Back to BACs? A High-Throughput, low cost BAC sequencing pipeline

Darren Heavens, Deepali Vasoya, Gawain Bennett, Heather Musk, James Lipscombe, Dharanya Sampath, Richard Leggett, Anthony Rogers, Robbie Waugh, Sarah Ayling and Matthew D. Clark

The Genome Analysis Centre, Norwich Research Park, United Kingdom

Abstract

Sequencing individual BACs from a minimal tile path (MTP) overcomes many assembly problems associated with heterozygosity, repeats, and duplications. This can be a huge benefit when sequencing **large, complex, repeat rich and polyploid genomes** such as bread wheat. Here we detail a scalable, **low cost (currently $8/BAC), high throughput pipeline to construct indexed libraries** from 2,304 BACs in a standard working day. The libraries can be pooled into **a single Illumina lane, sequenced, demultiplexed** and each BAC **individually assembled** to an average **contig N50 >45kbp**. We have validated this approach by sequencing the **barley 2H MTP** and are currently sequencing the **wheat 3DL MTP.** Our assemblies are already considerably better than common whole genome shotgun projects (contig N50 ~10kbp) and by adding further mate pair data we achieve average **scaffold N50 >75kb**.

Future work will focus on increasing the pooling strategy (current average sequence coverage per BAC is >200x), increasing the LMP insert to 10kbp and improving the normalisation protocol. We hope to improve the daily throughput to 3072 BACs (8x 384), generate single scaffolds for most BACs and further reduce the cost per BAC.

1. BAC DNA Prep

BACs are grown in 140µl 2xYT supplemented with chloramphenicol and processed on a Beckman FXp robot using a **modified alkali lysis protocol** based on CosMc bead technology (Beckman). ATP dependent DNase is employed to remove host *E. coli* DNA and **typical yields are 35ng of BAC DNA**, see Figure 1, with **95%+ purity** as determined by qPCR.

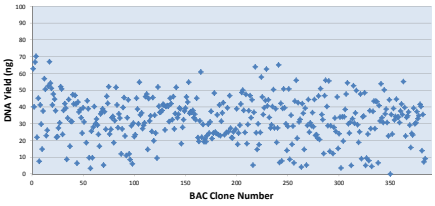
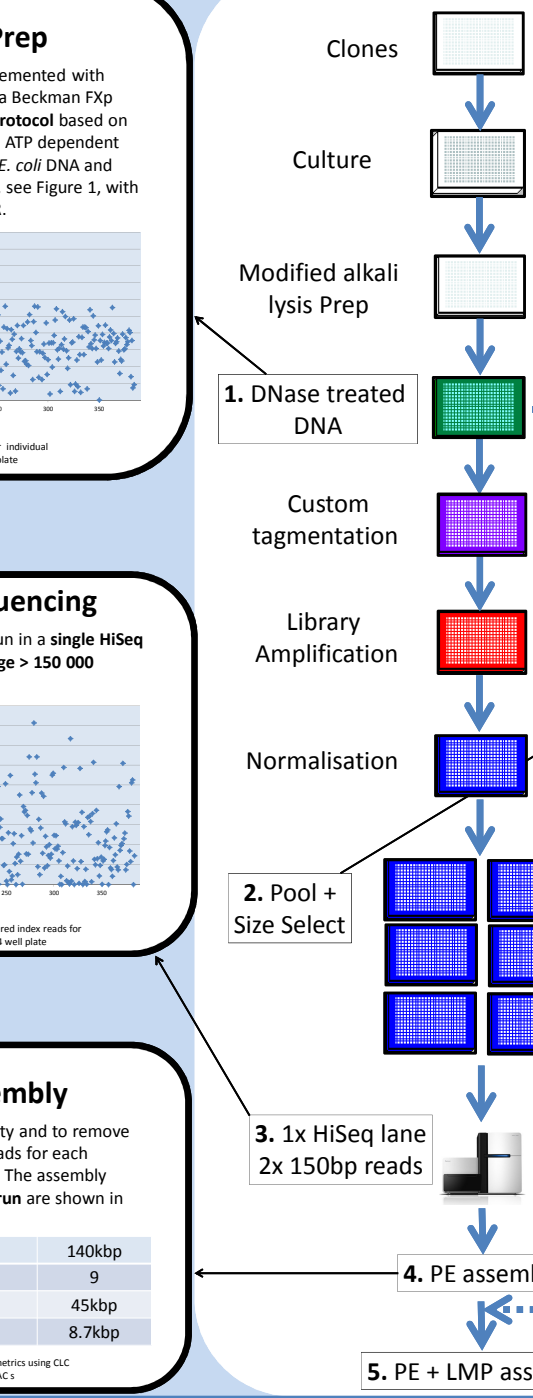


