# Chapter 1: Introduction to the evolution of transcriptional regulation in eukaryotes

## transcriptional regulation in eukaryotes

### 1.1.1 Brief overview of eukaryotic transcription regulation

In complex organisms, the proper establishment and maintenance of developmental programs and cell/tissue physiology is largely dependent on the precise regulation of gene expression patterns at the level of transcription (Lee and Young 2000; Levine and Tjian 2003; Davidson 2006). Transcription of protein-coding gene products is highly-tunable. At the most basal level, it requires the assembly of the RNA polymerase II (RNA pol II) holoenzyme complex of 10 to 12 proteins, to a set of general transcription factors (GTFs) to form a pre-initiation complex (PIC) onto the core promoter region (≤ 100 bp) to orient and direct the start of transcription (Smale and Kadonaga 2003; Juven-Gershon et al. 2008). Transcriptional activity can then be modulated by the action of activators and repressors; typically by sequence-specific transcription factors (TFs) and transcription cofactors that bind TFs, that can affect the rate of recruitment and stability of the RNA pol II complex as well as modify the local chromatin structure to promote or limit access to the core promoter and other regulatory regions by binding to DNA outside of the core promoter (Lee and Young 2000). Hence, the complex and variable mosaic of genes expressed in any given cell type, point in development, or in response environmental stresses and physiological cues, is largely regulated through the physical interaction between TFs and their binding sites.

### 1.1.2 Structure and organization of promoters and *cis*-regulatory modules

Promoters can be defined by their distance from transcription start sites as core or proximal promoters. The core promoter encompasses the general region of DNA required for the recruitment and assembly of the PIC and is typically within a hundred bases to the site(s) from where transcription initiates, while the proximal promoter can be located up to several hundred base pairs away and contain binding sites for TFs (TFBSs) that interact with the general transcription machinery to modulate specificity of transcription (Wray et al. 2003; Sandelin et al. 2007). The term “distal promoter” has also been used by some to describe the region upstream of the proximal promoter that contains additional regulatory elements that operate in a position and orientation-dependent manner.

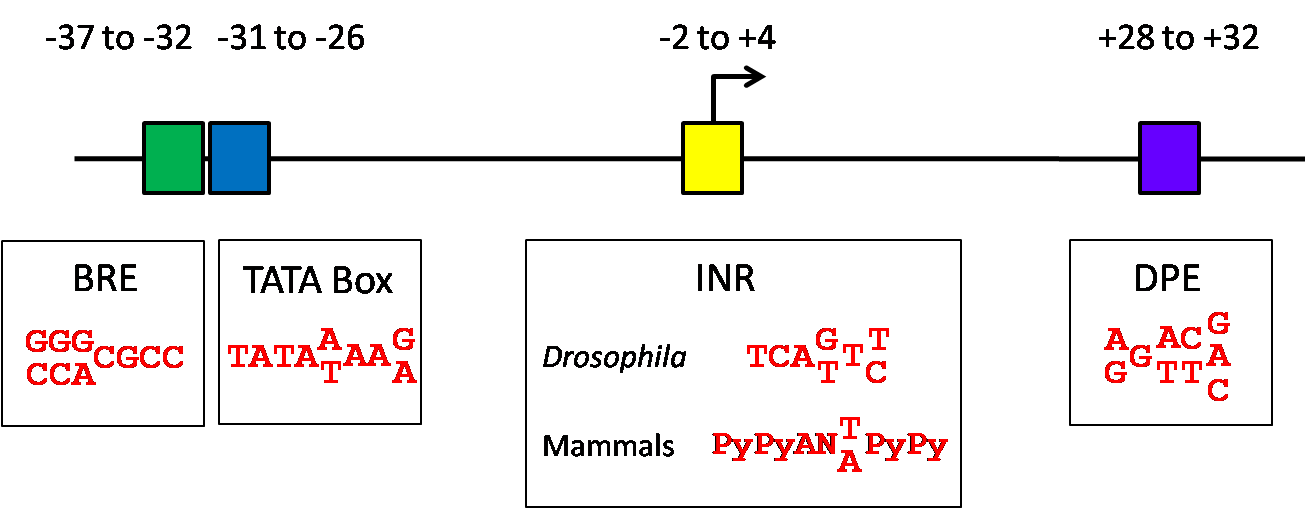


Figure 1.1 PROTOTYPICAL CORE PROMOTER MOTIFS (adapted from Smale and Kadonaga 2003).

Amongst the best studied core promoter elements is the TATA-box and its consensus TATAA sequence is found 25-30bp upstream of TSS in 10-20% of mammalian promoters (Gershenzon and Ioshikhes 2005). The TATA-binding protein (TBP), as its name suggests, binds to the TATA-box and is a component of TFIID complex, one of the GTFs in the PIC. Genes with promoters containing the TATA-box typically have tissue-specific expression (Schug et al. 2005) and a narrow distribution of TSSs (Carninci et al. 2006). The TATA-box often co-occurs with another core promoter element, the initiator sequence (INR), whose consensus YYANWYY overlaps with the TSS. These two sequence elements have been shown to be capable of independently recruiting the PIC and initiating transcription (Smale and Kadonaga 2003).

The BRE (TFIIB recognition element) is associated with a subset of TATA-containing promoters and has differing consensus sequences depending on its location relative to the TATA-box. The BRE can act to augment or dampen transcription in an unknown manner (Deng and Roberts 2007) and when found upstream (BREu) of the TATA-box, has a consensus of SSRCGCC and RTDKKKK when found downstream (BREd).

The downstream promoter element or DPE was originally described in *Drosophila* when purified FLAG-tagged TFIID was found to bind to both promoters containing and lacking the TATA-box (Burke and Kadonaga 1996). The TATA-less promoters contained a motif located +28 to +23 downstream of the TSS with a consensus of RGWYVT and acts synergistically with the INR element to recruit TFIID and promote PIC formation in a manner analogous to the TATA-box. Similarly, the MTE (motif ten element) which has a consensus of CSARCSSAAC in the +18 to +27 region, has also been reported to act cooperatively with the INR, even in lieu of a DPE or TATA-box. Although human basal transcription machinery has been shown to be able to activate transcription in TATA-less DPE- and/or MTE-containing promoters in *Drosophila* in a DPE-dependent or MTE-Inr dependent manner suggesting conservation of these mechanisms, the prevalence of these motifs and their consensus in humans is unknown (Burke and Kadonaga 1997; Smale and Kadonaga 2003).

It is important to note that none of these core promoter motifs or any other described motifs (Smale and Kadonaga 2003; FitzGerald et al. 2006; Juven-Gershon et al. 2008) are found in every promoter; instead, their frequencies between motifs and prevalence amongst different species are variable and not fully known. Moreover, the proteins that recognize and bind these motifs are ubiquitously-expressed, so the regulatory information embedded in the core promoter is typically not sufficient to modulate transcript levels or specify spatio-temporal transcription of genes and requires the additional combinatorial control of sequence-specific TFs.

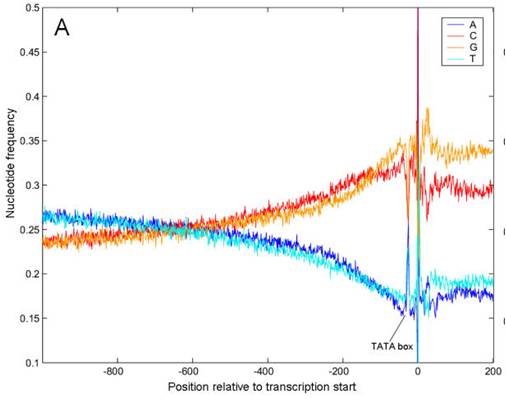


Figure 1.2 Nucleotide frequencies around the transcription start site in humans (Modified from Aerts et al, 2004)

In addition to a battery of motif combinations that define promoters, a characteristic feature of many vertebrate promoters is their association with CpG islands. The base content near the start of transcription has a distinctive shape marked by a rapid increase in G+C content (Aerts et al. 2003) (Figure 1.2). This pattern is correlated with the increased incidence of CpG dinucleotides which are normally found at less than a quarter of the expected frequency in the vertebrate genome (Gardiner-Garden and Frommer 1987). Non-methylated CpG dinucleotides are rare in the genome because the cytosine is typically methylated in vertebrates and spontaneous deamination of methylated cytosine results in thymine, resulting in a transition mutation (Bird 1980). The exceptions are CpG islands, which are stretches of DNA with relatively high CpG content over hundreds of bases that have somehow been protected from CpG methylation. CpG islands tend to be associated with promoters; in humans, promoters exhibit a bimodal distribution of CpG frequencies that divide into a high CpG class (72%) and a low CpG class (28%) associated with different properties (Saxonov et al. 2006). High CpG class promoters are associated with wide expression breadth and general housekeeping gene functions, while low CpG class promoters are more specific in tissue expression and function (Saxonov et al. 2006). The bimodal distribution of CpG frequencies in promoters has also been observed in other vertebrates but appears to be absent in invertebrates (Elango and Yi 2008). Hence, the presence of CpG islands is a good indicator of nearby genes generally in vertebrates, although it should also be noted that a non-negligible fraction of promoters have low CpG content.

In the simple model eukaryote *Saccharomyces cerevisiae* (Baker’s yeast), PIC formation and RNA pol II recruitment to specific promoters is dependent on the recognition and binding of TFs to the upstream activating sequences (UAS). The UAS is a region typically located several hundred bases upstream of the TSS that contains several binding sites for one or two TFs (Levine and Tjian 2003). In contrast, metazoans possess a more complex regulatory architecture in addition to the promoter region discussed above, involving multiple sequence structures called *cis*-regulatory modules (CRMs), each controlling a subset of the total catalogue of gene expression patterns at specific points in time and space.

Unlike promoters which are located just 5’ of genes and serves as a site for transcription machinery and initiation, CRMs are thought to function in a location and orientation-independent manner, and hence can be located far away from the gene(s) they act upon. CRMs are generally thought to encode the regulatory instructions that modulate the transcript levels of genes by altering chromatin structure and accessibility and/or recruiting transcriptional machinery to promoters. Each gene is controlled by multiple CRMs such that they are estimated to outnumber genes 5 to 10 fold (Davidson 2006). These modules are typically several hundreds of bases in length and contain clusters of TFBSs for on average 4 to 8 different TFs and function in a cell type-specific manner (Levine and Tjian 2003; Davidson 2006) (Figure 1.3). CRMs can be broadly classified as enhancers when they activate gene expression; insulators when they act as boundary elements to prevent inappropriate enhancer usage and silencers when they silence expression of nearby genes. Historically, these regulatory elements were identified by transgenic reporter assays; either for sequences that can activate transcription of the reporter regardless of position or orientation relative to the promoter (enhancers); or for sequences when placed between adjacent regulatory elements in a position-dependent manner, can block their function (enhancer-blocking insulators); or can shield the position effects of spreading heterochromatin (barrier insulators) (Bell et al. 1999; Phillips and Corces 2009). Several types of proteins interact with enhancer elements to facilitate their function, including chromatin remodelers that can re-position nucleosomes in an ATP-dependent manner (e.g. SWI/SNF complexes) as well as those that covalently modify the N-terminal tails of histones (e.g. histone acetyltransferases (HATs) and deacetylases) to decondense chromatin (Kingston and Narlikar 1999) and “mediator” complexes that can bridge sequence-specific TFs and the general transcription machinery (Heintzman and Ren 2009). These proteins/protein complexes are generally recruited by the activation domains of TFs as transcriptional co-activators. Meanwhile, the establishment of insulator domains in vertebrates genome-wide seems to require binding of CTCF, a highly-conserved and ubiquitously-expressed protein containing 11 zinc-fingers (Kim et al. 2007; Xie et al. 2007). Interestingly, it was recently observed that the intersection of CTCF-binding sites by ChIP-seq in different cell types was quite large, suggesting that most insulators act in a cell-type invariant manner (Cuddapah et al. 2009; Heintzman et al. 2009). In contrast, enhancers are

J:\Thesis\figures\chapter1\enhancer_model_visel_09.tif

Figure 1.3 (a) An example gene expressed in both the brain and limb is controlled by distant cis-regulatory sequences, but identification of these elements and their spatio-temporal activities (question marks) is challenging. (b) and (c) Tissue-specific CRMs containing binding sites for different TFs become active when all the required factors are expressed in the same tissue. TF binding and association with transcriptional co-activators is thought to enable CRM association with the promoter to activate transcription by RNA Polymerase iI (likely by looping). insulators prevent improper interactions between the enhancer and promoter (Visel et al, 2009).

### 

characterized by cell-type specific histone modification signatures (Heintzman et al. 2009) and enhancer-associated protein binding (e.g. HATs) (Visel et al. 2009). Irrespective of cell-type specificity, these regulatory sequences have a tendency to be located far away from genes, up to tens or hundreds of kb away from their targets and are thought exercise their function by forming chromatin loops allow direct physical interaction between distal regulatory elements and their target genes (enhancers) or between insulators (Miele and Dekker 2008). Although it has long been speculated that distal *cis*-regulation requiresloop formation and it is tempting to imagine how looping between insulators (perhaps through cohesin (Parelho et al. 2008; Wendt et al. 2008)) can occlude intervening regulatory elements or promote the recruitment and/or sequestering of transcription machinery to target genes as “transcription factories”, the precise mechanisms of long-range regulation remain elusive. Emerging technologies to capture physical interactions between genomic loci show promise in resolving higher-order chromatin interactions (Dekker et al. 2002; Dostie et al. 2006; Zhao et al. 2006) and is discussed later within this chapter. PIC formation and local remodelling of chromatin for access by transcriptional machinery. Transcriptional repressors are TFs that can repress transcriptional output, for example through competing with activators and occluding their binding sites or binding to silencer DNA elements to inhibit transcription, or binding to insulator DNA elements to block the effects of enhancers (Lee and Young 2000).

TFs typically contain at least a DNA-binding domain (e.g. homeodomain, SANT domain) or several (e.g. C2H2 zinc fingers), and an independent interaction domain to mediate protein-protein interactions. Some TFs additionally contain a nuclear localization signal (NLS) and/or a ligand-binding domain, such as those found in the nuclear receptor family, and their activity can be modulated by the binding of specific steroid hormones (Wray et al. 2003). TFs can be classified into families on the basis of the DNA-binding domain(s) they contain, which themselves are defined by their structure, primary sequence and mode by which they interact with DNA. Although the number of TFs can fluctuate dramatically from genome to genome, the repertoire of DBDs is relatively small. Most eukaryotic TFs contain one or more helix-turn-helix, helix-loop-helix, zinc finger or leucine zipper-type domains (Luscombe et al. 2000) through which sequence specific protein-DNA contacts are mediated (5-15 bp), typically in the major groove and sometimes with additional DNA contacts made by adjacent structures in the minor groove (Pabo and Sauer 1984). Detailed discussions of DBD structural groups, domain members and their mode of binding can be found elsewhere (Luscombe et al. 2000; Talukder 2007).

Specific recognition of target sites by TFs is a necessary component of the transcription process and is achieved within the cell somehow, despite the presence of on the order of 109 alternative sites present within eukaryotic genomes to sift through. Moreover, many TFs are paralogous members belonging to large families and bind similar consensus sequences but yet, are generally not functionally redundant. For instance, the Forkhead (Fox) family of TFs contain a highly-conserved “winged helix” domain and bind to a core RYMAAY consensus (Tsai et al. 2006); yet the members of this family generally have diverse expression patterns and participate in a wide variety of functions. At least 19 forkhead subfamilies have been identified and implicated in roles as varied as liver development (FoxA1-2), eye organogenesis (FoxC1-2), ovarian development (FoxL2), stress response, cellular proliferation and longevity (FoxO) and language development (FoxP2) (Tuteja and Kaestner 2007b; Tuteja and Kaestner 2007a), implying their tight and specific regulation.

It is likely that cells have evolved multiple non-exclusive strategies to achieve the required level of TF-DNA specificity. One way to accomplish this is through the regulation of the TFs themselves, such as restricting the expression of competing TFs and/or co-factors to different time-points or cell-types and/or adjusting their concentrations to favour binding of particular factors over others. However, perhaps the most common way to introduce specificity is through protein-protein interactions via homo- and heterodimerization of TFs, as well as interactions with co-factors (Amoutzias et al. 2008). Dimerization of TFs increases the length of the corresponding recognition sequence, with particular spacing and orientation requirements (e.g. sites in tandem or as palindromes), limiting the likelihood of spurious hits. Additionally, varying arrangements of binding sites and spacing may encourage cooperative binding of dimers of one TF while excluding binding of monomers of other TFs (Georges et al. 2009).

### 1.1.5 Challenges in identifying and defining TF and CRM interactions

Piecing together the interplay between TFs and their target genes within different cell-types and developmental time-points into coherent gene regulatory networks (GRNs) remains a formidable challenge. In the following sub-chapters, I discuss a number of experimental and computational techniques which can enable the collection of an assortment of “parts lists” from which node to edge relationships (TF to target gene or CRM) within GRNs can be inferred. However, functional characterization of these relationships is perhaps the largest bottleneck impeding the progress of GRN assembly as functions of the *cis-* and *trans-*regulatoryare components are context-dependent and difficult to experimentally assay.

## 1.2 Methods for identifying binding preferences of transcription factors

The importance in understanding TF function in the context of different gene regulatory programs cannot be overstated and roles for TFs have been implicated in many human diseases (Vaquerizas et al. 2009). Yet, most TFs remain uncharacterized (Vaquerizas et al. 2009), underscoring the need to collect the pieces of information crucial to defining TF function, namely, where TFs bind in the genome and what their target genes are under different spatial and temporal contexts. Advances in experimental methods have facilitated the study of protein-DNA interactions with greater ease and wider scope and the following section will describe some of the major techniques in use today.

### 1.2.1 Chromatin immunoprecipitation coupled with microarrays and/or high-throughput

### sequencing (ChIP-chip and ChIP-seq)

Chromatin immunoprecipitation coupled with microarray analysis and/or high-throughput sequencing is a commonly used *in vivo* TF-centric method for detecting genome-wide TF occupancy. ChIP-chip and ChIP-seq experiments typically begin with the fixing of the protein of interest to their DNA substrates within cells by the addition of crosslinking agent such as formaldehyde. The bound genomic DNA is then fragmented by sonication or enzymatic digestion and selected by immunoprecipitation using a specific antibody against the protein. Depending on the downstream detection method, the purified ChIP-DNA can be amplified, fluorescently-labelled and hybridized to microarrays containing genomic DNA along with the total non-crosslinked genomic DNA control (ChIP-chip), or directly sequenced by next-generation sequencing technologies such as Illumina sequencing (Bentley et al. 2008), Roche 454 pyrosequencing (Margulies et al. 2005) or SOLiD ABI sequencing (Shendure et al. 2005) (ChIP-seq). Array probes displaying a significant enrichment of ChIP to control fluorescent signal (ChIP-chip) or non-repetitive genomic regions with enriched read counts (ChIP-seq) correspond to regions likely harbouring binding sites for the interrogated protein.

Although ChIP-based methods are advantageous in that protein-DNA interactions are studied under physiological conditions and do not require *a priori* knowledge of the protein-DNA footprint, they also suffer from several experimental drawbacks and limitations. For instance, the success of ChIP-based experiments hinges upon the availability and specificity of an antibody against the protein of interest and although epitope tags can be introduced, it is often difficult to do so without disrupting expression levels, protein folding, or biochemical activity. Furthermore, indirect protein-DNA interactions can be mistakenly identified through formaldehyde-crosslinking of proteins to one another in addition to protein-DNA crosslinks. Nevertheless, ChIP-based methods have been successfully applied for example, to study the genomic occupancy of most of the TFs in yeast (Harbison et al. 2004) and to locate promoters, enhancers and insulators in the human genome (Kim et al. 2005; Heintzman et al. 2007; Kim et al. 2007; Heintzman et al. 2009).

Although the enormity of vertebrate genomes make ChIP-chip somewhat cost-prohibitive to sufficiently tile probes across the genome, the emergence and lowering costs of high-throughput sequencing offer a viable alternative. Visel et al. (Visel et al. 2009) recently demonstrated the utility of ChIP-seq for enhancer discovery by determining the occupancy of p300, an acetyltransferase and transcriptional co-activator found in most enhancer-associated protein complexes. 87% (75 out of 86) of a randomly-selected sample of random genomic regions associated with p300 peaks within embryonic mouse forebrain, midbrain and limb tissue each showed reproducible tissue-specific expression of the *LacZ* reporter in E11.5 embryos. Furthermore, up to 91% of the p300 occupied regions overlap sequence predicted to be under evolutionary constraints across vertebrates, supporting the notion that the ChIPed regions are of functional importance. Hence, ChIP-chip and ChIP-seq provide a powerful means for delineating the spatio-temporal activity of both *trans* and *cis*-regulatory components *in vivo*.

