

How we do GBS...

And what's next?

Kevin Murray

@kdmurray91

kevin@kdmurray.id.au

Borevitz Lab, ANU

MapNet GBS workshop

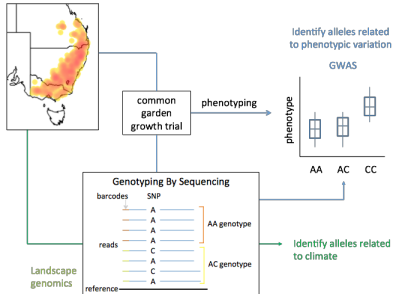
23 Oct 2014

- ▶ Bioinformatics RA
 - ▶ TraitCapture Project Developer
 - ▶ Genomics (Low level sequence analysis)
 - ▶ Phenomics (Image analysis)
 - ▶ Sample tracking, data standards, HPC
- ▶ Starting PhD in Bioinformatics of Evolutionary Genomics next year
- ▶ *"Data Intensive Biologist"*

- ▶ Bioinformatics RA
 - ▶ TraitCapture Project Developer
 - ▶ Genomics (Low level sequence analysis)
 - ▶ Phenomics (Image analysis)
 - ▶ Sample tracking, data standards, HPC
- ▶ Starting PhD in Bioinformatics of Evolutionary Genomics next year
- ▶ *“Data Intensive Biologist”*
- ▶ A **very** lapsed kiwi (from Napier)

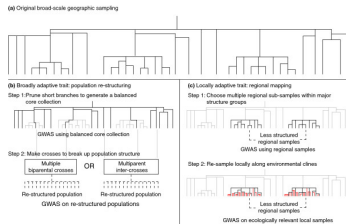
What do we use GBS for?

- ▶ Population and Landscape Genomics
- ▶ Population re-sampling
- ▶ Genotype verification
- ▶ Epigenomics - HpaII/Mspl



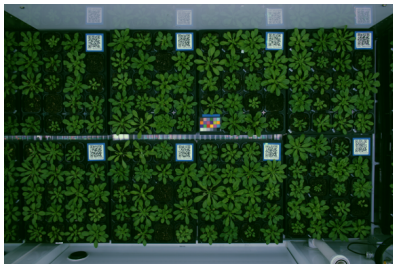
What do we use GBS for?

- ▶ Population and Landscape Genomics
- ▶ Population re-sampling
- ▶ Genotype verification
- ▶ Epigenomics - HpaII/Mspl



What do we use GBS for?

- ▶ Population and Landscape Genomics
- ▶ Population re-sampling
- ▶ Genotype verification
- ▶ Epigenomics - HpaII/Mspl



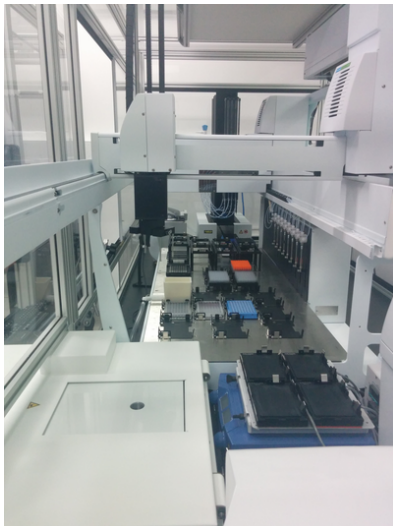
Library preparation

- ▶ We do all our library prep. in house
- ▶ Cost \approx \$10/sample including extraction
- ▶ Done in lots of 96
- ▶ Semi-automated wet-lab protocol



Library preparation

- ▶ We do all our library prep. in house
- ▶ Cost \approx \$10/sample including extraction
- ▶ Done in lots of 96
- ▶ Semi-automated wet-lab protocol



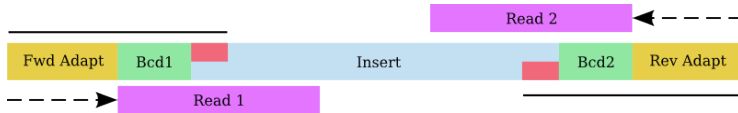
Library preparation

- ▶ We do all our library prep. in house
- ▶ Cost \approx \$10/sample including extraction
- ▶ Done in lots of 96
- ▶ Semi-automated wet-lab protocol



