Sequencing bacterial genomes with next generation technologies

A common requirement of bacterial genetics studies is the *de novo* determination of the complete genome sequence of a species or strain of interest[1]. Using Sanger technology, the 'best practice' protocols for achieving the complete sequencing of a bacterial genome involve generation of sequencing libraries with different sized inserts, sequencing the genome to several fold depth from these libraries, and then using assembly tools to infer a draft sequence. This draft sequence, often in many fragments, is then refined by additional sequencing to span gaps, before a final assembly is produced, ready for annotation

Next generation technology speeds up (and reduces the cost of) the first draft, data generation part of the bacterial genome sequencing process. It is relatively easy to generate many-fold coverage of genomes. Because the next generation technologies do not include a step that involves cloning in bacterial vectors, coverage is usually much more even, and is little affected by different AT/GC proportions in the genome.

For *de novo* sequencing of a bacterial genome, we thus recommend

(a) sequencing to ~10 fold coverage in 400 base Roche FLX Titanium reads

(b) sequencing to at least 20x and preferably 30x coverage from two Illumina GAII libraries (having different mean insert lengths, 250 and >400 bases) using 50 base paired end reads.

Assembly of these data with dedicated next-generation assemblers such as Velvet[1] and Newbler[2], will probably yield contigs with well over 98% of the genome in them. However these contigs (there could be as few as 10, and as many as 300 or more) will not be joined-up into a complete genome. The number of contigs that are recovered from an assembly of this kind depend very sensitively on the size of the genome (bigger genomes yield more contigs) and the repeat content of the genome (in the case of bacteria this means the content of IS elements and prophage in the main, although other sequence repeats can also cause 'breaks' in the assembly).

The strategy thereafter will depend on your needs for the project. If you want a complete, circular genome, additional sequencing will need to be done to link the contigs. Depending on the number of contigs and the kind of genome, this could involve performing PCR across the sequence gaps, and then directed Sanger sequencing of the PCR products, or generation of a larger-insert plasmid library and end sequencing a low coverage of clones to identify those that span expected gaps.

The GenePool is happy to supply quotes for draft sequencing and finishing of bacterial genomes, and for bioinformatic support for subsequent first-pass annotation. If your laboratory is not skilled in analyses and annotation, you might consider collaborating with GenePool bioinformaticians to perform detailed analyses of your target genome.

1 Zerbino DR, Birney E. Genome Res. 2008;18:821-9. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. 2 Newbler is proprietary software of Roche designed for FLX reads.

Contact us for more information and a quotation **genepool@ed.ac.uk**

The Gene Pool (The University of Edinburgh Sequencing Facility)
Ashworth Laboratories, King's Buildings, Edinburgh, Scotland EH9 3JT, UK

phone 0131 6513633 | email genepool@ed.ac.uk | http://genepool.bio.ed.ac.uk/