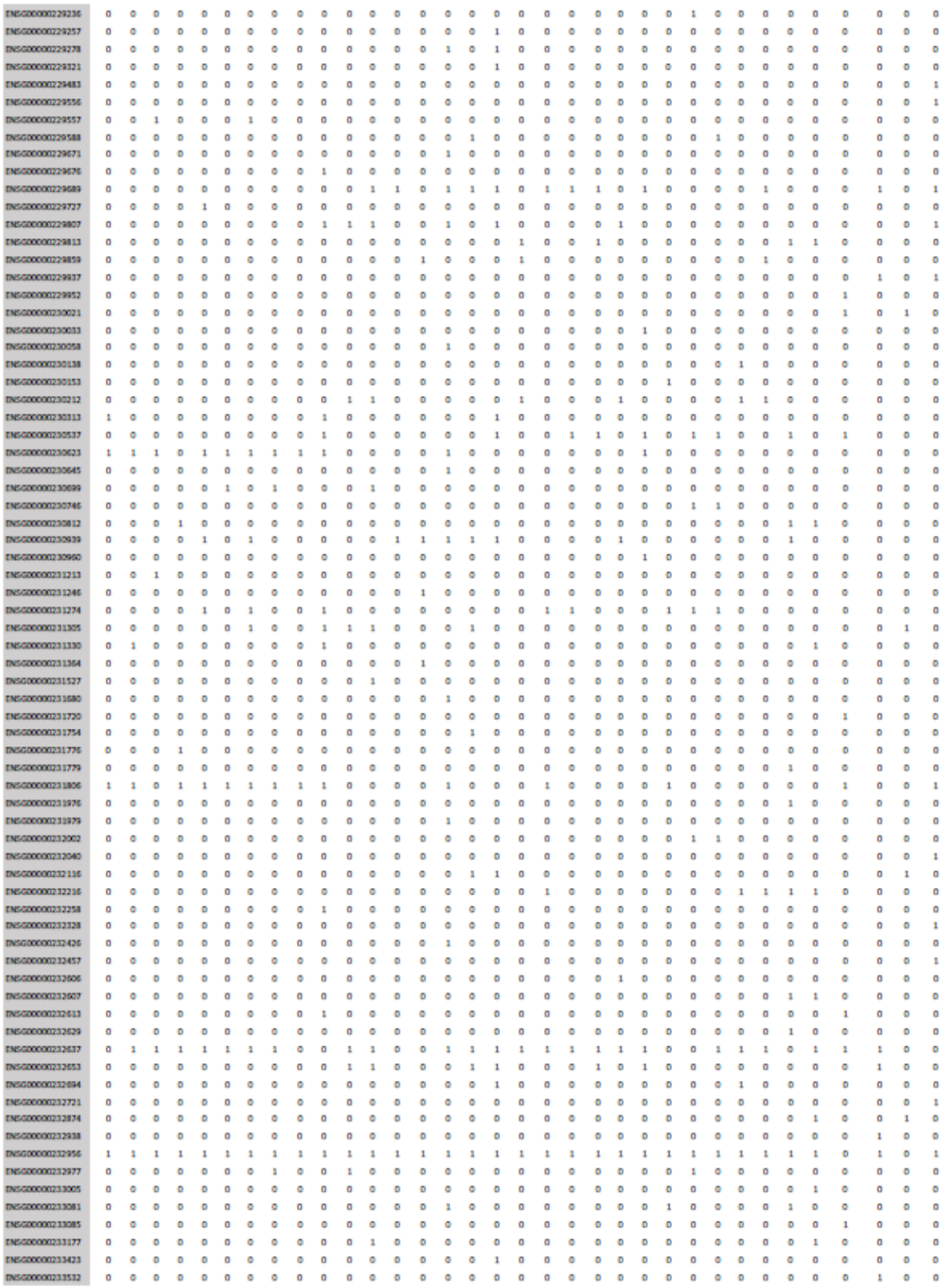
ENSG00000185730	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
ENSG00000185823	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185864	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
ENSG00000185897	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000185972	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000186009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000186051	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000186197	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000186198	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000186260	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1
ENSG00000186350	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000186399	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000186439	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000186487	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000186564	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000186810	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000186818	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000186919	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000186973	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000187185	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000187210	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000187268	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000187288	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	0	0
ENSG00000187325	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000187492	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000187550	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000187556	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000187569	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000187595	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000187642	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ENSG00000187688	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000187823	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000187867	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
ENSG00000187912	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
ENSG00000188011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
ENSG00000188033	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000188175	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ENSG00000188199	0	0	1	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1	0
ENSG00000188257	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
ENSG00000188263	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0
ENSG00000188283	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	1	1
ENSG00000188352	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	1	0	1	0	0	1	1	1	0	1	0	0
ENSG00000188385	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0
ENSG00000188389	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0
ENSG00000188511	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ENSG00000188525	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000188529	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
ENSG00000188626	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000188716	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000188732	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000188766	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000188897	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
ENSG00000188994	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000189001	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000189056	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000189058	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0
ENSG00000189099	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000189120	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000189143	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000189181	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000189182	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000189325	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
ENSG00000189332	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000189366	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000189369	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000196109	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	0	1	0	0	0	0
ENSG00000196110	0	0	0	0	0	1	1	1	1	0	1	1	1	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	1	1	1
ENSG00000196126	0	1	0	0	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000196132	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000196172	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000196350	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000196408	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
ENSG00000196415	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