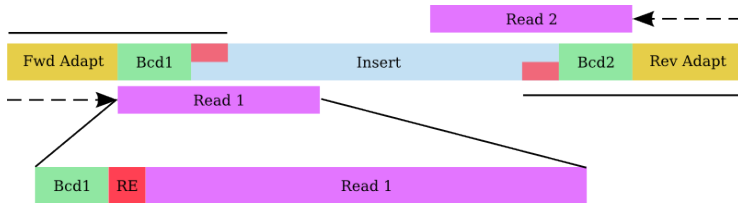
Library preparation protocol

- ▶ Modified protocol from Elshire et al. (2011)
- ▶ PE 101bp HiSeq 2500 reads
- ▶ We use the PstI as cutter
- ▶ Sequence 384-1152 samples/lane
- ▶ Need combinatorial barcoding (next slide)



Library preparation protocol

- ▶ Modified protocol from Elshire et al. (2011)
- ▶ PE 101bp HiSeq 2500 reads
- ▶ We use the PstI as cutter
- ▶ Sequence 384-1152 samples/lane
- ▶ Need combinatorial barcoding (next slide)



Combinatorial Barcoding

- ▶ We use dual in read barcodes
- ▶ Independent barcodes for R1 & R2
- ▶ Set of $96 \times 12 = 1152/\text{lane}$
- ▶ Must be balanced & staggered
- ▶ Hamming distance >2

TGCG
AGGAT
TTCAGA
CGCGGT
GAATTCA
CTACGGA
....
CCGGATAT
TTCCTGGA

X

TGCG
AGGAT
TTCAGA
CGCGGT
GAATTCA
CTACGGA
....
CCGGATAT
TTCCTGGA

=

CTCG	ATGAAAG	1_A1
TGCA	ATGAAAG	1_A2
ACTA	ATGAAAG	1_A3
AACT	ATGAAAG	1_A5
.....		
GCTGTGGA	CTTGCTT	12_H8
GTGAGGGT	CTTGCTT	12_H10
TATCGGGA	CTTGCTT	12_H11
TTCCTGGA	CTTGCTT	12_H12

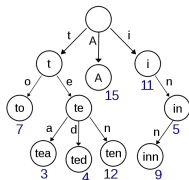
96 R1

12 R2

1152 pairs

De-multiplex with AXE

- ▶ Barcoding scheme requires advanced de-multiplexing
- ▶ Trie-based lookup algorithm
- ▶ Fast (PE lane in 5-10 mins)
- ▶ Highly accurate (99.6%TPR, 100%TNR)
- ▶ Open source at <http://git.io/kIhEZA>
- ▶ Benchmarking IPython notebook at <http://bit.ly/1sXrx4E>

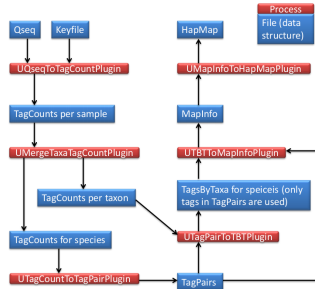


- ▶ Sample-tracking system manages FASTQs after demux:
 - ▶ QC reads (scythe, sickle, seqqs)
 - ▶ After QC, 16-mer globally unique ID re-added to reads
 - ▶ Auto-creates TASSEL KeyFile & working dir for new analysis
 - ▶ Auto-remove 16-mer unique ID from reads if not using TASSEL.

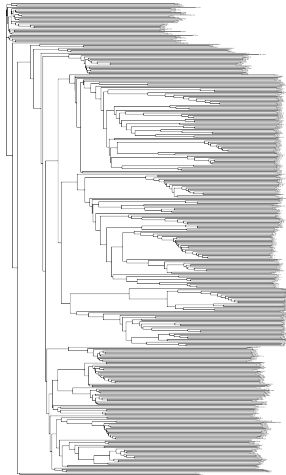
- ▶ Sample-tracking system manages FASTQs after demux:
 - ▶ QC reads (scythe, sickle, seqqs)
 - ▶ After QC, 16-mer globally unique ID re-added to reads
 - ▶ Auto-creates TASSEL KeyFile & working dir for new analysis
 - ▶ Auto-remove 16-mer unique ID from reads if not using TASSEL.
- ▶ DB links FASTQs to real world
 - ▶ Integrate w/ ALA
 - ▶ Provide user-friendly names to downstream analysis