### 1.2.2 *In vitro* selection (SELEX)

SELEX stands for systematic evolution of ligands by exponential enrichment and is one of the most commonly used *in vitro* methods to investigate the binding behaviours of proteins. Starting with a large library of random oligomers, the target protein is immobilized on a sepharose column and incubated with the library to partition bound and unbound oligomers through iterative rounds of selection, elution and amplification of bound materials to enrich for sequences with the highest affinity and specificity for the target (Oliphant et al. 1989).

### 1.2.3 One-hybrid methods

Yeast (Y1H) and bacterial one-hybrid (B1H) methods are alternative techniques for detecting protein-DNA interactions. Conceptually similar, both methods can make use of a library of reporter constructs linked to randomized binding sites (baits). The protein of interest is expressed as a fusion protein (prey) containing the GAL4 activation domain (Y1H) or the α-subunit of RNA polymerase (B1H) (Meng et al. 2005). Binding events between prey and baits are identified by the expression of the reporter gene. The B1H system has the additional advantage of generating highly complex binding site libraries (limited by the generation of unique clones) thanks to its high transformation efficiency relative to yeast although some eukaryotic proteins may not fold properly in bacteria (Meng and Wolfe 2006).

### 1.2.4 DNA immunoprecipitation with microarray detection (DIP-chip)

The widespread availability of microarrays has ushered in a new wave of rapid, array-based *in vitro* methods for inferring binding specificities, and among the first is a method known as DIP-chip (DNA immunoprecipitation with microarray detection) (Liu et al. 2005). Unlike ChIP, the DIP-chip method relies upon epitope-tagging the protein of interest and is not dependent on antibody availability. The tagged-protein is allowed to mix with sheared genomic DNA from the organism under study before affinity purification of the bound protein-DNA complexes. Subsequent amplification, fluorescent labelling and hybridization of the eluted bound DNA with differentially labelled total genomic DNA allows for detection by microarrays containing the genome of interest. In contrast to SELEX, a single-selection step avoids over-selection and the use of long genomic fragments (~600bp) promotes binding in the natural sequence context. However, larger genomes cannot be accommodated on a single array at current probe densities and moderate to small-sized genomes may suffer from insufficient coverage or low resolution. Currently, DIP-chip is most suited for study of protein binding to yeast-sized genomes or smaller.

### 1.2.5 Cognate site identifier (CSI)

As microarray technologies continually improve to facilitate a growing number of probe sequences, several groups have designed array-based platforms aimed at interrogating binding specificities of DNA-binding proteins and molecules across the entire sequence space up to length *k* in an unbiased manner. These techniques in principle should provide the means to obtaining a rank-ordered list of binding affinities to all possible sequences of length *k* for any protein or molecule of interest. One such example is the Cognate Site Identifier (CSI) array (Warren et al. 2006), comprising of over 100,000 self-complimentary palindromic sequences, designed to form hairpins displaying all possible 8-base sequences (about the typical length of a TFBS) in quadruplicate. As a proof-of-principle, Warren *et al.* (Warren et al. 2006) demonstrated the ability of the CSI technique to query the binding preferences of synthetic molecules of known specificity as well as a well-studied *Drosophila* TF Exd, showing the high-intensity binding of all variations of the known consensus sequences was highly reproducible between replicate assays. CSI assays provide a read-out of not only the rank-order preferences of a DNA-binding ligand to every possible 8-mer sequence, but also provide information on the relative contributions of each base at each position of a motif, including some flanking positions. This universal approach to comprehensively profile the binding affinity landscape of DNA-binding ligands and others like it such as protein binding microarrays (Bulyk et al. 2001; Mukherjee et al. 2004; Berger et al. 2006), should allow for the rapid collation of high-quality affinity data towards understanding how these ligands interpret the genome within cells.

### 1.2.6 Protein binding microarrays (PBMs)

Current iterations of protein binding microarrays (Berger et al. 2006), like CSI arrays, are designed to sample the entire sequence space in a compact and comprehensive manner. Using a design to maximally cover every 4k k-mer sequence in an overlapping manner known as the de Bruijn sequence, every 10-mer can be covered at least once and every 8-mer at least 32 times (palindromic 8-mers at least 16 times) by probes consisting of a common primer sequence and subsequences of the de Bruijn sequence on custom Agilent microarrays. This increased sequence coverage allows for robust measurements of relative binding affinities, yielding highly-reproducible sequence preferences, even on arrays bearing an independent de Bruijn sequence design. As probe densities on microarrays increase, these combinatorial sequence designs can be easily expanded to accommodate larger values of k, as well as house multiple independent de Bruijn sequence designs within a single array (Mintseris and Eisen 2006).

### 1.2.7 TF Binding affinity representations

Whether *in vivo* or *in vitro* methods are used to define TF binding sequences, appropriate models are required to represent their sequence binding preferences. A concise model is that of a consensus sequence, created by aligning binding sites for a TF and selecting a representative nucleotide (including IUPAC degenerate nucleotide symbols) at each position in the alignment. Position weight matrices (PWMs) improve on this simple representation by allowing for biases in base preferences at each position to be quantified (Stormo 2000). A position frequency matrix (PFM) is constructed first by calculating the fraction of each nucleotide at each position within the sequence alignment, before dividing by the nucleotide frequencies of the background distribution (e.g. the genomic frequency) and log-transformed to yield a PWM. A PWM can then be used to score sequences for matches by summing the scores at each position, resulting in a score describing the relative binding energy of the protein-DNA interaction (Berg and von Hippel 1987; Stormo 2000). However, using a matrix in this manner makes the assumption that each position within the binding site contributes independently to the overall binding activity such that the binding of one nucleotide does not affect the likelihood of another nucleotide binding in neighbouring positions. Although there is evidence of non-independence amongst positions in experimentally-determined binding affinity data (Bulyk et al. 2001; Man and Stormo 2001; Bulyk et al. 2002; Maerkl and Quake 2007; Badis et al. 2009), these additive matrix models are for the most part satisfactory approximations of the measured affinities (Benos et al. 2002). Others have developed more complicated models to capture dependencies between binding site positions using Markov models (Zhou and Liu 2004; Zhao et al. 2005; Sharon et al. 2008) and Bayesian networks (Ben-Gal et al. 2005; Pudimat et al. 2005) but none of these approaches have usurped the matrix model as the *de facto* motif representation; in part due to the unwieldy underlying mathematical concepts and lack of intuition for the average biologist to use, as well as the difficulty of connecting these alternative motif models to existing motif-scanning and scoring paradigms.

## 1.3 Strategies for identifying CRMs

The sparse distribution and short and degenerate nature of TFBSs make them very difficult to identify due to numerous random, spurious matches arising simply by chance (Wasserman and Sandelin 2004). Interestingly, a recent information theory-based analysis considering hundreds of experimentally-determined motifs suggests that eukaryotic TFBSs generally lack sufficient bits of information to specify a regulatory region within the genome, given its size (Wunderlich and Mirny 2009). The ratio to spurious sites to cognate sites would be too large to distinguish regulatory sequence by inspection alone, a phenomenon coined as the “futility theorem” by Wasserman and Sandelin (Wasserman and Sandelin 2004). Based on a hypothesis that clusters of spurious sites would be less likely to form by chance, Wunderlich and Mirny (Wunderlich and Mirny 2009) found that in any 1 kb region, the presence of 10 to 20 sites from 3 to 10 different TFs would be sufficiently informative to provide the necessary specificity for eukaryotic TFs to recognize their cognate sites. Hence, the arrangement of binding sites into clusters could theoretically be sufficiently informative to distinguish regulatory regions from the genomic background and many approaches aimed at finding CRMs discussed in the following section leverage this observation.

### 1.3.1 Comparative genomics to identify conserved non-exonic sequence

Comparative genomics relies on the underlying principle that the studied sequences are derived from a common ancestor and should reflect a combination of the ancestral sequence and the action of evolution. Genomes are shaped by a variety of evolutionary forces, each with distinguishing signatures. Functional sequences tend to be selectively constrained by negative (purifying) selection to evolve more slowly than sequences evolving under neutral drift, while positive (Darwinian) selection drives rapid change to a favoured state of increased fitness. These signatures come to light through comparative sequence analysis and indeed, some of the first examples of *cis*-regulatory sequences were identified in this manner. Emorine and colleagues (Emorine et al. 1983) sequenced and compared the human, mouse and rabbit immunoglobulin kappa light chain genes and identified a highly conserved 130bp region within an intron which was predicted and later confirmed to be an enhancer (Picard and Schaffner 1984). Such islands of conservation are known as “phylogenetic footprints” preserved over evolutionary time.

The sequencing of the human (Lander et al. 2001) and mouse genomes (Waterston et al. 2002) in the early 2000s paved the way for whole genome comparisons and allowed for the proportion of the mammalian genome under evolutionary selection to be estimated for the first time. Using ancestral repeat sequences and four-fold degenerate codon positions to infer the neutral substitution rate, the proportion of the genome that exceeded the level of conservation than would be expected given the neutral rate was estimated to be around 5% (Waterston et al. 2002). Considering that protein-coding genes comprise of merely 1.5% of the genome and another 1% accounting for UTRs (Waterston et al. 2002), characterization of the remaining 2.5% of conserved sequence of unknown function is of exceptional interest. As of February 2010, 1194 genomes have been sequenced and published, 123 of them from eukaryotes (Liolios et al. 2009), providing ample raw materials for genome comparisons, but which species should one compare and how many?

In general, aligning multiple species (>2) offers greater resolving power than pairwise comparisons alone, and allows for delineation of the ancestral state at each base using an outgroup. If the species are too closely related, functionally important sequences remain embedded amongst other parts of the genome due to lack of opportunity for neutral drift to separate them. On the other hand, if the species are too distantly related, genomes may have diverged too much to detect signals of functional elements (Boffelli et al. 2004). Of course, when phylogenetic footprinting is used, an implicit assumption is made about the common regulatory mechanisms in use across orthologous genes and where that is not true, closer species comparisons would be required.

The studies by Cliften et al. (Cliften et al. 2001; Cliften et al. 2003) and Kellis et al. (Kellis et al. 2003) are excellent examples of leveraging multiple alignments of four closely-related “sensu stricto” yeast species to enumerate sequence motifs conserved between them, resulting in 79 and 72 ((Cliften et al. 2003; Kellis et al. 2003), respectively) known and novel motifs, many of which have similar functional annotations orare found upstream of co-expressed genes or are bound by the same transcription factors. Cliften et al. (Cliften et al. 2003) additionally used two relatively more distant yeast species (*S. castellii*and *S. Kluyveri*) further distinguish and refine negatively-selected conservation signals from merely similarity due to recent shared ancestry.

In vertebrates however, comparative genomics is complicated by larger genomes, non-uniform mutation rates across different genomic regions and lineages (Waterston et al. 2002) and complex genome rearrangement events. However, several approaches have proven to be useful in detecting potential conserved regulatory sequences: extreme sequence conservation between closely-related species (Bejerano et al. 2004) and sequence conservation across species separated by extreme evolutionary distances (Aparicio et al. 1995; Woolfe et al. 2005; Pennacchio et al. 2006). These methods impose a strict requirement for sequence conservation, either for long stretches of exact sequence identity (Bejerano et al. 2004) or high conservation maintained over long divergence times (Aparicio et al. 1995; Woolfe et al. 2005; Pennacchio et al. 2006), such that recovered sequences are likely highly evolutionarily constrained. 481 “ultraconserved” elements (UCEs) with perfect sequence identity over 200 bp across human, mouse and rat (Bejerano et al. 2004) and 1373 conserved non-coding sequences between human and Fugu averaging 84.3% sequence identity over 100 bp (Woolfe et al. 2005) were identified by some of these methods. Interestingly, these elements are enriched amongst genes important in transcription regulation and development (“trans-dev” genes) and many show evidence of driving tissue-specific reporter expression (Woolfe et al. 2005; Pennacchio et al. 2006), including 61% (33/54) of human-Fugu elements that were also UCEs. While these methods appear to capture many *bona fide* regulatory sequences capable of driving specific expression patterns, they are highly specific for identifying sequences with uniquely rigorous constraints, likely important to preserve gene expression programs important in vertebrate development. Moreover, they generally lack sensitivity for sequences under less constraint; a general symptom of current sequence alignment methods. Additionally, lineage-specific gains and losses of regulatory sequence would elude detection and it should be noted that not all conserved sequence is functional and not all functional sequences are conserved (Elgar and Vavouri 2008).

### 1.3.2 Histone modification patterns and chromatin signatures as markers of enhancers and active transcription

DNA in cells is packaged into chromatin, which is comprised of octamers of core histones (H3, H4, H2A and H2B) wrapped with 147 base pairs of DNA (nucleosome) and intervening linker DNA. The core histones each possess an unstructured N-terminal “tail” which extends outwards from the chromatin fibre and can be covalently modified at particular residues (Luger and Richmond 1998). Recently, the existence of a “histone code” has been postulated (Strahl and Allis 2000), reflecting the diversity and complexity of histone marks associated with processes such as transcription (Zhang and Reinberg 2001; Kim et al. 2005; Pokholok et al. 2005; Xu et al. 2005), chromatin condensation (Wei et al. 1999; Roh et al. 2005), DNA repair and replication (Bradbury 1992; Aggarwal and Calvi 2004). Indeed, work from the Ren lab and others have shown that promoters and enhancers can be distinguished by specific chromatin signatures in humans (Kim et al. 2005; Heintzman et al. 2007; Wang et al. 2008; Heintzman et al. 2009; Hon et al. 2009). Although the complete repertoire of distinct chromatin signatures has yet to be catalogued, promoters are generally marked by trimethylation of histone H3, lysine 4 (H3K4me3) and correlates with gene expression and trimethylation of histone H3 lysine 27 (H3K27me3) correlates with gene repression, while enhancers are often marked with monomethylation of histone H3 lysine 4 (H3K4me1) and acetylation of histone H3 lysine 27 (H3K27ac) (Barski et al. 2007; Heintzman et al. 2007; Heintzman et al. 2009). Interestingly, Heintzman et al. (Heintzman et al. 2009) showed that enhancer signatures were mostly cell-type specific while promoter signatures and CTCF occupancy were not, and further predict the existence of 105 to 106 enhancers within the human genome.

### 1.3.3 Nuclease protection and crosslinking patterns as markers of open chromatin

The packaging of DNA into chromatin within eukaryotic cells generally renders most of the genome (including regulatory sequences) inaccessible by wrapping of nucleosomes, except in regions where local remodelling of chromatin structure displace and/or evict nucleosomes, such as for the activation of promoters for transcription (Jiang and Pugh 2009). In yeast, promoters typically have low nucleosome occupancy resulting in regions termed “nucleosome-free regions” (NFRs) which allows for preferential access by TFs, regardless of transcriptional activation (Bernstein et al. 2004; Lee et al. 2004; Sekinger et al. 2005). NFRs are also found in human promoters, but generally associated only with active promoters (Kim et al. 2005; Ozsolak et al. 2007). The regions where the local chromatin conformation is more “open”, tend to be particularly hypersensitive to nuclease digestion and inefficient at forming protein-DNA crosslinks (Gross and Garrard 1988; Giresi et al. 2007; Giresi and Lieb 2009).

It was first noted by Wu et al. (Wu et al. 1979) the existence of regions of chromatin in *Drosophila* that were particularly sensitive to digestion by the DNase I enzyme, with cleavage patterns distinct from that of micrococcal nuclease which produces nucleosome and multi-nucleosome-sized fragments. Keene et al. (Keene et al. 1981) subsequently mapped DNase I hypersensitive regions to the 5’ end of the four heat-shock genes (hsp 28, hsp 26, hsp23 and hsp22), at the 67B locus in *Drosophila* and hypothesized that the 5’ location of DNase I hypersensitive sites (DHSs) relative to genes may be generally associated with the active transcription of genes. Evidence for the association of hypersensitive sites with enhancer or promoter elements is now plentiful (Crawford et al. 2004; Sabo et al. 2004a; Sabo et al. 2004b; Crawford et al. 2006; Birney et al. 2007; Boyle et al. 2008), although the exact mechanisms used to mobilize nucleosomes at each promoter remains unknown. Chromatin remodelling complexes such as conserved members of the SWI/SNF family use ATP to facilitate movement of nucleosomes (Carlson and Laurent 1994; Kwon et al. 1994) and direct competition by regulatory proteins for particular DNA sequences may also destabilize nucleosomes. Regardless of the mechanism, genome-wide maps of open chromatin should enable indexing of active regulatory elements across multiple cell-types, tissues and timepoints and compliment other data types such as ChIP-chip/ChIP-seq and gene expression profiling towards understanding gene regulation complexity.

A number of new technological advances now allow for genome-wide interrogation of DHSs without resorting to cumbersome and low resolution Southern blots. DNase-chip (Crawford et al. 2006; Sabo et al. 2006) and DNase-seq (Boyle et al. 2008; Hesselberth et al. 2009) methods both increase the throughput and coverage over Southern blots by exploiting the coverage afforded by tiling microarrays and high-throughput sequencing. Regions preferentially digested by DNase I digested can be enriched by ligation of biotinylated adapters and captured on a strepadvidin column (Crawford et al. 2006) or size fractionated by a sucrose gradient (Sabo et al. 2006) or directly sequenced (Hesselberth et al. 2009). Approximately 16 to 46.5% of DHSs (depending on cell type) identified were proximal to TSSs, and coincided well with regulatory features denoted by Pol II and CTCF ChIPed regions as well as enriched binding of sequence specific transcription factors (Birney et al. 2007; Boyle et al. 2008), demonstrating the effectiveness of the approach.

A class of complimentary techniques to that of nuclease hypersensitivity was born out of modifications of the ChIP procedure, using crosslinking with formaldehyde to preserve histone-DNA interactions. Because the majority of crosslinkable interactions in the nucleus are between histones and DNA, chromatin complexes dominate the crosslinking profile and provide a snapshot of the current chromatin state *in vivo* (Brutlag et al. 1969; Solomon and Varshavsky 1985). Crosslinking occurs more frequently in closed chromatin than in open chromatin and this difference in efficiency can be exploited to enrich for DNA in open chromatin. Formaldehyde-assisted isolation of regulatory elements (FAIRE) (Giresi et al. 2007; Giresi and Lieb 2009) does so by sonication and separation of histone-complexed DNA and free DNA by phenol-chloroform extraction. A related technique, Sono-seq (Auerbach et al. 2009) is premised on the idea that crosslinked chromatin is more prone to breakage by sonication in regions where nucleosome occupancy is relatively low. Like FAIRE, Sono-seq relies upon sonication of crosslinked chromatin, reversal of crosslinks prior to phenol-chloroform extraction (unlike FAIRE) and size-selection prior to high-throughput sequencing. Yaragatti et al. (Yaragatti et al. 2008) describe a similar procedure to preferentially enrich for DNA in NFRs by first formaldehyde-crosslinking chromatin and permeabilizing nuclei, before adding of restriction enzyme HaeIII which cuts at GGCC (frequency of 1 in 300 bp in mouse genome) and separation from the permeabilized nuclei by centrifugation.

FAIRE and Sono-seq signals are both enriched near markers of active chromatin such as (RNA pol II ChIP signals, H3K4me1 and H3K4me2 marks, DHS and CpG islands) but Sono-seq DNA appears to be predominantly associated with active promoter regions and are not enriched relative to H3K4me3 modifications or distal H3K4me3 marks, DHSs or CpG islands. The method described by Yaragatti et al. (Yaragatti et al. 2008) captures known enhancer sequences and is coupled with a ligation-mediated PCR procedure and reporter assay to show that captured sequences are 20-fold enriched for enhancer activity relative to uncrosslinked digested DNA. Continued application of these techniques to diverse tissues and cell-types, normal and diseased will further our understanding of how chromatin structure and underlying primary sequence differences contribute to specify gene expression patterns.

### 1.3.4 *in silico* methods for CRM detection

Computational methods for CRM prediction that don’t rely on evolutionary conservation information typically come in two flavours; those that search for a collection of known motifs and those that search for motifs *de novo*. Both of these approaches exploit the modular organization of clustered arrays of TFBSs from multiple TFs in *cis*-regulatory modules by searching for groupings of motifs.

Early methods such as CIS-ANALYST (Berman et al. 2002) rely upon specifying an arbitrary threshold for number of motif hits within a specified window; while Cister (Frith et al. 2001) does not make motif calls based on thresholding, while additionally considers both the strength of motif hits and distances between them it still requires *ad hoc* specification of many other parameters such as CRM size and number of expected motifs. Despite the simplicity of CIS-ANALYST, Berman et al. (Berman et al. 2002) managed to recover 14 clusters of TFBSs overlapping 19 experimentally-defined CRMs known to regulate 9 genes involved in *Drosophila* embryonic development tabulated from the literature by specifying 13 matches from any combination of the TFs Bcd, Cad, Hb, Kr, or Kni in a 700bp window. However, much of the success of this method stemmed from knowing beforehand that these TFs act in concert to regulate the expression of the 9 chosen genes.

