CHAPTER 5

Evolution of Genomic Expression

Bernardo Lemos, Christian R. Landry, Pierre Fontanillas, Susan P. Renn, Rob Kulathinal, Kyle M. Brown, and Daniel L. Hartl

Introduction

Genomic regulation is key to cellular differentiation, tissue morphogenesis, and development. Increasing evidence indicates that evolutionary diversity of phenotypes—from cellular to organismic—may also be, in large part, the result of variation in the regulation of genomic expression.

In this chapter we explore the complexity of gene regulation from the perspective of single genes and whole genomes. The first part describes the major factors affecting gene expression levels, from rates of gene transcription—as mediated by promoter–enhancer interactions and chromatin modifications—to rates of mRNA degradation. This description underscores the multiple levels at which genomic expression can be regulated as well as the complexity and variety of mechanisms used. We then briefly describe the major experimental and computational biology techniques for analyzing gene expression variation and its underlying causes. The final section reviews our understanding of the role of regulatory variation in evolution, including the molecular evolution and population genetics of noncoding DNA, as well as the inheritance and phenotypic evolution of levels of mRNA abundance.

The Complex Regulation of Genomic Expression

The regulation of gene expression is a complex and dynamic process. It is not a simple matter to turn a gene on and off, let alone precisely regulate its level of expression. Regulation can be accomplished through various mechanisms at nearly every step of the process of gene expression. Furthermore, each mechanism may require a variety of elements, including DNA sequences, RNA molecules, and proteins, acting in combination to determine the final amount, timing, and location of functional gene product. The complexity of regulation is even more evident when it is considered in the context of evolution and from the standpoint of integrated gene expression across the genome. The dynamic process of genomic expression is not strictly fixed but can be context dependent, responding to cellular (genomic) or environmental influences. Through successive generations, the interplay of these mechanisms evolves, thus generating a selectively advantageous amount and/or location of functional gene product.

Most of the cellular elements regulating gene expression can be divided into two basic categories: *cis-* and *trans*-acting factors, from the Latin meaning "on the same side" and "on the opposite side," respectively. Both *cis*- and *trans*-acting regulatory elements may contribute to the various mechanisms of gene expression regulation. Strictly, *cis* and *trans* effects do not refer to the physical location of the regulatory element, but are rather operationally defined in terms of the way these regulatory elements segregate genetically with respect to the gene that is the target of the regulatory activity. Promoters, enhancers, regulatory introns, and 3′ regulatory sequences are examples of *cis*-regulatory elements. This is because these elements are located within the gene locus itself or in close proximity to it, such that they are generally inherited together as a unit (i.e., the probability of recombination between the regulatory elements and the gene's structural sequences is virtually zero).

Trans-acting factors include proteins and RNAs derived from distant sites in the genome that act as regulatory elements. These can be either on different chromosomes or on the same chromosome far away from the gene locus, such that they can be independently inherited (i.e., the recombination frequency is virtually 50%). Specific factors necessary for initiating or blocking transcription, or proteins that allow for appropriate gene-specific mRNA trafficking and stability, are just a few of the myriad *trans*-acting factors that contribute to gene expression.

The definition of *cis* and *trans* with respect to segregation is important because these two types of elements work in concert and often there is no clear functional (or even positional) distinction between *cis* and *trans*. *Trans*-acting factors bind to or interact with *cis*-regulatory sequences of the DNA and RNA. It is therefore the interplay of *cis*- and *trans*-acting loci that determines the amount of functional gene product. Indeed, this level of complexity and interaction provide the very substrate for the evolution of regulation as various *cis* and *trans* factors are brought together through segregation and recombination. When combined, elements that have evolved under differing selective pressures may produce novel phenotypes, thus offering novel substrates for natural selection.

Early work attempted to categorize genetic mutations affecting a particular phenotype as either structural or regulatory (e.g., Wilson et al. 1977). This separation was motivated by the intuition that some proteins have clearly defined structural roles (e.g., collagen), while the function of other proteins lies primarily in the regulation of other genes (e.g., muscle-specific transcription factors). However, it should be stressed that this classification of gene loci and their mutations breaks down for most genes. This is because many proteins (structural) act in *trans* to alter the expression level of other genes (regulatory), such that a single mutation can have both structural and regulatory consequences. Work by Yvert and colleagues (2003) clearly exemplifies this point. These authors mapped a large number of *trans*-acting loci affecting gene expression differences between two strains of yeast, and found that genes with a large variety of molecular functions, such as enzymes, signal transducers, and cytoskeleton-binding proteins, could influence gene expression levels. Accordingly, amongst *trans*-acting loci there is no specific enrichment for transcription factors.

In the following section, we outline the overall process of gene expression by indicating the mechanisms known to regulate gene product level, location, or timing at various stages of transcription, translation, and post-translation. Transcriptional regulation of gene expression can occur at the level of genomic DNA prior to transcription, or at the step of transcription, when the RNA is being produced. Post-transcription regulation occurs through several mechanisms affecting the processing, stability, and/or localization of the mRNA. The amount of functional gene product can also be regulated post-translationally. For each stage in the process of gene expression, we provide examples of a few well-studied mechanisms of regulation, and attempt to identify examples of both *cis*-acting and *trans*-acting elements involved in each of the regulatory mechanisms. Figure 5.1 gives examples of the different stages along the path to functional protein during which genomic expression may be regulated. In most instances, the evolutionary acquisition and consequences of these mechanisms have yet to be addressed. Most research regarding the *molecular evolution* of gene or genomic expression has focused on transcription factors (TF), promoter sequences, and TF-binding sites within promoters. Similarly, most research regarding the *phenotypic evolution* of genomic expression has focused on the evolution of the mRNA abundance phenotype, most often without an explicit connection to evolutionary variation in specific underlying mechanisms associated with variation in this phenotype.

Classical Transcriptional Regulation in *Cis* and *Trans*

Promoters, enhancers, repressors, transcription factors, and regulatory proteins

Promoters are among the most thoroughly studied and best understood regulators of gene expression (Ptashne and Gann 2002; Thomas and Chiang 2006). In eukaryotes, the promoter is defined as the DNA region within a few hundred base pairs upstream of the transcription start site, the section of DNA where the basic machinery of gene expression is assembled. The promoter region encodes various sequence motifs where transcription factors bind along with RNA polymerase II to initiate transcription (Smale and

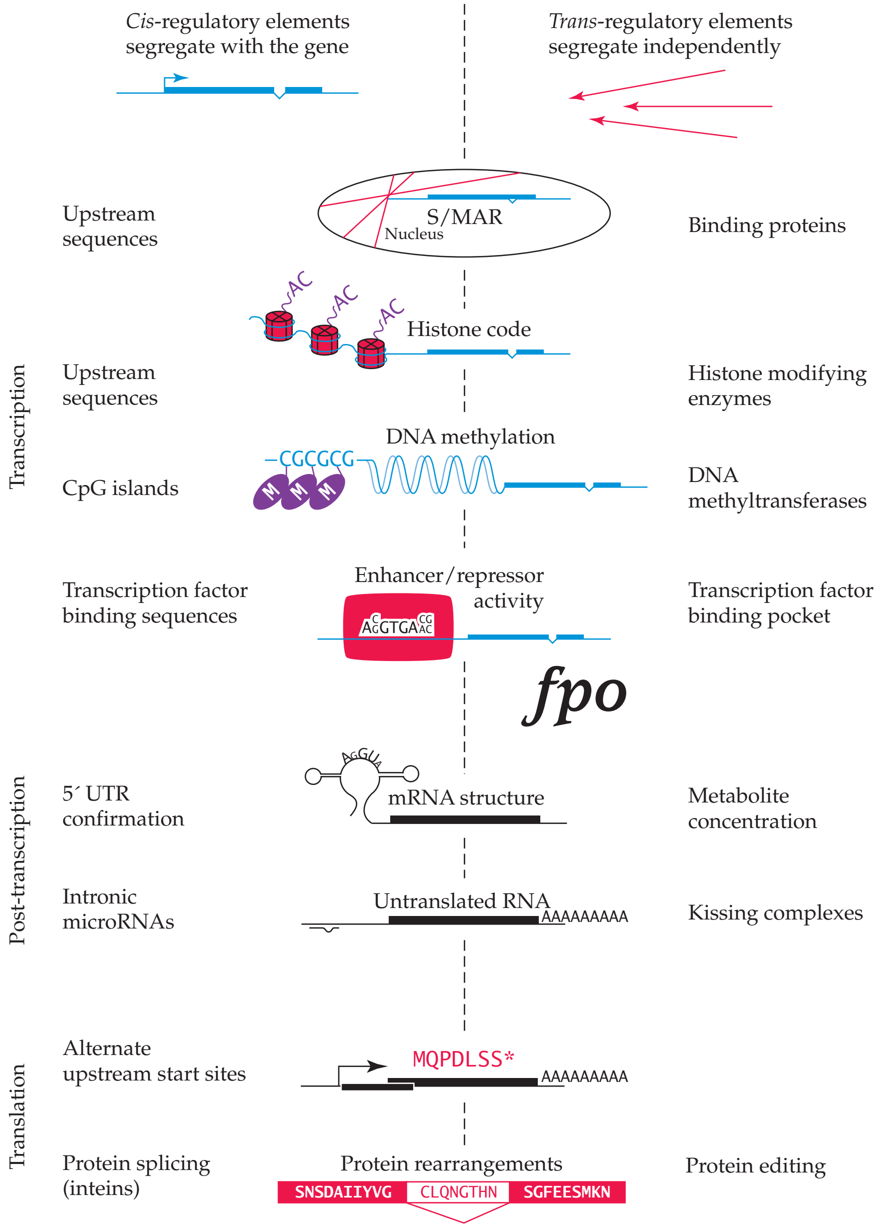

Figure 5.1 Some of the *cis* and *trans* elements associated with a variety of regulatory mechanisms acting at three levels of the gene expression cascade (see text for further details).

Kadonaga 2003). The initiation and rate of transcription depend on the interplay of many DNA and protein elements. Ultimately, a gene's transcription level reflects the interactions between various activating and inhibitory complexes assembled not only at the promoter region, but also at various enhancer and silencer sites along the chromatin (Wray et al. 2003).

The core promoter includes DNA elements that can extend about 35 nucleotides upstream and/or downstream of the transcription initiation site (Smale and Kadonaga 2003). The core is obviously the archetype of *cis*-regulatory factors. At least six core promoter elements have been discerned so far (Gershenzon et al. 2006): TATA box, Initiator (Inr), Downstream Promoter Element (DPE), TFIIB Recognition Element (BRE), Downstream Core Element (DCE), and Motif Ten Element (MTE). Although earlier studies suggested that the structure of the core promoter might be highly conserved throughout the eukaryotes, tremendous diversity is now evident. First, the sequence and the position of DNA motifs are variable within and between species (Smale and Kadonaga 2003; Bazykin and Kondrashov 2006; FitzGerald et al. 2006). Second, not all core promoter elements are systematically associated with all gene promoters; instead combinations of a subset of promoter elements are more frequently observed. In *Drosophila melanogaster*, TATA box, Inr, DPE, and MTE are found in, respectively, 16%, 66%, 22%, and 10% of the genes (Gershenzon et al. 2006). In mammalian promoters, the TATA box element is also present in a minority of genes, but shows substantial sequence conservation and is commonly associated with tissue-specific expression (Carninci et al. 2006). Surprisingly, the presence of the TATA box is also associated with elevated rates of gene expression divergence among yeast species (Tirosh et al. 2006). On the other hand, TATA-less promoters, which are often enriched in CpG islands, seem to be particularly rapidly evolving in mammals (Carninci et al. 2006).

The specific proteins that bind to the core promoters are perhaps the most fundamental *trans*-acting factors regulating gene expression. They include the general transcription factor TFIID, which recognizes the promoter and coordinates the assembly of the remaining general transcription factors (TFIIA, TFIIB, TFIIE, and TFIIH). TFIID is itself a large protein complex formed by the TATA-box binding protein (TBP) as well as several transcription-associated factors (TAFs). The latter include coactivators capable of propagating signals from distant enhancer or repressor elements to the promoter site. In a stereotyped sequence of protein binding, TBP, TFIID, TFIIA, and TFIIB must bind to the promoter region first in order to recruit RNA polymerase II, TFIIE, TFIIH, TFIIF, and other factors necessary to initiate transcription (Tjian 1996; Smale and Kadonaga 2003). These are the *trans*-acting partners to *cis*-occurring promoters and/or enhancer sequences. These proteins are remarkably conserved across the eukaryotes.