Downstream Analysis

- ▶ Variant calling:
 - ▶ TASSEL!!
 - ▶ We use UNEAK mostly
 - ▶ Ref-based pipeline w/
BWA-MEM

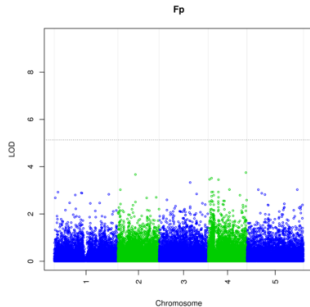


- ▶ Variant calling:
 - ▶ TASSEL!!
 - ▶ We use UNEAK mostly
 - ▶ Ref-based pipeline w/
BWA-MEM
- ▶ Post analysis in R
 - ▶ Custom filtering for
missing data/paralogy
 - ▶ H-Clust

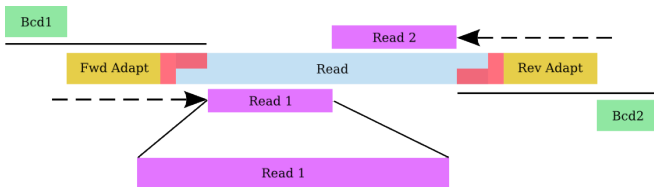


Downstream Analysis

- ▶ Variant calling:
 - ▶ TASSEL!!
 - ▶ We use UNEAK mostly
 - ▶ Ref-based pipeline w/
BWA-MEM
- ▶ Post analysis in R
 - ▶ Custom filtering for
missing data/paralogy
 - ▶ H-Clust
- ▶ Downstream:
 - ▶ Structure
 - ▶ BayENV
 - ▶ QTLRel

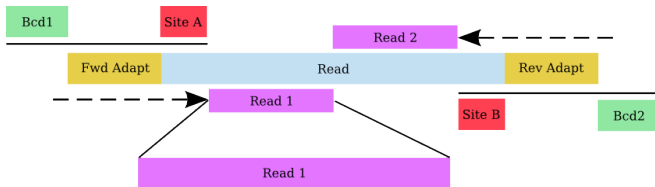


► Better adaptors



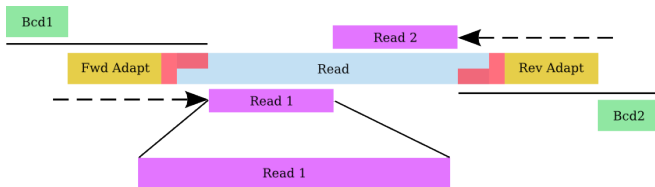
The Future!

- ▶ Better adaptors
- ▶ Home-brew NexTera
 - ▶ NextRAD
 - ▶ Low coverage WGS
 - ▶ Long Pseudo-reads



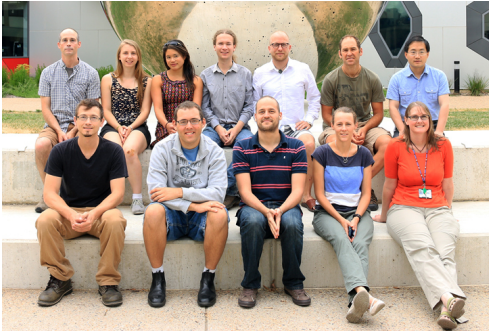
The Future!

- ▶ Better adaptors
 - ▶ Home-brew NexTera
 - ▶ NextRAD
 - ▶ Low coverage WGS
 - ▶ Long Pseudo-reads
- ▶ Informatics:
 - ▶ Imputation?
 - ▶ Paralog/Ploidy detection
 - ▶ Streaming Variant Calling (my PhD)



Thanks

- ▶ Justin Borevitz
- ▶ Comrades in Informatics
 - ▶ Aaron Chuah
 - ▶ Riyan Chen
 - ▶ Jared Streich
- ▶ Wet-lab Wizardry
 - ▶ Niccy Aitken
 - ▶ Norman Warthmann
- ▶ Rob for the invitation
- ▶ You all for listening!!



`git.io/TYrIFw`