The Evolution of Gene Regulatory Interactions

DAVID A. GARFIELD AND GREGORY A. WRAY

Changes in the timing and level at which genes are expressed are known to play an important role in evolution, but the mechanisms underlying changes in gene expression remain relatively obscure. Until quite recently, evolutionary biologists, like most biologists, tended to study single genes as isolated entities. These studies have added enormously to our understanding of biological evolution. But because gene regulation by its very nature involves interactions between two (or more) genes, researchers have missed a range of evolutionary phenomena that can be observed only at the level of networks of interacting genes. In this article, we consider the change in perspective that genomic technologies—particularly the advent of large-scale platforms for DNA sequencing, genotyping, and measuring gene expression—are bringing to evolutionary biology. We focus specifically on how these technologies can and are being used to increase our understanding of how and why gene expression evolves.

Keywords: evolution, gene networks, genomics, gene expression, gene regulation

The idea that changes in gene regulation can play an important role in evolution is not new. In an influential article published in 1975, Mary-Claire King and Allan Wilson argued that because the sequence and function of proteins isolated from humans and chimpanzees were so similar, something other than protein evolution per se must underlie the phenotypic differences between these two species. They posited that changes in the regulation of gene expression were responsible for more adaptive evolution than changes in the protein-coding regions of genes. King and Wilson's hypothesis was framed without reference to specific molecular mechanisms. However, there was growing evidence from the field of developmental genetics supporting the hypothesis that "small differences in the time of activation or in the level of activity of [even] a single gene" could have important evolutionary consequences owing to the extensive consequences that changes in regulatory interactions could have on developmental processes and, thus, organismal form and function.

In 1983, Rudolf Raff and Thomas Kaufman pushed the importance of changes in gene regulation for evolution further with their book *Embryos, Genes, and Evolution*. Drawing from hundreds of studies in developmental genetics, embryology, and evolutionary biology, Raff and Kaufman hypothesized that mutations affecting the function of regulatory genes are likely to underlie many evolutionary changes in morphology. Their reasoning was based on two observations about biological systems. The first is that targeted perturbations of biological systems often resulted in large-scale coordinated changes in organismal phenotypes. Clear examples of this coordination are homeotic mutations, such as those affecting the function of the transcription factor *Ubx* or its DNA binding sites. Mutations in any one of these genes leads to the transformation of one body segment into another (Gilbert 2006). These mutations are clear examples of how changes in just a handful of regulatory interactions could cause large-scale morphological changes.

Their second observation is that the effects of changes in regulatory genes or their target DNA could be "tuned" to affect only specific tissues. For enzymes and structural proteins, amino acid sequence directly dictates function, determining, for instance, the catalytic properties and substrate specificity of an enzyme. Changes in amino acid sequence thus directly affect protein function, and does so, *ceteris paribus,* in all tissues in which the gene is expressed. The effects of regulatory mutations, on the other hand, are indirect, affecting phenotype through the activity and expression of the genes that they regulate. For example, a regulatory mutation could change when and where in development cell migration or proliferation begins, thereby changing adult morphology without affecting the underlying cellular processes. Changes in the protein-coding sequence of genes underlying essential functions such as cell division or proliferation, in contrast, are highly *pleiotropic*; these changes affect every instance of cell

BioScience 60: 15–23. ISSN 0006-3568, electronic ISSN 1525-3244. © 2010 by American Institute of Biological Sciences. All rights reserved. Request permission to photocopy or reproduce article content at the University of California Press's Rights and Permissions Web site at *www.ucpressjournals.com/reprintinfo.asp*. doi:10.1525/bio.2010.60.1.6

division, often with dramatic and deleterious consequences. Thus, Raff and Kaufman argued, even though mutations occur in all genes, mutations affecting regulatory interactions were less likely to be deleterious than changes in essential protein-coding genes.

The hypotheses of King and Wilson and Raff and Kaufman together pointed to the potential significance of regulatory interactions in evolutionary change. In so doing, they directed attention away from single genes and toward networks of interacting genes. Although these ideas later proved to be remarkably prescient, at the time they were conceptual arguments with few clear supporting examples. For many years, a major obstacle prevented progress in studying the evolution of gene networks: It's easier to identify a gene than a gene interaction, and much easier to identify changes in gene coding sequences than changes in gene regulation. As a result, we know a lot more about how individual genes and proteins evolve than we do about how the interactions between genes evolve, and even less about the effects of changes in regulatory interactions on organismal fitness. The gap in our knowledge was enormous during the 1980s, but has since narrowed with the introduction of several methods.

The *Hox* paradox

The introduction of the polymerase chain reaction (PCR) and techniques for visualizing the location of proteins and mRNAs within whole cells and embryos during the 1980s made it possible to observe when and where within an embryo specific genes and their protein products were expressed, and observe what happened to these patterns of expression when other genes were experimentally modified. It also became possible to compare gene expression patterns between different species, and in the 1990s hundreds of studies were published comparing the expression of the same gene in embryos of diverse species.

