

5 Integrating Genomics into Evolutionary Theory

Gregory A. Wray

Many of the major advances in evolutionary biology have grown out of synthesis between disparate disciplines. Indeed, synthesis was present right from the beginning. Darwin was a consummate integrator of information: in formulating his theory of natural selection he drew key insights not only from the scientific literature and his own extensive observations of natural history, but also from geology and sociology. He also drew on medicine, plant and animal breeding, and the nascent fields of embryology and paleontology to provide material evidence in support of his ideas. Many decades later, the Modern Synthesis integrated fundamental advances from Mendelian and quantitative genetics into evolutionary thinking, while a more contemporary view of paleontology played a smaller but significant part of the integration. The Modern Synthesis is often considered the most pivotal era in the history of post-Darwinian thinking about evolution. Witness the title and topic of the present volume: the Modern Synthesis is the benchmark against which all other advances in evolutionary theory are measured (Pigliucci 2007).

A compelling case could be made, however, that information about the material basis for heredity has been just as transformative to evolutionary biology as the Modern Synthesis. Understanding the structure of DNA, the physical nature and variety of mutations, the molecular consequences of different kinds of mutations, and the mechanisms by which genes produce traits have all led to profound insights into evolutionary processes and mechanisms (Lynch 2007). Although the impact of molecular biology on evolutionary thinking was spread over several decades rather than concentrated into just a few years, the insights that have emerged from it are as profound as any that emerged from the Modern Synthesis. A prominent example is Kimura's Neutral Theory, which was motivated by empirical observations of genetic variation. The development of the Neutral Theory was not a natural extension of the Modern

Synthesis, and could not have happened in a pre-molecular era. Yet Kimura's ideas have utterly transformed how evolutionary biologists model and analyze evolutionary processes at a genetic level.

A century and a half after the publication of the *Origin of Species*, evolutionary biology is once again in a period of extraordinary integration and synthesis (Feder and Mitchell-Olds 2003; Rose and Oakley 2007; Pagel and Pomiankowski 2007). A quick perusal of the chapters in this book reveals that the impetus for this excitement is coming from several sources. Unquestionably, however, one of the most important of these is the availability of genome-scale data sets from many species and from many individuals within some species. As methods for gathering genomic data become more robust and as prices for doing so drop, consideration of these very large and rich data sets will become routine in every facet of evolutionary biology. The impact will be profound. In this chapter, I discuss some of the opportunities and challenges that the genomic era brings to evolutionary biology, and some of the ways current research into genome evolution is extending the Modern Synthesis.

Extending the Modern Synthesis to the Genome

Although genomics is one of the youngest branches of biology, two distinct phases in its history are already over. The first is the era when information was limited to genome sequences from just a handful of widely divergent species. The number of genome sequences is rising exponentially: prokaryotic genomes are being sequenced on a daily basis, and eukaryotic genome sequences appear almost weekly. Alignments of entire mammalian genomes and reconstructed ancestral genome sequences for internal nodes (Ma et al. 2006; Blanchette et al. 2004) signal the beginning of a new era in understanding how genomes evolve. Sequenced genomes from closely related species provide particularly appealing subjects for evolutionary analyses, and this information is now available for several clades (e.g., Kellis et al. 2003; Stark et al. 2007; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). It is possible to apply comparative methods in a serious way to sequences at the scale of tens of kb up to entire genomes (e.g., A. G. Clark et al. 2003; Doniger and Fay 2007; Hahn 2007), based on information that is accessible through the Web. As ultrahigh throughput sequencing technologies become more robust and affordable, it is becoming possible to generate whole-genome sequences and very large population samples of targeted regions at costs that a single lab group can contemplate.

The second phase in the history of genomics that is already over is the era when data consisted only of DNA sequences. Today, much of the excitement centers around functional data at a genome-wide scale. Microarrays, which measure mRNA levels for thousands of genes at once, were the first technology to provide data of this kind (Eisen et al. 1998). Evolutionary biologists began using microarrays as costs dropped, providing the first glimpses of evolutionary differences in gene expression throughout the genome (e.g., Oleksiak et al. 2002; Khaitovich et al. 2005). Ultrahigh-throughput DNA sequencing will largely supplant microarrays within the next few years, circumventing the need to design custom arrays for each species (a major impediment to comparative analyses) and providing the first comprehensive and unbiased sampling of gene expression. Several other kinds of genome-scale functional assays are making their way into evolutionary studies, including assays for alternate splicing (Calarco et al. 2007), binding sites for transcription factors (Moses et al. 2006; Odom et al. 2007), DNA methylation (Zhang et al. 2008), chromatin configuration (Babitt and Kim 2008), and microRNA binding sites (Wang et al. 2008). Although these remain expensive technologies to apply in a comparative context, prices continue to drop, and they will undoubtedly see increasing application by evolutionary biologists.

