DOI:
10.1038/nrgXXXX

MILESTONE 20

Putting it all together

You might remember this problem from your childhood: when you lose the top to your puzzle box, you are confronted with lots of pieces and no idea what they are supposed to look like when assembled. Genome sequencers faced the same dilemma when beginning large-scale DNA sequencing. They did the same thing that you might: they started at known landmarks and systematically built up the larger picture.

In order to assemble short stretches of DNA sequence from each read into a larger whole, particularly on a large scale, bioinformaticists developed algorithms that could take input directly from fluorescent sequencing machines. The earliest programs to achieve wide use were called Phred, Phrap and Consed, developed by Phil Green and colleagues. Phred initially went through the sequence reads and assigned a 'base call' to the chromatogram output from the machine. Phrap then assembled the list of bases from multiple reads into the most likely single path through the sequence. Users then viewed and edited the output with Consed, to generate higher-quality sequences as required. These programs were developed for, and used on, the public Human Genome Project.

Gene Myers and colleagues later developed an algorithm that used the end-pair information from sequencing subclones and could assemble larger sequences. They postulated that the whole genome could be cut into pieces, sequenced randomly and reconstructed given sufficient computational power. They demonstrated this approach on the genome of *Drosophila melanogaster* and famously went on to 'race' the publicly-funded Human Genome Project using, in the end, a combination of their whole-genome assembly methods and data from the public project. However, so-called shotgun whole-genome assemblies are now the method of choice for large genome projects, and the field has moved on to next-generation programs like Arachne, Atlas and PCAP, each using different algorithms.

Chris Gunter, Senior Editor, Nature

ORIGINAL RESEARCH PAPERS Ewing, B. & Green, P. Basecalling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998) | Ewing, B., Hillier, L., Wendl, M. & Green, P. Basecalling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998) | Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998) | Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 868–877 (2000)

FURTHER READING International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001) | Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001) | She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004)

WEB SITES

National Center for Biotechnology Information assembly information: http://www.ncbi.nlm.nih.gov/genome/guide/Assembly/Assembly.shtml

Phrap: http://www.phrap.org