From these studies came some important results supporting the significance of changes in gene regulatory interactions in evolution. The first was the realization that even animals with very different body plans share a common set of developmental regulators whose interactions are strikingly conserved. A key early discovery was that *Hox* genes, which were known to play a key role in patterning the primary body axis of fruit flies, were expressed in similar patterns and appeared to be playing similar regulatory roles in vertebrates and other animal phyla, despite the relatively large (> 70%) divergence in *Hox* gene sequences among phyla (Holland and Hogan 1988, Lemons and McGinnis 2006). *Hox* genes encode transcription factors, so other kinds of transcription factors were similarly examined, as were genes encoding other kinds of regulatory proteins, such as signaling molecules and their receptors. These studies often produced the same basic result: Homologous regulatory genes are present in distantly related groups of animals and, in many instances, seem to play similar regulatory roles even in animals with highly diverged body plans.

The second result was that although the same basic "tool kit" is deployed to construct a variety of body plans, many important morphological differences among related species were found to correlate strongly with changes in the expression pattern of these fundamental regulators, supporting Raff and Kaufman's claim that changes in developmental regulators play a role in morphological evolution. For instance, Averof and Patel (1997) examined the expression of *Hox* genes in crustacean embryos and found that shifts in the anterior expression segmental boundary of *Ubx/Abd-A* corresponded to changes in feeding versus walking limb morphologies in those segments between crustacean species. Similarly, Cohn and Tickle (1999) showed that the expression of several *Hox* genes was expanded along the body axis in pythons relative to mice, suggesting that the loss of limbs in snakes was due to an expansion of the "neck domain" rather than a loss of the genes encoding legs.

Taken together, these findings presented researchers with a paradox. On one hand, the basic machinery underlying early development, such as the *Hox* genes, is widely conserved among divergent phyla. But at the same time, these genes also underlie the development of distinct morphologies between more closely related species. The resolution of this "*Hox* paradox" is that the general role of many genes in patterning the embryo *has* been preserved, but the precise pattern of their expression or their influence on later events of development have both changed. These modifications are possible only through changes in regulatory interactions, whether mediated through changes in protein or nucleic acid sequences.

The number of regulatory differences affecting gene expression that distinguish crustacean and mouse development must be enormous, and sorting them out, much less identifying the relative contributions of selection and drift to regulatory change, is probably an intractable challenge. In contrast, cases where regulatory gene expression differs among more closely related species provide practical opportunities to delve into the genetic bases of regulatory interactions and figure out exactly how gene interactions change during the course of evolution. For example, several species of fruit fly show differences in wing coloration that are now known to result from mutations in the *cis*-regulatory region of the pigment gene *yellow* (Gompel et al. 2005). This study and many others support the contention that changes in *cis*-regulatory regions can have important consequences for adaptation and the evolution of novel traits through the effects of these mutations on gene expression (Wray 2007).

Changes in regulatory interactions, such as those that alter the expression of *yellow,* have largely been studied on a case-by-case basis. Several recent, genome-scale technologies offer opportunities to expand the scale of investigation from specific genes and gene interactions to (potentially, at least) all of the genes and interactions encoded in a given species' genome. There are few biological systems for which we understand the patterns of connections between genes in sufficient detail to be able to discuss the evolution of gene expression in the context of networks of local interactions. These new technologies are already providing insight about how regulatory interactions evolve and the sorts of impacts these regulatory changes have on organismal phenotypes. In the remainder of this article, we will discuss how these new technologies can be used to understand the causes and consequences of changes in gene regulatory interactions. We will focus on four specific questions: (1) How common is gene expression variation within and between species? (2) What types of genetic changes underlie changes in gene expression? (3) How does natural selection work to shape gene interactions? and (4) What kinds of changes in gene interactions produce trait differences?

How common is gene expression variation?

The first step in understanding how gene regulatory interactions have evolved between species is to ask two more basic questions: (1) How often do gene expression profiles differ between related species? and (2) How often are expression differences due to genetic versus nongenetic (i.e., environmental) factors? Assaying a gene's expression profile using traditional methods, such as Northern blots, is a slow and labor-intensive process. Even though many cases have been identified in which the expression profile of a gene clearly differs between two species, it remains difficult to place these results into a broader genetic context. Do the expression patterns of other genes change at the same time as the gene of interest? Is there variation between individuals within a species? How often do changes in gene expression evolve in general? Does this differ among different kinds of genes?