Do these genome-scale data sets offer anything to evolutionary biologists that analyses of single genes do not? After all, a gene-centric focus has proven hugely successful for more than half a century. The answer is in the affirmative for a couple of basic reasons. For one, there are interesting evolutionary phenomena that are apparent only at the scale of hundreds or thousands of genes. For instance, we would not know about the properties of very weak negative selection, including bias against certain codons (Akashi 1996; Nielsen et al. 2007) and spurious transcription start sites (Hahn et al. 2003; Froula and Francino 2007), were it not possible to examine sequences from whole genomes. Second, genome-scale data sets provide much more accurate and less biased information than any single gene, or even dozens of genes, can provide. Each gene is unique with respect to levels of variation and fixed differences, reflecting a distinct combination of the influences of negative selection to maintain the function of its product, any recent adaptive changes, the genomic region where it resides, and the presence of duplications nearby or far away (Gillespie 1992; Li 1997). Surveying thousands of genes provides a much clearer understanding of general trends by averaging across these distinct individual histories. Third, full genome

sequences make it much easier to study rare events, such as gene duplications and losses, transpositions and other large-scale rearrangements, changes in centromere positioning, and so forth. The vast majority of what we know about the origin and fate of mutations is based on the most abundant kind of mutation (single base substitutions) because they can be quantified within segments as tiny as a few kb or even less; in contrast, we know much less about the frequency with which rarer mutations arise and persist, the conditions that influence these processes, and their phenotypic and fitness consequences. At the scale of entire genomes, even rare events can be studied quantitatively. Fourth, comprehensive surveys of all genes allow one to screen for functional differences. For instance, microarrays measure which genes show the biggest differences in expression between species or populations or environments, a powerful complement to genetic association studies. Like quantitative genetics, this approach is unbiased and comprehensive, in the sense that it can query most or all genes rather than just a hand-picked set of genes; however, it is sometimes far cheaper than quantitative genetics and can be applied in many species for which breeding designs are not practical.

In studying multiple genome-scale data sets, evolutionary biologists are venturing into largely uncharted territory for both theory and data analysis. The opportunities are exciting, but the challenges are not trivial. To take just one example (but an important one): nearly all of population genetic theory is predicated on the assumption of minimal epistatic interaction among genes; in other words, the basic assumption is that one can ignore the rest of the genome when considering segregating variation in any given gene. Yet there is growing evidence that epistasis is pervasive (Gibson and Dworkin 2004). Genetic background often has a strong effect on the expressivity of a mutation, whether this is measured as an organismal trait or an intermediate phenotype such as gene expression (e.g., Brem and Kruglyak 2005). Effects of linkage are also generally ignored in studies of population genetics and molecular evolution. Another kind of interaction that is often ignored is gene duplication. One of the lessons from whole-genome sequencing has been the discovery that many genes are tandemly duplicated. The presence of a nearby paralog may affect patterns of nucleotide substitution by relieving functional constraint (e.g., Lynch et al. 2001; Hittinger and Carroll 2007).

The reality is that genes do not evolve in isolation, but rather in the context of the rest of the genome. Statistical methods for estimating the

magnitude of epistatic interactions at genomic scales are beginning to appear (Jannink and Jansen 2001; Yang et al. 2007; Pattin et al. 2008), but have not been widely applied. Multilocus models of selection that incorporate nonadditive contributions to fitness have been developed (Gavrilets and de Jong 1993; Beerenwinkel et al. 2007), but empirical applications again remain limited. Now that it is possible to obtain genotypes from hundreds or thousands of markers throughout the genome, it begins to be possible to apply these tests and models more widely.

The following sections discuss three extensions to the Modern Synthesis that are emerging out of the tumult and excitement of evolutionary genomics.

Updating Models of Genetic Information

The first extension from a genomics perspective is to update working models of genetic information. The dominant gene models that evolutionary biologists have used for decades are beginning to show their age. This is true of the gene models used in population genetics, molecular evolution, and evolutionary genetics—although for somewhat different reasons in each case. Progress in molecular biology has rendered the gene models used in all three of these areas of evolutionary biology outdated, and genomic data sets are pushing them past the breaking point. Each model is considered in turn below, along with some of the ways in which current research is providing fruitful extensions.

Population Genetics

The traditional gene model of population genetic theory is highly abstract. A gene is considered in isolation from the rest of the genome, on the assumption that the majority of genetic variation is additive. Gene-by-environment interactions are assumed to be minimal, and the effects of a mutation are assumed to be static even if new mutations arise or the environment changes. The abstract model also ignores information that is often available about specific genes, such as the function of its product and whether it belongs to a gene family. This is not to say that gene-by-environment interactions have never been studied, of course, nor that gene function is never taken into account (see, for example, Via and Lande 1985); the point is that such studies are the exception.

Simple models are convenient from a mathematical perspective, and abstraction was not only justifiable but necessary during the time of the Modern Synthesis when the physical nature of genes and mutations were

a complete mystery. But a simple, abstract gene model is becoming increasingly limiting in an era where we know an enormous amount about how genes function, how different kinds of mutations alter that function, how mutations can interact to influence traits, and how genes differ from one another in terms of function and trait associations.