Eukaryotic enhancer and repressor elements are additional sites that influence gene expression and may occur several hundred to several thousand base pairs from the promoter site. In many cases, the various enhancer or repressor elements act independently of each other. Each such enhancer

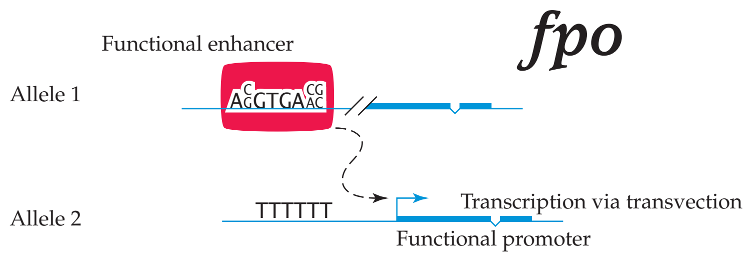

Figure 5.2 Transvection as demonstrated in Drosophila. The enhancer region from one allele can act in *trans* to affect the expression of the allele on the other chromosome. This has been demonstrated using one allele lacking a functional RNA polymerase binding site (so that transcription cannot be initiated) and another allele in which tissue-specific enhancer elements have been mutated. In combination, the functional portions of these genes are able to complement each other and transcription on one chromosome is directed by enhancer sequences on the paired chromosome.

or repressor receives and integrates various signals from regulatory proteins that recognize specific binding sites; these signals are subsequently transmitted to the transcriptional machinery located at the promoter. The interaction of all enhancer elements, repressor elements, and other factors at the promoter results in the precise regulation of the timing, location, and level of gene expression.

Although more has been learned about the role of promoters and enhancers in the regulation of transcription than about the role of any other regulatory element, the elusive phenomenon of transvection reminds us how little we know about even this most basic mechanism of gene regulation. Transvection was first described in 1954 by E. B. Lewis in the context of two mutations that complemented each other in spite of both being within the *Drosophila Ubx* locus (see Duncan 2002 and references therein). This was later recognized to arise from somatic pairing between one allele with a loss of function mutation in its regulatory sequence and another allele with a loss of function mutation in its coding sequence (Figure 5.2). It remains to be understood how regulatory elements in one allele regulate the expression of its homolog on the other chromosome.

Epigenetic Regulation and Chromatin Modifications

Appropriate chromatin conformation is required for access and binding of regulatory proteins to the DNA. The nucleosome is the fundamental repeating unit of chromatin. Made up of 146 base pairs of DNA wrapped around an octamer of conserved core histone proteins, nucleosomes are linked together to form a helical fiber. Each histone contains numerous sites for potential modifications, which have been hypothesized to act in a combinatorial code to mark a region for potential activation or silencing. These marks extend the information potential of the genetic code to provide a so-called epigenetic memory, which plays a major role in regulating cell fate decisions. These stable, epigenetic changes persist through mitosis and in some cases through meiosis. Consideration of epigenetic gene regulation has led to models for the inheritance of acquired epigenetic variations—models in which environmental stimuli induce heritable modifications that might result in adaptive responses to the stimuli (Jablonka and Lamb 1989; Jablonka and Lamb 2002; Gorelick 2005).

DNA methylation

DNA methylation is the most well-understood form of epigenetic gene regulation. First proposed in 1975 (Holliday and Pugh 1975), it has since been intensively characterized in mammals and plants (Jaenisch and Bird 2003; Scott and Spielman 2004). The establishment and maintenance of methylation is required for the normal development and cell differentiation of many organisms. In mammals, DNA methylation occurs predominantly at cytosines in CpG dinucleotides, and several enzymes (DNA methyltransferases) are responsible for de novo methylation and maintenance of methylation marks during mitosis and meiosis. Little is known, however, about the sequences and conditions that direct methylation activity. With respect to a gene under control of DNA methylation, the CpG islands and surrounding sequences that direct the specificity of methylation are *cis*-acting factors, while the methylation enzymes are *trans*-acting factors. In general, increased methylation is associated with down-regulation of gene expression (Wolffe and Matzke 1999), and unmethylated CpG islands lead to increased transcription, although exceptions to this have been described (e.g., Herman et al. 2003). In addition, DNA methylation states have been shown to be sensitive to environmental factors (Jaenisch and Bird 2003) and also to have long lasting effects on behavior (Weaver et al. 2004; Weaver et al. 2005; Feil 2006). Finally, DNA methylation is found in most transposable elements in Arabidopsis and primates (Lippman et al. 2004; Meunier et al. 2005), suggesting its role as a defense mechanism preventing the expression of these elements.

Yeast, worms, and flies have generally been thought to lack DNA methylation systems. In the case of fruit flies, this notion has been challenged by the availability of whole genome sequences. A single DNA methyltransferase has been identified in the genome of *Drosophila melanogaster* (Hung et al. 1999), and experimental work has detected low levels (<0.5%) of cytosine methylation in the fly (Lyko et al. 2000). Moreover, methylated sequences in the fly were found to be associated with CpT or CpA dinucleotides, in sharp contrast to the canonical CpG motif often found methylated in mammals (Kunert et al. 2003).

DNA methylation is particularly interesting in the context of parent-of-origin-dependent inheritance and genomic imprinting (Hajkova et al. 2002; Delaval and Feil 2004). In genomic imprinting, the expression of a subset of genes depends on "marked" maternal and paternal alleles that are recognized and differentially regulated by the transcriptional machinery. Methylation of CpG dinucleotides is the basis for the parent-specific mark. The evolutionary relevance of genomic imprinting has been examined in detail from a theoretical standpoint (see Wilkins and Haig 2003 and McDonald et al. 2005 for recent reviews). Taken together, these observations underscore the relevance of DNA methylation in the evolution of genomic expression.

Histone modifications

Five classes of eukaryotic histones are known (H1, H2A, H2B, H3, and H4), all of which are lysine/arginine rich and have a globular domain that facilitates their assembly with DNA to form chromatin. Histones also have a charged aminoNH_2 terminus (the so-called histone tail) that is subject to various post-translational modifications (e.g., acetylation, methylation, and phosphorylation), which influence chromatin structure and gene expression. It has been conjectured (Jenuwein and Allis 2001; Felsenfeld and Groudine 2003; Fischle et al. 2003) that the kind of histones, and their packing, location, and post-translational modification make up an "epigenetic code," which is used by the cell to specify a precise and stable gene expression profile.

Histone deacetylases, histone acetyltransferases, and histone methyltransferases make up some of the *trans*-acting elements involved in regulating histone modifications that ultimately influence gene expression. For instance, chromatin enriched for acetylated histones is generally thought to be "open" and accessible to transcription factors, thereby rendering its constituent genes transcriptionally active, or potentially so (Grewal and Moazed 2003). Conversely, chromatin enriched for non-acetylated histones is generally thought to be more "condensed," thereby making its constituent genes inaccessible to transcription factors and therefore silenced. Similarly, chromatin enriched for methylated histones is generally thought to be less transcriptionally active than other regions. Furthermore, DNA and histone methylation appear to maintain a repressed chromatin state in plants and vertebrates, although such links appear to be weaker in insects. Finally, we note that whereas histone modifications are known to persist through cell division, much of the histone code may be erased in meiosis; therefore, epigenetic memory is usually thought to result from other DNA and chromatin modifications.

As a cautionary note we emphasize that recent functional genomic work has questioned the specificity of the effects of histone modifications on gene expression. For instance, Dion and colleagues (2005) examined the effect of acetylation of four lysine residues in the tail of histone H4 in yeast. They constructed yeast strains containing up to three lysine-to-arginine mutations in the histone H4 tail, thus preventing acetylation while retaining the positive charge. All the single lysine-to-arginine substitutions showed quite similar gene expression changes, irrespective of the particular lysine altered. Similarly, all combinations of double mutants showed similar changes. This lead Dion and coworkers (2005) to propose that histone H4 acetylation has a simple cumulative effect on yeast gene expression, arguing against a more complex model in which combinations of acetylated lysines "code" for unique expression profiles. Furthermore, in *D. melanogaster*, a systematic mapping of methylation and acetylation of H3 and H4 histones suggested an "all-or-none" pattern (Shübeler et al. 2004; Liu et al. 2005): while the active genes tended to be marked by all the assayed modifications, the non-transcribed genes tended to have no histone marks. A similar, rather simple histone code has also been suggested for budding yeast (Kurdistani et al. 2004; Dion et al. 2005) and mammals (Bernstein et al. 2005). Nonetheless, yeast also shows a more subtle pattern of modification in which clustering of genes with similar patterns of acetylation distinguishes groups of coexpressed genes that are functionally related (Kurdistani et al. 2004). These studies are also modifying our views of the correlation between histone marks and gene expression by demonstrating that both hyper- and hypoacetylation of histones may be associated with gene activity.

Chromosome territories and nuclear architecture

Gene expression is most often studied from the standpoint of promoters, enhancers, suppressors, and local epigenetic modifications. Less frequently studied, however, is the impact of higher-order chromosome structure and nuclear organization on gene expression. Nevertheless, a growing awareness of the relevance of spatial chromosome dynamics in genomic expression, together with technological developments, has stimulated a surge in interest and research in this area (e.g., Bolzer et al. 2005; Harmon and Sedat 2005; Pickersgill et al. 2006).

The nucleolus is perhaps the best-known and most prominent structural feature in the nucleus of most plant and animal cells. This cytological structure is the site where the ribosomal DNA (rDNA) regions of several chromosomes come together and rRNA transcription takes place (see Santoro 2005 for a review). Interestingly, there is substantial natural genetic variation in the amount of methylation observed in rDNA regions of different Arabidopsis strains (Riddle and Richards 2002). This natural variation in rDNA methylation may serve as yet another source of genetic variation in gene expression upon which natural selection can act.

In a less well-described regulatory mechanism, eukaryotic genomes are functionally compartmentalized by attachment to the supporting nuclear matrix (Bode et al. 2003). The overall dynamics of this structure are mediated in part by elements of 300 base pairs to several kilobases named scaffold/matrix-attachment regions (S/MARs)(e.g., Heng et al. 2004). In conjunction, S/MAR-binding proteins act in *trans* to regulate gene expression. Genome-wide predictions identify a large number of S/MARs (Frisch et al. 2002; Glazko et al. 2003) associated with enhancement as well as repression of gene expression.

On a smaller scale, locus control regions (LCR) can be identified that coordinately regulate promoters of several related genes spread over hundreds of kilobases. This is achieved by close juxtaposition of different chromosomal regions in the nucleus. This mechanism was recently shown to be relevant in the regulation of genes involved in T-helper cell differentiation

(Spilianakis et al. 2005; Spilianakis and Flavell 2006). To address the evolutionary implications of such regulation, more information is needed about the genome-wide prevalence and impact of regulatory processes involving nuclear architecture, as well as a detailed mechanistic understanding of individual elements and their interactions. These processes undoubtedly depend on modification of chromatin structure, and they suggest that dispersed multigene complexes are coregulated in part by structural colocalization within the nucleus. How such structures vary across populations and species remains virtually unknown.

The mechanistic complexity of epigenetic gene regulation is perhaps best illustrated by paramutation—a phenomenon first described in peas (Bateson and Pellew 1915), most thoroughly studied in maize (Brink 1959; Brink 1973; Chandler et al. 2000; Hollick and Chandler 2001), and more recently discovered in mammals (Herman et al. 2003; Rassoulzadegan et al. 2006). Paramutation involves heritable changes in gene activity without changes in DNA sequence. The change in gene activity is mediated by heritable epigenetic modification induced by cross-talk between allelic loci (Figure 5.3). Paramutation and paramutation-like phenomena do not adhere to rules of Mendelian inheritance. Our knowledge of the underlying mechanisms is limited, but evidence suggests a complex interplay of many epigenetic processes, such as RNA silencing, physical pairing of homologous chromosomal regions, and chromatin modifications. The term paramutation has come to describe many phenomena in which communication between two alleles or homologous sequences establishes distinct, heritable epigenetic states (Chandler and Stam 2004; Stam and Mittelsten Scheid 2005).

