A comparison of 454 vs traditional sequencing approaches

Gemma M. Payne¹, John C. McEwan¹, Rudiger Brauning¹, Nessa H. O’Sullivan¹, Hutton Oddy², Frank W. Nicholas³, International Sheep Genomics Consortium⁴

Introduction

The International Sheep Genomics Consortium’s aims are to sequence the ovine genome, subsequently identify SNPs and produce a 60K SNP chip. The Roche 454 sequencing platform is a new technology based on pyrosequencing that promises to be cost effective and comparatively quick. Here we report on a pilot study to determine the utility and explore the advantages and disadvantages of 454 compared to traditional Sanger sequencing. Four ovine BACs were selected for sequencing with the 454 GS20. These BACs had previously been sequenced using Sanger technology, as part of the ENCODE project, they contain around 40% repetitive DNA and are typical of the ovine genome. In total, 31Mbp was produced from 454 sequencing, corresponding to 35-52X sequencing depth. Over 80% of the sequences were “clean” – i.e. no vector contamination, low complexity sequence, large runs of Ns or polyA. This represented 30-45X effective sequencing depth. 454 reads were assembled de novo into contigs using CAP3 and Newbler software programs. The Newbler assembler performed the best. Contigs were assembled with a guide genome into longer MELD sequences which covered 95.3-99.6% of the reference sequences. We found 20X sequencing depth sufficient to achieve this level of coverage. MELD sequences had near 100% identity to the original Sanger sequence. Our conclusion is that Roche 454 sequencing is a cost-effective alternative to Sanger sequencing for some sequencing applications.

Methods

BACs were isolated from the Chori-243 ovine BAC library. Inserts ranged from 143 to 215 kbp’s. These BACs were previously shotgun Sanger sequenced as part of the ENCODE (ENCYclopedia Of DNA Elements) project. Sanger sequenced BACs were used as reference sequences. BACs were prepared for sequencing at AgResearch and sequenced by 454 Life Sciences. In this study, sequencing depth is:

# 454 bases # reference sequence bases

Earlier experiments showed that Newbler performed better at assembling reads into contigs than both PHRAP and CAP3. Thus, the Newbler assembler was used for the majority of this study. As Newbler was not initially available, CAP3 was used to assemble and subsequently characterise sequence data for vector contamination, sequence complexity, runs of Ns and polyA’s. Sequence data used for Newbler assembly had previously been cleaned of bacterial contamination by 454 Life Sciences.

The in-house program MELD assembled contigs into a MELD sequence. A comparison of 454 vs traditional sequencing approaches

Conclusions

454 raw data is best assembled into contigs with the Newbler assembler. Subsequently, contigs can be assembled on a larger scale using the reference sequence or genome of a related species as a guide. This process produced high quality sequence, with coverage and identity to Sanger reference sequences approaching 100% for all four BACs. The majority of this coverage corresponded to unique contigs, with only a small percent due to multiple contigs (actg <3%). This infers there is little redundancy in the 454 sequence dataset. This is especially important for whole genome sequencing as it suggests 454 sequencing will give a good overall coverage of the genome. While we achieved much higher sequencing depth, re-sampling showed that a 20X depth is sufficient to obtain near maximal coverage of the reference sequence for BAC sized projects. In summary, this means that 454 sequencing, in conjunction with a reference framework, provides a cost-effective and efficient means of producing high quality usable sequencing data for the ovine genome project.