Unfortunately, it is difficult to gauge how well a simple, abstract gene model performs relative to one that incorporates more information. Part of the problem is that we have only a dim idea of the general extent of nonadditive interactions among segregating variants and of gene-by-environment interactions. What is clear, however, is that these kinds of assumption-violating interactions are not rare. Gene-by-environment interactions are pervasive (reviewed in West-Eberhard 2003). Although less well documented, epistasis apparently is also widespread (Gibson and Dworkin 2004; Hermisson and Wagner 2004; Azevedo et al. 2006). Population geneticists have modeled epistasis (e.g., Wagner and Mezey 2000; Schlosser and Wagner 2008), and quantitative genetics provides powerful tools for measuring its effects (e.g., H. Li et al. 2007; Aylor and Zeng 2008), but most empirical studies focus on main effects and relatively few have attempted to uncover the general extent of epistasis. The largest relevant data sets come from studies of gene expression at a genome-wide scale (Rockman and Kruglyak 2006). Recent studies have demonstrated the evolutionary impact of epistasis across genomes (e.g., Brem and Kruglyak 2005; Cooper et al. 2008) and that ecologically relevant environmental variation can influence the transcription of a large proportion of genes (e.g., Idaghdour et al. 2008; Sambandan et al. 2008). The fitness consequences of these effects are rarely known, but many are hypothesized to be adaptive, including stress responses, immune system function, induced defenses, and various forms of phenotypic plasticity (West-Eberhard 2003; Pigliucci 2005).

Limitations in the abstract gene model become clear when information about a specific gene is available. Most of these concern the added predictive power that this information can bring. When modeling the likely fate of a mutation, it can be helpful to know, for instance, that a gene is the product of a recent duplication and might be under relaxed selective constraint, or that it lies within an inversion, and the resulting lack of recombination might influence its fate independent of fitness consequences. Likewise, it is useful to know something about the function of the gene's product: immune system components and testis-expressed genes, for example, generally evolve faster than most other

tion has occurred in a histone gene is much more useful for predicting its fate than the most sophisticated population genetic model that ignores gene function. Similarly, mutations in different positions within a gene can also have very different fates, because they are more or less likely to alter the function of the protein and thereby affect fitness. This effect is very clear, for instance, when comparing the fate of mutations within the active site of an enzyme or the DNA binding domain of a transcription factor with the rest of the gene.

Relevant information about specific genes and mutations can be incorporated into population genetic models. One area where population genetic theory has already developed explicit models to take advantage of ancillary information is to distinguish between genes that reside on autosomes from those that reside on sex chromosomes. These models generally predict subtle effects, but analyses of genome-scale data sets allow one to test their predictions (e.g., Nachman and Crowell 2000; Lu and Wu 2005). Whether a gene resides on an autosome or a sex chromosome is a qualitative datum, but many other relevant kinds of information are quantitative. Extending population genetic models to incorporate these other kinds of information will require parameterization.

Increasingly, the necessary information is available. Genomic sequences are at hand for many of the species commonly studied by evolutionary biologists, so that it is often possible to obtain information about gene copy number, gene product function, and chromosomal position (and much more information than this for the major model organisms). In clades where the genomes of multiple species have been sequenced or where variation has been surveyed genome-wide, it is also possible to derive a quantitative expectation of the likely fate of a mutation over longer time scales. Genome-wide analyses are beginning to reveal distinct patterns among functional classes of genes in terms of linkage disequilibrium, mutational spectrum, and population structure within species (e.g., Voigt et al. 2006; R. M. Clark et al. 2007), as well as in terms of sequence substitution rates between species (e.g., A. G. Clark et al. 2003; Haygood et al. 2007). Although these studies are primarily exploratory, they provide the basis for building an expectation about the fate of alleles in populations that are parameterized for individual genes, functional classes of genes, and nucleotide positions within genes.

Molecular Evolution

In contrast to population genetics, the gene models used in studies of molecular evolution have recognized differences among mutations from

the outset. The traditional molecular evolution gene model consists of a sequence of DNA that begins with an ATG codon, ends with one of the stop codons, and contains several more codons in between. The distinction between synonymous and nonsynonymous nucleotide substitutions provides a crude model for interpreting the fitness consequences of mutations, and the ratio of these two classes of mutations has been a staple of analyses for decades (Gillespie 1992; W.-H. Li 1997; Hartl and Clark 2006). But this is a rough and imprecise model: many, perhaps the majority, of nonsynonymous substitutions have no fitness consequence because they don't affect protein function, while some synonymous substitutions have fitness consequences because they alter splicing or codon usage (e.g., Kimchi-Sarfaty et al. 2007).

Other kinds of mutations are generally ignored. Insertions and deletions in multiples of three bases need not alter the reading frame, although they will if they fall across an intron-exon junction. In-frame indels are almost always eliminated from alignments prior to quantitative analysis, even though they are at least as likely to affect fitness as nonsynonymous substitutions. Three other kinds of mutations—indels that shift the reading frame, premature stop codons, and changes in the position of the start codon—can all have large functional consequences because they alter protein length. These kinds of mutations are usually assumed to result in loss of protein function, and therefore to carry a large negative fitness component. Comparisons within and across genomes reveal that this is clearly not always the case (e.g., Ng et al. 2008).

An even larger deficiency in the codon-based gene model is that it ignores a large fraction of mutations that affect gene function. Transcriptional initiation is regulated by sequences that lie almost entirely outside coding sequences; transcripts are spliced to remove introns in a sequence-dependent manner; many genes utilize alternative transcription start sites; 3' untranslated regions often contain sites that regulate message stability and trafficking; and additional noncoding sequences regulate chromatin configuration or encode microRNA molecules that regulate transcript turnover or translation (Lewin 2007; Latchman 2007).

