Commentary Eight

The Expansion of Molecular Data in Evolutionary Biology

Joshua S. Rest

Near the end of his voyage on the *Beagle,* Darwin chose to study barnacles because of their peculiar characteristics and because of his instinct that their geological distribution held important insights (Love 2002). Darwin had a fantastic range of potential study subjects that he might have instead chosen to pursue. His observations spanned great biological variety, from howler monkeys, butterflies, parasitic orchids, and toucans to ceiba trees. He probably felt overwhelmed by the conflicting desires to travel and gather more data, versus the need to analyze, theorize, test, and write about his corpus of observations. When Darwin discovered barnacles possessing unexpected features, it may have seemed logical to collaborate and send them to a cirripedologist so that he might instead focus his efforts on the species problem. But there was no such specialist, so he devoted 8 years to becoming a barnacle specialist (Stott 2003). Eight years of meticulous dissection, classification, and correspondence with a network of other naturalists passed, while his essay about the origin of species sat locked in a drawer and stewed in his mind.

Today, the amount of molecular data available to evolutionary biology may seem as overwhelming as the data that Darwin had collected. Like the new organisms and habitats that Darwin observed on the *Beagle*, molecular data offers a seemingly unlimited source of characters for study. However, a single biologist cannot hope to master the diverse types of data now available. A cursory list of recently developed molecular data types includes next-generation sequencing based on emulsion PCR or bridge PCR, pyrosequencing, and chromatin immunoprecipitation followed by sequencing to assay acetylation, methylation, nucleosome positioning, or DNA–protein interactions. GenBank, the repository of sequencing data, has grown exponentially since its inception (Benson et al. 2009), and as of April 2010, contained over 279 billion base pairs. Other types of data have also grown rapidly; for example, the number of known protein–protein interactions in yeast has increased exponentially over time (He and Zhang

2009). The number of studies, the number of interactions detected per study, and the number of authors per study have all commensurately increased. The quantity of data will continue to grow, as it is now proposed to sequence 10,000 vertebrate genomes (Genome 10k 2009), and a systematic cataloging of genome-wide epigenetic traits in worms and flies is well underway (Celniker et al. 2009). This enormous increase in data production and analysis has been catalyzed by, and indeed relies upon, the escalation in processing power described by Moore's law and its corollaries related to data storage and memory (Bell et al. 2009).

All of these data types hold great promise for the future of evolutionary biology. Perhaps the most immediate benefit has been the quantity and quality of allelic polymorphism data from increased sequencing coverage of individuals, populations, and genes, allowing high-resolution determination of population structures (see Zhang, Chapter 4; Kolaczkowski and Kern, Chapter 6). Several of the high-throughput sequencing approaches are being used to identify potentially adaptive variation in wild populations and non-model organisms (Ellegren 2008; Hudson 2008). RNA sequencing allows identification of variation that is focused on genes and gene transcript levels in non-model organisms, because a reference genome is not required (Novaes et al. 2008). Sequencing technology is also used to assess genome-wide organizational, epigenetic, and protein-binding characteristics of DNA. Although such epigenetic and protein-binding analysis is currently practical only in model organisms, the number of species that can be considered to be models continues to increase, and the data generated in these model organisms can be mapped onto the genomes of non-model organisms to assess evolution of these characters. Using this approach across fungal species, for example, transcription factor binding sites involved in ribosomal regulation were shown to have been gained and lost via an intermediate stage during which redundant elements were present (Tanay et al. 2005). This approach has also been used to study the evolution of metabolic- and protein-interaction networks.

These molecular approaches are clearly powerful for evolutionary biologists, but an individual who wants to master a particular area outside her or his field usually cannot afford to spend 8 years becoming an expert in a new discipline. A first solution, of course, is to collaborate with a specialist. The degree of expertise available is demonstrated by the dramatic increase in both scientists and peer-reviewed journals: the number of journals has increased from only hundreds at the end of the nineteenth century to over 10,000 today (Mabe and Amin 2001). It is not hard to recall discoveries in which close collaboration between molecular and evolutionary biologists played a role. One recent example is the result from large-scale sequencing efforts that showed widespread purifying and positive selection acting on the *Drosophila* genome. Nearly two-thirds of mutations identified in noncoding regions are deleterious, and up to half of the amino acid substitutions and one-in-five noncoding substitutions are adaptive (Andolfatto 2005; Begun 2007). These observations challenge the prominence of the neutral theory for describing evolutionary patterns in *Drosophila*. In contrast, one can also think of scientific questions that have remained refractory and situations in which conflict between specializations or disciplines has not helped matters. For example, phylogenetic relationships and the timing of divergences within Neoaves have defied resolution, despite increasing amounts of sequence data, because there is conflict between molecular and paleontological data (e.g., Brown et al. 2008). A lack of interaction between molecular

evolutionists and paleontologists is partly to blame; for example, these two communities often interpret the meaning of paleontological dates differently (Brochu et al. 2004). It is clear that scientists in different fields will have to collaborate to find creative solutions to these difficult problems.

