

Comparison of assembly strategies for high throughput *de novo* sequencing of bacterial genomes

Bujie Zhan, Pernille K Andersen, Jakob Hedegaard, Christian Bendixen, Frank Panitz

AARHUS UNIVERSITY, Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, Tjele 8830, Denmark

Introduction

We sequenced *Actinobacillus pleuropneumoniae* serotype 2 (AP2) and serotype 6 (AP6) on Roche 454 Life Science and Illumina Solexa platforms. In order to get good assembly result for post genome annotation, several different *de novo* assembly methods have been tried. Efficient hybrid *de novo* assemble methods have been discovered. A preliminary genome annotation and comparative genome studies based on these hybrid assembly was also performed and confirmed the hybrid assembly accuracy.

Materials and methods

A 300 bp whole genome shotgun library for AP2 and AP6 was sequenced on Roche 454 GS FLX with 12X coverage respectively. Two paired-end genomic libraries with insert size 300 bp and 800 bp for Solexa sequencing respective for AP2 and AP6 gives a coverage 200X foreach microbe. *De novo* genome assembling was performed with different tools and strategies as showed in figure 1. Genome annotation was done based on contigs from *de novo* assembly and genes from related serotype AP3_[NC_010278] and AP7_[NC_010939] were mapped to assembly contigs by BLAT [1] search.

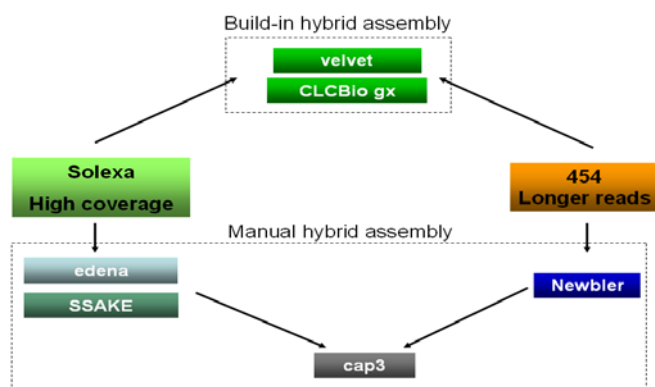


Figure.1 Different assembling strategies were tested.

Results

The best assembly result we got was as follows: 454 reads were assembled with Newbler [2]; subsequently all resulting contigs were extended with SSAKE [3] by using Solexa reads; these extended contigs were assembled again with CAP3 [4]. For results see table 1. In addition, we also used the commercially available software CLCBio Genomic Workbench [www.clcbio.com] for evaluation purpose.

Table.1 The manual hybrid assembly result

Genome	Total Bases (bp)	N50Size (bp)	Contigs #	Average Size (bp)
AP2	2236319	86685	49	45639
AP6	2321722	91035	49	47382

The results of all the other assembly methods with different sequencing data performed in this study are show in figure 2.

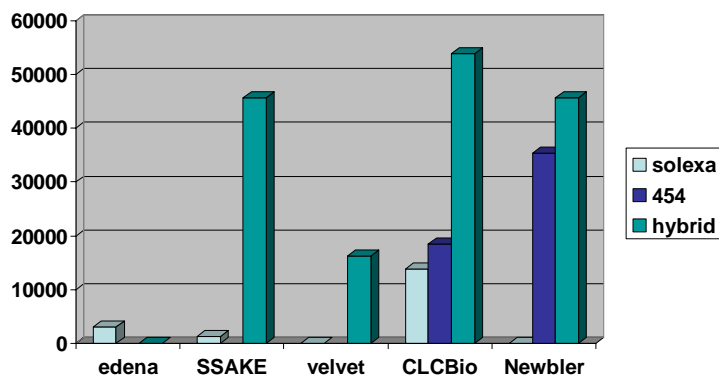


Figure 2. Because all assembly presented whole genomic content, here we take the average contigs size as a measurement of assembly quality. In spite of the lower coverage, 454 reads assembly always gives larger contig sizes. Not surprisingly, the hybrid assembly always gives the largest average contigs sizes.

Contigs related in table 1 were running though Glimmer [5] and EasyGene [6] pipeline for gene prediction, only those genes confirmed by both predictions were selected. All predicted gene number are list in table 2.

Table.2 The preliminary genome annotation summary

Genome	Protein coding genes	tRNA genes	rRNA genes	tmRNA genes
AP2	2044	54	5	1
AP6	2147	52	4	1

All genes from closely related serotype AP3 and AP7 were used to search homology sequences in AP2 and AP6 contigs; homology gene number among AP2, AP3 and AP7 is show as figure 3.

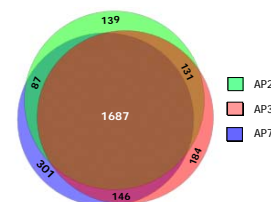


Figure3. Homology genes shared in predicted genes of AP2 and genes in AP3 and AP7

Gap closing PCR and sequencing is currently performed and the contig number is reduced to 37 by introducing 29 gap closing PCR sequences into assembly for AP2 so far.

Conclusions

An efficient manual hybrid *de novo* assembly strategy for 454 and Solexa sequences has been established as: contigs from 454 Newbler assembly were extended with Solexa reads by using SSAKE, and finally re-assemble with CAP3. This strategy achieves larger average contig size when compared with some build-in hybrid assembly tools and give highly accurate genomic sequences. Post genomic analysis like genome annotation and comparative genomic studies will therefore also benefit from this manual hybrid assembly.

References

- Kent, W. J. 2002, *Genome Research* 4: 656-664.
- Roche 454 Life Science
- Warren et al. (2007), *Bioinformatics* 23(4):500-1.
- Huang and Madan (1999), *Genome Res.* 9(9):868-77.
- A. L. Delcher et al, *Nucleic Acids Research* 27:23 (1999), 4636-4641.
- P. Nielsen and A. Krogh, *Bioinformatics*: 21:4322-4329, 2005.

Acknowledgments

We thank Rasmus Ory Nielsen for raw data processing. This project is supported by the Danish Research Council for Technology and Production Sciences (*Analysis of gene expression in infectious diseases – development of tools for tomorrow's therapeutics*) and the Danish Strategic Research Council (Nablit; grant no. 2106-07-0021). For further information, please contact Frank.Panitz@agrsci.dk or visit www.agrsci.dk/gbi.