New technologies for assaying gene expression are making it easier to gather the expression measurements needed to answer these questions using multiple genes, multiple individuals, and multiple species. The first major breakthrough came with the invention of microarrays (figure 1a). Many of the first studies using microarrays were carried out using yeast. It has long been known that some strains of yeast grow better under particular conditions. Fay and colleagues (2004) asked what kinds of genes allow some yeast strains to cope with heavy metals in their environment while others cannot. They raised different yeast strains on media containing increasing amounts of copper, and then measured gene expression using a microarray that assayed most known genes in the yeast genome. Although the expression of several hundred genes changed in response to copper in one or more of the strains, only 20 of these changes were correlated with resistance to copper across strains. Using a genomic approach, the researchers were able to narrow the search for the genetic basis for copper tolerance from the entire genome (> 7000 genes) to less than two dozen genes, including genes encoding proteins involved in stress response and metal binding that were differentially regulated in the strains most tolerant to copper. Importantly, it became clear in subsequent studies that none of the genes identified was sufficient alone to produce copper tolerance. Fay and colleagues (2004) were thus able to show that while copper negatively affected the growth of all yeast strains, the species as a whole carried some genetically based variation in gene expression that allowed it to adapt to changes in the amount of copper found in its environment, though no changes affecting single genes appeared sufficient to drive the evolution of copper tolerance.

Another early application of microarrays to an evolutionary problem involved the killifish, *Fundulus heteroclitus*, a member of the minnow family whose distribution extends from the Gulf of Mexico to Maine. Several publications had previously shown that killifish populations are locally adapted to warm- and cold-water temperatures across this broad geographic range (reviewed in Hochachka and Somero 2002). A study by Oleksiak and colleagues (2002) used microarrays to investigate how many and what kinds of genes differ in expression between and among fishes from warm- and cold-water populations. Their results were striking. As many as 18% of all loci differed significantly in their expression between individuals from the same population, suggesting that natural populations can harbor enormous variation in gene expression—far more than had been expected—on which selection could potentially act. But at the same time, there were not vastly more differences in gene expression between locations as compared with within locations, suggesting that a significant amount of the observed gene expression variation between species may be the result of drift rather than natural selection.

Microarrays have now been used in a variety of evolutionary and ecological studies. The results support the same basic conclusion: Gene expression patterns can be enormously variable between species and populations, and these differences can play adaptive roles. But at the same time, much of the difference in gene-expression profiles between species may be the result of neutral evolution or environmental differences, highlighting the need for carefully designed experiments incorporating multiple biological replicates.

Microarrays have also been used in developmental biology studies aimed at inferring how genes interact during development. One of the best-studied gene regulatory networks in development is the network of the sea urchin embryo. The initial interactions within this network took decades to elucidate using traditional techniques for measuring gene expression, such as Northern blots and quantitative PCR. Microarray and similar technologies have vastly increased the speed at which interactions are discovered and added to the network. For example, by subjecting developing embryos to chemical agents that perturb embryonic axis formation, and then measuring the expression of thousands of genes on microarrays, researchers in Germany (Poustka et al. 2007) were able to identify dozens of potential interactions for the first time. Other studies have gone further, using targeted knock-downs of specific genes, followed by genome-wide expression measurements to elucidate the basic structure of previously unknown gene regulatory networks (Imai et al. 2006, Su et al. 2009).

Although microarrays are enormously useful in understanding genetic regulatory networks, they have several

