SYSTEMATIC CHARACTERIZATION OF *CIS*-REGULATION IN *C.ELEGANS* USING EVOLUTIONARY CONSERVATION

by
Donavan Cheng

A dissertation submitted to the Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
September, 2009

ABSTRACT:

Comparative genomics approaches for *cis*-regulatory element detection typically rely on sequence alignment, even though recent studies show modest overlap (~50%) between confirmed regulatory elements and regions of high sequence alignability. This dissertation focuses on developing alignment-independent approaches for detecting conserved *cis*-regulatory elements and modules and is organized in three parts: In the first study, we present Flipper, a novel alignment-independent Gibbs sampling based algorithm which uses over-representation *and* evolutionary conservation equally to detect conserved DNA regulatory elements *ab initio* from orthologous sequence. Flipper performs up to 23% better than existing methods at recovering seeded motifs from synthetic test data and also recovers more known motifs from yeast, worm and fly ChIP-chip data. To discover novel regulatory motifs, we ran Flipper on promoters of sets of coexpressed genes in *C.elegans*. We focused on the ribosomal protein (RP) gene cluster, as it is highly coexpressed but yet little is known about its regulation. Flipper detected 22 motifs associated with the RP promoters , where four motifs (M546, M313, M540 and M439) were significantly conserved and specific to the RP gene cluster in *C.elegans* and its relatives *C.remanei*, *C.briggsae*, and *C.brenneri*. In our second study, we used a promoter::mCherry transcriptional reporter assay to test our predicted motifs for function. M546 severely abrogated mCherry expression when mutated in 8 out of 11 tested promoters and similarly, M313 was necessary for promoter function in 4 of 9 cases, M540 in 3 of 7 cases and M439 in 1 of 3 cases respectively. In a promoter "transplant" experiment, we demonstrated that M546 and M540 are functionally conserved and are necessary for *C.briggsae* promoters to drive mCherry expression in *C.elegans*. M546 and M540 occur in a large number of non-ribosomal promoters and we show that M546 is also necessary for function in the *mcm-7* promoter, even though its expression profile is markedly different from RPs. In the third study, we demonstrate that rules governing the organization of *cis*-regulatory elements in modules, in terms of relative spacing, positioning and orientation constraints, can also be conserved across species. Using this information, we discover a strong, conserved spacing and orientation bias in pairs of co-occurring M546 and M540 sites in RP promoters. Using a "sequence swap" experiment, we disrupted the spacing between M546 and M540 sites and showed that it has a severe effect on *rps-7* promoter function. We show that a large number of non-ribosomal promoters contain M546 and M540 sites because these sites reside in an arm of the CELE2 transposon, which happened to insert itself in these promoters. Interestingly, the M546-M540 pair in these promoters do not obey the RP spacing constraint and these promoters are not enriched in any common GO annotations, while other non-ribosomal promoters containing M546-M540 sites with the RP spacing constraint are strongly enriched for growth and development GO annotations ($p < 10^{-9}$), which are consistent with the need for RP biogenesis. In summary, using an alignment independent approach, we have identified conserved *cis*-regulatory elements necessary for RP gene expression in *C.elegans*, with the M546 and M540 motifs possibly part of a regulatory module that is involved in more general regulation of growth and early development processes.

Advisor:
Michael A. Beer, PhD
Thesis readers:
Michael A. Beer, PhD, Joel S. Bader, PhD
Thesis committee:
Michael A. Beer, PhD, Joel S. Bader, PhD, Rachel Karchin, PhD, Geraldine Seydoux, PhD and Randall R. Reed, PhD

ACKNOWLEDGEMENTS:

I would like to acknowledge my advisor, Michael A. Beer, for his guidance and help, without which this dissertation would not be possible. Thank you for your support and insight – although we may not see eye-to-eye all the time, your point of view is always appreciated and I do cherish our interactions.

I would also like to thank the members of my thesis committee, Drs. Bader, Karchin, Seydoux and Reed, for taking time off their busy schedules to participate in my thesis defense, proposal and in other discussions.

Thank you to all members of the Beer lab, past and present: Ariel Hsu, Cecilia Ng, Navneeta Bansal (our technicians), Yan Qi, Jun Kyu Rhee, Mahmoud Ghandi, Rahul Karnik, Dongwon Lee (my graduate student labmates), JT Chiang, Patrick Nguyen, Andrew Pao (!), Tuo Li, Murat Bligel, Rohit Dayal, JP Cardenas (our undergrads).

I would also like to thank everyone else whom I have interacted with, that have inspired me and given me guidance: Betty Doan, Ashish Kapoor, Guo N. Huang and Dr. Aravinda Chakravarti.

Finally, I would like to thank my friends for their emotional support through these five years, which for better or worse, has made into who I am today. Thank you, Shawn, Eric, Kelvin, Alvina, Robin, Tim and everyone else I know. Thank you, Tito for your patience and bearing with me this past year – I just hope you have enough patience to bear with me for an even longer time to come! Most of all, thank you Debbie and Figaro, you have been a constant source of support these four years and it is because of you that I can return to our apartment every night and call it our home.

DEDICATION:

This thesis is dedicated to two people I love very much - my father, Clifford Cheng and my mother, Margaret Kwok.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: Overview and Thesis Organization

1.1 Overview

Initial analysis of the draft human genome sequence in 2001 estimated that only 1.5 to 2% of the genome codes for proteins[1], while up to 5% of the total genome is under selective pressure and conserved compared to mouse[2]. This indicates that majority of functional sequence under selective pressure is non-coding and probably of regulatory significance. Indeed, the importance of *cis*-acting regulatory sequences is underscored by studies showing that evolutionary changes in gene expression can mostly be attributed to changes in *cis*-regulation[3-6] and that relatively little innovation has occurred in the protein-coding component of vertebrate genomes[7]. These results fueled the development of the ENCODE initiative, which was set up with the initial goal of characterizing all non-coding regulatory elements in 1% of the human genome[8], but has since been extended to cover the rest of the genome. Comparative genomic approaches are a key effort in this endeavor, because functional regulatory sequences are under strong selective pressure and can be detected based on their conservation across species.

The goal of this dissertation research is to highlight multiple ways evolutionary conservation can be used to aid the identification of *cis*-regulatory elements and modules. This thesis emphasizes a systems biology approach, where algorithm predictions are used to generate experimentally testable hypotheses and experimental results are fed back to refine algorithm development. In this dissertation, we develop an alignment independent method for regulatory motif prediction, which we use to identify novel *cis*-regulatory elements associated with ribosomal protein promoters in *C.elegans*. We demonstrate that *cis*-regulatory elements follow a conserved pattern of organization in regulatory modules and show that conservation of these organizational constraints can be used to identify novel *cis*-regulatory modules.

1.2 Thesis Organization

To provide background information on *C.elegans* and emphasize the importance of *cis*-regulatory elements in the control of gene expression, we provide a brief introduction to the biology of *C.elegans* in Chapter 2 and also review different aspects of *cis*-regulation: transcription factor (TF) binding sites, nucleosome positioning signals etc. Because most of the experimental results in this thesis concern the regulation of ribosomal proteins, the last section in Chapter 2 reviews the literature on ribosomal proteins and their role in disease causation.

Chapter 3 is a review of computational methods developed to detect motifs from sequence. It is split into two sections: a) algorithms that identify statistically over-represented motifs, b) "phylogenetic" algorithms that incorporate conservation into the motif search.

Chapters 4, 5 and 6 represent the scientific body of work for this dissertation. In the first half of Chapter 4, we introduce the motivation for using an alignment independent approach for identifying conserved motifs and provide details on the structure of the Flipper algorithm which we developed. In the second half, we compare Flipper against other motif discovery algorithms on synthetic and biological data and report novel motifs predicted from coexpression data in *C.elegans*. Chapter 5 focuses on sequence motifs predicted from the ribosomal protein gene cluster. Using an *in vivo* promoter::mCherry transcriptional reporter assay, we test and confirm the function of 4 predicted motifs in 23 ribosomal protein promoters. In Chapter 6, we show that aspects of regulatory element organization in *cis*-regulatory modules (CRMs) can be conserved and can be used to detect CRMs from sequence. By this criteria, we show that two of our predicted motifs, M546 and M540, function as a module in ribosomal proteins and disruption of module organization affects promoter function *in vivo.*

Chapter 7 summarizes the results of this research, discusses the overall impact these results have on the field of genomic research and provides suggestions for extending this research in other directions.

Chapter 2: Background
Cis-regulatory elements in *C.elegans*

2.1 *C.elegans* as a model organism for studying transcriptional regulation

Since Sydney Brenner performed the first studies on *C.elegans* in the 1970s [9,10], *C.elegans* has become a popular model organism for genetics, developmental biology and neuroscience. Part of the appeal of working with *C.elegans* is its short generation time and ease of maintenance and propagation in the laboratory. The lifespan of *C.elegans* is about two weeks and worms proceed through four larval moults (stages L1, L2, L3 and L4) before maturing as adult. The time from hatching to maturity is about 2 to 3 days. *C.elegans* is also unique from other popular model metazoans because they exist in two sexes, male and hermaphrodite, and can either reproduce by selfing or by sexual reproduction. Hermaphroditic reproduction is advantageous in generic studies because it allows researchers to easily obtain self-crosses and segregate homozygous individuals. Another distinguishing feature of *C.elegans* biology is that worms have a deterministic developmental plan. Adult worms have a fixed number of cells (972) and the entire *C.elegans* cell lineage has been painstakingly mapped through every cell division from zygote to adult organism[11]. In fact, several microRNA genes, i.e. lin-14 were first indentified based on their classification as "heterochronic genes" – genes that disrupted the timing of cell division events and the stereotypic cell fate of the resulting daughter cells[12].

We are particularly interested in *C.elegans* as a model organism for studying transcriptional regulation because it has a transparent body structure which makes it particularly amenable to *in vivo* imaging experiments.

Transcriptional reporter assays are commonly used to test sequence elements for regulatory potential. In these experiments, sequence believed to have regulatory activity is fused to a fluorescent reporter (i.e. GFP) and reporter expression is used as a readout of sequence regulatory potential. Transgenic *C.elegans* are generated bearing this reporter construct either using microinjection or microparticle bombardment. In microinjection, construct DNA is diluted with *C.elegans* genomic DNA and injected using a syringe into the gonad of an adult hermaphrodite. Ideally, the construct DNA

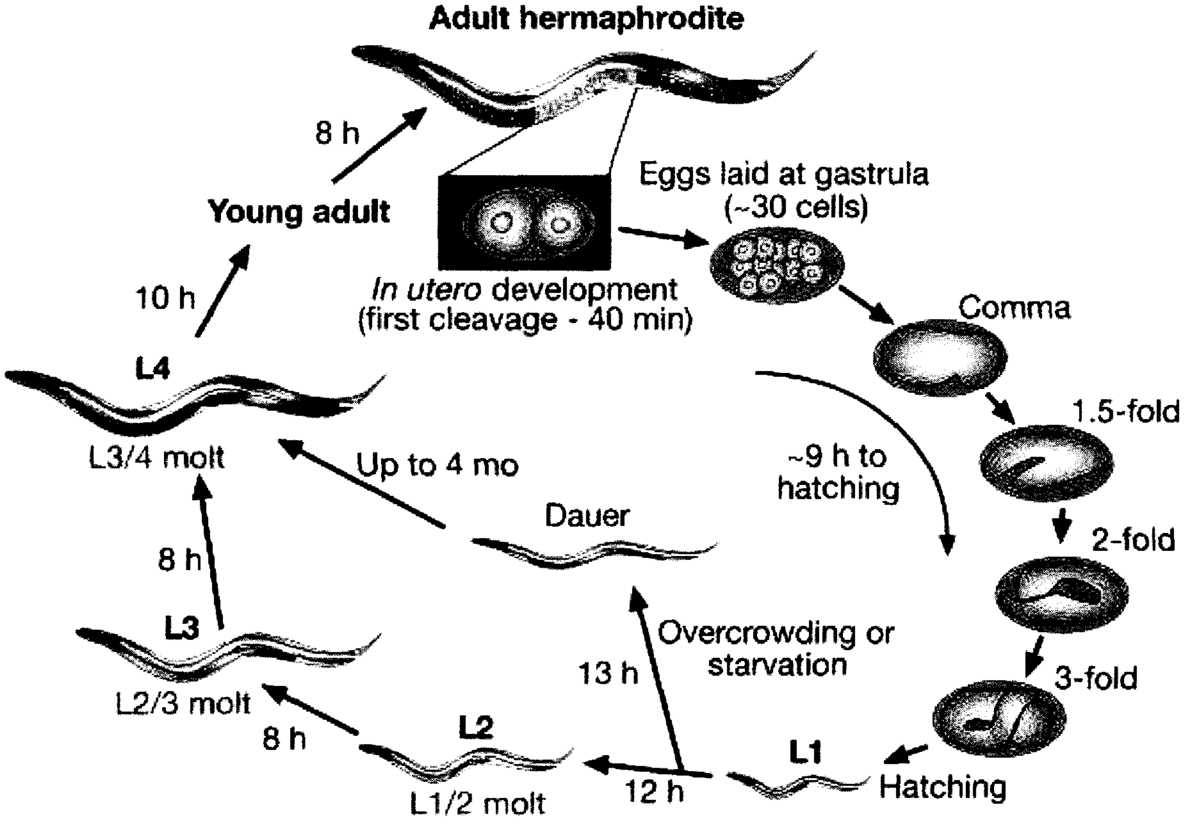


Figure 2.1: Lifecycle of the nematode *C.elegans*[13].

should integrate itself into the gamete genome and be inherited by the resulting offspring. Typically though, because the injected construct DNA is of high copy number, it is integrated with low frequency and is likely passed onto the offspring as an extrachromosomal array instead[14]. Extrachromosomal arrays are not inherited reliably and can be lost with successive generations. In addition, genes on extrachromosomal arrays are silenced in the *C.elegans* germline as part of a proposed host defence mechanism[15] – a consequence of which being that germline driven reporter expression can be only observed with integrated constructs.

Microparticle bombardment was developed as an alternative technique to microinjection for generating transgenic *C.elegans*[16]. Briefly, construct DNA is precipitated onto micron sized gold beads, which are accelerated to high speeds by the release of pressurized gas in the bombardment chamber and are propelled into the gonads of waiting worms. In this shotgun approach, DNA is released from the gold beads upon entering the gonad and either forms extrachromosomal arrays or integrates into the host genome. This occurs at very low probability, which is why high densities of worms are typically used in such an experiment and a selectable marker gene is used to identify successfully transformed worms. The main advantages microparticle bombardment claims to have over microinjection are its ease of use and its ability to generate single copy construct integrants with higher efficiency.

2.2 Overview of *cis*-regulatory elements

Cis-regulatory elements describe the general class of DNA sequence elements that affect the regulation of genes located on the same strand of DNA (action in *cis*). Examples of *cis*-regulatory elements include a) transcriptional factor binding sites (TFBS), b) chromatin insulators, c) nucleosome positioning signals and d) structural DNA elements. This section provides examples of mechanism of action for each category of *cis*-regulatory elements with an emphasis on regulation in *C.elegans.* We expect quite a bit of overlap between these categories, i.e. nucleosome positioning signals are likely functional due to their ability to affect DNA structure etc.

2.2.1 Transcription factor binding sites

Transcription factors regulate gene expression by recruiting chromatin remodeling complexes and other general transcription factors that interact with and recruit RNA PolII to the core promoter. In doing so, transcription factor binding is involved in specifying the temporal and spatial control of gene expression. Examples of this in *C.elegans* include studies conducted on the *myo-2* and *mec-3* promoters, which drive expression in the pharynx and touch receptor neurons respectively[17,18]. Two *cis*-regulatory elements, originally named B and C were identified to be necessary for *myo-2* promoter function, but in different tissue types. B is recognized by the homeodomain TF, *ceh-22* and drives gene expression in pharyngeal muscle. C is recognized by the forkhead TF *pha-4* and is responsible for gene expression in all other pharyngeal tissues. Similar to *myo-2* but underscoring more complex regulation, *mec-3* encodes a transcription factor that dimerizes with the POU domain TF UNC-86 to bind to its own promoter and regulate its own expression. This interaction is essential for *mec-3* expression in touch receptor neurons and can also be found in the promoters of other genes also expressed in these neurons.

Studies of transcription factor mediated regulation in *C.elegans* was encouraged by development of promoter::GFP transcriptional reporter assays. Two types of transcriptional reporters are commonly used: i) fusions with the native promoter where mutations can be made in putative TF binding sites to determine function, ii) minimal promoter-enhancer fusions, to determine if putative enhancers are sufficient for directing gene expression on their own. Given the relatively small size of the worm genome, efforts are currently ongoing to characterize the regulatory potential of all non-coding sequences in the "promoter-ome" by generating large-scale promoter::GFP transcriptional fusions using Gateway cloning technology and introducing these constructs in *C.elegans* using biolistic methods[19].

The interaction between the TF DNA binding domain and promoter sequence forms the basis for TF binding specificity. There are several families of TFs classified based on homology of their DNA binding domain. While members of the same TF class recognize a similar core motif, differences in protein structure not in the DNA binding domain contribute to significant diversity in binding affinity and fine tune the exact motif recognized[20]. Examples of TF families in *C.elegans* and some of their well-characterized members include: basic helix-loop-helix TFs that recognize the E-box sequence CANNTG, zinc finger TFs, homeodomain TFs, forkhead transcription factors *i.e. daf-16*, GATA transcription factors *i.e. elt-2*, lim domain transcription factors *i.e. mec-3* and POU domain transcription factors *i.e. unc-86.*

2.2.3 Chromatin insulators

Chromatin insulators establish bounds for regulatory activity by acting as enhancer blockers, ensuring that enhances do not influence the expression of neighbouring genes based purely on physical proximity[21]. While examples of chromatin insulators exist in *D.melanogaster*, perhaps the most famous example of chromatin insulators is the vertebrate insulator protein CTCF. CTCF derives its name from binding three direct repeats of the sequence, CCCTC, upstream of the chicken *c-myc* gene. The CTCF gene encodes an eleven zinc-finger DNA binding protein which uses different zinc fingers in binding different sites[22,23]. This may partly account for the multi-faceted regulatory role of CTCF, not only as a chromatin insulator, but also as an activator and repressor.

Two main mechanisms have been proposed to explain the action of CTCF as an insulator. Firstly, pairs of interacting CTCF proteins bound to DNA may create chromatin loops and hence, functional domains of gene expression. If genes and enhancers are located on separate loops, the increased physical distance may diminish the ability of the enhancer to exert regulatory influence on the genes. CTCF may also interact with nuclear matrix, using it as a proxy to form chromatin loops. As evidence of this interaction, Dunn et al have identified CTCF as a nuclear matrix associated region (MAR) binding protein[24]. Alternatively, CTCF may set up functional chromatin domains by acting as a barrier to the spread of heterochromatin. For example, histones adjacent to the CTCF chicken HS4 insulator are constitutively acetylated at H3 K9, which may prevent them from acquiring repressive marks i.e. methylation at H3 K9, and in doing so, halt the spread of heterochromatin from upstream loci[25].

Unlike vertebrates which primarily rely on CTCF for insulator function, several insulator elements and their corresponding binding proteins have been discovered in *D.melanogaster*. These include the *gyspy* insulator element originally identified from the *gypsy* retrotransposon, which binds proteins Su(Hw), Mod(mdg4)2.2, as well as CP190,0 and the *scs* and *scs'* sequences that bind ZW5 and BEAF-32 proteins respectively. Expectedly, the drosophila homolog of CTCF is also functional, binding to the *Fab-8* insulator element identified in the *drosophila Abdominal-B Abd-B* gene of the bithorax complex[26]. Whereas vertebrate CTCF has been found to interact with majority of identified insulators, drosophila insulator proteins do not show homology to each other and bind exclusive DNA sequences, indicating that there is considerably more insulator diversity in *D.melanogaster* than vertebrates. The reasons for such diversity can only be speculated at – one hypothesis is that vertebrates have lost all insulator regulation with the exception of CTCF. This may be possible, given that vertebrate genomes are larger and regulatory influence of enhancers may be delimited by distance rather than insulator based mechanisms.

2.2.3 Nucleosome positioning signals

Eukaryotic DNA is packaged into nucleosomes *in vivo*, which consist of 146 bp of DNA sequence wrapped around a histone octamer. DNA sequences display different affinities for nucleosome formation *in vitro* and a number of nucleosome reconstitution experiments have shown that motifs containing 3 base pairs of A and T, or G and C nucleotides separated by a 2 base pair linker, i.e. (A/T)3NN(G/C)3NN, are superior at nucleosome formation *in vitro*[27]. Repeated every 10 base pairs, this motif can be thought of as two submotifs, a (A/T)3NN trimer motif repeated with a 10 base pair periodicity that is 5 base pairs out of phase with a (G/C)3NN trimer motif. Interestingly, in follow-up studies, the (G/C)3NN motif was determined to be unfavorable for nucleosome formation on its own, suggesting that the other submotif, i.e. (A/T)3NN, may largely be responsible for the high affinity for nucleosome binding in the original motif[28]..

Affinity for nucleosome formation is relevant to transcriptional regulation *in vivo*, because nucleosome bound sequence may be protected from binding of other transcriptional regulators. It has also been established that lysine (and to a lesser extent, arginine) residues of histone tails can be modified through acetylation, methylation and phosphorylation, to modulate the affinity of nucleosomes for DNA sequence[29]. A model is emerging whereby transcription factors and other regulators compete with nucleosomes for binding of regulatory sequence[30]. As such, many recent papers have focused on using high-throughput methods, such as nucleosome IP followed by tiling array hybridization to query the *in vivo* occupancy of nucleosomes in the genome, and determine if particular sequence motifs display affinity for well-positioned nucleosomes, similar to the situation *in vitro*. Most of such studies have focused on yeast and occasionally metazoans, i.e. *C.elegans*[31,32,33,34].

Motivated by previous *in vitro* data, Segal *et al* asked if nucleosome occupied sequences are enriched for dinucleotide sequence motifs[31]. The authors isolated mononucleosomes from yeast and studied 199 associated sequences for motifs. Somewhat expectedly, the authors found the most significant motif to be similar to the *in vitro* result: 10 bp periodic AA/TT/AT dinucleotides that oscillate out of phase with GC dinucleotides with the same periodicity. Segal *et al* subsequently used the dinucleotide motif model to predict *in vivo* nucleosome organization – their algorithm used a dynamic programming approach to learn nucleosome positions with consideration of possible steric interactions between neighboring nucleosomes. Using five different approaches to validate their model predictions, i.e. comparing against experimental mapped nucleosome positions, performing *in vivo* and *in vitro* assays to verify model predictions, the authors claimed their model predictions were able to explain up to 50% of observed nucleosome positions *in vivo*.

In the first paper to use tiling arrays to experimentally interrogate *in vivo* nucleosome positions, Yuan et al used positions of isolated mononucleosomes to train a Hidden Markov Model (HMM) for predicting sequences as nucleosome-bound, fuzzy or nucleosome-free[32]. In a subsequent study, Lee *et al* used a higher resolution tiling array to improve nucleosome occupancy calls[33]. The data confirmed previous observations about nucleosome organization, i.e. intergenic regions are relatively nucleosome depleted compared to coding regions, nucleosome occupancy in promoter sequence is inversely correlated with gene expression etc. Lee et al grouped genes into four clusters based on the nucleosome organization pattern in their promoters and sought to identify sequence features capable of explaining the clustering patterns. Instead of a dinucleotide model, the authors used linear regression to weight sequence features based on their ability to explain the observed data. The sequence features with the highest weight values included periodic features, i.e. propeller twist, tip and tilt. Features that confer sequence rigidity, i.e. AAAA, also received greater weighting, presumably due to their role in nucleosome exclusion. Their algorithm was trained on chromosomes 1 to 8, and generated predictions more strongly correlated with observations on chromosomes 9 to 16 than the dinucleotide model of Segal *et al.* Interestingly, the authors note that the 199 sequences used for algorithm prediction in Segal *et al* have a nearly random distribution of occupancy ratios and do no appear to correspond to well-positioned nucleosomes. Lee *et al* propose that this discrepancy may be because genome wide nucleosome organization is due to 'exclusion signals', whereas strongly positioned nucleosome and other local 'translation and rotation' settings are directed by periodic dinucleotide motifs. *Cis*-regulatory elements directing nucleosome positioning may thus belong to either 'periodic, rotational' or 'rigid, exclusion' categories.

2.2.4 Structural DNA elements

Studies on *cis*-regulatory motif identification have focused on depicting sequence motifs as a linear sequence of alphabet letters, A, C, G and T. This can be a misleading simplification, especially since DNA is a chemical molecule with 3D structure and base pair changes can cause changes in the DNA backbone, as well as major and minor groove conformations. Reinforcing this view, transcription factors recognize not specific base pair combinations, but rather the solvent accessible surface the sequence element presents.

In a recent paper, Parker et al show that single base substitutions in sequence elements can have a wide range of effects on the DNA structural profile, such that sequence words differing in a single position can have wildly different structures whereas sequence words with few bases in common may have very similar structures[35]. The solvent accessible surface of DNA can be interrogated by judging the hydroxyl radical cleavage pattern of DNA. Using structural similarity as a metric, the authors scanned the genome for sequence elements that have conserved structural profiles in a comparative genomics approach. They called their algorithm Chai and compared its ability to predict regulatory elements against a previously developed algorithm based on conserved sequence identity (binCons). The authors demonstrate that structural conservation recovers more functional regulatory elements than sequence conservation, with up to 78% of DNase I hypersensitivity sites and 59% of predicted enhancer sites recovered with Chai. The authors did not report the number of predicted sites not overlapping DNase I hypersensitivity sites, which may provide insight into the role of conserved DNA structural domains not necessarily directly involved in active transcription.

As further evidence that functional structural profiles are under selective pressure, the authors studied mutations that affect the affinity of TF binding to its binding site. They looked at binding sites for a zinc finger TF, Zif268 and an archeal transcriptional regulatory Ss-LrpB, and observed that mutations that disrupt the motif structural profile generally decrease the affinity of TF binding. While high affinity motifs have similar structural profiles, low affinity motifs differ in structural profile substantially. Furthermore, in studying phenotype-associated SNPs, the authors also show that noncoding SNPs associated with a phenotype are significantly correlated with changes in structural profile, as opposed to neutral SNP variants, suggesting that changes in DNA structure through polymorphism can result in functional differences.

The authors demonstrate that 12% of the human genome is under selective pressure based on conserved structural features, compared to a previous 5% by sequence conservation. Assuming considerable overlap between these two estimates, this suggests a doubling in regulatory sequence not previously thought to be functional, based strictly on sequence conservation. These structural features may correspond to known classes of *cis*-regulatory sequences, i.e. transcription factor binding sites, nucleosome exclusion sequences etc, or more interestingly, hint at new classes of structural DNA elements with novel regulatory roles.

2.3 Examples of evolutionary changes in *cis*-regulatory elements

In a comparative genomics study, it is estimated that whereas 5% of the human genome is under selective pressure, only 30% of this sequence (1.5% of the genome total) is protein-coding[2]. This implies that the majority of conserved genome features are non-coding and under strong selective pressure because of their innate regulatory potential. It has been proposed that changes in *cis*-regulation and not innovation in *trans* acting factors, are the main driving force behind evolutionary variation. Several supporting reasons for this hypothesis include: i) since protein coding sequence accounts for 1.5% of the human genome, by pure chance, most mutations will occur in non-coding regions ii) *trans*-acting factors regulate multiple targets and mutations in the coding sequence for these factors are likely to have severe phenotypic consequences. In contrast, mutations in *cis*-regulatory elements have a milder, more localized effect since by definition, they only affect expression of genes located on the same strand.

A pair of studies by Duncan Odom and colleagues serve as strong supporting evidence for this hypothesis[36,6]. In the first study, Odom et al ChIPed liver-specific transcription factors that are conserved between human and mouse.

In comparing genome wide localization data, they observed that a substantial number of binding sites failed to align in a multiple alignment, even though the motif recognized by both human and mouse TFs was the same. While misalignment may account for some of these observations, the data suggest substantial species specific differences in TF binding locations exist and are likely a result of innovation in *cis*-regulation. An even stronger case for this argument was made in the second study, where Odom and colleagues studied TF binding events in a mouse model of trisomy 21. The mouse used carried a human copy of chromosome 21 in addition to its two existing orthologous copies, which provided a rare opportunity to study TF binding to a human chromosome in an environment where all the *trans*-acting factors are exclusively mouse. Interestingly, mouse TFs bound to the human chromosome in a pattern that almost perfectly reproduced binding of human TFs in human cells, confirming that most evolutionary differences in binding are driven by changes in *cis*-regulatory sequence.

In a commentary, Rockman and Stern explain this hypothesis by reasoning that natural selection favors the least pleiotropic route to phenotypic evolution[37]. They refer to work by Jeong et al, where the authors identified mutations at a *cis*-regulatory element of a pigmentation enzyme *tan*, responsible for causing pigmentation differences between *D.yakuba* and its sister species *D.santomea*[5]. Linkage studies had identified the *tan* locus as a significant contributor to the pigmentation difference, but since *tan* genes in both species encode identical proteins, the change in phenotype is likely due to changes in *tan* regulation instead.

Jeong and colleagues identified a *cis*-regulatory element 3-4 kb upstream of transcription start in *D.melanogaster* that drives *tan* expression in the abdominal region. As evidence of its relevance to the regulation of *tan* expression, the authors demonstrated that a transgene containing *tan* downstream of the *cis*-regulatory element can rescue the pigmentation defect in *D.santomea*. Even more convincingly, in a survey of natural isolates of *D.santomea*, the authors identified not one, but three mutations in the *cis*-regulatory element, indicating three nonfunctional regulatory alleles arose independently as a result of natural selection. Since *tan* is a pleiotropic enzyme that is expressed in many tissues, the work by Jeong et al is strong evidence for *cis*-regulatory evolution as a route to achieve specific changes in *tan* expression without affecting its roles in other tissues.

A number of similar studies highlight cases where phenotype changes caused by *cis*-regulatory changes confer a selective advantage and are selected for in evolution. For example, temporal changes in *lactase* expression are due to mutations in a *cis*-regulatory element that has arisen independently in multiple human subpopulations[38]. Another study demonstrated that expression of DARC, a chemokine receptor, is absent in red blood cells due to a *cis*-regulatory mutation that otherwise does not affect DARC expression in other cell types. The lack of DARC expression in red blood cells confers malaria resistance and a consequent selective advantage to human subpopulations in West Africa possessing this *cis*-regulatory mutation[39] .

In summary, *cis*-regulatory evolution can be thought of as an expedient route for nature to select for "fine-tuning" mutations that can effect specific changes without affecting overall protein function or structure. Since a majority of the genome under selection is non-coding, a significant proportion of phenotype changes in disease or evolution is likely due to *cis*-regulatory variation. In spite of this, most GWAS groups have focused on identifying disease-causing coding mutations, mainly because of the additional work and difficulty associated with functional studies needed to characterize the effect of non-coding mutations. The ENCODE project is focused on cataloguing all *cis*-regulatory elements in the human genome, to enable researchers to better characterize functional *cis*-regulatory mutations from false positives. With similar motivation, algorithms developed for *cis*-regulatory element discovery can help researchers parse the non-coding genome into functional vs. non-functional regions, with the objective of determining the effect of non-coding mutations on gene expression.

2.4 Defects in ribosomal protein regulation can cause phenotypic abnormalities.

A number of motif finding algorithms search for over-represented sequence motifs in the promoters of coexpressed genes. It is presumed that if a group of genes share a common gene expression profile, they are likely to be regulated by a common mechanism. In this regard, the ribosomal protein genes are ideal for studies of transcriptional coregulation since they are tightly coexpressed across a wide variety of tissue types, developmental stages and organisms.

Studies focused on the transcriptional regulation of ribosomal proteins have lagged, mainly because they have been viewed as uninteresting housekeeping genes that are expressed at high levels ubiquitously in all tissue types. Furthermore, ribosomal proteins are essential for organism viability and efforts to study loss-of-function phenotypes have been impeded by the fact that knockouts are almost universally embryonic lethal. This lack of attention is unfortunate, especially since deficiencies in ribosomal protein production can have clinically relevant phenotypic consequences. A classic example of this is Diamond-Blackfan syndrome (DBA), where 25% of unrelated patients have mutations in the *rps-19* gene[40]. DBA is characterized by defects in haematopoiesis, especially in the erythoid lineage. Other clinical manifestations of DBA include increased susceptibility to cancer, developmental abnormalities and decreased stature. To underscore difficulties with studying ribosomal protein defects, while mice homozygous for a *rps-19* null allele are not viable, heterozygous mice do not appear to have discernable phenotypic defects[41].

While traditional knockout approaches have not been successful in studying ribosomal protein regulation, work in zebrafish with morpholino knockdown has provided insight into effects of ribosomal protein misregulation[42,43]. In their study, Danilova and colleagues injected an antisense morpholino into zebrafish to inhibit translation of *rps-19*[42]. They demonstrated that using their assay, *rps-19* expression could be modulated in a dose-dependent manner that is controlled by the amount of morpholino injected. The authors were able to recapitulate phenotypic traits of DBA in zebrafish and further demonstrated that deficiencies in levels of RPS-19 caused dysregulation of p53 and associated factor deltaNp63, presumably due to increased nucleolar stress. Zebrafish lines targeted by morpholino knockdown express higher levels of p53 and increased apoptosis may be the cause for many developmental defects observed in DBA. Interestingly, the authors were able to rescue phenotypic defects in *rps-19* morpholino knockdowns by inhibiting p53, suggesting that p53 antagonists may be promising drugs for the treatment of DBA.