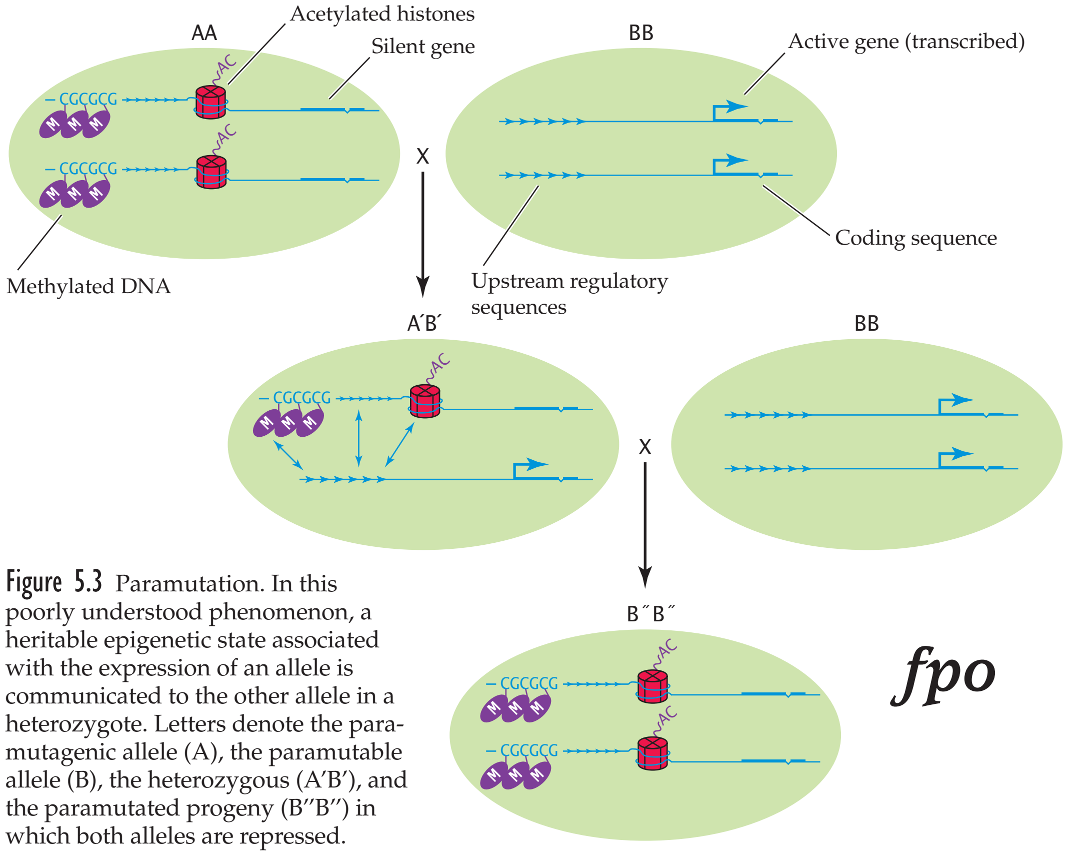

Figure 5.3 Paramutation. In this poorly understood phenomenon, a heritable epigenetic state associated with the expression of an allele is communicated to the other allele in a heterozygote. Letters denote the paramutagenic allele (A), the paramutable allele (B), the heterozygous (A′B′), and the paramutated progeny (B″B″) in which both alleles are repressed. Other F2 products of the backcross of the F1 A′B′ to BB are not shown, but would occur depending on the efficiency of the paramutation.

Post-Transcriptional Regulation

Once transcription has begun, the cell has a variety of post-transcriptional mechanisms to regulate the final amount of functional gene product. In this section we briefly outline some of the best-known mechanisms of post-trancriptional regulation of genomic expression, all of which involve the complex interaction of *cis*- and *trans*-acting factors.

The initial and most basic form of post-transcriptional regulation involves the premature termination of transcription. This gene regulatory mechanism is common in bacteria (Merino and Yanofsky 2005) and has been best studied in the transcription of the HIV-I genome in host cells (Kessler and Mathews 1992). It may also play a role in gene regulation in eukaryotes, where premature termination of transcription typically results from the nascent mRNA folding into secondary structures that are recognized by the cellular apparatus (Muhlrad and Parker 1994; Arigo et al. 2006).

When a full-length RNA transcript is made, it is modified in several ways that determine its cellular fate. Immediately after transcription, a methylated guanine nucleotide is added to the 5′ end of all mRNA transcripts (5′ methylguanosine cap). This feature of mature mRNA is important for the initiation of translation (Alberts et al. 2002). On the 3′ end of the immature transcript, the RNA is cleaved at a specific site and a poly(A) tail is added. This poly(A) tail, usually about 100–200 nucleotides long, is important for regulating the export of the mRNA out of the nucleus, regulating the stability and half-life of the transcript, and ensuring efficient translation at the ribosome (Ross 1995; Alberts et al. 2001).

After transcription is complete, RNA splicing removes noncoding introns and joins together neighboring exons. Alternative splicing, by joining together different exons to create the mature transcript, can produce many different proteins from a single gene. An extreme example of this is the Drosophila *Dscam* axon guidance receptor, which can potentially generate 38,016 different protein isoforms (Graveley 2005; Crayton et al. 2006). Appropriately spliced and edited transcripts are then regulated further by transport to appropriate parts of the cell; this intracellular transport is often mediated by the untranslated 5′ and 3′ ends of mRNAs.

In a regulatory mechanism know as mRNA editing, the nucleotide sequence of the transcript can be changed at specific places. Ordinarily, this process leads to the production of protein variants differing in one or a few amino acids, but, in its most dramatic manifestations, can determine the splicing and ultimate cellular location of the final gene product (Maas and Rich 2000). An example of mRNA editing is the modification of adenosine to inosine, which is recognized as guanosine by the cellular machinery (Maas and Rich 2000; Barbon et al. 2003). This modification is mediated by the *trans*-regulatory activity of adenosine deaminases on an mRNA editing-site sequence, which represents the *cis*-regulatory component of this mechanism of regulation.

In addition, mRNA degradation plays an important role in post-translational gene regulation. For example, the rate of mRNA degradation has been shown to vary widely among genes (Wang et al. 2002; Foat et al. 2005). Degradation of functional, mature mRNA is regulated by mRNA binding proteins and is specified by various features, including sequence motifs and mRNA secondary structures, usually in the 3′ untranslated region of the mRNA (Ross 1995). Another recent study found that genes with tightly folded 5′ untranslated regions may have lower rates of translation, lower protein and mRNA abundances, and shorter half-lives (Ringner and Krogh 2005).

Typically found in the 5′ untranslated regions of bacterial mRNAs, riboswitches are structural elements that regulate gene expression post-transcriptionally (Winkler et al. 2002; Tucker and Breaker 2005; Winkler and Breaker 2005). Riboswitches regulate gene expression by binding to small metabolites, without the involvement of other co-factors.

Finally, small regulatory RNAs, known as microRNAs, interact with a complex set of cellular machinery to regulate the translation of targeted mRNAs in eukaryotes. Jacob and Monod (1961) proposed that untranslated RNAs might regulate gene expression in the lac operon. This idea was discredited with the discovery of protein transcription factors, and was largely forgotten until the discovery of microRNAs (Lau et al. 2001; Lee and Ambros 2001). These represent an entire class of genes producing small (21 bases) untranslated RNAs (Fire et al. 1998) in worms, flies, vertebrates, and plants (Bartel 2004). The microRNAs are transcribed from larger RNA genes and cleaved to their active form, in which they bind to target mRNAs either precisely (primary mode of action in plants) or with a few base pairs of mismatch, targeting the mRNA either for silencing or degradation. The target sequences in the mRNA can be considered as cis-acting elements, whereas the microRNAs and the enzymes that process them, as well as the proteins that perform the degradation, constitute the trans-acting factors in this mechanism of regulation.

In summary, a lot has been learned about the various mechanisms that regulate the differential expression of genomes and their genes. In some cases, where the mechanisms are well-described and the relevant genes identified, the availability of whole genome sequences allows for a rapid assessment of the conservation and phylogenetic distribution of these genes. Nevertheless, how these mechanisms impact the evolution of RNA abundance and genomic expression remains largely uncharacterized.

Measuring Attributes of Genomic Expression with Experimental and Computational Tools

The conservation of the genetic code across most of the tree of life facilitates the identification of protein-coding genes from raw DNA sequences. In contrast, the precise identification of regulatory regions and mechanisms has remained a difficult and elusive task. Unlike protein-coding sequences, regulatory domains are not readily identified by standard landmarks such as start and stop sites, open reading frames, and the splice sites that delineate introns from exons. Regulatory regions also do not possess characteristic genome-wide particularities such as similar codon biases or parallel rates of divergence among codon sites. In addition, regulatory modules are irregularly localized across the genome and the regulatory "code" appears to be considerably more degenerate than the genetic code.

Despite these inherent difficulties, researchers have made substantial inroads into identifying the sequences that control gene expression in a temporal, spatial, and quantitative manner. The recent integration of empirical analyses of gene expression and the high-throughput computational analysis of transcriptional regulation on a genome-wide scale is beginning to reveal how genomic regulation affects the transition from genotype to phenotype.

Typically, regulatory elements and motifs are found using two approaches, both of which are ideally combined with experimental validation of the sequences identified. First, homologous noncoding regions from different species are compared, and regions with unexpected conservation are targeted as being candidates for regulatory activity. Second, sequences from genes sharing a particular attribute (e.g., coregulation, similar rate of mRNA decay, and so forth) are compared in order to identify sequence features associated with that particular attribute. The most often used attribute is coregulation across a set of environmental treatments. In the next sections we outline major experimental and computational techniques used in studies of genomic regulation.

Experimental approaches

One of the most common experimental approaches used to identify and verify candidate regulatory regions is to assay for altered regulatory activity after mutation, a procedure generally known as "promoter bashing." Typically, putative regulatory regions are cloned, mutated, and then checked for differences in gene expression by reporter assays either in vivo or in vitro (Stanojevic et al. 1991).

A technique that has become central to delineating the precise region of transcription factor binding is DNAse footprinting (Brenowitz et al. 1986). In DNAse footprinting, a binding site sequence is bound to its transcription factor in vitro in order to protect it from partial DNAse cleavage. Partial sequences of the protected fragments are then determined on a sequencing gel and the precise location of binding identified. Electrophoretic mobility shift assays (EMSA) can also be employed to identify transcription factor binding sites. EMSA works on the principle that bound DNA migrates at a different rate than unbound DNA. After two decades of footprinting experiments, there are now species-specific databases, such as the Drosophila DNase I Footprint Database (Bergmann et al. 2004), and more general databases, such as ORegAnno (Montgomery et al. 2006), that serve as important curated repositories for information on transcription factor binding sites.

The principles of transcriptional control have been elucidated by detailed studies on individual genes. However, the global architecture of the regulatory network only began to be understood with the advent of microarray-based methods (van Steensel and Henikoff 2003). These techniques permit a genome-wide mapping of protein–DNA interactions, chromatin packaging, and epigenetic modifications such as DNA methylation and histone modifications. We outline these technologies and their main contributions to our understanding of regulatory networks.

GENE EXPRESSION LEVELS Measuring gene expression level is the most fundamental requirement for studying the evolution of genomic expression. Several techniques are available for this purpose, widely varying in terms of cost, throughput, time investment, and practicality. Most recent techniques make use of array technologies in which the abundance of a given message is assessed by hybridizing a labeled cDNA sample to spotted microarrays.

DNA microarrays can vary extensively in the length of the DNA sequence in each spot (from cDNA clones, to single-exon PCR products, to medium-sized oligonucleotides of about 60–70 nucleotides, to very short oligonucleotides of less than 35 nucleotides), as well as in terms of genomic coverage. Tiling arrays, for instance, can cover both protein-coding as well as non-protein-coding sequences and provide a high-resolution sliding window view of expressed sequences in a particular genomic region or—at lower resolution—over the entire genome (Mockler et al. 2005). Use of tiling arrays has uncovered a large number of noncoding expressed sequences, many of which are transcribed from intergenic sequences. Figure 5.4 summarizes the important steps in the microarray analysis of gene expression levels. A limitation of oligonucleotide arrays is that the intensity of the signal may be overly sensitive to sequence mismatches. This will cause difficulties if there is genetic variation between the samples being contrasted, an obvious issue in evolutionary comparisons.