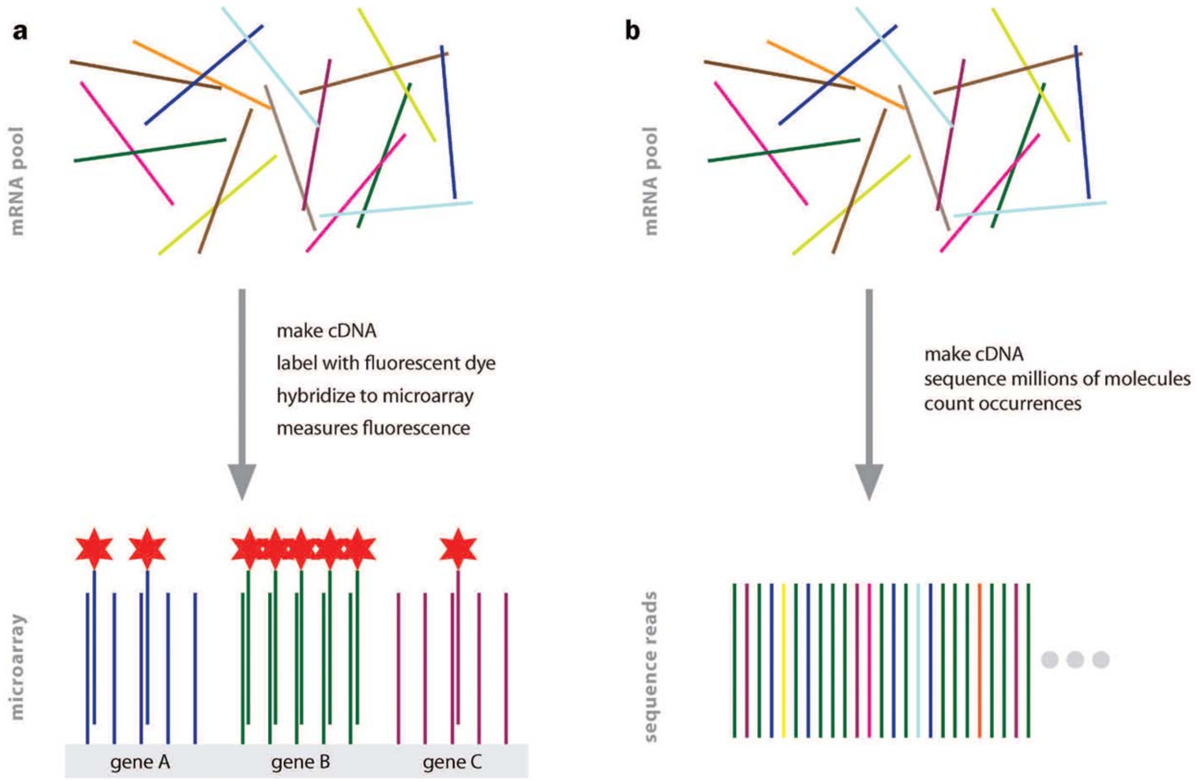

Figure 1. High-throughput gene expression measurements. Most techniques for measuring gene expression begin with the isolation of the mRNA molecules from cells, tissues, or organisms of interest. These mRNA molecules are then reverse-transcribed to get a sample of complementary DNA (cDNA) molecules (DNA copies of the original mRNAs), using methods that preserve the original proportions of the various mRNAs. In this graphic, each colored line represents a different mRNA molecule. Two methods have been developed that can measure the relative abundance of tens of thousands different mRNAs simultaneously. (a) The first such method is based on microarrays, small glass slides onto which thousands of microscopic spots have been printed that contain short segments of DNA that are complementary to the mRNAs under investigation. The cDNA that is to be measured is labeled with a fluorescent dye and hybridized to the DNA microarray. A laser is used to excite the dyes. The level of fluorescence is proportional to the number of labeled transcripts bound at a region of the slide. (b) An alternative approach is to make use of high-throughput sequencing technologies (box 1). The number of times that the cDNA corresponding to a particular gene is read is proportional to its abundance within the overall pool. Since it is now possible to sequence millions of such "tags" in a single run, the precision of measurements is quite good. There is no need to know the precise sequence of the genes under investigation, because the cDNA sequence is determined at the same time as its concentration is being measured. This is an important advantage in evolutionary studies, where full-genome sequences may not be available for some of the species under investigation.

important drawbacks (figure 1a). First, because microarrays rely on exact or near-exact matches between the messenger RNA (mRNA) transcripts of the genes being studied and the strands of DNA (oligonucleotides) attached to the chip, the sequence of the genes under investigation must be known in advance. Any sequence difference between the transcripts under investigation and the probes on the microarray can interfere with hybridization, making this approach difficult to apply to evolutionary comparisons within species with even moderate levels of genetic variation, and much less useful for comparisons among species.

Newer technologies relying on ultra high-throughput DNA sequencing (figure 1b, box 1) circumvent these problems (Wang et al. 2009). The basic method of high-throughput sequencing involves drawing an individual mRNA from the total pool of transcripts and sequencing it to identify which gene it came from. This is done millions of times per sample, providing information about a gene's sequence while also measuring the relative abundance of its message—independent of any prior knowledge about the gene sequence. These new technologies also have the advantage of being able to accurately estimate the concentration of even low-abundance mRNAs. This feature is of particular importance to the study of gene regulatory networks, where many functional components, such as transcription factors, are commonly expressed at levels as low as 10 molecules per cell. While these technologies are new, they are already turning up exciting findings, such as the revelation that there are a significant number of noncoding RNAs and antisense transcripts in the genome (He et al. 2008). The functional consequences of changes in these noncoding RNA transcripts is an active area of investigation.

What kinds of molecular interactions change over evolutionary time?

Genome-scale technologies are also being used to identify many of the specific kinds of changes in gene interactions that underlie modifications in gene expression. Gene networks are often drawn as lines connecting gene names, which gives the impression that all interactions are of the same nature. In reality, genes can interact in many different ways: The protein product of one gene may influence the expression of another gene, the protein products of two different genes may form a complex whose function depends on both, one protein may phosphorylate or cleave another protein, and so forth.

Detecting and characterizing each of these kinds of molecular interactions requires a different functional assay, which presents a significant practical challenge. The yeast two-hybrid assay was one of the earliest genome-scale functional assays developed to identify interactions, and led to the first graphical representations of such interactions at a genome-wide scale (box 2). However, yeast two-hybrid assays have drawbacks: They are more labor-intensive than most genome-scale approaches, and they generate a large proportion of false positives. Although several studies have identified evolutionary changes in protein interactions in single genes, genome-scale comparisons of protein-protein network organization remain a challenge for evolutionary studies.