In an unrelated study, Uechi *et al* extended the work by Danilova *et al*l by targeting more ribosomal proteins in zebrafish by morpholino knockdown. While knockdown of other ribosomal proteins generated similar phenotypic consequences as *rps-19* knockdown, there was substantial variation in the severity and tissue specificity of the resulting phenotypes across all tested ribosomal proteins. Assuming the authors controlled for efficiency of morpholino targeting so that each ribosomal protein was knocked down by a similar extent, these results indicate that not all ribosomal proteins are required in the same cell types, implicating finer scale transcriptional regulation in conferring tissue specific expression.

Deficiencies in ribosomal protein expression can also cause phenotypic consequences in *D.melanogaster*. The resulting *minute* phenotype is characterized by prolonged development, thin bristles, poor fertility and viability. Loci associated with the *minute* phenotype were first identified in 1985 by Kongsuwan et al[44], and since then, over 50 *minute* loci have been identified in the *D.melanogaster* genome – 15 of these loci have been characterized molecularly and have been shown to overlap ribosomal protein genes. Marygold *et al* undertook a more extensive characterization of *minute* loci and showed that of 65 defined *minute* loci, 64 of them correspond or likely correspond to genes encoding cytoplasmic ribosomal proteins[45]. The one locus that did not overlap a ribosomal protein genes instead overlapped a translation initiation factor, underscoring the overall relationship between defects in translation and the *minute* phenotype.

Ribosomal protein mRNA levels affect the severity of the resulting *minute* phenotype in a dose-dependent manner. By mobilizing a *P*-element insertion in the 5' UTR of the *rps3* gene, Saeboe-Larssen *et al* were able to create two alleles of this gene with different sized excisions at the *P* element insertion site[46]. The weaker allele resulted in 15% decrease in *rps3* mRNA production and resulted in a moderate *minute* phenotype, while the stronger allele reduced mRNA production by 60% and resulted in a severe *minute* phenotype. This is evidence that the severity of the minute phenotype is related to differences in ribosomal protein levels, and not loss-of-function in some ribosomal protein genes. Null ribosomal protein alleles are embryonic lethal and deficiency in ribosomal protein expression is probably tolerated to a limited extent for certain genes, resulting in similar phenotypes in flies, zebrafish and human (*minute* and DBA respectively). Since changes in ribosomal protein expression are likely caused by mutations in transcriptional regulation, efforts focused on identifying *cis* sequence elements and *trans* acting factors involved in this regulation, would be of great value in understanding the biology of diseases caused by inefficient protein biosynthesis, as well as the development of potential therapies.

Chapter 3: Background
Overview of motif discovery methods

The aim of this section is to provide the reader with an overview of motif finding algorithms commonly used for sequence analysis. These algorithms can be classified based on the criteria used for determining motif significance (over-representation vs. evolutionary conservation), as well as their ability to detect individual *cis*-regulatory elements or groups of elements organized in *cis*-regulatory modules. This section is thus organized into two subsections: a) algorithms that rely exclusively on over-representation to find individual motifs and b) "phylogenetic" algorithms that use both over-representation and conservation to find individual motifs

3.1 Over-representation based approaches for motif discovery

Motif finding algorithms were originally developed to identify common sequence signatures shared by coding regions of genes with shared homology. These methods identify motifs based on their over-representation, i.e. how often they occur in the input sequence set compared to the frequency expected by chance. Motifs are typically assessed for significance by computing the likelihood they were generated by a null distribution of background nucleotide frequencies.

Since the concept of motif finding was first introduced, a number of motif discovery methods have been developed, which try to distinguish themselves from each other by emphasizing incremental improvements in methodology. For example, algorithms are divided by the approach used to represent motifs: a) a word-based approach, where the algorithm uses over-represented *k*-mer words to generate a consensus motif (i.e. Weeder[47]) and b) a frequency matrix approach, where matrix entries represent the probability of observing a particular base at a given position within the motif (i.e. AlignACE[48] and MEME[49]). In addition, algorithms differ based on the statistical learning method used: MEME uses expectation-maximization, whereas AlignACE, Gibbs Recursive Sampler[50] and MotifSampler[51] use Gibbs Sampling. To improve performance, algorithms such as MotifSampler, also use markov models to more accurately account for higher order dependencies in background nucleotide distributions.

In this subsection, we briefly introduce the methodology of the following algorithms: AlignACE, MEME, BioProspector[52], CONSENSUS[53] and Weeder. We also describe work by Tompa *et al*[54] which compared various over-representation algorithms on a common test set, to obtain a benchmark for algorithm performance.

AlignACE

AlignACE[48], along with the Gibbs Recursive Sampler[50] and MotifSampler[51], are Gibbs sampling based algorithms that use position specific weight matrices to represent motifs. AlignACE searches for the set of sequence elements that optimize the posterior probability of being members of a common motif. Because position weight matrices (PWMs) can be thought of as a product of independent multinomial distributions (one for each matrix column), a closed form expression for the posterior probability can be obtained by using the Dirchlet distribution as a prior. In AlignACE, the authors assumed a flat, uninformative Dirichlet prior and set the pseudocounts of the distribution to reflect background nucleotide frequencies. Motif over-representation is computed in terms of a likelihood ratio or MAP score, comparing the posterior probability of observing the motif model vs. the background null model given the data.

AlignACE follows a "sample-and-update" scheme in its motif search. At the beginning of each iteration, existing sites are used to build a PWM of the motif. Next, site locations are cleared and input sequences are rescored using the PWM. The score reflects the posterior probability that the observed site is a instance of the motif and similar to a Metropolis-Hastings decision rule, is used to weight site admission into the overall site configuration. While the sampling approach is stochastic, AlignACE is not strictly a Gibbs sampling algorithm since updates to the configuration are not step-wise incremental in the number of sites. More precisely, AlignACE can be thought of as a stochastic expectation-maximization classifier, where individual sites are classified as "motif" vs. "background" depending on their posterior probability scores.

MEME

MEME[49] is an expectation-maximization (EM) based algorithm developed by Bailey and Elkan and is based on an earlier algorithm called MM for fitting parameters in a two-component finite mixture model. (In a motif discovery problem, the two components are the motif and background classes respectively.)

MEME extends MM for the discovery of multiple motifs on the same sequence set by masking previously discovered motif sites with repeated application of MM.

MEME optimizes the log likelihood of observing the data *X*, the missing data *Z*, given model parameters θ (for the PWMs) and λ (for the mixing parameter).

$$logL(\theta, \lambda|X, Z) = \sum_{i=1}^{n} \sum_{j=1}^{2} Z_{ij} log(p(X_i|\theta_j)\lambda_j)$$

Following the two-component mixture model, sites are assigned group membership Z_{ij} to either the motif class or the background class. Group membership for each site is unknown and is treated by the EM algorithm as missing information to be estimated in the E step. This is estimated by computing the posterior probability of observing a particular class, given the site data and current estimates of the mixing parameters and PWMs for each class. The following update rule for Z_{ij} is used:

$$Z_{ij} = \frac{p(X_i|\theta_j)\lambda_j}{\sum_{k=1}^{2} p(X_i|\theta_k)\lambda_k}$$

In the M step, the algorithm uses group membership probabilities from the M step to estimate values for the mixing parameters and the PWMs. The following update rules are used for λ , the mixing parameter and θ, the PWM:

$$\lambda_j = \sum_{i=1}^{n} \frac{Z_{ij}}{n} \qquad f_{ij} = \frac{c_{ij} + \beta_j}{\sum_{k=1}^{L} c_{ik} + \beta}$$

, where f_{ij} is the ij^{th} element of the PWM, c_{ij} is the observed count for base *j* in that particular position and β represents a pseudocount.

MEME differs from AlignACE because it does not use a sampling based approach. Instead, every possible site is scored for membership into either motif or background classes in a deterministic approach. The motif search is thus highly dependent on the initialization seed and may be caught easily in local optima. In addition, MEME does not have the functionality to consider gapped motifs, unlike AlignACE. This may impact its ability to discover binding sites for transcription factors such as Gal4, which has three informative columns on each side of the motif separated by a gap of 10 columns.

BioProspector

Similar to AlignACE, BioProspector[52] is a Gibbs sampling based motif discovery algorithm that also uses PWMs to represent motifs. Instead of sampling sites weighted by their posterior probability, sites are scored according to a likelihood ratio: $A_x = Q_x / P_x$, where Q_x is the probability the site was generated by the motif model and P_x the probability the site was generated by a background model. BioProspector also uses a modified decision rule with high and low thresholds for site sampling: Sites scoring higher than a threshold value T_H are added with 100% probability, whereas sites scoring lower than a threshold T_L_are discarded. One site is chosen from each promoter from a pool of sites that score between threshold values, with probability proportional to $A_x - T_L$. BioProspector appears to use some form of simulated annealing as well – T_L is set to 0 at the beginning of each search and is linearly increased until it is $T_H / 8$ at the end of the search. As such, the structure of the algorithm does not adhere strictly to the Gibbs sampling method and appears to be rather ad hoc in its design. No rationale was provided in the original paper for the choice of method.

A unique feature of BioProspector is that it is capable of modeling motifs with an internal gap. It does so by allowing motifs to contain two binding blocks, each with its own PWM, separated by a fixed gap length. Palindromic motifs reduce the requirement of two PWMs to one. This feature of BioProspector allows it to look for motifs corresponding to TFs that dimerize in their binding. Significance of motif scores are determined using nonparametric Monte Carlo statistics. BioProspector defines a motif score as the product of the number of observed motif sites and the weighted information content of the motif:

$$\text{Motif Score} = N \times exp\left(\sum_{positions} \sum_{nucleotides} q_{ij} log(\frac{q_{ij}}{p_j})\right)$$

Where N is the number of aligned segments in the motif, q_{ij} is the probability of observing nucleotide *j* at position *i* in the PWM and p_j is the probability of observing nucleotide *j* from the background distribution.

M independent and identically distributed sequence sets are generated under an input sequence probability model and motifs discovered on these random control sets are used to derive a null distribution of motif scores, from which a *p*-value can be assigned to determine motif significance.

CONSENSUS

CONSENSUS[53] is an algorithm developed by Hertz and Stormo, which differs from algorithms in this review in that it uses an information content metric as opposed to likelihood ratios or posterior probability scores, to assess motif over-representation.

$$I_{seq} = \sum_{j=1}^{L} \sum_{i=1}^{A} f_{ij} \ln \frac{f_{ij}}{p_i}$$

where f_{ij} is the frequency of observing base i at position j in the PWM, p_i is the background probability of observing base i.

The authors describe their information metric as a normalized log-likelihood ratio and relate it to the Kullback-Leibler information or relative entropy. To determine if the information value of a particular motif is significant, the authors provide two methods to calculate a p-value for their statistic: (i) method of large deviation statistics and (ii) a numeric method. The authors also provide a method for estimating the expected number of possible sites for a given motif. The product of the p-value and expected count measures yields the expected frequency of an information content score and is used to determine if it is statistically significant.

In CONSENSUS, motif sites are treated as sequences in an alignment and the aim of the algorithm is to maximize the alignment information content. To do so, the authors used a greedy framework, where the highest scoring site that maximizes motif information is always added to the alignment. Whereas the original implementation of CONSENSUS required the user to specify the desired motif width, the authors also presented an alternative version called WCONSENSUS where the motif width was allowed to vary, depending on a user-specified standard deviation bias.

The use of a greedy framework in CONSENSUS poses several problems. In general, greedy algorithms are sensitive to the initialization seed and are often stuck in local optima. More importantly, CONSENSUS uses repeated pairwise sequence comparisons to determine if a site should be included in an alignment. This exhaustive comparison approach may work for short promoters in bacterial genomes, but can become computationally intractable for longer promoters in larger genomes.

Weeder

Other methods reviewed in this subsection use position weight matrices to represent motifs, partly because the alternative representation which uses *k*-mer words, poses challenges to algorithm implementation. To find over-represented *k*-mers in the input sequence set, one would first have to consider 4^k possible word possibilities and in the most simple example, count all occurrences in the input sequence set for each *k*-mer and ask if the number of occurrences is statistically significant. This becomes very difficult for large values of *k*.

Pavesi *et al* developed a method of finding over-represented *k*-mers that is computationally manageable for small values of *k* (typical of the length of TF binding sites), allowing a certain level of mutations or mismatches in *k*-mers found.

Weeder relies on a suffix tree approach where searches for valid *k*-mers can be interpreted as tree-traversal problems. Similar to a greedy search, the tree-traversal "weeds" out paths that do not satisfy the maximum number of mutations requirement set by the user. While this prevents the number of possible paths for traversal from growing exponentially, the authors admit a possible downside of this approach is that it only considers a subset of all possible occurrences. Furthermore, implicit in this approach is the assumption that all mutations are equivalent in that they are sequence mismatches. To an extent, this reflects a drawback of the *k*-mer approach, since it cannot model position specific differences in information content unlike PWM approaches.

Benchmark for algorithm performance

Even though we have introduced four algorithms in this subsection, many more motif finding algorithms are available for download, each emphasizing its strengths over the others. To provide users of these algorithms with some clarity on which ones perform better, Tompa et al put together test sequence sets where TF binding sites were seeded into promoter sequence, and compared various motif finding algorithms on their ability to retrieve the seeded sites. Tompa *et al*[54] used thirteen motif finding algorithms in his survey, including AlignACE, ANN-Spec, CONSENSUS, GLAM, Improbizer, MEME, MITRA, MotifSampler, oligo/dyad-analysis, QuickScore, SeSiMCMC, Weeder and YMF.

To ensure the test sets reflect the biological situation *in vivo*, the authors used curated TF binding sites from TRANSFAC and considered three types of background sequence: i) the actual promoter sequence, ii) randomly chosen promoter sequence from the same genome and iii) artificial sequence generated by a 3rd order markov model. The test sequence sets were then made available to the authors of each algorithm for motif discovery, so no arguments can be made that an algorithm performed poorly because it was used incorrectly. At the same time, the authors could exercise their own discretion in pre- or post-processing algorithm predictions, spend as little or as much time as they desired at fine tuning their results, which can introduce considerable bias that should be taken into account when viewing the results of the cross-algorithm comparison.

Tompa *et al* made several observations based on the results of the comparison: Most strikingly, they noticed many algorithms performed better at recovering yeast motifs than any other organism. In an effort to be more comprehensive, Tompa *et al* used a variety of metrics, i.e. sensitivity, positive predictive value, nucleotide level correlation, to assess performance. Ranking of algorithm performance did not change much when different metrics were considered. Weeder clearly had the best performance compared to all other algorithms. In their discussion, Tompa *et al* attributed Weeder's success to its "judiciousness". Weeder was run in a "cautious" mode where only the strongest motifs were reported. Tompa's performance metrics were also affected by algorithm calls of "no motif found" – Weeder was very careful in declaring when no motif was discovered in a dataset and may have had an advantage over other algorithms in this respect.

3.2 Phylogenetic algorithms search for over-represented conserved motifs

A core paradigm of comparative genomics is that functional sequences are under strong selective pressure and are likely to be conserved across related species. Following the same logic, functional *cis*-regulatory elements should be conserved in promoters of orthologous genes and this conservation can be used as an orthogonal source of information for motif detection. Several algorithms have been developed to include conservation in their motif search, with a majority of algorithms favoring an alignment based approach for assessing conservation, i.e. PhyloGibbs[55], PhyME[56], CONVERGE[57]. The aim of this subsection is to provide an introduction to a number of these algorithms, as well as comment on their efficacy.

CSC scoring

Instead of developing an algorithm that incorporates conservation into the motif search, Li *et al* used conservation to filter motifs reported from an over-representation based search[58]. In their CSC framework, MEME[49] or a similar over-representation based algorithm is first used to identify over-represented motifs in the data. The significance threshold for MEME is deliberately set low to increase sensitivity at the expense of false positives. MEME is run on the input sequences from the anchor species, as well as orthologous sequences from related species.

The motifs reported by MEME are called marginally significant motifs (MSMs) and are later merged with other MSMs reported from the orthologs. MSMs are merged in pairwise fashion, so MSMs reported on each species are merged in pairs to form motifs common to two species, paired motifs are merged with a third motif to form motifs common to three species and so on. The authors develop their own test statistic to determine motif conservation, which is a simple count of the number of groups of orthologous genes containing instances of the ancestral motif. The authors also present a method for calculating test statistic significance, which yields a *p* value representing the chance of observing at least as many orthologous gene groups containing conserved motif sites.

Since the CSC framework does not integrate conservation into the motif search, it may miss motifs that are weakly over-represented in each species but still significantly conserved when compared across species. Furthermore, the authors weighted conservation of motifs in alignments based on a phylogenetic tree inferred from well-aligned sequences in orthologous promoters. The authors claim that the inferred rates of divergence model neutral evolution, but it is unclear if the well-aligned regions used for inference are indeed neutral to promoter function.

PhyloGibbs

PhyloGibbs[55] is a Gibbs sampling based motif finding algorithm that uses sequence alignment to assess motif conservation. More specifically, PhyloGibbs requires input sequences to be aligned by diAlign[59] and parses the alignment output into blocks of alignable vs. unalignable sequences. Motifs are represented as windows positioned over these blocks and while motif instances are permitted to occur in unaligned blocks, instances occurring in aligned blocks are given a higher score and added to the model with higher probability.

Central to the PhyloGibbs algorithm is its evolutionary model, which assumes that mutations occurring in transcription factor binding sites are fixed with probability proportional to its frequency in the PWM. Similarly, mutations in nonfunctional regions are fixed according to background frequencies.

$$T(\alpha|\beta, w_i, q) = q\ \delta_{\alpha,\beta} + (1-q)w_{\alpha i}$$

The above equation represents the probability that at position *i* of the PWM, base α will be mutated to base β, given that the weight matrix component for position *i* is w_i and $q = e^{-\gamma t}$ is the probability no mutation occurred in time *t*. The evolutionary model is incorporated into a likelihood score, which represents the probability that sequences from multiple species within the same motif "window" are related descendants of a common ancestral motif site:

$$P(W_i|w_i) = \sum_{\alpha} w_{\alpha i} \prod_{i \in W_i} T(s_j|\alpha, w_i, q_j)$$

, where the term in *T( )* represents the contribution from the evolutionary model. The authors state that expanding this term will result in a polynomial of N+4 terms.

To yield a closed form expression for the integral, the authors use the following mononomial approximation and fit parameters c and λ, demanding that the first and second moments of the approximation match the original function.

$$P(W_i|w_i) \sim c \prod_{\alpha} w_{\alpha}^{\lambda_{\alpha}}$$

This approximation is attractive, since it is amenable to the multinomial distribution and a closed form expression for the integrals can be written in terms of the gamma function representation, as below (c and λ as in the approximation, γ is a pseudocount parameter.)

$$\int \prod_i P(W_i|w_i)P(w_i)dw \sim \frac{\Gamma(4\gamma)}{\Gamma(4\gamma + \sum_{i,\alpha} \lambda_{\alpha i})} \prod_{\alpha} \frac{\Gamma(\gamma + \sum_i \lambda_{\alpha i})}{\Gamma(\gamma)} \prod_i c_i$$

Simply put, the posterior probability score for motifs occurring in well-aligned windows is upweighted compared to motifs in unaligned regions, depending on its sequence conservation.

PhyloGibbs also employs a simulated annealing strategy called "track and anneal" to identify optimal motif configurations. From personal experience, it appears that scoring motifs differently based on their placement in alignable regions gives PhyloGibbs an advantage in terms of rejecting false positives. PhyloGibbs typically reports few motifs from data, but the few reported motifs are often true positives. The only downside with PhyloGibbs is that it relies on alignment to assess conservation. In doing so, it assumes that *cis*-regulatory elements are associated with regions of alignment similarity. This is not necessarily true, especially since transcription factor binding sites are often short, degenerate and of varied composition. In fact, while PhyloGibbs does report fewer false positives, it appears to do so at the expense of reduced sensitivity. PhyloGibbs may have missed some true positives because these motifs failed to overlap well-aligned regions.

PhyME

PhyME[56] is similar to PhyloGibbs[55] in that it also searches for motifs in prealigned sequence and uses a similar evolutionary model where mutations in binding sites are fixed according to position weight matrix probabilities. The difference is that whereas PhyloGibbs is a Gibbs sampling based stochastic optimization algorithm, PhyME relies on expectation-maximization, similar to MEME, to converge to motif optima. Sinha *et al* also claim that PhyloGibbs is adapted to finding motifs for species related by a star shaped phylogeny, whereas PhyME can be more generally applied to any phylogenetic tree.

PhyME requires pre-aligned sequence, usually from M-LAGAN, as input. It extracts blocks of ungapped alignment that meet a minimum size and percent identity threshold and labels these regions as "well-aligned". The algorithm is built around a HMM framework, where each promoter alignment can be thought of as being generated from a motif sequence model, θ or background sequence model, θ_b. Input sequence can be parsed base-by-base, where either motif or background models are chosen given transition probabilities $p_m=p$ or $p_b=1-p$. PhyME optimizes an objective function, which is the log-likelihood ratio of generating the sequence by a motif model vs. background.

$$F(S,\Theta) = log\left(\frac{Pr(S|\theta)}{Pr(S|\theta_b)}\right)$$

Computing *F(S, θ)* relies on the sequence probability, *Pr(s | W)*, which is the probability of generating a subsequence sampled from the PWM θ or background model θ_b. Depending on the location of the sequence in an aligned vs. non-aligned region, *s* can be written as an alignment of sequence from multiple species. PhyME scores motif instances in aligned regions differently by incorporating a term for the evolutionary model, where ψ represents the phylogeny of the compared species, μ_σ represents the neutral rate of mutation between the ancestor and the species σ, *W* is the PWM matrix and s_σ is the nucleotide from species σ in the alignment ψ.

$$Pr(\psi|W,k) = \sum_{\alpha\in\Sigma} W_{k\alpha} \prod_{s_\sigma\in\psi} (\mu_\sigma W_{ks_\sigma} + (1-\mu_\sigma\delta_{\alpha s_\sigma}))$$

PhyME uses a Baum-Welch algorithm in expectation-maximization framework to train parameters θ_m and *p* of the HMM. Two parameters are computed in the E-step: A_i, the expected number of instances of motif *i*, and $E_{k\psi}$, the expected number of times the alignment ψ is sampled at position *k*. These parameters are used in the M-step to update estimates of *p* and θ_m: for example, $p = A_m / (A_m + A_b)$. PhyME uses the Newton method to solve for parameters in the M-step.

PhyloCON

PhyloCON[60] is an extension of the greedy algorithm of CONSENSUS to discover motifs in conserved promoter regions. Similar to PhyME and PhyloGibbs, PhyloCON relies on a pre-alignment to determine which regions in the promoter are conserved, but no mention was made in the paper stating a preference towards a particular alignment program. PhyloCON extends the pairwise merging framework of CONSENSUS: First, multiple motif profiles are generated for conserved regions. Next, profiles from different orthologous groups are compared and common profiles are merged based on their similarity. PhyloCON uses a statistic called the Average Log Likelihood Ratio (ALLR) to assess profile similarity and attempts to identify high scoring pairs (HSPs) for merging.

The ALLR was developed as follows: The log likelihood ratio, LLR measures the relative likelihood the observed site was generated by the motif model vs. background.

$$LLR = \sum_{b=A..T} n_b \ln \frac{f_{bi}}{p_b}$$

, where f_{bi} represents the frequency of observing base b at position i and p_b represents the corresponding probability using the background model.

In comparing two profiles, if the profiles were generated by a common motif, then scoring by base pair frequencies should be interchangeable between profiles. That is, the distribution from profile 1 should score profile 2 highly and vice versa. The ALLR is a weighted average of the "interchanged" LLRs:

$$ALLR = \frac{\sum_{b=A..T} n_{bj} \ln \frac{f_{bi}}{p_b} + \sum_{b=A..T} n_{bi} \ln \frac{f_{bj}}{p_b}}{\sum_{b \in A..T}(n_{bi} + n_{bj})}$$

High scoring pairs are considered members of a common motif. Like CONSENSUS, a greedy algorithm is used to merge high scoring pairs with other profiles, so profiles are grown incrementally in a pairwise fashion. PhyloCON does not distinguish between closely vs. distantly related species in its comparison, all species are assumed to be equally diverged and related via a star shaped topology. Strictly speaking, PhyloCON does not simultaneously search for over-represented and conserved motifs – like CSC scoring in reverse, it searches for over-represented motifs in regions of high sequence alignability. Unlike CONSENSUS, no mention was made in the paper regarding methods for determining the significance of ALLR scores.

Ensemble centroid sampling

The method developed by Newberg *et al*[61] presents two improvements over phylogenetic Gibbs sampling methods such as PhyloGibbs: the use of a felsenstein algorithm to model binding site divergence and the use of ensemble centroid estimation. While the method continues to rely on multiple alignment, the authors use a Felsenstein tree-likelihood algorithm[62] to compute the joint probability of an aligned set of nucleotides, for each column of the motif. The Felsenstein algorithm recursively traverses a user-supplied phylogenetic tree to obtain a joint probability of nucleotides on its leaves and is an alternative to the evolutionary model of

PhyloGibbs and PhyME, where mismatches in an aligned column of a motif are assumed to be mutations fixed by probabilities in a PWM. Similar to AlignACE, the algorithm optimizes for a posterior probability of the overall motif configuration, i.e. the MAP score, which uses a product multinomial distribution to parameterize the motif model and assumes a dirichlet distribution as a prior. The difference is that incorporating likelihoods from the Felsenstein tree likelihood algorithm causes the MAP score to be rewritten as follows:

$$P(R|A) \sim \int P(R, \Theta, \Theta_0|A) P(\Theta) P(\Theta_0) d\Theta d\Theta_0$$

$$P(R, \Theta, \Theta_0|A) \sim \left(\prod_{j \in background} Fels(R_{1,\ldots,N,j}|\Theta_0) \prod_b \Theta_{0b}^{\alpha_b - 1} \right) \prod_{l=1}^{w} \left(\prod_{j \in \mathrm{motif}\, l} Fels(R_{1,\ldots,N,j}|\Theta_l) \prod_b \Theta_{lb}^{\beta_{lb} - 1} \right)$$

Where R is the sequence data to be scored and A is an indicator variable, $a_j = 1$ if a motif starts at position j and is 0 otherwise. Θ, Θ_0 are the motif and background models respectively and α, β are background and motif pseudocounts.

Since this expression cannot be integrated analytically, the authors use importance sampling to obtain a solution. The downside of this approach is that it is often computationally intensive and to overcome this, the authors propose using an ensemble-based "centroid" method to estimate a solution. Instead of proceeding with as many iterations of importance sampling until motif convergence, the algorithm samples through an initial burn-in period of 2000-3000 iterations and tracks alignments for the next 8000-10000 iterations. The tracked alignments are considered part of an "ensemble" and the algorithm identifies as a solution, the "centroid" alignment which possesses the minimum distance to all other alignments in the set[63]. Distance between alignments is measured by taking the overlap between all pairs of sites in each alignment.

To an extent, by reporting the centroid solution, Newberg *et al* use binding site overlap to assess motif convergence. It should be noted that the centroid solution may not have the highest MAP score, but is guaranteed to contain the most number of sites in common with all other alignments in the ensemble. Of the methods reviewed, the method by Newberg *et al* appears to be most similar to PhyloGibbs, since PhyloGibbs also uses an "anneal-and-track" strategy as opposed to the ensemble tracking method of Newberg *et al.* Since PhyloGibbs is more adapted to solving problems given a star shaped phylogeny, the method of Newberg *et al* may be limited in analyzing species related by complicated phylogenies.

Chapter 4:
Development of an alignment independent algorithm for the prediction of *cis*-regulatory elements using evolutionary conservation

4.1 ABSTRACT

The study of comparative genomics has been motivated by the use of sequence conservation across species as a means to predict *cis*-regulatory elements, such as TF binding sites (TFBS). While highly conserved elements likely represent functional elements, recent studies have directly assayed TF binding in multiple species by ChIP and identified a large number of functional binding sites that fail to align. Motivated by these results, we have developed Flipper, a Gibbs-sampling based algorithm that searches for conserved, over-represented sequence elements without relying on multiple alignment. Flipper searches for conserved motifs by sampling sites in groups and requiring them to be distributed across orthologous promoters. We benchmark the performance of Flipper against traditional as well as conservation-based motif finding methods, using synthetic sequence data and *in vivo* binding data. On synthetic sequence test sets, we observe that Flipper performs up to 23% better than competing methods, especially when the compared species are sufficiently diverged so that alignability of TFBS is affected. Flipper is also capable of recovering known TF binding motifs from ChIP-chip and microarray expression datasets for yeast, worm and fly TFs.

Using Flipper to generate motif predictions on sets of coexpressed genes in *C.elegans,* we identified ELT-2 as a key regulator of genes encoding peptidases and other lysozymal components, as well as four novel motifs, M546, M540, M313 and M439, associated with ribosomal protein promoters.

4.2 INTRODUCTION

The identification and functional characterization of the regulatory elements which control the expression of genes remains a significant challenge. The mRNA expression level of a gene is typically determined by combinations of short (6-15bp) *cis*-regulatory sequence elements which are specifically recognized by transcription factors (TFs) and mediate interactions between these TFs and the RNA Polymerase II complex. The medical significance of gene regulation is underscored by the observation that the most common mutations leading to increased susceptibility to human cancers are chromosomal translocations which result in the apposition of a gene to the regulatory regions of another gene[64]. Despite its broad importance, until recently we have had limited ability to detect these non-coding regulatory elements or predict their function.

The availability of whole genome DNA sequence from an ever growing number of organisms has motivated renewed interest in developing novel computational and experimental approaches for determining and characterizing transcriptional regulatory elements on a genomic scale. To test the hypothesis that predictive regulatory elements could be identified from genome-wide expression data and DNA sequence, we developed a probabilistic Bayesian computational framework for systematically inferring combinatorial transcriptional regulatory logic[65]. While this method performed very well in *S. cerevisiae*, identifying regulatory mechanisms in multicellular organisms, and especially mammalian systems, will require significant algorithmic improvements and extensive experimental validation.

Two complementary observations allow the possibility of identifying regulatory elements from genomic sequence. (1) Over-representation: Most biological processes require the simultaneous participation of many gene products, at appropriate concentrations. Thus, transcriptional regulatory networks have evolved to direct the coexpression of genes at the mRNA level, and whole genome expression measurements (e.g. microarray data) consistently identify large sets of coexpressed genes in response to external, developmental, or genetic perturbations. Perhaps the easiest way to achieve coexpression is for genes to contain similar transcription factor (TF) binding sites or *cis*-regulatory elements, and many approaches attempt to use over-representation to identify these sequence elements in coexpressed genes. If a subset of genomic loci (or genes) are found to be bound by a common factor (or coexpressed), then specific DNA sequences which are responsible for this binding (or coexpression) will be enriched in these loci, and can be detected by pattern recognition algorithms, such as Gibbs sampling or expectation maximization . (2)Evolutionary constraint: If the proper expression of a gene contributes to fitness, mutations in regulatory elements should be under negative selection, or phylogenetically conserved. Thus similar DNA elements should be present in the regulatory regions of orthologous genes in closely related species

Several recently developed algorithms make use of both over-representation and evolutionary conservation to search for motifs. PhyloGibbs[55] is a Gibbs sampling based algorithm that searches for conserved motifs by sampling site configurations consistent with a multiple alignment of the input sequences. Sampled sites are scored using a weight matrix, with the added consideration that orthologous sequences in the same alignment window are related through an evolutionary model that assumes that mutations in binding sites are fixed based on their weight matrix frequencies. Another algorithm, PhyME[56] uses an Expectation-Maximization approach to train a Hidden Markov Model for parsing the input sequences into motif sites vs. background. Similar to PhyloGibbs, PhyME requires a pre-alignment step to determine regions of high sequence similarity, which it uses to evaluate motif site conservation. Other examples of phylogenetic motif finders include PhyloCON[60] and EMnEM[66]. All these algorithms share a common paradigm: orthologous sequences are separately aligned using a global alignment program (e.g. ClustalW[67] , DIALIGN[59] , M-LAGAN[68] , or MULTIZ[69]) before a pattern recognition algorithm is employed to search for overrepresented sequence features on the results of the alignment. A shortcoming of this approach is that *cis*-regulatory elements are typically rare, short, varied in composition and can be easily missed in a global alignment of large scale sequence homology. Also, the results of the motif search are limited by the method used to generate the initial alignment, the parameters of which must be optimized depending on the phylogeny of the sequences to be aligned.

The importance of an approach that does not rely on alignment is underscored by results from the ENCODE pilot project[8], which focused on regulatory factor binding studies over 1% of the human genome, including 18 sequence specific transcription factors, components of the general transcription machinery, histone composition and modification, and DNaseI sensitivity. In addition, the ENCODE consortium generated sequence from 28 vertebrates over these human regions, and used four sets of multisequence alignment programs to identify orthologous sequences in every other species. This analysis found that 5% of the ENCODE regions were 'evolutionarily constrained' with high confidence, the median length of which was 19bp. Yet, surprisingly, only 50% of the experimentally determined regulatory factor binding regions were 'constrained' by this measure. This might be due to the fact that regulatory regions are actually evolving relatively rapidly, but it seems more likely that multiple alignment cannot exhaustively detect conserved binding sites. In further support of this position, sequences which experimentally drive expression are often not detectable by standard alignment techniques . Other recent work has highlighted alignment uncertainty as a potential source of error in a wider class of genomic analyses . Since regulatory changes are postulated to be a strong driving force for evolutionary changes across phylogeny, it is possible that some of this lack of evolutionary constraint across functional regions may also be evidence of positive selection.

In this study, we have developed an alignment independent, Gibbs-sampling based algorithm, Flipper, which uses both over-representation *and* evolutionary conservation equally to detect sequence elements (putative TF binding sites) in orthologous regulatory regions. Position specific weight matrices (PWMs) are used to model motif composition. PWMs are significantly more general and realistic descriptions of transcription factor binding sites than *k*-mers, which are used by some alternative approaches[70-73]. We focus on generating predictions on four sequenced and closely related yeast, nematode, and fly genomes to minimize true regulatory divergence and present novel motifs associated with ribosomal protein promoters in *C.elegans*.