Two sequencing-based methods are still used to study genomic expression, but are being superseded by array technologies, even for nonmodel organisms. Expressed sequence tags (ESTs) were often used in early studies of genome-wide gene expression. ESTs are sequences obtained by random sequencing of clones from cDNA pools and were mostly produced in parallel with genomic projects in model organisms (Hatey et al. 1998). Because cDNA pools are often normalized by subtractive hybridization, such that low copy transcripts are enriched and highly abundant transcripts are removed, they should generally be avoided as a guide to mRNA transcript abundance. Serial analysis of gene expression (SAGE) is yet another technique for estimating levels of gene expression, which makes use of the fact that a sequence of about 15 nucleotides already contains enough information to identify (tag) the genomic sequence in which it is embedded. In this technique, a large number of DNA molecules, each consisting of a series of short expressed tags linked together in a single chain, are sequenced (Yamamoto et al. 2001). Each tag is then assigned to its gene of origin, and the number of times a tag for a given gene is found is used as a measure of the gene's expression level.

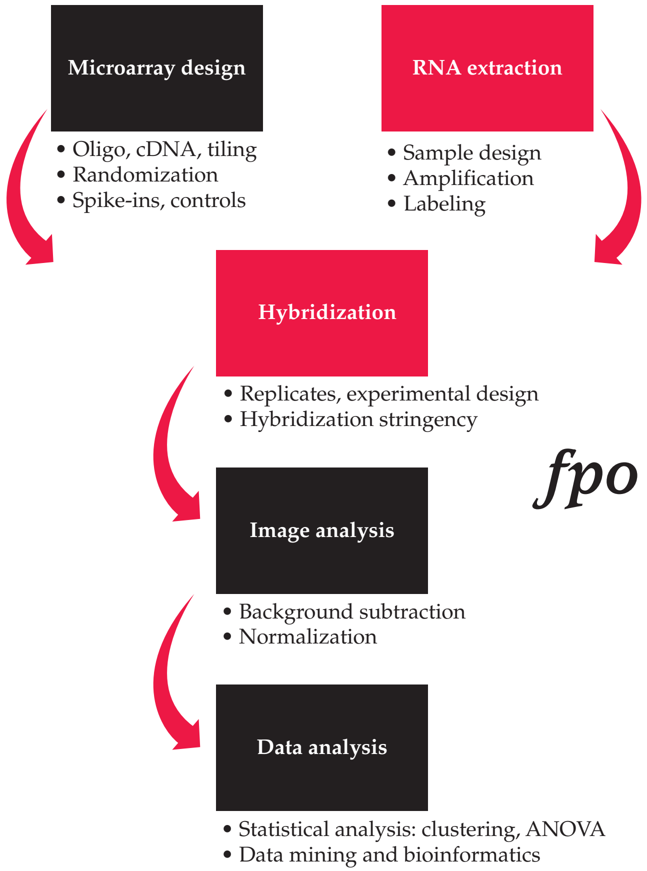

Figure 5.4 Microarray experimental design. The main steps of a typical experiment are highlighted. Specific procedural operations are also listed. Steps performed away from the lab bench are labeled in red. Specific control DNA might be spotted in the array if spike-ins of foreign RNA of known concentration is to be used for data normalization.

PROTEIN–DNA INTERACTIONS The use of DNA microarrays is not limited to the measurement of RNA abundance. They can also be used to explore other aspects of gene expression that have to do with the state of the gene being expressed (Figure 5.5). These techniques have been instrumental in deciphering the genomic architecture of gene expression, and promise much

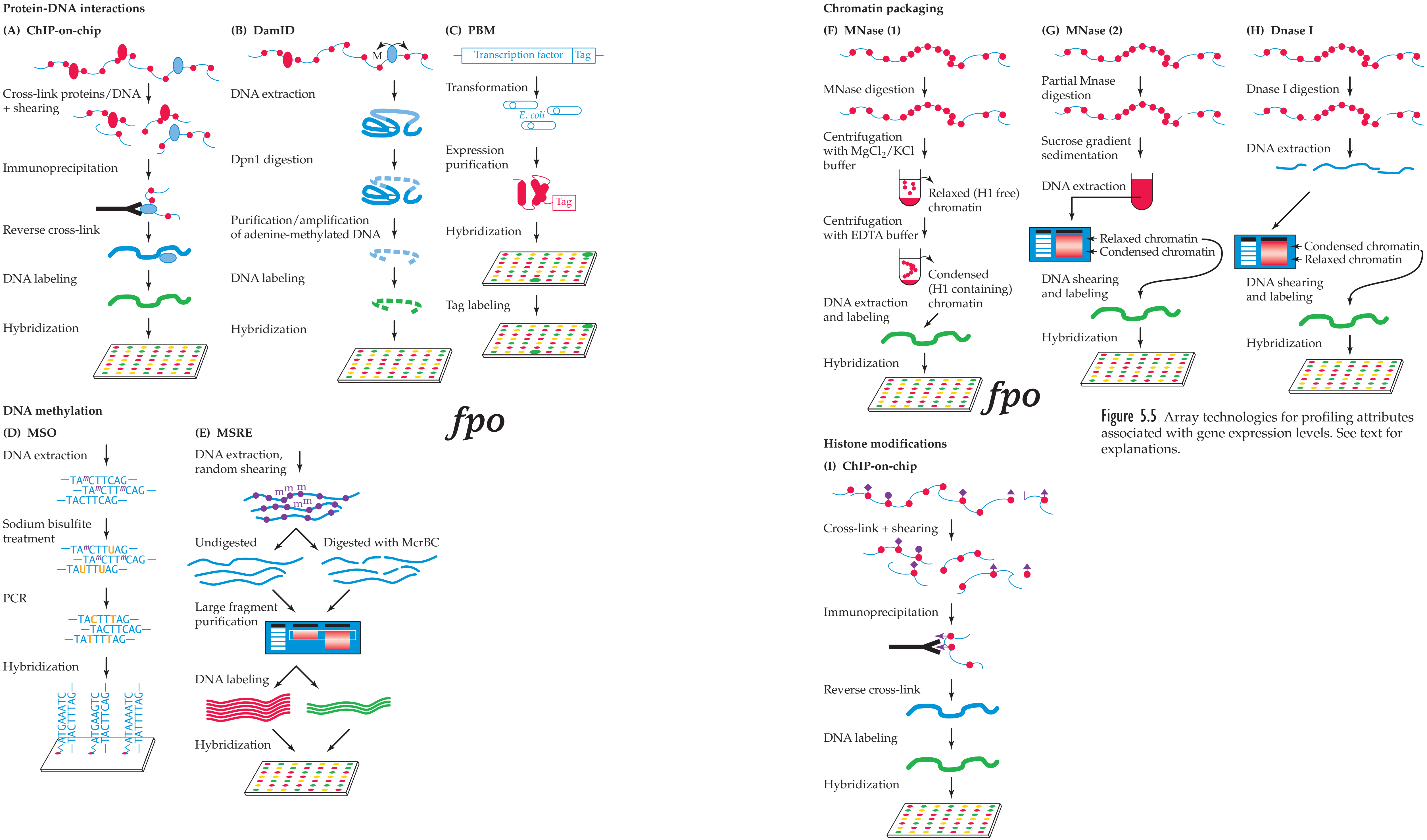

Figure 5.5 Array technologies for profiling attributes associated with gene expression levels. See text for explanations.

for future use in evolutionary biology. ChIP-on-chip is the more commonly used method to analyze protein–DNA interactions. The technique combines chromatin immunoprecipitation (ChIP) and microarray analysis (Weinmann and Farnham 2002). Cells are treated with a chemical agent (typically formaldehyde) that cross-links the protein complexes in situ to DNA. The chromatin is then fragmented and immunoprecipitated using specific antibodies that recognize the protein of interest. To identify the DNA sequence of the binding site, the cross-link is reversed, and the DNA fragments are labeled with fluorescent dye and hybridized to microrrays. Two other, complementary approaches, DamID (van Steensel and Henikoff 2000) and PBM (protein-binding microarray; Mukherjee et al. 2004), employ direct in situ labeling (methylation) of the bound DNA and in vitro identification of transcription factor binding sites, respectively.

With these approaches, basic questions about the regulatory network can be answered. For instance, how many genes are under control of one particular transcriptional factor? In *S. cerevisiae*, systematic studies of 106 transcription factors showed that, on average, about 40 genes are targeted by any given factor, the upper limit being 180 genes (Lee et al. 2002). In higher eukaryotes, it seems that some transcription factors also interact with even larger sets of promoters (Orian et al. 2003). A complementary question is: how many transcriptional factors link to a particular gene? More than one third of the genes in *S. cerevisiae* are bound by two and more factors (Lee et al. 2002), and some genes can have more than 12 transcription factors bound. These figures are expected to be underestimates because they depend on an arbitrarily chosen, conservative statistical threshold. In addition, almost all studies have focused on the upstream regions of genes, but downstream regions and introns also play an important role in the regulation of gene expression (e.g., Martone et al. 2003). These techniques have not yet been used to ask how transcription factor interaction patterns change between related species.

CHROMATIN PACKAGING Microarray technologies have also been used to investigate chromatin states along chromosomal regions. One set of methods takes advantage of the resistance of condensed chromatin to nuclease digestion, followed by separation of different-sized chromatin fragments by sedimentation in a sucrose gradient. The relaxed chromatin in each fraction is extracted after fractionating in an agarose gel (Gilbert et al. 2004). The DNA fragments can then be labeled and hybridized to microarrays. A second approach uses DNase I and separation in agarose gels to isolate condensed or relaxed chromatin (Sabo et al. 2006). Such studies have revealed that relaxed chromatin is tightly correlated with high gene density in the chromosomes (e.g., Gilbert et al. 2004). Perhaps more surprisingly, no correlation was found between gene expression level and the distribution of relaxed chromatin. Genes may be active even in condensed chromatin regions, and inactive in relaxed regions (Gilbert and Bickmore 2006).

DNA METHYLATION Another DNA modification that affects a gene's transcriptional state is DNA methylation. Several techniques have been developed to map the distribution of 5-methylcytosine (^{m5}C) in eukaryotic genomes. In the MSO (methylation-specific oligonucleotide microarray) technique, genomic DNA is treated with sodium bisulfite, which converts unmethylated, but not methylated cytosine into uracil (Adorján et al. 2002). During subsequent PCR, the DNA polymerase reads uracil as thymine and cytosine as guanine. To discriminate methylated and unmethylated cytosine at specific nucleotide positions in the original DNA, a specially designed oligonucleotide microarray is used: it contains a set of oligonucleotides with different combinations of guanine (to detect unmethylated cytosine) or adenine (to detect methylated cytosine) substituted at the cytosine positions (Adorján et al. 2002; Gitan et al. 2002). Another method uses methylation-sensitive restriction enzymes (MSRE). For instance, the enzyme McrBC cuts methylated, but not unmethylated DNA. After shearing, McrBC-treated genomic DNA samples are depleted in the high molecular-weight fraction if the original DNA was methylated. Microarray hybridizations can then identify the methylated fragments. Many variations of this approach have been reported (e.g., Lippman et al. 2005).

Computational approaches

With the recent availability of genome sequences and transcription profiles, along with advances in the field of computational biology, researchers have begun to successfully investigate regulatory regions at structural and functional levels. The power to detect, describe, and model conserved motifs both within and between species has increased substantially. Researchers are no longer restricted to analyzing a small subset of sequences from a particular gene (usually promoter regions), but can now apply motif-finding principles to large regions around many genes (Nardone et al. 2004). The combination of computational and empirical tools has resulted in powerful approaches to the discovery of regulatory regions across genomes. There are now a large number of online resources for the analysis of regulatory sequences, including searchable databases (Table 5.1) and computational biology tools (Table 5.2).

While finding promoters in a genome is assisted somewhat by knowing the positions of genes, promoter discovery from raw nucleotide sequences is, in many ways, a much more difficult task than finding protein-coding sequences. Promoters comprise a large and diverse set of sequences and show no clear defining signature. Also, because many mammalian genes have large noncoding 5′ exons, promoters are located at variable distances from the gene that they regulate. As a result, promoter prediction algorithms must strike a balance between finding real regions of interest and falling victim to false positives. Fortunately, empirical studies of gene regulatory elements have generated a large and diverse template for computational biologists to construct genome-scanning algorithms (e.g., Fickett and Hatzigeorgiou 1997).