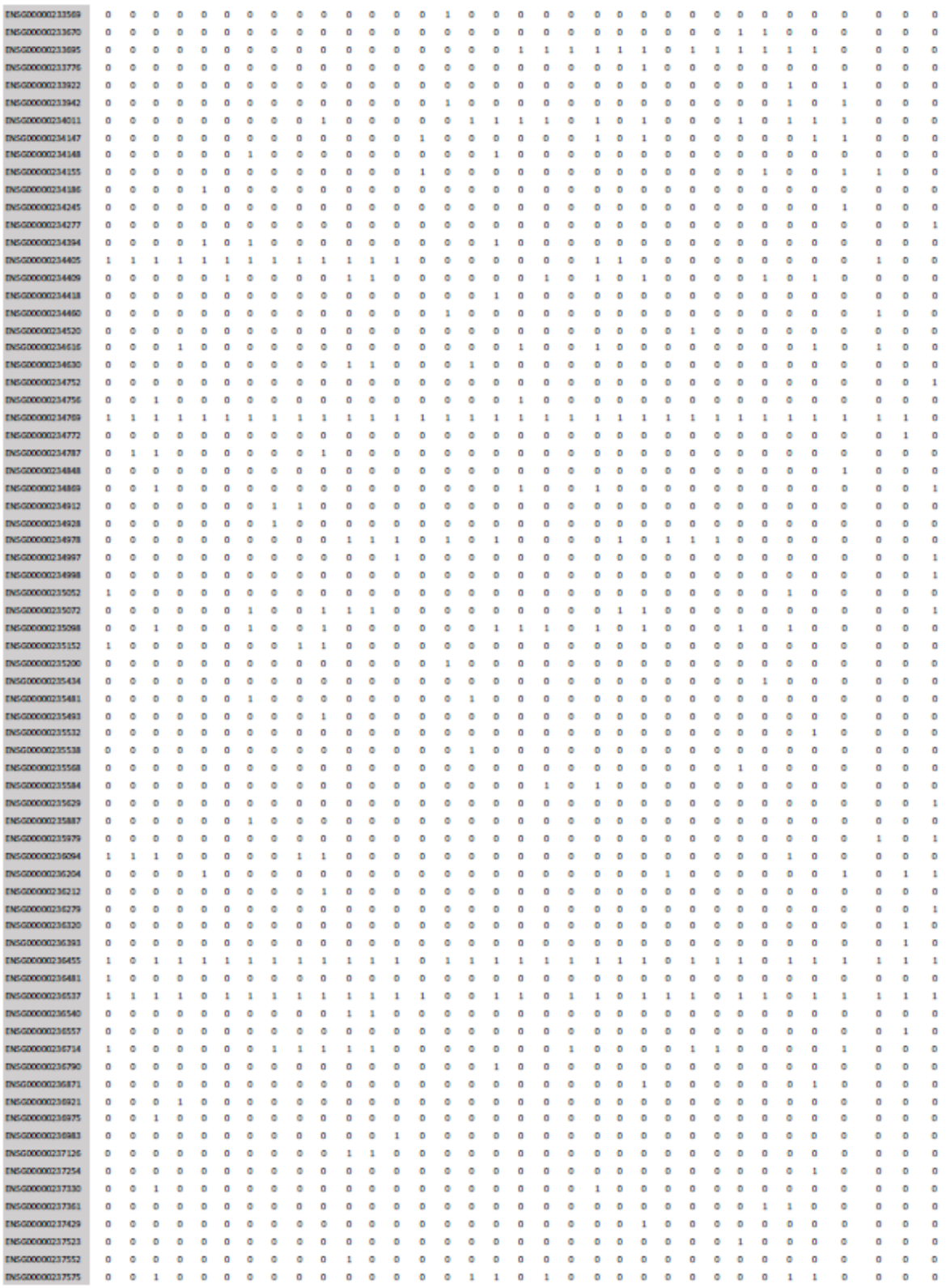












ENSG00000237585	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000237807	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
ENSG00000237836	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	1	1	0	0	0	1
ENSG00000237844	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000237877	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000237899	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
ENSG00000237927	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ENSG00000238009	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
ENSG00000238083	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0	0
ENSG00000238099	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000238113	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000238142	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000238158	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ENSG00000238178	0	1	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000238186	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000238217	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000238244	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000240045	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000240468	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000240497	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000240567	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000240907	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000241111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000241151	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ENSG00000241224	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000241666	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000241743	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000241770	0	0	0	0	0	1	1	0	1	1	1	1	1	0	1	0	0	0	0	1	0	0	0	1	1	1	0	1	0	0	0	0	0	1
ENSG00000242258	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000242366	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
ENSG00000242384	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000242498	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000242866	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000243290	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	1	1	1	0	0	0	0
ENSG00000243629	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ENSG00000243716	1	1	1	0	1	1	1	0	0	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
ENSG00000243955	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000243961	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	1	1	0	1	0	0	0	1	0	0	1
ENSG00000244117	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	1	0	0	0	0
ENSG00000244209	1	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	0	0	1
ENSG00000244215	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000244567	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000244599	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000244649	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000244694	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000245910	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
ENSG00000245923	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000245954	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000246705	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000246898	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
ENSG00000247595	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1
ENSG00000248103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ENSG00000248119	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000248161	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000248168	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ENSG00000248184	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000248202	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000248255	0	0	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000248272	1	1	0	1	1	0	0	0	0	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0
ENSG00000248323	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ENSG00000248663	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000248740	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000248787	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	0	1
ENSG00000248841	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1	0
ENSG00000248874	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000248994	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0
ENSG00000249069	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
ENSG00000249137	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000249201	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000249307	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
ENSG00000249378	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000249406	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000249425	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