Most of these various kinds of functional noncoding sequences have been studied by molecular biologists for decades. Regulatory sequences have received relatively little attention in studies of molecular evolution, but this is becoming increasingly difficult to justify (Chen and Rajewsky 2007; Wray 2007; Carroll 2008). Genome sequence comparisons between

species have estimated that a roughly comparable number of functionally constrained sites lie in noncoding and coding regions of eukaryotic genomes (Shabalina and Kondrashov 1999; Shabalina et al. 2001; Andolfatto 2005). In humans there may be more segregating mutations that alter transcriptional regulation than mutations affecting protein sequence (Rockman and Wray 2002), and positive selection on 5' noncoding regions (just one part of the regulatory landscape) was apparently at least as extensive as positive selection on all coding sequences during human origins (Haygood et al. 2007). Evidence is accumulating that the traditional focus on coding sequences misses out on a large fraction of mutations of adaptive significance within a genome. And this may be as much a qualitative as a quantitative blind spot: coding and noncoding mutations may contribute differentially to particular kinds of traits, such as morphology, reproduction, or immune function (Carroll 2008; Haygood et al. submitted).

Quantitative Genetics

The gene model of quantitative genetics is based simply on physical locations within the genome. The subjects of quantitative genetic study are appropriately called loci (positions) rather than genes, because they may or may not reside within a gene. Until quite recently, identifying the causal variants that define quantitative trait loci (QTL for short) has been quite difficult except in unusual cases. As a result, quantitative genetics has largely ignored the molecular consequences of causal mutations (biochemical consequence if coding, regulatory consequence if not), whether certain kinds of mutations are more likely to produce trait or fitness consequences than others, and how causal mutations actually alter organismal traits of interest.

Recently, however, the focus of quantitative genetics is shifting away from simply identifying QTL that are devoid of any functional context and toward a model of identifying genes with testable functional involvement and precise mutational bases that can be experimentally investigated. This has been made possible by the ability to carry out high-resolution mapping from large numbers of genetic markers (10^4 – 10^6) distributed across the genome, which in turn makes feasible the identification of causal mutations (also known as QTN, or quantitative trait nucleotides) in a growing number of cases. Knowing what kinds of genes and mutations contribute to adaptation in organismal traits adds a whole new dimension to evolutionary genetics, for instance, providing insights into how mutations produce trait differences and whether

parallel traits have parallel genetic bases (e.g., Shapiro et al. 2004; Prud'homme et al. 2006; Tishkoff et al. 2007).

Another extension is the use of scans for positive selection (described earlier), which provide a valuable complement to traditional quantitative genetic approaches by identifying genes that may be involved in adaptation throughout the genome. There are many cases where quantitative genetic approaches cannot be applied, for instance, when hybrids can't be made between species or generation times are too long. In such cases, genome-wide scans for positive selection provide one of the few comprehensive and unbiased approaches to identifying the genetic basis for trait differences among species.

The conventional gene models of population genetics, quantitative genetics, and molecular evolution all need to be, and are being, extended in order to accommodate new information. Much of the impetus comes from outside of evolutionary biology, primarily from advances in molecular biology but increasingly from genome-scale data sets. The inescapable fact is that we now live in a world where data sets are immensely larger than they were just a few years ago. Larger scales of data provide not only a more accurate, but also a more comprehensive, view of evolutionary processes than ever before. Importantly, this brings both new information and new challenges. On the one hand, larger data sets allow quantification of rare processes and the identification of spatial organization with the genome. On the other hand, leveraging these much larger data sets will require further extensions to both the theory and the analytical tools of evolutionary biology (Singh 2003; Lynch 2007; Pagel and Pomiński 2007).

Moving Beyond the Gene-in-a-Bubble Approach

The second extension to the Modern Synthesis that is emerging from genomics involves integrating approaches from different branches of evolutionary biology. It seems intuitively obvious that answering complex questions in evolutionary biology will sometimes require drawing on a combination of methods from phylogenetics, population genetics, and evolutionary genetics, molecular evolution, evolutionary ecology, and evolutionary developmental biology. In practice, however, most publications in evolutionary biology draw on the methods of just one of these areas.

Why is this so? Following the Modern Synthesis, evolutionary biology became increasingly gene-centric with attention focused on the gene as

the primary unit of selection (Hamilton 1963; Williams 1966; Dawkins 1976). About the same time, it became possible to sequence first proteins, and later DNA, technologies which utterly transformed evolutionary biology. And in so doing, this strongly reinforced the gene-centric perspective. The single gene became the prime unit of analysis for population genetics, and molecular evolution in particular. These two notably active and fertile areas of evolutionary biology during the last third of the twentieth century developed rather different gene models (see previous section).

But they shared one thing in common: the gene in their gene-centric models resided within a conceptual bubble, hermetically sealed away from environmental influences, regulatory processes, trait associations, and interactions with the rest of the genome. Empiricists worked with inbred lines, reared organisms under uniform environmental conditions, sequenced a single gene of interest, and focused on DNA to the exclusion of traits; theorists built models and developed statistical tests aligned with this gene-in-a-bubble approach. (In contrast, research in quantitative genetics often incorporated gene interactions and environmental influences.)

