A Data Torrent Raises Questions about Research Strategy

Never before have scientists been able to generate so many data about the genomes of an organism in such little time,” state Jonathan H. Stillman and Eric Armstrong in their valuable Overview article on how genomics are transforming our understanding of responses to climate change (this issue). Indeed, rapid advances in next-generation DNA sequencing (NGS) techniques are generating data at a staggering rate, such that even the venerable GenBank, with its billions of archived sequences, is unable to store them. Although NGS machines are expensive, their use enables the production of genomic data quickly and at low cost (per base) from fresh, frozen, ethanol-preserved, and archival museum specimens, as well as from degraded DNA, as found in, for example, frozen mammoths and Neanderthal remains. Computer-intensive bioinformatics techniques allow the assembly of long stretches of sequence from the many short stretches output by the machines.

The techniques, first used in biomedical studies and agribusiness, are now providing important information about evolution and how local populations respond to ecological changes, including climate change, in a wide variety of species. NGS allows the examination of expressed genes and epigenetically modified DNA as well as general nuclear DNA. It also allows the identification of thousands of genetic markers, including single-nucleotide polymorphisms, thus making it easier to identify loci of particular interest. Stillman and Armstrong detail several examples of how such investigations have apparently confirmed—or, no less impressive, disconfirmed—ideas about the basis of population responses to climate change. Yet it would be a mistake to think that ecological genetics has reached the promised land. Studies using NGS data are much easier for species that have reference genomes, of which there are still only a few hundred for eukaryotes. And like the proverbial drunkard who searches for lost keys under the lamppost—because that is where the light is—researchers could be misled by the glee of having so many data. It seems likely that most causal mechanisms of adaptation and evolution lie still in the dark beyond, a case expertly laid out by Matthew V. Rockman in Evolution (http://io.aibs.org/rockman).

In a nutshell: Most convenient approaches are likely to discover relatively large genetic effects on phenotypes rather than small ones—but there is no reason to believe that the large ones are typical or even the most important in evolution or adaptation. Publication bias and trait selection can exacerbate the problem. Data from large medical studies are discouraging: It seems important traits are affected by huge numbers of loci. More than (possibly many more than) 7000 single-nucleotide polymorphisms predict the occurrence of schizophrenia. The 10 strongest associations with blood pressure found in a genome-wide study jointly explain 1 percent of the variation. Pervasive polygenic effects might explain why, for example, there are surprisingly few known nucleotides affecting wing shape in the very well-studied Drosophila.

The challenges emphasize, rather than deny, the importance of the techniques Stillman and Armstrong discuss. Progress is surely possible, and NGS will be vital. But optimal research strategies demand serious thought. What glitters in an initial study (with apologies to Rockman) may not be gold.

TIMOTHY M. BEARDSLEY
Editor in Chief