4.3 METHODS

4.3.1 Computing the sequence alignability of ChIPed TFBS

We assumed promoters identified in ChIP and expression assays were bona fide TF targets[74-79]. We used datasets from three species: yeast, worm and fly. In all cases, the binding specificity of the TF had been predetermined using independent experimental methods and was converted into a Position Weight Matrix (PWM) representation. We scanned every possible site in a target promoter using a log-likelihood ratio, comparing the probability the site was generated via the PWM model vs. background. We assumed the target site to be the highest scoring site within the target promoter. We used diAlign[59] to perform multiple alignments of target promoters and their respective orthologous regions in related species. For *S.cerevisiae,* we used *S.paradoxus, S.mikatae* and *S.bayanus* as related species. For *C.elegans,* we used *C.briggsae, C.remanei* and *C.brenneri* as related species. Similarly, for *D.melanogaster,* we used *D.yakuba*, *D.ananassae* and *D.pseduoobscura*. We scored the alignability of the target site using a Sum-of-Pairs (SP) statistic. To calculate the SP score of a site, for each position in the site, we enumerated all possible species pairings and counted the number of pairings that yielded exact sequence matches. Pairings with gaps were treated as mismatches. The overall SP score for a site was computed as the average over all positions in the site.

4.3.2 Algorithm Structure

Flipper finds multiple motifs on the input sequence set by repeatedly initiating new searches from randomly selected Position Weight Matrices (PWMs). After iterations of sampling, the motif search converges to a local optima when the overall configuration MAP score does not increase. The resulting motif is compared against other motifs archived from previous search attempts for redundancy and is added as a new motif to the archive if it is sufficiently dissimilar (pearson correlation < 0.8). We used code from the AlignACE program as a template for implementing our algorithm. We preserved some of the data structures and modules (i.e. column sampling, motif archival and similarity comparison), but reimplemented the sampling and search initialization steps to include considerations for multiple species. Flipper was implemented using C++.

Flipper searches for multiple non-overlapping motifs by reinitiating independent motif searches from random site configurations. Each motif search can be generally separated into the following steps: (i) search initiation, (ii) group sampling , (iii) MAP score and motif update and (iv) motif archival.

Search initialization

Every motif search is initiated from a random starting PWM and site configuration. Since the space of possible PWMs is large, using uninformed initialization points, i.e. arbitrarily generated random PWMs, would require considerable numbers of restarts before the PWM space is adequately sampled. We overcome this problem by using picking a site at random from either the anchor sequences or their orthologs, to seed the initial PWM. We scan all anchor species sequences and their orthologs using this initial PWM and group the highest scoring sites into orthologous groups, which will form the initial site configuration.

Theoretical basis for motif convergence

The theoretical foundations of a Gibbs sampling approach to motif discovery were clearly laid out in Neuwald *et al*, 1995[80] and we use their notation here. The parameters of the model are 1) an unobserved indicator vector $\xi = (\xi_1, \xi_2, .., \xi_L)$, where ξ_i is an indicator variable that equals 1 if a motif is present at position *i* and zero otherwise and 2) the current frequencies of bases in the PWM model Θ. In Neuwald *et al* and previous approaches[80,81], a sampling strategy was developed to identify the most probable models of the sequence as either being generated by a background model Θ_0 or motif model Θ depending on the current state of the indicator vector ξ. The predictive update version of the Gibbs sampler presented in Neuwald *et al* samples ξ iteratively based on the probability of a site given the current PWM model. The convergence properties of this Gibbs sampling approach have been classically established and rely on the aperiodicity and irreducibility of the underlying Markov chain, which require that all possible states of ξ can be reachable with finite probability from any initial state[82-84]. We designed our implementation of the data augmentation procedure from Neuwald *et al*, to ensure that these conditions are met, by allowing each possible sampling of ξ to accessible with finite probability in a fixed number of steps from any other state. When these criteria are satisfied, the Markov Chain of sampled states (ξ_{ij}, Θ) converges to the stationary distribution $P(\xi_{ij}|S)$ as the number of iterations approach infinity. The predictive distribution for the parameters given the sequence set *S*, can be integrated over the motif parameter space (Θ_0, Θ) and is

$$P(\xi|S) = \int P(S|\xi)P(\Theta_0)P(\Theta)P(\xi)d\Theta_0 d\Theta d\xi$$

Group sampling

Flipper differs from other Gibbs-sampling based motif finding algorithms by requiring sites to be added to the overall configuration in orthologous groups. Here, $\xi = (G_1, G_2, .., G_n)$, that is, G_i is a group of sites distributed over promoter *i* and its orthologs and $G_i = (\xi_{i1}, \xi_{i2}, .., \xi_{im})$ respectively, where ξ_{ik} is an indicator variable for a site in the kth ortholog of the ith promoter. Since the space of possible site combinations is large, motif convergence by undirected group sampling would be prohibitively slow, requiring many rounds of sampling. Instead, we use a directed approach by weighting sampled groups by the product of individual site conditional probabilities:

$$Pr(G_i|\Theta) = Pr(\xi_{i1}, .., \xi_{im}|\Theta) \sim \prod_{k=1}^{M} Pr(\xi_{ik}|\Theta)$$

$$Pr(\xi_{ik}|\Theta) = \prod_{i=1}^{L} \prod_{j \in A..T} \theta_{ij}^{\delta(s_{ijk})}$$

Where the *ith* promoter has been chosen from a set of N input promoters, there are M related orthologous promoters including the anchor species, Θ represents the position weight matrix model for the motif, each *s* term represents a site within the group *G*, and $\delta(s_{ijk})$ is an indicator function that evaluates to 1 for the residue *j* at position *i* in the *k*th ortholog and zero otherwise.

In practice, a promoter is picked at random from the input set of *N* promoters and is scored, along with its orthologs, using the current PWM. Each site in every promoter is assigned a score corresponding to the probability it was generated by the PWM vs. background. Sites are then selected for inclusion in a sampled group *G*, based on their score relative to all other sites in the same promoter. This scheme selects for groups of conserved sites since sampled groups are constrained to contain sites distributed across orthologous promoters, one site per ortholog.

Groups are sampled without prior knowledge of pre-existing groups already in the site configuration. As a result, a newly sampled group can overlap with an existing group. Depending on the overlap, there are three possible "moves": (i) addition, if there is no overlap, (ii) exchange, if there is partial overlap and (iii) removal, if there is complete overlap and an existing group is effectively "resampled".

MAP score and motif update

Each move is an incremental, "one-group" change to the site configuration. The impact of each move can be evaluated by comparing the MAP (Maximum A Posteriori) score of the updated configuration against the previous configuration. Similar to AlignACE, the MAP score is a log likelihood ratio that compares the probability of observing the motif Θ vs. background Θ_0, given groups of sites in the configuration $(G_1, G_2, .. ,G_p)$.

$$\text{MAP} = log\frac{Pr(\Theta|G_1,..,Gp)}{Pr(\Theta_0|G_1,..,Gp)}$$

The posterior probability $Pr(\Theta|G_1,..,G_p)$ is represented by the following integral:

$$Pr(\Theta|G_1,G_2,..,G_p) = \int Pr(G_1,G_2,..,G_p|\Theta)Pr(\Theta)d\Theta$$

, where $Pr(G_1,G_2,..,G_p|\Theta)$ is the probability the set of sampled groups was generated by the motif model Θ and $Pr(\Theta)$ is the prior probability for all possible PWM motifs.

Since PWMs are product multinomial distributions, we can parameterize the prior distribution using a Dirichlet distribution, which is conjugate to the product multinomial, to obtain a closed form expression for the integral:

$$Pr(G_1,..,G_p|\Theta) = \prod_{j=1}^{L} \frac{\prod_{k=1}^{A..T}\Gamma(u_k)}{\prod_{k=1}^{A..T}\Gamma(\theta_{jk}+u_k)} \frac{\Gamma(\sum_{k=1}^{A..T}\theta_{jk}+u_k)}{\Gamma(\sum_{k=1}^{A..T}u_k)}$$

Where θ_{jk} is the frequency of base *k* at position *j* in the PWM and u_k is the pseudocount for base *k* in the Dirichlet distribution, assumed to be constant for all positions *j*. Background frequencies for bases in the genome are used as pseudocounts to parameterize the Dirichlet distribution. We use a zero order markov model for the background distribution as an approximation. The algorithm can be extended to consider higher order background models as well.

We use a Metropolis-Hastings decision rule to determine if a move should be accepted: if the move improves the MAP score, it is accepted with 100% probability, otherwise, it is accepted with probability proportional to the decrease in MAP score. Thus selected non-optimal moves are permitted, preventing the algorithm from being caught in local optima. The threshold for accepting non-optimal moves is moderated using simulated annealing.

Convergence by simulated annealing

To avoid being caught in local optima, we employ a simulated annealing approach in our decision rule to accept non-optimal changes to the overall site configuration. Non-optimal changes are accepted with probability:

$$Pr(\text{update}) = \exp\left(\frac{MAP_{new} - MAP_{old}}{T_{anneal}}\right)$$

T_{anneal} determines the frequency at which non-optimal moves are permitted. When a motif search is initiated, $T_{anneal}$is set high to permit more frequent non-optimal moves. At large values of T_{anneal}, the denominator for the exponential term dominates and the update probability is skewed close to 1, even if the move decreases MAP significantly. The initial value of T_{anneal} is specified by the user with the parameter “ap_temp” (default value: 2). In practice, we have noticed setting T_{anneal} too high initially can displace the search trajectory far from its initial starting location, causing it to be preferentially attracted to the strongest motifs, at the expense of missing weaker true positive motifs.

T_{anneal} is updated every iteration to reflect an exponential annealing schedule. That is,

$$T_{anneal}^{i+1} = \alpha \cdot T_{anneal}^{i} \text{ where } 0 < \alpha < 1$$

The value of alpha is also specified by the user with the parameter “ap_alpha” (default value: 0.5). A higher value of alpha decreases the speed of annealing, vice versa. When T_{anneal} becomes small, MAP score differences are emphasized and non-optimal moves become less frequent. As a result, the motif search becomes increasingly greedy, converging on the closest local optimum. Multiple rounds of group and site sampling are possible within one iteration of the algorithm, but all potential changes to the configuration evaluated within the same iteration use a common value for T_{anneal}.

Motif similarity comparison and archival

The motif model at the end of every iteration is compared against all other motifs archived from previous searches. If the current motif is greater than 80% similar to an existing archived motif, the current search is aborted and reinitialized.

We evaluate motif similarity using a Pearson correlation coefficient because it is a simple first approximation and is amenable to PWMs. The Pearson correlation coefficient can be expressed as:

$$\rho = \frac{1}{N}\sum_{k=1}^{N} \frac{\overline{\theta_{Xk}\theta_{Yk}} - \overline{\theta_{Xk}}\,\overline{\theta_{Yk}}}{\sqrt{VAR(\theta_{Xk})}\sqrt{VAR(\theta_{Yk})}}$$

Where N is the total number of informative columns, θ_{Xk} refers to the k^{th} column of the PWM for motif X, $\overline{\theta_{Xk}}$ refers to the mean of all entries in that column, $VAR(\theta_{Xk}) = \overline{\theta_{Xk}^2} - \overline{\theta_{Xk}}^2$ is the variance of the column. The motif search is stopped after the MAP score does not increase despite a fixed number of previous iterations. At that point, if the motif is sufficiently dissimilar from previous archived motifs, it is added to the archive and reported at the end of program execution.

Technical implementation of the algorithm

Search Initialization

As input, the user specifies the total number of times Flipper reinitiates a new search by using the ap_nruns option (default: ap_nruns=100). The user also inputs FASTA formatted sequences from an anchor species, for example promoters from a set of coexpressed genes in *C.elegans*, as well as orthologous sequences from related species, for example *C.briggsae* and *C.remanei.* Orthologous sequence assignment is predetermined by the user and must be arranged in separate files in FASTA format in the same order as their corresponding sequences in the anchor sequence file. If an ortholog cannot be assigned to an anchor sequence for another species, it is noted by a blank placeholder label ">NOT FOUND", followed by a blank line to indicate no sequence.

Site sampling

In our tests on synthetic sequence data, we noticed some instances where a group had been added to the overall site configuration that was "off-by-one", i.e. containing all but one of the seeded sites for a particular ortholog group. Given enough iterations, an overlapping group with all the seeded sites including the one previously missed, would be sampled and the non-optimal "off-by-one" group would thus be replaced, but this required repeated sampling and thus convergence to the optimal motif was slow.

To speed up motif convergence, we implemented "site sampling" in our algorithm, where a site within an existing group in the configuration was chosen at random and resampled. Similar to group sampling, resampled sites are picked at random, weighted by their PWM score. The MAP score of the configuration with the resampled site is compared against the score of the previous configuration, and updates to the configuration are accepted with probability.

Search algorithm stop criteria

Motif convergence is evaluated in terms of the configuration MAP score. Since changes that improve MAP score are accepted with probability 1 and non-optimal changes accepted with probability relative to their impact on MAP score, convergence to a locally optimal motif occurs when no changes within a fixed number of sampling rounds can improve the MAP score. The iteration time window for an improvement in MAP score can be specified by the user by setting the "-minpass" option (default value: 240). From our tests on synthetic sequence data, we observe relatively fast convergence to a motif within approximately 80 to 100 iterations, depending on the difficulty of the motifs seeded.

Column sampling

Like AlignACE, Flipper allows gaps to be inserted in motifs by differentiating between motif length and the number of informative positions within a motif. The user can specify the number of informative positions by setting the "-numcols" option (default value: 10). Motifs are assumed to be ungapped at first, but gaps can be inserted by exchanging existing motif positions for neighboring positions, if the exchange improves the information content of the PWM at that position. The algorithm considers neighboring columns within one motif length upstream or downstream and imposes a binomial coefficient penalty to limit excess growth of motif length. Neighboring positions are sampled weighed by their information content, and non-optimal column exchanges are permitted with probability.

4.3.3 Generation of synthetic sequence test sets

We developed a Perl script to generate a list of synthetic promoters that could be seeded with predetermined motif sites. As an initial step, background sequences are generated randomly using a zeroth order Markov model with 33% GC content. Sequence length is normally distributed with a specified mean and variance. Five PWMs, each ten columns wide, were generated at random. We changed the number of degenerate columns, as well as the polarization of each column, to simulate motifs with varying difficulties. (Similar to other algorithm tests on synthetic data[55,56], polarization refers to non-specific deviations from the consensus residue.) Each sequence is seeded with a single instance of each PWM. We refer to this initial sequence set as the "ancestral" sequence set. We simulated evolution on the ancestral set, allowing point mutations, insertions and deletions to occur, to obtain multiple "descendant" sequence sets. The rates of mutation can be modified by the user to simulate varying rates of divergence. Our evolution model is similar to the HKY85 model of nucleotide evolution[85], in that all four nucleotides are present at different frequencies. However, instead of modeling transitions and transversions with distinct rates, we allowed point mutations occurring in unseeded sequence to be fixed according to the zeroth order Markov background model and mutations occurring in seeded sequence to be fixed according to the PWM model that generated the relevant site. Mutation rates were modeled to proportional to the evolutionary distance between species.

4.3.4 Ortholog assignment via coding sequence homology

Homologs to protein-coding genes in *S.cerevisisae* were determined for *S.paradoxus, S.mikatae* and *S.bayanus* using reciprocal best BLAST hits. Orthologous promoters for *S.cerevisiae* were defined to be sequence 800 bp upstream of homologous genes. Similarly in *C.elegans*, we defined promoters as 2000 bp upstream of translation start. For *C.elegans,* we used *C.briggsae, C.remanei* and *C.brenneri* for species comparison. Orthologous promoters were defined similarly by looking for regions upstream of homologous genes. Translation start served as a proxy for transcription start, since it is often better annotated. Furthermore, the difference between transcription and translation start may be small for compact genomes, i.e. yeast.

4.3.5 Ortholog assignment via whole genome alignments

Because most ChIP regions in *D.melanogaster* do not fall within 2000 bp upstream of translation start, we could not use our previous approach based on coding sequence homology to determine orthologous regions. Instead, we downloaded pairwise whole genome alignments from the UCSC genome browser between *D.melanogaster* and *D.yakuba, D.ananassae* and *D.pseudoobscura* etc. The downloaded whole genome alignments were in net file format, having being generated from individual chain files using the UCSC *axtChain* and *axtNet* software packages[86]. For genomic coordinates identified as being ChIP bound in Li *et al* [79], _we extracted corresponding coordinates in the compared species and assigned them as orthologous regions. To avoid spurious assignments due to misalignment, we required coordinates to mapped onto contiguous blocks of sequence on the same chromosome in the compared species. Furthermore, since the mapped coordinates may extend into regions of gapped alignment for which it may be unclear how many gaps to include in determining region boundaries, in such cases regions are mapped by the largest "subsequence" that is flanked by two regions of ungapped aligment and boundaries are extended from these flanking regions by linear extrapolation to match lengths from the initial mapped coordinates.

To ensure both methods of ortholog assignment are consistent, we mapped ChIP region coordinates from *C.elegans* TF datasets with both coding sequence homology, as well as whole genome alignment methods. As expected, we found that the agreement between the two methods was high, with close to 90% overlap between coordinates mapped by both methods. In general, we were able to map more ChIPed regions from the anchor species (*C.elegans, D.melanogaster*) using whole genome alignment, since it does not require the ChIPed region to be within 2000 bp upstream of translation start of a gene and unambiguously assigned as the promoter of that gene. Ortholog assignment via whole genome alignment will likely be more applicable to larger genomes with longer intergenic regions.

4.3.6 Input parameters for compared algorithms

Algorithms were used to recover known TF binding specificities from synthetic and experimental sequence datasets and were evaluated in terms of their true positive rate (percent seeded sites recovered) and false positive rate (percent unseeded sites reported). Motif similarity to the known binding consensus was assessed using a Pearson correlation coefficient. As a control, we ran the algorithms on random promoter sequence sets, scored the reported motifs for similarity to the known motif and used the range of scores as a null distribution. The null distribution was parameterized as Gaussian and was used to assign a p-value to the significance of the motif hit found with the original test set.

The following parameters were used on synthetic sequence test sets:

a) AlignACE[48]: -gcback 0.33 –expect 40

b) MEME[49]: -dna -mod zoops –maxsize 200000 -nmotifs 60 -w 10 -wnsites 0.8 -nomatrim -revcomp -maxiter 500 -text –nostatus

c) BioProspector[52]: -W 10 –n 60 –r 60

d) CONSENSUS[53]: -a consensus/alphabet -c1 -L 10 -n 75 -pt 5

e) PhyME[56]: -w 10 –revcompW –nmotifs 40 –nsites 10 –ot 0.5 –niter 100 –nseediter 50

f) PhyloGibbs[55]: -D 1 -m 10 -N 1 -c -1 -z 40 -T 0.8

g) PhyloCON[60]: -a consensus/alphabet -o1 -s 2

The following parameters were used on *in vivo* binding datasets:

a) PhyloGibbs[55]: -m 10 -N 3 -c -1 -z 20 -T 0.8

b) PhyME[56]: -w 10 -revcompW -nmotifs 20 -ot 0.5 -niter 100 -nseediter 50

4.3.7 Prediction of novel motifs from sets of coexpressed *C.elegans* genes

We downloaded measurements of gene expression levels from multiple stages in *C.elegans* development (embryonic and adult) from Hill et al[87] and Baugh et al[88]. We augmented this expression data with measurements of tissue specific gene expression using SAGE obtained from McKay et al[89]. Expression profiles were normalized prior to *k*-means clustering. We ran Flipper on promoters from each of the resulting 30 clusters, using orthologous regions from *C.briggsae, C.brenneri* and *C.remanei.* Reported motifs were assessed for their specificity using the following approach. PWMs were used to scan all promoters in *C.elegans* genome for motif sites. Promoters containing sites exceeding a scan score cutoff were filtered by their overlap with known ribosomal protein genes. A *p*-value for the significance of the overlap was computed using a hypergeometric null distribution. The scan score cutoff was varied and set to maximize the ovelap significance *p* –value. Promoters with scores exceeding the optimal cutoff were considered to contain high confidence motif sites. Motifs were also ranked by their *p*-values, which measured their specificity for ribosomal protein genes.

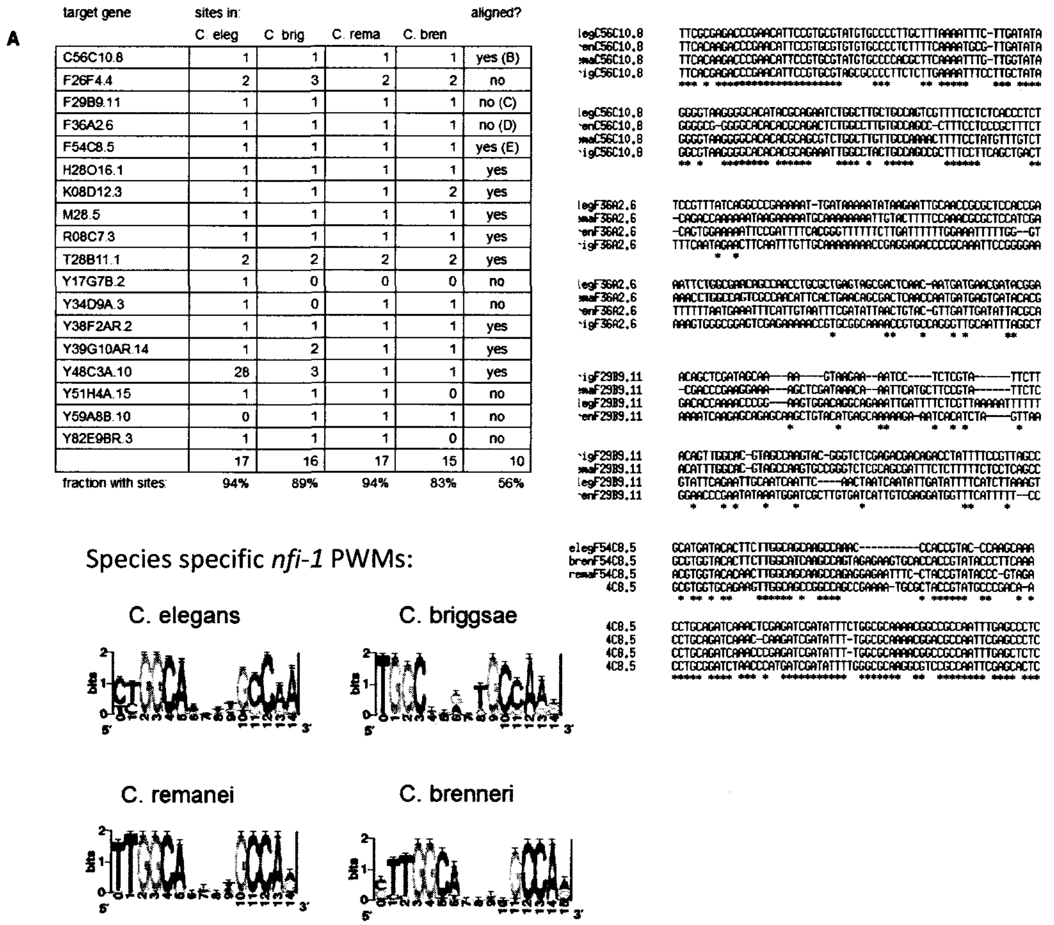
The distribution of scan scores for all promoters in the genome wide scan was also plotted and parameterized using a symmetric normal null distribution. GO terms associated with promoters containing high confidence motif sites were extracted from a GO term database located at www.geneontology.org[90].

4.4 RESULTS

4.4.1 TF binding sites have binding affinities conserved across species, but often fail to align.

Previous attempts to incorporate sequence conservation into motif search have commonly employed an alignment-based framework. i.e. PhyloGibbs[55,56 60]. This approach assumes that the input sequences are reasonably alignable, and that TF binding sites (TFBS) are likely to be enriched in blocks of aligned sequence. Given the uncertainty regarding conservation of functional TF binding sites raised by the ENCODE pilot project, we decided to investigate the degree to which TF binding sites as confirmed by experiment (i.e. chromatin immunoprecipitation or ChIP) align across multiple species.

Towards this end, we studied functional binding sites for a well characterized *C.elegans* TF, *nfi-1*[78] . The authors identified 30 genomic loci bound with high confidence by *nfi-1*, the homolog of the four vertebrate NFI genes (NFIA, NFIB, NFIC, and NFIX in humans). Both the mouse and human NFI transcription factors have been very well characterized and their binding site has been identified by *in vitro* selection with the palindromic consensus TTGGCA---TGCCAA . Nineteen (19) of the 30 ChIPed loci reside in intergenic regions for which orthologous neighboring genes are clearly identifiable in related nematode species, *C.briggsae, C.brenneri and C.remanei.* We ran AlignACE on the ChIPed sequences and their orthologs, to find over-represented sequence elements in each species independently. As expected, in each species we found a strong consensus to the mammalian NFI binding site, as shown in Figure 4.1, with some minor variations. Fig 4.1A shows the distribution of NFI sites over the 19 genes in each of the four species. Running on each species independently, AlignACE recovers the NFI binding site in 86% of the nematode orthologous target genes. We have also run alignment programs (ClustalW, DIALIGN, M-LAGAN, and MULTIZ) on these sets of orthologous genes. As shown in Fig 4.1A, only 56% of the binding sites aligned over all four species (rightmost column in Fig 4.1A),. Examples where all NFI sites aligned are shown in Fig 4.1B and 4.1E (using ClustalW format for readability, the results from other alignment programs are similar, as shown below). Examples where sites are present but do not align are shown in Fig 4. 1C and 4.1D. Typically sites fail to align when sequence similarity aligns a distal portion of the intergenic region at the expense of the shorter (~15bp) NFI site. In Fig 4.1C the sites in *elegans* and *remanei* align, but *briggsae* and *brenneri* are missed, in Fig 4.1D only sites in *briggsae* and *remanei* align.



A

target gene	sites in: C. eleg	C brig	C. rema	C. bren	aligned?
C56C10.8	1	1	1	1	yes (B)
F26F4.4	2	3	2	2	no
F29B9.11	1	1	1	1	no (C)
F36A2.6	1	1	1	1	no (D)
F54C8.5	1	1	1	1	yes (E)
H28O16.1	1	1	1	1	yes
K08D12.3	1	1	1	2	yes
M28.5	1	1	1	1	yes
R08C7.3	1	1	1	1	yes
T28B11.1	2	2	2	2	yes
Y17G7B.2	1	0	0	0	no
Y34D9A.3	1	0	1	1	no
Y38F2AR.2	1	1	1	1	yes
Y39G10AR.14	1	2	1	1	yes
Y48C3A.10	28	3	1	1	yes
Y51H4A.15	1	1	1	0	no
Y59A8B.10	0	1	1	1	no
Y82E9BR.3	1	1	1	0	no
	17	16	17	15	10
fraction with sites	94%	89%	94%	83%	56%

Species specific *nfi-1* PWMs:

```
legC56C10.8     TTCGCGAGACCCGAACATTCCGTGCGTATGTGCCCCTTGCTTTAAAATTTC-TTGATATA
enC56C10.8      TTCACAAGACCCGAACATTCCGTGCGTGTGTGCCCCTCTTTTCAAAATGCG-TTGATATA
maC56C10.8      TTCACAAGACCCGAACATTCCGTGCGTATGTGCCCCACGCTTCAAAATTTG-TTGGTATA
igC56C10.8      TTCACGAGACCCGAACATTCCGTGCGTAGCGCCCCTTCTCTTGAAAATTTCCTTGCTATA
                *** * ************************      ***     ** *****    *** ****

legC56C10.8     GGGGTAAGGGGCACATACGCAGAATCTGGCTTGCTGCCAGTCGTTTTCCTCTCACCCTCT
enC56C10.8      GGGGCG-GGGGCACACACGCAGACTCTGGCCTTGTGCCAGCC-CTTTCCTCCCGCTTTCT
maC56C10.8      GGGGTAAGGGGCACACACGCAGCGTCTGGCTTGTTGCCAAAACTTTTCCTATGTTTGTCT
igC56C10.8      GGCGTAAGGGGCACACACGCAGAAATTGGCCTACTGCCAGCCGCTTTCCTTCAGCTGACT
                ** *   ******** *******     **** * *****     *******          **

legF36A2.6      TCCGTTTATCAGGCCCGAAAAAT-TGATAAAAATATAAGAATTGCAACCGCGCTCCACCGA
maF36A2.6       -CAGACCAAAAAATAAGAAAAATGCAAAAAAAAATTGTACTTTTCCAAACGCGCTCCATCGA
enF36A2.6       -CAGTGGAAAAATTCCGATTTTCACGGGTTTTTCTTGATTTTTTGGAAATTTTTGG--GT
igF36A2.6       TTTCAATAGAACTTCAATTTGTTGCAAAAAAAAACCGAGGAGACCCCGCAAATTCCGGGGAA
                      *  *

legF36A2.6      AATTCTGGCGAACAGCCAACCTGCGCTGAGTAGCGACTCAAC-AATGATGAACGATACGGA
maF36A2.6       AAACCTGGCCAGTCGCCAACATTCACTGAACAGCGACTCAACCAATGATGAGTGATACACG
enF36A2.6       TTTTTTAATGAAATTTCATTGTAATTTCGATATTAACTGTAC-GTTGATTGATATTACGCA
igF36A2.6       AAAGTGGGCGGAGTCGAGAAAAACCGTGCGGCAAAACCGTGCCAGGGTTGCAATTTAGGCT
                                         *       **    *    * *      **

igF29B9.11      ACAGCTCGATAGCAA---AA----GTAAGAA--AATCC----TCTCGTA------TTCTT
maF29B9.11      -CGACCCGAAGGAAA---AGCTCGATAAACA--AATTCATGCTTCCGTA------TTCTC
legF29B9.11     GACACCAAAACCCGG---AAGTGGACAGGCAGAAATTGATTTTCTCGTTAAAAATTTTTT
enF29B9.11      AAAATCAAGAGCAGAGCAAGCTGTACATGAGCAAAAAGA-AATCACATCTA----GTTAA
                                  *      *     **       *  * *        *

igF29B9.11      ACAGTTGGCAC-GTAGCCAAGTAC-GGGTCTCGAGACGACAGACCTATTTTCCGTTAGCC
maF29B9.11      ACATTTGGCAC-GTAGCCAAGTGCCGGGTCTCGCAGCGATTTCTCTTTTTCTCCTCAGCC
legF29B9.11     GTATTCAGAATTGCAATCAATTC----AACTAATCAATATTGATATTTTCATCTTAAAGT
enF29B9.11      GGAACCCGAATATAAATGGATCGCTTGTGATCATTGTCGAGGATGGTTTCATTTTT--CC
                  *    * *    *    *         *                **     *

elegF54C8.5     GCATGATACACTTCTTGGCAGCAAGCCAAAC----------CCACCGTAC-CCAAGCAAA
brenF54C8.5     GCGTGGTACACTTCTTGGCATCAAGCCAGTAGAGAAGTGCACCACCGTATACCCTTCAAA
remaF54C8.5     ACGTGGTACACAACTTGGCAGCAAGCCAGAGGAGAATTTC-CTACCGTATACCC-GTAGA
4C8.5           GCGTGGTGCAGAAGTTGGCAGCCGGCCAGCCGAAAA-TGCGCTACCGTATGCCCGACA-A
                * ** * **    ******* *  ****             * *******  **    * *

4C8.5           CCTGCAGATCAAACTCGAGATCGATATTTCTGGCGCAAAACGGCCGCCAATTTGAGCCCTC
4C8.5           CCTGCAGATCAAAC-CAAGATCGATATTT-TGGCGCAAAACGGACGCCAATTCGAGCCCTC
4C8.5           CCTGCAGATCAAACCCGAGATCGATATTT-TGGCGCAAAACGGCCGCCAATTTGAGCTCTC
4C8.5           CCTGCGGATCTAACCCATGATCGATATTTTGGGCGCAAGGCGTCCGCCAATTCGAGCACTC
                ***** **** *** * ************ ******* ** ******** **** ***
```

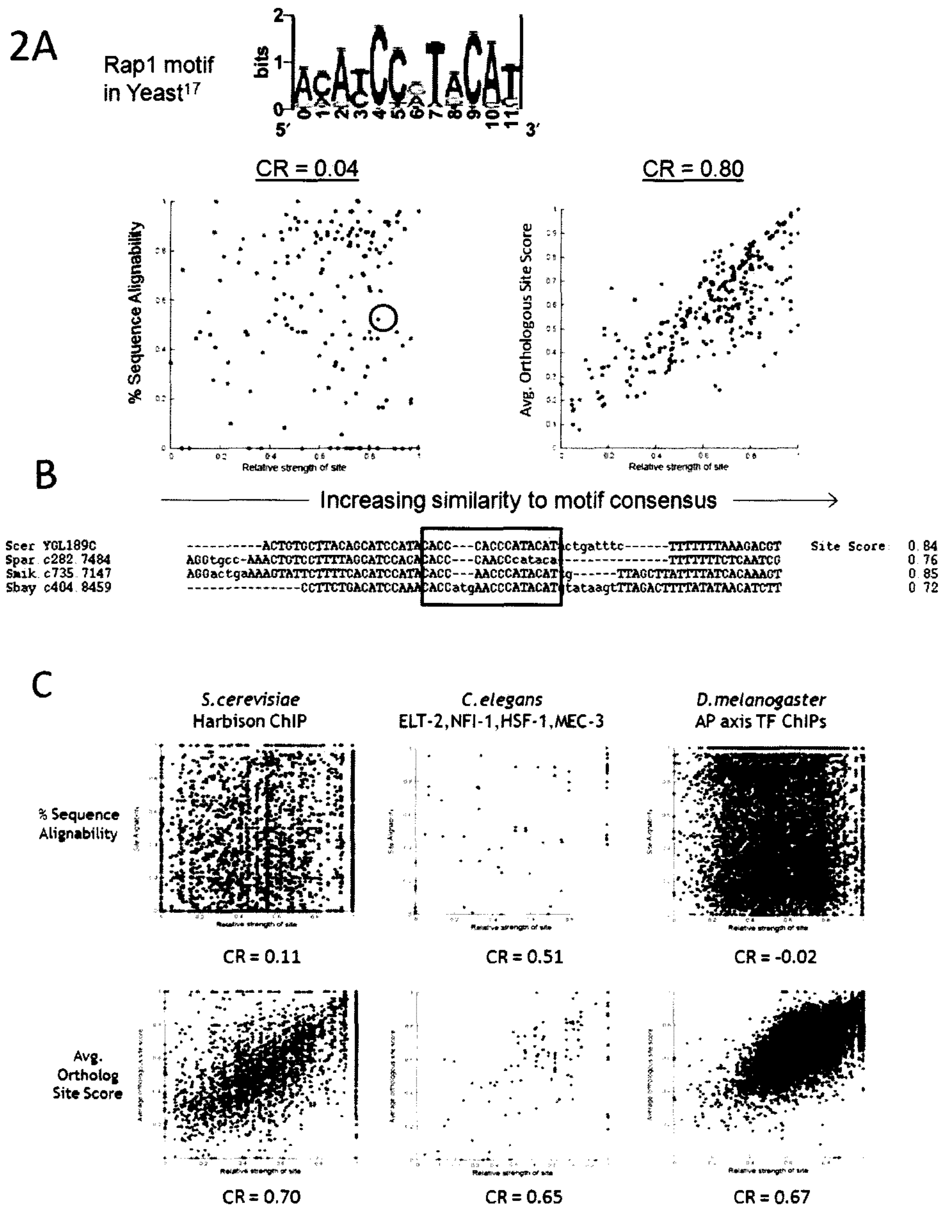
Figure 4.1: A) Binding sites detected in experimentally determined, *nfi-1* bound *C.elegans* genes and nematode orthologs, and species specific matrices for *nfi-1.* 80-90% of the orthologous regions contain conserved sites, yet only 56% are alignable by standard algorithms (B-E). Only the relevant subset of the intergenic regions is shown.