Current methods typically expand upon this basic premise by using more rigorous approaches (such as probabilistic models or HMMs) to maximize the PWM score within a sequence window of a given size for the input combination (Bailey and Noble 2003; Johansson et al. 2003) or the best combination (Rajewsky et al. 2002; Aerts et al. 2003) of PWMs and sequences, as determined by the each algorithm. A related class of programs can search for clustering of over-represented motifs within windows in the input set of co-regulated or co-expressed sequences, relieving the need to pre-specify the expected TFBSs in the CRM predictions (GuhaThakurta and Stormo 2001; Zhou and Wong 2004). These approaches are also often extended to make use of phylogenetic footprinting and evolutionary conservation to guide CRM predictions (Sinha et al. 2004; Blanchette et al. 2006; Warner et al. 2008).

### 1.3.5 Validating and characterizing regulatory sequences

An important but often underperformed step in the regulatory sequence discovery process is the experimental validation of sequence functionality in an *in vivo* setting. This is because the classical reporter assay techniques used to interrogate regulatory sequence function are labour-intensive and low-throughput, severely limiting the growth of the functional sequence catalogue. Moreover, reporter assays only provide a read-out of putative regulatory sequence activity under the conditions and timepoints studied and is usually limited to detecting enhancer activity. Nevertheless, moderate-throughput transient expression assays using an *Hsp68-LacZ* reporter construct have been successfully used to interrogate putative regulatory sequences for tissue-specific enhancer activity in mouse embryos (Poulin et al. 2005; Pennacchio et al. 2006; Visel et al. 2008) and similarly in zebrafish embryos (Woolfe et al. 2005; McEwen et al. 2006).

Chromosome conformation capture (3C), a technique developed by Dekker et al. (Dekker et al. 2002) and its variants (circularized chromosome conformation capture, 4C and carbon-copy chromosome conformation capture, 5C) enable the detection of physical interaction events between distantly-positioned sequences such as enhancers to promoters. Although this method cannot be used to determine whether putative regulatory sequences have functional roles in regulating the transcription activity of genes, it is useful in delineating the frequency of interactions between specific genomic regions of interest. Like ChIP, 3C and its derivatives depend upon the crosslinking of proteins to other proteins and DNA located in close proximity within nucleus by the addition of formaldehyde to cells. Enzymatic digestion separates crosslinked from non-crosslinked fragments before ligation and PCR amplification of crosslinked fragments using primers against the region of interest. Interaction frequencies between regions of interest can be compared for enrichment relative to neighbouring sequences, in order to ascertain genuine from happenstance interactions due to random collisions between DNA segments. The addition of a ChIP step to enrich for sonicated crosslinked fragments specifically coupled with a protein of interest can also give insight into its participation in long-range interactions in ChIP-loop (3C coupled with ChIP) (Cai et al. 2006; Kumar et al. 2007) or ChIA-PET (similar to ChIP-loop, using sonication instead of enzymatic digestion to fragment and sequenced using paired end tags) (Fullwood et al. 2009) assays.

## 1.4 Evolution of gene expression and regulation

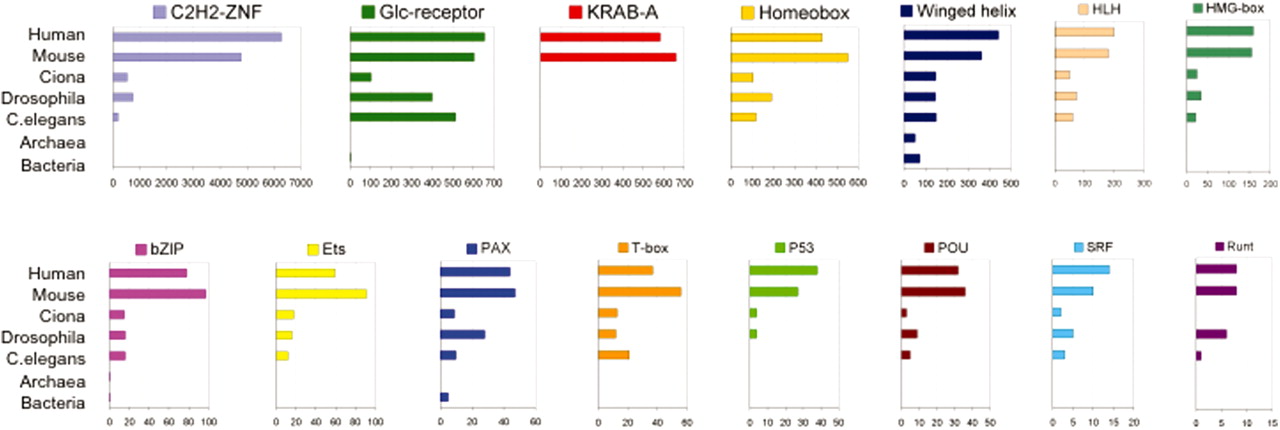
The idea that transcriptional regulatory evolution is a major contributor to generating intra- and interspecies phenotypic diversity (Britten and Davidson 1969; Britten and Davidson 1971; King and Wilson 1975) is rapidly gaining acceptance as mounting evidence suggest that specific regulatory changes underlie many functional adaptations found in natural populations (Belting et al. 1998; Shapiro et al. 2004; Gompel et al. 2005; Hay and Tsiantis 2006; Prud'homme et al. 2006; McGregor et al. 2007; Jeong et al. 2008; Linnen et al. 2009; Rebeiz et al. 2009; Wittkopp et al. 2009; Chan et al. 2010). However, the relative contributions of evolutionary changes of *cis* and *trans*-regulatory components accounting for intra- and interspecies gene expression variation is still a matter of debate, largely due to the difficulty in directly assaying and quantifying such changes as their functions are context-dependent. As such, exploration of gene expression variation is often the proxy for studying transcriptional regulatory evolution.

Starting with a null model of neutral evolution, which predicts expression divergence as simply the result of a clock-like accumulation of neutral mutations as a function of time (Khaitovich et al. 2004); typical approaches to studying gene expression evolution look for evidence of deviation from the neutral model. One way to quantify this deviation is to use an estimate of the neutral mutation rate, such as the variance of intraspecies gene expression and ask whether the interspecies expression variance differs from this rate, scaled by evolutionary time (Rifkin et al. 2003). It then follows that under a model of stabilizing selection, expression should be consistent within and between species as changes with negative functional consequences are removed, while sizable deviations from neutrality predicts positive selection. However, testing these models with expression data is challenging and should be approached with caution, as differences in data processing and model assumptions can lead to contradictory results (Su et al. 2002; Yanai et al. 2004; Liao and Zhang 2006; Yanai et al. 2006).

### 1.4.1 Evolutionary changes in *trans*-acting factors

The genotype to phenotype relationship of protein coding sequence is fairly straightforward; a string of amino acids can be predicted from the sequence and then domain structures from the protein sequence, leading to an inference of gene function. In contrast, the relationship of regulatory sequences to particular phenotypes is much harder to pinpoint as the associated transcriptional profile readout is context-dependent and linked to the expression of the regulating TFs that bind them, which in turn is dependent on the regulatory sequences associated with those factors and so on in a cascading manner. Hence, given that changes in TF primary sequence, structure (e.g. protein-protein interaction domain) or expression affecting its specificity would affect multiple loci, it is likely more difficult to select for them relative to changes in regulatory sequences. Indeed, an exhaustive mutation analysis in a basic helix-loop-helix TF found that novelty in binding specificity by amino acid changes is extremely limited and likely not a viable means for diversification of TF function (Maerkl and Quake 2009). Furthermore, experimental evidence supports the relative invariance in TF function such that in cases where an orthologous TF has been tested for functional equivalence to the endogenous TF of a distantly-related species, most but not all tested functional roles of the endogenous TF were equivalently recapitulated by the ortholog (Lutz et al. 1996; Xue and Noll 1996; Hanks et al. 1998) or by a paralog (Hanks et al. 1995; Greer et al. 2000). TFs important in development in particular, have strongly conserved functions maintained throughout bilateria, such as Hox TFs in embryonic anterior-posterior body plan specification, Nkx2-5 in heart/visceral mesoderm development and Pax-6 in eye development (Veraksa et al. 2000), as present-day TFs descended from a core set in a common ancestor over 500 million years ago. However, it should be noted that exceptions do exist, as the alteration in expression of a *trans*-acting factor BCL11A was found to underlie the expression divergence at the β-globin locus during mouse and human development (Sankaran et al. 2009).

Much of the novelty in the course of vertebrate TF evolution stems from duplication events, including whole-genome duplications (McLysaght et al. 2002; Dehal and Boore 2005), contributing to an expansion of TF families in vertebrates (Figure 1.4) from ~200 TFs in yeast to ~1500-2000 TFs in humans (Vaquerizas et al. 2009). Tandem duplications involving C2-H2 zinc-finger proteins in particular, have expanded the size of the family such that is one of the largest gene families in mammals (Tadepally et al. 2008). Following the classical gene duplication model (Ohno 1970), the consequence of TF paralogy is likely an increase in the freedom of the duplicates to evolve new expression patterns, interaction partners or binding specificities (neo-functionalize) since the ancestral TF can maintain the proper function. For example, the duplication of *Runt,* a TF involved in skeletogenesis has been recently shown to be correlated the emergence of bony tissues in vertebrates following the divergence from non-vertebrate chordates, where single-copy *Runt* plays a role in cartilage formation (Hecht et al. 2008). These observations support the notion that *trans*-regulatory changes are likely relatively rare events requiring duplications or an equivalent interacting factor to generate redundancy in order to minimize pleiotropy. Corroborating evidence demonstrating that sequence, rather than the nuclear environment is likely the predominant dictator of the location of TF binding, transcription initiation and expression output come from several experiments (Wittkopp et al. 2004; Wilson et al. 2008; Tirosh et al. 2009), including one using aneuploid mouse hepatocytes carrying human chromosome 21 (Wilson et al. 2008), pointing to regulatory sequence changes largely underlying interspecies expression differences.



1.4 Comparison of the DBD repetoire in different taxa corresponding to the largest domain families in mammals, demonstrating large expansions in these families in vertebrates relative to invertebrates (Nowick and stubbs, 2010).

### 1.4.2 Evolvability of *cis*-regulatory sequence elements and CRMs

In stark contrast to the rarity and general forbiddance of *trans*-regulatory mutations, it is now commonly believed that the *cis*-regulatory component of GRNs is highly flexible and evolvable and that the gain and loss of TFBSs provides a central means for gene expression variation and divergence (Dermitzakis and Clark 2002; Costas et al. 2003; Moses et al. 2006; Borneman et al. 2007; Doniger and Fay 2007; Ettwiller et al. 2008). To date, over 100 phenotypic trait differences in humans can be traced to *cis*-regulatory mutations (Wray 2007), including lactose tolerance (Tishkoff et al. 2007) and creative dancing (Bachner-Melman et al. 2005).

The short nature of TFBSs enable the abolishment of TF binding or facilitate TF switching by a simple point mutation, deletion or insertion and the generation new TFBSs in the same manner (Wray et al. 2003). Modularity in CRM organization also presumably allows for changes within independent modules, thereby affecting only a subset of the transcriptional profile and allowing for the fine-tuning of expression programs. However, the functional consequence of mutations at each individual site can only be tested through exhaustive experimentation and a mutation may ultimately be selectively neutral, given the flexibility of CRM architecture, if a compensatory site arises elsewhere (binding site turnover). In fact, despite the broad conservation of TF function and targets, a considerable proportion of binding events detected by ChIP (e.g. 2/3 in (Odom et al. 2007)) do not align (Borneman et al. 2007; Odom et al. 2007) in different species. The apparent mobility of TFBSs over the course of evolution challenges the effectiveness of mere sequence conservation detection as evidence for functional relevance, particularly in light of reports that TF recognition sequences may not be generally more conserved relative to background flanking sequences, even at highly bound regions (Li et al. 2008).

### 1.4.3 Re-wiring of transcriptional networks

Recently, genome-wide studies have raised a growing awareness of transcriptional rewiring as an avenue for achieving the necessary flexibility for making specific evolutionary changes to regulatory sequences of target genes one at a time while maintaining the overall transcriptional output of the circuit (Ihmels et al. 2005; Tanay et al. 2005; Tsong et al. 2006; Martchenko et al. 2007; Hogues et al. 2008; Tuch et al. 2008). As eluded to in the previous section, redundancy provided by equivalent interacting factors could theoretically provide the necessary leeway to coordinate and optimize regulatory changes to co-expressed gene modules, likely through an intermediate state (True and Haag 2001; True and Carroll 2002). Evidence for TF swapping comes from the comparison of the mating-type regulation pathway in yeasts. Conserved co-expression of mating type **a**-specific genes occurs in **a**-cells in both *S. cerevisiae* and C*. albicans* but is achieved through repression by the **α**-2 repressor in *S. cerevisiae* and activation through **a**-2 in *C. albicans*. Intriguingly, the circuit output is the same although achieved through different switches, likely facilitated by a common interaction partner Mcm1 (Tsong et al. 2006). Even a tightly conserved module involving ribosomal protein (RP) genes appears to have undergone a switch in a regulator since the divergence of *S.cerevisiae* and *C. albicans* from a common ancestor. RP genes in *S. cerevisiae* are controlled by a complex involving the TF Rap1, Hmo1p, Ifh1 and Fhl1 while in *C. albicans*, Hmo1 does not appear to be required and instead, the TF Tbf1, rather than Rap1 is highly associated with RP promoters (Hogues et al. 2008). These examples are consistent with a model in which some functional redundancy afforded by the cooperative binding of TFs can allow for evolutionary alterations in expression levels of these factors, binding site affinities and interaction strength to be sampled without negatively affecting the overall regulatory circuit output and in turn, provide the necessary flexibility to initiate subsequent specific gene expression changes when required.

## 1.5 Project overview

When Britten and Davidson and King and Wilson (Britten and Davidson 1969; Britten and Davidson 1971; King and Wilson 1975) put forth their theories on evolution of phenotypic novelty and linking them regulatory system changes, the capacity to experimentally test these hypotheses was extremely limited. Fast forwarding thirty years, the explosion of genome sequencing projects and development of technologies such as microarrays, high-throughput sequencing and comparative genomics in the past decade has truly enabled the exploration of the evolutionary forces that shape gene expression and regulation genome-wide. The goal of the project outlined here is to use these genomic approaches to explore the evolution of vertebrate gene expression at the level of *cis*-regulatory sequences and *trans*-acting factors.

Chapter 2 describes my efforts in comparing sequence and gene expression profile similarities and differences across five diverse vertebrate species, representing the major branches of the vertebrate phylogeny from mammals to avians/reptiles and to amphibians and bony fishes. I defined several measures of expression conservation and used them to determine the most likely conservation pattern of spatial gene expression across common tissues along the vertebrate phylogeny. I then looked for relationships between gene expression conservation and other genic features, such as protein sequence similarity, tissue specificity and non-exonic sequence conservation. Aspects of this work were assisted by my collaborators, specifically, Dr. Tomas Babak, Dr. Gordon Chua, Dr. Ralph Zirngibl and Dr. Michael Ratcliffe in the collection of animal tissues. Gerald Quon applied the variance-stabilizing normalization (Huber et al. 2002) to the microarray data and performed the enhancer element locator (Hallikas et al. 2006) analysis. Dr. Michael Brudno developed and applied a novel sequence alignment approach to generate sequence alignments in orthologous non-coding regions across the five vertebrate species.

In Chapter 3, I examine the *trans*-regulatory aspect of vertebrate gene expression regulation and describe a collaborative effort between members of the Hughes lab and Dr. Martha Bulyk’s lab at the Harvard Medical School to assemble a large index of mouse TF sequence binding preferences using a comprehensive protein-binding microarray design to achieve this goal. My specific roles in this project involved PBM data processing, deriving informative motif models and exploring the mammalian genome using this index. In addition, the same chapter also describes my role in the bioinformatic analysis of PBM-derived index of yeast TF binding specificities and their relationship to yeast promoter architecture. The PBM experiments were all performed by my collaborators in the Hughes and Bulyk labs and specific people are acknowledged within the chapter for their contributing work, particularly Dr. Michael Berger and Dr. Quaid Morris in the application of their respective motif algorithms (Berger et al. 2006; Chen et al. 2007) and Dr. Lourdes Peña-Castillo, with whom I developed the regression-based motif evaluation framework described in Chapter 3.

In Chapter 4, I re-visit the *cis*-regulatory component of vertebrate gene regulation and describe my work in using state-of-the-art high-throughput sequencing technologies to experimentally resolve the sequences likely active in the regulation of gene expression in vertebrate tissues using the FAIRE technique (Giresi et al. 2007). This was motivated by the result of the analysis detailed in Chapter 2 which suggests that current sequence alignment approaches are generally incapable of detecting conservation at the level of individual binding sites at these phylogenetic distances and that binding site turnover likely plays a large role in regulatory evolution.

To close, Chapter 5 describes the ongoing efforts my collaborators in Dr. Jason Lieb’s group at the University of North Carolina at Chapel Hill and I are engaged in towards exploring the evolution and organization of *cis*-regulatory sequences active in vertebrate tissues and developing predictive computational models to discriminate functional regulatory from non-functional sequences within genomes with Dr. Anna Goldenberg.

# 

# Chapter 2: Using microarrays to determine the breadth of gene expression conservation in vertebrate tissues

## 2.1 Abstract

Vertebrates share the same general body plan and organs, possess related sets of genes, and rely on similar physiological mechanisms, yet show great diversity in morphology, habitat and behaviour. Alteration of gene regulation is thought to be a major mechanism in phenotypic variation and evolution, but relatively little is known about the broad patterns of conservation in gene expression in non-mammalian vertebrates.

## Introduction

Vertebrates all share a body plan, gene number and gene catalog (Lander et al. 2001; Waterston et al. 2002; Consortium 2004; Jaillon et al. 2004) inherited from a common progenitor, but so far it has been unclear to what degree gene expression is conserved. King and Wilson (King and Wilson 1975) initially posited that phenotypic differences among primates are mainly due to adaptive changes in gene regulation, rather than to changes in protein-coding sequence or function, and this idea has accumulated supporting evidence in recent years (Ferea et al. 1999; Cooper et al. 2003; Shapiro et al. 2004; Wittkopp et al. 2004; Gompel et al. 2005; Prud'homme et al. 2006; Miller et al. 2007). Recent work has indicated that gene expression evolves in a fashion similar to other traits, where in the absence of selection, random mutations introduce variants within a population (Anand et al. 2003; Wray et al. 2003; Khaitovich et al. 2004; Lim et al. 2004; Shapiro et al. 2004; Yanai et al. 2004; Hammock and Young 2005; Zhang et al. 2007). Changes negatively affecting fitness are probably eliminated by purifying selection: core cellular processes seem to be co-expressed from yeast to human (Stuart et al. 2003), and conservation of the expression of individual genes in specific tissues has been observed across distantly related vertebrates (Yanai et al. 2004; Gershenzon and Ioshikhes 2005; Marza et al. 2005; Chervenak et al. 2006), perhaps reflecting requirements for patterning and development as well as conserved functions of organs, tissues and cell types. Conversely, changes that benefit fitness (for example, under new ecological pressures) may become fixed: changes in gene expression are believed to underlie many differences in morphology, physiology and behaviour and, indeed, subtle differences in gene regulation can result in spatial and temporal alterations in transcript levels, with phenotypic consequences at the cell, tissue and organismal levels (King and Wilson 1975; Wray et al. 2003). The degree to which stabilizing selection constrains directional selection and neutral drift across the full vertebrate subphylum is, to our knowledge, unknown.

Comparative genomic analyses provide a perspective on the evolution of both cis- and trans-regulatory mechanisms, and they are often used as a starting point for the identification of regulatory mechanisms. One estimate, using collinear multiple-genome alignments, suggested that roughly a million sequence elements are conserved in vertebrates (particularly among mammals, which represent the majority of sequenced vertebrates) (Bejerano et al. 2004; Siepel et al. 2005; Woolfe et al. 2005; Pennacchio et al. 2006), with most being non-exonic (Siepel et al. 2005), and a series of studies have demonstrated the cis-regulatory potential of the most highly conserved nonexonic elements (for example, (Woolfe et al. 2005; Pennacchio et al. 2006; Visel et al. 2008)). Another study (Thomas et al. 2003) found that only 29% of nonexonic mammalian conserved bases are evident in chicken, and that nearly all aligning sequence in fish overlaps exons, raising the possibility that gene regulatory mechanisms may be very different among vertebrate clades. Absence of conserved sequence does not imply lack of regulatory conservation, however, as many known cis-regulatory elements seem to undergo rapid turnover (Dermitzakis and Clark 2002; Odom et al. 2007), and there are examples in which orthologous genes have similar expression patterns despite apparent lack of sequence conservation in regulatory regions (Oda-Ishii et al. 2005). As further evidence of pervasive regulatory restructuring in vertebrate evolution, an analysis (Sanges et al. 2006) that accounted for shuffling (non-collinearity) of locally conserved sequences suggested that the number of conserved elements may be several fold higher than collinear alignments detect, particularly between distant vertebrate relatives, such as mammals and fish.

