Piecing it all together

The whole-genome shotgun method has assembled a high-quality draft mouse sequence. Future projects will wed the shotgun's speed and economy to established, map-based methods, says Declan Butler.

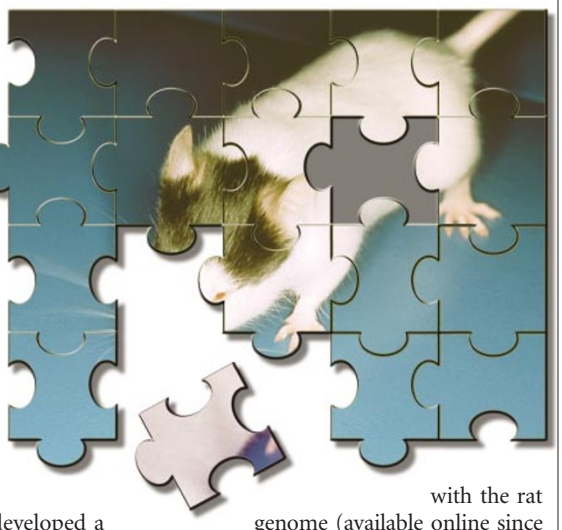
"Tossed genome salad." With these and other disparaging words[1], leading members of the public human genome project last year derided the draft assembly of the human genome produced by Celera Genomics of Rockville, Maryland[2]. Without access to sequence[3] and mapping[4] information produced by the public project, its leaders claimed, Celera would have been nowhere[5].

But these arguments, with their culinary images about who had the superior recipe for sequencing, obscured the fact that future mammalian genome projects would incorporate ingredients of both team's approaches. Despite the continuing debate over whether it succeeded with the human genome, Celera's approach, dubbed the whole-genome shotgun (WGS), was adopted by the publicly funded researchers for their encore, the draft mouse genome. The resulting sequence[6] surpasses the quality of last year's public draft human genome[3].

"The lesson of the mouse assembly is that it is possible to get a good draft using a whole-genome shotgun," says George Weinstock, co-director of the Baylor College of Medicine Human Genome Sequencing Center in Houston, Texas.

The gold standard for sequencing a mammalian genome is the clone-by-clone method, used to produce the public version of the human genome. Here, the genome is chopped into chunks up to several hundred thousand base pairs long, and cloned into bacterial artificial chromosomes (BACs). Each BAC is sequenced by shattering its cloned DNA into smaller fragments and then reassembling the sequence using computer algorithms that match the overlapping ends. Because BACs can be mapped onto the genome's chromosomes by looking for markers called sequence-tagged sites (STSs), the entire sequence is gradually built up. The downside is the cost, time and effort of constructing and sequencing in depth the library of 20,000 or more BACs that is needed to map the entire genome.

The WGS approach avoids this hassle. Rather than producing a BAC map, the genome is instead blasted into millions of fragments, which are sequenced and reassembled to produce a series of sequence 'scaffolds'. These can then be mapped directly onto the chromosomes using STSs. The problem is that mammalian genomes contain millions of repeated sequences. For the assembly algorithms, the challenge is similar to completing a jigsaw that comprises mostly blue sky.



The mouse assembly's success stems from several factors, including the availability of sophisticated public-domain assembly software. At the Whitehead Institute Center for Genome Research at the Massachusetts Institute of Technology, computational biologists have developed a package called ARACHNE, and the Wellcome Trust Sanger Institute at Hinxton, near Cambridge, UK, has produced an equivalent known as Phusion. "They use a whole bunch of computational tricks," says Eric Lander, director of the Whitehead genome centre.

Squeaky clean

Improvements in sequencing technology, giving longer and more accurate sequence reads, have also allowed the mouse genome team to make abundant use of 'mate pairs' — sequences from either end of a DNA segment of known length. If mate pairs subsequently end up the wrong distance apart within the final assembly, it suggests a problem.

Most significantly, perhaps, it has been possible to nail the WGS scaffolds to excellent genetic[7] and physical[8] maps of the mouse genome. The latter was created in part by using the human genome as a 'cheat sheet'. Mice genes align well with human ones, and large blocks exist in which gene order is the same in both genomes. The BACs will now be sequenced in detail over the next couple of years to produce a clean, finished sequence in which gaps are closed and errors corrected.

Although the mouse assembly confirms the utility of the WGS, in itself it says little about the optimal hybrid strategy. But along with the rat genome (available online since 25 November at www.hgsc.bcm.tmc.edu), computational biologists now have enough WGS and clone-by-clone data to try out assemblies using different combinations of each. The best strategy for future projects will depend on the organism, whether a draft or finished sequence is needed, and species' genetic variability.

Looking further ahead, similarities in gene order across mammalian genomes should enable researchers to sequence BAC clones lightly, before comparing them to the human and mouse sequences to construct rough physical maps. With these in hand, it should be possible for researchers interested in comparative genomics to dip into sequence-specific regions of interest in several different species.

The WGS, meanwhile, seems set to become the method of choice for producing rough drafts of large genomes. Clone-by-clone sequencing will remain the *haute cuisine* for finishing genomes, but the WGS's fast food is now firmly on the menu. ■

Declan Butler is *Nature*'s European correspondent.

1. Butler, D. *Nature* **409,** 747–748 (2001).
2. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
3. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
4. International Human Genome Mapping Consortium *Nature* **409,** 934–941 (2001).
5. Waterston, R. H., Lander, E. S. & Sulston, J. E. *Proc. Natl Acad. Sci. USA* **99,** 3712–3716 (2002).