A second solution for evolutionary biologists who want to take advantage of publicly available molecular data is to become familiar with a few of the most essential computational tools. As the number of complete genomes, transcriptomes, and associated data increases, so do the analytical tools and aptitude required for analysis. A decade or two ago, with the advent of phylogenetics, many evolutionary biologists became amateur computational biologists. Today, the genomic era requires a high level of analytical sophistication. In particular, large-scale collection of data comes along with a commensurate proportion of errors and requires careful statistical analysis to account for multiple comparisons. Storing, parsing, tabulating, and visualizing genome-scale datasets can be a significant technical challenge (Bell et al. 2009), and the data often require substantial refinement before comparative analysis can even be considered. Novel data sets and hard problems require dedicated research by experts in statistics, bioinformatics, and computer science. Through time, these tools will be integrated into packages for more general use. At present, processing, statistical analysis, and visualization of biological and genomic data are being centralized in the R software environment, and learning this language is currently an efficient way for all biologists to access shared libraries of data-oriented tools. R and its associated libraries are especially useful for visualizing large and complex datasets in ways that can lead to meaningful understanding of the nature of the data.

As magical as this data revolution may seem, however, evolutionary biology will be ill served by merely increasing the amount of data collected. Let us not forget that Linnaeus and Lamarck had gathered copious observations on natural history, but only Darwin actually derived a sound scientific explanation for the data. Even in the jungle of molecular genomics data, careful thinking, good hypotheses, and broad knowledge of organisms will remain central to actual discovery. Data collection and analytical details are only important to the extent that they complement these skills. In this way, how evolutionary biologists *ask* questions will remain fundamentally the same.

Literature Cited

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437: 1149–1152.

Begun, D. J., A. K. Holloway, K. Stevens, and 10 others. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* 5: e310.

Bell, G., T. Hey, and A. Szalay. 2009. Beyond the data deluge. *Science* 323: 1297–1298.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, and 2 others. 2009. Genbank. *Nucl. Acids Res.* 37: D26–31.

Brochu, C. A., C. D. Sumrall, and J. M. Theodor. 2004. When clocks (and communities) collide: Estimating divergence time from molecules and the fossil record. *J. Paleontol.* 78: 1–6.

Brown, J. W., J. S. Rest, J. Garcìa-Moreno, and 2 others. 2008. Strong mitochondrial DNA support for a cretaceous origin of modern avian lineages. *BMC Biol.* 6: 6.

Celniker, S. E., L. A. L. Dillon, M. B. Gerstein, and 15 others. 2009. Unlocking the secrets of the genome. *Nature* 459: 927–930.

Ellegren, H. 2008. Sequencing goes 454 and takes large-scale genomics into the wild. *Mol. Ecol.* 17: 1629–1631.

Genome 10k Community of Scientists. 2009. Genome 10k: A proposal to obtain whole-

genome sequence for 10,000 vertebrate species. *J. Hered.* 100: 659–674.

He, X. and J. Zhang. 2009. On the growth of scientific knowledge: Yeast biology as a case study. *PLoS Comp. Biol.* 5: e1000320.

Hudson, M. E. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol. Ecol. Resour.* 8: 3–17.

Love, A. C. 2002. Darwin and *Cirripedia* prior to 1846: Exploring the origins of the barnacle research. *J. Hist. Biol.* 35: 251–289.

Mabe, M. and M. Amin. 2001. Growth dynamics of scholarly and scientific journals. *Scientometrics* 51: 147–162.

Novaes, E., D. R. Drost, W. G. Farmerie, and 4 others. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9: 312.

Stott, R. 2003. *Darwin and the Barnacle.* Faber and Faber, London.

Tanay, A., A. Regev, and R. Shamir. 2005. Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. *Proc. Natl. Acad. Sci. USA* 102: 7203–7208.