Isolating genes from the environment, the rest of the genome, and trait associations is a powerful approach for addressing certain kinds of questions, because it factors out "messiness" of various kinds. Treating genes in isolation also made sense in an era when relevant data about influences external to a single gene were sparse. But the gene-in-a-bubble approach is ultimately limiting. Real organisms live in environments that vary in space and time; alleles function in a diversity of genetic backgrounds; and genes affect traits that have real ecological consequences.

An important extension at this point is popping the bubble surrounding the single gene and placing it into a broader and more biologically realistic context. This requires examining a gene from the perspectives of population genetics, evolutionary ecology, quantitative genetics, evolutionary developmental biology, and more. Genomic data sets provide an impetus to break down the boundaries between these traditionally distinct disciplines by offering a common currency for analysis and modeling. Genome sequences, for example, facilitate the development of genetic markers for quantitative genetics, provide data for developing accurate background models of sequence evolution testing, produce data that allow testing for selection throughout the genome, furnish the information necessary for carrying out functional analyses such as microarrays, allow inference of changes in regulatory sequences such as enhancers

and microRNAs, and identify candidate genetic differences that may explain trait differences.

Integrating approaches among traditionally distinct disciplines within evolutionary biology can yield unique insights. Studies are beginning to appear that identify which genes are involved in the evolution of a particular trait, discern whether these genes have been under positive or balancing selection, reveal how changes in the function of these genes alter the trait of interest, and measure the impact of the trait consequences in the natural environment. Examples of studies that integrate some or all of these perspectives include the evolution of reduced armor in freshwater populations of three-spined sticklebacks (Shapiro et al. 2004, 2006), the evolution of wing and abdominal color pattern within the genus *Drosophila* (Gompel et al. 2005; Prud'homme et al. 2006; Jeong et al. 2008), and the evolution of lactose tolerance and malaria resistance during very recent human evolution (Hamblin and Di Rienzo 2000; Enattah et al. 2002; Tishkoff et al. 2007).

Extending Analyses across Scales of Genetic Organization

The third kind of extension of the Modern Synthesis inspired by genomics is applying both theoretical and analytical approaches to studying evolutionary processes across the full scale of genetic organization. Although the focus has been at the scale of a single gene for several decades, the lower and upper bounds of this scale are increasingly accessible to study by evolutionary biologists. At the smallest end of the scale are single mutations, while at the largest lie whole genomes. In between fall other less commonly discussed but important scales of genetic organization: haplotypes and genic partitions (exons, introns, 5' and 3' untranslated regions, and regulatory regions) lie between mutations and genes in scale, while gene networks and chromosomes occupy distinct organizational dimensions at a scale between genes and whole genomes.

The challenge is to extend both the theoretical and the analytical approaches of population genetics, quantitative genetics, and molecular evolution across this entire range of scales. Increasingly, data sets are appearing that either put a strain on traditional methods of evolutionary analysis or simply can't be analyzed within existing frameworks. This problem is particularly acute at the whole-genome scale. Very large data sets are often subjected to far more statistical comparisons than traditional evolutionary studies (often by several orders of magnitude).

Correcting for multiple comparisons is clearly important, but off-the-shelf methods designed for other purposes may be inappropriate because their assumptions are violated by the nature of the data. In addition, there is rarely a clear understanding of the distribution of trait measures, and the appropriate statistical test is not always obvious. Finally, missing information and variation in data quality are often a much larger problem with genomic data sets than in traditional evolutionary analyses. Visual checks on data quality are simply not possible because of scale, and statistical tests that assume complete data sets are often inappropriate.

Gene networks are a second important scale of genetic organization where it will be necessary to extend traditional methods of evolutionary analysis. Understanding how genetic variation affects network function, for instance, will require modeling networks to provide testable predictions, as well as a way to incorporate the structure of interactions among genes when analyzing results. Some existing modeling approaches (based, for instance, on Boolean, Bayesian, or algebraic topology methods) and statistical methods (e.g., path analysis) provide promising possibilities. A special challenge for association studies is how to incorporate information about known interactions within gene networks, as this violates assumptions of independence that are part of traditional approaches (Lynch and Walsh 1998); a few studies have developed and applied methods to do this (e.g., Walsh et al. 2008).

Not all the challenges in extending evolutionary theory and analysis to other scales of genetic organization are statistical: data representation and visualization present their own problems. No one wants to read a table that is 1,000 lines long (*E. coli*-sized), much less 20,000 lines long (human genome-sized). A common visualization tool is the "heat map," which represents a value for each gene, typically a test statistic or functional measure, as a color value. These displays can pack thousands of results into a small space, albeit at the cost of severe data reduction and the ability to read the names of individual genes. There are also challenges in building databases that provide ready access to results so that analyses can be conducted by other investigators.