Table 5.1 Online regulatory resources: searchable databases

Database	Description	Web address
Promoters[a]		
DoOP	Database of orthologous clusters of promoters	http://doop.abc.hu/
EPD	Eukaryote POLII promoter database	http://www.epd.isb-sib.ch/
Worm DB	*C. elegans* promoter database	http://rulai.cshl.edu/cgi-bin/CEPDB/home.cgi
Mammal DB	Mammalian promoter databases	http://rulai.cshl.edu/CSHLmpd2/
SCPD	Promoter database for *S. cerevesiae*	http://rulai.cshl.org/SCPD/index.html
PlantProm DB	Plant promoter database	http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom
Motifs[b]		
Transterm	Translational signal database (mRNA motifs)	http://guinevere.otago.ac.nz/transterm.html
PLACE	Plant *cis*-acting regulatory DNA elements database	http://www.dna.affrc.go.jp/htdocs/PLACE/
Transcription Factors[c]		
TRANSFAC	TF database	http://www.gene-regulation.com/pub/databases.html#transfac
ooTFD	Object-oriented TF database	http://www.ifti.org/ootfd/
ProteinLounge TFdb	TF database	http://www.proteinlounge.com/trans_home.asp
MIRAGE	Resource for the analysis of gene expression	http://www.ifti.org/
PRODORIC	Prokaryote database of gene regulation	http://prodoric.tu-bs.de/
TFdb	RIKEN mouse TF database	http://genome.gsc.riken.jp/TFdb/
AGRIS	Arabidopsis gene regulatory information server	http://arabidopsis.med.ohio-state.edu/

Table 5.1 *Continued*

Database	Description	Web address
Transcription Factors[c] *(continued)*		
RARTF	RIKEN Arabidopsis TF database	http://rarge.gsc.riken.go.jp/rartf/
RiceTFDB	Rice TF database	http://ricetfdb.bio.uni-potsdam.de/
DBTBS	Transcriptional regulation database in *B. subtilis*	http://dbtbs.hgc.jp/
RegulonDB	*E. coli* K-12 database for transcriptional regulation	http://www.cifn.unam.mx/Computational_Genomics/regulondb/
Ecoli TFDB	*E. coli* TF database	http://bayesweb.wadsworth.org/binding_sites/
Transcription Factor Binding Sites[d]		
FlyReg	Drosophila DNase I footprint database	http://www.flyreg.org/
ORegAnno	The open regulatory annotation	http://www.bcgsc.ca:8080/oregano/Index.jsp
JASPAR	TF binding profile database	http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl
MAPPER	Multi-genome analysis of positions and patterns	http://bio.chip.org/mapper
DNASTAR	TF binding site database	http://www.dnastar.com/web/r50.php
TFSEARCH	Search TF binding sites	http://www.cbrc.jp/research/db/TFSEARCH.html
TESS	Predicting transcription binding sites	http://www.cbil.upenn.edu/tess/

[a] These publicly available databases contain a curated list of promoter regions from various species. Data can be downloaded in bulk or individually visualized by searchable IDs such as GenBank accession numbers or gene names.

[b] Motifs of various kinds are available from these interactive databases. Motifs can be searched against specific sequences and genomes.

[c] A useful set of curated databases of known transcription factors (TFs) and their protein domains. TFs may be downloaded in bulk. Some sites offer in-house Blast portals.

[d] These databases, some species-specific, are generated from literature reports of footprinting experiments. Transcription factors, their binding sites in the genome, and sometimes also their binding affinities, are reported. Depending on the website, phylogenetic conservation of the binding sites may also be reported.

TRANSCRIPTION FACTORS BINDING SITES AND SEQUENCE MOTIF DISCOVERY The experimental elucidation of binding sites for individual transcription factors has provided computational biologists with an indispensable tool. From a set of known aligned binding sites, positional matrices (sometimes called profiles or position-specific scoring matrices) of base pair frequencies can be generated. Different algorithms use various implementations of this signal-based data, including position weight matrix (PROMOTER2.0) and neural net (ProScan). Motifs can be short and contiguous or bipartite and long. The latter includes, for instance, palindromic sequences separated by a spacer element that is usually variable in length. Regulatory sequences can be readily visualized using a sequence logo format, which transforms matrix data into visual information.

In contrast to these signal-based methods, another way to find sequence motifs is based purely on the sequence content. Characteristic patterns of conserved sequences may be found among coregulated genes or among orthologous sequences of different species, for example, overrepresentation (relative to random noncoding sequence) of putative sequence motifs. The human genome contains approximately 1850 distinct transcription factors and the number of potential combinations of any number of these acting upon a particular gene is enormous.

COMPARATIVE GENOMICS AND PHYLOGENETIC FOOTPRINTING A powerful and popular approach to decoding regulatory sequences is comparative genomics. By finding sequences that are conserved across species, one can

Table 5.2 Online regulatory resources: search tools

Data base	Description	Web address
Promoter Prediction[a]		
McPromoter	The Markov Chain Promoter Prediction Server	http://genes.mit.edu/McPromoter.html
NNPP	Promoter Prediction by Neural Network	http://www.fruitfly.org/seq_tools/promoter.html
TRES	Comparative Promoter Analysis	http://bioportal.bic.nus.edu.sg/tres/
PromoterWise	Compares 2 DNA sequences, ideal for promoters	http://www.ebi.ac.uk/Wise2/promoterwise.html
PromoSer	Batch retrieval of proximal promoters	http://biowulf.bu.edu/zlab/PromoSer/
Motif Searching[b]		
MOTIF Search	Search motifs	http://motif.genome.jp/
EZRetrieve	Sequence retrieval tool	http://siriusb.umdnj.edu:18080/EZRetrieve/
Possum	Detect *cis*-elements in DNA sequences	http://zlab.bu.edu/~mfrith/possum/
CorePromoter	Core-Promoter Prediction Program	http://sciclio.cshl.org/genefinder/CPROMOTER/
Gene Express	Analysis of genomic regulatory sequences	http://wwwmgs.bionet.nsc.ru/systems/GeneExpress/
Phylogenetic Footprinting[c]		
Phylofoot	Portal to phylogenetic footprinting	http://www.phylofoot.org/
UCSC	UCSC Genome Browser	http://genome.ucsc.edu/cgi-bin/hgGateway
TraFaC	Finds conserved *cis*-elements across species	http://trafac.cchmc.org/trafac/index.jsp
PipMaker	Aligns similar regions of sequence	http://pipmaker.bx.psu.edu/pipmaker/
VISTA	Suite of programs that aligns genomic sequences	http://genome.lbl.gov/vista/index.shtml
LAGAN	Comparative genomic alignment programs	http://lagan.stanford.edu/lagan_web/index.shtml
FootPrinter	Phylogenetic footprinting of orthologous sequences	http://bio.cs.washington.edu/software.html
Bayesaligner	Phylogenetic footprint using a Bayesian approach	http://bayesweb.wadsworth.org/cgi-bin/bayes_align12.pl
TF and TF Binding Sites[d]		
ConSite	FTF binding sites via aligned genomic sequence	http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite
TFSEARCH	Transcription factor search	http://www.cbrc.jp/research/db/TFSEARCH.html
MSCAN	Find functional clusters of TF binding sites	http://mscan.cgb.ki.se/cgi-bin/MSCAN
Weeder Web	TF binding sites in sequences via co-regulated genes	http://159.149.109.16
SITECON	Conserved physicochemistry in TFBS alignments	http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/
POBO	TF binding site verification with bootstrapping	http://ekhidna.biocenter.helsinki.fi:9801/pobo/
DTFAM	Explores TF associations through text-mining	http://research.i2r.a-star.edu.sg/DRAGON/TFAM/
Fly Enhancer	Finds clusters of binding sites in Drosophila	http://flyenhancer.org/Main
AliBaba	Prediction of transcription factor binding sites	http://www.alibaba2.com/

[a] A set of promoter prediction software publicly available online. Your sequence of interest can be uploaded onto each webserver in order to identify promoter regions using a wide variety of approaches.

[b] Motifs in uploaded sequences can be detected using a multitude of methods from these websites. These tools represent just a handful of available online resources to identify motifs.

[c] These online alignment sites allow one to find conserved sequences in regulatory regions. Some of these websites already contain precomputed alignments of regulatory regions from sequenced genomes.

[d] These sites contain tools that allow you to search your sequence of interest for transcription factors and their binding sites.

quickly infer whether they are functionally important without using costly molecular or biochemical procedures. Phylogenetic footprinting aims to find functionally important regulatory regions by identifying conserved orthologous sequences (Gumucio et al. 1993; Hardison et al. 1997). This approach has been very successful with the advent of complete genome sequences from model genetic organisms. Deep-rooted phylogenetic taxa can be used to find invariant regions indicative of constrained function, and closely related sister taxa can be used to find regions of sequence conservation, as well as genus-specific regulatory units, among species with a more shallow ancestry (i.e., phylogenetic shadowing; Boffelli et al. 2003).

Since the highlighting of conserved regions has become such an important tool, the major genomic databanks have started to provide genome browsers that allow one to easily visualize conserved regions. NCBI's Map Viewer (National Center for Biotechnology Information), EMBL-EBI (European Molecular Bioinformatics Laboratory), and the UCSC Genome Browser (University of California at Santa Cruz) possess excellent graphical interfaces for users to search for conserved orthologous regions. In addition, VISTA and PiPMaker are popular phylogenetic footprinting tools.

The Evolution of Genomic Expression: What Do We Know?

Evolutionary genomics is beginning to address the molecular evolution of regulatory sequences and the phenotypic evolution of mRNA abundance. Ultimately, this approach may provide a comprehensive understanding of the magnitudes and patterns of evolutionary variation in regulatory sequences and relevant phenotypes. Transcription is the first step in the mapping of genetic variation to higher-level phenotypes and, therefore, diversity in gene expression levels is one of the most direct phenotypic outcomes of regulatory variation. This diversity is now being widely documented and interesting patterns are being discovered. Understanding how genomic regulation translates into variation in gene expression levels is thus the first step towards understanding variation in more complex organismic features. Because of this, gene expression levels are likely to become a model phenotype for testing current methods and assumptions in our understanding of the dynamics of polymorphism and divergence in natural populations, including ecologically relevant and disease-related variation. However, virtually nothing is known about evolutionary variation in attributes that affect variation in gene expression (e.g., rates of mRNA degradation, levels of methylation, nucleosome positioning, and so forth). In this section we review our current understanding of the patterns and processes affecting the evolution of genomic expression.

Genomic expression and morphological evolution

We know a lot more about variation in the coding sequence of genes than about variation in regulatory regions or in gene expression patterns. Yet variation in gene expression is likely to account for a large fraction of the phenotypic diversity observed within and between species. Abundant examples of regulatory variation contributing to phenotypic diversity at the morphological, behavioral, and physiological levels illustrate the role of regulatory evolution in phenotypic and adaptive diversification.

An early and now classical demonstration of the relevance of regulatory variation in evolution was provided by Cherry, Case, and Wilson in 1978. These authors took metrics of shape typically used by systematists interesting in distinguishing between species of frogs to measure morphological differences between humans and chimpanzees. The striking result was that, morphologically, human and chimps are much more different from each other than are species of frogs belonging to different suborders. This result stands out because, while suborders of frogs show considerable differences in protein-coding DNA sequences, humans and chimps are remarkably similar to each other at the DNA level. This suggests a substantial role for regulatory evolution in the human–chimp divergence.

In another example, the pattern of hairs on the first instar larva varies among closely related species of the *Drosophila melanogaster* group and the evolution of *cis*-regulatory elements of the *ovo/shaven-baby* gene is responsible for hair patterning in *Drosophila sechellia,* distinguishing it from its closest relative (Sucena et al. 2003). Similarly, the cuticular hydrocarbon pheromones involved in mating preference in *D. melanogaster* display geographic variation in the 5,9-heptacosadiene/7,11-heptacosadiene ratio. This polymorphism is caused by a deletion in the promoter region of a desaturase gene, changing its expression pattern and causing knock-on effects on pheromone production (Takahashi et al. 2001). Closely related species of cichlid fishes have different visual spectral sensitivities, which probably has important consequences for the foraging behavior of these fishes and their mate choice (based on male coloration). Differences in visual sensitivity are often achieved by shifts in chromophore usage or in opsin coding sequences, but in this case they are caused by changes in opsin gene expression (Carleton and Kocher 2001). Finally, the colors and patterns of eyespots on butterfly wings are yet another system in which a connection between variation in the expression of specific genes and higher-level evolutionary changes has been established (Brunetti et al. 2001; Beldade and Brakefield 2002).