Trans-acting factors (transcription factors or TFs) also show examples of striking conservation, such as among the homeotic factors, and diversifying selection (Huntley et al. 2006). Studies comparing expression patterns between human and chimpanzee liver found that TF genes were enriched among the genes with greatest human-specific increase in expression levels (Gilad et al. 2006; Blekhman et al. 2008), supporting arguments for alteration of trans-regulatory architecture as a driving evolutionary mechanism (Wagner and Lynch 2008). On the other hand, in the Drosophiladevelopmental transition, expression of transcription factor genes is more evolutionarily stable than expression of their targets, on average (Rifkin et al. 2003). The fact that enhancers will often function similarly in fish and mammals, even when the enhancer itself is not conserved, indicates that mechanisms underlying cell-specific and developmental expression are likely to be widely conserved across vertebrates (Venkatesh and Yap 2005; Fisher et al. 2006).

Global trends in conservation of gene expression, conservation of cis-regulatory sequence and relationships between the two are not completely understood (Wray et al. 2003; Venkatesh and Yap 2005; Wagner and Lynch 2008), partly because the cis-regulatory 'lexicon' (that is, how TF binding sites combine to form enhancers) remains mostly unknown, testing individual enhancers is tedious and expensive, and many vertebrates are not amenable to genetic experimentation. These issues are of both academic and practical consequence: in addition to our curiosity about the origin and distinctive characteristics of the human species, primary sequence conservation is widely used to identify regulatory mechanisms. We reasoned that expression profiling data from species spanning much greater phylogenetic distance than humans and mice, and thus having greater opportunity for both neutral drift and positive selection, would allow assessment of the degree of conservation of tissue gene expression among all vertebrates, and a comparison of the conservation of expression to the conservation of nonexonic primary sequence. Here, we describe a survey of gene expression in adult tissues and organs in the main vertebrate clades: mammals, avians/reptiles, amphibians and fish. Our analyses demonstrate that core tissue-specific gene expression patterns are conserved across all major vertebrate lineages, but that the correspondence between conservation of expression and amount of conserved nonexonic sequence is weak overall, at least at a level that is detectable by current alignment approaches.

## **2.3 Results**

### 2.3.1 Microarray profiling reveals tissue-specific gene expression in *Gallus gallus*, *Xenopus tropicalis* and *Tetraodon nigroviridis,* consistent with functional roles of the tissues

To comprehensively survey the extent of gene expression conservation in tissues of five diverse vertebrate species, I used custom-designed, species-specific Agilent 60-mer microarrays to examine the expression all known and predicted genes of *Gallus gallus* (chicken), *Xenopus tropicalis* (frog) and *Tetraodon* *nigroviridis* (pufferfish) in twenty tissues in each respective organism (Table 2.1). These tissues were collected with the assistance of Tomas Babak, Gordon Chua, Ralph Zirngibl and Michael Ratcliffe.

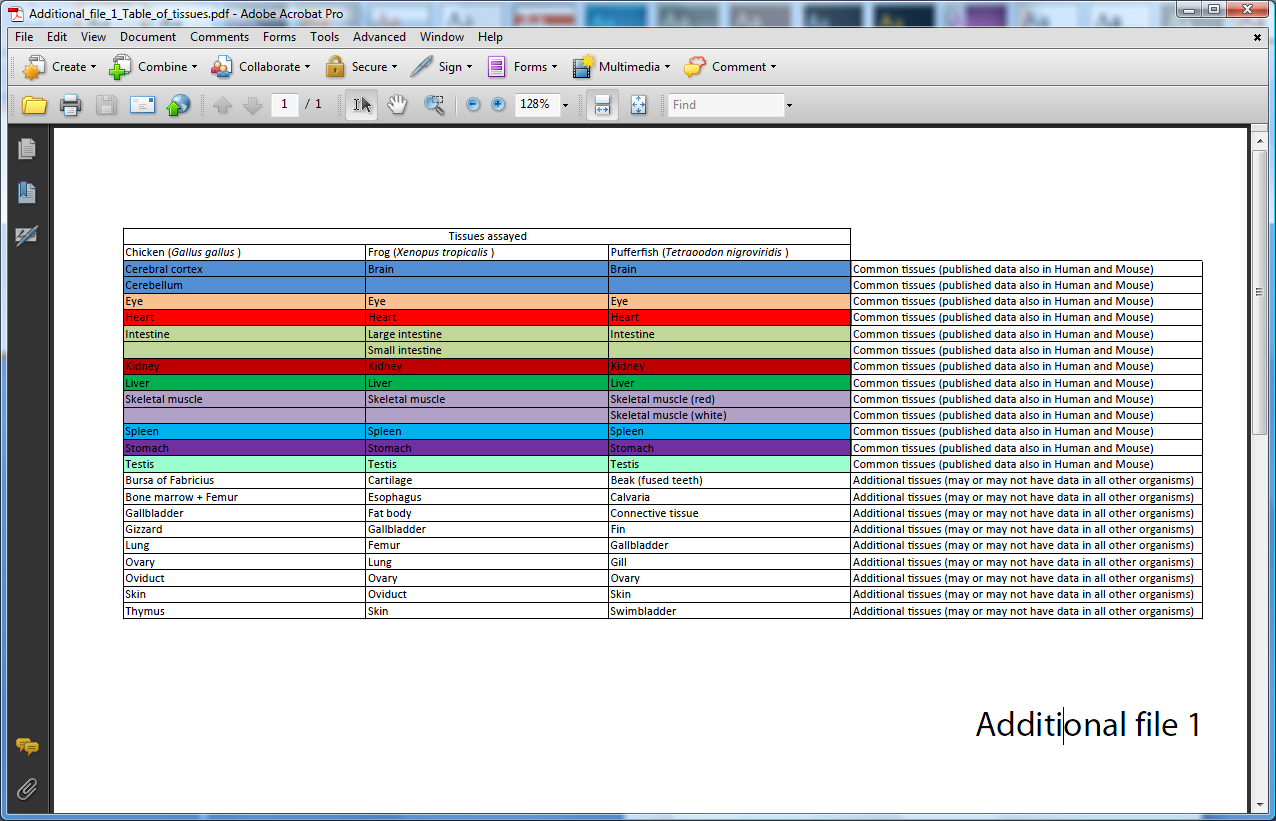
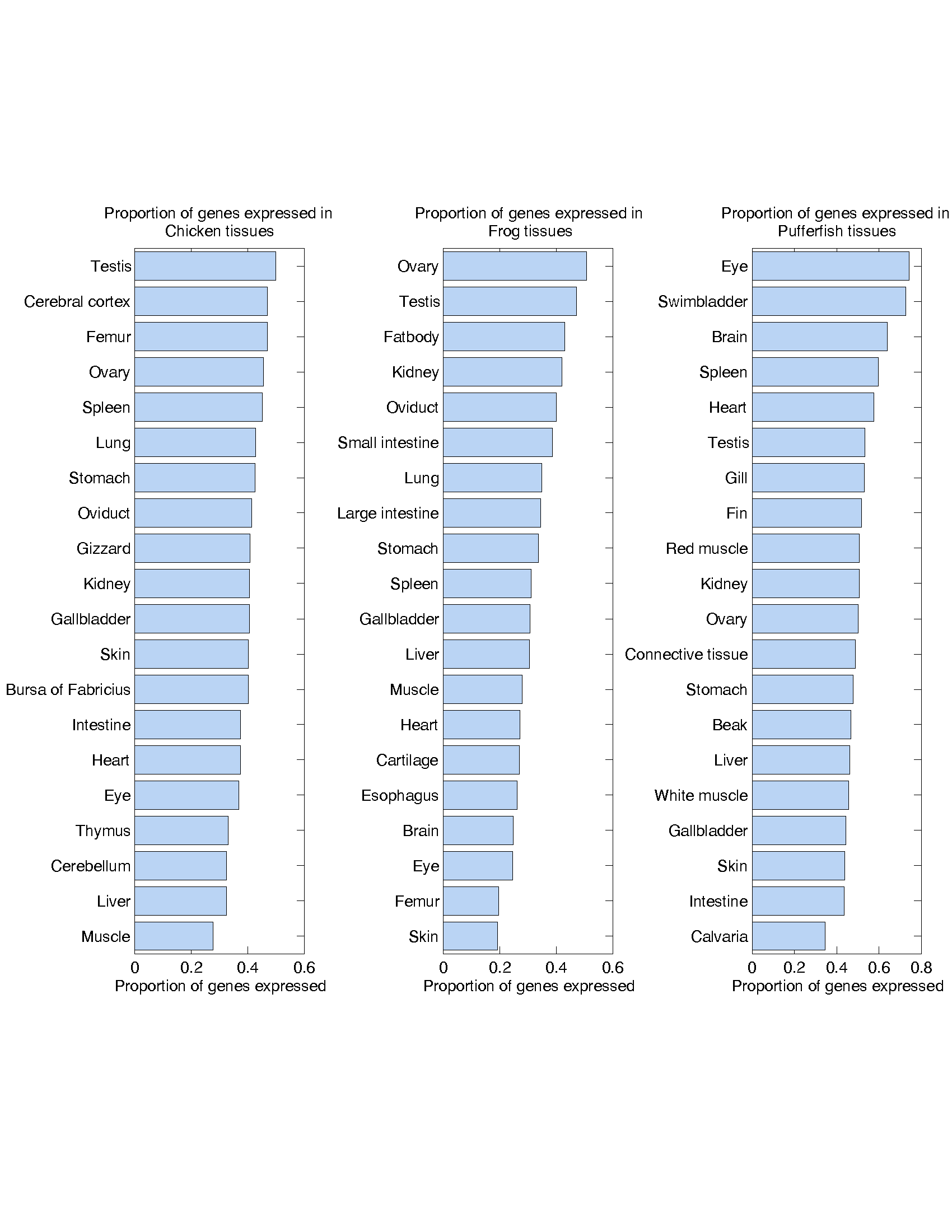


Table 2.1 Tissues assayed by novel microarrays in this study

Using a previously-described procedure to define expression (Zhang et al. 2004), 72%, 77% and 89% of genes on the arrays were detected in at least one of the tissues profiled in chicken, frog and pufferfish respectively, with 40%, 33% and 52% of genes detected in each tissue, on average (Figure 2.1), comparable to previously reported figures of 52% and 59% in human and mouse (Su et al. 2004). The percentages presented here may be lower because many predicted genes were included on the arrays.

Because the sets of genes and tissues profiled in each species were not identical, I first examined the data sets separately, by two-dimensional hierarchical clustering and Gene Ontology (GO) enrichment analyses. As evident in Figure 2.2, there are clusters of genes whose expression is highest within a tissue or a set of functionally-related tissues. Genes with particular GO annotations were preferentially expressed within each tissue, in many cases consistent with tissue function. For instance, genes associated with visual perception [GO:0007601] are preferentially expressed in the eyes of all three species, while genes associated with lipid metabolic process [GO:0006629] were preferentially expressed in the gallbladder, liver, intestine and stomach of all three species. The full 

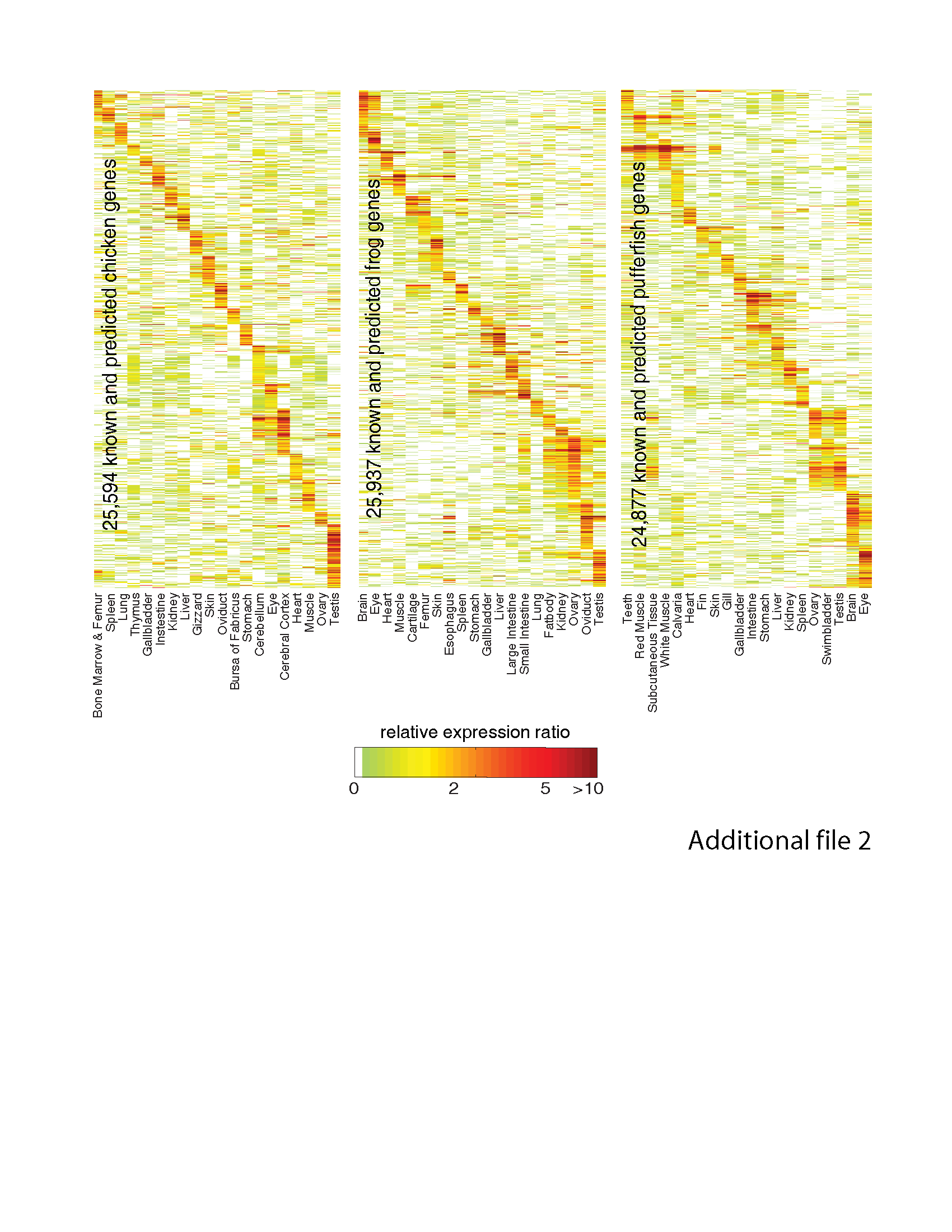
**FIGURE 2.1 BREAKDOWN OF THE PROPORTION OF ALL GENES IN EACH SPECIES THAT ARE EXPRESSED WITHIN EACH TISSUE.**

set of GO annotations and enrichment p-values can be found at the comparative vertebrate gene expression website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/vertebrate\_ expression).

For each dataset, I clustered the relative expression ratios of each gene within each of the 20 tissues profiled and ordered the rows and columns independently, breaking the high-level branches to rearrange the clustered data to obtain a diagonal appearance as described in (Zhang et al. 2004). The resulting clustergrams (Figure 2.2) from these analyses show that prominent tissue-specific expression patterns are found in all vertebrates.

### 2.3.2 Tissue-specific gene expression is broadly conserved across vertebrates

To examine gene expression in a broad range of vertebrates, I collected a compendium of gene expression datasets, consisting of previously published datasets for human (Schadt et al. 2004) and mouse (Zhang et al. 2004), and newly generated datasets containing 20 tissues each from chicken (Gallus gallus), frog (Xenopus tropicalis) and pufferfish (Tetraodon nigroviridis). To ask whether tissue-specific gene expression patterns are conserved among vertebrates, I focused on 1-1-1-1-1 orthologs (genes that are present in a single unambiguous copy in each of the five genomes), because genes that have undergone duplication events are subject to different constraints from singletons (Gu et al. 2005; Chung et al. 2006). Among 4,898 1-1-1-1-1 orthologs found by Inparanoid (O'Brien et al. 2005), 3,074 were measured by microarrays in all ten common tissues of chicken, frog, pufferfish, and mammals (human and mouse combined expression – see Section 2.5.5 and 2.5.6). The expression profiles of these 3,074 genes in analogous and functionally related tissues in different species were more similar than they were to those of unrelated tissues from the same species (Figure 2.3), even for pufferfish, which diverged from the other vertebrates in our study roughly 450 million years ago (Mya), well before the divergence of frog (about 360 Mya) or chicken (about 310 Mya) (Hedges 2002). Despite differences in cognition and behaviour between humans and other species, overall gene expression in the brain is most similar across the species studied compared with expression in other tissues (median expression ratio Pearson correlation (r) = 0.63),



**FIGURE 2.2 CLUSTERGRAMS SHOW THE MICROARRAY DATASETS IN CHICKEN, FROG AND PUFFERFISH, DISPLAYED AS RELATIVE EXPRESSION RATIOS FOR EACH GENE WITHIN EACH OF THE 20 TISSUES PROFILED.**

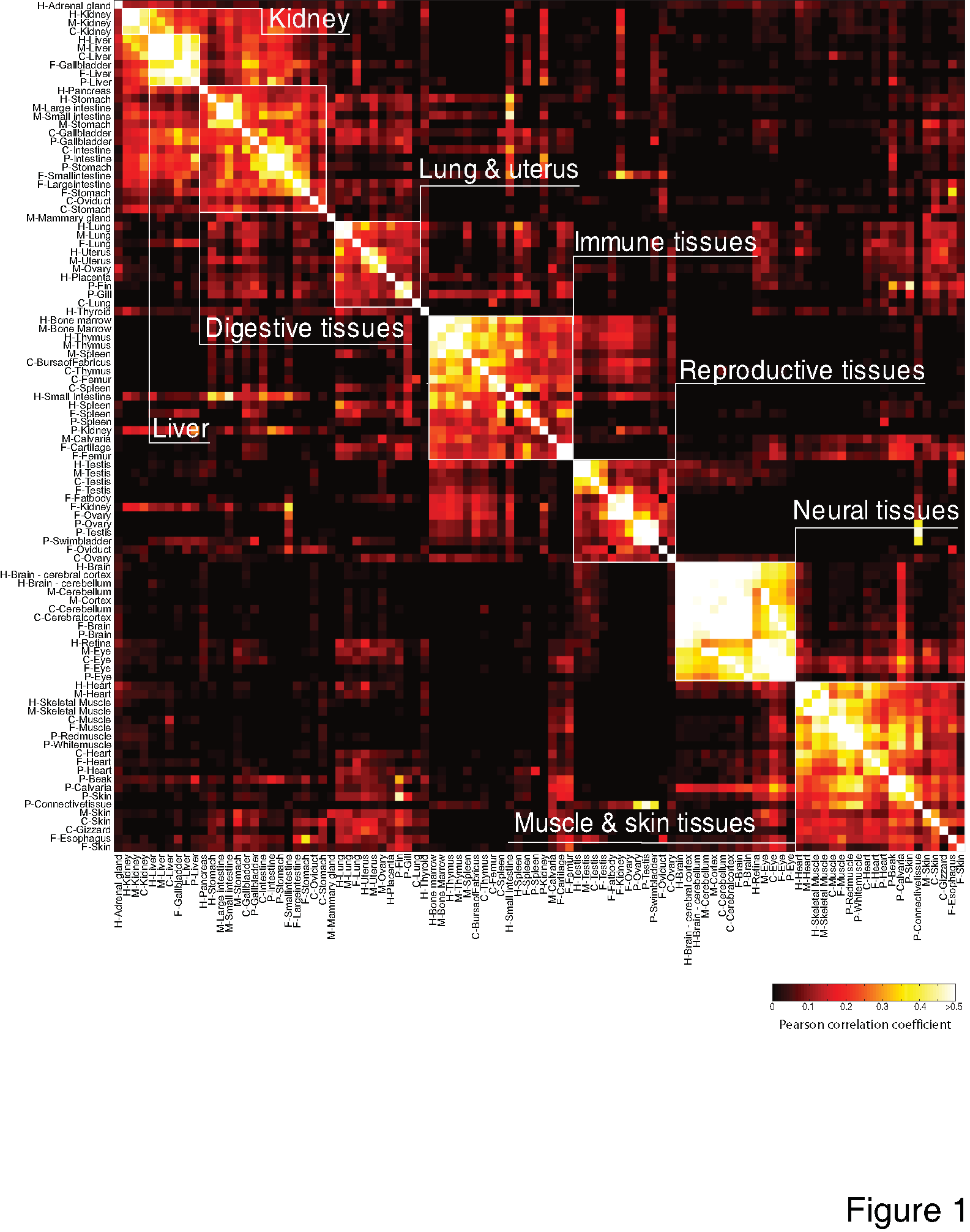


Figure 2.3 Comparison of tissue expression profiles among five diverse vertebrates. Clustered heat map of the all-versus-all Pearson correlation matrix between 20 tissues in each of human (H), mouse (M), chicken (C), frog (F) and pufferfish (P) over all 3,074 1-1-1-1 orthologs. Analogous and functionally-related tissues are boxed in white, demonstrating the cross-specices similarity of those tissues on the basis of their gene expression profiles.

consistent with a previous study comparing human and chimpanzee (Khaitovich et al. 2005). The relatively low divergence of gene expression in brain is hypothesized to be due to constraints imposed by the participation of neurons in more functional interactions than cells in other tissues (Khaitovich et al. 2006). In contrast, gene expression in the kidney was most dissimilar between species (median expression ratio Pearson r= 0.21), possibly reflecting evolution of kidney function (see Discussion). A dendrogram for the ten common tissues (with the same tissue measured in all five datasets; Figure 2.4) shows clear segregation of the data for heart/muscle, eye, central nervous system (CNS), spleen, liver and stomach/intestine. Only the testis and kidney datasets are split, each into two groups, with pufferfish and/or frog forming the outlying group. Examining the GO biological process terms enriched in tissues among the 3,074 genes are also generally conserved across the five species (the full table of enriched terms is available through the project website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/vertebrate\_expression). Following these observations, I conclude that programs of tissue-specific expression are broadly conserved among vertebrates.

