nature genetics

Sequence specificity

The technology of DNA sequencing has come a long way since the days of slab gels. Graduate students and postdocs no longer crouch over light boxes to make a call on each nucleotide. The use of fluorescence detection in automated sequencing, along with appropriate metrics and algorithms to ensure sequence quality and proper assembly, put whole-genome sequences within the reach of large public and private efforts. Two new approaches to sequencing promise to wrest these projects from such large-scale sequencing centers and, eventually, to make them part of the routine work of individual labs and clinics. The company 454 Life Sciences has reported a 'sequencing-by-synthesis' approach that has been used to sequence the *Mycoplasma genitalium* genome (25 Mb) in a stunning four hours (*Nature*, advance online publication 31 July 2005; doi:10.1038/nature03959). At the same time, a group led by George Church has established the use of epifluorescence microscopy in automated sequencing, achieving quality high-throughput reads at a fraction of the usual cost (*Science*, published online 4 August 2005; doi:10.1126/science1117389). Hamish Scott provides additional context in his meeting report on page 1019 of this issue.

Given the continued investment, both public and private, in improved sequencing technology, it seems likely that, at some point, cost and labor will no longer be barriers to even the most extravagant sequencing experiment. This is in some ways a reversal of the state of the field when the Human Genome Project was in its planning stages. At the time, the available methods were easily outpaced by the imaginations of many geneticists, who dreamed of the things they could do with a database full of whole-genome sequences. No longer. The current sequencing capacity of the genomics community, though still productively engaged in the thriving field of comparative genomics, may soon be in need of some new business. And the new technologies may for a time outstrip even the most fertile imaginations. When asked about these 'sequencing revolutions' in an interview for *The Scientist*, Jonathan Eisen commented, "People aren't even thinking what they can do with it yet." Aravinda Chakravarti adds, "Huge reams of data are nice but what we all want are crisp, reliable and, most importantly, exciting and fun inferences from the data. There are still too few scientists who know enough about both biology and mathematics to make this goal a reliable certainty."

These new advances are only beginning to be implemented, and even 'old-fashioned' ABI capillary sequencers may find new, more targeted uses outside the ongoing whole-genome sequencing programs. As an example of this smaller-scale approach, the National Human Genome Research Institute (NHGRI) recently posted a request for information (RFI) on a potential new program for the identification of mendelian disease genes by genomic sequencing (http://grants.nih.gov/grants/guide/notice-files/NOT-HG-05-006.html). According to Adam Felsenfeld, Program Director at the NHGRI for large-scale sequencing, the institute's existing sequencing portfolio is a broad one. It encompasses the collaboration with the National Cancer Institute to sequence cancer genomes, a thriving effort in comparative and evolutionary genomics centered around invertebrate genomes, and of course an ongoing effort to annotate the human genome, most notably by the ENCODE consortium. The working group for this new medical sequencing program, led by David Altshuler, Les Biesecker and Jeffrey Bodkin, has been charged with making recommendations for programs that would assist individual investigators who are working on projects of "significant and direct medical interest." Although many proposals will be considered—the NHGRI has approximately 12 billion bases of sequencing capacity per month at its disposal—the most promising one at present is directed at mendelian disease genes.

What is the rationale behind the RFI? The current statistics in OMIM suggest that there are still potentially thousands of mendelian disorders whose molecular basis is unknown. According to Biesecker, there is a perception that many efforts to map and positionally clone the genes underlying these disorders are hung up at the fine mapping stage, either because the interval is too large for an investigator to sequence or because the causative lesion is in a regulatory element some distance from the candidate gene. He emphasizes that this information is very difficult to come by, as National Institutes of Health project databases typically summarize the initial plans for each funded project but not their progress (or lack thereof) toward completion. Felsenfeld estimates that such directed sequencing might be usefully applied to 50–100 of these projects, although the working group is relying on the community's response to the RFI to come up with a firm number. Comment will be accepted until 4 November. There will also be an NHGRI-sponsored workshop on the proposal at the upcoming meeting of the American Society of Human Genetics on 28 October.

Ultimately it will be up to the NHGRI's council to decide whether to go forward with this program, which Felsenfeld estimates will cost approximately $25,000–50,000 for each project. The four papers identifying mendelian disease genes in this issue of *Nature Genetics* all seem to have proceeded in a straightforward manner, but many other gene identifications might be hastened by the availability of targeted medical sequencing.

It's also clear from the RFI, and from conversations with Felsenfeld and Biesecker, that this is merely a first step. Conceptually, the genetics of human mendelian disease is a *fait accompli*. The bigger challenge is the complex trait, and the strategic use of sequencing capacity should be a useful adjunct to the marker-rich whole-genome association studies that are now in the pipeline. The successful use of a central resource for targeted sequencing in mendelian disease would nonetheless set an important precedent for its use in association studies. It would also strengthen the case for a creative, hypothesis-driven use of this most powerful technology, even as technical advances will generate more data, more easily, than ever before. ■