Molecular evolution of regulatory sequences

The study of protein-coding sequences has been boosted by the explosion of comparative data coupled to new statistical methods for analyzing and interpreting coding sequences. Current methods have incorporated a number of factors regarding rates of coding sequence variation (e.g., transition/transversion ratios, position heterogeneity), and have allowed several genome-wide analyses of the selective forces acting on protein sequences (e.g., Nielsen et al. 2005). Most of these analyses were based on models and statistical tests on the ratio of nonsynonymous (*dN*) to synonymous (*dS*) nucleotide substitutions. It is noteworthy that genome-wide analyses have also challenged the very assumptions that underlie using *dS* as a reliable proxy for the neutral mutation rate (Wyckoff et al. 2005), which suggests that classical interpretations may need to be reevaluated.

Historically, the molecular evolutionary analysis of regulatory sequences has largely remained outside the mainstream of such analyses of coding sequences. This is not because the relevance of regulatory evolution has been underappreciated, but is rather due to the major challenges inherent to the evolutionary analysis of noncoding DNA. First, despite the large number of mechanisms of gene regulation already described, qualitatively novel mechanisms are still being discovered. These elusive mechanisms relate to mRNA genes, microRNAs, S/MARs, and nuclear organization, to cite a few. Similarly, several ultra-conserved noncoding DNA sequences have been identified (e.g., Bejerano et al. 2004) whose functional role is unknown. Hence, developments in the last ten years have uncovered a variety of regulatory mechanisms whose evolution has yet to be investigated, let alone modeled and fully incorporated into mainstream studies of molecular evolution. It is clear, nevertheless, that a huge variety of elements and regulatory phenomena influence genomic expression, and a truly comprehensive theory for the evolution of regulatory sequences should include all these mechanisms.

Second, little is known about the evolution even of noncoding DNA with apparently obvious relevance to genome expression. For example, there is now a vast amount of data regarding promoter functioning and its regulation by other elements such as enhancers. This is particularly well-illustrated in the case of gene regulation in the human and mouse immune systems. In spite of this wealth of information, a comprehensive understanding of gene regulation from the perspective of promoter activity is still missing. As a consequence, although promoters have long been recognized as the site for transcriptional regulation, few evolutionary analyses have been carried out.

In spite of these challenges, a few recent studies have attempted to define structural features of regulatory sequences (Dermitzakis et al. 2003; Chin et al. 2005), as well as to develop metrics for measuring regulatory divergence of these sequences (Castillo-Davis et al. 2004; Chin et al. 2005). These studies are also complemented by analysis of patterns of substitution in noncoding DNA within and between species (Andolfatto 2005). All in all, these studies provide new venues to explore the evolution of regulatory sequences, and suggest promising directions for future research.

Dermitzakis and coworkers (2003), for instance, found that conserved noncoding sequences are often under stronger selective constraint than proteins and noncoding RNAs. Furthermore, the patterns of evolutionary variation in conserved noncoding sequences are distinguishable from those observed for protein-coding sequences. Substitutions in noncoding sequences were more clustered along the sequence compared to those in protein-coding sequences (Dermitzakis et al. 2003), presumably because purifying selection pressure is unevenly distributed along regulatory sequences. Moreover, noncoding sequences showed more symmetric rates of divergence (i.e., A → T and T → A, or C → G and G → C) than coding sequences. Recent work by Moses and colleagues (2003) attempted to characterize the pattern of evolution within transcription factor binding sites in yeast, which were shown to evolve more slowly than background sequences. In addition, they found substantial position-specific variation in rates of sequence evolution within transcription factor binding sites. While some positions within transcription binding motifs were highly conserved, rates of evolution in less important sites could not be distinguished from background rates. Furthermore, Moses and colleagues (2003) found a strong correlation between positional rate variation in a single genome and that observed between genomes. Work by Chin and coauthors (2005) is also illustrative. These authors used a hidden Markov model (HMM) to break down promoters into selectively neutral regions and evolutionarily constrained regions under purifying selection. The latter contained an overabundance of regulatory motifs. Chin and coauthors (2005) estimated that about 30 percent of the promoter sites in yeast are evolving under purifying selection, whereas the remaining 70 percent are accumulating mutations at the neutral rate. Another interesting study by Keightley and coworkers (2005) compared the extent to which sequence conservation differs between two pairs of species. They found that the conservation of noncoding sequences upstream (5′) of the coding region was substantially greater in mouse–rat comparisons than in human–chimpanzee comparisons. Based on this observation, the authors argued that purifying selection on regulatory variation has been less efficient in primates than in rodents.

Stabilizing selection, positive selection, and neutrality of gene expression levels

It has long been realized that stabilizing selection (i.e., purifying selection) is a pervasive force in the evolution of higher-order morphological phenotypes, as well as protein-coding sequences. Gene expression levels are no exception to this pattern, and increasing evidence suggests a fundamental role for stabilizing selection in restricting evolutionary variation in transcript levels (Denver et al. 2005; Jordan et al. 2005; Lemos et al. 2005b; Rifkin et al. 2005).

The high conservation of gene expression levels across species is particularly striking in view of the ample supply of mutations expected to influence gene expression across evolutionary timescales, as measured by the experimental accumulation of mutations and their effects on gene expression levels (Figure 5.6). In particular, Denver and colleagues (2005) and Rifkin and colleagues (2005) used mutation accumulation lines of worms and flies, respectively, to experimentally estimate the neutral mutation rate for gene expression levels at about 10^{-5}. Although this rate is about two orders of magnitude below the typical value found for a number of morphological and enzyme activity traits (Lynch 1988), mutation accumulation lines still have more dispersion in mRNA abundances than is observed across genotypes segregating in natural populations (Denver et al. 2005). Accordingly, it has been suggested that gene expression divergence between yeast gene dupli-

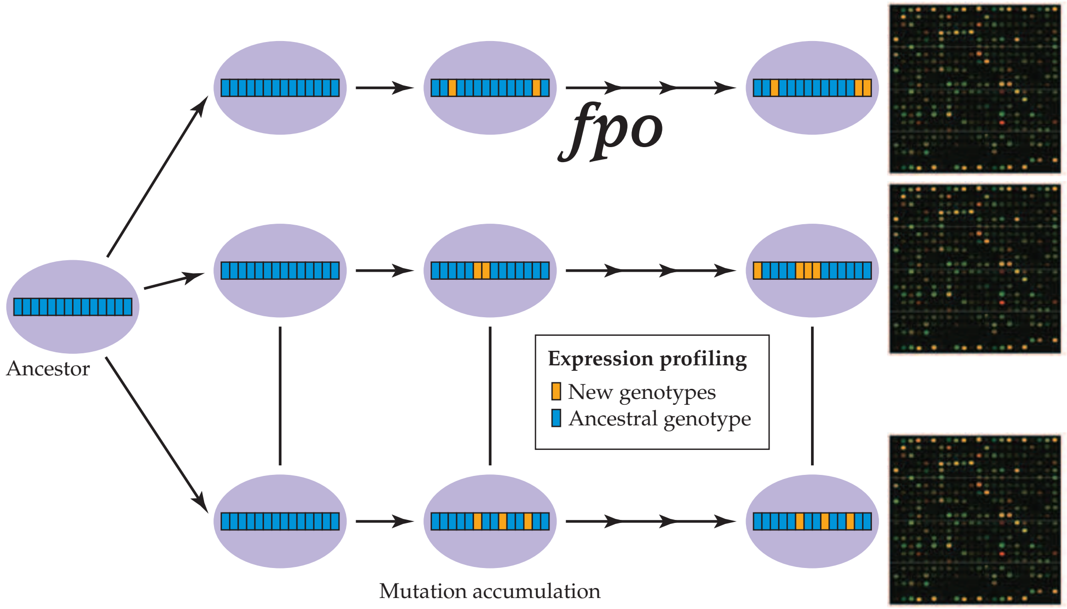

Figure 5.6 Mutation accumulation design for studying the neutral rate of regulatory divergence.

cates (Oakley et al. 2005) or between orthologous genes in different species (Lemos et al. 2005b) does not follow a phylogenetic model, in that closely related genes are not more likely than distantly related genes to show similar expression levels. Instead, the comparisons suggest that gene expression evolution proceeds so rapidly that the magnitude of divergence quickly saturates—due solely to the high mutation rates associated with mRNA abundances—even in the absence of diversifying or positive natural selection. Diversifying and positive selection would only further increase the rate of gene expression evolution, thereby leading to an even more rapid saturation of the evolutionary signal. Indeed, Ferea and colleagues (1999) and Toma and colleagues (2002) showed that large differences in gene expression can be accomplished in only a few generations of artificial selection.

These conclusions are in sharp contrast with a suggestion that gene expression variation may be unconstrained (Khaitovich et al. 2004). This suggestion was motivated by the finding that the divergence of translated mRNAs did not seem to be lower than that of transcribed pseudogenes, whose mRNA is not capable of producing a functional protein. However, results from comparisons between translated mRNAs and untranslated mRNAs derived from pseudogenes should generally be interpreted with caution. This is because untranslated mRNAs that happen to be expressed across timescales as long as that observed between species are unlikely to be nonfunctional.

An important question whose answer is not yet completely apparent regards the major forces that *produce* evolutionary differences in gene expression levels. Although the prevalence of stabilizing selection appears to be beyond doubt, it remains to be established whether gene expression differences between species or populations arise mainly through fixation by positive selection of distinct expression alleles in different environmental contexts or, alternatively, through fixation by random genetic drift of selectively equivalent expression states. The neutral theory of molecular evolution developed by Kimura (Kimura 1983) and others promoted the view that most segregating variation within species, as well as most fixed differences in protein sequences between species, arise from selectively equivalent alleles whose small differences have negligible effects on organismal fitness. Consequently, most differences observed between species and populations would result from random fixation of equivalent or nearly equivalent alleles. It should be stressed that the prevalence of stabilizing selection on transcription levels does not imply that a gene's mRNA abundance is at its optimum or that it has a negligibly small environmental or mutational variance.

Furthermore, we must note a few shortcomings of many analyses of gene expression divergence and polymorphism. First, most statistical methods for identifying gene expression differences test each gene independently of all others, one gene at a time. However, genes are often coordinately regulated as gene expression modules and these modules, rather than the genes themselves, may often be the relevant targets for evolutionary analysis.

In fact, evolutionary biologists have long debated the proper multivariate description of biological complexity, a problem by no means restricted to gene expression level (Lande and Arnold 1983). It is therefore important to consider whether any gene-by-gene metric of gene expression variability and divergence is an adequate descriptor of the biological complexity. This is because gene expression profiles might be more meaningfully examined as a single complex character (e.g., a network) instead of a simple collection of single-gene similarities and differences. The quantification of biological divergence and polymorphism on a truly systemic scale remains elusive, and the development of biologically meaningful multivariate descriptors of gene expression states remains a challenge. Such descriptors would integrate across entire pathways, functional groups, and interrelated modules.

The normalization of mRNA abundance data across samples is another difficult area in the analysis of gene expression variation across evolutionary timescales. This is because normalization methods commonly assume that total mRNA abundance is constant across samples or, in other words, that only a small number of genes differ in expression levels between samples (Quackenbush 2002). This is a fundamental problem because evolutionary comparisons often involve a large number of differences in gene expression levels; therefore, the assumption of similar mean expression level across samples may not always hold. The potentially confounding effect of sequence divergence on estimates of mRNA abundance is another fundamental issue when analyzing and interpreting polymorphism and divergence data. Along these lines, Gilad and colleagues (2005) examined how normalization procedures interact with sequence divergence to produce biased estimates of gene expression levels. They found significant effects even between samples that are as similar as humans and chimpanzees.

What constrains or promotes evolutionary variation in gene expression levels?

Understanding the constraints imposed upon evolutionary variation in gene expression levels is a major research goal of evolutionary genomics. This goal includes not only identification of the sources of constraints on expression levels but also, equally challenging, reconstruction of the selective landscape underlying evolutionary variation in gene expression levels. Although direct causation is elusive and hard to establish, several factors have so far been shown to be associated with evolutionary variation in gene expression levels. Furthermore, we note that, whatever the determinants of gene expression polymorphism and divergence might be, there seem to be remarkable commonalities in the patterns and relative magnitudes of evolutionary variation in gene expression and protein-coding sequences—so much so that a positive correlation between these two modes of evolution can be detected (Nuzhdin et al. 2004; Khaitovich et al. 2005; Lemos et al. 2005b; Liao and Zhang 2006).