### 2.3.3 Thousands of individual tissue-specific gene expression events are conserved across all vertebrate clades

I next sought to quantify the conservation of expression of individual genes. To this end, I used two conceptually simple measures intended to capture different aspects of conservation of expression. The first asks how often specific gene expression events (instances in which gene X is expressed in tissue Y) are conserved across all vertebrates. I refer to this as the “binary measure” because, to simplify statistical analysis, I considered a fixed proportion of the normalized, ranked microarray intensities of genes in each tissue to be expressed ('1'), and analyzed the data using several such proportions (1/6, 1/5, 1/4, 1/3, 1/2. The binary matrices can be found at the project website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/vertebrate\_expression). I then asked how often a gene is expressed in all species in a given tissue (that is, a fully conserved expression “event”). The proportion of conserved expression events at different thresholds ranges from 3% to 19.3% of all possible expression events, among the 3,074 1-1-1-1-1 orthologs (Figure 2.5a), and the proportion of genes with at least one conservation event ranges from 11% to 49.5% (Figure 2.5b), in all cases clearly exceeding permuted (negative control) datasets. On the basis of the spread between blue and orange bars in Figure 2.5, about 10% of the 30,740 possible gene expression events are conserved among all vertebrates, and at least 20% of all 1-1-1-1-1 orthologs participate in at least one such event. This measure probably underestimates the conservation of gene expression, because we surveyed only ten tissues and because I did not considered lack of expression across all species to represent an example of conserved expression.

The second measure I used was Pearson correlation across the ten common tissues. As with the binary measure, I found that gene expression across tissues between real 1-1-1-1-1 orthologs is more similar than randomly matched genes in pairwise comparisons between species (Figure 2.6 shows all pairwise comparisons, and also the median of pufferfish versus all other species, to provide a summary of overall conservation). The difference between the real and random (permuted) lines in Figure 2.6 indicates that roughly 20% of all 1-1-1-1-1 orthologs display conserved expression – a proportion comparable to that obtained using the binary measure. In fact, at r= 0.4, the apparent false discovery rate is similar to that obtained with the 1/3 cutoff using the binary measure (27.4% versus 34.5%), as is the number of genes classified as having conserved expression (843 versus 1,062). The overlap between these two sets of genes is higher than expected at random

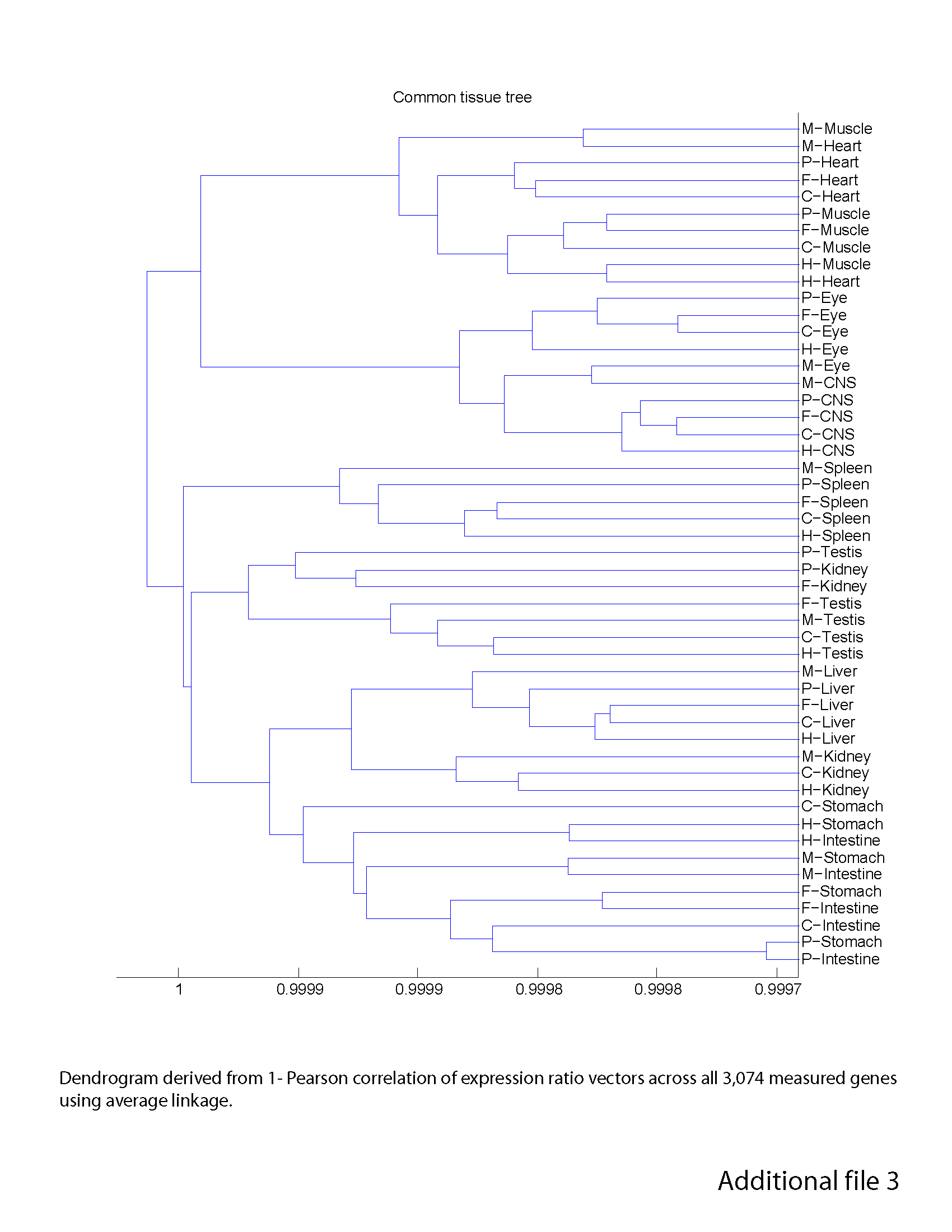


Figure 2.4 Dendrogram of correlations among ten common tissues, using 1 – Pearson correlation and average linkage over 3,074 genes.

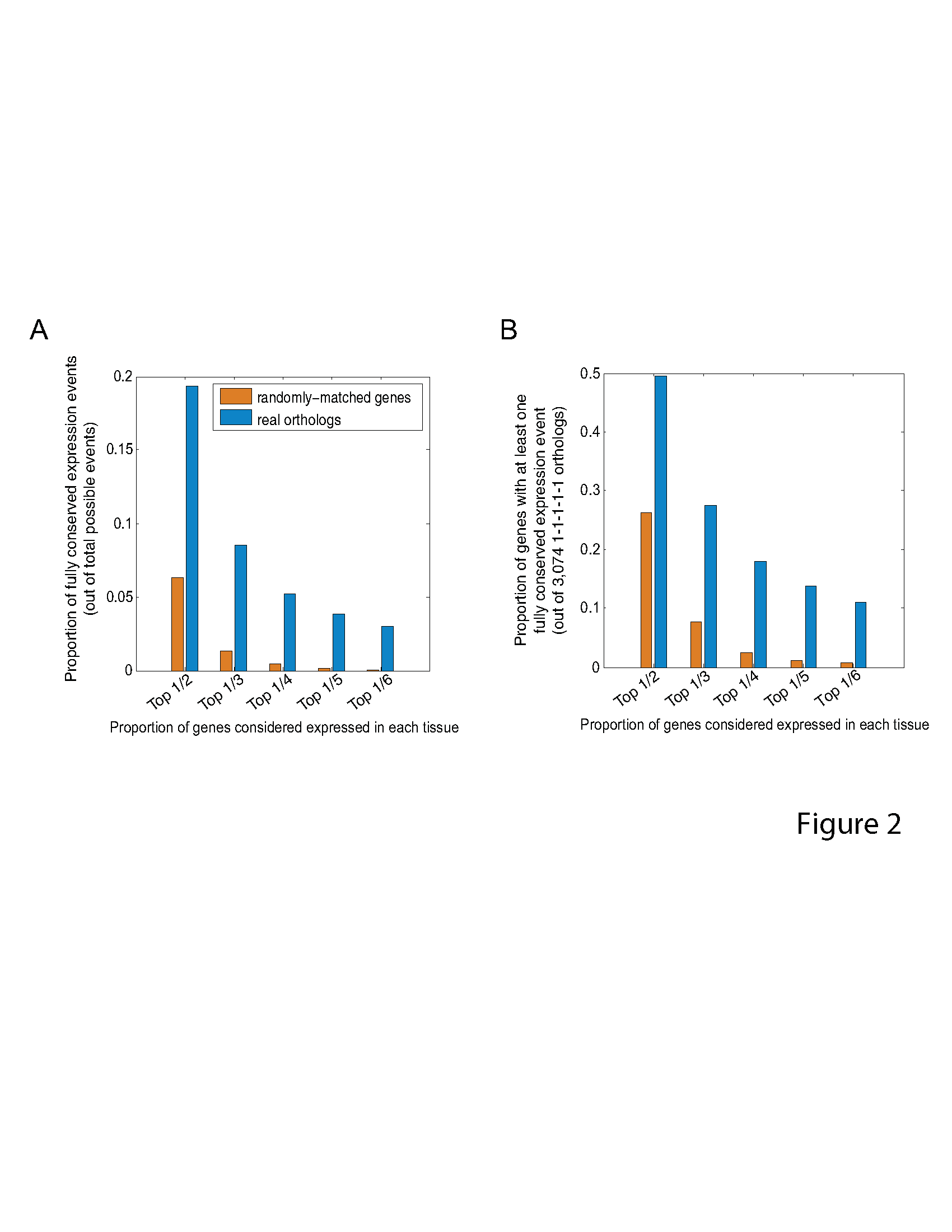


Figure 2.5 Conservation of gene expression using the binary measure. (a) Proportion of conservation events out of total possible conservation events at different thresholds using the binary model. (b) Proportion of genes with at least one conservation event among the ten common tissues out of all 3,074 measured genes using the binary model.

(417 versus 291 at random); however, it is far from absolute, indicating that the definition of conserved expression influences conclusions regarding conservation of expression.

Regardless of the method of comparison the same essential conclusion is reached: a major component of tissue gene expression has apparently remained intact since the common ancestor of all vertebrates. A large fraction of genes is encompassed; between the two measures (the binary measure and the Pearson measure), 48.4% of all 1-1-1-1-1 orthologs (1,488/3,074) scored as having conserved expression at about 30% apparent false discovery rate. Thus, in just the ten common tissues we analyzed, gene expression is at least partially conserved for at least a third of all unique orthologs (48.4% × 0.7 = 33.9%) by at least one of the two definitions of conservation. The expression of these 1,488 genes in modern-day lineages is shown in Figure 2.7. Most of these genes

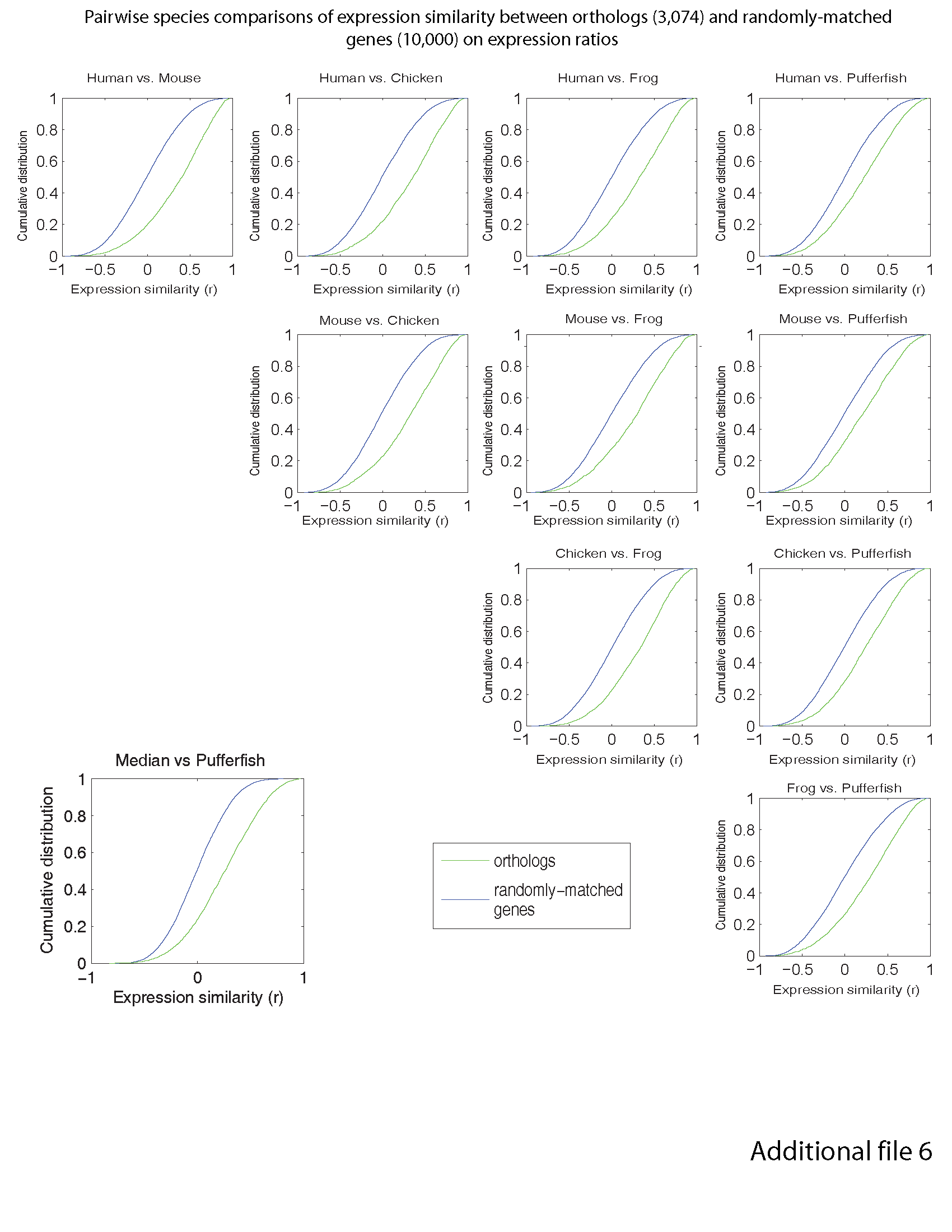


Figure 2.6 The cumulative distributions show the proportion of all 3,074 genes with Pearson r (normalized intensities) below the value shown on the horizontal axis, for real orthologs (green) and randomly matched genes (blue).

have tissue-specific patterns of expression, indicating that the genes we are identifying are not simply ubiquitously expressed housekeeping genes.

Although the focus of my study was to identify conserved gene expression patterns, our data are consistent with previous findings that divergence of gene expression scales with evolutionary time (Khaitovich et al. 2004; Zhang et al. 2007)  when averaged over all genes (Figure 2.8a) or all tissues (Figure 2.8b; the same trend is apparent in Figure 2.7 and Figure 2.6). Individual tissue expression profiles show different evolutionary trajectories, however (Figure 2.8c), presumably reflecting diversity in constraints on tissue function.

### 2.3.4 Conservation of expression does not correlate with proportion or amount of conserved nonexonic sequence

I next asked what gene properties correlate with conservation of expression among the 3,074 measured unique orthologs. I considered the following gene properties: those that are contained in my data, that is, median expression level and Shannon entropy as a measure of tissue specificity and preferential expression in individual tissues; GO annotations; and sequence properties, that is, length of gene, size of encoded protein, presence of a DNA-binding domain (for known and predicted TFs), sequence conservation of encoded protein (pairwise BLASTP bit score) and amount of conserved nonexonic sequence (measured in several ways) (Tables found at the project website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/vertebrate\_expression) see Section 2.5.13 for details).

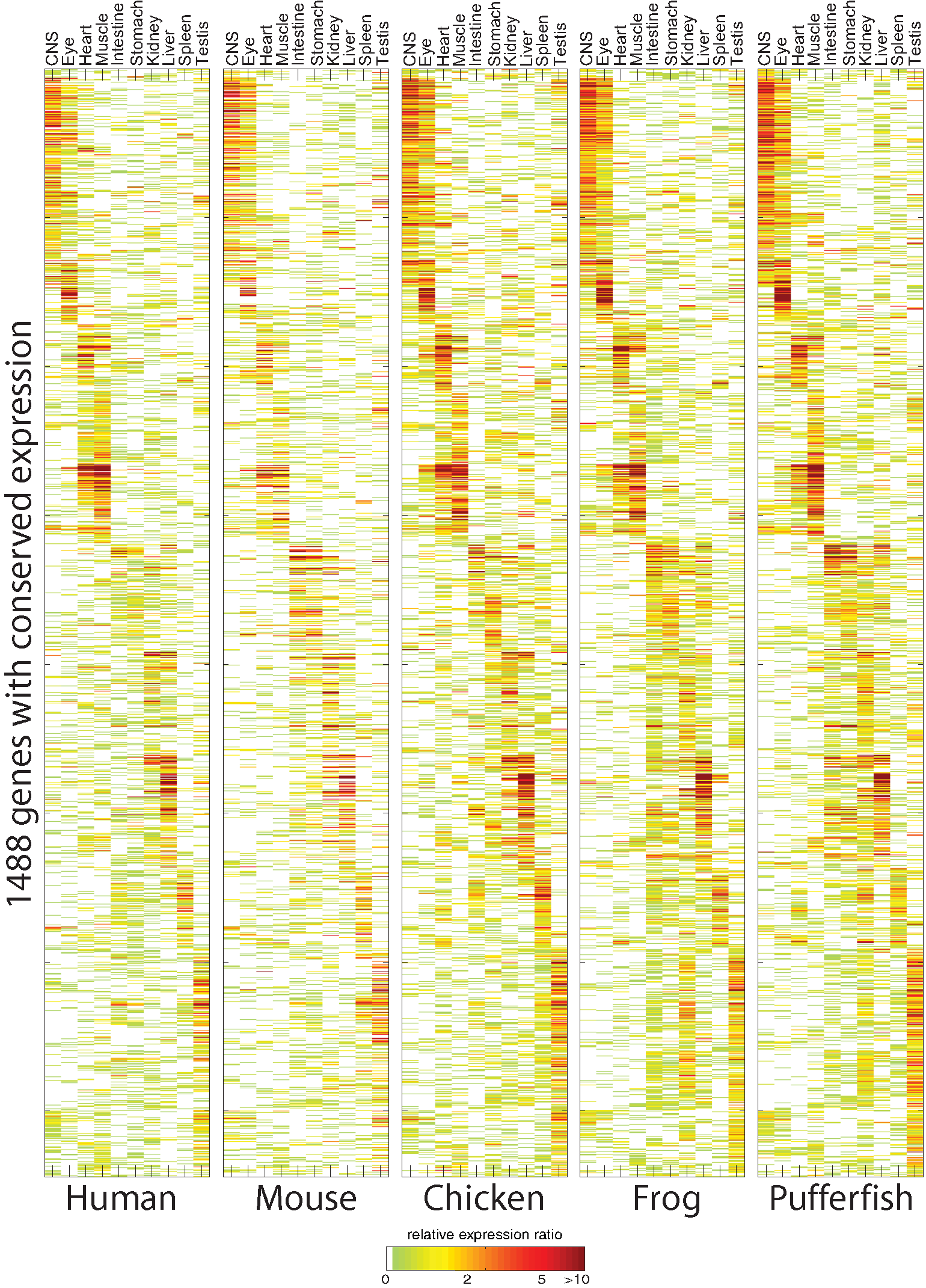


Figure 2.7 A core conserved vertebrate tissue transcriptome. Expression ratios of the measured and predicted expression patterns of 1,488 1-1-1-1-1 orthologs as described in the text and Materials and methods are shown. Two-dimensional hierarchical agglomerative clustering using a distance metric of 1 – Pearson correlation followed by diagonalization was applied to the pufferfish data and the same row and column order was forced onto the other data sets.