Significantly more progress has been made in studying the evolution of protein-DNA interactions across genomes. The key technology, called ChIP-chip or ChIP-Seq (where ChIP stands for chromatin immunoprecipitation), involves chemically attaching proteins to DNA in live cells, fragmenting the DNA, then recovering fragments bound by a specific protein to find out where in the genome the protein binds. This approach has also benefited from genome-scale technologies (box 2). A study by Odom and colleagues (2007) used this approach to compare the binding locations across mouse and human genomes for four transcription factors that regulate gene expression in the liver. They found that all four transcription factors retained the same sequence preferences when binding to DNA in the two species. Remarkably, however, the specific sites where the transcription factors bound in the genome were often different, with 41% to 89% of binding sites present in only one species, depending on the transcription factor.

These results demonstrate that specific intermolecular interactions can turn over even in cases where their functional consequences are evolutionarily conserved. Similar results have also been observed for protein-DNA interactions in yeast and fruit flies (Jin et al. 2001, Brem et al. 2002), suggesting that changes in protein-DNA binding sites may not be a rare phenomenon even on relatively short evolutionary time scales. These hint that regulatory interactions may be more dynamic than the evolution of protein function, but they also point to an important lacuna in our knowledge:

Box 1: Ultra high-throughput DNA sequencing technologies.

For more information see Shendure and Ji (2008).

Roche/454: A small biotechnology company called 454 developed the first "next-generation" or ultra high-throughput DNA sequencing technology to reach the market (the company has since been acquired by Roche). This technology uses pyrosequencing, a process that had been developed earlier by a company called Biotage. The sequence of single-stranded template DNA is determined by the addition of free nucleotides (A, C, T, G) one at a time in the presence of DNA polymerase, luciferase, additional enzymes, and a sequencing primer. The successful incorporation of a complementary base releases a photon that can be detected by a sensitive camera. By observing which bases cause the release of the photon, the sequence of the original template can be reconstructed. 454 technology scales up pyrosequencing to hundreds of thousands of simultaneous reactions by isolating individual DNA templates on beads, amplifying them with emulsion PCR (polymerase chain reaction), and placing them in microscopic wells within a hexagonal array. Average sequencing reads are up to 450 base pairs, by far the longest of any "next-gen" sequencing method. As such, this approach is well suited for de novo sequencing of large genomes and the discovery of genetic variants within populations.

Illumina/Solexa: Another small biotechnology company called Solexa (subsequently acquired by Illumina) developed the second next-generation DNA sequencing technology that is based on a rather different approach. Template DNA is sheared and ligated to universal adaptors, which are then attached at a very low concentration to a surface. These bridges of DNA (attached to the plate on both ends) are amplified by PCR into about 40 million colonies. Sequencing is carried out by the addition of fluorescently labeled, reversible terminators (3' modified A, C, T, G), along with primers to the adaptor sequences and DNA polymerase. A laser excites the fluorescence of each colony and the color is read with a camera. The 3' terminator is removed by enzymatic reaction, and the process is then repeated to generate reads about 35 base pairs in length. This technology produces a very large number of short reads, and as such is particularly well suited for measuring gene expression (figure 2b) or protein-DNA binding (see box 2, ChIP-Seq).

Applied Biosystems /Agencourt: The third "next-gen" technology to reach the market was developed by a company called Agencourt (since acquired by Applied Biosystems and sold under the trade name SOLiD). It uses completely different chemistry to sequence DNA based on ligation. Template DNA is sheared, universal adaptors are added, and then this library is amplified. Sequencing is carried out by the sequential ligation of 9-base, single-stranded oligonucleotides, each bearing a different fluorescent label on its central base (nnnnAnnnn, nnnnTnnnn, nnnnCnnnn, nnnnGnnnn). Oligonucleotides that match the template sequence hybridize, and a laser and camera system is used to identify which base matches. The process is then repeated using a series of initial primers, each located one base further back on the template, and so forth. This approach currently generates up to 300 million reads of about 50 base pairs, and has similar applications to Solexa technology.

Box 2. Methods for assessing direct molecular interactions throughout the genome.

Yeast two-hybrid assays: This approach can identify proteins that interact with each other by testing their affinities in living yeast cells. This is usually done with reference to a particular protein of interest (the "bait"), which is tested against all the other proteins in the genome. A library is constructed from the DNA of the organism of interest by fusing each gene to one member of a pair of proteins needed to activate a reporter gene, such as *LacZ*. The gene encoding the bait protein is fused to the other member of the activator pair. The library of genes encoding potential interactor proteins is then transfected into yeast cells, with one construct per cell. In any cell where the cloned protein binds to the bait protein, the two halves of the activator will be brought close enough to one another to drive the expression of the reporter gene. Yeast colonies expressing the reporter gene can be identified visually (they will be blue, for instance, if expressing *LacZ* as a reporter). The DNA of the clone contained by positive cells is then isolated and sequenced to reveal which protein is interacting with the bait protein.