Figure 1: Graph showing the DNA yield for individual BAC DNA preps across a 384 well plate



2. BAC Paired End (PE) Library Construction

Libraries are constructed following an **in-house, custom protocol based on the Nextera library construction kit** (Illumina). They are subjected to amplification using different combinations of 48 barcoded P5 and 48 barcoded P7 primers. Post amplification **library yields are around 15ng/ µl in 10 µl**, see Figure 2.

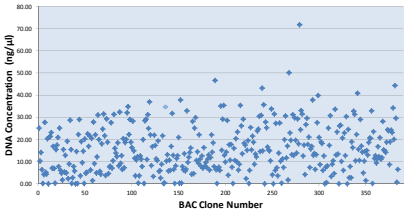


Figure 2: Graph showing the concentration of individual BAC libraries across a 384 well plate

Libraries are then **normalised** using MagQuant beads (GC biotech) and up to **2304 BACs** pooled, purified and **size selected** on a Sage Scientific Blue Pippin.

3. BAC Pool Sequencing

Up to 2304 BACs at a time can be run in a **single HiSeq lane**. For a 1536 BAC pool we **average > 150 000 filtered reads per BAC**, see Figure 3

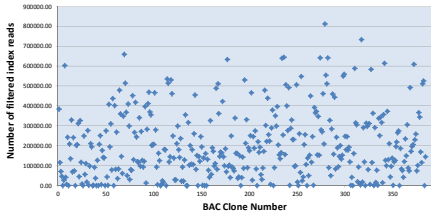


Figure 3: Graph showing the number of filtered index reads for individual BAC libraries across a 384 well plate

4. BAC PE Assembly

Sequence reads are filtered for quality and to remove host *E. coli* sequences and the PE reads for each individual BAC assembled using CLC. The assembly metrics for a **1536 Barley BAC pool run** are shown in Table 1.

Average Assembled Content	140kbp
Average Contig Number per BAC	9
Average N50	45kbp
Average Contig Size	8.7kbp

Table 1: Table showing the average assembly metrics using CLC on 2x 150bp reads across 1536 BAC s

5. BAC Long Mate Pair (LMP) Construction and Scaffolding

Pools of 384 BAC DNA samples are subjected to standard LMP library construction using the Nextera LMP kit (Illumina). To merge data sets the PE sequences are first assembled using ABYSS and the LMP data then added using SOAP. For one example adding the LMP plate data improves the **average N50 of the assembly across a 384 well plate to >75kbp from 17kbp, see Figure 4**. In some cases adding LMP data allows BACs to be **assembled into single contigs,** see Figure 5.

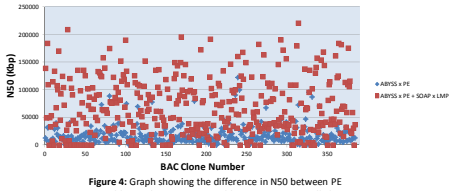


Figure 4: Graph showing the difference in N50 between PE assemblies and PE + LMP assemblies across a 384 well plate

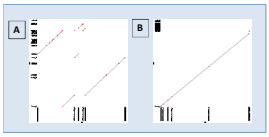


Figure 5: Dot plots showing the CLC assembly for a single BAC with PE data (A) and with added LMP data (B). X-axis is 454 based references, Y-axis our pipeline.

Acknowledgements

We thank Dave Baker, Tom Barker and Rachel Piddock (TGAC) for help and advice in quantifying libraries and running the Illumina Sequencers, Darren Waite (TGAC) for running the primary analysis pipeline and Bernardo Clavijo (TGAC) for advice on sequence assembly . We also thank Nils Stein and Sebastian Beier at IPK for sharing their sequence analysis pipeline with us and other members of the Wheat and Barley International Sequencing Consortia.

TGAC
The Genome Analysis Centre™

BBSRC

Greater Norwich Development Partnership

We are here!!!
Come and talk to us!
Find us at Booth 324

Contact Details

Email:darren.heavens@tgac.ac.uk
Tel: +44 (0) 1603450838
Email:matt.clark@tgac.ac.uk
Tel: +44 (0) 1603450138
www.tgac.ac.uk