ENSG00000249464	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000249523	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0
ENSG00000249548	1	1	1	0	1	0	1	0	0	1	1	1	0	0	0	1	0	0	0	1	0	1	1	1	1	0	0	0	0	1	0	0	0
ENSG00000249550	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000249566	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000249743	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000249803	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000249862	1	0	1	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000249945	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000249994	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000250131	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0
ENSG00000250166	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250210	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
ENSG00000250237	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250266	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250288	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1	0	1	0	0	0	0	1	0	0	0	1
ENSG00000250328	1	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250360	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250456	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000250490	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250508	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250509	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250584	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250682	1	1	1	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250689	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250722	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250729	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250763	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ENSG00000250829	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
ENSG00000250934	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250937	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000250986	1	1	0	0	0	1	1	0	1	1	0	0	0	1	0	1	0	0	0	1	0	0	1	1	1	0	1	0	0	1	0	0	1
ENSG00000250993	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000251039	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
ENSG00000251179	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000251230	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000251417	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000251448	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0
ENSG00000251531	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000251532	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000251555	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000251565	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000251574	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000252419	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000253154	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000253300	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000253317	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000253364	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000253669	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
ENSG00000253686	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000253696	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
ENSG00000253715	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000253718	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000253745	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000253768	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000253830	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000253877	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000253888	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000253931	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ENSG00000253959	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000253967	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000253983	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000254038	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000254164	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000254202	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000254416	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000254429	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000254610	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000254656	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000254731	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000254739	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	1	0	0	1	0
ENSG00000254744	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0	0

ENSG00000254802	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
ENSG00000254812	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1	1
ENSG00000254859	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	0
ENSG00000254872	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
ENSG00000254874	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000254919	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000254969	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0
ENSG00000255007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000255090	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ENSG00000255320	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
ENSG00000255470	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000255516	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000255524	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000255545	0	0	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000255633	0	0	0	0	0	1	1	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000255691	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ENSG00000255760	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0
ENSG00000255772	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ENSG00000255774	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ENSG00000255775	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000255874	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000255980	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000255994	1	1	0	0	1	1	1	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000256039	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
ENSG00000256045	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000256061	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000256162	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ENSG00000256166	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000256235	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0
ENSG00000256529	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	0	1	1	1	1	0	0	1	1	1	1	0	0	0	0
ENSG00000256540	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
ENSG00000256575	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000256618	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000256672	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
ENSG00000256733	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ENSG00000256848	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0
ENSG00000256850	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
ENSG00000256892	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
ENSG00000256955	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000256969	1	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	
ENSG00000256995	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
ENSG00000257048	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
ENSG00000257061	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	
ENSG00000257082	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
ENSG00000257083	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000257084	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	
ENSG00000257086	1	1	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	1		
ENSG00000257253	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000257435	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000257513	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
ENSG00000257599	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000257817	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000258017	0	1	0	0	1	0	0	0	0	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	
ENSG00000258066	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000258122	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	
ENSG00000258302	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	
ENSG00000258366	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
ENSG00000258399	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	
ENSG00000258410	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
ENSG00000258460	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
ENSG00000258593	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
ENSG00000258602	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	
ENSG00000258609	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000258616	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000258675	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
ENSG00000258701	0	0	0	0	1	1	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	
ENSG00000258733	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000258743	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000258744	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000258761	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	
ENSG00000258791	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	
ENSG00000258809	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ENSG00000258844	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Supplementary Table B-1. A list of all highly expressed methylated promoter genes and their expression status (0 = unexpressed; 1 = expressed) across 34 tissue types.

Appendix C: Supplemental Material from Chapter 4

Domain + Motif

TRAIN → TEST ↓	Fly	Human
Fly	68	
Human	63	74

Domain + Motif + DDI + DLI

TRAIN → TEST ↓	Fly	Human
Fly	70	
Human		79

Domain + Motif + DDI + DLI (with weighted kernels)

TRAIN → TEST ↓	Fly	Human
Fly	72	
Human		

Supplemental Figure C-1. AUROC with 5 fold cross-validation of pairwise SVM method. The figure depicts 3 distinct sets of kernel combinations input as features to the pairwise SVM method, and the resulting within and cross-species accuracies obtained. *Analyses pertaining to this figure are still in progress.*

BIBLIOGRAPHY

Alkhatib SG, Landry JW. 2011. The Nucleosome Remodeling Factor. *FEBS Lett* **585**: 3197–3207.

Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. 2013. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**: 360–363.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–61.

Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T. 2012. FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res* **40**: 580–584.

Askew C, Sellam A, Epp E, Hogues H, Mullick A, Nantel A, Whiteway M. 2009. Transcriptional regulation of carbohydrate metabolism in the human pathogen Candida albicans. *PLoS Pathog* **5**.

Baker CR, Tuch BB, Johnson AD. 2011. Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc Natl Acad Sci U S A* **108**: 7493–8.

Banerjee M, Thompson D. 2008. UME6, a novel filament-specific regulator of Candida albicans hyphal extension and virulence. *Mol Biol Cell* **19**: 1354–1365.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299–308.

Bannister AJ, Kouzarides T. 2011. Regulation of chromatin by histone

modifications. *Cell Res* **21**: 381–395.

Bannister AJ, Kouzarides T. 2005. Reversing histone methylation. *Nature* **436**: 1103–1106.

Bartholomew CR, Suzuki T, Du Z, Backues SK, Jin M, Lynch-Day M a., Umekawa M, Kamath a., Zhao M, Xie Z, et al. 2012. Ume6 transcription factor is part of a signaling cascade that regulates autophagy. *Proc Natl Acad Sci* **109**: 11206–11210.

Bartova E, Krejci J, Harnicarova A, Galiova G, Kozubek S. 2008. Histone modifications and nuclear architecture: a review. *J Histochem Cytochem* **56**: 711–721.

Basso K, Margolin A a, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. 2005. Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37**: 382–390.

Baylin S, Bestor TH. 2002. Altered methylation patterns in cancer cell genomes: Cause or consequence? *Cancer Cell* **1**: 299–305.