Parisi and coworkers (2003) and Ranz and coworkers (2003) found extensive differences in whole-organism transcriptional profiles of male and female adult fruit flies. These authors showed that about half the genome is differentially expressed between the sexes and argued for the relevance of sex-dependent evolution of gene regulation. Moreover, by classifying genes as male- or female-biased (i.e., gene expression higher in males or females, respectively), Ranz and colleagues (2003) and Meiklejohn and colleagues (2003) showed that male-biased genes have higher levels of gene expression polymorphism and divergence than female-biased and unbiased genes.

The functional class or biological process of a gene product is another attribute relevant to evolutionary variation in gene expression levels. Expression variation in genes from functional classes closely related to transcriptional regulation (e.g., transcription factors) might be expected to influence gene expression more strongly than variation in genes from functional classes more distantly related to transcriptional regulation (e.g., metabolic enzymes). This prediction was verified in a number of studies (e.g., Rifkin et al. 2003; Lemos et al. 2005b), suggesting that the expression levels of transcription factors are indeed more tightly controlled than the expression levels of metabolic enzymes. Classically, two sets of genes have been shown to evolve rapidly at the level of the protein-coding sequence: immune system–related genes and male reproduction–related genes. As discussed above, genes that tend to be more expressed in males also tend to show higher levels of polymorphism and divergence in expression. Genes of the immune system also appear to show the same trend towards faster regulatory evolution. Genes of the major histocompatibility complex (MHC) have been known for a long time to be under balancing selection (reviewed in Bernatchez and Landry 2003), which acts to maintain high levels of polymorphism in the coding sequences of these genes. A recent study (Loisel et al. 2006) on the evolutionary history of *cis*-regulatory regions of a MHC gene (DQA1) in primates shows that balancing selection also acts on transcription factor binding sites to maintain functional nucleotide variation with consequences on gene regulation.

Physical attributes such as the number of protein–protein interactions and mRNA abundance may also be relevant for determining the magnitude of evolutionary variation in gene expression levels. For instance, proteins that interact might impose mutual stoichiometric constraints on the amount of variation permitted in their concentrations. This is because a change in the concentration of one protein might result in a stoichiometric cost in its interacting partners. Following this, it is expected that the concentration of proteins whose function depends on direct interaction with a large number of partners should be more evolutionarily constrained than proteins with fewer interacting partners. This prediction has been confirmed using evolutionary variation in gene expression levels as a proxy for evolutionary variation in protein concentration (Lemos et al. 2004). Absolute mRNA abundances may be another physical factor highly relevant to evolutionary variation in gene expression levels. It has been suggested that highly expressed genes might show higher levels of expression polymorphism and divergence (Lemos et al. 2005a). However, it remains unclear to what extent this observation depends on the particular metric used and the accuracy of gene expression assays across a wide range of absolute values (Lemos et al. 2005b). Finally, it can be predicted that genes not essential for organismal survival will be less constrained to vary in expression level. This prediction is supported by the observation that nonessential genes show greater genetic variation for gene expression than essential genes among natural isolates of wine yeasts (Landry et al. 2006). Finally, we note that the presence of a TATA-box motif in the promoter of genes has been positively associated with a gene's level of gene expression polymorphism and divergence, whereas TATA-less genes show decreased levels of evolutionary variation (Tirosh et al. 2006). Figure 5.7 shows some attributes associated with evolutionary variation in gene expression levels, as well as the sign of the effect.

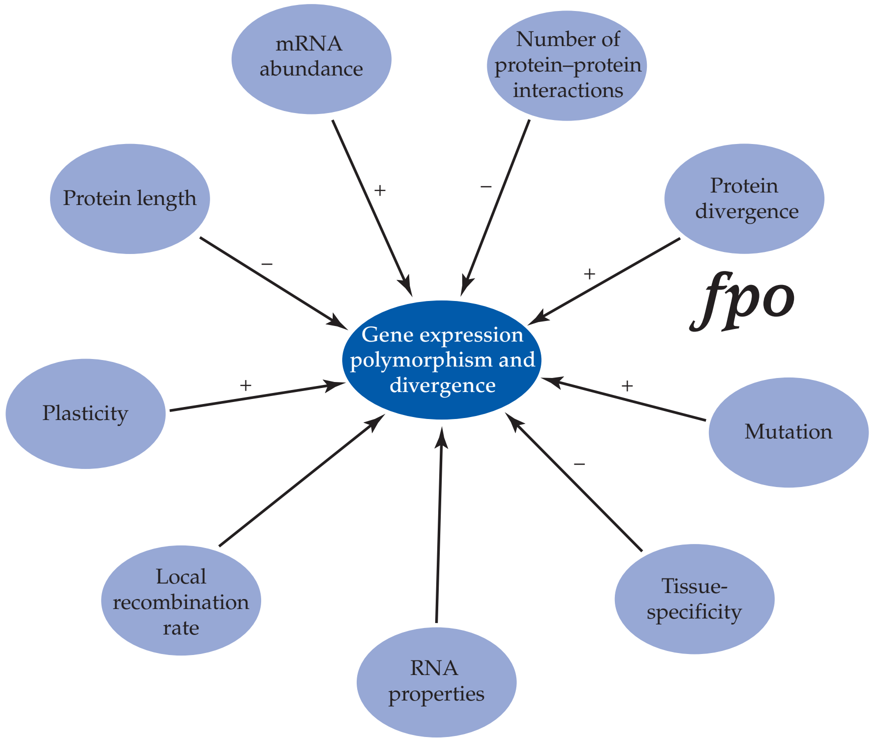

Figure 5.7 Genomic attributes associated with evolutionary variation in gene expression levels. Negative and positive associations are noted.

Inheritance of gene expression levels: Regulatory variation in cis *and* trans

An understanding of the evolution of gene regulation and particularly of the evolutionary forces that control and direct it—mutation, genetic drift, and natural selection—requires the study of genetic variation for gene

Table 5.3 Some genomic studies reporting levels of polymorphism in gene expression[a]

Organism	Number of genotypes and / or individuals	Number of genes assayed	Number of genes showing significant variation	Reference	Tissue	Platform	Statistics
Insects							
Drosophila melanogaster	8 inbred strains	~5,000	218–928 between pairs of strains	Meiklejohn et al. 2003	Whole flies, males	cDNA	Bagel
Drosophila simulans	10 heterozygous strains	~8,000	1,136 ($P < 0.05$); 218 ($P < 0.001$)		Whole flies, males	Affymetrix	ANOVA
Fungi							
Saccharomyces cerevisiae	9 strains	~6,000, whole genome	241 ($P < 0.01$)	Fay et al. 2004	Cells grown in rich medium	Oligo array	ANOVA
S. cerevisiae	4 strains	~6,000, whole genome		Townsend et al. 2003	Cells grown in rich medium	cDNA	Bagel
Fish							
Fundulus heteroclitus/F. grandis	10/5 individuals within populations	~1,000	~161 ($P < 0.01$)	Oleksiak et al. 2002	Heart	cDNA	ANOVA
Atlantic salmon	12 individuals			Aubin-Horth et al. 2005			
Mammals							
Human	35 CEPH cell lines	~5,000		Cheung et al. 2003	Lymphoblastoid cells	cDNA	
Plants							
Arabidopsis thaliana	5 accessions	~8,000	1525 ($P < 0.01$)	Chen et al. 2005	Leaf tissue	Affymetrix	ANOVA
Worm							
C. elegans	5 natural isolates	~5,500	118 ($P < 0.01$)	Denver et al. 2005	Whole worms	cDNA	ANOVA

[a] mRNA abundances are measured across genotypes raised on a controlled environment.

expression in natural populations. Large-scale gene expression profiling of organisms as diverse as yeast, fruit flies, and humans provides an unequivocal result: heritable genetic variation for gene expression level is abundant in nature (Table 5.3).

Since gene expression levels represent multifactorial quantitative traits, they can be studied using the tools and theoretical models developed by quantitative genetics. A formal quantitative genetic study of mRNA abundance thus requires a good understanding of its genetic architecture. The genetic architecture of a trait refers to its characterization in terms of the direct effects of genes and environment, as well as the genetic and environmental interactions affecting the trait expression. The first aspect of the genetic architecture of transcriptional variation is the contribution of *cis*- and *trans*-acting genetic variation. Several factors have contributed to the recent surge of interest in the distinction between *cis* and *trans* effects (e.g., Pastinen and Hudson 2004; Wittkopp et al. 2004; Landry et al. 2005). For instance, genetic variation in *cis*-acting elements is thought to be less likely to have pleiotropic effects than genetic variation in *trans*-acting molecules. In contrast, a single transcription factor may regulate the expression level of dozens of genes such that a mutation in it may affect many of its targets. Furthermore, *cis*-regulatory variation is also of particular interest because its effect on gene expression is potentially easier to identify as individual regulatory mutations than *trans*-regulatory variation. This is because individual mutations in *cis* are clear QTL candidates, whereas *trans* effects often represent aggregate of effects across multiple sites dispersed through the genome and thus much harder to identify. (Yvert et al. 2003).

The first question one might ask about the architecture of gene expression variation is how much variation is found in *cis* and in *trans*? This has been assessed in vivo using traditional genetics and a variety of novel molecular biology tools. The first approach combines gene expression profiling with genetic markers in F2 progenies to map linkages (expression QTL or eQTL) of the transcription phenotypes. *Cis*-acting eQTLs are identified by binning the genome in small regions (physical or genetic distance) and locating each eQTL relative to the gene being regulated. If they both fall in the same bin, the eQTL is said to be *cis*-acting. Depending on the density of markers used and the number of meiosis events analyzed, the size of the bin may vary and thus limit the resolution of this approach. In the case of the study conducted by Morley and colleagues (2004), these bins were five megabases long. The effects of identified *cis*-regulatory variants can be confirmed by in vitro approaches such as transient transfection assays (e.g., Rockman and Wray 2002). This method has attracted much attention

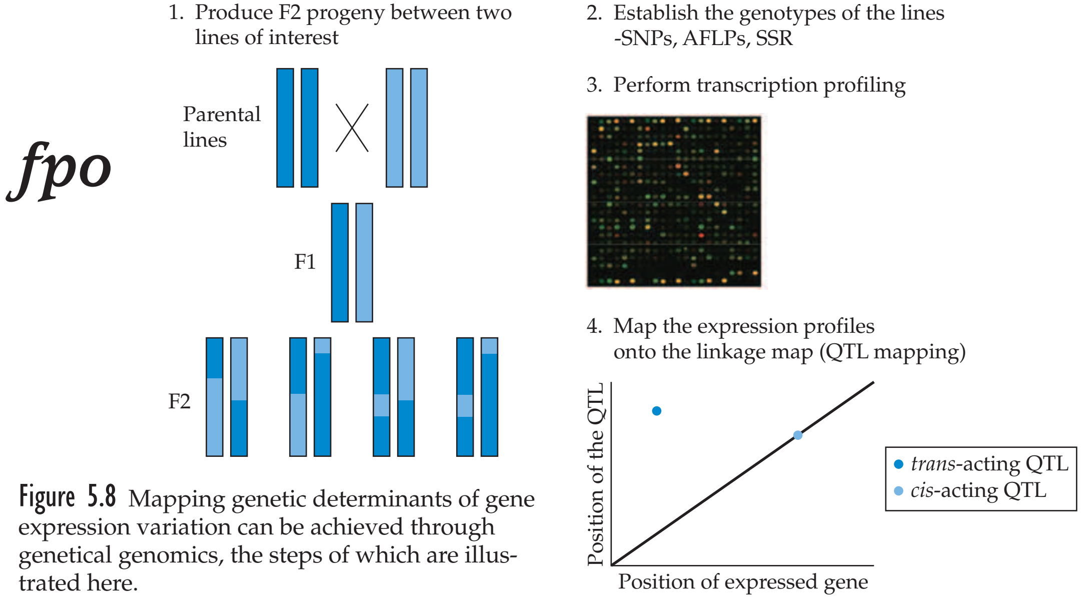
fpo
1. Produce F2 progeny between two lines of interest
Parental lines
F1
F2
2. Establish the genotypes of the lines -SNPs, AFLPs, SSR
3. Perform transcription profiling
4. Map the expression profiles onto the linkage map (QTL mapping)
Position of the QTL
Position of expressed gene
● *trans*-acting QTL
● *cis*-acting QTL

Figure 5.8 Mapping genetic determinants of gene expression variation can be achieved through genetical genomics, the steps of which are illustrated here.

recently and has given rise to a field named "genetical genomics" (de Koning and Haley 2005) (Figure 5.8).