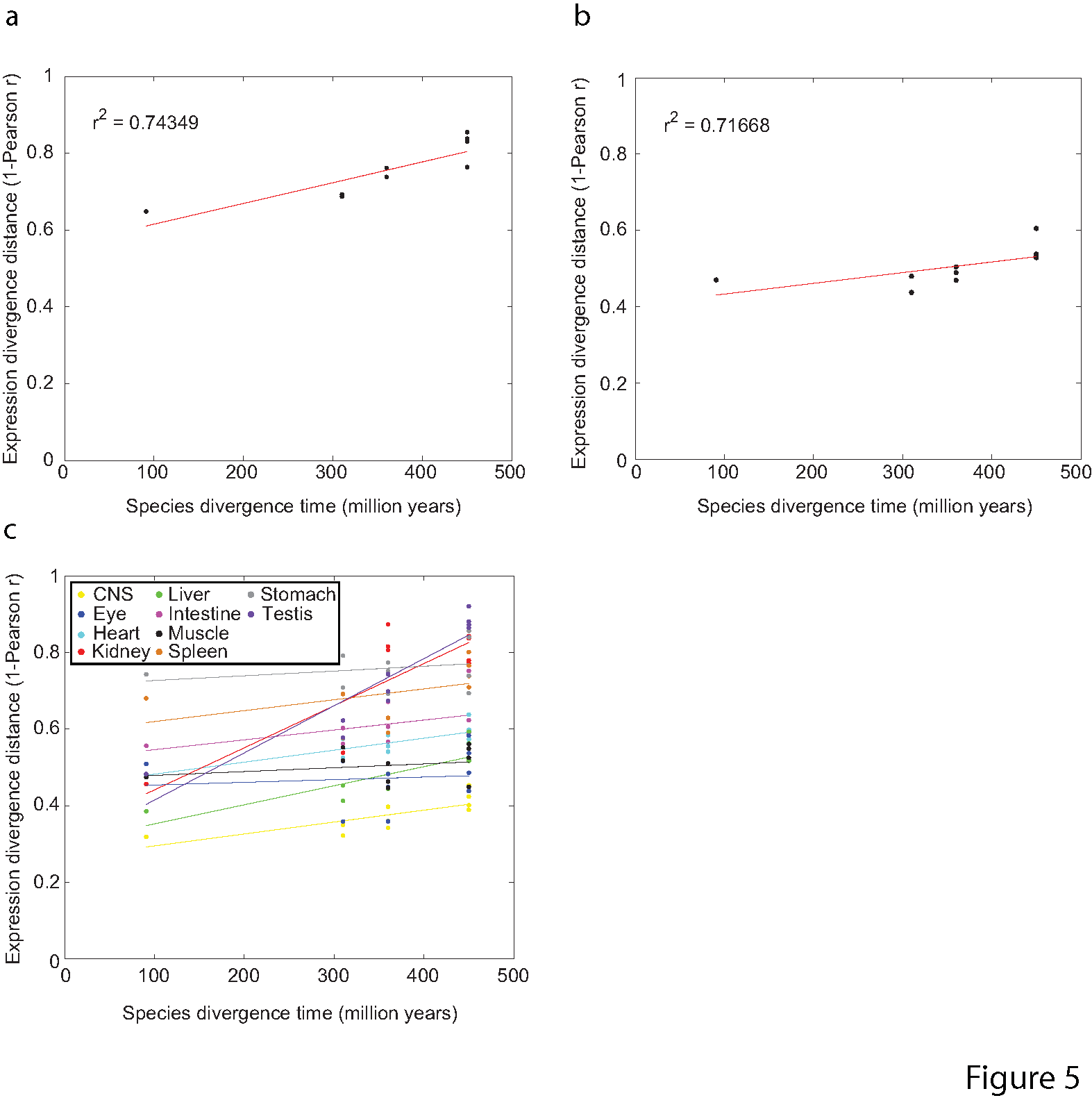


Figure 2.8 Comparison of gene expression conservation to evolutionary distance. The scatter plots show expression distance as 1 – Pearson correlation, using median-subtracted asinh values (comparable to ratios). (a) Median pairwise correlation over all genes; each point represents a pair of species. (B) Median pairwise correlation over all tissues; each point represents a pair of species. (C) Individual pairwise correlations over tissues, as indicated with colours; each point represents a single tissue in a single pair of species. Estimated species divergence times were obtained from HEDGES(Hedges 2002).

Several observations emerged from this analysis. First, the genes with the highest expression similarity between species are most often genes expressed in a highly tissue-specific manner in tissues with specialized functions. Although the Pearson correlation is heavily influenced by extreme values, thus giving higher weight to tissue-specific pairs, most of these high scoring genes were also classified as conserved by our binary measure. Among the 50 genes with highest median pairwise Pearson correlation of expression are structural components of the eye lens, liver-synthesized proteins involved in the complement system and blood coagulation, and neurotransmitter receptors and transporters. This observation is supported by the GO categories enriched among genes with high expression similarity, such as synaptic transmission (GO:0007268), visual perception (GO:0007601), wound healing (GO:0042060) and muscle development (GO:0007517) (Wilcoxon-Mann-Whitney test (WMW) p-values 1.55 × 10-4, 2.36 × 10-3, 2.24 × 10-3and 4.98 × 10-5, respectively; full *p*-value table available at the project website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/vertebrate\_expression). In contrast, I did not find any evidence that the expression of TFs (228 of the 3,074 measured orthologs) is more or less conserved than that of non-TFs, in contrast to previous reports of both higher (Gilad et al. 2006) and lower (Rifkin et al. 2003) rates of evolution of TF expression. A slightly lower proportion of TFs did seem to show conservation events relative to non-TFs using the binary measure, but this difference is due to the fact that TFs are expressed in fewer tissues: the difference is not seen when comparing TFs and non-TFs with similar overall expression levels (data not shown).

It is widely believed that conserved nonexonic sequence often serves a cis-regulatory function, and it follows that a larger amount of conserved nonexonic sequence might correlate with a higher probability of conserved expression. However, I found that the correspondence was very weak: for example, for the binary model, we obtained Spearman correlations of -0.086 and 0.0029 with the number of nonexonic bases in Phastcons conserved regions (Siepel et al. 2005) and in ultraconserved elements (Bejerano et al. 2004) (UCEs), respectively; for the Pearson model, these correlations were 0.054 and 0.0075, respectively. Similar results were obtained when proportion of bases replaced number of bases (Figure 2.9a, b). The handful of outlying points in the upper right of Figure 2.9b includes several TFs, a subset of which are known to have an exceptional degree of nonexonic sequence conservation (Bejerano et al. 2004).

I reasoned that pervasive shuffling might obscure most of the cis-regulatory elements, particularly in pufferfish. In order to address this possibility, Dr. Michael Brudno developed a technique similar to that of Sanges et al. to detect shuffled conserved sequence elements (SCEs), which may be non-collinear, across the five species (see Section 2.5.12 for more detail). Among the total 4,898 1-1-1-1-1 orthologs, I identified 491,028, 457,074, 79,001, 54,134 and 11,731 SCEs in human, mouse, chicken, frog and pufferfish with median lengths of 164, 80, 68, 68 and 65 nucleotides, respectively. These SCEs showed good overlap with those in Sanges *et al.*(Sanges et al. 2006) (75.5% of the sequences in Sanges *et al.*(Sanges et al. 2006) are within regions I aligned were identified as SCEs in my analysis) and they were calibrated to minimize false positives. However, I still did not observe a strong relationship between the degree of conservation and the proportion or number of aligned bases in each species (median Spearman correlation: -0.062 and 0.042 for binary and Pearson models, respectively, versus proportion of aligned nonexonic bases in each species; Figure 2.9c, d; similar correlations are obtained with number of aligned non-exonic bases).

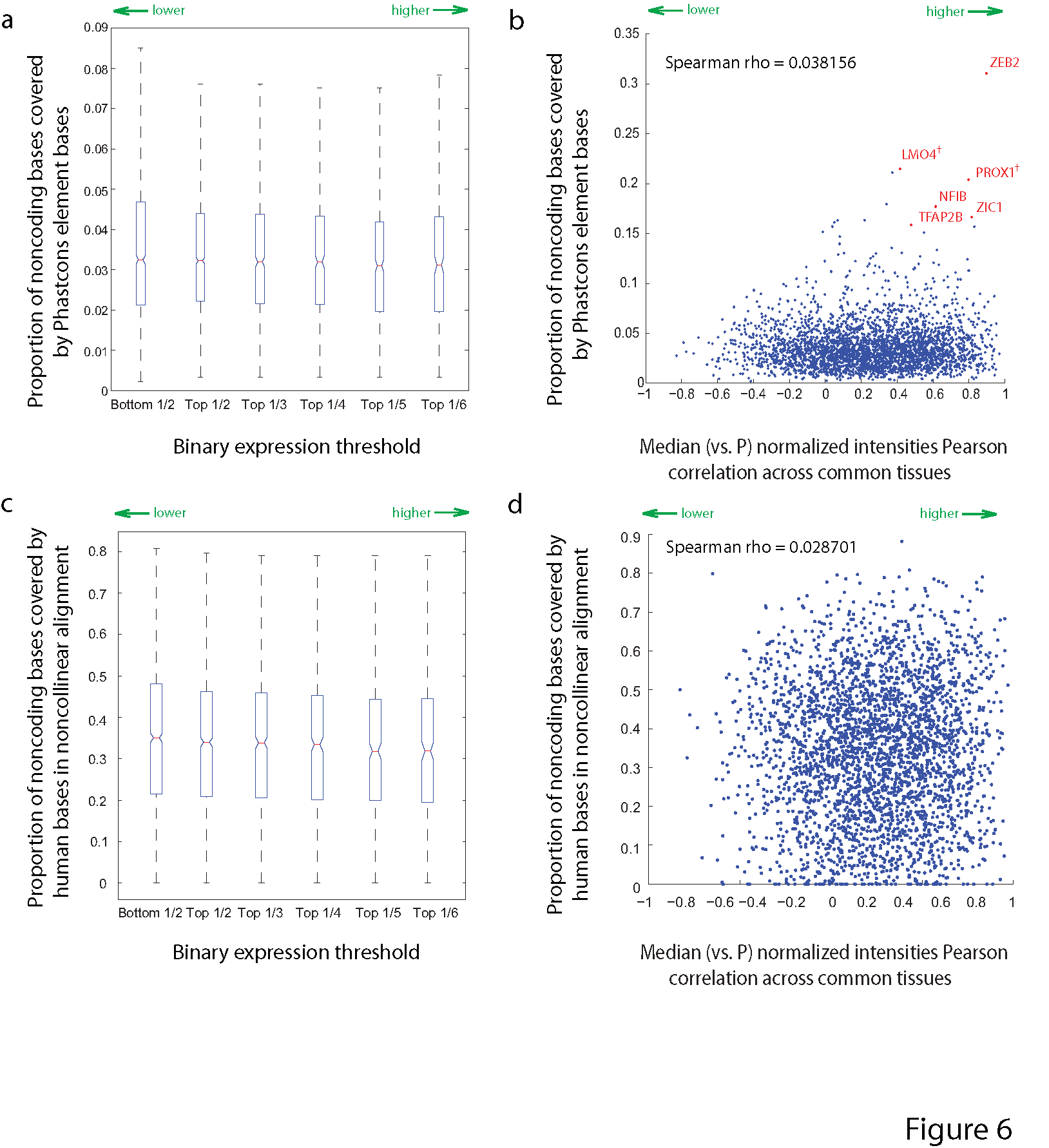


Figure 2.9 Relationships between expression similarity between orthologous genes and amount of conserved nonexonic sequence. Proportion of conserved nonexonic sequence defined as Phastcons elements (a, b) and human bases in non-collinear alignment (c, d) compared against the conservation of gene expression by the binary measure (a, c) and Pearson measure (normalized intensities versus pufferfish (P) (b, d) (see text and materials and methods for details). Selected TFs are indicated in (b) (see text). Probable TFs as determined by their ensembl gene descriptions, but were not identified by our domain analyses are indicated by †. Spearman rho refers to the spearman correlation coefficient.

I also examined the correlations between nonexonic sequence conservation and expression correlation at varying evolutionary distances from human. Although correlations remain weak (Figure 2.10a), I did find that genes in the highest quartile of sequence conservation had a significantly higher distribution of expression correlation than those in the lowest quartile, for all pairwise comparisons except human versus pufferfish (Figure 2.10b). However, in all comparisons, there are many genes with little sequence conservation and high expression correlation, and vice versa. In fact, among the 173 genes with the most highly conserved expression in our study by both measures we applied (those in the top 1/6 by the binary measure and with median Pearson r≥ 0.5); most (102) have no nonexonic conserved sequence in fish, on the basis of our SCEs. The expression of these 102 genes in the ten common tissues in the representatives of all modern lineages is shown in Figure 2.11.

Because TF binding sites are degenerate, it is conceivable that these genes have a high number of conserved TF binding sites, despite their lack of primary sequence conservation. To examine this possibility, Gerald Quon used the Enhancer Element Locator (Hallikas et al. 2006) (EEL) program to align TF binding sites defined by 138 motif models downloaded from the JASPAR database (Sandelin et al. 2004). Over all 4,804 aligned human/pufferfish ortholog pairs, the number of genes that scored highly using EEL was only slightly higher with real ortholog pairs than with randomly assigned orthologs with similar amounts of nonexonic associated sequence in both genomes (p= 0.24, Kolmogorov-Smirnov test; see Materials and methods and Figure 2.12) and there is almost no correlation between EEL score and conservation of expression (EEL score against median versus pufferfish normalized intensity Pearson r= 0.022). I conclude that the regulatory architecture of the vast majority of genes has diverged beyond recognition by any current approaches, despite the apparently very similar regulatory output in many cases, and the likelihood that at least some orthologous TFs are functioning in the same tissues.

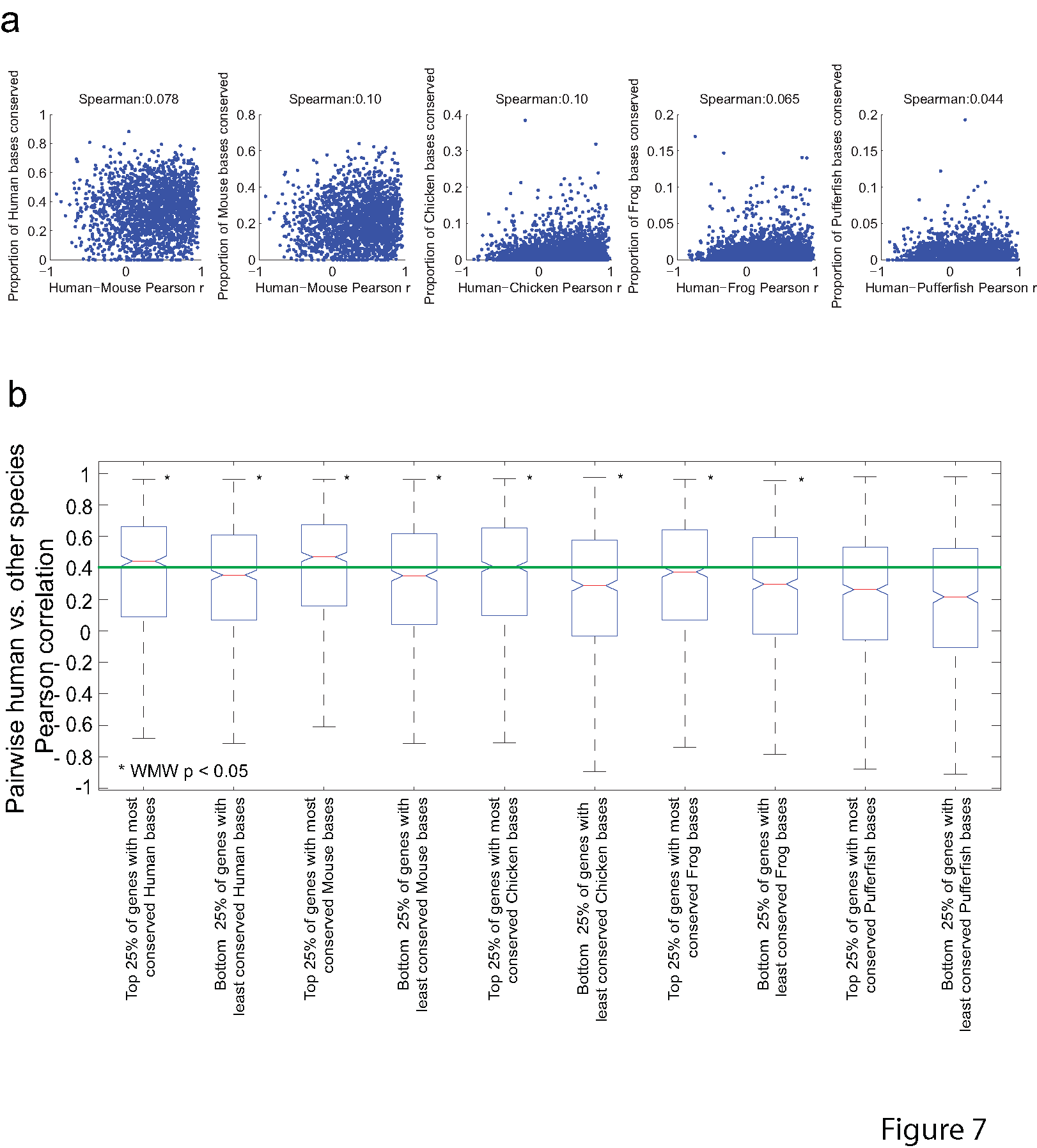


Figure 2.10 Low correlation between conservation of gene expression and amount of conserved nonexonic sequence is largely independent of evolutionary distance. (a) Scatter plots show the proportion of bases conserved in SCE alignments versus Pearson correlation (ratios) for individual genes. (B) Box plots show the distribution of pearson correlations for genes in the top and bottom quartiles of number of conserved bases. Asterisks indicate significant differences between the top and bottom quartiles.

## **I:\Thesis\Figure_8_Highly_conserved_genes_without_sequence_conservation_in_pufferfish.png**

Figure 2.11 Expression of 102 genes with highly conserved expression across all vertebrate lineages, but no detectable nonexonic sequence conservation between pufferfish and frog, chicken, mouse, or human. Mouse and human expression profiles are merged to represent mammals. Gene identifiers and descriptions for human were downloaded from Ensembl(Hubbard et al. 2009).

## J:\Thesis\figures\chapter2\Additional_file_9_EEL_score_CDFs.tif

Figure 2.12 Cumulative distribution of EEL scores for real and permuted orthology between human and pufferfish.

## 2.4 Discussion

My data provides a resource of large-scale gene expression data in tissues of three non-mammalian vertebrates and demonstrate that there is conservation of core vertebrate tissue gene expression. My analysis almost certainly underestimates the proportion of genes with conserved expression patterns, because it was focused on only ten large adult tissues in mature animals in captivity. Nonetheless, my results already provide an index of several thousand highly conserved tissue gene expression events and a picture of core gene expression in major tissues and organs of the common progenitor of all vertebrates, which most likely resembled the expression patterns shown in Figure 2.7.

In my analysis, I found that some biological processes emerged as more highly conserved than others. Genes involved in the more conserved processes on the whole tend to be preferentially expressed in tissues with a limited number of cell types (brain, eye, liver, heart and muscle) that carry out specialized functions particular to those tissues. This finding is consistent with the notion that mechanisms underlying important biological processes should be conserved across taxa (Whitehead and Crawford 2006). It is likely that the conservation of gene expression in these tissues extends beyond the base of vertebrates; coexpression of neuronal genes, for example, is observed as far as nematodes (Stuart et al. 2003). Genes expressed in tissues subject to greater environmental influence (such as intestine, stomach and spleen) may be more likely to take on new roles and diverge in expression as means of adaptation. I find gene expression similarity in the testis across vertebrates to be relatively low compared with other tissues, in support of earlier observations of accelerated evolution of testis gene expression in Drosophila (Meiklejohn et al. 2003)and primates (Khaitovich et al. 2005), and consistent with mating competition and speciation. I also note that relatively low conservation of gene expression in the kidney is consistent with its divergent function; in teleosts, the kidney is a major lymphoid organ (Zapata et al. 2006).

My finding that the correlation between the amount of conserved nonexonic sequence and conservation of gene expression is low underscores the apparent malleability of the cis-regulatory 'lexicon' (Dermitzakis and Clark 2002; Sanges et al. 2006; Wilson et al. 2008). It is easy to rationalize cases in which there is conserved sequence but little or no conservation of expression in my study, because it is possible that the majority of conserved sequence identified is either not cis-regulatory, is functioning as cis-regulatory in a context that I have not measured and/or regulates neighbouring genes. What is most striking is that many genes with the highest conserved tissue-specific expression have almost no nonexonic conserved sequence outside mammals. Divergence in trans-regulatory architecture does not provide a satisfying explanation for this observation: although there are examples in which the binding specificity of TFs evolves (Bustamante et al. 2005; Lopez-Bigas et al. 2008), as a general rule the individual monomeric TF sequence specificities seem to be unchanged over large evolutionary distances (Berger et al. 2008), and DNA-contacting residues are often the most conserved (Luscombe and Thornton 2002). The fact that conservation of TF expression is comparable to that of other genes with similar levels of tissue specificity also supports the notion that many ancestral vertebrate regulatory mechanisms are still in use. Moreover, enhancers can be functional across large evolutionary distances (for example, human reporters in zebrafish), even when the enhancers are not conserved, or are below the threshold of detection by current alignment techniques (Fisher et al. 2006).