Ben-Hur A, Noble WS. 2005. Kernel methods for predicting protein-protein interactions. *Bioinformatics* **21**: i38–i46.

Bensen ES, Filler SG, Berman J. 2002. A Forkhead Transcription Factor Is Important for True Hyphal as well as Yeast Morphogenesis in Candida albicans. *Eukaryot Cell* **1**: 787–798.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Arthur L, Ecker JR, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**: 1045–

1048.

Bhardwaj N, Kim PM, Gerstein MB. 2010. Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Sci Signal* **3**: ra79.

Bhattacharya A, Warner JR. 2008. Tbf1 or not Tbf1? *Mol Cell* **29**: 537–8.

Bird A. 1992. The essentials of DNA methylation. *Cell* **70**: 5–8.

Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**: 91–99.

Bird AP. 1984. DNA methylation versus gene expression. *J Embryol Exp Morphol* **83 Suppl**: 31–40.

Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D. 2014. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res* **42**: D396–D400.

Bock JR, Gough D a. 2001. Predicting protein–protein interactions from primary structure. *Bioinformatics* **17**: 455–460.

Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A, Cedar H. 1994. Sp1 elements protect a CpG island from de novo methylation. *Nature* **371**: 435–8.

Breiman L. 2001. Random Forests. *Eur J Math* **45**: 5–32.

Brown TA. 2002. *Genomes*. 2, illustr ed. BIOS Scientific.

Bulger M, Groudine M. 2010. Enhancers: The abundance and function of

regulatory sequences beyond promoters. *Dev Biol* **339**: 250–257.

Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res* **26**: 301–314.

Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, et al. 2015. The BioGRID interaction database: 2015 Update. *Nucleic Acids Res* **43**: 470–478.

Chen G, Jensen ST, Stoeckert CJ. 2007. Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol* **8**: R4.

Connelly CF, Wakefield J, Akey JM. 2014. Evolution and Genetic Architecture of Chromatin Accessibility and Function in Yeast. *PLoS Genet* **10**.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of transcription start sites from nascent RNA supports a unified architecture of mammalian promoters and enhancers. *Nat Genet* **46**: 1311–1320.

Creighton CJ, Morgan M, Gunaratne PH, Wheeler DA, Gibbs RA, Gordon Robertson A, Chu A, Beroukhim R, Cibulskis K, Signoretti S, et al. 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**: 43–49.

Crespo J, Powers T. 2002. The TOR-controlled transcription activators GLN3, RTG1, and RTG3 are regulated in response to intracellular levels of glutamine. *Proc Natl Acad Sci U S A* **99**: 6784–6789.

Crespo JL, Hall MN. 2002. Elucidating TOR signaling and rapamycin action: lessons from Saccharomyces cerevisiae. *Microbiol Mol Biol Rev* **66**: 579–

591.

Csárdi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal Complex Syst* **1695**: 1695.

Dearden PK. 2015. Origin and evolution of the enhancer of split complex. *BMC Genomics* **16**: 712.

De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei C-L, Natoli G. 2010. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol* **8**: e1000384.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes Dev* **25**: 1010–1022.

Deaton AM, Webb S, Kerr ARW, Illingworth RS, Guy J, Andrews R, Bird A. 2011. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res* **21**: 1074–1086.

Dekker J, Marti-Renom MA, Mirny LA. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**: 390–403.

Dennis G, Sherman BT, Hosack D a, Yang J, Gao W, Lane HC, Lempicki R a. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**: P3.

Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, Born D, Speck T, Krüger D, Grebnev G, Kuban M, et al. 2014. The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res* **42**: D259-66.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012.

Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–80.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schübeler D. 2015. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**: 575–579.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox T V, Davies R, Down TA, et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38**: 1378–85.

Elango N, Yi S V. 2008. DNA methylation and structural and functional bimodality of vertebrate promoters. *Mol Biol Evol* **25**: 1602–1608.

Elfving N, Chereji R V., Bharatula V, Björklund S, Morozov A V., Broach JR. 2014. A dynamic interplay of nucleosome and Msn2 binding regulates kinetics of gene activation and repression following stress. *Nucleic Acids Res* **42**: 5468–5482.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216.

Feng D, DuMontier C, Pollak MR. 2015. The role of alpha-actinin-4 in human

kidney disease. *Cell Biosci* **5**: 44.

Field Y, Fondufe-Mittendorf Y, Moore IK, Mieczkowski P, Kaplan N, Lubling Y, Lieb JD, Widom J, Segal E. 2009. Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nat Genet* **41**: 438–445.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: The protein families database. *Nucleic Acids Res* **42**: 222–230.

Florence B, Guichet A, Ephrussi A, Laughon A. 1997. Ftz-F1 is a cofactor in Ftz activation of the Drosophila engrailed gene. *Development* **847**: 839–847.

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, Von Mering C, et al. 2013. STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**: 808–815.

Frankel AD, Kim PS. 1991. Modular structure of transcription factors: Implications for gene regulation. *Cell* **65**: 717–719.

Gaston K, Jayaraman P-S. 2003. Transcriptional repression in eukaryotes: repressors and repression mechanisms. *Cell Mol Life Sci* **60**: 721–741.

Giannattasio S, Liu Z, Thornton J, Butow R a. 2005. Retrograde response to mitochondrial dysfunction is separable from TOR1/2 regulation of retrograde gene expression. *J Biol Chem* **280**: 42528–35.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental

transcriptome of Drosophila melanogaster. *Nature* **471**: 473–479.

Guichet A, Copeland JWR, Erdélyi M, Hlousek D, Závorszky P, Ho J, Brown S, Percival-Smith A, Krause HM, Ephrussi A. 1997. The nuclear receptor homologue Ftz-F1 and the homeodomain protein Ftz are mutually dependent cofactors. *Nature* **385**: 548–552.

