A Large Genome Centre Core Pipeline Refresh

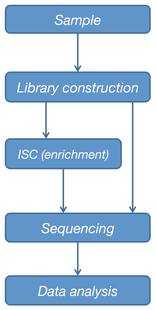
Scott Thurston*, **Peter Ellis, Lesley Shirley, Carol Scott, James Glover, Benjamin Farr, Jamieson Lovell, Richard Rance, Tony Cox**

Wellcome Trust Sanger Institute, Hinxton, Cambridge, U.K.
****Presenting author***

Background

Overview of High-throughput DNA Sequencing

The current operation employs 15 laboratory staff running multiple high-throughput Illumina sequencing pipelines to produce in excess of 1 petabase of data per year.



Up to 20,000 DNA samples arrive each month
- QC: DNA quantitation (picogreen); SNP QA check
- Re-formatting (cherry-pick) and normalisation

Up to 16,000 DNA libraries are prepared each month
- WGS (450 bp) or targeted (150 bp)
- Typically 6x PCR cycles
- PCR-free option

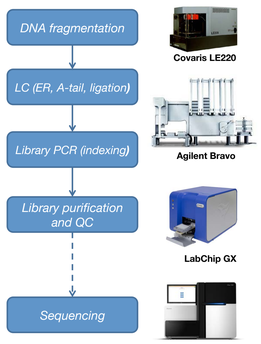
Agilent SureSelect – indexed sequence capture (ISC)

High-throughput sequencing underpinned by Illumina SBS tech
- HiSeq X, HiSeq 4000, HiSeq 2500, MiSeq
- 75 or 150 base PE sequencing + index

Analysis pipeline
- Data QC & storage to iRODS

DNA library construction (2010-2017)

Our high-throughput pipelines have been continually improved over the years but a step change in workflow design was necessary to prepare us for future projects.



Covaris LE220

Agilent Bravo

LabChip GX

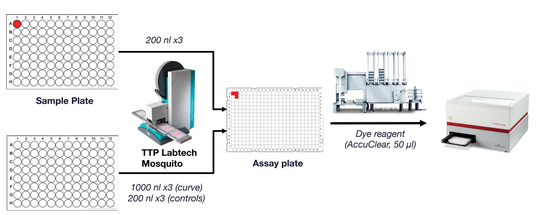
Why we loved this pipeline
- Robust at scale
  - Range of genomes; mammalian, microbial
  - Wide range of DNA sources
- Modest DNA input (500 ng)
- High quality DNA libraries
- Capacity matched to demand and seq capacity
- Cost-effective

Why we needed to upgrade
- Increase in science with limited DNA availability
- Capacity increase; FTE/capital outlay NOT the solution
- Increased demand for PCR-free (1000 ng input)
- Library costs exceed DNA Seq costs for many projects

Improving DNA quantitation: A key enabler

A key step towards overhauling and streamlining our workflows was to improve the quality of plates entering our high-throughput pipelines.

What are the benefits of this new QC?
- Less than 1 µl of sample is required
- Wider linear dynamic range than previous QC methods (0.03-200 ng/µl of stock DNA)
- Reduced turnaround time and costs compared to previous QC methods
- Positive displacement pipetting mitigates the viscoelastic properties of high molecular-weight DNA
- Improved cherry-picking and increased first-time pass rate



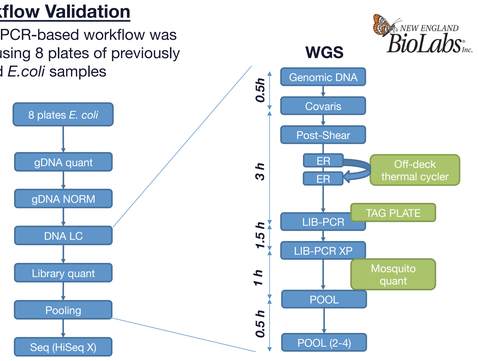
PCR-Based LC

The challenge
- Develop and implement a streamlined PCR-based LC process
- Reduce DNA input requirements by at least 50%
- Reduce per sample costs by ~50%
- No reduction in capacity nor increase in FTE requirement

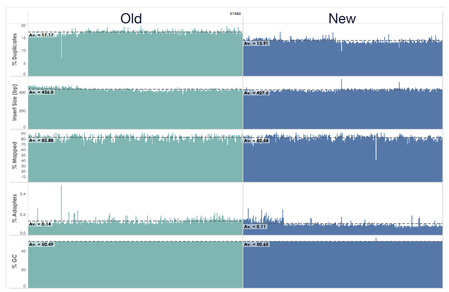
Development of a streamlined, automated PCR-based workflow
- New NEBNext Ultrall workflow is ~30% faster enabling 2x PCR-based LC runs per Bravo per day
- Less demanding of FTE resource than the previous Sanger workflow
- Standard DNA input requirement reduced 2.5x to 200 ng
- Workflow is compatible with whole genome and targeted sequencing
- Streamlined DNA library quantitation utilising same method as novel gDNA assay

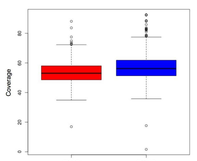
LC Workflow Validation

The entire PCR-based workflow was validated using 8 plates of previously sequenced *E.coli* samples