ChIP-CHIP and ChIP-Seq: This method identifies where a particular protein is bound to DNA in living cells, and is used to identify which genes a given transcription factor regulates. Cells are fixed with formalin to bind the transcription factors to the specific section of DNA they are currently regulating, and the DNA is sheared into relatively short segments. An antibody specific for the protein of interest is used to capture protein-DNA complexes. The DNA is then stripped from the protein and identified either by hybridization to a microarray consisting of all regions of the genome under investigation or, increasingly, by sequencing the pool of DNA using high-throughput technologies such as Solexa sequencing (see box 1). This identifies specific segments that are functional regulatory sites in that cell type under specific environmental conditions.

We have little information about the consequences of most of these changes on gene expression or organismal traits. Addressing this issue represents an important next step for understanding how gene regulatory interactions evolve.

Where does natural selection act in the genome?

Microarrays and other high-throughput methods for measuring gene expression have revealed that changes in gene expression are very common among individuals and among species, and there is evidence that these changes are due, at least in part, to changes in gene regulatory regions over even short evolutionary time scales. Still, we know very little about the evolutionary mechanisms driving these changes. Presumably, some gene expression differences are the result of natural selection, whereas perhaps the majority of differences are due to the effects of random mutation and drift.

Because positive selection, negative selection, and drift (neutral evolution) leave different patterns of change in genomes, DNA sequence comparisons can be used to infer the extent to which natural selection and drift have influenced the evolution of specific genes or regions of DNA known (or presumed) to be involved in regulating gene expression. Several statistical tests have been developed for this purpose, but most have at their core a common idea: comparing the rate at which sequence changes accumulate in a gene or region of interest to a (nearby) region known or presumed to be evolving neutrally. If a region of interest shows a significantly higher rate of change than a neutral region, then this constitutes evidence of positive selection. If the region shows a significantly lower rate of change, then negative selection is most likely acting against changes in this region. Similar rates of change suggest that the region of interest is evolving neutrally.

Traditionally, these tests for selection have been applied to individual genes and regulatory elements. However, researchers are increasingly turning their attention to surveys of natural selection at the level of whole genomes to ask questions about broad evolutionary trends—trends that cannot be seen from the perspective of individual genes or regulatory elements. The first genome-wide scans focused on changes in protein-coding regions of the genome. Clark and colleagues (2003) and Nielsen and colleagues (Nielson 2005, Nielson et al. 2005) carried out some of the first genome-wide scans for positive selection, comparing chimpanzees and humans. They found that genes involved in immune responses are, on average, more likely to show evidence of adaptation than most other categories of genes. This result fits expectations: The immune system is under constant siege by pathogens and therefore continuously adapts to meet new challenges. Surprisingly, though, genes involved in neural development and neural function as a group did not show any evidence of adaptation. Other studies subsequently confirmed these findings, but left open the question of what kinds of mutations contributed to the dramatic evolution of human brain size and cognitive function.

One possibility was the suggestion by King and Wilson (1975) that the basis of many adaptations would be found in regulatory interactions. Haygood and colleagues (2007) modified the approach used for protein-coding regions and used it to scan for positive selection acting on likely regulatory regions during the evolutionary divergence of humans and chimpanzees. The results highlighted two categories of genes whose regulatory regions were disproportionately likely to show evidence of positive selection on the branch leading to humans, but not on the branch leading to chimpanzees: genes involved in diet and metabolism and genes involved in neural development and cognitive function. These categories make sense in the light of human-specific traits, since both our diet and cognition are distinct outliers among the great apes. This study also found that more genes show evidence for adaptive changes in regulatory regions of DNA than in protein-coding regions, supporting King and Wilson's idea that changes in gene regulation play an important role in adaptation.

Until relatively recently, whole-genome sequences were available only for model organisms such as *Drosophila* and *Caenorhabditis elegans*. But as the cost of large-scale sequencing projects has decreased, the genomes of more and more nonmodel organisms are being sequenced. This has been accomplished largely through refinements in the same basic chemistry for sequencing DNA that was invented more than 30 years ago, but the trend in whole-genome sequencing is set to accelerate as the ultra high-throughput sequencing approaches mentioned above are brought to bear on genomes. Three of these methods are being marketed (box 1), and more are in development. Each of these "next-generation" sequencing technologies uses a different underlying chemistry, and they all differ from Sanger's method, which has dominated molecular biology for the past two decades. What they have in common is immense scale. Whereas the DNA sequencers used for the human genome project produced approximately 60 thousand bases in a single run, ultra high-throughput sequencers can produce a billion or more. The cost per base to sequence DNA on these instruments is a tiny fraction of the cost of using the previous generation of technology.

What kinds of changes in gene interactions produce trait differences?

Another area where the application of new technologies to genome-scale data sets has had an impact is in identifying which genes and mutations influence the evolution of particular traits. Even in cases in which it is clear that a gene's expression profile has changed, or in which there is a strong signature of positive selection, it is rarely obvious which organismal traits are affected. For more than a century, it has been clear that multiple genes influence variation in many traits, including diverse aspects of morphology, behavior, and physiology (Lynch and Walsh 1998). The field of quantitative genetics developed powerful statistical methods to estimate the number of genes that influence a particular trait. This sort of inference requires genetic markers with known locations in the genome: Correlations between the genotype at a particular marker and a trait of interest indicate that the two lie near each other on the same chromosome.