Guillaumet-Adkins A, Richter J, Odero MD, Sandoval J, Agirre X, Catala A, Esteller M, Prósper F, Calasanz M, Buño I, et al. 2014. Hypermethylation of the alternative AWT1 promoter in hematological malignancies is a highly specific marker for acute myeloid leukemias despite high expression levels. *J Hematol Oncol* **7**: 4.

Guo Y, Yu L, Wen Z, Li M. 2008. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res* **36**: 3025–3030.

Ha M, Kim E-D, Chen ZJ. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc Natl Acad Sci U S A* **106**: 2295–2300.

Haaf T, Schmid M. 1991. Chromosome topology in mammalian interphase nuclei. *Exp Cell Res* **192**: 325–332.

Habib N, Wapinski I, Margalit H, Regev A, Friedman N. 2012. A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. *Mol Syst Biol* **8**: 619.

Han H, Cortez CC, Yang X, Nichols PW, Jones P a., Liang G. 2011. DNA methylation directly silences genes with non-CpG island promoters and

establishes a nucleosome occupied promoter. *Hum Mol Genet* **20**: 4299–4310.

Hanna-Rose W, Hansen U. 1996. Active repression mechanisms of eukaryotic transcription repressors. *Trends Genet* **12**: 229–234.

Harmston N, Lenhard B. 2013. Chromatin and epigenetic features of long-range gene regulation. *Nucleic Acids Res* **41**: 7185.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–74.

Heffer A, Shultz JW, Pick L. 2010. Surprising flexibility in a conserved Hox transcription factor over 550 million years of evolution. *Proc Natl Acad Sci U S A* **107**: 18040–5.

Heffer A, Xiang J, Pick L. 2013. Variation and constraint in Hox gene evolution. *Proc Natl Acad Sci U S A* **110**: 2211–6.

Hendzel MJ, Wei Y, Mancini MA, Van Hooser A, Ranalli T, Brinkley BR, Bazett-Jones DP, Allis CD. 1997. Mitosis-specific phosphorylation of histone H3 initiates primarily within pericentromeric heterochromatin during G2 and spreads in an ordered fashion coincident with mitotic chromosome condensation. *Chromosoma* **106**: 348–360.

Hinnebusch AG. 2005. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* **59**: 407–450.

Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, Purisima E, Nantel A,

Whiteway M. 2008. Transcription factor substitution during the evolution of fungal ribosome regulation. *Mol Cell* **29**: 552–62.

Hoivik E a., Witsoe SL, Bergheim IR, Xu Y, Jakobsson I, Tengholm A, Doskeland SO, Bakke M. 2013. DNA Methylation of Alternative Promoters Directs Tissue Specific Expression of Epac2 Isoforms ed. E. Ballestar. *PLoS One* **8**: e67925.

Homann OR, Dea J, Noble SM, Johnson AD. 2009. A phenotypic profile of the Candida albicans regulatory network. *PLoS Genet* **5**: e1000783.

Hon GC, Hawkins RD, Ren B. 2009. Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet* **18**: R195-201.

Huang Q, You Z, Zhang X, Zhou Y. 2015. Prediction of protein-protein interactions with clustered amino acids and weighted sparse representation. *Int J Mol Sci* **16**: 10855–69.

Ihmels J, Bergmann S, Gerami-Nejad M, Yanai I, McClellan M, Berman J, Barkai N. 2005. Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* **309**: 938–40.

Illingworth RS, Bird AP. 2009. CpG islands - “A rough guide.” *FEBS Lett* **583**: 1713–1720.

Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP. 2010. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome ed. W. Reik. *PLoS Genet* **6**: e1001134.

Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo

K, Rongione M, Webster M, et al. 2009a. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41**: 178–186.

Irizarry RA, Wu H, Feinberg AP. 2009b. A species-generalized probabilistic model-based definition of CpG islands. *Mamm Genome* **20**: 674–680.

Isalan M, Lemerle C, Michalodimitrakis K, Horn C, Beltrao P, Raineri E, Garriga-Canut M, Serrano L. 2008. Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**: 840–845.

Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**: 318–356.

Jenuwein T, Allis CD. 2001. Translating the histone code. *Science (80-)* **293**: 1074–1080.

Jones KA, Kadonaga JT, Rosenfeld PJ, Kelly TJ, Tjian R. 1987. A cellular DNA-binding protein that activates eukaryotic transcription and DNA replication. *Cell* **48**: 79–89.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**: 484–492.

Jones PA, Baylin SB. 2007. The Epigenomics of Cancer. *Cell* **128**: 683–692.

Jonkers I, Lis JT. 2015. Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**: 11–13.

Kadonaga JT, Carner KR, Masiarz FR, Tjian R. 1987. Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. *Cell* **51**: 1079–1090.

Kadonaga JT, Jones KA, Tjian R. 1986. Promoter-specific activation of RNA polymerase II transcription by Sp1. *Trends Biochem Sci* **11**: 20–23.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. 2009. Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**: D767--D772.

Kim I, Lee H, Han SK, Kim S. 2014. Linear motif-mediated interactions have contributed to the evolution of modularity in complex protein interaction networks. *PLoS Comput Biol* **10**: e1003881.

Kim T, Hemberg M, Gray J, Costa A. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182–187.

King M, Wilson A. 1975. Evolution at two levels in humans and chimpanzees. *Science (80-)* **188**: 107–116.

Kireeva ML, Walter W, Tchernajenko V, Bondarenko V, Kashlev M, Studitsky VM. 2002. Nucleosome remodeling induced by RNA polymerase II: loss of the H2A/H2B dimer during transcription. *Mol Cell* **9**: 541–552.

Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, Fulton LL, Dooling DJ, Ding L, Mardis ER, et al. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* **490**: 61–70.

Kouzarides T. 2007. Chromatin modifications and their function. *Cell* **128**: 693–705.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. 2015. Integrative analysis

of 111 reference human epigenomes. *Nature* **518**: 317–330.

Langsrud Ø. 2005. Rotation tests. *Stat Comput* **15**: 53–60.

Lapointe F-J, Garland J. 2001. A Generalized Permutation Model for the Analysis of Cross-Species Data. *J Classif* **18**: 109.

Larsen F, Gundersen G, Lopez R, Prydz H. 1992. CpG islands as gene markers in the human genome. *Genomics* **13**: 1095–1107.

Latchman DS. 1997. Transcription factors: an overview. *Int J Biochem Cell Biol* **29**: 1305–1312.

Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, et al. 2010. Dynamic changes in the human methylome during differentiation. *Genome Res* **20**: 320–331.

Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, Whiteway M. 2010. Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS Biol* **8**: e1000329.

Lavoie H, Hogues H, Whiteway M. 2009. Rearrangements of the transcriptional regulatory networks of metabolic pathways in fungi. *Curr Opin Microbiol* **12**.

Lay FD, Liu Y, Kelly TK, Witt H, Farnham PJ, Jones PA, Berman BP. 2015. The role of DNA methylation in directing the functional organization of the cancer epigenome. *Genome Res* **25**: 467–477.

Levy S, Hannenhalli S. 2002. Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome* **13**: 510–4.

Li J-T, Hou G-Y, Kong X-F, Li C-Y, Zeng J-M, Li H-D, Xiao G-B, Li X-M, Sun X-W. 2015. The fate of recent duplicated genes following a fourth-round whole

genome duplication in a tetraploid fish, common carp (Cyprinus carpio). *Sci Rep* **5**: 8199.

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schübeler D. 2011. Identification of genetic elements that autonomously determine DNA methylation states. *Nat Genet* **43**: 1091–1097.

Linehan W, Srinivasan R, Schmidt L. 2010. The genetic basis of kidney cancer: a metabolic disease. *Nat Rev Urol* **7**: 277–285.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315–322.

Liu H, Liu Y, Zhang J-T. 2008. A new mechanism of drug resistance in breast cancer cells: fatty acid synthase overexpression-mediated palmitate overproduction. *Mol Cancer Ther* **7**: 263–70.

Liu PZ. 2005. even-skipped is not a pair-rule gene but has segmental and gap-like functions in Oncopeltus fasciatus, an intermediate germband insect. *Development* **132**: 2081–2092.

Lodish H. 2008. *Molecular Cell Biology*. 4th editio. W. H. Freeman.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**: 1632–1635.

Lutz B, Lu HC, Eichele G, Miller D, Kaufman TC. 1996. Rescue of Drosophila labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved. *Genes Dev* **10**: 176–184.

Lynch M. 2006. The origins of eukaryotic gene structure. *Mol Biol Evol* **23**: 450–68.

Lynch VJ, Wagner GP. 2008. RESURRECTING THE ROLE OF TRANSCRIPTION FACTOR CHANGE IN DEVELOPMENTAL EVOLUTION. *Evolution (N Y)* **62**: 2131–2154.

Maeso I, Irimia M, Tena JJ, Casares F, Gómez-Skarmeta JL. 2013. Deep conservation of cis-regulatory elements in metazoans. *Philos Trans R Soc Lond B Biol Sci* **368**: 20130020.

Mallick J, Whiteway M. 2013. The evolutionary rewiring of the ribosomal protein transcription pathway modifies the interaction of transcription factor heteromer Ifh1-Fhl1 (interacts with forkhead 1-forkhead-like 1) with the DNA-binding specificity element. *J Biol Chem* **288**: 17508–19.

Mancini-DiNardo D, Steele SJS, Ingram RS, Tilghman SM. 2003. A differentially methylated region within the gene Kcnq1 functions as an imprinted promoter and silencer. *Hum Mol Genet* **12**: 283–294.

Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. 2016. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Meth* **13**: 366–370.

Martchenko M, Levitin A, Hogues H, Nantel A, Whiteway M. 2007. Transcriptional rewiring of fungal galactose-metabolism circuitry. *Curr Biol* **17**: 1007–13.

Martino D, Saffery R. 2015. Characteristics of DNA methylation and gene expression in regulatory features on the Infinium 450k Beadchip. 1–7.

Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**: 29–59.

Masuda H, Zhang D, Bartholomeusz C, Doihara H, Hortobagyi GN, Ueno NT. 2012. Role of epidermal growth factor receptor in breast cancer. *Breast Cancer Res Treat* **136**: 331–345.

Matys V, Kel-Margoulis O V, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, et al. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**: D108-10.

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**: 253–257.

McGinnis N, Kuziora MA, McGinnis W. 1990. Human Hox-4.2 and Drosophila deformed encode similar regulatory specificities in Drosophila embryos and larvae. *Cell* **63**: 969–976.

Menafra R, Brinkman AB, Matarese F, Franci G, Bartels SJJ, Nguyen L, Shimbo T, Wade P a., Hubner NC, Stunnenberg HG. 2014. Genome-wide binding of MBD2 reveals strong preference for highly methylated loci. *PLoS One* **9**: 1–12.

Mendizabal I, Yi S V. 2016. Whole-genome bisulfite sequencing maps from multiple human tissues reveal novel CpG islands associated with tissue-specific regulation. *Hum Mol Genet* **25**: 69–82.

Mirny LA. 2010. Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci* **107**: 22534–22539.

Moarii M, Boeva V, Vert J-P, Reyal F. 2015. Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics* **16**: 873.