Considerable progress has been made toward addressing these challenges. The analysis of genomic data sets is the primary area of research for increasing numbers of statisticians, mathematicians, and computer scientists. Some of their research has direct applications for evolutionary analyses, including false discovery rate to correct for multiple comparisons (Storey and Tibshirani 2003), hidden Markov models for modeling sequence evolution (Siepel and Haussler 2004), and appropriate statisti-

cal frameworks for association studies (Pritchard et al 2000). The transfer of expertise is not all in one direction: phylogenetic methods are routinely (but unfortunately not universally) applied in mainstream genomic studies to distinguish orthology from paralogy. Although phylogenetic inference is more computationally intensive than pattern-matching methods such as reciprocal best BLAST hits, it generally produces a lower error rate. On the data representation front, visual displays are gradually evolving beyond the heat maps that became almost ubiquitous following the invention of microarrays. Finally, databases have begun to move beyond simply serving as data repositories, and are setting standards for organizing and archiving information.

Evolutionary biology is transitioning from an era of data limitation to one of data abundance, and even superabundance, in a limited but growing number of areas. Put simply, the challenge is shifting away from how to gather data and toward how to analyze, integrate, and make sense of very large data sets (Singh 2003).

Summary and Prospects

A century and a half after the publication of the *Origin of Species*, evolutionary biology is entering a time of extraordinary expansion. The foundation that Darwin built was tested and vastly strengthened, initially with the elucidation of transmission genetics and later with the nature of the genetic material. Today we are in the midst of another exciting, and potentially equally transformative, period in the history of evolutionary biology. Information from molecular biology, developmental biology, and, most recently, genomics is prompting substantial changes to the gene-centric view that emerged during and shortly after the Modern Synthesis. I have argued that three specific extensions are under way already. First, enormous advances in understanding how genes function and how they work together to produce developmental and physiological processes are prompting substantial and highly informative updates to the gene models used in evolutionary studies. Second, technological advances now allow us to study genes in the context of the rest of the genome and the environment rather than as isolated entities, revealing much about gene interactions, phenotypic plasticity, and developmental roles. And third, expanding analyses beyond genes as the focal unit is allowing researchers to study the evolution of the hereditary material at a wide range of scales, from single mutations through gene networks to

The advent of genome-scale data sets is prompting a wide range of exciting new studies. The results are revealing evolutionary phenomena of which we were previously unaware, such as codon bias and biases in the distribution of positive selection on coding and noncoding sequences depending on gene function. Genomic data are also being used to evaluate predictions that were formerly untestable, such as genome-wide estimates of epistasis and the long-term fate of gene duplications. Technological advances are providing entirely new ways to identify the genetic basis for trait evolution through whole-genome association studies and genome-scale functional assays such as microarrays. Applying the traditional approaches of population genetics, evolutionary genetics, and molecular evolution to genomic data sets poses nontrivial challenges. Balanced against these challenges, however, are extraordinary opportunities to better understand evolutionary processes and mechanisms. Theorists and quantitative biologists, in particular, are entering a period of exceptional opportunity as data sets expand in scale, scope, and kind.

It is too soon to know how extensively and in what ways genomic data will influence our understanding of evolution. But one point is already clear: these data are building upon and extending the robust framework of the Modern Synthesis in ways that we could not have imagined even a decade ago.

Acknowledgments

David Garfield and Jenny Tung provided many helpful comments. My research is supported by grants from the National Science Foundation and the National Institutes of Health.

References

- Akashi H (1996) Molecular evolution between *Drosophila melanogaster* and *D. simulans*: Reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* 144: 1297–1307.
- Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437: 1149–1152.
- Aylor DL, Zeng ZB (2008) From classical genetics to quantitative genetics to systems biology: Modeling epistasis. *PLoS Genetics* 4: 1000029.
- Azevedo L, Suriano G, van Asch B, Harding RM, Amorim A (2006) Epistatic interactions: How strong in disease and evolution? *Trends in Genetics* 22: 581–585.
- Babbitt GA, Kim Y (2008) Inferring natural selection on fine-scale chromatin organization in yeast. *Molecular Biology and Evolution* 25: 1714–1727.

- Beerenwinkel N, Pachter L, Sturmfeis B, Elena SF, Lenski RE (2007) Analysis of epistatic interactions and fitness landscapes using a new geometric approach. *BMC Evolutionary Biology* 7: 60.
- Blanchette M, Kent WJ, Riemer C, Elmski L, Smit AFA, Roskin KM, Baertisch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* 14: 708–715.
- Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 genes: expression traits in yeast. *Proceedings of the National Academy of Sciences of the USA* 102: 1572–1577.
- Calarco JA, Saltzman AL, Ip JY, Blencowe B (2007) Technologies for the global discovery and analysis of alternative splicing. *Advances in Experimental Medicine and Biology* 623: 64–84.
- Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* 134: 25–36.
- Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics* 8: 93–103.
- Clark AG, Gnanowski S, Nielsen R, Thomas PD, Kejarawal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferrera S, Wang G, Zheng XG, White TJ, Sainsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302: 1960–1963.
- Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen HM, Frazer KA, Huson DH, Schölkopf B, Nordborg M, Raetsch G, Ecker JR, Weigel D (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317: 338–342.
- Cooper TF, Remold SK, Lenski RE, Schneider D (2008) Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in *Escherichia coli*. *PLoS Genetics* 4: e35.
- Dawkins R (1976) *The Selfish Gene*. New York: Oxford University Press.
- Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Computational Biology* 3: 932–942.
- Dworkin I, Gibson G (2006) Epidermal growth factor receptor and transforming growth factor- β signaling contributes to variation for wing shape in *Drosophila melanogaster*. *Genetics* 173: 1417–1431.
- Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA* 95: 14863–14868.
- Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I (2002) Identification of a variant associated with adult-type hypolactasia. *Nature Genetics* 30: 233–237.
- Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics* 4: 649–655.
- Froula JL, Francino MP (2007) Selection against spurious promoter motifs correlates with translational efficiency across bacteria. *PLoS ONE* 2: e745.
- Gavrilets S, de Jong G (1993) Pleiotropic models of polygenic variation, stabilizing selection, and epistasis. *Genetics* 134: 609–625.
- Gibson G, Dworkin I (2004) Uncovering cryptic genetic variation. *Nature Reviews Genetics* 5: 681–691.
- Gillespie JH (1992) *The Causes of Molecular Evolution*. New York: Oxford University Press.
- 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.