Because a large number of *nfi-1* sites fail to align, they would be missed by motif finding algorithms that rely on sequence alignment to assess binding site conservation. To determine the extent to which these results are specific only to *nfi-1*, we studied experimentally confirmed binding sites for a number of TFs from *S.cerevisiae, C.elegans* and *D.melanogaster*, asking if sequence alignability can serve as a proxy for binding site conservation. We examined the correlation between TF binding site strength and sequence alignability, expecting that if alignability were a good proxy for conservation, stronger binding sites would be more alignable.

In *S.cerevisisae*, we used binding data for 44 yeast TFs[74], as well as additional ChIP data specific to Rap1[91]. Briefly, using diAlign, we performed multiple alignments of target promoters using their orthologs from *S.bayanus, S.mikatae* and *S.paradoxus.* We also scanned the promoters using a position weight matrix (PWM) derived from the known TF binding consensus. Assuming the TF target site is the highest scoring site in each promoter, we compute the correlation between its PWM score and its alignability, measured using a sum-of-pairs (SP) statistic (see Methods). Fig 2C shows a poor correlation between binding site strength and site sequence alignability ($\rho = 0.10$) for binding sites of all yeast TFs. For example in Fig 2A, the correlation with site alignability for Rap1 binding sites from Lieb et al, is 0.04 and we observe a number of strong Rap1 sites that fail to align. We repeated this analysis for ChIP and microarray expression datasets of *C.elegans* TFs, ELT-2[75], MEC-3[76], HSF-1[77] and NFI-1[78], and ChIP datasets for

D.melanogaster TFs involved in AP axis patterning[79]. From Fig 2C, we observe that the correlations between TF binding site strength and binding sites sequence alignability were also low (ρ = 0.51 for *C.elegans* and ρ = -0.02 for *D.melanogaster*), suggesting that conservation as determined by sequence alignment, is a poor predictor for TF binding.

As an alternative, we asked if TF binding affinities are conserved, i.e. if there is correlation between the strength of a site and the presence of another strong site somewhere in each of the orthologous promoters. Similarly,orthologous promoters were scanned using the same PWM and the highest PWM score taken to be the ortholog of the target site. We computed the correlation between the PWM score of the original target site and the average score for its orthologs across all compared species. We found a consistent improvement in correlation for all *S.cerevisiae, C.elegans* and *D.melanogaster* datasets (ρ=0.70, ρ=0.65 and ρ=0.67) over previous correlations with alignability. These results suggest that while a large number of functional binding sites fail to align and are missed by multiple alignment algorithms, they can be detected by broader measures of conservation.



```
Scer YGL189C       ----------ACTGTGCTTACAGCATCCATACACC---CACCCATACATActgatttc------TTTTTTTAAAGACGT   Site Score  0.84
Spar c282 7484     AGGtgcc-AAACTGTCCTTTTAGCATCCACACACC---CAACCcataca--------------TTTTTTTCTCAATCG                 0.76
Smik c735 7147     AGGactgaAAAGTATTCTTTTCACATCCATACACC---AACCCATACATtg------TTAGCTTATTTTATCACAAAGT                0.85
Sbay c404 8459     ---------------CCTTCTGACATCCAAACACCatgAACCCATACATtataagtTTAGACTTTTATATAACATCTT                  0.72
```

Figure 4.2: A) Rap 1 binding site strength is poorly correlated with site sequence alignability (cr=0.04), but more strongly correlated with the co-occurrence of similarly scoring sites in orthologous regions (cr=0.80). Correlations were computed in terms of a Pearson Correlation coefficient. B) YGL189C is an example of a promoter that has a strong Rap1 binding site that does not align well. The binding site proper is highlighted by the red box. C) Our previous result is not unique to Rap1 - binding site strength is better correlated with orthologous site scores than site sequence alignability for other datasets as well: ChIP datasets for 44 yeast TFs (cr=0.70 vs. cr=0.11), ChIP and expression datasets for 5 worm TFs (cr=0.65 vs. cr=0.51) and ChIP datasets for 6 fly TFs involved in patterning of the AP axis (cr=0.67 vs. cr=-0.02).

4.4.2 Flipper iteratively samples groups of conserved sites across orthologous regions to update a PWM model of the motif.

We have developed a Gibbs sampling algorithm for *ab initio* motif discovery, Flipper, to search for over-represented and conserved motifs in groups of orthologous promoters without relying on sequence alignment to assess conservation. Instead, if a promoter contains a conserved binding site, we expect another strong site to be present somewhere in each orthologous promoter. Our algorithm searches for conserved binding sites by this definition, by sampling groups of sites in orthologous intergenic regions (one site in each region), rather than single sites in individual regions.

An outline of the algorithm structure is presented in Fig 4.3. As input, Flipper takes the promoter sequences from a set of coexpressed genes and corresponding orthologous promoters from related species. To begin the motif search, a site is chosen randomly at first to initialize a PWM model. Using a Gibbs sampling approach, Flipper searches for a PWM model that maximizes the posterior probability of observing the motif, given the data. It does so by repeatedly sampling groups of sites for admission to the overall site configuration. A sampled group is assigned a posterior probability score that reflects the likelihood of observing the motif, given the sites within the group. The posterior probability score affects how frequently a group is sampled, with high scoring groups being sampled more frequently than lower scoring ones. A sampled group can change the overall site configuration in three possible scenarios – it can be added to the configuration if it does not overlap an existing group, replace an existing group that it partially overlaps, or be removed from the configuration if it is an existing group that happens to be resampled.

The overall site configuration is evaluated using a MAP score, which computes the posterior probability of observing the motif, given the sites in the configuration. Changes made to the site configuration are accepted using a Metropolis-Hastings decision rule: if the change improves the MAP score of the configuration , the configuration is updated; if not, the configuration is updated with probability, according to a simulated annealing schedule. Details regarding algorithm structure and technical implementation were previous described in Methods.

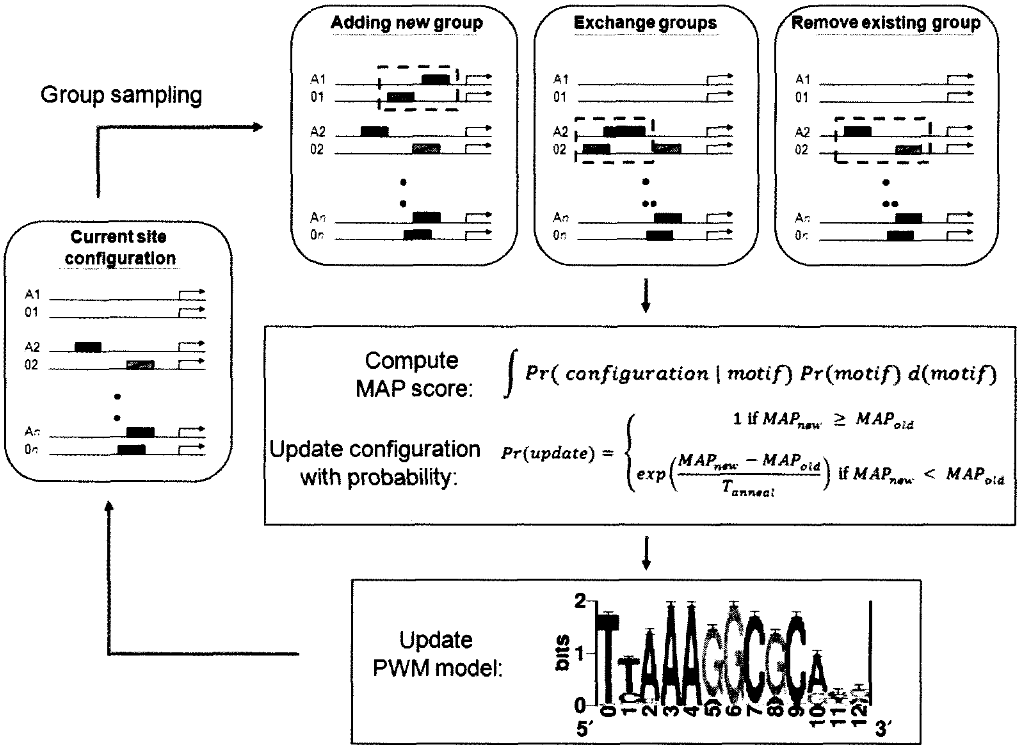

Group sampling
Adding new group
Exchange groups
Remove existing group
Current site configuration
A1
01
A2
02
An
On
Compute MAP score:
$\int Pr(\ configuration \mid motif)\ Pr(motif)\ d(motif)$
Update configuration with probability:
$Pr(update) = \begin{cases} 1 \text{ if } MAP_{new} \geq MAP_{old} \\ exp\left(\frac{MAP_{new} - MAP_{old}}{T_{anneal}}\right) \text{ if } MAP_{new} < MAP_{old} \end{cases}$
Update PWM model:
bits
2
1
0
5′
3′
0 1 2 3 4 5 6 7 8 9 10 11 12

Figure 4.3: Outline of the Flipper algorithm. The input consists of *n* promoters from the anchor species (A1, A2 … A*n*) and orthologous promoters from a related species (O1, O2 … O*n*). Sites are sampled in groups, distributed across orthologous promoters. Three scenarios are possible in "group sampling": a) the sampled group does not overlap with any existing group in the configuration and is considered for addition, b) the sampled group contains some sites that overlap with an existing group, and is considered for exchange, c) the sampled group is an existing group, and is considered for removal from the configuration. Dashed lines highlight events a) b) and c). The likelihood that the configuration was generated by the PWM vs. background is computed and integrated using a dirichlet prior to obtain the MAP score. MAP scores of the new configuration are computed and compared relative to the previous configuration. The configuration is updated if the sampled group improves overall MAP. A sampled group that lowers configuration MAP is accepted with probability, based on a simulated annealing algorithm, indicated by T_{anneal}. The PWM is updated to reflect changes in site configuration and is used to score new groups in the next round of "group sampling".

4.4.3 Flipper outperforms other algorithms on synthetic sequence test sets.

To evaluate the performance of Flipper against other competing motif discovery algorithms, we compared algorithm performance on generated synthetic sequence data. Briefly, PWMs were generated and motif instances were seeded into background sequence of an ancestral sequence set. Simulated evolution was performed on the background sequence set to generate descendant sequence sets. The algorithms were assessed on their ability to recover seeded binding sites from the descendant sequence sets. To model degeneracy in TFBS, we varied the information content in the generated PWMs by changing the column degeneracy and the column polarization (percentage of consensus-matching bases in that particular position). The point mutation and in-del rates were also varied to simulate differences in evolutionary divergence between species.

Flipper was compared against two classes of algorithms: a) 'non-phylogenetic' algorithms that search for over-represented sites, i.e. AlignACE[48], MEME[49], CONSENSUS[53] and BioProspector[52] and b) 'phylogenetic' algorithms that use sequence alignment to bias their motif search, i.e. PhyloGibbs[55], PhyME[56] and PhyloCON[60]. Algorithms were evaluated for their sensitivity, measured by the percentage of seeded bases successfully recovered, as well as their false discovery rate (FDR), measured by the percentage of unseeded bases falsely reported as positive.

Flipper vs. Non-Phylogenetic algorithms.

An advantage of having more orthologous sequence available for comparison is the benefit of a greater signal-to-noise ratio. We wanted to investigate if Flipper could perform better against conventional "non-phylogenetic" algorithms, which could be run on pooled sequence datasets from multiple species without any explicit consideration of the underlying phylogeny. The algorithms were used to recover moderately degenerate motifs from sequence datasets, where the number of species compared was increased from 2 to 5. The average length of sequences generated was 2000 bp. We generated PWMs for motifs containing 4 degenerate columns with a polarization of 85% for non-degenerate columns. For these parameters, the information content of the generated motifs was 55%. We repeated our simulations 10 times using different random seeds. From Fig 4.4A and 4.4B, we observe Flipper consistently outperformed competing nonalignment techniques for all numbers of species compared (average AUC for Flipper = 77%, average AUC for other algorithms = 67%). We also observed that the performance of Flipper improved as more species were compared, while the performance of non-alignment algorithms remained constant.

Flipper vs. Alignment based algorithms.

Because algorithms such as PhyloGibbs[55], PhyME[56] and PhyloCON[60] depend on an external alignment pre-processing step to determine sequence conservation, we hypothesized that their performance might be susceptible to changes in input sequence alignability. We performed a simulation to model a situation where the descendant species become increasingly divergent, by increasing the probability of mutation from 0.6 to 0.8. Using a

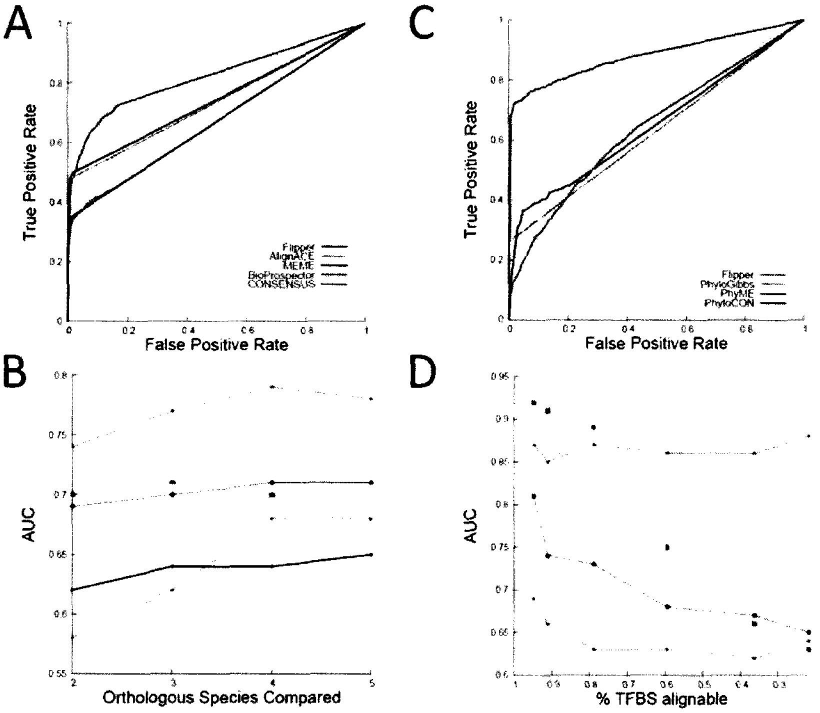


Figure 4.4:
Flipper outperforms competing algorithms on synthetic sequence test sets. A) and B) compare the performance of Flipper against conventional motif finding algorithms: AlignACE, MEME. BioProspector and CONSENSUS. C) and D) show comparisons against alignment-based phylogenetic motif finding algorithms: PhyloGibbs, PhyME and PhyloCON. A) ROC curve comparing percent seeded sites successfully recovered (y-axis) vs. percent unseeded sites incorrectly reported (x-axis) when 4 species are used for comparison. B) AUC (area under the ROC curve) measures overall algorithm performance as the number of orthologous species compared increases from 2 to 5. C) ROC curve comparing percent seeded sites successfully recovered (y-axis) vs. percent unseeded sites incorrectly reported (x-axis) when probability of mutation was increased to 0.8 and average TF binding site alignability was decreased to 22%. D) AUC is measured for each algorithm, simulating scenarios where mutation rates are increased from 0.6 to 0.8 and resulting TF binding site alignability is decreased from 95% to 22%.

sum-of-pairs statistic to score alignments of the input sequences, we found that this reduced the overall alignability of seeded motif sites from 95% to 22%.

When Flipper, PhyloGibbs, PhyME and PhyloCON were used to recover seeded sites in these sequence sets, we observed that the AUC for alignment based algorithms fell sharply as the alignability dropped past 70% (Fig 4C and 4D). In contrast, the AUC for Flipper remained constant at 87%, confirming that our sampling strategy allows it to be robust to changes in sequence alignability.

4.4.4 Flipper recovers biologically relevant motifs from experimental data.

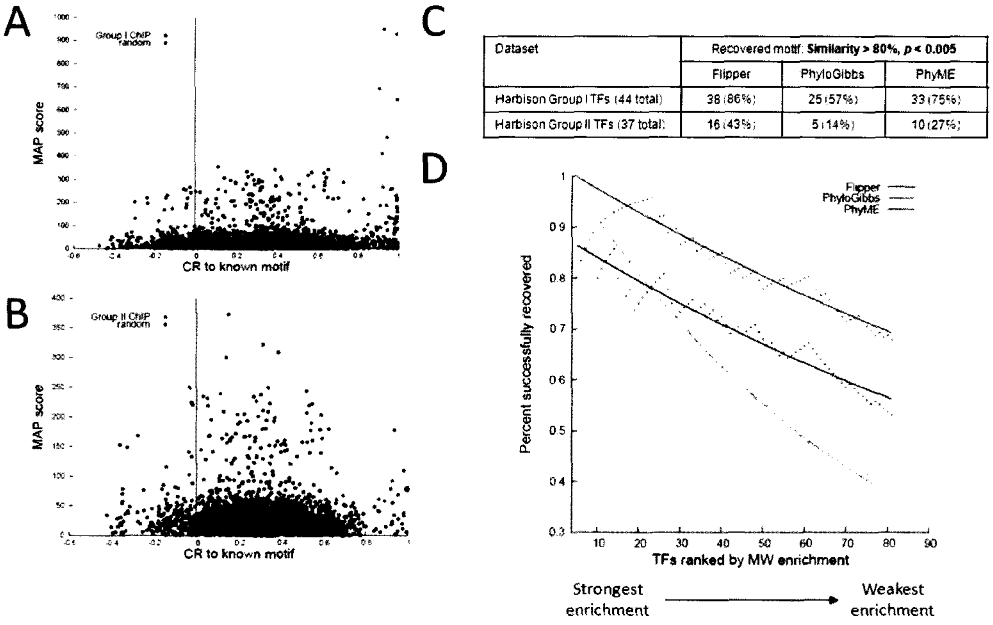
We used Flipper, PhyloGibbs and PhyME to recover known TF binding specificities from promoters that had been identified as TF targets either through ChIP or microarray expression experiments. We used datasets from *S.cerevisiae, C.elegans* and *D.melanogaster.* Motifs reported by each algorithm were converted into PWMs and were ranked by their similarity to the known literature consensus. The algorithms were also run on a control set of randomly selected *S.cerevisiae* promoters and respective orthologs. We considered an algorithm successful, if it reported a motif on the test set that at least 80% similar to the known consensus and significantly more similar than other motifs reported on the control set ($p < 0.005$).

For *S.cerevisiae*, we used ChIP-chip data[74] for 102 yeast transcription factors, 81 of which had a previously determined binding consensus. Using a suite of 6 different motif finders, the authors were able to recover the known binding site from the ChIPed sequences for 44 regulators (Group I); we refer to the remaining

37 TFs as Group II. We used Flipper, PhyloGibbs and PhyME to recover binding motifs for datasets from Group I and II, using *S. paradoxus*, *S. mikitae* and *S. bayanus* as related species.

We observed that Flipper was more successful than PhyloGibbs and PhyME at recovering the known motif from datasets in both groups. Specifically, Flipper was successful in 38 out of 44 cases from Group I, compared to 25 and 30 for PhyloGibbs and PhyME. Flipper was also successful in 16 out of 37 cases in Group II, compared to 5 and 10 for PhyloGibbs and PhyME.

We asked how the performance of Flipper, PhyloGibbs and PhyME compared as motif enrichment in the sequence sets decreased. We measured motif enrichment by performing a Mann Whitney U test, comparing promoters from the ChIP dataset against control sets of randomly chosen promoters. We grouped together the five highest ranked datasets based on their motif enrichment *p* values and asked how the algorithms had performed for these datasets. We then repeatedly expanded this group, by adding with each iteration next highest ranked dataset by motif enrichment. In doing so, we could compare average algorithm performance as datasets with weaker "signal" were admitted (Fig 4.5D). We observe that Flipper recovers the known motif with high sensitivity for datasets with strong motif enrichment, which decreases gradually with weaker datasets. In contrast, the performance of PhyloGibbs drops rapidly as with weaker datasets, indicating that its success may be restricted to only datasets with strong motif enrichment.

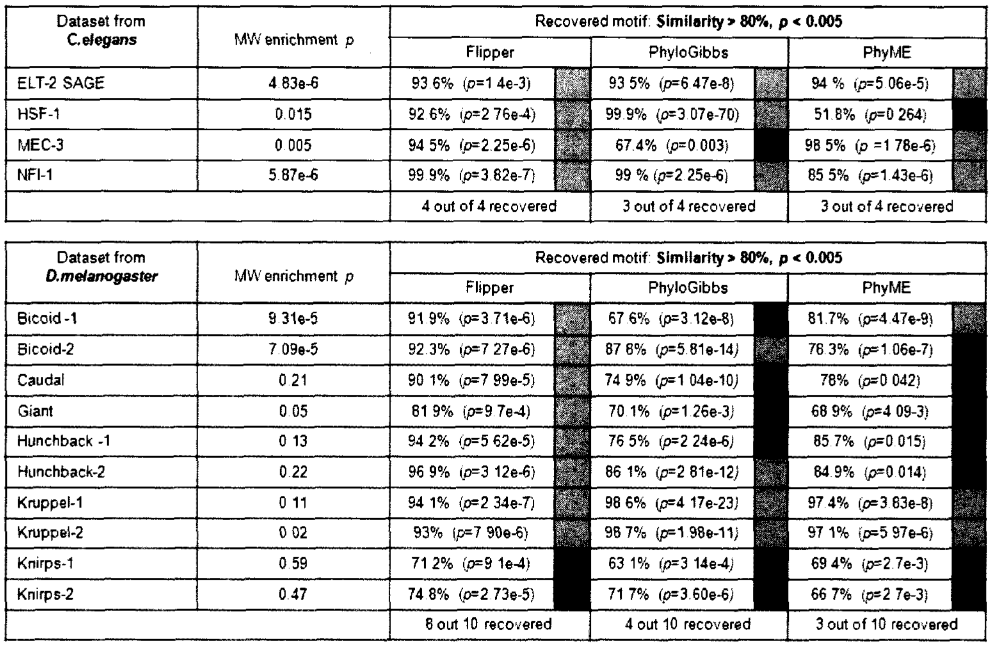


Dataset	Recovered motif: **Similarity > 80%, *p* < 0.005**		
	Flipper	PhyloGibbs	PhyME
Harbison Group I TFs (44 total)	38 (86%)	25 (57%)	33 (75%)
Harbison Group II TFs (37 total)	16 (43%)	5 (14%)	10 (27%)

Figure 4.5:
Flipper recovers known motifs from ChIP datasets of yeast TFs in Harbison et al. A) and B) plot the MAP score and similarity of motifs successfully recovered by Flipper in red, on A) 44 datasets where computational methods in Harbison et al had succeeded in recovering the known motif, and B) 37 datasets where the same computational methods failed. As a control, Flipper was also run on control datasets with size-matched sequences randomly chosen from the yeast genome. The MAP and similarity scores for all motifs reported on control datasets are plotted in black. C) The number of motifs successfully recovered within a significance threshold (similarity > 80%, $p < 0.005$) are reported for Flipper, PhyloGibbs and PhyME. D) The fraction of successfully recovered motifs is evaluated for Flipper, PhyloGibbs and PhyME as average MW enrichment within the dataset cluster decreases. Exponential fits to the actual datapoints are included to emphasize trends in the data. P values reported are medians of 5 independent control datasets.

For *C.elegans*, we used datasets listing genes identified as targets of ELT-2[75], MEC-3[76], HSF-1[77,78] and NFI-1[78], from a variety of ChIP-chip, SAGE and microarray expression experiments. Promoters of these genes were defined as sequence up to 2000 bp upstream of translation start. Orthologs of *C.elegans* promoters were taken to be sequence upstream of homologous genes in related species, *C.briggsae, C.brenneri* and *C.remanei.* We also performed a Mann-Whitney (MW) U test on the target promoters, asking how enriched they were for the known binding motif of the respective TF. A summary of all MW enrichment values, reported motif similarities and their significance can be found in Fig 4.6. Using an 80% similarity cutoff and a *p*-value cutoff of 0.005, we observe that while Flipper successfully recovers the known motif on all four datasets, PhyloGibbs and PhyME recover the known motif for 3 out of 4 datasets.

For *D.melanogaster*, we used ChIP-chip datasets for TFs involved in AP axis patterning in early development: Bicoid, Caudal, Giant, Hunchback, Kruppel and Knirps[79]. We used sequences from *D.yakuba*, *D.ananassae* and *D.pseudoobscura* for comparison. We took the peak score probe regions from the top 50 binding regions from each ChIP as input sequences for motif discovery. Because only 106 of 485 probe regions (22%) were within 2000 bp upstream of a gene, we were unable to use our previous definition of a promoter to define orthologous regions. Instead, we directly mapped orthologous regions using pairwise whole genome alignments between *D.melanogaster* and each of its three relatives. Orthologous promoters mapped using whole genome alignment and coding sequence homology have sequence overlap of at least 80% (*dmel-dyak* = 87% overlap, *dmel-dana* = 84% overlap, *dmel-dpse* = 82% overlap), suggesting that both methods of mapping orthologous regions are consistent. We observed that Flipper was able to recover the known motif for 8 out 10 cases, compared to 4 for PhyloGibbs and 3 for PhyME.



Dataset from **C.elegans**	MW enrichment *p*	Recovered motif: **Similarity > 80%, *p* < 0.005**		
		Flipper	PhyloGibbs	PhyME
ELT-2 SAGE	4.83e-6	93.6% (*p*=1.4e-3) 🟩	93.5% (*p*=6.47e-8) 🟩	94 % (*p*=5.06e-5) 🟩
HSF-1	0.015	92.6% (*p*=2.76e-4) 🟩	99.9% (*p*=3.07e-70) 🟩	51.8% (*p*=0.264) 🟥
MEC-3	0.005	94.5% (*p*=2.25e-6) 🟩	67.4% (*p*=0.003) 🟥	98.5% (*p* =1.78e-6) 🟩
NFI-1	5.87e-6	99.9% (*p*=3.82e-7) 🟩	99 % (*p*=2.25e-6) 🟩	85.5% (*p*=1.43e-6) 🟩
		4 out of 4 recovered	3 out of 4 recovered	3 out of 4 recovered

Dataset from **D.melanogaster**	MW enrichment *p*	Recovered motif: **Similarity > 80%, *p* < 0.005**		
		Flipper	PhyloGibbs	PhyME
Bicoid -1	9.31e-5	91.9% (*p*=3.71e-6) 🟩	67.6% (*p*=3.12e-8) 🟥	81.7% (*p*=4.47e-9) 🟩
Bicoid-2	7.09e-5	92.3% (*p*=7.27e-6) 🟩	87.8% (*p*=5.81e-14) 🟩	78.3% (*p*=1.06e-7) 🟥
Caudal	0.21	90.1% (*p*=7.99e-5) 🟩	74.9% (*p*=1.04e-10) 🟥	78% (*p*=0.042) 🟥
Giant	0.05	81.9% (*p*=9.7e-4) 🟩	70.1% (*p*=1.26e-3) 🟥	68.9% (*p*=4.09-3) 🟥
Hunchback -1	0.13	94.2% (*p*=5.62e-5) 🟩	76.5% (*p*=2.24e-6) 🟥	85.7% (*p*=0.015) 🟥
Hunchback-2	0.22	96.9% (*p*=3.12e-6) 🟩	86.1% (*p*=2.81e-12) 🟩	84.9% (*p*=0.014) 🟥
Kruppel-1	0.11	94.1% (*p*=2.34e-7) 🟩	98.6% (*p*=4.17e-23) 🟩	97.4% (*p*=3.83e-8) 🟩
Kruppel-2	0.02	93% (*p*=7.90e-6) 🟩	98.7% (*p*=1.98e-11) 🟩	97.1% (*p*=5.97e-6) 🟩
Knirps-1	0.59	71.2% (*p*=9.1e-4) 🟥	63.1% (*p*=3.14e-4) 🟥	69.4% (*p*=2.7e-3) 🟥
Knirps-2	0.47	74.8% (*p*=2.73e-5) 🟥	71.7% (*p*=3.60e-6) 🟥	66.7% (*p*=2.7e-3) 🟥
		8 out 10 recovered	4 out 10 recovered	3 out of 10 recovered

Figure 4.6:
Flipper successfully recovers known motifs from ChIP, SAGE and microarray expression datasets in *C.elegans* and *D.melanogaster.* Green boxes indicate datasets in which an algorithm recovered a motif that is at least 80% similar to the known binding specificity ($p < 0.005$). Red boxes indicate datasets where an algorithm was unsuccessful. Significance of reported motif hits was assessed by comparing against motifs reported by the algorithms on control datasets of size-matched sequences randomly selected from the *C.elegans* or *D.melanogaster* genomes. P values reported are medians of 5 independent control datasets.

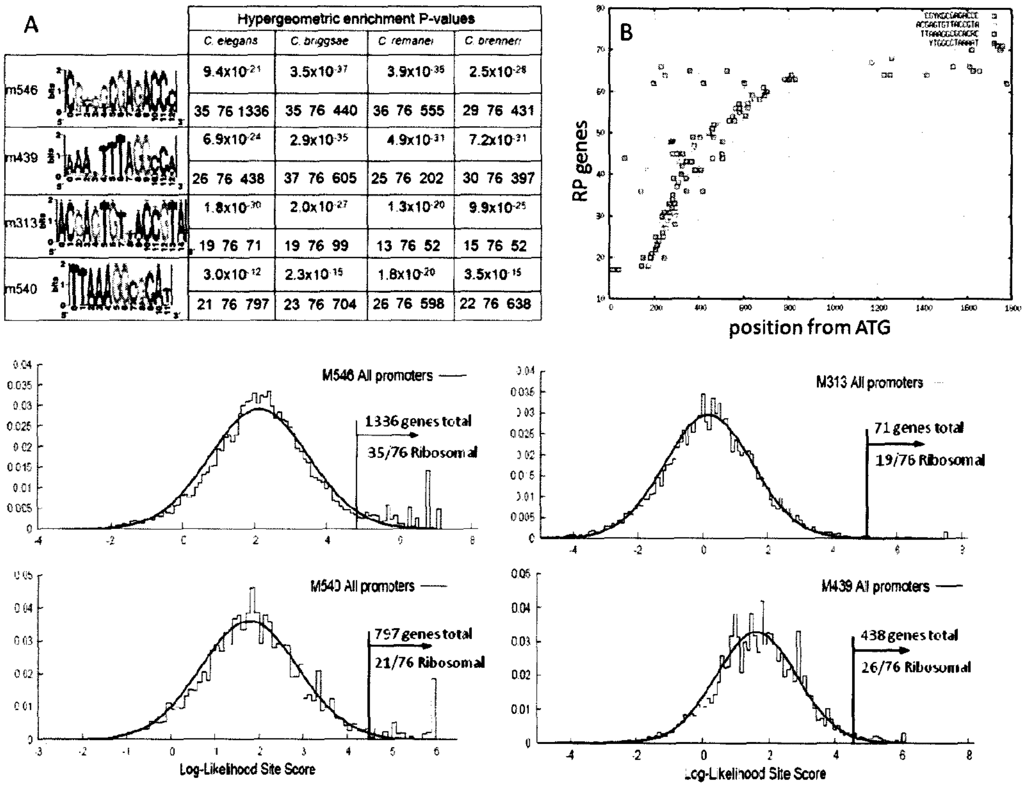
4.4.5 Novel *cis*-regulatory elements from *C.elegans* coexpression data

Having validated our algorithm on on control datasets of known TF targets, we next asked if we could use Flipper to discover novel sequence motifs from promoters of coexpressed genes in *C.elegans*. Using *k*-means clustering, we clustered genes based on their expression profiles from microarray and SAGE datasets[87-89]. We used a pairwise Pearson correlation coefficient to assess the degree of coexpression between genes. Orthologs for promoters were mapped using coding sequence homology in *C.briggsae, C.remanei* and *C.brenneri.* We used Flipper, PhyloGibbs and PhyME for motif discovery on the promoter sequence sets. Reported motifs were evaluated for cluster specificity using a p-value generated from a hypergeometric distribution.