Moazed D. 2001. Common Themes in Mechanisms of Gene Silencing. *Mol Cell* **8**: 489–498.

Moore AD, Bornberg-Bauer E. 2012. The dynamics and evolutionary potential of domain loss and emergence. *Mol Biol Evol* **29**: 787–96.

Müeller-Storm HP, Sogo JM, Schaffner W. 1989. An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge. *Cell* **58**: 767–777.

Nagarajan RP, Zhang B, Bell RJ a, Johnson BE, Olshen AB, Sundaram V, Li D, Graham AE, Diaz A, Fouse SD, et al. 2014. Recurrent epimutations activate gene body promoters in primary glioblastoma. *Genome Res* **24**: 761–774.

Newell-Price J, Clark AJL, King P. 2000. DNA Methylation and Silencing of Gene Expression. *Trends Endocrinol Metab* **11**: 142–148.

Nicholls S, Straffon M, Enjalbert B, Nantel A, Macaskill S, Whiteway M, Brown AJP. 2004. Msn2- and Msn4-like transcription factors play no obvious roles in the stress responses of the fungal pathogen Candida albicans. *Eukaryot Cell* **3**: 1111–23.

Ntini E, Järvelin AI, Bornholdt J, Chen Y, Boyd M, Jørgensen M, Andersson R, Hoof I, Schein A, Andersen PR, et al. 2013. Polyadenylation site–induced

decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol* **20**: 923–928.

Ozonov E a, van Nimwegen E. 2013. Nucleosome free regions in yeast promoters result from competitive binding of transcription factors that interact with chromatin modifiers. *PLoS Comput Biol* **9**: e1003181.

Park Y, Marcotte EM. 2012. Flaws in evaluation schemes for pair-input computational predictions. *Nat Methods* **9**: 1134–1136.

Patel NH, Ball EE, Goodman CS. 1992. Changing role of even-skipped during the evolution of insect pattern formation. *Nature* **357**: 339–342.

Peterson CL, Laniel M-A. 2004. Histones and histone modifications. *Curr Biol* **14**: R546–R551.

Pfreundt U, James DP, Tweedie S, Wilson D, Teichmann SA, Adryan B. 2009. FlyTF: Improved annotation and enhanced functionality of the Drosophila transcription factor database. *Nucleic Acids Res* **38**: 443–447.

Phatnani HP, Greenleaf a L. 2006. Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev* **20**: 2922–2936.

Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N, et al. 2006. PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics* **7**: 365.

Pollard SM, Stricker SH, Beck S. 2009. A Shore Sign of Reprogramming. *Cell Stem Cell* **5**: 571–572.

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory

evolution. *Proc Natl Acad Sci U S A* **104 Suppl**: 8605–8612.

Pryszcz LP, Huerta-Cepas J, Gabald?n T. 2011. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res* **39**: e32–e32.

Ptashne M. 2014. The Chemistry of Regulation of Genes and Other Things. *J Biol Chem* **289**: 5417–5435.

Ptashne M, Gann A. 1997. Transcriptional activation by recruitment. *Nature* **386**: 569–577.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Rantala JK, Edgren H, Lehtinen L, Wolf M, Kleivi K, Vollan HKM, Aaltola A-R, Laasola P, Kilpinen S, Saviranta P, et al. 2010. Integrative functional genomics analysis of sustained polyploidy phenotypes in breast cancer cells identifies an oncogenic profile for GINS2. *Neoplasia* **12**: 877–888.

Robertson KD. 2000. DNA methylation: past, present and future directions. *Carcinogenesis* **21**: 461–467.

Rokas A, Hittinger CT. 2007. Transcriptional rewiring: the proof is in the eating. *Curr Biol* **17**: R626-8.

Romano L a, Wray G a. 2003. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. *Development* **130**: 4187–4199.

Roy S, Wapinski I, Pfiffner J, French C. 2013. Arboretum: Reconstruction and analysis of the evolutionary history of condition-specific transcriptional

modules. *Genome Res* 1039–1050.

Saenko S V, Marialva MS, Beldade P. 2011. Involvement of the conserved Hox gene Antennapedia in the development and evolution of a novel trait. *Evodevo* **2**: 9.

Sarda S, Hannenhalli S. 2015. High throughput identification of cis-regulatory rewiring events in yeast. *Mol Biol Evol* **32**: 3047–3063.

Saxonov S, Berg P, Brutlag DL. 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**: 1412–1417.

Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. 2012. Hippie: Integrating protein interaction networks with experiment based quality scores. *PLoS One* **7**: 1–8.

Schirm S, Jiricny J, Schaffner W. 1987. The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity. *Genes & Dev* **1**: 65–74.

Schwalb B, Michel M, Zacher B, Fruhauf K, Demel C, Tresch A, Gagneur J, Cramer P. 2016. TT-seq maps the human transient transcriptome. *Science (80-)* **352**: 1225–1228.

Schwarz LJ, Fox EM, Balko JM, Garrett JT, Kuba MG, Estrada MV, González-Angulo AM, Mills GB, Red-Brewer M, Mayer IA, et al. 2014. LYN-activating mutations mediate antiestrogen resistance in estrogen receptor–positive breast cancer. *J Clin Invest* **124**: 5490–5502.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. 2003.

Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**: 166–76.

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. 2007. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A* **104**: 4337–41.

Shilpa V, Bhagat R, Premalata CS, Pallavi VR, Ramesh G, Krishnamoorthy L. 2014. Relationship between promoter methylation & tissue expression of MGMT gene in ovarian cancer. *Indian J Med Res* **140**: 616–623.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Sinha AU, Meller J. 2007. Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics* **8**: 82.

Smale ST, Baltimore D. 1989. The “initiator” as a transcription control element. *Cell* **57**: 103–113.