- Hahn MW (2007) Detecting natural selection on cis-regulatory DNA. *Genetica* 129: 7–18.
- Hahn MW, Stajich JE, Wray GA (2003) The effects of selection against spurious transcription factor binding sites. *Molecular Biology and Evolution* 20: 901–906.
- Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. *American Journal of Human Genetics* 66: 1669–1679.
- Hamilton WD (1963) Evolution of altruistic behavior. *American Naturalist* 97: 354–356.
- Hartl DL, Clark AG (2006) *Principles of Population Genetics*. 4th rev. ed. Sunderland, MA: Sinauer.
- Haygood R, Babbitt CC, Fedrigo O, Wray GA (submitted) Positive selection in the human genome exhibits strong contrasts between coding and noncoding sequences.
- Haygood RH, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA (2007) Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature Genetics* 39: 1140–1144.
- Hermisson J, Wagner GP (2004) The population genetic theory of hidden variation and genetic robustness. *Genetics* 168: 2271–2284.
- Hittinger CT, Carroll SB (2007) Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* 449: 677–681.
- Idaghdour Y, Storey JD, Iadallah SI, Gibson G (2008) A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan amazighs. *PLoS Genetics* 4: e1000052.
- Jannink J-L, Jansen R (2001) Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* 157: 445–454.
- Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB (2008) The evolution of gene regulation underlies a morphological difference between two *Drosophila* sister species. *Cell* 132: 783–793.
- Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241–254.
- Khairatovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Pääbo S (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309: 1850–1854.
- Kimchi-Sarfaty C, Oh JM, Kim JW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315: 525–528.
- Latchman DS (2007) *Eukaryotic Transcription Factors*. 5th ed. Burlington, MA: Academic Press.
- Lewin B (2007) *Genes IX*. Boston: Jones and Bartlett.
- Li H, Gao G, Li J, Page GP, Zhang K (2007) Detecting epistatic interactions contributing to human gene expression using the CEPH family data. *BMC Proceedings* suppl. 1: S67.
- Li W-H (1997) *Molecular Evolution*. Sunderland, MA: Sinauer.
- Liu Y, Duan W, Paschall J, Saccone NL (2007) Artificial neural networks for linkage analysis of quantitative gene expression phenotypes and evaluation of gene x gene interactions. *BMC Proceedings* suppl. 1: S47.
- Lu J, Wu CI (2005) Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proceedings of the National Academy of Sciences of the USA* 102: 4063–4067.
- Lynch M (2007) *Origins of Genome Architecture*. Sunderland, MA: Sinauer.
- Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* 159: 1789–1804.