Our analysis clustered 169 genes as being enriched in genes coding for proteolytic enzymes, i.e. peptidases and hydrolases (p = 4.93e-12). Flipper reports a motif –ACTGATAA- for this cluster, which matches the known binding site for ELT-2 (similarity = 95%). Since ELT-2 specifies intestinal development, our result implies that genes coding for proteolytic enzymes are expressed in the intestine, consistent with their role in digestive processes.

Our analysis also identified the ribosomal protein (RP) cluster as being tightly co-expressed ($P < 10^{-216}$). Flipper predicted 22 motifs that are over-represented and highly conserved in the promoters of RP genes. We decided to investigate which selected motifs were most specific to the RP cluster. We scanned all promoters in the genome for the presence of the motif, scoring sites by computing their likelihood ratio for being generated by a PWM sequence motif model vs. a background sequence model. Promoters were ranked based on the scores of their respective maximum scoring sites. We varied the cutoff for this score to maximize the hypergeometric P-value for enrichment in the RP gene set. The most enriched motifs found were M546, M313, M540, and M439, shown in Fig 4.7. The hypergeometric P-value is shown in red in Fig 4.7A, and below that are three numbers for each species: the number of RP sequences in which at least one site is found, the number of RP sequences (76 in each species), and the number of total genes in the genome with a site above the optimal cutoff. Although the enrichment in RPs is quite significant, these motifs are found in many other genes. We also observed these motifs co-occur in spatially localized clusters of sites in ribosomal protein promoters, as shown in Fig. 4.7B, reminiscent of Drosophila enhancers.

From a distribution of all scores from the genome wide scan (Fig 4.7C), we observe more promoters contain sites scoring higher than the specificity cutoff than expected by a symmetric null distribution. Expectedly, this set included many RP promoters. We found enriched GO terms for these promoters that included "GO:0040007 Growth" and "GO:0007275 development".



	Hypergeometric enrichment P-values			
	C. elegans	*C. briggsae*	*C. remanei*	*C. brenneri*
m546	9.4x10^{-21}	3.5x10^{-37}	3.9x10^{-35}	2.5x10^{-28}
	35 76 1336	35 76 440	36 76 555	29 76 431
m439	6.9x10^{-24}	2.9x10^{-35}	4.9x10^{-31}	7.2x10^{-31}
	26 76 438	37 76 605	25 76 202	30 76 397
m313	1.8x10^{-30}	2.0x10^{-27}	1.3x10^{-20}	9.9x10^{-25}
	19 76 71	19 76 99	13 76 52	15 76 52
m540	3.0x10^{-12}	2.3x10^{-15}	1.8x10^{-20}	3.5x10^{-15}
	21 76 797	23 76 704	26 76 598	22 76 638

Figure 4.7:

A) Computationally identified conserved motifs in four nematode species and their specificity P-values in each species (red). The numbers below each P-value indicate the number of RPs which contain the motif, the number of RPs, and the total number of genes with a motif above the most specific cutoff. B) Sites for these motifs occur in spatially localized clusters in RP promoters: RP genes are sorted by the mean position of the cluster, and motif positions relative to ATG are shown. C) Genome wide scans for motif instances show that genes containing motif sites above the specificity cutoff are enriched in these sites more than expected given a symmetric null distribution.

4.5 DISCUSSION

Motivated by the lack of correlation between TF binding site strength and binding site alignability, we developed Flipper, a motif finding algorithm that searches for over-represented and conserved motifs without using sequence alignment. We tested Flipper against traditional, non-phylogenetic algorithms (i.e. AlignACE, MEME) as well as alignment-based phylogenetic methods (i.e. PhyloGibbs, PhyME). Using synthetic sequence test sets and biological datasets, we showed that Flipper outperforms competing methods with an improvement of up to 23%. We used Flipper to generate motif predictions on sets of genes clustered based on SAGE and microarray expression data and discovered 4 novel motifs associated with ribosomal protein promoters: M546, M313, M439 and M540.

One aim of this study is to develop Flipper as a general bioinformatics resource, an algorithm for detecting conserved motifs that is unconstrained by sequence alignment. In designing the Flipper algorithm, emphasis was placed on enhancing the robustness of the motif search. Motifs are represented using position weight matrices (PWMs) instead of consensus sequences, with allowance for gaps within motifs. Importantly, Flipper does not rely on sequence alignment to assess the conservation of sampled sites. In a survey of TF binding sites directly assayed from ChIP experiments, we observed that sequence alignability generally poorly correlates with binding site affinity, with many strong candidates for TF binding failing to align (Fig 4.2). Instead, conservation is reflected in the organization of

TF binding sites across species, with species of similar binding affinity recurring in orthologous promoters. With this measure of binding site conservation, Flipper samples from orthologous promoters in groups, with one site from each related promoter. Groups containing sites that score consistently well using the motif PWM are admitted into the overall site configuration with high probability.

Flipper requires sampled groups to contain one site from an orthologous promoter in each species, regardless of the evolutionary distance between compared species. Given there may be significant divergence in binding sites between distantly related species, this requirement may be overly strict, resulting in some bona fide sites being undetected. In practice however, most motif finding algorithms suffer from poor specificity and the increased stringency in site sampling may yield a net benefit for Flipper in rejecting more false positives. Furthermore, DNA binding domains of TFs evolve much slower than their respective binding sites, implying that the recognized motif may not change significantly even between distantly related species. Flipper can be extended to weight contributions of sites within a group by inter-species evolutionary distance similar to other methods, but this would require approximating rates of evolutionary divergence, as well as models of species phylogeny.

An advantage of comparative genomics is increased signal-to-noise ratio for motif detection, due to the availability of additional sequence information from orthologs. While over-representation based methods, such as AlignACE and MEME, do not have the functionality to handle orthology between the input sequences, motif discovery with these methods can nevertheless benefit from the increased signal-to-noise ratio by considering orthologous promoters as independent sequences. It should thus be noted that in comparing Flipper against non-phylogenetic algorithms (i.e. AlignACE and MEME) on synthetic sequence datasets, we had tried to 'level the playing field' by running the compared algorithms on a pooled set of sequences from all species. Flipper maintains an advantage over these algorithms, presumably because it explicitly considers input sequences to be related by orthology.

From our validation studies on synthetic data, we also confirmed that alignment-based phylogenetic algorithms are susceptible to decreased site alignability, with performance being severely affected when binding site alignability is lower than 70%. This implies that alignment based methods may be limited in predicting binding sites in larger genomes, such as worm and fly, especially since *in vivo* binding sites in these organisms, as assayed by ChIP, are on average only 40 to 50% alignable.

We demonstrated that Flipper also maintains an advantage over alignment-based phylogenetic algorithms on biological sequence sets, derived from ChIP and expression-based experiments. Interestingly, of the three studied species, Flipper had the biggest advantage on yeast and fly ChIP datasets, where binding site strength was markedly less correlated with site alignability than the strength of orthologous binding sites (Fig 4.2).

From our experience, we observed that while Flipper can accommodate gaps in the final motif, PhyloGibbs and PhyME perform better at recovering ungapped motifs. For example, with the worm datasets, where PhyloGibbs and PhyME recovered the ELT-2 motif easily but failed to recover the MEC-3 and HSF-1 motifs which were gapped. Since DNA binding domains of most TFs recognize short DNA sequences and rely on dimerization to increase specificity, it is likely that a number of binding motifs *in vivo* will contain gaps. This can pose a challenge to alignment programs, and by association, alignment-based motif finding algorithms, especially if TF dimerization accommodates variable gap length, which may not be conserved across species.

We also observed that PhyloGibbs and PhyME perform better on datasets where motif enrichment is strong. For example, we noted that the performance of PhyloGibbs deteriorates more rapidly than Flipper for datasets with decreasing motif enrichment, as measured with a Mann-Whitney U test (Fig 4.5). This trend is replicated in worm and fly datasets as well – PhyloGibbs failed to recover the known motif for 3 out of 6 fly datasets where the motif enrichment was fair/poor ($0.1 < p < 0.5$), whereas PhyME failed to recover 4 out 6 of these motifs. One possibility could be datasets with weaker motif enrichment may contain binding sites that are less alignable. PhyloGibbs, in particular, is more sensitive than PhyME to weaker datasets. Our experience with PhyloGibbs is that it has a high threshold for false positive rejection, which could compromise its performance on datasets where true binding sites are too weak or occur too infrequently to meet this threshold.

In summary, we have shown that a substantial number of TF binding sites *in vivo* do not align well in multiple alignment and as a result, can be missed by alignment-based motif discovery tools. Flipper does not use multiple alignment and instead assesses motif conservation by constraining sampled sites to be distributed across orthologous promoters. Flipper joins a number of other methods in adopting an alignment independent approach[92], and we are hopeful that it can be a robust tool for motif discovery in mammalian genomes.

Chapter 5:
Motifs predicted by Flipper are necessary for proper regulation of *C.elegans* ribosomal protein expression

5.1 ABSTRACT

Ribosomal proteins (RPs) are among the most co-expressed sets of genes in the genome, but little work has done to characterize their regulation. We previously reported six novel motifs associated with RP promoters in *C.elegans* and in this study, use promoter::mCherry transcriptional reporters to determine if these motifs are functional *in vivo*. We tested four motifs, M546, M540, M313 and M439 in a total of 23 RP promoters for function and discover that they are each functional in a subset of promoters. To determine if our motifs are functionally conserved, we performed a promoter transplant experiment, where we use RP promoter orthologs from *C.briggsae* to drive mCherry expression in *C.elegans* and ask if the native mCherry expression would be affected if the *C.briggsae* motif was mutated. We tested M546 and M313 in the orthologs of *rpl-2, rps-7, rpl-14* and *rpl-19* and verified that the motifs are necessary for *C.briggsae* promoter function. In a genome wide survey of ribosomal motif sites, we observed a substantial number of non-ribosomal promoters containing high confidence M546 and M540 sites. We tested the M546 motif in the *mcm-7* promoter and show that it is also necessary for promoter function. In summary, our promoter::mCherry reporter assay validates motif predictions made by Flipper and M546, our most significant motif, is functional in both RP and non-ribosomal promoters, suggesting it may be recognized by a more general regulator of growth and development processes.

5.2 INTRODUCTION

The ribosome is a highly specialized functional unit for translation and is composed of both protein and rRNA components. Since ribosomal proteins exist in a complex with precise stoichiometric requirements, it is not unexpected that ribosomal proteins are the among some of the most tightly coexpressed genes, consistent in yeast, worms and human. Few studies on ribosomal protein regulation have been performed, because ribosomal proteins are commonly regarded as housekeeping genes and are thus uninteresting to disease biology. Supporting this view is the observation that ribosomal protein knockouts generally result in a severe embryonic lethal phenotype. Interestingly though, heterozygotes for the knockout allele may not display any phenotype, since feedback mechanisms can allow transcription from the functional allele to compensate for the other null allele[41]. This suggests that regulatory mutations that affect the levels of ribosomal protein production may yield more relevant phenotypes than complete loss-of-function mutations in coding sequence.

In *D.melanogaster*, mutations in ribosomal protein loci are associated with a *minute* phenotype – flies have prolonged development, thin bristles, poor viability and fertility[44,45]. To show that differences in expression levels can impact the severity of the phenotype, Saeboe-Larssen *et al* mobilized a *p* element insertion in the 5' UTR of the *rps3* gene and generated two alleles with 15% and 60% reductions in mRNA production respectively[46]. Severity of the minute phenotype correlated with reduced *rps3* levels in a dose dependent manner, suggesting that regulatory mutations affecting ribosomal protein expression may contribute strongly to the *minute* phenotype. As a more direct test of causation, ribosomal protein levels can be modulated experimentally in Zebrafish using morpholino knockdown[42,93]. While ribosomal protein knockdown causes developmental defects in general, there is considerable tissue specific variation between the effect of individual ribosomal protein knockdowns, indicating that different tissue types have specific requirements for ribosomal protein expression. Tissue specific regulatory mechanisms may have evolved to meet these requirements. *Rps19* is particularly relevant to studies of disease, since defects in this gene are associated with Diamond-Blackfan Anemia (DBA) in humans. DBA is characterized mainly by defects in haematogenesis especially in the erthyroid lineage, that manifest in early infancy. Using a zebrafish model, Danilova *et al* observed that knockdown of *rps19* caused stress on the translation process and in turn, caused upregulation of p53 and the apoptotic response[42]. The authors claim that dysregulation of p53 pathways may result in DBA and were able to alleviate the DBA phenotype by suppressing p53. This study offers insight into downstream effects of depleting ribosomal protein levels and fits into an overall picture where haploinsufficiency of ribosomal protein genes can result in cancer[94], presumably due to dysregulation of p53 pathways in response to translation stress.

Since ribosomal proteins are so highly coexpressed and strongly conserved across species, we attempt to use Flipper, an algorithm we developed for discovering conserved over-represented sequence motifs, to identify *cis*-regulatory elements in the promoters of ribosomal protein genes. We use the promoter sequences of 76 ribosomal proteins in *C.elegans* for our analysis and use related nematodes *C.brenneri*, *C.remanei* and *C.briggsae* for comparison. We validate our predictions *in vivo* using promoter::mCherry transcriptional reporters and further verify that these *cis*-regulatory elements are functionally conserved by testing motif occurrences in *C.briggsae* promoters within a *C.elegans* regulatory context.

5.3 METHODS

5.3.1 Promoter::mCherry transcriptional reporter assay

We used promoter driven fluorescent reporter expression to interrogate the function of our predicted motifs. We used three fluorescent reporter systems: promoter::GFP, promoter::GFP-NLS and promoter::H1::mCherry to assay promoter function. Constructs for the promoter::GFP and promoter::GFP-NLS reporter were adapted from the pPD95.77 and pPD121.83 vectors from the Fire Lab vector kit respectively, whereas constructs for the promoter::H1::mCherry reporter were adapted from the pJIM20 construct (gift from the Waterston lab). Promoters were defined to be sequence 2000 bp upstream of translation start, or up to the nearest gene boundary. Genomic DNA was isolated from the canonical WT *C.elegans* strain, WRN2 and promoter fragments were amplified from the genomic DNA template using Phusion, a high-fidelity PCR polymerase system (New England Biolabs). Amplified promoter fragments were cloned into pJIM20 in a sticky-blunt ligation (labeled primers with T4-PNK (New England Biolabs), digested with *XbaI* and ligated using T4-DNA Ligase (New England Biolabs) overnight at 16°C.) Constructs adapted from the Fire Lab vector kit (i.e. from pPD95.77 and pPD121.83) were further modified to contain a mini-gene cassette for the *unc-119* gene, cloned from the MM051 plasmid (gift from the Maduro lab). The *unc-119* cassette was included as a selection marker, intended for rescue selection in a *Δunc-119* background.

Mutant promoters were generated using recombinant fusion PCR[95]. Briefly, motifs were scrambled by shuffling base pair positions to preserve overall base pair composition. At the same time, bases are shuffled so that the mutated region now contains a restriction enzyme recognition site not previously in the promoter. This enables easy confirmation by restriction enzyme digest that the motif site in the promoter has indeed been mutated.

5.3.2 Generation of transgenic *C.elegans* by microparticle bombardment

We adapted a protocol for microparticle bombardment from the Praitis lab[16]. We used microspherical gold beads 1.0 micron in diameter (InBIO Gold, Australia) as a DNA carrier. Prior to DNA loading, gold beads are vigorously vortexed and sonicated to disperse any clumps, washed repeatedly with 70% Ethanol and sterile water before being suspended in 50% Glycerol. Approximately 10 μg of DNA is added to 70 μl of gold suspension. The DNA is precipitated onto the gold beads by adding 100 μl of 2.5M $CaCl_2$ and 40 μl of 0.1M spermidine. The mix is vortexed regularly during each addition to avoid clumping of the gold beads. After adding spermidine and vortexing, the gold beads are spun down and resuspended in two wash steps: 300 μl of 70% Ethanol, followed by 300μl of 100% Ethanol. A final suspension of gold beads in 100% Ethanol is used for the actual bombardment. We used a PDS1000 bombardment chamber (BioRAD) for microparticle bombardment. DP38(*Δunc-119, ed 3)* worms were grown to high densities in liquid culture, settled and plated onto Agar plates for bombardment.

After bombardment, worms are set aside for two weeks before being inspected for successful transformants. Transformants were selected based on their ability to enter dauer and the recovery of motility. Transformed worms were individually transferred onto plates for subsequent genomic DNA extraction and PCR.

To determine if the transformed worms carry the appropriate reporter constructs, we extracted genomic DNA from individual transformed worms and perform PCR using primers specific to the vector backbone that should not yield product in the absence of an integrated construct.

5.4 RESULTS

5.4.1 Sequence alignability does not imply promoter element function

In comparative genomics studies, multiple alignments are typically used to determine regions of high sequence conservation in promoters. Conserved regions are assumed to be enriched in *cis*-regulatory elements, which may be recognition sites for transcription factor binding or structural elements that affect chromatin organization[35]. While there is a strong relationship between element conservation and biological function, the use of sequence alignment to assess conservation is less clear. Results from the ENCODE pilot study[8] showed that whereas 5% of studied regions were 'evolutionarily constrained' with high confidence, only 50% of the experimentally determined regulatory factor binding regions were 'constrained' by this measure, implying that multiple alignment cannot detect all functional binding sites. Furthermore, in a computational survey of multiple alignments of ChIPed regions, we previously showed that sequence alignability poorly correlates with the strength of TF binding sites located within these regions. To determine the extent to which sequence alignability can inform the search for functional sequence motifs, we decided to assay the function of highly alignable sequence elements using an *in vivo* transcriptional reporter assay. Briefly, promoter::mCherry reporter constructs were used to generate transgenic *C.elegans* with microparticle bombardment. Transgenic worms were imaged for mCherry expression and expression patterns were compared between constructs containing the native WT promoter sequence, against constructs with the candidate sequence element mutated. We focused on ribosomal protein promoters since they are highly transcribed from and widely expressed, allowing easy detection of changes caused by mutated sequence elements. Using diAlign[59], we performed multiple alignments of ribosomal protein promoters with their orthologs from *C.briggsae, C.brenneri* and *C.remanei*. Sequence alignability for each base was assessed using a sum-of-pairs (SP) statistic. A moving average using a 50 bp window was applied to "smooth" the SP score values, to determine regions of high sequence alignability. For the *rpl-2* promoter, we tested three 19 bp regions: (-557:-538), (-313:-294) and (-124:-105) bp upstream of translation start (Fig. 5.1). In addition, we tested a strong M546 site previously identified by a genome wide scan for high confidence M546 sites, 377 bp upstream of translation start of *rpl-2*.

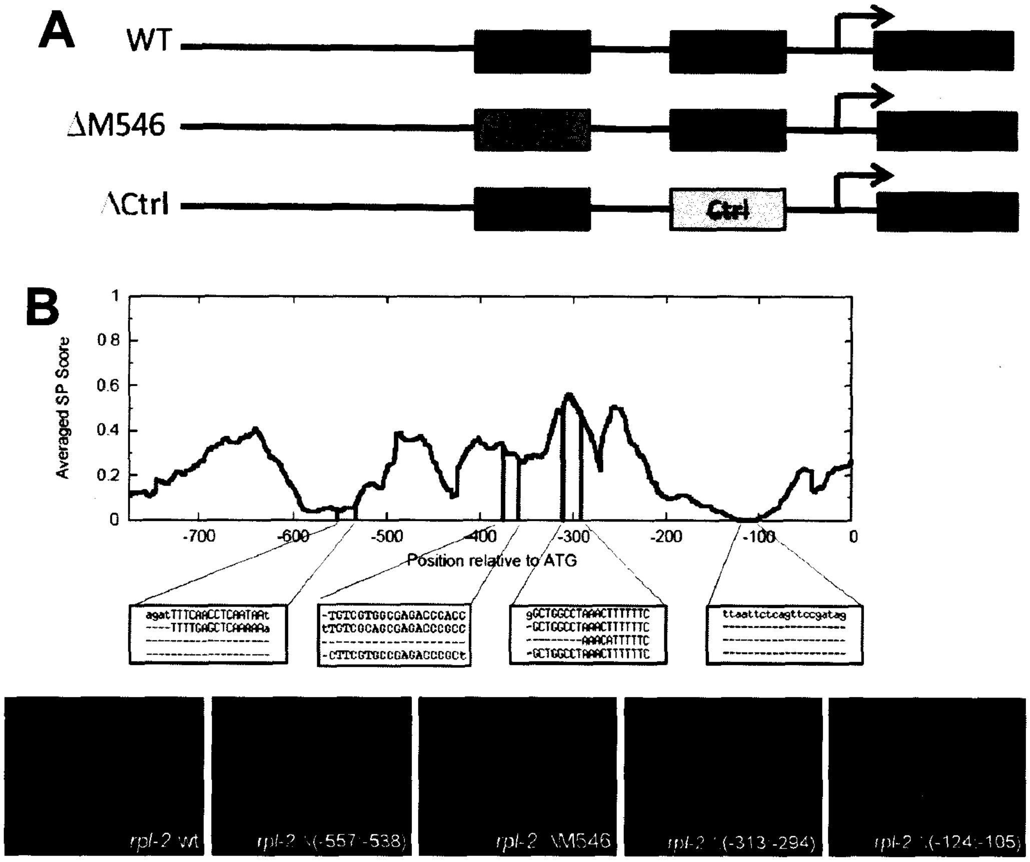


Figure 5.1:

A) Outline of promoter::mCherry transcriptional reporter assay, comparing native WT promoters constructs against versions with candidate motif sites and control sequence elements mutated. B) Averaged SP score is plotted across the length of the *rpl-2*promoter. Mutated sequences, (-557:538), (-313:294), M546, (-124:-105), are indicated in red and green with examples of multiple alignment stated below. Representative pictures of transgenic lines carrying WT vs. mutant promoters are shown.

Candidate sites for mutation were chosen based on sequence alignability and proximity to ATG. We hypothesized that sequences closer to translation start may have a higher probability of being functional despite low sequence conservation. Site (-124:-105) was closest to ATG but did not align well. Site (-313:-294) was positioned proximal to ATG with similar distance as the M546 site and was the most alignable sequence element in the promoter. We expected site (-557:-538) to be a negative control when mutated, since it was neither close to ATG or strongly conserved.

We observed that out of the four elements tested, M546 was the only element that affected *rpl-2* promoter function when mutated. The WT *rpl-2* promoter drove mCherry expression in a wide range of cell types as expected, whereas mutating the M546 site in the promoter diminished overall mCherry expression, to the extent of complete abrogation in most cell types in the midsection of the worm. As expected, mutating site (-557:-538) had no effect on mCherry expression, as did mutating site (-124:-105), suggesting that proximity to ATG does not imply function, especially when the sequence elements are poorly conserved. Strikingly, site (-313:-294) which was the most highly alignable sequence element in the promoter, did not affect mCherry expression when mutated, suggesting that sequence alignability cannot be used as a sole predictor of promoter element function.

5.4.2 Motifs predicted by Flipper are necessary for ribosomal protein promoter function

We had previously developed a Gibbs sampling based, phylogenetic motif discovery algorithm, Flipper, which did not rely on multiple alignment to assess sequence element conservation. Flipper reported 22 motifs on promoters of the ribosomal protein (RP) cluster. We decided to further investigate the six most over-represented motifs by MAP score and chose 4 motifs: M546, M313, M540 and M439 for experimental validation. To determine if these motifs were functional *in vivo*, we generated transgenic *C.elegans* lines carrying promoter::mCherry transcriptional reporter constructs – one containing mCherry driven by the intact WT copy of the promoter, the other with a candidate motif site mutated by scrambling. Scrambling motifs allows us to preserve the overall GC content of the promoter, as well as maintain the positioning of other sequences relative to ATG.

In total, we tested M546, M313, M540 and M439 for function in 23 ribosomal protein promoters and found that M546 was the strongest motif of the four, being necessary for promoter function in 8 out of 11 tested cases. In comparison, M313 was necessary in 4 out of 9 cases, M540 in 3 out of 7 cases and M439 in 1 out of 3 cases. Our experimental results are summarized in Fig 5.2 with representative pictures of transgenic lines in Fig 5.3 . We had scanned all promoters in the genome using PWMs of these motifs and picked these promoters for testing, since they all contained motif sites scoring much higher than expected by chance given a symmetric null distribution.

In general, we observed mCherry expression in a wide variety of tissues that was reproducible across all WT RP promoters tested, consistent with their tight coexpression. Interestingly, RP promoter driven expression is not constant over all developmental stages: we observed stronger mCherry expression in a wider variety of tissue types in L2 and L3, which progressively declined as the worms aged and approached adulthood. This suggests that RP promoter activity may be temporally regulated and RPs may not be as ubiquitously expressed as previously thought.

Different motifs had varying effects on mCherry expression when mutated – whereas mutating M546 consistently abolished mCherry expression to the point of autofluorescence, mutating M540 and M313 had a weaker effect. Transgenic lines containing promoters lacking these motifs occasionally had reduced mCherry expression, but often retained mCherry expression in select cell types. For example, mutating M313 in the *rpl-19* and *rpl-34* promoters abolished mCherry expression in almost all cell types, but residual mCherry expression remained in cells in the head and tail region of the worm. From the genomic distribution of M313, we observe that M313 is highly specific to RP promoters, unlike M546 and M540. A possible hypothesis is that M313 is recognized by a RP specific factor that helps recruit more general regulators that in turn recognize M546 and M540. M313 mediated recruitment may occur in selected cell types, so cells relying on alternative *cis*-regulatory elements may retain mCherry expression despite M313 mutation.

We were unable to find a canonical motif that was by itself consistently necessary for function in every promoter tested, even though M546 comes close to achieving this status. This strongly suggests that RP promoter expression results from the combinatorial action of multiple *cis*-regulatory elements and the trans-acting factors that bind them. The higher order *cis*-regulatory logic encodes redundancy in RP promoter regulation, which is probably necessary since ribosomal proteins are essential to proliferation and biosynthesis in general. Furthermore, we had used a strict significance cutoff in determining motif sites for testing. Promoters that tested negative for M546 function may contain functional sites for other motifs, i.e. M313 and M540, which were not tested because they did not score higher than the required thresholds. Given the combinatorial nature of *cis*-regulatory logic, our motif elements may function in regulatory modules and may not have an effect on promoter function when mutated alone.

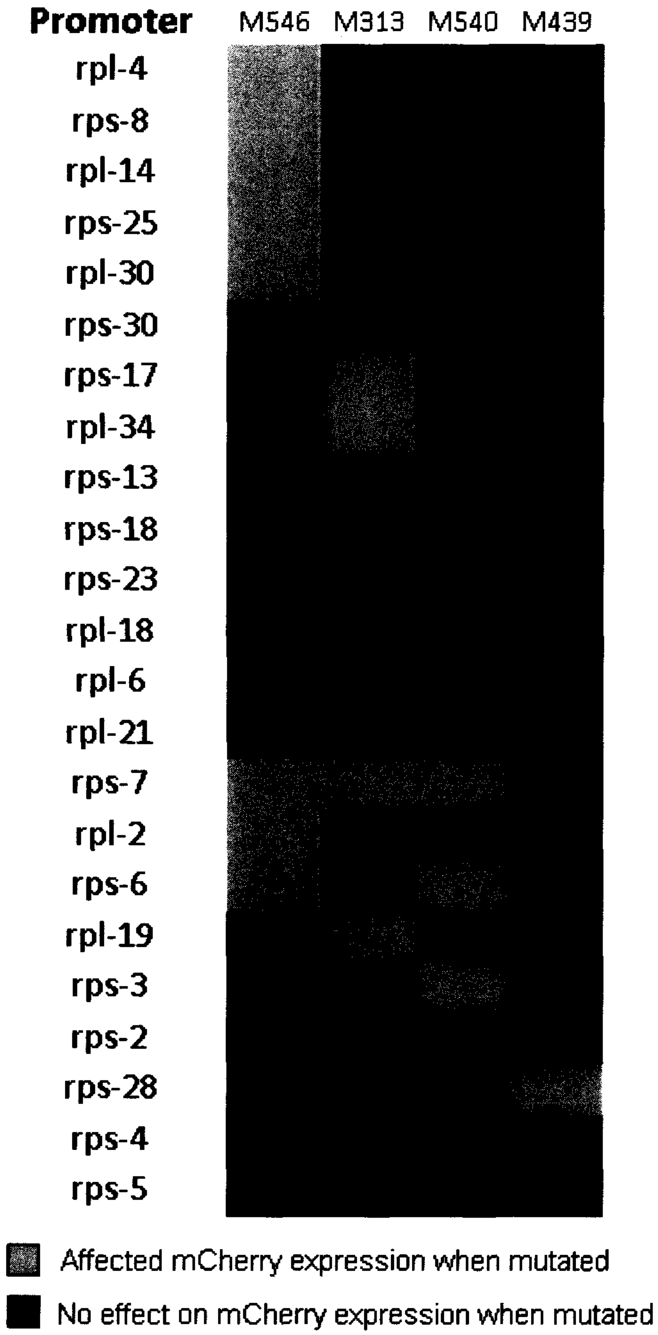


Figure 5.2:

M546, M313, M540 and M439 are functional in a subset of ribosomal protein promoters. M546 is functional in 8 out of 11 tested promoters, M313 is functional in 4 out of 9 tested promoters, M540 is functional in 3 out of 7 tested promoters and M439 is functional in 1 of 2 tested promoters.

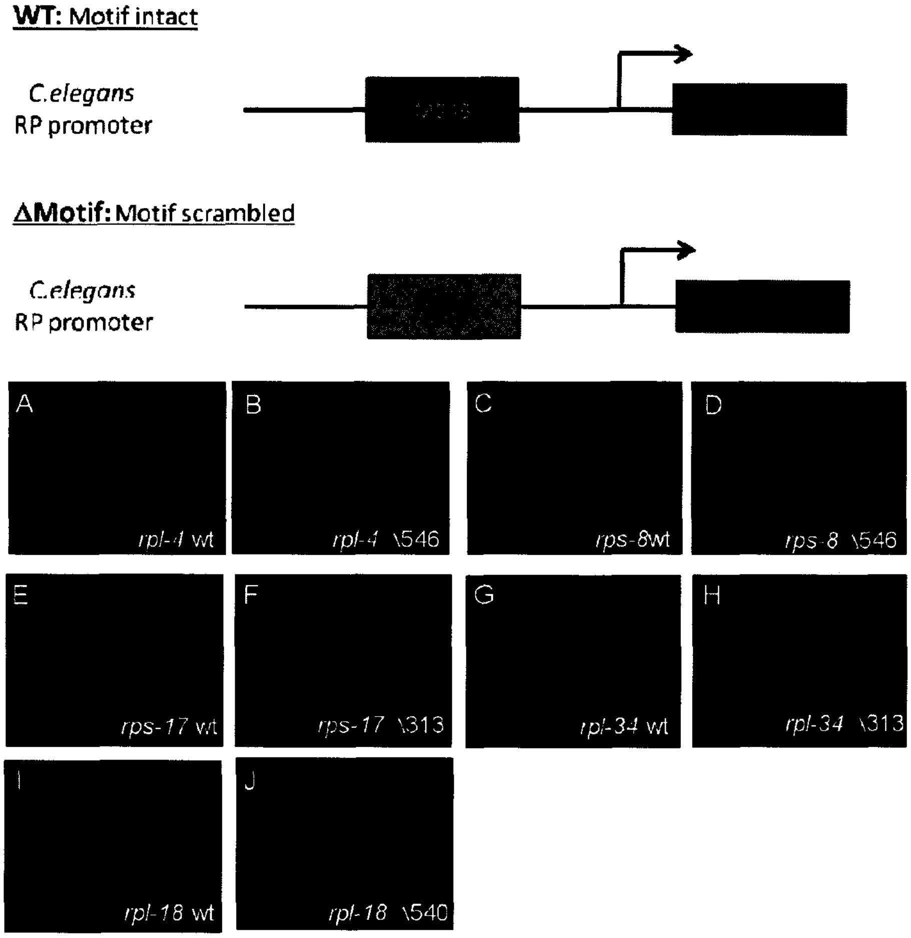


Figure 5.3:
mCherry expression of wildtype (WT) ribosomal regulatory regions show consistent temporal and spatial coexpression and expression is abrogated when predicted *cis*-regulatory elements are mutated by scrambling. A) mCherry expression using native *rpl-4* promoter. B) mCherry expression with M546 site in *rpl-4* mutated. C) mCherry expression using native *rps-8* promoter. D) mCherry expression with M546 site in *rps-8* mutated.E) mCherry expression using native *rps-17* promoter. F) mCherry expression with M543 site in *rps-17* mutated. G) mCherry expression using native *rpl-34* promoter. H) mCherry expression with M543 site in *rpl-34* mutated. I) mCherry expression using native *rpl-18* promoter. J) mCherry expression with M540 site in *rpl-18* mutated.

5.4.3 M546, M313 are necessary for *C.briggsae* ribosomal promoter function in *C.elegans*

Since M546, M313 and M540 were identified on the basis of their functional conservation, we asked if *C.briggsae* orthologs of ribosomal protein promoters could drive mCherry expression in *C.elegans* and if corresponding instances of M546, M313 and M540 in these promoters were necessary for their expression. We created transgenic *C.elegans* lines carrying promoter::mCherry constructs for promoters of *cbg05588*, *cbg21343*, *cbg03771* and *cbg12479* - *C.briggsae* orthologs of *rpl-2*, *rps-7*, *rpl-14* and *rpl-19* genes in *C.elegans.* Even though the average sequence identity between *C.elegans* and *C.briggsae* promoters was only 30%, all of the tested *C.briggsae* promoters were capable of driving mCherry expression in a pattern similar to their *C.elegans* orthologs, suggesting that *C.elegans* and *C.briggsae* share a common mechanism for regulating RP promoter expression.