Understanding the mechanisms underlying conservation of expression patterns is a major challenge in our understanding of evolution, and of genome function and gene regulation: even within a single genome, it is difficult to find cis-regulatory modules shared by co-expressed genes, indicating that there are many ways to achieve the same expression output. I propose that my catalog of conserved (and non-conserved) expression will be useful to test ideas regarding enhancer definition. In particular, I predict that the small size of pufferfish genes and knowledge of the expression of TFs in each tissue may facilitate searches for enhancers on the basis of density and arrangement of TF binding sites, rather than primary sequence alignment (Oda-Ishii et al. 2005; Segal et al. 2008). These and other techniques require experimental data for training and testing, and I have now provided such data for tens of thousands of genes across the vertebrate lineage, including several thousand unique orthologs.

## 2.5 Materials and methods

### 2.5.1 Design of microarrays for known and predicted genes of 3 non-mammalian vertebrates

60-mer probes were designed against the set of known and predicted mRNAs compiled from version 37 of the Ensembl (Hubbard et al. 2009) and JGI databases (http://genome.jgi-psf.org/Xentr4/Xentr4.home.html) corresponding to 24,877, 25,594 and 25,937 known and predicted loci in Tetraodon, chicken and X. tropicalis, respectively, resulting in 41,533, 41,534 and 41,523 probes. These probes were designed to have minimal cross-hybridization potential and Tm-balanced, with one or more probes mapping to each transcript. The probes were synthesized onto the arrays by direct deposition ink-jet technology and manufactured as 44K oligonucleotide microarrays (Agilent Technologies, Santa Clara, USA).

### 2.5.2 Tissue sources and collection

Five chickens (three young females, one in-lay adult female and one male rooster) were obtained from Sunnybrook Health Sciences Centre, Toronto, Canada. The bursa of Fabricius and thymus were dissected from the young females, while the oviduct and ovaries were dissected from the in-lay adult female and the testis dissected from the rooster. Approximately 75 adult female and five adult male frogs were obtained from Nasco (Atkinson, USA), housed in 20 L aquariums at 20°C and fed a diet of Aquamax Grower 600 trout food (Purina Mills, Gray Summit, USA). Exactly 100 green spotted pufferfish, of unknown age and sex, described as caught wild in Malaysia, were purchased from Aquarium Services Warehouse Outlets (Thornhill, Canada) and housed in a single 120 gallon (453 L) aquarium tank at room temperature. Their markings, size (about 6–7 cm) and behavior matched descriptions of *Tetraodon nigroviridis*. The fish were fed frozen brine shrimp twice a day. Animal handling and euthanization procedures were performed according to protocols approved by the University of Toronto. Tissues collected from each species are listed in Table X. All fish and frogs were anaesthetized by adding 35 ml of a 1:10 clove oil (Rougier Pharma, Mirabel, Canada):ethanol mixture to 1.5 L tank water containing a single animal. Following cessation of movement (15–20 s), they were euthanized by decapitation. Chickens were euthanized by a barbiturate injection followed by swift decapitation.

### 2.5.3 mRNA isolation and cDNA synthesis

Following euthanization of each animal, tissues were dissected immediately (within 10 min), washed in 1X PBS, snap-frozen in liquid nitrogen and stored at -80oC until use. Total RNA was extracted by homogenizing each tissue in Trizol reagent (Invitrogen, Carlsbad, USA), following the manufacturer’s protocol and its integrity checked on 1% agarose-formaldehyde gels. mRNA was purified on oligo(dT) cellulose beads (New England Biolabs, Ipswich, USA) and at least 1 µg of poly(A)+ RNA was used for subsequent cDNA synthesis, using a 0.25:0.75 mix of T18VN oligmer and random N(9) primers. The mixture was incubated at 65oC for 5 min to denature and chilled on ice for 4 min before incubating at room temperature for 10 min to anneal. The reverse transcription reaction contained a 1:1 mixture of 5-(3-aminoallyl) thymidine 5’-triphosphate (Sigma-Aldrich, St. Louis, USA) and thymidine triphosphate, in addition to an equal amount of dATP, dCTP and dGTP, as well as buffer and SuperScript II enzyme (Invitrogen, Carlsbad, USA), which was incubated at 40 min at 42oC. The cDNA products were bound and washed 3 times on QIAquick PCR purification columns (Qiagen, Hilden, Germany) with 80% ethanol and eluted with water.

### 2.5.4 Labeling and hybridization

Following reverse transcription, purified DNA was labeled with N-hydroxy succinimide esters of Cy3 or Cy5 (GE Healthcare Bio-sciences, formerly Amersham Biosciences, Little Chalfont, UK) according to the manufacturer’s recommendations and then quenched with hydroxylamine. Cy3- and Cy5-labelled pairs of cDNA samples were combined and purified from excess dye molecules using QIAquick columns. A hybridization buffer mixture containing 1M NaCl, 0.5% sodium sarcosine, 50 mM methyl ethan sulfonate, pH 6.5 33% formamide and 40 µg of salmon sperm DNA was added to the labeled cDNAs to a final volume of 0.5 mL and injected into an Agilent hybridization chamber. Hybridizations were carried out for 20 to24 h at 42oC in a rotating hybridization oven (Robbins Scientific Corporation, Sunnyvale, USA). Following hybridization, the arrays were washed by rocking for 30s in a buffer containing 6X SSPE and 0.005% sarcosine, followed by a second wash for 30s in 0.06X SSPE before scanning with a 4000A microarray scanner (Molecular Devices, formerly Axon Instruments, Sunnyvale, USA). Each tissue was assayed in duplicate, in fluor-reversed pairs onto different microarrays and the data from each single-channel was averaged (average Pearson correlation between replicate arrays = 0.89).

### 2.5.5 Microarray data processing and normalization

TIFF image files outputted by the 4000A scanner were quantitated using GenePix software (Molecular Devices, Sunnyvale, USA). Local spatial artifacts were removed by spatial detrending, using high-pass filtering with 10% or greater outliers. Variance stabilizing normalization (Huber et al. 2002) was applied to normalize all single channels to each other and arcsinh-transformed. The median value was selected for probes mapping to the same transcript and relative expression ‘ratios’ were calculated by subtracting the median asinh intensity across all tissues (negative values were zeroed). All new microarray data were uploaded to the Gene Expression Omnibus (GEO) database (Barrett et al. 2009), with accession numbers [GEO:GSE12974, GEO:GSE12975, GEO:GSE12976], and are also found on the project website (http://hugheslab.ccbr.utoronto.ca/ supplementary-data/vertebrate\_expression/).

### 2.5.6 Compilation of published human and mouse microarray data sets

Microarray-based expression data for 55 mouse tissues (Zhang et al. 2004) and 60 human tissues (Schadt et al. 2004) were downloaded and mapped to Ensembl version 37 identifiers by aligning probe sequences to Ensembl transcript sequences using BLAT (Kent 2002). For comparisons restricted to just the unique 1-1-1-1-1 orthologs, human and mouse expression data were combined to maximize the number of genes and tissues without a missing value. These profiles were first normalized by VSN (Huber et al. 2002) before combining together.

### 2.5.7 Definition of unique 1-1-1-1-1 orthology between genes from five vertebrate species

Inparanoid (O'Brien et al. 2005) was used to analyze each possible pairwise all-versus-all protein BLAST (Altschul et al. 1990) comparison between all known proteins for each of the five vertebrate species in version 37 of the Ensembl database to delineate pairwise gene orthology relationships. These relationships were then assembled into unique 1-1-1-1-1 ortholog groups across the five species using custom Perl scripts in an approach analogous to that of (Alexeyenko et al. 2006)*.*

### 2.5.8 Gene clustering and gene ontology analysis

GO annotations were downloaded from Ensembl BioMart (Kasprzyk et al. 2004) for each species. Annotations for chicken, frog and pufferfish **(**Tetraodon**)** were further supplemented by mapping the corresponding mouse annotations by any type of orthology as defined by Ensembl. Annotations were up-propagated and terms with few or too many annotated genes were removed as described previously (Zhang et al. 2004). All GO WMW analyses performed across the set of unique orthologs were done with human annotations.

### 2.5.9 Definition of expression conservation events in each gene and tissue using the binary measure

Each set of orthologs in each tissue in each species was split on the basis of their measured normalized expression intensities according to the following thresholds: top 1/2, top 1/3, top 1/4, top 1/5 and top 1/6. Because the human and mouse datasets were designed independently from those of the other three species, there were orthologous genes with missing measurements. In order to facilitate comparison between as many unique orthologs between the five species as possible, variance stabilizing normalization (Huber et al. 2002) was applied to the human and mouse orthologs in order to make them comparable, under the assumption that gene expression is mostly conserved between the mammals relative to the rest of the vertebrate phylogeny. The human and mouse ortholog data were combined by averaging to obtain a set of 3,074 unique orthologs with measured microarray data across ten common tissues.

### 2.5.10 Calculating Shannon entropy as a measure of tissue specificity

Shannon entropy, which measures the degree of overall tissue specificity of a gene, was calculated as described by (Schug et al. 2005). Briefly, the relative expression of a gene gin a tissue trelative to its expression given in Ntissues is defined as:

=

Equation 1

where wg, tis the expression level of the gene gin tissue t. The Shannon entropy of a gene's expression distribution is then calculated as:

Equation 2

This value is expressed in bits and ranges from zero to log2(10) genes expressed in a single tissue and uniformly expressed in all the common tissues examined, respectively.

### 2.5.11 Definition of overlap between nonexonic regions associated with human 1-1-1-1-1 orthologs and Phastcons elements and UCEs

Locations of Phastcons elements (Siepel et al. 2005) and ultraconserved sequences (UCEs) (Bejerano et al. 2004) were downloaded from the hg17 version of the UCSC genome browser (Kuhn et al. 2009) and from the supplementary websites (http://compgen.bscb.cornell.edu/~acs/ conservation/ and http://users.soe.ucsc.edu/~jill/ultra.html, respectively) The number of Phastcons elements and UCEs (and the number of bases) that overlapped all nonexonic sequences, including intergenic sequences, up to 50 kb upstream and downstream of each human 1-1-1-1-1 ortholog was tabulated using custom Perl scripts. The proportion of Phastcons and UCE bases covered in non-coding regions was calculated as the number of bases out of 50 kb of flanking bases upstream and downstream of each gene with coding regions masked out.

### 2.5.12 Alignment of non-coding regions of orthologous genes

All repeat-masked (Smit 1996-2004) intronic, 3' untranslated region and intergenic non-coding sequence upstream and downstream of protein-coding sequences in orthologous groups that I have identified using the Inparanoid algorithm (O'Brien et al. 2005) were downloaded from Ensembl (Hubbard et al. 2009) version 37. A multiple global alignment across all of the genomic sequences within each set of unique orthologs was built using LAGAN (Brudno et al. 2003a) to identify conserved non-coding elements in all species. Up to 50 kb of upstream and downstream sequence, to the point that the transcript of another annotated gene was encountered was used.

A conservation cutoff of 55% in a 50 nucleotide window was initially applied to search for conserved regions between the human and the orthologous genomes. After removing conserved elements that were annotated as exonic, each sequence element in the alignment that was conserved only in a subset of the genomes was extracted and the most parsimonious ancestral reconstruction of each of the sequences was built using Fitch's algorithm (Fitch 1971), treating the gap character as a fifth symbol. This was then used to search against the other orthologous genome(s) using the CHAOS aligner (Brudno et al. 2003b) with very sensitive parameters. The Smith-Waterman threshold was varied between 40 and 60 (60 is the default conservation setting for the BLASTZ alignment program (Schwartz et al. 2003); this setting is both relatively sensitive and very specific, and thus almost all hits above this threshold will be real). This homology filtering step was used to identify non-collinear conserved sequences that may have changed position and orientation relative to exons over evolutionary time. In contrast to the method used by (Sanges et al. 2006), both the mouse and chicken sequence was used as a homology filter. Using the chicken as the base organism, the false positive rate was significantly lowered (only one decoy (i.e. permuted gene identity) sequence with a score above 45), while the true positive rate was unchanged (all of the human/frog and human/pufferfish conserved regions recovered with mouse as the base were also recovered with pufferfish (Tetraodon) as the base).

This approach takes advantage of the conservation of order of the conserved elements when no rearrangements have taken place, and the flexibility of aligning less conserved regions that have been shuffled around in the genome. By running Fitch's algorithm (Fitch 1971) on the aligned sequences unlike the (Sanges et al. 2006) approach of simply using the mouse sequence as input to CHAOS) the sensitivity of our alignment technique was increased. In particular, it was easier to align Tetraodonsequence to the common ancestor of human and chicken genomes than to either genome individually.

### 2.5.13 Compilation and correlation of genic features across unique orthologs

A measure of protein sequence conservation between two species was derived by performing pairwise BLASTP (Altschul et al. 1990) between the protein-coding sequences (downloaded from Ensembl BioMart (Kasprzyk et al. 2004)) of an orthologous gene pair and retaining the bit score. The median bit score was taken as a measure of protein sequence conservation over all species. Average and maximum expression level and the expression rank within tissues were calculated in the expected manner for each gene across the ten common tissues. The number of bases in an aligned conserved element (aligned as described below) was obtained by summing the number of bases in each species within a five-way gapped alignment between all species. The total number of aligned bases in an aligned conserved element is the sum of counts in each species. TF genes within the set of unique orthologs were defined by the presence of a DNA-binding domain in the mouse protein sequence in the Pfam database (Finn et al. 2008). The list of TF genes I compiled can be found here: http://hugheslab.ccbr.utoronto.ca/ supplementary-data/vertebrate\_expression/Additional\_file\_11\_List\_of\_probable\_transcription\_ factors.txt. Conservation of expression was measured by either the binary or Pearson measures, both of which yield a real value (an integer between zero and six in the case of the binary method). With the exception of GO annotations and TF identities, all measures for each gene are found here: http://hugheslab.ccbr.utoronto.ca/supplementary-data/vertebrate\_expression/Additional\_file\_ 7\_3074\_measured\_orthologs\_feature\_matrix.xls. Comparisons of properties with real values were made by Spearman p-value, and these are also found here: http://hugheslab.ccbr.utoronto.ca/ supplementary-data/vertebrate\_expression/Additional\_file\_7\_3074\_measured\_orthologs\_feature\_ matrix.xls and shown graphically here: http://hugheslab.ccbr.utoronto.ca/supplementary-data/vertebrate\_expression/Additional\_file\_12\_Gene\_feature\_table\_clustergram.pdf.

### 2.5.14 Enhancer Element Locator

The nonexonic sequence associated (as defined for the multiple sequence alignment algorithm) with 4,804 human/fish ortholog pairs were scanned for conserved TF binding sites by applying the Enhancer Element Locator (Hallikas et al. 2006) (EEL) program using the default parameters to perform a local pairwise TF binding site alignment. 94 (of 4,898) ortholog pairs in which one of the orthologs appears in the intron of another gene were not aligned. The motif models input into EEL were the 138 models returned by JASPAR webserver (Bryne et al. 2008) on 27 February 2009. The score of only the best alignment from each orthologous gene group was captured and was used to construct a distribution of 4,804 gene group alignment scores. A shuffled (negative control) distribution of EEL scores was also constructed by attempting to align the TF binding sites between the human non-coding sequences in each of the ortholog pairs and those of six non-orthologous Tetraodongenes. These genes were selected among the other 4,803 non-orthologous paired Tetraodongenes to be the six with the most similar amount of non-coding sequence as the proper ortholog, under the constraint that three of the genes had to have less non-coding sequence and three had to have more non-coding sequence. As evidence that the EEL analysis is detecting some very limited degree of conservation, it can be noted that the distribution of EEL scores is slightly higher in real than in randomly assigned orthologs starting around an EEL score of 175; in particular, 1.3% (63 of 4,084) of all EEL scores can be estimated to be non-random by subtracting the proportion of random scores above 175 (39.8%) from the proportion of real scores above 175 (41.1%).

Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B. 2003. Computational detection of cis -regulatory modules. *Bioinformatics* **19 Suppl 2**: ii5-14.

Aggarwal BD, Calvi BR. 2004. Chromatin regulates origin activity in Drosophila follicle cells. *Nature* **430**(6997): 372-376.

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL. 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**(14): e9-15.

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

Amoutzias GD, Robertson DL, Van de Peer Y, Oliver SG. 2008. Choose your partners: dimerization in eukaryotic transcription factors. *Trends in biochemical sciences* **33**(5): 220-229.

Anand S, Wang WC, Powell DR, Bolanowski SA, Zhang J, Ledje C, Pawashe AB, Amemiya CT, Shashikant CS. 2003. Divergence of Hoxc8 early enhancer parallels diverged axial morphologies between mammals and fishes. *Proceedings of the National Academy of Sciences of the United States of America* **100**(26): 15666-15669.

Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *Proceedings of the National Academy of Sciences of the United States of America* **92**(5): 1684-1688.

Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrancois P, Struhl K, Gerstein M, Snyder M. 2009. Mapping accessible chromatin regions using Sono-Seq. *Proceedings of the National Academy of Sciences of the United States of America* **106**(35): 14926-14931.

Bachner-Melman R, Dina C, Zohar AH, Constantini N, Lerer E, Hoch S, Sella S, Nemanov L, Gritsenko I, Lichtenberg P et al. 2005. AVPR1a and SLC6A4 gene polymorphisms are associated with creative dance performance. *PLoS genetics* **1**(3): e42.

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science (New York, NY* **324**(5935): 1720-1723.

Bailey TL, Noble WS. 2003. Searching for statistically significant regulatory modules. *Bioinformatics* **19 Suppl 2**: ii16-25.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA et al. 2009. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic acids research* **37**(Database issue): D885-890.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**(4): 823-837.

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

Bell AC, West AG, Felsenfeld G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**(3): 387-396.

Belting HG, Shashikant CS, Ruddle FH. 1998. Modification of expression and cis-regulation of Hoxc8 in the evolution of diverged axial morphology. *Proceedings of the National Academy of Sciences of the United States of America* **95**(5): 2355-2360.

Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I. 2005. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* **21**(11): 2657-2666.

Benos PV, Bulyk ML, Stormo GD. 2002. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* **30**(20): 4442-4451.

Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.

Berg OG, von Hippel PH. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* **193**(4): 723-750.

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET et al. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**(7): 1266-1276.

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **24**(11): 1429-1435.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America* **99**(2): 757-762.

Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL. 2004. Global nucleosome occupancy in yeast. *Genome biology* **5**(9): R62.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* **8**(7): 1499-1504.

Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET Thurman RE et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D et al. 2006. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome research* **16**(5): 656-668.

Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y. 2008. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS genetics* **4**(11): e1000271.

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

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. 2007. Divergence of transcription factor binding sites across related yeast species. *Science (New York, NY* **317**(5839): 815-819.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**(2): 311-322.

Bradbury EM. 1992. Reversible histone modifications and the chromosome cell cycle. *Bioessays* **14**(1): 9-16.

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. *Science (New York, NY* **165**(891): 349-357.

-. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Quarterly review of biology* **46**(2): 111-138.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S. 2003a. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome research* **13**(4): 721-731.

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S. 2003b. Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19 Suppl 1**: i54-62.

Brutlag D, Schlehuber C, Bonner J. 1969. Properties of formaldehyde-treated nucleohistone. *Biochemistry* **8**(8): 3214-3218.

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research* **36**(Database issue): D102-106.

Bulyk ML, Huang X, Choo Y, Church GM. 2001. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America* **98**(13): 7158-7163.

Bulyk ML, Johnson PL, Church GM. 2002. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res* **30**(5): 1255-1261.

Burke TW, Kadonaga JT. 1996. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev* **10**(6): 711-724.

-. 1997. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev* **11**(22): 3020-3031.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437**(7062): 1153-1157.

Cai S, Lee CC, Kohwi-Shigematsu T. 2006. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nature genetics* **38**(11): 1278-1288.

Carlson M, Laurent BC. 1994. The SNF/SWI family of global transcriptional activators. *Current opinion in cell biology* **6**(3): 396-402.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics* **38**(6): 626-635.

Chan YF, Marks ME, Jones FC, Villarreal G, Jr., Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science (New York, NY* **327**(5963): 302-305.

Chen X, Hughes TR, Morris Q. 2007. RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics* **23**(13): i72-79.

Chervenak AP, Basu P, Shin M, Redmond LC, Sheng G, Lloyd JA. 2006. Identification, characterization, and expression pattern of the chicken EKLF gene. *Dev Dyn* **235**(7): 1933-1940.