- Lynch M, Walsh B (1998) Genetics and Analysis of Quantitative Traits. Sunderland, MA: Sinauer.
- Ma J, Zhang LX, Suh BB, Raney BJ, Burhans RC, Kent WJ, Blanchette M, Haussler D, Miller W (2006) Reconstructing contiguous regions of an ancestral genome. *Genome Research* 16:1557-1565.
- Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB (2006) Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLoS Computational Biology* 2:1219-1231.
- Nachman MW, Bauer VT, Crowell SL, Aquadro CF (1998) DNA variability and recombination rates at X-linked loci in humans. *Genetics* 150:1133-1141.
- Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156:297-304.
- Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busan DA, Strausberg RL, Venter JC (2008) Genetic variation in an individual human exome. *PLoS Genetics* 4:e1000160.
- Nielsen R, Dumont YLB, Hubisz MJ, Aquadro CF (2007) Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Molecular Biology and Evolution* 24:228-235.
- 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.
- Pagel M, Romiankowski A, eds (2007) Evolutionary Genomics and Proteomics. Sunderland, MA: Sinauer.
- Patin KA, White BC, Barney N, Gui J, Nelson HH, Kelsey KT, Andrew AS, Karagas MR, Moore JH (2008) A computationally efficient hypothesis testing method for epistasis using multifactor dimensionality reduction. *Genetic Epidemiology* 33:87-94.
- Pigliucci M (2005) Evolution of phenotypic plasticity: Where are we going now? *Trends in Ecology and Evolution* 20:481-486.
- Pigliucci M (2007) Do we need an extended evolutionary synthesis? *Evolution* 61:2743-2749.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.
- Prud'homme B, Carroll SB (2006) Monkey see, monkey do. *Nature Genetics* 38:740-741.
- Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD, True JR, Carroll SB (2006) Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* 440:1050-1053.
- Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222-234.
- Rockman MV, Kruglyak L (2006) Genetics of global gene expression. *Nature Reviews Genetics* 7:862-872.
- Rockman MV, Wray GA (2002) Abundant raw material for cis-regulatory evolution in humans. *Molecular Biology and Evolution* 19:1991-2004.
- Ronald J, Brem RB, Whittle J, Kruglyak L (2005) Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genetics* 1:213-222.
- Rose MR, Oakley TH (2007) The new biology: Beyond the Modern Synthesis. *Biology Direct* 2:30.
- Sambandan D, Carbone MA, Anholt RRH, Mackay TEC (2008) Phenotypic plasticity an genotype by environment interaction for olfactory behavior in *Drosophila melanogaster*. *Genetics* 179:1079-1088.
- Schlösser G, Wagner GP (2008) A simple model of co-evolutionary dynamics caused by epistatic selection. *Journal of Theoretical Biology* 250:48-65.
- Shabalin SA, Kondrashov AS (1999) Pattern of selective constraint in *Celegans* and *C. briggsae* genomes. *Genetical Research* 74:23-30.
- Shabalina SA, Ourgutov AY, Kondrashov VA, Kondrashov AS (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends in Genetics* 17:373-376.
- Shapiro MD, Bell MA, Kingsley DM (2006) Parallel genetic origins of pelvic reduction in vertebrates. *Proceedings of the National Academy of Sciences of the USA* 102:13753-13758.
- Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter L, Kingsley DM (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 428:717-723.
- Siipola A, Haussler D (2004) Combining phylogenetic and hidden Markov models in biologic sequence analysis. *Journal of Computational Biology* 11:413-428.
- Singh RS (2003) Darwin to DNA: molecules to morphology: The end of classical population genetics and the road ahead. *Genome* 46:938-942.
- Stark A, Liu MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, Ruby JG, Brennecke J, Hodges E, Hinrichs AS, Caspi A, Park SW, Han MY, Maeder ML, Polansky BJ, Robson BE, Aerts S, van Helden J, Hassan B, Gilbert DG, Eastman S, Pritchard JK, Wray GA, Deloukas P (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* 39:31-40.
- Via S, Lande R (1985) Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution* 39:505-522.
- Voight BF, Kudaravalli S, Wen XQ, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biology* 4:659-659.
- Wagner GP, Mezey J (2000) Modeling the evolution of genetic architecture: A continuum of alleles model with pairwise A x A epistasis. *Journal of Theoretical Biology* 203:163-175.
- Wash T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Mahotra D, Bhandari A, Stray SM, Rippey CF, Rocanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Melzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320:539-543.
- Wang XW, Gu J, Zhang MQ, Li YD (2008) Identification of phylogenetically conserved microRNA cis-regulatory elements across 12 *Drosophila* species. *Bioinformatics* 24:165-171.
- West-Eberhard MJ (2003) *Developmental Plasticity and Evolution*. New York: Oxford University Press.

- Williams GC (1966) Natural selection, the costs of reproduction, and a refinement of Lacks' Principle. *American Naturalist* 100: 687–690.
- Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* 8: 206–216.
- Yang J, Zhu J, Williams RW (2007) Mapping the genetic architecture of complex traits in experimental populations. *Bioinformatics* 23: 1527–1536.
- Zhang X, Shiu SH, Cai A, Borevitz JO (2008) Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. *PLoS Genetics* 4: e1000032.

6 Complexities in Genome Structure and Evolution

Michael Purugganan

In 1977, the first genome was sequenced—the viral species ϕ X174, a mere 5.4 kb in length—and this sequencing was a landmark in biology (Sanger, Nicklen, and Coulson 1977). It was 18 years later, in 1995, when the genome of a living organism was first announced, the 1.8 Mb *Haemophilus influenzae* bacterial genome (Fleischmann et al. 1995), and in 2001, we crossed another major milestone when it was announced that the human genome sequence—3 Gb in length—had been completed by two groups (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). Today, whole-genome sequencing projects proliferate ever faster; as of early 2008, more than 180 genome sequences had been completed across all major kingdoms, and genome sequencing technology has advanced to the point that (depending on the size of the organism's genome) single investigator laboratories can contemplate sequencing the genome of their favorite organism. And the frenzy is not confined to model species; in early 2008, the 2.2 Gb platypus genome was completed (Warren et al. 2008), as was the 372 Mb papaya genome (Ming et al. 2008).

The advent of genomics also made possible the study of comparative genomics, possible, which, taken in an evolutionary framework, allows us to examine the diversification of genome structure and the function of their component genes. Genomics has provided enabling technologies for the evolutionary sciences, producing large-scale information about genome structure and function across multiple species. This has resulted in expanding knowledge of the complete inventory of genes found in organismal genomes, and in the process has begun to bring to light several issues that continue to excite the interests of molecular evolutionists. In this chapter, I will discuss four issues that modern evolutionary biology has either learned or needs to grapple with in the age of genomics. These topics are only a handful of the myriad opportunities