"Near," however, is a relative term. With a few dozen markers, the distance between the closest marker and the causal gene can be tens of millions of base pairs apart, space enough for hundreds of genes. The more markers one can survey, the finer the physical mapping. For most of the 20th century, markers were limited to genes that produced obvious trait differences, such as eye color or wing shape in fruit flies. These classical genetic markers were extremely useful, but few in number. This limitation began to disappear with the introduction of molecular methods into studies of evolutionary biology during the 1980s. Any sequence difference can be used as a genetic marker, provided one can read, or genotype, that region of the genome. The ability to sequence DNA opened up the ability to survey many more markers, but introduced two new limitations, cost and time. Genotyping classical genetic markers is fast and cheap: One looks at individuals and records their appearance. In contrast, genotyping molecular markers typically requires generating a separate DNA sequence for each marker in each individual. Costs and labor rise almost linearly with the number of markers surveyed, and most studies by evolutionary biologists surveyed at most a few hundred markers. For an organism like a mouse, this means narrowing down a causal mutation to an area that contains several dozen to a hundred genes. Thus, these methods are not well suited for identifying exactly which genes and mutations are responsible for phenotypic differences.

Recent technological improvements in genotyping now make it possible to genotype over a million genetic markers at once, allowing for multiple markers per gene even in relatively large genomes, such as our own. The most commonly used technology (figure 2) requires knowing in advance where genetic variation is distributed in the genome. While this limits the use of this technology to well-characterized systems, such as humans, the increasing availability of genome-wide sequence data from multiple individuals of the same species will increase the extent to which this technology can be employed, and may replace this technology as it becomes possible to sequence entire individuals de novo.

Genome-wide genotyping has been successful in identifying a number of loci affecting susceptibility to diseases such as type II diabetes. The most common outcome of these association studies is that trait variation is influenced by interactions between large numbers of genes. In two studies from the Wellcome Trust (Lettre et al. 2008, Weedon et al. 2008), more than 400,000 genetic markers were surveyed in several thousand individuals to identify genes underlying differences in height. Strikingly, the 10 loci with the largest contribution to height collectively explain only about 2% of the genetic basis for differences in height between individuals. Thus it appears that thousands of DNA variants influence this most basic and highly heritable of anatomical traits—a large number of gene interactions by any reckoning—with no single genetic variant having an overwhelming effect on height. This result emphasizes just how important it is to study the evolution of gene interactions rather than genes (even many genes) in isolation.

More recently, researchers have begun to use genome-wide genotyping to identify genetic variants that affect gene expression (Gilad et al. 2008). A paper by Brem and colleagues (2002) exemplifies the power of this approach. In this study, the researchers crossed two strains of yeast to look for correlations between DNA polymorphisms and gene expression. They identified 570 genes whose expression was influenced by one or more loci. These expression-influencing loci fell largely into two classes. The first category consisted of loci that influenced the expression of only a single gene. These loci were located in *cis* (physically close) to the gene that they influenced, presumably residing in a region of regulatory DNA such as a promoter or enhancer. The other category

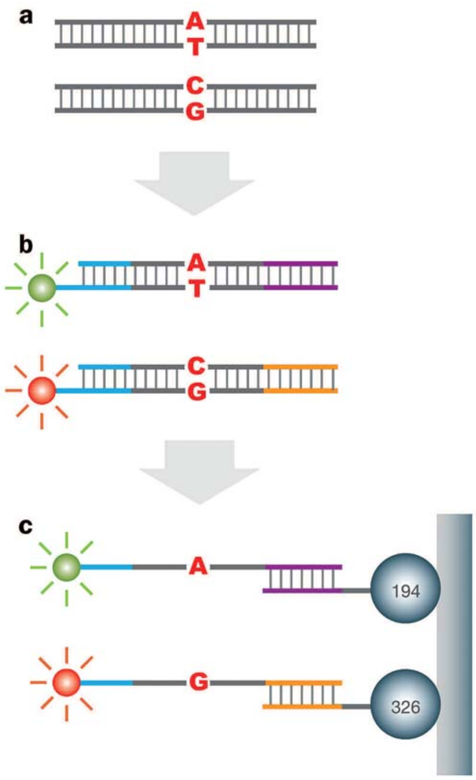

Figure 2. Technology behind genome-wide genotyping. One of the most widely used platforms for high-throughput genotyping is the GoldenGate assay marketed by Illumina. For each potentially polymorphic base of interest in the genome, three primers are introduced into the reaction mix. The first primer binds to both a specific glass bead with an address label and to a nonvariable region of the gene of interest. (The address label is etched into the glass bead.) The other two primers are specific to one of the two potential bases at a variable site, and each is labeled with a different fluorescent dye. The beads now contain both a bound complementary DNA molecule and a fluorescent tag whose color depends on which of the polymorphic bases is present in that sequence. A laser and camera system is then used to simultaneously measure the fluorescence of an individual bead and the address label on that bead. The color of the fluorescence reveals which of two potential bases is present at the polymorphic site.