Another approach relies on the fact that *cis*-acting variation is a property of an allele of a gene. In an individual heterozygous for an exonic SNP (single nucleotide polymorphism), *cis*-regulatory divergence can be estimated by measuring the relative concentration of mRNAs containing the two alternative nucleotides, usually using genomic DNA as a control. Since the two alleles and their *cis*-regulatory elements share the same pool of *trans*-acting factors, unequal abundance of transcripts of the two alleles would suggest the presence of genetic variation acting in *cis*. In cases where crosses are performed between two inbred lines or closely related species, the divergence in gene expression level between parental lines can be compared to the difference between alleles in the F1 generation. Any difference between the parental lines that is not assigned to *cis* divergence is then assigned to divergence in *trans*. This approach was used in a study of hybrids between *Drosophila melanogaster* and *D. simulans*, which showed that 28 of 29 genes studied had divergent *cis*-regulatory elements (Wittkopp et al. 2004).

The two approaches just described have been used to assess the relative contribution of *cis* and *trans* factors in a variety of species. Because these studies sampled different numbers of genotypes, used different numbers of markers, and used different molecular techniques, it is difficult to compare their results directly (e.g., de Koning and Haley 2005). However, the results lead to some general conclusions. First, *cis*-acting eQTLs typically represent a smaller proportion of the total (~30%) than *trans*-acting eQTLs. Second, eQTLs that have the strongest effects tend to be *cis*-acting. For instance, increasing the stringency for statistical significance of eQTL effects increases the proportion of *cis*-acting eQTLs observed (e.g., Schadt et al. 2003). Finally, there are clusters of eQTLs in *trans* that affect the expression of large numbers of genes—in other words, there are portions of chromosomes that affect a larger number of genes than expected by chance alone. For instance, eight such clusters were identified in a cross between two strains of yeast (Brem et al. 2002). In the human lymphoblastoid cell lines studied by Morley and colleagues (2004), two regions of five megabases each contained six or more eQTLs out of the 142 most significant ones.

Genotype-by-environment interactions, sex-biased genes, and epistasis

Another important aspect of the genetic architecture of quantitative traits is how alleles at different loci interact with each other. Given that gene expression regulation involves numerous molecular interactions, one might expect epistasis for fitness (i.e., the contribution of nonadditive gene interactions to fitness) to be an important factor in how selection acts on gene expression. Consider, for instance, the effect of a mutation in a *cis*-regulatory sequence. If the effect of this mutation depends on the genetic background (variation in *trans*, for example), we predict an epistatic interaction.

Mathematical models of gene expression regulation have predicted that epistasis will be an important factor contributing to gene expression variation (Gibson 1996, Landry et al. 2005). In their review of *cis*-regulatory variation in humans, Rockman and Wray (2002) identified many cases of *cis*-by-*trans* interactions, where the effect of a *cis* variant depends on the genetic background. They also identified *cis*-by-*cis* interactions, where the effect of a *cis* variant depends on other *cis*-acting variants. Many interactions fell into these two categories in recent large-scale studies of the genetic architecture of the yeast, eucalyptus, and Drosophila transcriptomes (Brem et al. 2005; Kirst et al. 2005; Landry et al. 2005). Finally, a survey of allelic expression in interspecific hybrids of Drosophila reveals that many *cis*-by-*trans* interactions have accumulated since the divergence between *Drosophila melanogaster* and *Drosophila simulans* (Landry et al. 2005). Several of the approximately 30 genes studied displayed a pattern consistent with *cis*-by-*trans* interaction, which means that the divergence in gene expression between the two species was smaller, or in the opposite direction, than the divergence between alleles measured in the hybrid background. All these studies have used crosses between closely related species to identify *cis*-by-*trans* interactions, and it remains to be shown that the same interactions are common contributors to epistatic genetic variation within species.

As mentioned above, genetic variation for gene expression is extensive in nature, providing abundant raw material for evolution. Since this genetic variation will be parsed by natural selection, any factor that influences this variation will affect the course of evolution. Nongenetic sources of variation in gene expression, such as the environment or development, may also contribute to modification of gene expression. Most importantly, these effects may interact with the genotypes to shape the amount of genetic variation observed. Surveys have revealed that genetic variation in gene expression can be dependent on the sex (genotype-by-sex interaction), the environment (genotype-by-environment interaction, or genetic variation for phenotypic plasticity), and developmental stage (age-by-genotype interaction). For instance, genetic variation among strains of *D. melanogaster* is dependent on the sex in which it is measured; some genes that display genetic variation in gene expression in males may not be variable in females (Jin et al. 2001). If an eQTL experiment were performed for those genes that show genotype-by-sex interaction, different numbers and locations of eQTLs would be identified in each sex. Similarly, Landry and colleagues (2006) showed that genetic variation in gene expression among strains of *Saccharomyces cerevisiae* depends on the environment in which it is measured and that substantial genotype-by-environment interactions are evident (Figure 5.9). Understanding how these interactions are maintained, mechanistically and evolutionarily, will therefore be important in deciphering the forces acting on the regulation of gene expression.

fpo

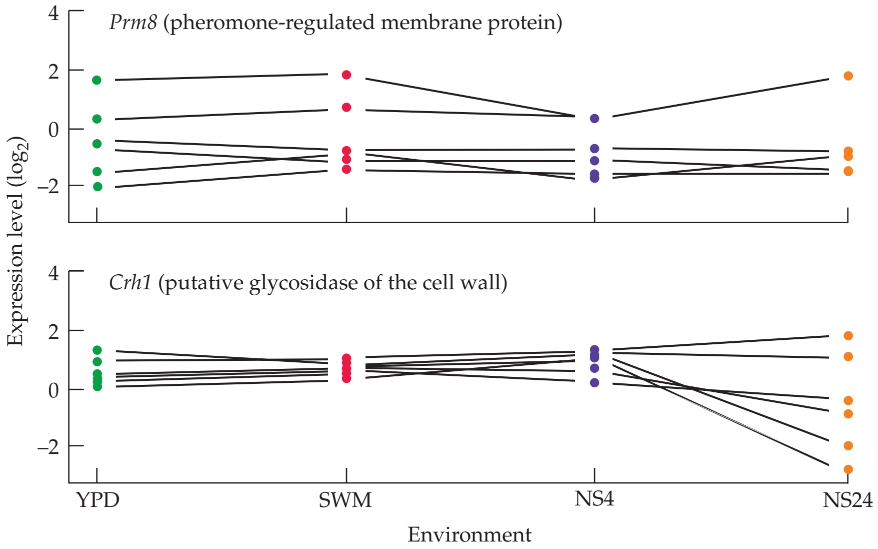

Figure 5.9 Plasticity and genotype-by-environment interaction in gene expression. The expression levels of two genes, *prm8* and *crh1*, were measured in six different yeast strains grown in four different environments. *Prm8* (top graph) shows genetic variation for gene expression, but the six strains show the same gene expression differences across growth conditions and, therefore, no genotype-by-environment interaction. *Crh1* (bottom graph) shows genotype-by-environment interaction, in that gene expression in the six strains responds differently to different environments (e.g., compare NS4 to NS24). YPD, SWM, NS4, and NS24 are different growth media. (Data from Landry et al. 2006.)

Gene Regulatory Networks, Subnetworks, and Modules

Gene regulatory networks are yet another level of organization of gene expression and evolution. A network can be represented as a graph—a collection of nodes connected by edges, interacting as a system (Barabasi and Oltvai 2004). A subnetwork is a subset of the whole network. A module is a coherent subnetwork whose structure and function is largely independent of interactions with members of other subnetworks. These general concepts have been fruitfully applied at many levels of the biological hierarchy, from molecules to species. Usually, the nodes represent biological units (proteins, cells, organs, individuals, and so forth), and the edges represent interactions between them. For example, a trophic network depicts prey/predator relationships and the energy flow through a food web. The species are the nodes and the edges are "who eats whom" interactions. At the other end of the scale, a gene regulatory network describes the transcriptional and translational web and its regulation (by proteins, RNAs or environmental signals) within a cell. In this case, the nodes are genes and the edges represent regulatory interactions between pairs of genes or a shared regulatory element.

Biological systems are complex and the information must often be simplified to gain useful insights. In the simplest case, networks are considered as Boolean objects, represented by uniform nodes and connected by undirected edges (for more complex models, see Proulx et al. 2005). Boolean nodes have only two discrete states, *on* or *off*, and the nodes interact through logical or Boolean functions. In this modeling framework, only the topology of the network is retained. Even with these simplifications, network behaviors are still extremely rich. A straightforward and well-established indicator of network topology is the distribution of the connectivity (the number of edges of a particular node). A second indicator is the clustering coefficient, which measures the degree of connection between nodes connected to the same specific node. Nodes densely connected to each other define clusters, or modules. In a gene regulatory network, the connectivity and the modularity provide direct insights about important concepts in evolutionary biology, like epistasis, canalization, and plasticity. Wagner (1996) modeled the evolution of transcriptional regulatory networks and concluded that more densely connected networks are more insensitive to disruption by mutations.

Much excitement has been generated by high-throughput experimental techniques such as genome-wide expression profiling and location analysis (ChIP-on-chip) because these techniques promise to allow for rapid reconstruction of regulatory networks. Large-scale gene perturbation experiments generate valuable information about the number of genes whose expression is affected by an environmental or a gene perturbation (mutation, overexpression or knockout). Regrettably, such perturbation experiments cannot distinguish direct from indirect interactions. On the other hand, protein–DNA interaction data are produced at a slower rate, but directly identify the binding sites of transcriptional factors. Accordingly, Lee and coworkers (2002) performed a genome-wide analysis to determine the binding distribution of 106 known regulatory proteins along the yeast genome. The results suggest that about ten percent of these regulatory genes are autoregulated. This proportion seems substantially smaller than in *E. coli*, where most genes (52% to 74%) are autoregulated (Shen-Orr et al. 2002). Also, about 37 percent of the yeast regulators are involved in feed-forward loop motifs; these contain a regulator controlling a second regulator that acts together with the first one to bind a common target gene. In *E. coli*, this type of motif was found to control about 240 genes.

In another approach, Stuart and colleagues (2003) and Bergmann and colleagues (2004) used published perturbation experiments to compare the network topologies of evolutionarily distant organisms such as *A. thaliana*, *C. elegans*, *D. melanogaster*, *E. coli*, *H. sapiens*, and *S. cerevisiae*. They found that, for all these organisms, the connectivity is distributed as a power law, with negative exponents of similar magnitudes. These power law connectivity distributions indicate that most genes have few connections while a few genes have many connections. Moreover, there is a significant enrichment of highly connected genes as compared to random networks. Power law distributions have been attributed to dynamically evolving networks and to systems that are optimized to provide robust performance in uncertain environments. For gene networks, gene duplication could be the proximal mechanism explaining this enrichment of connected genes (Teichmann and Babu 2004). They also found that expression networks are highly clustered. In yeast, the network comprises from 5 to 100 independent gene modules, depending on the analysis methods and an arbitrarily defined threshold. However, modules and interactions may vary significantly between organisms.

Conclusions

Clearly, there are myriad regulatory interactions and mechanisms that play a role specifying the location, timing, and level of gene expression. This provides a rich stage where hypotheses can be clearly stated and the variables of evolutionary genetics (epistasis, pleiotropy, plasticity, and so forth) can be more concretely defined. Also, the impact of a number of attributes on variation in genomic expression can be assessed. Accordingly, investigators have described many determinants of variation in genomic expression within and between species, including membership in functional classes, pattern of sex and tissue biases, properties of the protein–protein interaction network, protein attributes, and mRNA abundance itself. Also being disentangled are the effects of mutations and selection on levels of gene expression polymorphism and divergence. Interpreting data integrated from disparate sources is likely to remain a key challenge to understanding the overall picture of the regulation and evolution of genomic expression.