Chung WY, Albert R, Albert I, Nekrutenko A, Makova KD. 2006. Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. *BMC bioinformatics* **7**: 46.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science (New York, NY* **301**(5629): 71-76.

Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M. 2001. Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome research* **11**(7): 1175-1186.

Consortium ICGS. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**(7018): 695-716.

Cooper TF, Rozen DE, Lenski RE. 2003. Parallel changes in gene expression after 20,000 generations of evolution in Escherichiacoli. *Proceedings of the National Academy of Sciences of the United States of America* **100**(3): 1072-1077.

Costas J, Casares F, Vieira J. 2003. Turnover of binding sites for transcription factors involved in early Drosophila development. *Gene* **310**: 215-220.

Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS. 2006. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature methods* **3**(7): 503-509.

Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R, Bouffard G, Young A, Masiello C, Green ED, Wolfsberg TG et al. 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proceedings of the National Academy of Sciences of the United States of America* **101**(4): 992-997.

Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. 2009. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome research* **19**(1): 24-32.

Davidson EH. 2006. *The regulatory genome : gene regulatory networks in development and evolution*. Academic, Burlington, MA ; Oxford.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**(10): e314.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science (New York, NY* **295**(5558): 1306-1311.

Deng W, Roberts SG. 2007. TFIIB and the regulation of transcription by RNA polymerase II. *Chromosoma* **116**(5): 417-429.

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* **19**(7): 1114-1121.

Doniger SW, Fay JC. 2007. Frequent gain and loss of functional transcription factor binding sites. *PLoS computational biology* **3**(5): e99.

Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C et al. 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* **16**(10): 1299-1309.

Elango N, Yi SV. 2008. DNA methylation and structural and functional bimodality of vertebrate promoters. *Mol Biol Evol* **25**(8): 1602-1608.

Elgar G, Vavouri T. 2008. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet* **24**(7): 344-352.

Emorine L, Kuehl M, Weir L, Leder P, Max EE. 1983. A conserved sequence in the immunoglobulin J kappa-C kappa intron: possible enhancer element. *Nature* **304**(5925): 447-449.

Ettwiller L, Budd A, Spitz F, Wittbrodt J. 2008. Analysis of mammalian gene batteries reveals both stable ancestral cores and highly dynamic regulatory sequences. *Genome biology* **9**(12): R172.

Ferea TL, Botstein D, Brown PO, Rosenzweig RF. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **96**(17): 9721-9726.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL et al. 2008. The Pfam protein families database. *Nucleic acids research* **36**(Database issue): D281-288.

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science (New York, NY* **312**(5771): 276-279.

Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic zoology* **20**(3): 406-441.

FitzGerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C. 2006. Comparative genomics of Drosophila and human core promoters. *Genome biology* **7**(7): R53.

Frith MC, Hansen U, Weng Z. 2001. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* **17**(10): 878-889.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**(7269): 58-64.

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. *J Mol Biol* **196**(2): 261-282.

Georges AB, Benayoun BA, Caburet S, Veitia RA. 2009. Generic binding sites, generic DNA-binding domains: where does specific promoter recognition come from? *Faseb J*.

Gershenzon NI, Ioshikhes IP. 2005. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* **21**(8): 1295-1300.

Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440**(7081): 242-245.

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research* **17**(6): 877-885.

Giresi PG, Lieb JD. 2009. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods (San Diego, Calif* **48**(3): 233-239.

Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature* **433**(7025): 481-487.

Greer JM, Puetz J, Thomas KR, Capecchi MR. 2000. Maintenance of functional equivalence during paralogous Hox gene evolution. *Nature* **403**(6770): 661-665.

Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57**: 159-197.

Gu X, Zhang Z, Huang W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proceedings of the National Academy of Sciences of the United States of America* **102**(3): 707-712.

GuhaThakurta D, Stormo GD. 2001. Identifying target sites for cooperatively binding factors. *Bioinformatics* **17**(7): 608-621.

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

Hammock EA, Young LJ. 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science (New York, NY* **308**(5728): 1630-1634.

Hanks M, Wurst W, Anson-Cartwright L, Auerbach AB, Joyner AL. 1995. Rescue of the En-1 mutant phenotype by replacement of En-1 with En-2. *Science (New York, NY* **269**(5224): 679-682.

Hanks MC, Loomis CA, Harris E, Tong CX, Anson-Cartwright L, Auerbach A, Joyner A. 1998. Drosophila engrailed can substitute for mouse Engrailed1 function in mid-hindbrain, but not limb development. *Development (Cambridge, England)* **125**(22): 4521-4530.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**(7004): 99-104.

Hay A, Tsiantis M. 2006. The genetic basis for differences in leaf form between Arabidopsis thaliana and its wild relative Cardamine hirsuta. *Nature genetics* **38**(8): 942-947.

Hecht J, Stricker S, Wiecha U, Stiege A, Panopoulou G, Podsiadlowski L, Poustka AJ, Dieterich C, Ehrich S, Suvorova J et al. 2008. Evolution of a core gene network for skeletogenesis in chordates. *PLoS genetics* **4**(3): e1000025.

Hedges SB. 2002. The origin and evolution of model organisms. *Nature reviews* **3**(11): 838-849.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**(7243): 108-112.

Heintzman ND, Ren B. 2009. Finding distal regulatory elements in the human genome. *Curr Opin Genet Dev* **19**(6): 541-549.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39**(3): 311-318.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS et al. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods* **6**(4): 283-289.

Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, Purisima E, Nantel A, Whiteway M. 2008. Transcription factor substitution during the evolution of fungal ribosome regulation. *Molecular cell* **29**(5): 552-562.

Hon G, Wang W, Ren B. 2009. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS computational biology* **5**(11): e1000566.

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L et al. 2009. Ensembl 2009. *Nucleic acids research* **37**(Database issue): D690-697.

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

Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, Gordon L, Branscomb E, Stubbs L. 2006. A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome research* **16**(5): 669-677.

Ihmels J, Bergmann S, Gerami-Nejad M, Yanai I, McClellan M, Berman J, Barkai N. 2005. Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science (New York, NY* **309**(5736): 938-940.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A et al. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**(7011): 946-957.

Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB. 2008. The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. *Cell* **132**(5): 783-793.

Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10**(3): 161-172.

Johansson O, Alkema W, Wasserman WW, Lagergren J. 2003. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics* **19 Suppl 1**: i169-176.

Juven-Gershon T, Hsu JY, Theisen JW, Kadonaga JT. 2008. The RNA polymerase II core promoter - the gateway to transcription. *Current opinion in cell biology* **20**(3): 253-259.

Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. 2004. EnsMart: a generic system for fast and flexible access to biological data. *Genome research* **14**(1): 160-169.

Keene MA, Corces V, Lowenhaupt K, Elgin SC. 1981. DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. *Proceedings of the National Academy of Sciences of the United States of America* **78**(1): 143-146.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**(6937): 241-254.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome research* **12**(4): 656-664.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science (New York, NY* **309**(5742): 1850-1854.

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

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Paabo S. 2004. A neutral model of transcriptome evolution. *PLoS Biol* **2**(5): E132.

Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B. 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**(6): 1231-1245.

Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. 2005. A high-resolution map of active promoters in the human genome. *Nature* **436**(7052): 876-880.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science (New York, NY* **188**(4184): 107-116.

Kingston RE, Narlikar GJ. 1999. ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. *Genes Dev* **13**(18): 2339-2352.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M et al. 2009. The UCSC Genome Browser Database: update 2009. *Nucleic acids research* **37**(Database issue): D755-761.

Kumar PP, Bischof O, Purbey PK, Notani D, Urlaub H, Dejean A, Galande S. 2007. Functional interaction between PML and SATB1 regulates chromatin-loop architecture and transcription of the MHC class I locus. *Nature cell biology* **9**(1): 45-56.

Kwon H, Imbalzano AN, Khavari PA, Kingston RE, Green MR. 1994. Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex. *Nature* **370**(6489): 477-481.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD. 2004. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature genetics* **36**(8): 900-905.

Lee TI, Young RA. 2000. Transcription of eukaryotic protein-coding genes. *Annual review of genetics* **34**: 77-137.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. *Nature* **424**(6945): 147-151.

Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL et al. 2008. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* **6**(2): e27.

Liao BY, Zhang J. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* **23**(3): 530-540.

Lim MM, Wang Z, Olazabal DE, Ren X, Terwilliger EF, Young LJ. 2004. Enhanced partner preference in a promiscuous species by manipulating the expression of a single gene. *Nature* **429**(6993): 754-757.

Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE. 2009. On the origin and spread of an adaptive allele in deer mice. *Science (New York, NY* **325**(5944): 1095-1098.

Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC. 2009. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*.

Liu X, Noll DM, Lieb JD, Clarke ND. 2005. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome research* **15**(3): 421-427.

Lopez-Bigas N, De S, Teichmann SA. 2008. Functional protein divergence in the evolution of Homo sapiens. *Genome biology* **9**(2): R33.

Luger K, Richmond TJ. 1998. The histone tails of the nucleosome. *Curr Opin Genet Dev* **8**(2): 140-146.

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

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

Lutz B, Lu HC, Eichele G, Miller D, Kaufman TC. 1996. Rescue of Drosophila labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved. *Genes Dev* **10**(2): 176-184.

Maerkl SJ, Quake SR. 2007. A systems approach to measuring the binding energy landscapes of transcription factors. *Science (New York, NY* **315**(5809): 233-237.

-. 2009. Experimental determination of the evolvability of a transcription factor. *Proceedings of the National Academy of Sciences of the United States of America* **106**(44): 18650-18655.

Man TK, Stormo GD. 2001. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res* **29**(12): 2471-2478.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Martchenko M, Levitin A, Hogues H, Nantel A, Whiteway M. 2007. Transcriptional rewiring of fungal galactose-metabolism circuitry. *Curr Biol* **17**(12): 1007-1013.

Marza E, Barthe C, Andre M, Villeneuve L, Helou C, Babin PJ. 2005. Developmental expression and nutritional regulation of a zebrafish gene homologous to mammalian microsomal triglyceride transfer protein large subunit. *Dev Dyn* **232**(2): 506-518.

McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, Elgar G. 2006. Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. *Genome research* **16**(4): 451-465.

McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, Payre F, Stern DL. 2007. Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature* **448**(7153): 587-590.

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

Meiklejohn CD, Parsch J, Ranz JM, Hartl DL. 2003. Rapid evolution of male-biased gene expression in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **100**(17): 9894-9899.

Meng X, Brodsky MH, Wolfe SA. 2005. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature biotechnology* **23**(8): 988-994.

Meng X, Wolfe SA. 2006. Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nature protocols* **1**(1): 30-45.

Miele A, Dekker J. 2008. Long-range chromosomal interactions and gene regulation. *Molecular bioSystems* **4**(11): 1046-1057.

Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, Shriver MD, Kingsley DM. 2007. cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell* **131**(6): 1179-1189.

Mintseris J, Eisen MB. 2006. Design of a combinatorial DNA microarray for protein-DNA interaction studies. *BMC bioinformatics* **7**: 429.

Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB. 2006. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS computational biology* **2**(10): e130.

Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature genetics* **36**(12): 1331-1339.

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

Oda-Ishii I, Bertrand V, Matsuo I, Lemaire P, Saiga H. 2005. Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians Halocynthia roretzi and Ciona intestinalis. *Development (Cambridge, England)* **132**(7): 1663-1674.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature genetics* **39**(6): 730-732.

Ohno S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin.

Oliphant AR, Brandl CJ, Struhl K. 1989. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Molecular and cellular biology* **9**(7): 2944-2949.

Ozsolak F, Song JS, Liu XS, Fisher DE. 2007. High-throughput mapping of the chromatin structure of human promoters. *Nature biotechnology* **25**(2): 244-248.

Pabo CO, Sauer RT. 1984. Protein-DNA recognition. *Annu Rev Biochem* **53**: 293-321.

Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T et al. 2008. Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**(3): 422-433.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**(7118): 499-502.

Phillips JE, Corces VG. 2009. CTCF: master weaver of the genome. *Cell* **137**(7): 1194-1211.

Picard D, Schaffner W. 1984. A lymphocyte-specific enhancer in the mouse immunoglobulin kappa gene. *Nature* **307**(5946): 80-82.

Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E et al. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**(4): 517-527.

Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio LA. 2005. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* **85**(6): 774-781.

Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, True JR, Carroll SB. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* **440**(7087): 1050-1053.

Pudimat R, Schukat-Talamazzini EG, Backofen R. 2005. A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics* **21**(14): 3082-3088.

Rajewsky N, Vergassola M, Gaul U, Siggia ED. 2002. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC bioinformatics* **3**: 30.

Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB. 2009. Stepwise modification of a modular enhancer underlies adaptation in a Drosophila population. *Science (New York, NY* **326**(5960): 1663-1667.

Rifkin SA, Kim J, White KP. 2003. Evolution of gene expression in the Drosophila melanogaster subgroup. *Nature genetics* **33**(2): 138-144.

Roh TY, Cuddapah S, Zhao K. 2005. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev* **19**(5): 542-552.

Sabo PJ, Hawrylycz M, Wallace JC, Humbert R, Yu M, Shafer A, Kawamoto J, Hall R, Mack J, Dorschner MO et al. 2004a. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America* **101**(48): 16837-16842.

Sabo PJ, Humbert R, Hawrylycz M, Wallace JC, Dorschner MO, McArthur M, Stamatoyannopoulos JA. 2004b. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proceedings of the National Academy of Sciences of the United States of America* **101**(13): 4537-4542.

Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A et al. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature methods* **3**(7): 511-518.

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

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **8**(6): 424-436.

Sanges R, Kalmar E, Claudiani P, D'Amato M, Muller F, Stupka E. 2006. Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome biology* **7**(7): R56.

Sankaran VG, Xu J, Ragoczy T, Ippolito GC, Walkley CR, Maika SD, Fujiwara Y, Ito M, Groudine M, Bender MA et al. 2009. Developmental and species-divergent globin switching are driven by BCL11A. *Nature* **460**(7259): 1093-1097.

Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America* **103**(5): 1412-1417.

Schadt EE, Edwards SW, GuhaThakurta D, Holder D, Ying L, Svetnik V, Leonardson A, Hart KW, Russell A, Li G et al. 2004. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome biology* **5**(10): R73.

Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ, Jr. 2005. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome biology* **6**(4): R33.

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

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **451**(7178): 535-540.

Sekinger EA, Moqtaderi Z, Struhl K. 2005. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Molecular cell* **18**(6): 735-748.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter D, Kingsley DM. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**(6984): 717-723.

Sharon E, Lubliner S, Segal E. 2008. A feature-based approach to modeling protein-DNA interactions. *PLoS computational biology* **4**(8): e1000154.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, NY* **309**(5741): 1728-1732.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**(8): 1034-1050.

Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED. 2004. Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC bioinformatics* **5**: 129.

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

Smit A, Hubley, R and Green, P. 1996-2004. RepeatMaker Open-3.0 [<http://www.repeatmasker.org]>.

Solomon MJ, Varshavsky A. 1985. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proceedings of the National Academy of Sciences of the United States of America* **82**(19): 6470-6474.

Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* **16**(1): 16-23.

Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* **403**(6765): 41-45.

Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, NY* **302**(5643): 249-255.

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **99**(7): 4465-4470.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101**(16): 6062-6067.

Tadepally HD, Burger G, Aubry M. 2008. Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. *BMC evolutionary biology* **8**: 176.

Talukder S. 2007. Exploring the DNA-binding specificities of mouse transcription factors. p. 111 leaves. University of Toronto, 2007.

Tanay A, Regev A, Shamir R. 2005. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **102**(20): 7203-7208.

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**(6950): 788-793.

Tirosh I, Reikhav S, Levy AA, Barkai N. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science (New York, NY* **324**(5927): 659-662.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics* **39**(1): 31-40.

True JR, Carroll SB. 2002. Gene co-option in physiological and morphological evolution. *Annual review of cell and developmental biology* **18**: 53-80.

True JR, Haag ES. 2001. Developmental system drift and flexibility in evolutionary trajectories. *Evolution & development* **3**(2): 109-119.

Tsai KL, Huang CY, Chang CH, Sun YJ, Chuang WJ, Hsiao CD. 2006. Crystal structure of the human FOXK1a-DNA complex and its implications on the diverse binding specificity of winged helix/forkhead proteins. *J Biol Chem* **281**(25): 17400-17409.

Tsong AE, Tuch BB, Li H, Johnson AD. 2006. Evolution of alternative transcriptional circuits with identical logic. *Nature* **443**(7110): 415-420.

Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD. 2008. The evolution of combinatorial gene regulation in fungi. *PLoS Biol* **6**(2): e38.

Tuteja G, Kaestner KH. 2007a. Forkhead transcription factors II. *Cell* **131**(1): 192.

-. 2007b. SnapShot: forkhead transcription factors I. *Cell* **130**(6): 1160.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**(4): 252-263.

Venkatesh B, Yap WH. 2005. Comparative genomics using fugu: a tool for the identification of conserved vertebrate cis-regulatory elements. *Bioessays* **27**(1): 100-107.

Veraksa A, Del Campo M, McGinnis W. 2000. Developmental patterning genes and their conserved functions: from model organisms to humans. *Molecular genetics and metabolism* **69**(2): 85-100.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**(7231): 854-858.

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature genetics* **40**(2): 158-160.

Wagner GP, Lynch VJ. 2008. The gene regulatory logic of transcription factor evolution. *Trends in ecology & evolution (Personal edition)* **23**(7): 377-385.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics* **40**(7): 897-903.

Warner JB, Philippakis AA, Jaeger SA, He FS, Lin J, Bulyk ML. 2008. Systematic identification of mammalian regulatory motifs' target genes and functions. *Nature methods* **5**(4): 347-353.

Warren CL, Kratochvil NC, Hauschild KE, Foister S, Brezinski ML, Dervan PB, Phillips GN, Jr., Ansari AZ. 2006. Defining the sequence-recognition profile of DNA-binding molecules. *Proceedings of the National Academy of Sciences of the United States of America* **103**(4): 867-872.

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

Waterston RH Lindblad-Toh K Birney E Rogers J Abril JF Agarwal P Agarwala R Ainscough R Alexandersson M An P et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915): 520-562.

Wei Y, Yu L, Bowen J, Gorovsky MA, Allis CD. 1999. Phosphorylation of histone H3 is required for proper chromosome condensation and segregation. *Cell* **97**(1): 99-109.

Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T et al. 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**(7180): 796-801.

Whitehead A, Crawford DL. 2006. Variation within and among species in gene expression: raw material for evolution. *Mol Ecol* **15**(5): 1197-1211.

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, Fisher EM, Tavare S, Odom DT. 2008. Species-specific transcription in mice carrying human chromosome 21. *Science (New York, NY* **322**(5900): 434-438.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. *Nature* **430**(6995): 85-88.

Wittkopp PJ, Stewart EE, Arnold LL, Neidert AH, Haerum BK, Thompson EM, Akhras S, Smith-Winberry G, Shefner L. 2009. Intraspecific polymorphism to interspecific divergence: genetics of pigmentation in Drosophila. *Science (New York, NY* **326**(5952): 540-544.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**(1): e7.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**(3): 206-216.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **20**(9): 1377-1419.

Wu C, Bingham PM, Livak KJ, Holmgren R, Elgin SC. 1979. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell* **16**(4): 797-806.

Wunderlich Z, Mirny LA. 2009. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet* **25**(10): 434-440.

Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. 2007. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proceedings of the National Academy of Sciences of the United States of America* **104**(17): 7145-7150.

Xu F, Zhang K, Grunstein M. 2005. Acetylation in histone H3 globular domain regulates gene expression in yeast. *Cell* **121**(3): 375-385.

Xue L, Noll M. 1996. The functional conservation of proteins in evolutionary alleles and the dominant role of enhancers in evolution. *Embo J* **15**(14): 3722-3731.

Yanai I, Graur D, Ophir R. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *Omics* **8**(1): 15-24.

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

Yaragatti M, Basilico C, Dailey L. 2008. Identification of active transcriptional regulatory modules by the functional assay of DNA from nucleosome-free regions. *Genome research* **18**(6): 930-938.

Zapata A, Diez B, Cejalvo T, Gutierrez-de Frias C, Cortes A. 2006. Ontogeny of the immune system of fish. *Fish Shellfish Immunol* **20**(2): 126-136.

Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E et al. 2004. The functional landscape of mouse gene expression. *Journal of biology* **3**(5): 21.

Zhang Y, Reinberg D. 2001. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev* **15**(18): 2343-2360.

Zhang Y, Sturgill D, Parisi M, Kumar S, Oliver B. 2007. Constraint and turnover in sex-biased gene expression in the genus Drosophila. *Nature* **450**(7167): 233-237.

Zhao X, Huang H, Speed TP. 2005. Finding short DNA motifs using permuted Markov models. *J Comput Biol* **12**(6): 894-906.

Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U et al. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics* **38**(11): 1341-1347.

Zhou Q, Liu JS. 2004. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* **20**(6): 909-916.

Zhou Q, Wong WH. 2004. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences of the United States of America* **101**(33): 12114-12119.