Similar to the RP promoters in *C.elegans*, we observed severely reduced mCherry expression when the M546 elements in the *CBG05588*, *CBG03771* and *CBG21343* promoters were mutated. Likewise, mutating the M313 element in the *CBG12479* promoter had a similar effect,recapitulating the expression pattern in *rpl-19* mutant promoters where mCherry expression was abolished in most tissues with the exception of some cells in the head and tail. Examples of these expression patterns can be found in Fig 5.4. These results were experimental validation that our motifs were indeed functionally conserved across species. In all cases, mutating our motif elements in *C.briggsae* RP orthologs recapitulated the corresponding effect seen in *C.elegans* mutant RP promoters, suggesting that not only are our motifs recognized by conserved *trans*-acting regulators, the functional interaction between *cis*-regulatory elements to direct expression is also conserved.

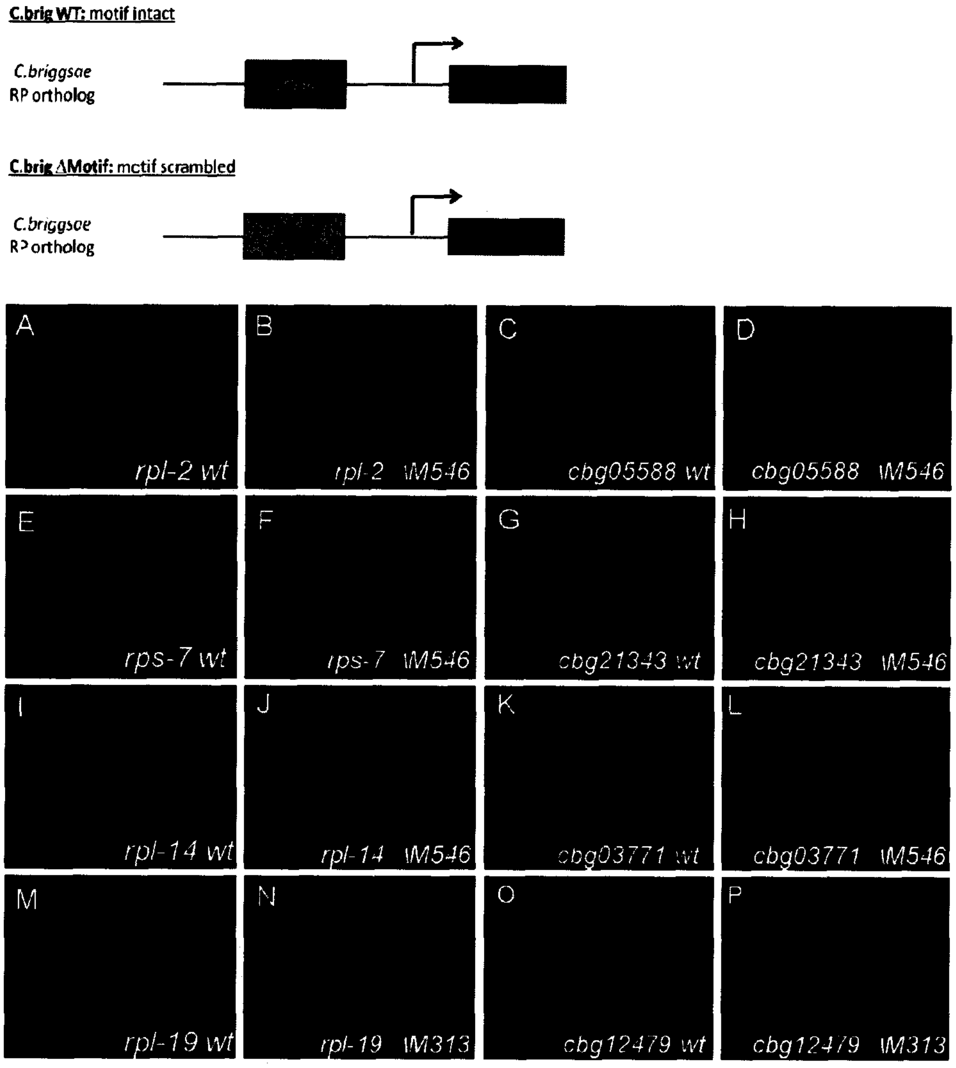


Figure 5.4:
Orthologous ribosomal protein promoters from *C.briggsae* can recapitulate mCherry expression patterns in *C.elegans*, but depend on M546 and M313 motifs for function. A) and B) Expression using native *rpl-2* promoter vs. with M546 site mutated. C) and D) Expression using native *CBG5588* promoter, ortholog of *rpl-2* in *C.briggsae.* vs with M546 site mutated. E) and F) Expression from native *rps-7* promoter vs. with M546 site mutated. G) and H) Expression from native *CBG21343* promoter, ortholog of *rps-7* in *C.briggsae* vs. with M546 site mutated. I) and J) Expression from native *rpl-14* promoter vs. with M546 site mutated. K) and L) Expression from native *CBG03771* promoter, ortholog of *rpl-14* in *C.briggsae* vs. with M546 site mutated. M) and N) Expression from native *rpl-19* promoter vs. with M313 site mutated. O) and P) Expression using native *CBG12479* promoter, ortholog of *rpl-19* in *C.briggsae* vs. with M313 site mutated.

5.4.4 M546 is necessary for the function of non-ribosomal protein promoters

We previously performed a genome wide scan of promoters, asking which promoters contain high confidence instances of our RP motifs and if our motifs were specific to the RP promoter cluster only. We observed that whereas M313 was highly specific to RP promoters, M546 and M540 were found in a large number of non-ribosomal protein promoters. More interestingly, we observed that M546 and M540 elements tend to co-occur in pairs in these promoters.

Since M546 was functional in 8 out of 11 RP promoters tested *in vivo*, we asked if it could likewise be functional in non-ribosomal promoters that were unrelated in their expression profile to RP genes. To test this, we looked at the *mcm-7* gene, which codes for a protein that binds origins of replication and is involved in licensing specific genomic loci for initiation of DNA replication[96]. Previous studies have shown that transgenic reporter lines with the native *mcm-7* promoter express GFP mostly in the embryos and our results confirm this, where we observe strong mCherry expression during embryonic development. M546 and M540 sites co-occur in the *mcm-7* promoter, whereas M313 is absent.

In constructs where a copy of M546 is mutated in the *mcm-7* promoter, we observe severely reduced mCherry expression, often to the point of auto-fluorescence, implying that M546 is also necessary for regulating *mcm-7* expression. Despite the difference in expression profile, both *mcm-7* and RP genes are involved in growth and proliferative responses, leading us to propose that M546 may be recognized by a more general set of transcription factors that may regulate genes involved in growth and development. Since M313 is specific to RP promoters, it may bind a RP-specific TF that recruits the regulators binding M546 and M540. Another motif specific to *mcm-7* and other mcm genes may serve a similar function.

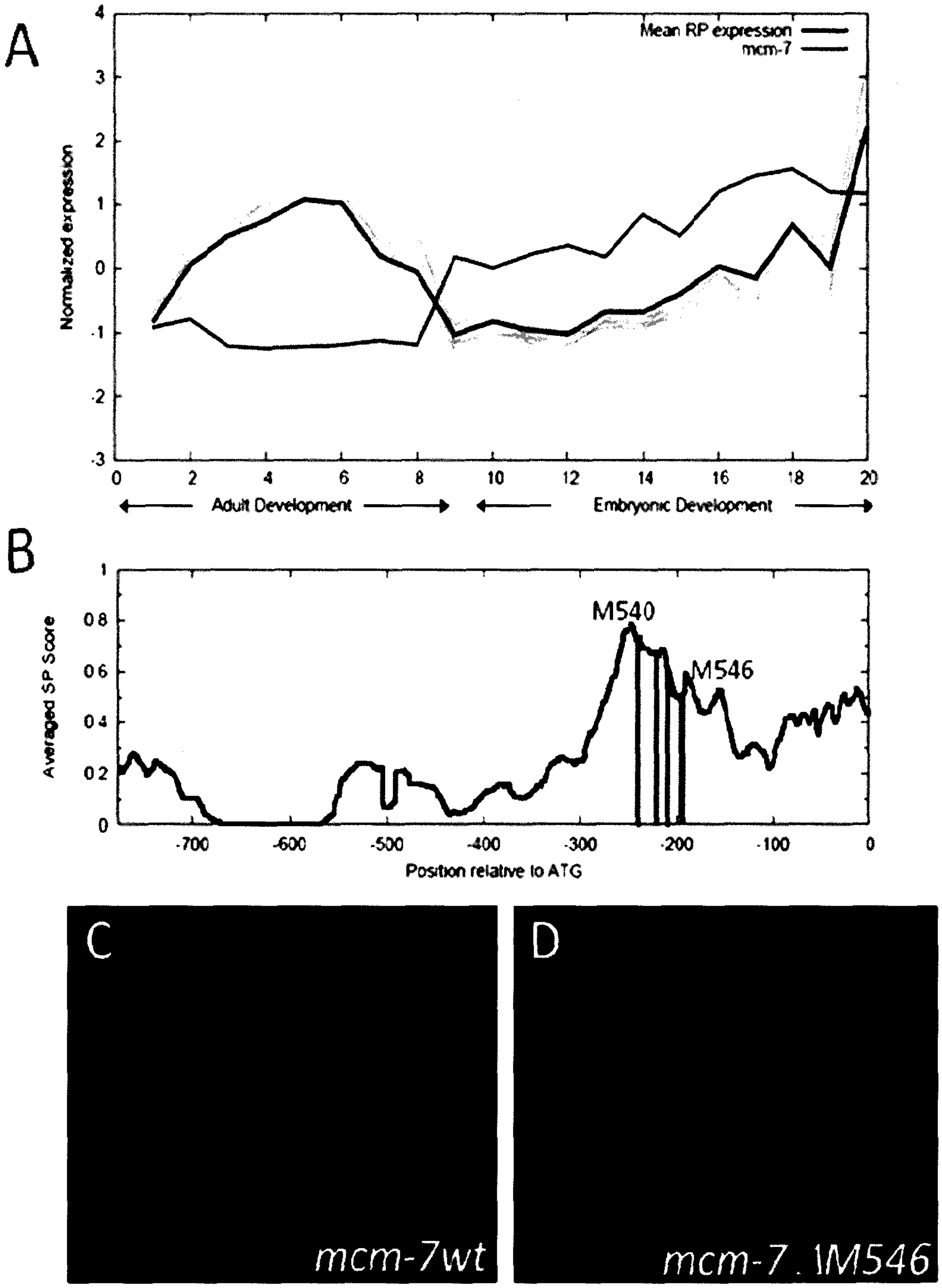


Figure 5.5:
A) mcm-7 expression does not fit the mean expression profile of ribosomal proteins and occurs predominantly during embryonic development. B) M546 and M540 co-occur in the *mcm-7* promoter . C) and D) mCherry expression in embryos is abolished when the M546 site in *mcm-7* is mutated.

5.4.5 RP motifs co-occur in stereotypic patterns in promoters

To investigate if RP promoters are regulated by some mechanism of higher order *cis*-regulatory logic, we asked if they co-occur in stereotypic patterns in RP promoters. We used the four tested motifs, i.e. M546, M313, M540, M439, for this analysis. We scanned all RP promoters using the motif PWM, normalized and transformed the resulting scores using a logistic function so that the maximum and minimum scores are 1 and 0 and the median score is 0.5. We represent the presence or absence of a motif in each promoter using a score vector, where each vector entry corresponds to the normalized score of the highest scoring site for each motif. Promoters were grouped by motif co-occurrence by hierarchical clustering of their associated score vectors.

The results of the hierarchical clustering are presented in a heat map in Figure 5.6. Our analysis shows that there does not appear to be a universal pattern of motif co-occurrence that is shared by all RP promoters. Instead, there appears to be five distinct subpatterns of motif co-occurrence in RP promoters, suggesting that different modes of combinatorial regulation may act to produce a common expression profile. A subset of RP promoters do not contain strong sites for any motif, suggesting other RP motifs identified by Flipper may regulate them. Multiple *cis*-regulatory modules may exist to encode redundancy in the regulation of RP expression, which is expected since RPs are essential for protein biosynthesis and general cellular function. The organization of motifs into higher order clusters may account for observations in our reporter assay where no single motif is consistently required for function in all promoters. For example: the M540 element may not be functional in the *rpl-2* promoter because a strongly conserved M439 site is present closer to ATG in that promoter. Similarly, rps-13 contains a strong, conserved M439 site that may be functional instead of M313. We performed the same clustering analysis on groups of random promoters and did not see any patterns of motif cooccurence. Motif presence in our heatmap is associated with columns that are four species wide, indicating that patterns of motif organization are conserved in orthologs. Our clustering results generate testable rules regarding the organization of motif elements in putative regulatory modules, which can be easily validated using our transcriptional reporter setup.

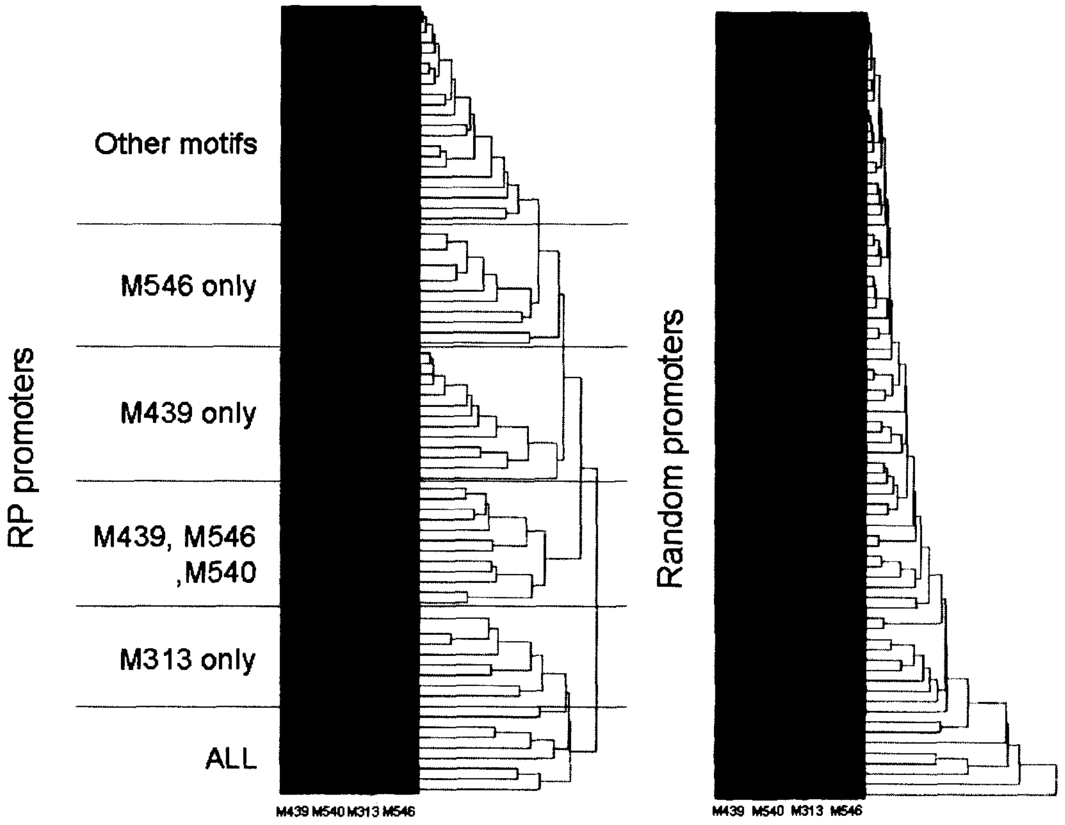
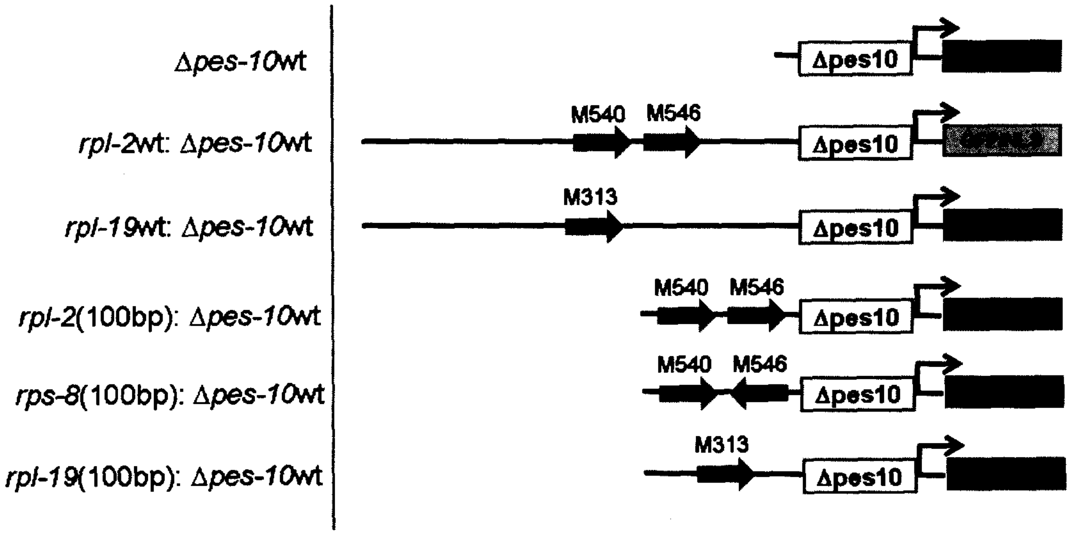


Figure 5.6:
Heat map depicting patterns of motif presence / absence in all RP promoters. An equal number of random promoters was selected from all promoters in the *C.elegans* genome. 10 sets of random promoters were chosen, the clustering result of one set is shown above for reference.

To determine if the motif combinations indicated by our clustering results are functional, we used a Δ*pes-10* minimal promoter to assay sequence fragments containing motif combinations for enhancer activity. The objective of the experiment was to identify particular motif combinations capable of driving mCherry expression in lieu of the complete promoter sequence. We tested cooccuring pairs of M546-M540 motif sites in the *rpl-2* and *rps*-8 promoters, as well as a M313 site in the *rpl-19* promoter. The results are shown in Fig. 5.7. We used an empty vector construct as a negative control, i.e. Δ*pes*-10 minimal promoter only. We observed considerable background expression, especially in the ventral nerve cord, head and tail sections, indicating there may be cryptic sequence elements in the vector backbone capable of autoactivating mCherry expression. For positive controls, we tested constructs where the complete *rpl-2* and *rpl-19* promoter sequences were appended to a GFP-NLS and mCherry CDS respectively. As expected, the observed expression profile for these constructs matched previously observed RP promoter driven patterns. None of the tested motifs (and combinations) were able to recapitulate mCherry expression from the complete promoter – *rpl-2* and *rps-8* constructs containing M546 and M540 motif pairs were capable of activating promoter expression, but did not produce significant differences from the background expression pattern observed with the empty vector control. Similarly, the construct containing only the M313 site did not yield any reporter expression.

These results clearly indicate that the regulatory circuit for RP expression requires more motif interactions than the combinations tested in our experiment. Referring to Fig. 5.6, we observe that M546-M540 motif pairs tend to co-occur with M439 motifs, suggesting that M439 may be the missing motif needed to complete the regulatory circuit. Whereas our clustering results show that the *rpl-19* promoter contains only M313, our experimental results indicate that M313 is clearly incapable of reproducing the expression patterns associated with the complete promoter sequence. It is likely that other motifs predicted by Flipper (aside from M546, M540 and M439) with lower MAP score interact with M313 to regulate *rpl-19* expression. We intend to include these motifs in our clustering for a more complete analysis of combinatorial regulation.



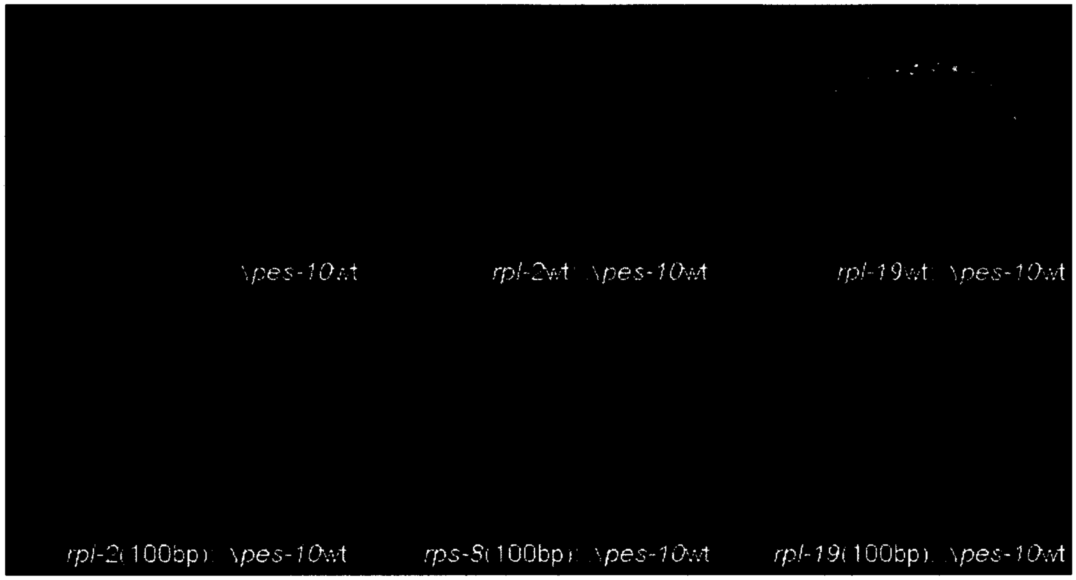


Figure 5.7:
RP motifs cannot recapitulate WT promoter expression by themselves. Δ*pes-10* minimal promoter constructs were used in an enhancer test assay. Negative control (without promoter), positive control (intact WT promoter for *rpl-2* and *rpl-19*) constructs were tested for comparison. 100 bp of sequence surrounding M546-M540 motif pairs in the *rpl-2* and *rps-8* promoters, M313 motif in the *rpl-19* promoter were tested for enhancer activity.

5.5 DISCUSSION

As an application of Flipper to predict novel sequence motifs, we used Flipper to find novel regulatory motifs in ribosomal protein (RP) promoters. The RPs in *C. elegans* are tightly coexpressed across development, concomitant with lineage specific proliferation and translational requirements. While often considered to be housekeeping genes, we have shown that RPs are tightly regulated temporally throughout development and in a tissue specific manner in worms (see Figure 5.7), mice, and in human tumors. The right panel in Figure 5.8 shows the expression of human ribosomal proteins across prostate tissue samples from Singh *et al* [97]. The left panel shows the distribution of all correlation coefficients between pairs of these genes, and is strongly shifted toward coexpression (+1), while random genes selected from this dataset exhibit no net correlation (in green).

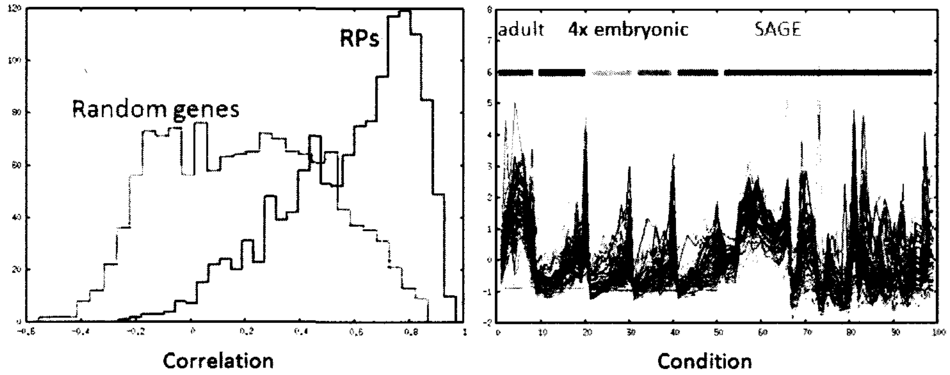


Figure 5.8:
Ribosomal proteins (RPs) in *C.elegans* are more tightly coexpressed than random genes. The degree of coexpression is consistently high across developmental stages (adult and embryonic) and in various tissue types (SAGE).

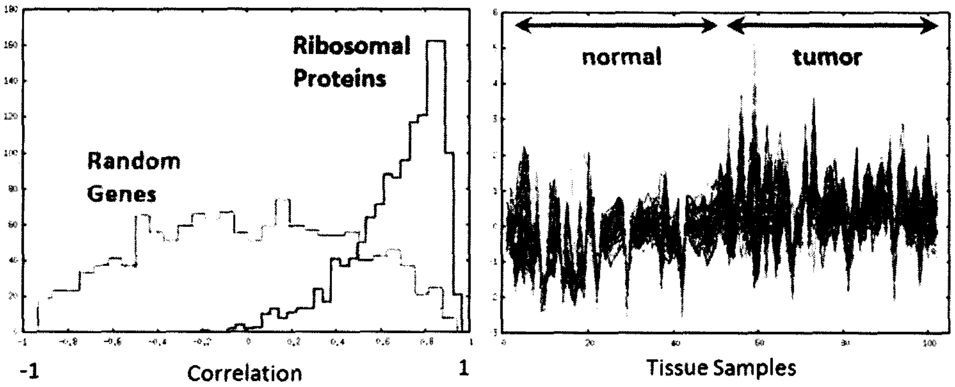


Figure 5.9:
Ribosomal proteins (RPs) in human are significantly more correlated in expression than random genes. Almost all RP genes follow a common expression profile that is consistent across normal and tumor prostate tissue samples.

Because transcription of ribosomal protein genes occurs in a coordinated manner at higher levels than other transcription events, it has been suggested that their regulation may involve specialized mechanisms that cannot be generalized to other genes. In spite of this, there is precedent for transcription factors regulating ribosomal protein expression. For example, Rap1 is a general transcription factor that can either activate or repress transcription, that binds promoters of a significant number of ribosomal protein genes in yeast[91,98]. Although Rap1 does not have a clear ortholog in *C.elegans*, it is possible that a homologous TF exists in *C.elegans* that regulates ribosomal protein transcription similarly.

We have identified several highly significant conserved motifs in the ribosomal promoters of C. elegans, and we have functionally validated several of these novel elements, showing that they are required for the proper coexpression of the RPs. In particular, M546 is necessary for *C.briggsae* RP promoters to drive expression n *C.elegans*, suggesting it may be involved in an evolutionarily conserved mechanism of transcriptional regulation. If M546 sites function in a variety of ribosomal and non-ribosomal promoters, M313 and M540 may act as ribosomal-specific, accessory motifs that recruit a general transcriptional regulator that recognizes M546 to the promoter to initiate transcription. This is consistent with the finding that all motif mutations observed in our reporter assay are inactivating mutations, suggesting that the regulators specific for the motifs are activators of transcription. It would be interesting to see if this model of regulation applies for non-ribosomal promoters that also contain an M546 site. We could ask if any motifs co-occur with M546 in non-ribosomal promoters and investigate if genes containing both M546 and accessory motif are involved in similar pathways or programs.

The tight regulatory control of RPs is beginning to be corroborated by recent findings that improper expression of ribosomal proteins and ribosomal rRNA leads to diverse developmental defects: 1) heterozygous RP mutant flies exhibit a "minute" phenotype, (homozygous mutants are embryonic lethal)[41] 2) reduced rRNA expression leads to defects in *glp-1/notch* signaling and the meiotic-mitotic transition in the nematode gonad, (possibly due to insufficient levels of *delta* ligand)[99], and 3) decreased levels of rRNA in the *bap28* mutant lead to increased apoptosis in the zebrafish CNS and embryonic lethality[100]. Direct evidence of the consequences of ribosomal protein misregulation have been reported in flies: decreased levels of RPS3 due to a p-element insertion in its promoter were shown to lead to the minute phenotype and defects in gonadogenesis[46].

We hypothesize that defects in ribosomal protein regulation and reduced ribosomal protein expression levels will have diverse phenotypes in a wide range of translationally challenged cells, especially in combination with other genetic defects. The pathways and TFs which regulate ribosomal protein biosynthesis in mammals and in *C. elegans* are currently unknown, and are difficult to study by classical techniques because they are essential. We are hopeful that further studies of these elements will identify the mechanisms of coexpression of RPs in *C. elegans*, and yield insight into mammalian RP coexpression.

Chapter 6:
Detecting *cis*-regulatory modules using evolutionary conservation

6.1 ABSTRACT

Cis-regulatory elements often co-occur in spatially localized clusters in promoters, known as *cis*-regulatory modules (CRMs). Computational methods were originally developed to detect over-represented CRMs from sequence and more recent approaches incorporate sequence conservation to improve search accuracy, but none of these algorithms model organizational relationships between *cis*-regulatory elements in CRMs. We propose that these organizational relationships, reflected in the relative spacing, positioning and orientation biases between elements, are functionally conserved between species and can be useful in CRM prediction. As a proof-of-principle, we show that PAC and RRPE elements are related via conserved spacing and positioning biases. Applying similar analysis to ribosomal motifs predicted by Flipper, we discover a strong spacing bias of 8/11 bp between pairs of M546 and M540 elements. Using a "sequence-swap" experiment, we demonstrate that the 8 bp spacing constraint is necessary for promoter function in the *rps-7* promoter, but the 11 bp spacing is not functional in the *rpl-2* promoter. We further show that non-ribosomal promoters containing M546-M540 pairs can be divided into two classes – those with 8/11 bp spacing and others with 20/21 bp spacing. We determine that the 20/21 bp spacing is the result of a CELE2 transposon insertion in these promoters. Interestingly, non-ribosomal promoters with 8/11 bp spacing are enriched for growth and development GO annotations, whereas promoters with 20/21 bp spacing are not significantly enriched in any categories. In conclusion, our results suggest that M546 and M540 constitute a *cis*-regulatory module involved in early growth and development.

6.2 INTRODUCTION

In general, trans acting factors that recognize *cis*-regulatory elements do not act independently – instead, they interact cooperatively with other binding partners to recruit chromatin remodeling complexes, enhance overall affinity of binding etc. The 'footprint' of these protein-protein interactions is reflected in the organization of *cis*-regulatory elements in the promoter. Functional elements do not occur individually, rather they co-occur in spatially localized clusters in promoters. Well characterized examples of *cis*-regulatory modules include clusters of transcription factor binding sites in the developmental gene regulatory network (GRN) of the sea urchin[101] and modules involved in regulating the segmentation genes of *D.melanogaster* [102,103].

Experimental identification of *cis*-regulatory modules and the underlying regulatory logic can be laborious and time consuming, due to the combinatorial number of possibilities for testing. As such, several algorithms have been developed to detect *cis*-regulatory modules from sequence information alone, i.e. Ahab[102], Stubb[104], cisModule[105], EMCModule[106], ModuleDigger[107] and GibbsModule[108] etc. The challenge in *de novo cis*-regulatory module detection is the simultaneous identification of PWMs for motifs involved and the number and types of motifs involved in each module. As a work-around, Ahab[102] uses a predefined set of input motifs and fits the probabilities of observing weight matrices vs. background in the module sequence within a maximum likelihood framework. Stubb[104] uses a similar approach, but has the added ability to incorporate an evolutionary model of binding site conservation by comparing sequence from a second, closely related ortholog, i.e. *D.melanogaster vs. D.pseudoobscura.* EMCModule[106] also discovers CRMs from an input collection of PWMs, but distinguishes itself from other algorithms by modeling the preference of TFs for binding partners as a transition matrix (each element models the dependence between motif *i* and *j*) and parameterizing the distance between binding sites as a truncated geometric distribution. The algorithm uses Gibbs sampling to learn parameters for the transition matrix and geometric distribution, uses a forward-backward dynamic programming approach to learn the binding site locations and an 'evolutionary Monte Carlo' approach (hence EMC) to determine which input motifs should be included in the CRM. A truly *de novo* approach, cisModule[105] eschews the input motif collection approach and uses a hierarchical mixture model to simultaneously detect CRM positions within promoter sequence and the positions and compositions of the binding sites within them. A two level hierarchical model is used, where at the first level, input sequences are a mixture of background and CRM components, at the second level, CRMs are a mixture of motifs and CRM-specific background. At each level, the authors use a dynamic programming approach to determine CRM and motifs positions by recursion. PWMs for motifs are learned by aligning the most frequently sampled sites for the *k*th motif in the CRMs.

Just like their component *cis*-regulatory elements, functional *cis*-regulatory modules are likely to be conserved across species as well. Indeed, the Stubb[104] algorithm mentioned above was an example to include two-species comparisons in CRM discovery. A recent algorithm, GibbsModule[108], extends this further to *N* species comparison where CRM conservation is leveraged upon results from multiple sequence alignment. Briefly, CRMs are modeled not as a collection of TF binding sites, but rather sequence centered about a core TF binding site. The sequence window is 200 bp, 100 bp each side of the core motif. The algorithm learns a PWM for the core motif and each iteration samples *N* sites in the input sequences, as well as its orthologs, as putative motif instances. Each binding site and its surrounding sequence is taken as a CRM and CRMs from the anchor species are pairwise aligned against ortholog CRMs. The authors emphasize that CRM sequences are not aligned using a traditional Smith-Waterman approach. Since mutations can change the sequence structure but not affect functional interactions, the authors uses a method called Module Alignment, which iteratively performs local alignment and masks the best aligned segment in successive iterations. The CRM conservation score is a cumulative sum of individual best local aligned segment scores. The PWM of the core motif is updated from the most conserved CRM pairs and used for the next iteration. While a step in the right direction, this approach may be limited to CRM discovery in cases where a core motif *is* anchoring the module. Ideally, CRMs should also be compared across *N* species – the authors used a pairwise approach instead, presumably to simplify computation.