Smale ST, Kadonaga JT. 2003. The RNA Polymerase II Core Promoter. *Annu Rev Biochem* **72**: 449–479.

Smith Z, Chan M, Mikkelsen T, Gu H. 2012. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**: 339–344.

Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**: 613–626.

Sproul D, Kitchen RR, Nestor CE, Dixon JM, Sims AH, Harrison DJ, Ramsahoye BH, Meehan RR. 2012. Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biol* **13**: R84.

Sproul D, Nestor C, Culley J, Dickson JH, Dixon JM, Harrison DJ, Meehan RR, Sims AH, Ramsahoye BH. 2011. Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. *Proc Natl Acad Sci U S A* **108**: 4364–4369.

Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelhing-Sun S, Lee K, et al. 2014. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**: 365–370.

Storey JD. 1995. A direct approach to false discovery rates. *J R Stat B* **64**: 479–498.

Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* **403**: 41–45.

Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, et al. 2012. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet* **8**.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**: 465–476.

Tanay A, Regev A, Shamir R. 2005. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A* **102**: 7203–8.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462–470.

Thompson D a, Roy S, Chan M, Styczynsky MP, Pfiffner J, French C, Socha A, Thielke A, Napolitano S, Muller P, et al. 2013. Evolutionary principles of modular gene regulation in yeasts. *Elife* **2**: e00603.

Trojer P, Reinberg D. 2007. Facultative Heterochromatin: Is There a Distinctive Molecular Signature? *Mol Cell* **28**: 1–13.

Tsankov A, Yanagisawa Y, Rhind N, Regev A, Rando OJ. 2011. Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. *Genome Res* **21**: 1851–62.

Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. 2010. The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol* **8**.

Tuch BB, Li H, Johnson AD. 2008. Evolution of Eukaryotic Transcription Circuits. *Science (80-)* **319**: 1797–1800.

Tucker KL. 2001. Methylated Cytosine and the Brain: A New Base for Neuroscience. *Neuron* **30**: 649–652.

Ucar D, Beyer A, Parthasarathy S, Workman CT. 2009. Predicting functionality of protein-DNA interactions by integrating diverse evidence. *Bioinformatics* **25**: i137-44.

van Dongen S, Abreu-Goodger C. 2012. Using MCL to extract clusters from

networks. *Methods Mol Biol* **804**: 281–95.

van Eijk KR, de Jong S, Boks MPM, Langeveld T, Colas F, Veldink JH, de Kovel CGF, Janson E, Strengman E, Langfelder P, et al. 2012. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics* **13**: 636.

Van Vlodrop IJH, Niessen HEC, Derks S, Baldewijns MMLL, Van Criekinge W, Herman JG, Van Engeland M. 2011. Analysis of promoter CpG island hypermethylation in cancer: Location, location, location! *Clin Cancer Res* **17**: 4225–4231.

Vishnoi A, Sethupathy P, Simola D, Plotkin JB, Hannenhalli S. 2011. Genome-wide survey of natural selection on functional, structural, and network properties of polymorphic sites in Saccharomyces paradoxus. *Mol Biol Evol* **28**: 2615–27.

Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. 2014. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol* **15**: R37.

Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* **34**: 1692–1699.

Wan J, Oliver VF, Wang G, Zhu H, Zack DJ, Merbs SL, Qian J. 2015. Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. *BMC*

Genomics **16**: 49.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007a. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* **23**: i549-58.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007b. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**: 54–61.

Weatheritt RJ, Gibson TJ. 2012. Linear motifs: Lost in (pre)translation. *Trends Biochem Sci* **37**: 333–341.

Weingarten-Gabbay S, Segal E. 2014. A shared architecture for promoters and enhancers. *Nat Genet* **46**: 1253–1254.

Weirauch MT, Hughes TR. 2010. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet* **26**: 66–74.

Wilks C, Cline MS, Weiler E, Diehkans M, Craft B, Martin C, Murphy D, Pierce H, Black J, Nelson D, et al. 2014. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database (Oxford)* **2014**: 1–10.

Wittkopp PJ, Kalay G. 2011. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**: 59–69.

Woodcock CL, Ghosh RP. 2010. Chromatin Higher-order Structure and Dynamics. *Cold Spring Harb Perspect Biol* **2**: a000596–a000596.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**: 206–16.

Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP. 2010. Redefining CpG islands using hidden Markov models. *Biostatistics* **11**: 499–514.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**: 650–659.

Yellaboina S, Tasneem A, Zaykin D V., Raghavachari B, Jothi R. 2011. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res* **39**: D730–D735.

You Z-H, Chan KCC, Hu P. 2015. Predicting Protein-Protein Interactions from Primary Protein Sequences Using a Novel Multi-Scale Local Feature Representation Scheme and the Random Forest. *PLoS One* **10**: e0125811.

You Z-H, Zhu L, Zheng C-H, Yu H-J, Deng S-P, Ji Z. 2014. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinformatics* **15 Suppl 1**: S9.

Yu CY, Chou LC, Chang DT. 2010. Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics* **11**: 167.

Yu Y, Li W, Su K, Yussa M, Han W, Perrimon N, Pick L. 1997. The nuclear hormone receptor Ftz-F1 is a cofactor for the Drosophila homeodomain protein Ftz. *Nature* **385**: 552–555.

Zhang S-W, Hao L-Y, Zhang T-H. 2014. Prediction of Protein–Protein Interaction

with Pairwise Kernel Support Vector Machine. *Int J Mol Sci* **15**: 3220–3233.

Zhong Y, Holland PWH. 2011. The dynamics of vertebrate homeobox gene evolution: gain and loss of genes in mouse and human lineages. *BMC Evol Biol* **11**: 169.

Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. *Trends Genet* **24**: 481–484.

Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, et al. 2013. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**: 477–481.