consisted of loci that influenced multiple genes and were found in *trans* (far from) the genes that they influenced. Intriguingly, the multiple genes influenced by a single *trans* locus were very often functionally related, suggesting that changes in a single regulatory gene can have a coordinated effect on several aspects of an organism's phenotype. Brem and colleagues (2005) carried out a follow-up study to identify expression-affecting loci whose influence was seen only in the presence of specific genetic backgrounds. They were able to identify hundreds of loci whose influence was visible only in the presence of a specific genotype at another locus, demonstrating that many of the regulatory interactions separating these two strains of yeast depend on multiple, interacting genes.

Conclusions

Changes in gene regulation lie at the heart of many important phenotypic differences between species. New technologies are allowing us to explore the consequences and causes of changes in gene expression in ways never before possible, and are opening up new perspectives on how evolution proceeds. However, there are still significant challenges facing biologists interested in the evolution of gene regulation, not the least of which is how to deal with the enormous quantity of data produced by new technologies. The high levels of intraspecific variation in gene expression as well as the polygenic nature of many traits of interest will also pose challenges, and will require biologists to employ sufficiently large sample sizes and carefully controlled experiments to make full use of the information provided by the new technologies.

Acknowledgments

Thanks to Courtney Babbitt, Jeni Croce, and Jenny Tung for helpful comments and discussion.

References cited

Averof M, Patel NH. 1997. Crustacean appendage evolution associated with changes in *Hox* gene expression. Nature 388: 682–686.

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752–755.

Brem RB, Storey JD, Whittle J, Kruglyak L. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436: 701–703.

Clark AG, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302: 1960–1963.

Cohn MJ, Tickle C. 1999. Developmental basis of limblessness and axial patterning in snakes. Nature 399: 474–479.

Fay JC, McCullough HL, Sniegowski PD, Eisen MB. 2004. Population genetic variation in gene expression is associated with phenotypic variation in *Saccharomyces cerevisiae.* Genome Biology 5: R26.

Gilad Y, Rifkin SA, Pritchard JK. 2008. Revealing the architecture of gene regulation: The promise of eQTL studies. Trends in Genetics 24: 408–415.

Gilbert SC. 2006. Developmental Biology. Sinauer.

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

Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nature Genetics 39: 1140–1144.

He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. 2008. The antisense transcriptomes of human cells. Science 322: 1855–1857.

Hochachka PW, Somero GN. 2002. Biochemical Adaptation. Oxford University Press.

Holland PW, Hogan BL. 1988. Spatially restricted patterns of expression of the homeobox-containing gene *Hox* 2.1. during mouse embryogenesis. Development 102: 159–174.

Imai KS, Levine M, Satoh N, Satou Y. 2006. Regulatory blueprint for a chordate embryo. Science 312: 1183–1187.

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster.* Nature Genetics 29: 389–395.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188: 107–116.

Lemons D, McGinnis W. 2006. Genomic evolution of *Hox* gene clusters. Science 313: 1918–1922.

Lettre G, et al. 2008. Identification of ten loci associated with height highlights new biological pathways in human growth. Nature Genetics 40: 584–591.

Lynch M, Walsh B. 1998. Genetics and Analysis of Quantitative Traits. Sinauer.

Nielsen R. 2005. Molecular signatures of natural selection. Annual Review of Genetics 39: 197–218.

Nielsen R, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biology 3: 976–985.

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

Oleksiak MF, Churchill GA, Crawford DL. 2002. Variation in gene expression within and among natural populations. Nature Genetics 32: 261–266.

Poustka AJ, Kuhn A, Groth D, Weise V, Yaguchi S, Burke RD, Herwig R, Lehrach H, Panopoulou G. 2007. A global view of gene expression in lithium and zinc treated sea urchin embryos: New components of gene regulatory networks. Genome Biology 8: R85.

Raff RA, Kaufman TC. 1983. Embryos, Genes, and Evolution: The Developmental-genetic Basis of Evolutionary Change. Macmillan.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nature Biotechnology 26: 1135–1145.

Su YH, Li E, Geiss GK, Longabaugh WJ, Kramer A, Davidson EH. 2009. A perturbation model of the gene regulatory network for oral and aboral ectoderm specification in the sea urchin embryo. Developmental Biology 329: 410–421.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63.

Weedon MN, et al. 2008. Genome-wide association analysis identifies 20 loci that influence adult height. Nature Genetics 40: 575–583.

Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. Nature Reviews Genetics 8: 206–216.

David A. Garfield (dag23@duke.edu) and Gregory A. Wray (gwray@duke.edu) are with the Department of Biology and Institute for Genome Sciences and Policy at Duke University, in Durham, North Carolina.