What the EMCModule approach alluded to and is lacking in many of the other approaches mentioned above, is that CRMs are not purely collections of motifs, but contain an intrinsic organizational structure that is characteristic and relevant to its role in regulation. EMCModule attempted to model these organizational constraints by parameterizing the distribution of distances between sites, as well as modeling the dependency between motifs in a CRM. Beer *et al*[65] demonstrated that modeling distance, positioning and orientation constraints in co-occuring sets of TF binding sites can improve one's ability to predict which genes are coexpressed. In this study, we extend this idea by proposing that these organizational constraints, by virtue of their functional relevance, may also be conserved across species. As a proof-of-principle, we study the organization of PAC and RRPE elements in promoters of yeast and related species and extend our analysis to ribosomal motifs previously discovered using Flipper. We show that pairs of M546 and M540 motifs have characteristic spacing and orientation biases, and demonstrate that these constraints are functional *in vivo.*

6.3 RESULTS

6.3.1 PAC-RRPE organizational constraints are conserved across species.

Previous studies have suggested that transcription factors binding PAC and RRPE elements may interact synergistically and this interaction is reflected in the organization of PAC and RRPE in *cis*-regulatory modules[65,109-111]. For example, genes with PAC in their promoter located closer to translation start than RRPE, are significantly more correlated in their expression. In other words, PAC and RRPE are part of a regulatory module where PAC is positionally constrained relative to RRPE. We asked if PAC-RRPE pairs present with characteristic spacing and orientation biases, which may reflect secondary organizational constraints of the cis-regulatory module.

We applied *k*-means clustering to gene expression data from yeast[112,113] and obtained a cluster of 112 genes involved in rRNA processing, that are enriched in PAC and RRPE elements. We scanned promoter sequences using PWMs representing the known PAC and RRPE binding motifs and picked the highest scoring site within each promoter as its representative motif site. Promoters were aligned so the putatively regulated gene was located at the 3' end. Spacing differences were counted in base pairs for each PAC-RRPE pair so that if PAC were located closer to translation start than RRPE, the difference would be recorded as positive - negative spacing for RRPE located closer to translation start than PAC. Plotting the observed spacing differences in a histogram, we observe a distinct bias for positive spacing, confirming previous observations of positional bias. In addition, PAC and RRPE elements tend to be located very closely to each other with a clear spacing bias of 0 to 5 bp.

To determine if this spacing bias is conserved in PAC-RRPE pairs in orthologs, we repeated our analysis using promoters of orthologous genes in *S.paradoxus*, *S.mikatae* and *S.bayanus*. Plotting the results in histograms, we observe a similar bias towards positive spacing in orthologous PAC-RRPE pairs. In addition, spacing differences in PAC-RRPE pairs follow a similar distribution in orthologs with peaks at 0-5 bp. To determine if the similarity in distribution was due to conservation of spacing differences, we plotted the spacing difference of a PAC-RRPE pair in a *S.cerevisiae* promoter against the spacing difference of a PAC-RRPE pair in its corresponding promoter ortholog. A scatter plot of the results are shown in Fig 6.1. In general, most points (representing spacing difference comparisons) lie on the y=x diagonal, indicating that in general, there is little change in spacing differences between orthologous PAC-RRPE pairs. Furthermore, most points lie in the upper right quadrant (positive-positive), indicating that there is conservation of the positioning bias as well. There is especially strong conservation of the 0-5 bp spacing bias.

To determine if this conservation is significant, we repeated our analysis on randomly selected promoters instead. While we observed an increase in off-diagonal points, a substantial number of points remained on the y-x diagonal, suggesting that even in unrelated promoters, there is little change in promoter composition between orthologs due to the small size and compact nature of the genomes compared. We observe a similar concentration of points in the 0-5 bp range for random promoters, but this is substantially lower than the peak observed for coexpressed genes, confirming that the 0-5 bp spacing bias in coexpressed genes is indeed significantly conserved.

We next asked if PAC-RRPE pairs were constrained in terms of their relative orientation, in addition to their relative spacing and position. Since neither PAC nor RRPE are palindromic, there are four orientation configurations possible. We asked what percentage of PAC-RRPE pairs follow a particular orientation configuration and if the observed counts differed significantly from counts expected from repeating the analysis on a set of random promoters. At the same time, we compared orthologous PAC-RRPE pairs to investigate if biases in relative orientation are also conserved across species. Our analysis results are represented in a two-way table in Fig 6.2.

We did not observe a substantial difference between counts observed on coexpressed promoters vs. random promoters, suggesting that relative orientation does not affect the interaction of the TFs recognizing the PAC and RRPE elements. Interestingly, in spite of this, a substantial number of orientation configurations are conserved for PAC-RRPE pairs in promoters of coexpressed genes, more than expected by comparison to random promoters. While there does not appear to be an orientation bias, orientation configurations could still be conserved , because promoter orthologs of coexpressed genes are more likely to be conserved than orthologs of unrelated random genes.

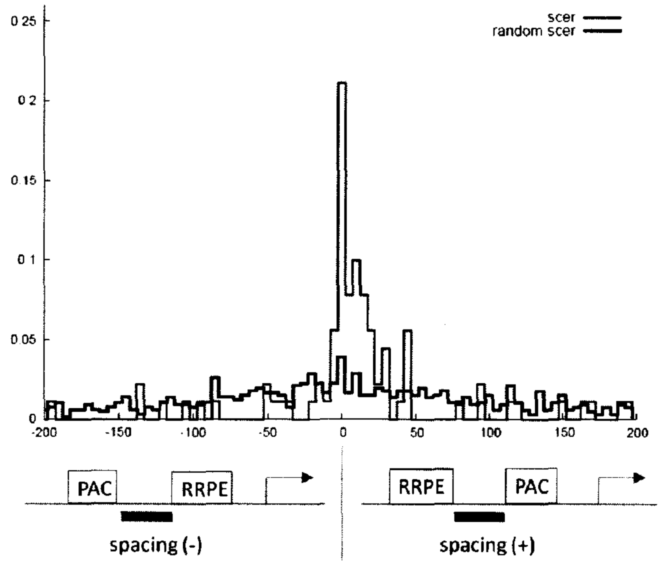
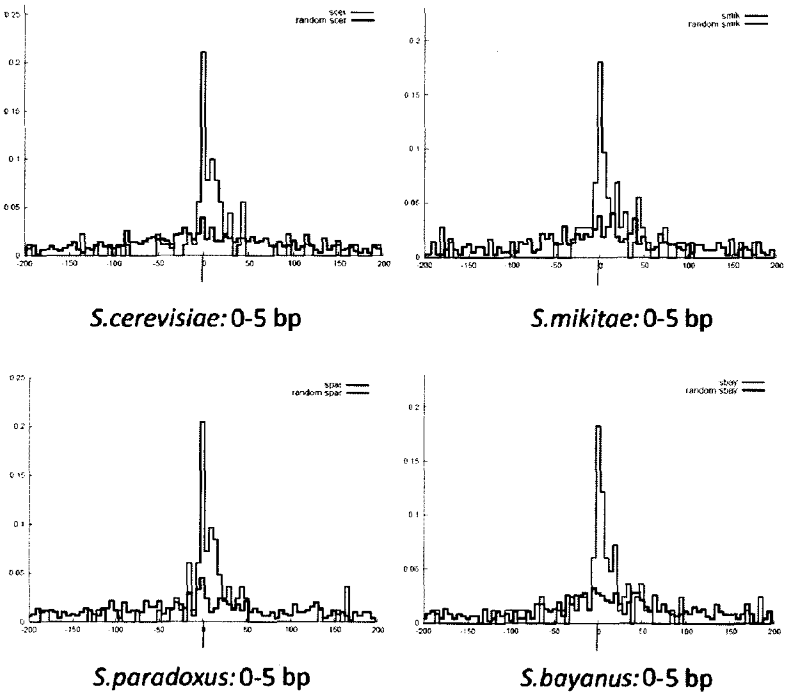


Figure 6.1: Distribution of spacing differences between PAC and RRPE elements in promoters of coexpressed genes. Positive: PAC closer to translation start than RRPE. Negative: RRPE closer to translation start than PAC.



S.cerevisiae: 0-5 bp

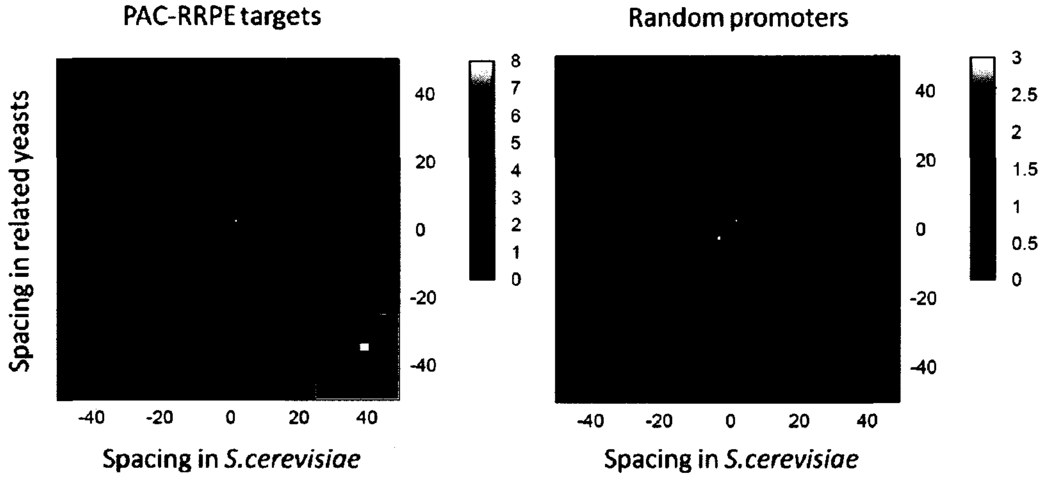
S.mikitae: 0-5 bp

S.paradoxus: 0-5 bp

S.bayanus: 0-5 bp

Figure 6.2:

Distribution of spacing differences between PAC and RRPE elements in promoters of coexpressed genes and their orthologs. There is consistent over-representation of the 0-5 bp spacing difference among all species.



→ PAC

→ RRPE

PAC-RRPE targets

Orientation in *S.cer*

Orientation in orthologs

	→→	→←	←→	←←
	26%	28%	15%	31%
→→	30%	19%	14%	36%
→←	16%	53%	20%	11%
←→	23%	13%	28%	36%
←←	16%	7%	17%	60%

Random promoters

Orientation in *S.cer*

Orientation in orthologs

	→→	→←	←→	←←
	26%	24%	23%	27%
→→	32%	22%	24%	22%
→←	23%	25%	27%	24%
←→	23%	24%	29%	24%
←←	23%	23%	23%	31%

Figure 6.3:
Comparison of PAC-RRPE spacing and orientation differences between *S.cereivisiae* promoters and their respective orthologs. Significance is assessed by repeating analysis on a negative control set of unrelated promoters.

6.3.2 M546-M540 pairs have characteristic spacing and orientation bias.

Using PAC-RRPE as an example, we showed that aspects of *cis*-regulatory module organization, reflected in characteristic spacing, orientation and positioning biases, can be conserved across orthologous comparisons. We next asked if similar organizational biases existed between pairs of RP motifs (predicted by Flipper) and if we could use this information to enable us to discover RP *cis*-regulatory modules.

We performed a similar analysis *vis a vis* the PAC-RRPE elements: RP promoters were scored using PWMs for RP motifs, the highest scoring site was chosen as a representative and particular spacing, positioning and orientation biases were counted for pairs of RP motifs. As a control, we repeated our analysis on promoters of unrelated genes and asked if our observed counts with RP promoters were significantly higher than expected with random promoters. We used orthologs from *C.brenneri*, *C.remanei* and *C.briggsae* for comparison. Of all the pairs of RP motifs considered, M546 and M540 exhibited the most obvious organizational biases. This is not entirely unexpected given our results from a genome wide scan of motif occurrences, where we observed M546 and M540 co-occurring as pairs in a large number of promoters, not all of them ribosomal.

To visualize the organization of M546-M540 pairs, we plotted the distribution of their spacing differences as histograms in Fig 6.4, where spacing is taken as positive if M546 is located closer to translation start than M540 and negative otherwise. Our results indicate that there is a strong positioning bias for

M546 being located closer to translation start than M540. In addition, we observe two peaks in the spacing distribution at 8 and 11 bp, suggesting that M546 and M540 may share characteristic spacing bias similar to PAC-RRPE. M546-M540 pairs in RP orthologs share similar positioning and spacing biases, indicating that these organizational constraints are indeed conserved. In contrast, we did not observe any peaks in distributions from random promoters (median count: 1).

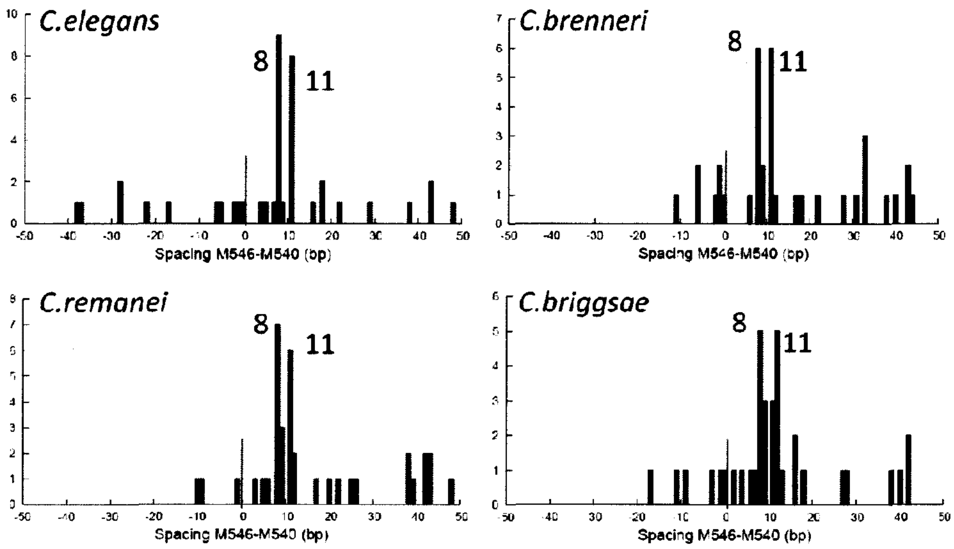
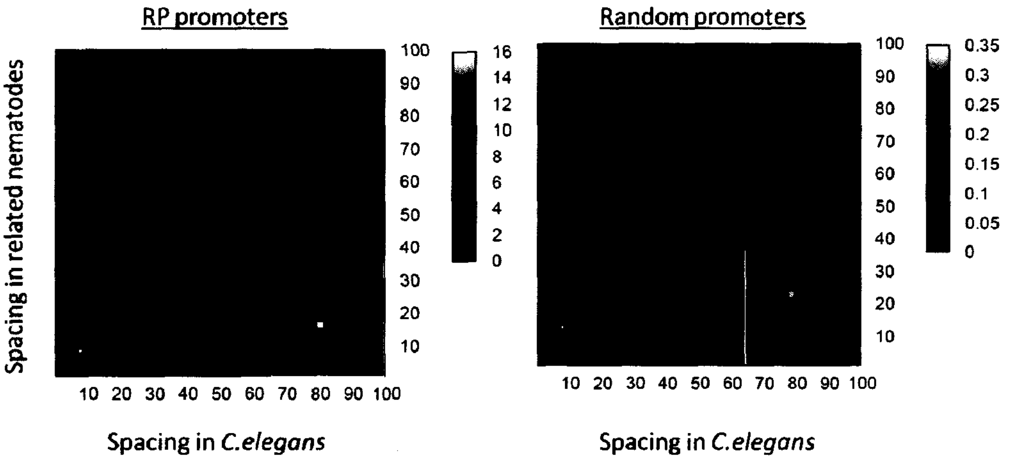


Figure 6.4:
Distribution of spacing differences between M546 and M540 pairs in RP promoters from *C.elegans* and their orthologs in *C.brenneri*, *C.remanei* and *C.briggsae*. A line in each plot at 0 bp marks the transition from positive to negative spacing. Spacing differences greater than 50 bp are omitted for presentation, but are not significant in number.

Following our analysis with PAC-RRPE, we compared spacing differences between M546-M540 pairs in *C.elegans* promoters against corresponding pairs in orthologs. Since there is a strong bias towards positive spacing, we represent all spacing differences by their absolute value in the scatter plot. We observe that similar to PAC-RRPE, most points in our comparison lie on the y=x diagonal in the scatter plot, indicating that the similarity in distribution between species is indeed due to conservation of the spacing bias between M546-M540 pairs in orthologs. Furthermore, there is strong conservation of the 8/11 bp spacing bias, more than expected by repeating the same analysis using random unrelated promoters. We used 10 sets of random promoters and averaged the result. More points in the random promoter comparison were located off-diagonal and hence not conserved. The 8/11 bp spacing bias appeared in the random promoter comparison as well, but at frequencies substantially lower than with RP

promoters.



➡ M546

➡ M540

RP promoters

Orientation in *C.elegans*

Orientation in orthologs	➡ ➡	➡ ⬅	⬅ ➡	⬅ ⬅
	21%	17%	**47%**	15%
➡ ➡	39%	21%	16%	24%
➡ ⬅	17%	30%	27%	27%
⬅ ➡	9%	8%	**71%**	12%
⬅ ⬅	26%	15%	37%	22%

Random promoters

Orientation in *C.elegans*

Orientation in orthologs	➡ ➡	➡ ⬅	⬅ ➡	⬅ ⬅
	18%	28%	26%	28%
➡ ➡	31%	30%	28%	22%
➡ ⬅	31%	29%	20%	26%
⬅ ➡	25%	29%	32%	25%
⬅ ⬅	31%	26%	25%	26%

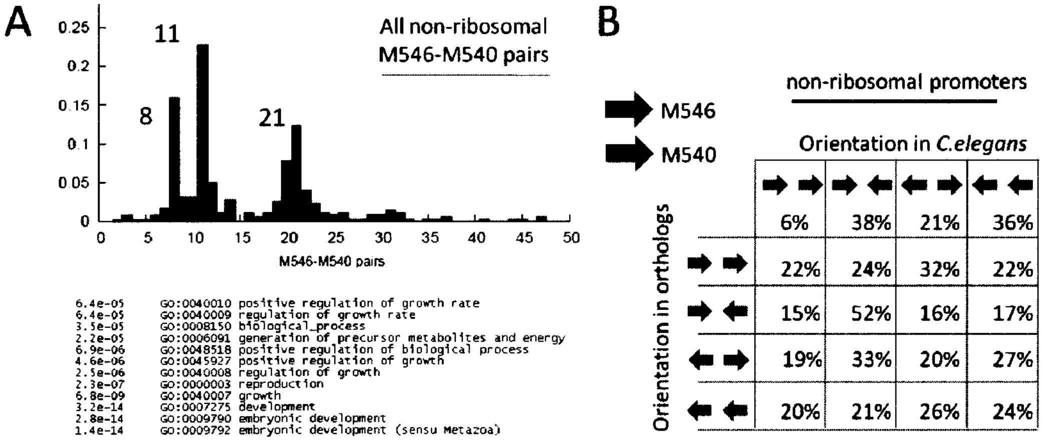
Figure 6.5:
A) Spacing differences between M546 and M540 sites in RP promoters are compared against spacing in RP orthologs. We repeated the analysis using a set of random promoters as a control. B) Relative orientation differences between M546 and M540 pairs in RP promoters and their orthologs, compared against random promoters.

Similar to our analysis with PAC and RRPE, we asked if M546 and M540 sites are constrained in terms of their relative orientation. We observed that up to 47% of all M546-M540 pairs exhibited a particular orientation bias. We determined the significance of this observation by performing a Chi-squared goodness of fit test against the expected distribution on random promoters and obtained a *p* value of 3.7e-3. Furthermore, there appears to be strong conservation of this orientation bias: 71% of M546-M540 pairs in orthologous promoters exhibit the same orientation bias if it is present in the anchor species promoter. Interestingly enough, the orientation and spacing bias of M546-M540 pairs appears to be related: 17 of 76 RP promoters contain M546-M540 pairs spaced 8/11 bp apart, but 16 out of 17 of these promoters (94%) exhibit this orientation bias, compared to the background rate of 47% among all RP promoters. This suggests that the spacing and orientation biases may be part of the same general organizational rule for this regulatory module.

6.3.3 M546-M540 pairs in non-ribosomal promoters overlap with CELE2 insertions.

In a genome wide scan for M546 and M540 sites in *C.elegans* promoters, we had identified many non-ribosomal promoters that contain high confidence sites for both motifs. We had tested M546 in one of these promoters for function (*mcm-7*) and confirmed that it was indeed needed for regulating non-ribosomal expression. Because M546 and M540 tended to co-occur as pairs in non-ribosomal promoters, we asked if they possessed different organizational constraints which would distinguish them from pairs in RP promoters.

We filtered M546 and M540 sites in all promoters based on their log-likelihood ratio scores, requiring that high confidence sites score at least 5 log10 fold higher than expected by a background model. This resulted in 571 non-ribosomal promoters that contain strong M546 and M540 pairs. We analyzed spacing and orientation constraints for these pairs, but did not examine positional bias because of the ambiguity in assigning regulated genes for divergent intergenic regions.



6.4e-05	GO:0040010	positive regulation of growth rate
6.4e-05	GO:0040009	regulation of growth rate
3.5e-05	GO:0008150	biological_process
2.2e-05	GO:0006091	generation of precursor metabolites and energy
6.9e-06	GO:0048518	positive regulation of biological process
4.6e-06	GO:0045927	positive regulation of growth
2.5e-06	GO:0040008	regulation of growth
2.3e-07	GO:0000003	reproduction
6.8e-09	GO:0040007	growth
3.2e-14	GO:0007275	development
2.8e-14	GO:0009790	embryonic development
1.4e-14	GO:0009792	embryonic development (sensu Metazoa)

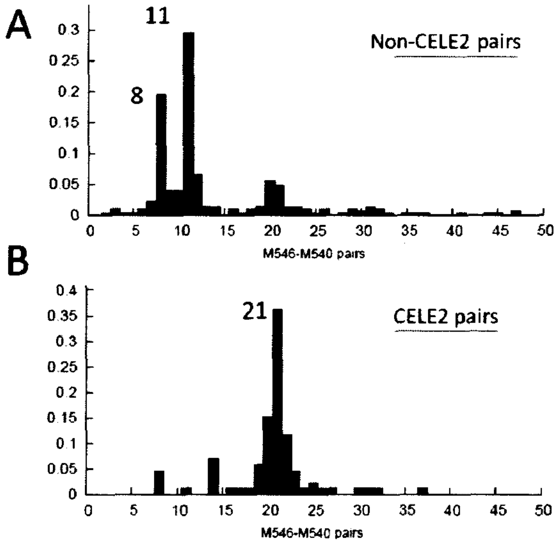
	→ →	→ ←	← →	← ←
	6%	38%	21%	36%
→ →	22%	24%	32%	22%
→ ←	15%	52%	16%	17%
← →	19%	33%	20%	27%
← ←	20%	21%	26%	24%

Figure 6.6:
A) Spacing differences between M546 and M540 pairs in non-ribosomal promoters. GO terms enriched in these promoters are listed below. Significance of enrichment was computed using a hypergeometric distribution. B) Relative configurations of M546-M540 pairs in RP promoters and their respective orthologs.

Our results are shown in Fig 6.6. We observe three modes of possible spacing bias between M546 and M540 site pairs: 8, 11 and 21 bp. The 8/11 bp spacing is similar to that of RP promoters but the 21 bp spacing appears to be unique to the non-ribosomal promoters only. In terms of relative orientation, we did not observe any marked bias towards any one particular orientation constraint. We had expected a slight bias towards the orientation bias seen in RP motif pairs, but did not observe this with non-ribosomal pairs. To determine if non-ribosomal genes associated with high confidence M546 and M540 pairs are enriched for common functions, we performed a GO term enrichment analysis where enrichment significance is determined using p values derived from a hypergeometric distribution. Enriched GO terms generally fall in the category of growth (p=6.8e-9) and development (p=3.2e-14). This is somewhat expected given that RPs are required in biosynthetic pathways and mechanisms for their regulation may extend to other genes involved in similar growth and proliferative programs.

Upon closer inspection, we noticed that a significant number (206 out of 571) of selected non-ribosomal promoters overlapped with the transposon CELE2 in *C.elegans*. CELE2 is non-autonomous DNA transposon and belongs to the family of Miniature Inverted-Repeat Transposable Elements (MITEs) and happens to contain two seed sites that strongly resemble the M546 and M540 motifs in its terminal inverted repeat region[114]. These seed sites are spaced 21 bp apart, which could account for the added bias in spacing observed. Expectedly, after sorting the non-ribosomal promoters by their overlap with CELE2, we observed that promoters overlapping CELE2 contained M546-M540 pairs exclusively spaced 20/21 bp apart, whereas pairs in non-overlapping promoters were spaced 8/11 bp apart.

Transposons have been proposed as a mechanism for spreading seed sites for regulatory elements by previous studies: a) transposases have been "domesticated" for use in genome recombineering because they recognize specific sequence signatures[115-117], b) transposable elements contain seed sites that sufficiently resemble TF binding motifs so point mutations that increase their resemblance are positively selected for[118]. We asked if CELE2 may be performing a similar function by inadvertently spreading M546 and M540 seed sites to non-ribosomal associated promoters. To answer this question, we looked for GO terms enriched in either a) non-ribosomal promoters containing CELE2 insertions and b) non-ribosomal promoters lacking CELE2 insertions. Whereas non-ribosomal promoters lacking CELE2 insertions remain enriched for growth and development GO categories, e.g. GO:0048513 organ development (p = 3e-5), GO:0048806 genitalia development (p =2.1e-5), GO:0040007 growth (p = 2.5e-10), promoters that contain strong M546-M540 pairs by virtue of CELE2 insertion are *not* enriched in any common GO categories.



7.7e-05	GO:0050896	response to stimulus
5e-05	GO:0007548	sex differentiation
4.3e-05	GO:0040035	hermaphrodite genitalia development
3.9e-05	GO:0048856	anatomical structure development
3e-05	GO:0048513	organ development
3e-05	GO:0005622	intracellular
2.1e-05	GO:0048806	genitalia development
1.8e-05	GO:0040010	positive regulation of growth rate
1.8e-05	GO:0040009	regulation of growth rate
2.7e-06	GO:0050789	regulation of biological process
1.5e-06	GO:0045927	positive regulation of growth
1.4e-06	GO:0048518	positive regulation of biological process
4.6e-07	GO:0040008	regulation of growth
2.8e-07	GO:0008150	biological_process
2.5e-10	GO:0040007	growth
9.8e-12	GO:0000003	reproduction
1.3e-16	GO:0009790	embryonic development
7.1e-17	GO:0009792	embryonic development (sensu Metazoa)
9.5e-18	GO:0007275	development

0.004	GO:0006091	generation of precursor metabolites and energy
0.0036	GO:0016491	oxidoreductase activity
0.0021	GO:0044272	sulfur compound biosynthesis
0.0016	GO:0006118	electron transport
0.0014	GO:0004096	catalase activity

Figure 6.7:
Distribution of spacing differences between M546 and M540 sites, as well as enriched GO terms are shown for A) non-ribosomal promoters that lack CELE2 insertions and B) non-ribosomal promoters that contain CELE2 insertions.

Our results suggest that promoters do not gain M546-M540 associated regulation by virtue of CELE2 insertion. Although the terminal ends of the CELE2 transposon contain strong seed pairs for M546 and M540, it is possible that the associated trans-acting regulators are unable to bind to these motifs and form a successful interaction because the spacing difference between the sites is 20 bp as opposed to 8/11 bp. In addition, non-ribosomal genes with M546-M540 pairs spaced 8/11 bp are associated with growth and development GO annotations that are consistent with the need for RPs in such processes. This indicates that the M546-M540 regulatory module is functional in these promoters, despite their marked difference in expression profile from RPs, reinforcing the hypothesis that M546-M540 may be recognized by a set of general developmental regulators, i.e. chromatin remodeling complexes, that instead rely on secondary TFs recognizing gene-cluster specific *cis*-regulatory elements to confer specificity of gene expression.

6.3.4 Characteristic spacing between M546 and M540 is necessary for promoter function.

To determine if the spacing bias between M546 and M540 pairs is necessary for promoter function, we performed the following "sequence swap" experiment where we exchanged the position of M540 in the *rps-7* promoter with an unrelated sequence 50 bp upstream of its original position. In doing so, we would increase the effective distance between the M546 and M540 elements, changing the spacing from 8 bp to 58 bp. The "sequence swap" design has the added advantage of preserving overall promoter composition, which would have been changed if extraneous sequence were inserted between M546 and M540 elements instead. It is possible that *cis*-regulatory elements overlapping junctions of sequence exchange are disrupted in this design, which is why we deliberately selected a sequence from a region of low alignment conservation as the control sequence for exchange, to minimize this risk. Fig 6.8 shows the sequence alignability across the *rps-7* promoter and outlines steps in the overall mutagenesis scheme.

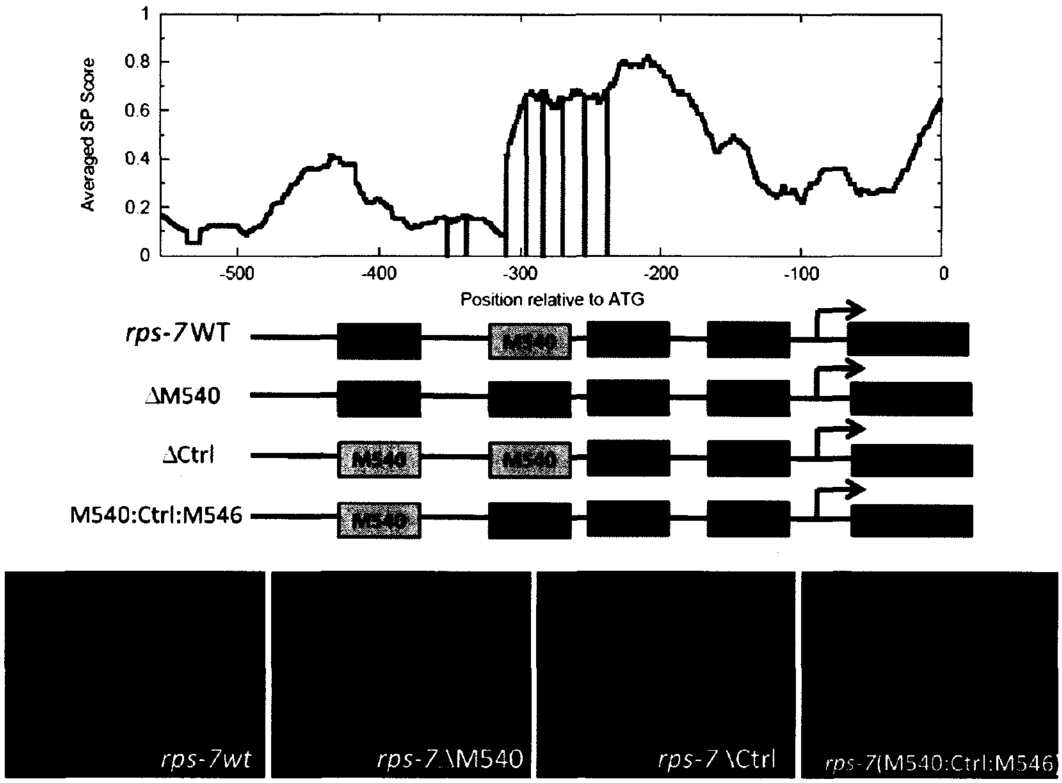


Figure 6.8:
Scheme for testing M546-M540 spacing constraint in the *rps-7* promoter. Sequence alignability in terms of SP score is plotted across the entire promoter and the control sequence element, M540, M546 and M313 are highlighted in red, green, blue and pink respectively. ΔCtrl: control element replaced by M540, ΔM540: M540 site replaced by control element, M540:Ctrl:M546: M540 and control elements exchanged positions.

We exchange positions of M540 and the control sequence element (CTRL) by first mutating either M540 to CTRL (ΔM540), or CTRL to M540 (ΔCTRL). The ΔM540 promoter effectively has two copies of CTRL, whereas the ΔCTRL promoter has two copies of M540. Our results show that M540 is necessary for promoter function, since mCherry expression from the ΔM540 construct is severely reduced. This recapitulates previous results where we scrambled the M540 motif, instead of replacing it with another sequence from the same promoter. mCherry expression from the ΔCTRL construct is unchanged from WT, confirming that the CTRL sequence element is indeed neutral to promoter function. To obtain the final construct, we mutated the original CTRL to M540 in the ΔM540 construct, or the original M540 to CTRL in the ΔM540 construct. The phenotype with the change in spacing was more subtle but still noticeable – mCherry expression was diminished in most tissues of the worm with residual expression in hypodermal cells.

We performed the same "sequence swap" experiment, this time using the *rpl-2* promoter. M546 and M540 sites in the *rpl-2* promoter are spaced 11 bp apart with a different orientation constraint from the sites in the *rps-7* promoter. In addition, the *rpl-2* and *rps-7* promoters contain different combinations of motifs – for example, *rps-7* contains a strong M313 site whereas *rpl-2* does not, which may imply that the regulatory logic represented by these combinations is different.

The results of this experiment are represented in Fig 6.9. We had previously mutated the M540 element in the *rpl-2* promoter by scrambling and found this had no effect on mCherry expression. We observed a similar result with the *rpl-2:*ΔM540 construct, indicating that the M540 element was by itself not necessary for promoter function. mCherry expression from the ΔCTRL construct was unaffected even though the CTRL sequence element in the *rpl*-2 promoter was more alignable than its counterpart in the *rps-7* promoter, reinforcing the idea that not all highly alignable sequences are functional. Finally, unlike our previous observations with the *rps-7* promoter, we did not observe any change in mCherry expression when the positions of the M540 and CTRL sequence elements were exchanged. There are several possible reasons: Firstly, the M540 element in the *rpl-2* promoter may be a false positive site, since it does not affect mCherry expression when mutated. Increasing the distance between M546 and a neutral sequence element is less likely to affect promoter function. Secondly, the 11 bp spacing in the *rpl-2* promoter may not be functional as opposed to the 8 bp spacing in the *rps-7* promoter. We will need to perform more "sequence swap" experiments with other promoters in order to verify this hypothesis. Lastly, M546 and M540 are in a different orientation in the *rpl-2* promoter compared to *rps-7* and different logic rules may be encoded by the modules in the *rpl-2* promoter vs. *rps-7* promoter.