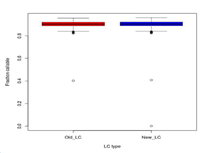


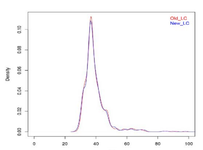
- The new workflow met all pipeline metric performance targets
- Data comparison indicated no loss of quality with new, streamlined workflow





Analysis of data sets between workflows shows a high level of consistency for:
- Uniformity of coverage
- Fraction of the genome that is callable

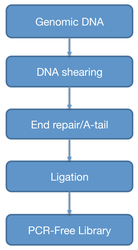


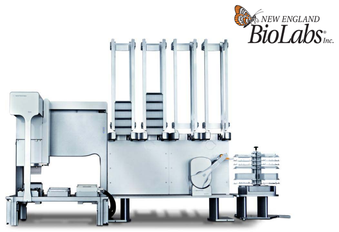


PCR-Free LC

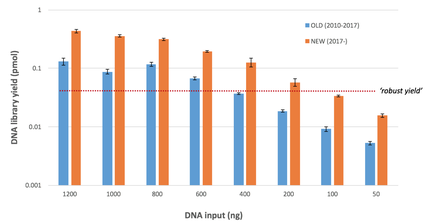
The challenge
- Develop and implement a streamlined PCR-free LC process
- Reduce DNA input requirements by 50%
- No reduction in capacity nor increase in FTE requirement
- Align process with workflows implemented for PCR-based LC
- Implement Unique Dual Indexing (UDI)

Development of a streamlined, automated PCR-Free workflow
- New NEBNext Ultrall PCR-Free workflow is 2x faster end-to end and allows 2x PCR-Free LC runs per Bravo per day
- PCR-Free workflow adopts new gDNA quant and normalisation steps

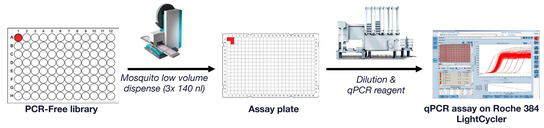




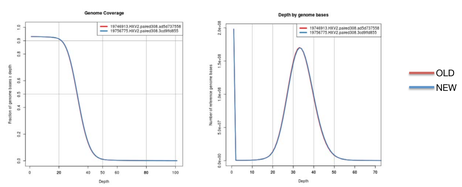
- DNA input requirements are reduced from 1000 ng to 500 ng



- Library QC is dramatically streamlined in new LC workflow (~5x faster)
- Low volume dispensing minimises loss of precious library to measurement process



- New PCR-Free workflow matches data quality of previous workflow
- Genotype concordance between methods >99%



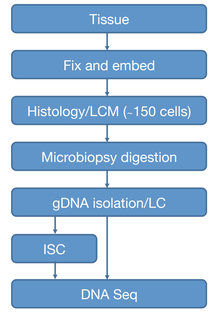
- Unique dual indexes (UDI) are incorporated in to each PCR-FREE library
- UDIs allow us to filter 'contaminating' reads generated by index hopping
- First release UDI set comprises 96 x 96 8-base barcoded adapters

LCMB LC

The challenge
- Sanger's cancer and somatic mutation group aim to investigate clonal dynamics and mutational signatures in all tissues
- Requires laser capture microdissection (LCM) of just a few hundred cells
- DNA input is approximately 1 ng
- A high-throughput pipeline required for many thousands of samples
- Whole genome amplification methods NOT allowed

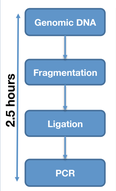
LCM biopsy e.g. ~150 cells → **High quality WG or targeted Seq data**

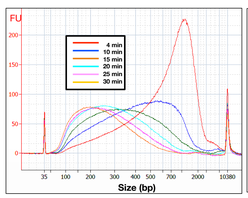
Optimising LCM sample prep for low input LC workflow
- Sample prep and LC workflow were co-developed

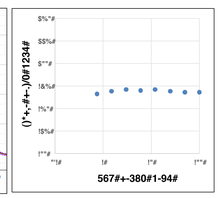


- **Tissue fixation**
- **Embedding compound**
- **Staining protocols**
- **Laser capture microdissection**
- **Biopsy digestion buffers**
- **gDNA isolation rolled in with LC**

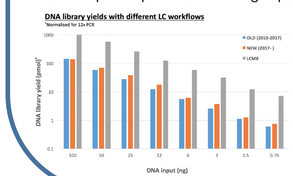
Ultra low DNA input LC is enabled by NEBNext Ultra II FS
- Ultra II FS is a novel LC reagent from New England BioLabs (NEB)
- No requirement for Covaris shearing (enzymatic)
- Fragment profile is dependent on reaction time (e.g. 10 min for WGS profile)
- Fragment profile is independent of DNA input

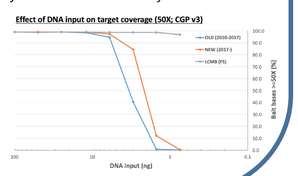






- The LCMB pipeline produces 10-15x more DNA library compared with other methods
- DNA input requirements for high quality data are dramatically reduced





Conclusions

2017 has seen a major overhaul of our high-throughput LC pipelines
- Improved quality of gDNA plates entering DNA library construction
- Faster, less expensive processes with no loss of data quality
- Reduction in DNA input requirements for PCR-based & PCR-Free workflows
- First steps to implementation of unique dual indexes for majority of DNA libraries
- Implementation of a new workflow (LCMB) capable of producing high quality whole genome and targeted human seq data from a few hundred cells