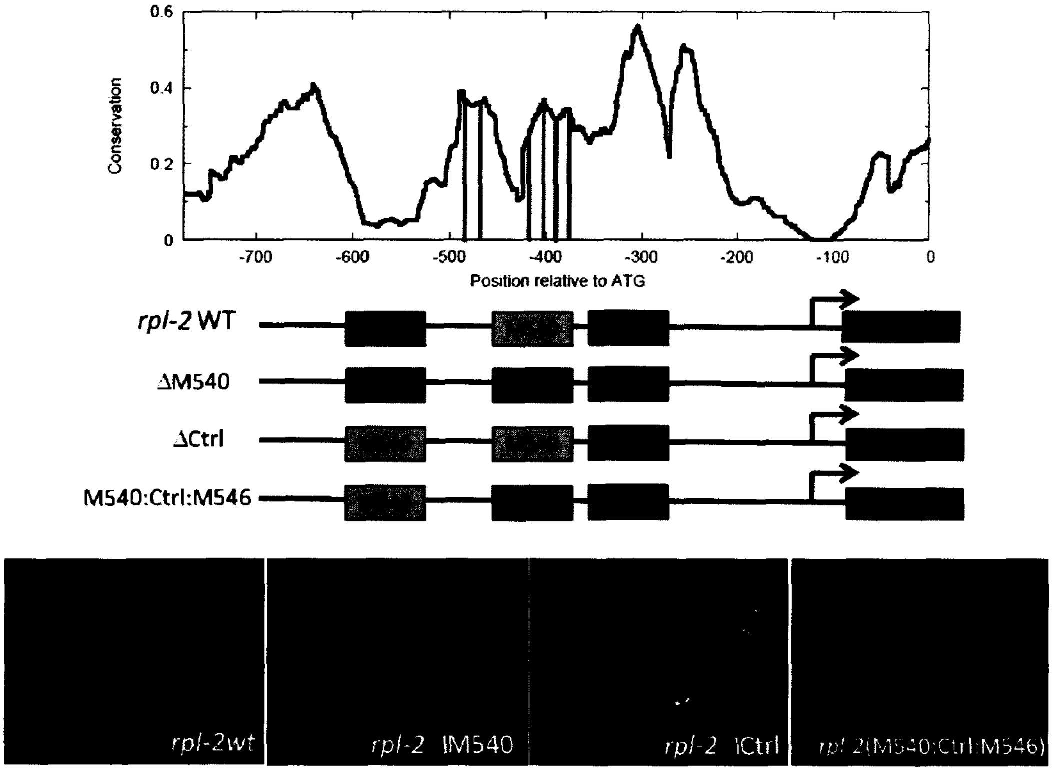


Figure 6.9:
Scheme for testing M546-M540 spacing constraint in the *rpl-2* promoter. Sequence alignability plots and mutation scheme are similar to Figure 6.8. M540 and M546 are highlighted in green, and blue respectively. Spacing between M546 and M540 sites in *rpl-2* is 11 bp, as opposed to 8 bp in *rps-7*.

6.4 DISCUSSSION

Using PAC-RRPE as an example of a well-characterized *cis*-regulatory module, we have shown that aspects of *cis*-regulatory module organization can be represented by characteristic spacing, orientation and positioning biases exhibited by pairs of *cis*-regulatory elements and that these organizational constraints can be strongly conserved when comparing corresponding *cis*-regulatory modules in orthologous promoters. Using conservation as a filter, we find similar spacing and orientation biases between two RP motifs, M546 and M540, we had previously identified using Flipper. M546 and M540 share a strong spacing bias of 8/11 bp and we observe similar spacing bias for M546-M540 pairs in non-ribosomal promoters as well. Interestingly, a substantial number of non-ribosomal promoters contain strong sites for M546 and M540 by virtue of CELE2 insertion, but M546-M540 pairs derived from this mechanism are spaced 20/21 bp apart, as opposed to 8/11 bp. We show that whereas non-ribosomal genes containing CELE2 insertions do not share common GO annotations, other non-ribosomal genes lacking CELE2 insertions are strongly enriched in growth and development GO annotations, suggesting that M546-M540 associated regulation is strongly tied to the 8/11 bp spacing constraint. Using a "sequence swap" experiment to perturb the spacing between M546 and M540 in the *rps-7* and *rpl-2* promoters, we demonstrate that the 8 bp, but not 11 bp spacing constraint is necessary for promoter function.

While methods have been developed to learn *cis*-regulatory logic rules from sets of over-expressed genes[65], there has not been an attempt to determine if these rules are conserved in *cis*-regulatory modules present in orthologous promoters. Our analysis is a step in this direction and suggests conservation can be an orthogonal source of information to over-representation for the detection of *cis*-regulatory modules. Following similar principles as Flipper, we can design a *de novo cis*-regulatory module discovery algorithm that instead of sampling groups of conserved sites, samples pairs of sites that are distributed across orthologous promoters and share characteristic spacing, positioning or orientation biases. Analogous to the difference between *k*-mer approaches and position weight matrices (PWMs), we can allow degeneracy in the organizational constraints by representing them using a similar frequency matrix approach. We can learn the set of most informative counts by similar Gibbs sampling approach. We predict that alignment-based approaches may have limited success in *cis*-regulatory module detection, since *cis*-regulatory modules with characteristic but flexible spacing constraints (i.e. M546 and M540 pairs with a bimodal 8/11 bp spacing bias) will likely score poorly in a multiple alignment that emphasizes overall alignment sequence identity.

Many studies of *cis*-regulatory element function using transcriptional reporter assays have focused on testing individual elements as opposed to testing interactions between elements. Our study is unique in this regard and we intend to extend it by asking if the orientation bias observed between M546 and M540 is functional as well. Additionally, the presence of strong seed sites for M546 and M540 motifs in the terminal inverted repeat region of CELE2 poses some interesting questions for further study. Firstly, why are M546-M540 pairs spaced by 8/11 bp functional, but not corresponding pairs spaced 20 bp apart? Since a turn of the DNA helix is approximately 10 bp, it could be that increasing the spacing bias to 20 bp places the factors that bind M546 and M540 on different sides of the helix, consequently disrupting their interaction. It would be interesting if we could confer M546-M540 associated regulation onto a non-ribosomal promoter by decreasing the spacing between its M546 and M540 sites from 20 bp to 8 bp for example. This would be even stronger evidence arguing that the interaction between M546 and M540 is real and functional. Secondly, despite the increased spacing bias, the seed sites for M546 and M540 are individually strong seed sites for these motifs. How then has *C.elegans* adapted to such profligate spreading of M546 and M540 seed sites across its genome? Perhaps M546 and M540 are by themselves not sufficient to effect change in transcription because they bind general transcriptional regulators that require recruitment by specific TFs. Since CELE2 is specific to *C.elegans* and not other nematodes, it would also be interesting to ask if genes containing CELE2 insertions are expressed differently from their orthologs in related species.

In summary, we have demonstrated that the organization of *cis*-regulatory elements in modules is conserved and can be used to predict *cis*-regulatory modules from patterns of co-occuring motifs. We can use conservation as an orthogonal source of information to design an algorithm similar to Flipper, for *cis*-regulatory module discovery. As a parallel study, we can use known motifs corresponding to characterized TFs in yeast, and ask if they co-occur in conserved patterns in promoters of co-expressed genes. We expect this analysis to report the PAC-RRPE pair as a regulatory module and hopefully uncover other novel organizational patterns corresponding to regulatory modules. We can extend a similar analysis to *C.elegans*, whereby motifs predicted by Flipper can be used instead of a library of known motifs. In doing so, we hope to uncover *cis*-regulatory modules associated with other sets of coexpressed genes other than RPs. It should be noted that we may have been successful at detecting conserved *cis*-regulatory modules for RPs because they represent a particularly specialized group of genes that have changed little over the course of evolution. Genes involved in adaptive responses are likely to be regulated in species specific fashion and our approach may be unsuccessful at detecting conserved organizational rules for *cis*-regulatory elements in these promoters. This shortcoming can be partially overcome by restricting our analysis to pairs of co-occuring motifs – orthologous promoters acquire differing numbers of motifs through evolution but functional pairwise interactions between motif pairs should remain conserved. As more data is collected from high throughput assays measuring gene expression changes and transcription factor binding, it is becoming more apparent that most changes in gene expression result from changes in *cis*-regulation[4,6] and is likely due to changes in *cis*-regulatory element organization. Understanding the *cis*-regulatory logic underlying this organization will undoubtedly be necessary for elucidating the effects changes in organization will have on gene expression. Our conservation based approach identifies functional organizational constraints for *cis*-regulatory modules in *C.elegans* RP promoters and we are confident that it can be extended to detect *cis*-regulatory modules in larger genomes, such as mouse and human.

Chapter 7: General Discussion

7.1 SUMMARY OF RESULTS

The overall goal of this dissertation is to demonstrate that functional aspects of *cis*-regulation (i.e. motifs, modules, organizational constraints) are under strong selective pressure and hence, are likely to be evolutionarily conserved across related species.

A recurrent theme that emerges from our analysis is that regulatory conservation is poorly measured by the degree of sequence identity in multiple alignment. We first encountered this issue in the first study, where we observed that TF binding site alignability did not correlate with binding site strength. Motivated by this observation, we developed an alignment-free motif detection algorithm, Flipper, that leveraged conservation by sampling groups of sites distributed across orthologous promoters. By testing on synthetic data, as well as *in vivo* binding data, we showed that Flipper outperformed alignment-based phylogenetic algorithms, especially if the compared species were sufficiently diverged so that binding site alignability was affected.

The difference between function and sequence alignability resurfaced in our promoter::mCherry transgenic experiments. We had mutated sequence elements predicted by Flipper to be motif sites and observed changes in mCherry expression. In comparison, we mutated control sequence elements that were more sequence alignable and observed no change in mCherry expression. These experimental results are consistent with our expectation from our computational simulations, that sequence alignability is an overall poor predictor of function.

Another key focus of this dissertation involves using conservation as a tool to discover mechanisms regulating the coexpression of ribosomal proteins (RPs) in *C.elegans*. While loss-of-function mutations in RP genes can result in embryonic lethality, defects in regulation can deplete cellular RP levels, cause translation related stress and result in dysregulation of apoptotic and developmental pathways[42]. We discovered 22 motifs associated with RP promoters and showed that 76 RP genes could be grouped into five clusters based on patterns of motif co-occurrence using the top four motifs. Using promoter::mCherry transgenic experiments, we confirmed the function of RP motifs in a subset of RP promoters. Furthermore, we demonstrated experimentally that our predicted motifs were indeed functionally conserved, by testing motifs in *C.briggsae* RP orthologs within a *C.elegans* regulatory context. Interestingly, our strongest motif, M546, was also functional in a non-ribosomal promoter, leading us to propose that M546 may be recognized by a general transcriptional regulator involved in growth and early development.

The final focus of our research involved applying conservation towards the discovery of *cis*-regulatory modules (CRMs). Specifically, we hypothesized that spatial and orientation biases displayed by *cis*-regulatory elements in a module should be conserved, because they reflect functional interactions between their respective regulators that are similarly conserved across species. Scoring for conserved organization constraints for our predicted RP motifs, we detected a strong spacing bias between M546 and M540 elements of 8/11 bp. In a "sequence swap" experiment, we verified that the 8 bp spacing between M546 and M540 elements was functional in the *rps-7* promoter, but the 11 bp spacing was not functional in the *rpl-2* promoter. We discovered that many non-ribosomal promoters contain M546 and M540 sites by virtue of a CELE2 transposon insertion. Interestingly, the terminal inverted region of CELE2 contains M546 and M540 elements spaced by 20/21 bp and non-ribosomal promoters with the insertion are not enriched for any common GO annotations. In contrast, non-ribosomal promoters without CELE2 insertion have their M546 and M540 sites spaced by 8/11 bp and are strongly enriched for growth and development GO annotations. It appears that CELE2 does not 'spread' M546-M540 associated regulation throughout the genome, presumably because the M546-M540 pair it carries does not obey the correct spacing constraint.

7.2 SIGNIFICANCE OF THESIS RESEARCH

The advent of next generation sequencing technologies has dramatically lowered the cost of whole genome sequencing and we fully expect the numbers of newly sequenced species to increase rapidly through the next few years, fueling the need for more sophisticated analyses in the field of comparative genomics. The research conducted in this dissertation joins a number of studies[36,119] in warning against literal interpretation of sequence alignability as a proxy for conserved function. It also offers alternative methods for predicting functionally conserved sequence elements, for example: identifying conserved motifs through their distribution in groups of orthologous promoters, identifying conserved CRMs by testing if organizational relationships between the component *cis*-regulatory elements are conserved. Flipper is a general bioinformatics resource that can be shared with other labs interested in discovering novel motifs in the promoters of their favourite genes.

As a proof-of-concept, we focused on investigating mechanism of ribosomal protein (RP) regulation in *C.elegans*. Our analysis demonstrated that there may be five or more different *cis*-regulatory modules present in the promoters of 76 RP genes, implying that the RP genes are under tighter and more complicated regulation than previously thought. Mutating M313 disrupted mCherry expression in a subset of tissue types, suggesting that RP expression may even be regulated at the tissue type level. We hope to apply similar approaches to study the cis-regulation of RPs in humans. Recent studies have shown that depleting RP levels in tissues can result in translationally-challenged phenotypes, that are relevant to human diseases such as Diamond-Blackfan Anemia (DBA) and even cancer[40,94,120]. Since RP expression levels have an immediate impact on the levels of RP proteins in the cell, identification of *cis*-regulatory elements involved in RP expression may help in the characterization of non-coding variants in RP loci that are associated with DBA and cancer related phenotypes. From an extended perspective, if *trans*-acting factors binding these *cis*-regulatory elements (and their respective upstream pathways) can be identified, they may be strong candidates for potential 'agonist' therapies, i.e. design compounds to stimulate or deinhibit these pathways in order to ramp up RP expression in translationally challenged patients.

Our findings regarding the *cis*-regulation of RPs also raise questions regarding the evolution of RP expression in nematodes and the evolution of *cis*-regulation in general. In a recent publication, Bourque et al[118] demonstrated that a substantial number of species specific TF binding sites are associated with transposable elements and proposed that transposons may have a role in spreading seed sites for TF binding motifs across the genome. This hypothesis has been previously proposed by McClintock and colleagues[121,122]. Our results on CELE2 and its seed sites for M546 and M540 draw interesting parallels: Instead of spreading M546/M540 associated regulation, non-ribosomal promoters containing CELE2 insertions are not enriched in any common GO annotations, perhaps because the M546-M540 seed pair does not satisfy the 8/11 bp spacing constraint. Since RP are expressed widely and at high levels, the effects of spreading RP associated *cis*-regulation to other promoters may cause unwanted ectopic or over-expression of downstream genes, which can result in severe cellular dysregulation. There may be a selective pressure on the CELE2 M546-M540 seed pair to retain its nonfunctional spacing constraint. Since CELE2 is specific to *C.elegans*, gene expression could be assayed for *C.elegans* and compared against related species, to determine to what extent acquiring a CELE2 insertion has affected the expression of *C.elegans* genes compared to their orthologs. This would provide further insight into how transposons have affected the evolution of *cis*-regulation.

7.3 FUTURE DIRECTIONS

Having demonstrated that aspects of *cis*-regulatory element organization in CRMs are conserved, a clear next step would be to incorporate this information into an algorithm that can search for CRMs with conserved organizational rules. One possibility would be to extend the Flipper algorithm, so that it samples not groups of sites distributed across orthologous promoters, but groups of CRMs distributed across orthologs, with the added constraint that component regulatory elements obey certain organizational rules characteristic to the CRM. CRMs could be assessed based on the sum of individual component motif MAP scores, so repeated sampling will select for CRM over-representation. Because sampled CRMs are constrained to be distributed across orthologs and have distinct organizational rules, the algorithm can be thought of as a constrained optimization process. To prevent the CRM search from being caught in local optima, we can parameterize spacing constraints with a geometric distribution, positioning and orientation constraints with a Bernoulli distribution etc. We can employ a method similar to EMCModule[106] to learn these parameters from the data within a Bayesian framework. A foreseeable concern with this approach is that it may be too strict in requiring perfect conservation of CRM structure across species, i.e. this approach would fail to detect a situation where two strong motifs co-occur in an anchor species promoter, but three weaker motifs occur in its corresponding CRM ortholog. A possible solution to this could involve modifying the algorithm to detect conserved pairwise motif interactions in CRMs, which could be extended to detect groups of motif interactions.

In terms of experimental directions, there remains much work to be done in order to fully characterize the transcriptional mechanisms regulating RP expression. We showed that the RP promoters follow five patterns of motif organization, but we have not determined the regulatory logic behind these patterns of motif co-occurrence. At first glance, we could create constructs where motifs are mutated according to all possible permutations of motifs in each pattern, but this would be too laborious and time consuming. A minimal promoter system might be an easier approach – where motifs are added sequentially in a minimal promoter until a particular combination is sufficient to recapitulate RP promoter driven mCherry expression.

More experiments need to be performed to strengthen the claim that the spacing constraint between M546 and M540 pairs is indeed functional *in vivo*. For example, we observed that the 11 bp spacing constraint was not functional in the *rpl-2* promoter, but this was a case where M540 was not functional either. An interesting experiment would be to reduce the 20/21 bp spacing between M546 and M540 in a non-ribosomal promoter to 8/11 bp. If the seed pair was truly non-functional because of the incorrect spacing, then 'restoring' function by changing the spacing would be a convincing argument for the function of the spacing bias. Other future experimental directions include performing a EMSA gel mobility shift assay to determine if any proteins are bound to probes encoding our sequence motifs, and if successful, performing a yeast[123] or bacterial[124] one-hybrid assay to determine possible *trans*-acting binding partners.

REFERENCES:

1. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
2. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).
3. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. *Nature* **430**, 85-8 (2004).
4. Coller, H. A. & Kruglyak, L. Genetics. It's the sequence, stupid! *Science* **322**, 380-1 (2008).
5. Jeong, S. et al. The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. *Cell* **132**, 783-93 (2008).
6. Wilson, M. D. et al. Species-specific transcription in mice carrying human chromosome 21. *Science* **322**, 434-8 (2008).
7. Clamp, M. et al. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* **104**, 19428-33 (2007).
8. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
9. Brenner, S. The genetics of Caenorhabditis elegans. *Genetics* **77**, 71-94 (1974).
10. Sulston, J. E. & Brenner, S. The DNA of Caenorhabditis elegans. *Genetics* **77**, 95-104 (1974).
11. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol* **100**, 64-119 (1983).
12. Ruvkun, G., Wightman, B., Burglin, T. & Arasu, P. Dominant gain-of-function mutations that lead to misregulation of the C. elegans heterochronic gene lin-14, and the evolutionary implications of dominant mutations in pattern-formation genes. *Dev Suppl* **1**, 47-54 (1991).
13. Varki A, C. R., Esko JD, Freeze HH, Stanley P, Bertozzi CR, Hart GW, Etzler ME. *Essentials of GlycoBiology* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY, 2009).
14. Stinchcomb, D. T., Shaw, J. E., Carr, S. H. & Hirsh, D. Extrachromosomal DNA transformation of Caenorhabditis elegans. *Mol Cell Biol* **5**, 3484-96 (1985).
15. Kelly, W. G., Xu, S., Montgomery, M. K. & Fire, A. Distinct requirements for somatic and germline expression of a generally expressed Caernorhabditis elegans gene. *Genetics* **146**, 227-38 (1997).
16. Praitis, V., Casey, E., Collar, D. & Austin, J. Creation of low-copy integrated transgenic lines in Caenorhabditis elegans. *Genetics* **157**, 1217-26 (2001).
17. Okkema, P. G. & Fire, A. The Caenorhabditis elegans NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Development* **120**, 2175-86 (1994).
18. Xue, D., Finney, M., Ruvkun, G. & Chalfie, M. Regulation of the mec-3 gene by the C.elegans homeoproteins UNC-86 and MEC-3. *Embo J* **11**, 4969-79 (1992).
19. Dupuy, D. et al. A first version of the Caenorhabditis elegans Promoterome. *Genome Res* **14**, 2169-75 (2004).
20. Badis, G. et al. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-3 (2009).
21. Gerasimova, T. I. & Corces, V. G. Chromatin insulators and boundaries: effects on transcription and nuclear organization. *Annu Rev Genet* **35**, 193-208 (2001).
22. Burcin, M. et al. Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Mol Cell Biol* **17**, 1281-8 (1997).
23. Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387-96 (1999).

24. Dunn, K. L., Zhao, H. & Davie, J. R. The insulator binding protein CTCF associates with the nuclear matrix. *Exp Cell Res* **288**, 218-23 (2003).

25. Dunn, K. L. & Davie, J. R. The many roles of the transcriptional regulator CTCF. *Biochem Cell Biol* **81**, 161-7 (2003).

26. Gerasimova, T. I., Lei, E. P., Bushey, A. M. & Corces, V. G. Coordinated control of dCTCF and gypsy chromatin insulators in Drosophila. *Mol Cell* **28**, 761-72 (2007).

27. Shrader, T. E. & Crothers, D. M. Artificial nucleosome positioning sequences. *Proc Natl Acad Sci U S A* **86**, 7418-22 (1989).

28. Wang, Y. H. & Griffith, J. D. The [(G/C)3NN]n motif: a common DNA repeat that excludes nucleosomes. *Proc Natl Acad Sci U S A* **93**, 8863-7 (1996).

29. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074-80 (2001).

30. Buck, M. J. & Lieb, J. D. A chromatin-mediated mechanism for specification of conditional transcription factor targets. *Nat Genet* **38**, 1446-51 (2006).

31. Segal, E. et al. A genomic code for nucleosome positioning. *Nature* **442**, 772-8 (2006).

32. Yuan, G. C. et al. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309**, 626-30 (2005).

33. Lee, W. et al. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* **39**, 1235-44 (2007).

34. Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P. & Fire, A. Z. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res* **16**, 1505-16 (2006).

35. Parker, S. C., Hansen, L., Abaan, H. O., Tullius, T. D. & Margulies, E. H. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324**, 389-92 (2009).

36. Odom, D. T. et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**, 730-2 (2007).

37. Rockman, M. V. & Stern, D. L. Tinker where the tinkering's good. *Trends Genet* **24**, 317-9 (2008).

38. Tishkoff, S. A. et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39**, 31-40 (2007).

39. Tournamille, C., Colin, Y., Cartron, J. P. & Le Van Kim, C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* **10**, 224-8 (1995).

40. Draptchinskaia, N. et al. The gene encoding ribosomal protein S19 is mutated in Diamond-Blackfan anaemia. *Nat Genet* **21**, 169-75 (1999).

41. Matsson, H. et al. Targeted disruption of the ribosomal protein S19 gene is lethal prior to implantation. *Mol Cell Biol* **24**, 4032-7 (2004).

42. Danilova, N., Sakamoto, K. M. & Lin, S. Ribosomal protein S19 deficiency in zebrafish leads to developmental abnormalities and defective erythropoiesis through activation of p53 protein family. *Blood* **112**, 5228-37 (2008).

43. Uechi, T. et al. Deficiency of ribosomal protein S19 during early embryogenesis leads to reduction of erythrocytes in a zebrafish model of Diamond-Blackfan anemia. *Hum Mol Genet* **17**, 3204-11 (2008).

44. Kongsuwan, K. et al. A Drosophila Minute gene encodes a ribosomal protein. *Nature* **317**, 555-8 (1985).

45. Marygold, S. J. et al. The ribosomal protein genes and Minute loci of Drosophila melanogaster. *Genome Biol* **8**, R216 (2007).

46. Saeboe-Larssen, S., Lyamouri, M., Merriam, J., Oksvold, M. P. & Lambertsson, A. Ribosomal protein insufficiency and the minute syndrome in Drosophila: a dose-response relationship. *Genetics* **148**, 1215-24 (1998).

47. Pavesi, G., Mauri, G. & Pesole, G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* **17 Suppl 1**, S207-14 (2001).

48. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**, 939-45 (1998).

49. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).

50. Thompson, W., Rouchka, E. C. & Lawrence, C. E. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* **31**, 3580-5 (2003).

51. Thijs, G. et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**, 1113-22 (2001).

52. Liu, X., Brutlag, D. L. & Liu, J. S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 127-38 (2001).

53. Hertz, G. Z. & Stormo, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563-77 (1999).

54. Tompa, M. et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**, 137-44 (2005).

55. Siddharthan, R., Siggia, E. D. & van Nimwegen, E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* **1**, e67 (2005).

56. Sinha, S., Blanchette, M. & Tompa, M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5**, 170 (2004).

57. MacIsaac, K. D. et al. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7**, 113 (2006).

58. Li, X., Zhong, S. & Wong, W. H. Reliable prediction of transcription factor binding sites by phylogenetic verification. *Proc Natl Acad Sci U S A* **102**, 16945-50 (2005).

59. Morgenstern, B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211-8 (1999).

60. Wang, T. & Stormo, G. D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19**, 2369-80 (2003).

61. Newberg, L. A. et al. A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics* **23**, 1718-27 (2007).

62. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368-76 (1981).

63. Ding, Y., Chan, C. Y. & Lawrence, C. E. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *Rna* **11**, 1157-66 (2005).

64. Yustein, J. T. & Dang, C. V. Biology and treatment of Burkitt's lymphoma. *Curr Opin Hematol* **14**, 375-81 (2007).

65. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185-98 (2004).

66. Moses, A. M., Chiang, D. Y. & Eisen, M. B. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac Symp Biocomput*, 324-35 (2004).

67. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).

68. Brudno, M. et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**, 721-31 (2003).

69. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-15 (2004).

70. Chan, C. S., Elemento, O. & Tavazoie, S. Revealing posttranscriptional regulatory elements through network-level conservation. *PLoS Comput Biol* **1**, e69 (2005).

71. Elemento, O. & Tavazoie, S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol* **6**, R18 (2005).

72. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338-45 (2005).

73. Elemento, O., Slonim, N. & Tavazoie, S. A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* **28**, 337-50 (2007).
74. Harbison, C. T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104 (2004).
75. McGhee, J. D. et al. The ELT-2 GATA-factor and the global regulation of transcription in the C. elegans intestine. *Dev Biol* **302**, 627-45 (2007).
76. Zhang, Y. et al. Identification of genes expressed in C. elegans touch receptor neurons. *Nature* **418**, 331-5 (2002).
77. GuhaThakurta, D. et al. Identification of a novel cis-regulatory element involved in the heat shock response in Caenorhabditis elegans using microarray gene expression and computational methods. *Genome Res* **12**, 701-12 (2002).
78. Whittle, C. M., Lazakovitch, E., Gronostajski, R. M. & Lieb, J. D. DNA-binding specificity and in vivo targets of Caenorhabditis elegans nuclear factor I. *Proc Natl Acad Sci U S A* **106**, 12049-54 (2009).
79. Li, X. Y. et al. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* **6**, e27 (2008).
80. Neuwald, A. F., Liu, J. S. & Lawrence, C. E. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* **4**, 1618-32 (1995).
81. Lawrence, C. E. et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208-14 (1993).
82. Fishman, G. *Monte Carlo: Concepts, Algorithms and Applications* (Springer, New York, 1996).
83. Liu, J. S. *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2004).
84. Robert, C. P., Casella, G. *Monte Carlo Statistical Methods* (Springer-Verlag, New York, 2004).
85. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160-74 (1985).
86. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* **100**, 11484-9 (2003).
87. Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G. & Brown, E. L. Genomic analysis of gene expression in C. elegans. *Science* **290**, 809-12 (2000).
88. Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L. & Hunter, C. P. Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development* **130**, 889-900 (2003).
89. McKay, S. J. et al. Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. *Cold Spring Harb Symp Quant Biol* **68**, 159-69 (2003).
90. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
91. Lieb, J. D., Liu, X., Botstein, D. & Brown, P. O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28**, 327-34 (2001).
92. Ward, L. D. & Bussemaker, H. J. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics* **24**, i165-71 (2008).
93. Uechi, T. et al. Ribosomal protein gene knockdown causes developmental defects in zebrafish. *PLoS One* **1**, e37 (2006).
94. Amsterdam, A. et al. Many ribosomal protein genes are cancer genes in zebrafish. *PLoS Biol* **2**, E139 (2004).
95. Hobert, O. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic C. elegans. *Biotechniques* **32**, 728-30 (2002).
96. Blow, J. J. & Dutta, A. Preventing re-replication of chromosomal DNA. *Nat Rev Mol Cell Biol* **6**, 476-86 (2005).

97. Singh, D. et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-9 (2002).

98. Brindle, P. K., Holland, J. P., Willett, C. E., Innis, M. A. & Holland, M. J. Multiple factors bind the upstream activation sites of the yeast enolase genes ENO1 and ENO2: ABFI protein, like repressor activator protein RAP1, binds cis-acting sequences which modulate repression or activation of transcription. *Mol Cell Biol* **10**, 4872-85 (1990).

99. Voutev, R., Killian, D. J., Ahn, J. H. & Hubbard, E. J. Alterations in ribosome biogenesis cause specific defects in C. elegans hermaphrodite gonadogenesis. *Dev Biol* **298**, 45-58 (2006).

100. Azuma, M., Toyama, R., Laver, E. & Dawid, I. B. Perturbation of rRNA synthesis in the bap28 mutation leads to apoptosis mediated by p53 in the zebrafish central nervous system. *J Biol Chem* **281**, 13309-16 (2006).

101. Davidson, E. H. & Levine, M. S. Properties of developmental gene regulatory networks. *Proc Natl Acad Sci U S A* **105**, 20063-6 (2008).

102. Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E. D. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* **3**, 30 (2002).

103. Schroeder, M. D. et al. Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol* **2**, E271 (2004).

104. Sinha, S., Liang, Y. & Siggia, E. Stubb: a program for discovery and analysis of cis-regulatory modules. *Nucleic Acids Res* **34**, W555-9 (2006).

105. Zhou, Q. & Wong, W. H. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A* **101**, 12114-9 (2004).

106. Gupta, M. & Liu, J. S. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci U S A* **102**, 7079-84 (2005).

107. Sun, H. et al. ModuleDigger: an itemset mining framework for the detection of cis-regulatory modules. *BMC Bioinformatics* **10 Suppl 1**, S30 (2009).

108. Xie, D., Cai, J., Chia, N. Y., Ng, H. H. & Zhong, S. Cross-species de novo identification of cis-regulatory modules with GibbsModule: application to gene regulation in embryonic stem cells. *Genome Res* **18**, 1325-35 (2008).

109. Nguyen, D. H. & D'Haeseleer, P. Deciphering principles of transcription regulation in eukaryotic genomes. *Mol Syst Biol* **2**, 2006 0012 (2006).

110. Boorsma, A., Lu, X. J., Zakrzewska, A., Klis, F. M. & Bussemaker, H. J. Inferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression. *PLoS One* **3**, e3112 (2008).

111. Pilpel, Y., Sudarsanam, P. & Church, G. M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* **29**, 153-9 (2001).

112. Spellman, P. T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**, 3273-97 (1998).

113. Gasch, A. P. et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**, 4241-57 (2000).

114. Surzycki, S. A. & Belknap, W. R. Repetitive-DNA elements are similarly distributed on Caenorhabditis elegans autosomes. *Proc Natl Acad Sci U S A* **97**, 245-9 (2000).

115. Agrawal, A., Eastman, Q. M. & Schatz, D. G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **394**, 744-51 (1998).

116. Miller, W. J., McDonald, J. F., Nouaud, D. & Anxolabehere, D. Molecular domestication--more than a sporadic episode in evolution. *Genetica* **107**, 197-207 (1999).

117. Volff, J. N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**, 913-22 (2006).

118. Bourque, G. et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18**, 1752-62 (2008).

119. Borneman, A. R. et al. Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815-9 (2007).
120. Lai, K. et al. Many ribosomal protein mutations are associated with growth impairment and tumor predisposition in zebrafish. *Dev Dyn* **238**, 76-85 (2009).
121. Mc, C. B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36**, 344-55 (1950).
122. Feschotte, C., Jiang, N. & Wessler, S. R. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* **3**, 329-41 (2002).
123. Deplancke, B. et al. A gene-centered C. elegans protein-DNA interaction network. *Cell* **125**, 1193-205 (2006).
124. Meng, X., Brodsky, M. H. & Wolfe, S. A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat Biotechnol* **23**, 988-94 (2005).

VITA

DONAVAN CHENG

Education:

Johns Hopkins University School of Medicine, Baltimore, MD
Ph.D, Biomedical Engineering 2004 -
2009

University of Toronto, Toronto, Canada
B.A.Sc, Engineering Science (Biomedical Engineering Option) 2000 -
2004

Awards:

NSERC Postgraduate PGSD3 Fellowship 2006 -

2009

NSERC Undergraduate Research Fellowship 2004

Bernard E. Etkin Medal of Excellence in Fluid and Solid State Dynamics 2003

Peer-Reviewed Publications:

Cheng DTS, Bansal N, Lee D, Beer MA. "Conserved motifs identified by a novel alignment-free algorithm control *C. elegans* ribosomal protein coexpression" *Submitted.*

Viswanathan S, Davey RE, Cheng D, Raghu RC, Lauffenburger DA, Zandstra PW. "Clonal evolution of stem and differentiated cells can be predicted by integrating cell-intrinsic and -extrinsic parameters." Biotechnol Appl Biochem. 2005 Oct;42(Pt 2):119-31