How we do GBS. . .

And what's next?

Kevin Murray
@kdmurray91
kevin@kdmurray.id.au

Borevitz Lab, ANU

MapNet GBS workshop
23 Oct 2014

About Me

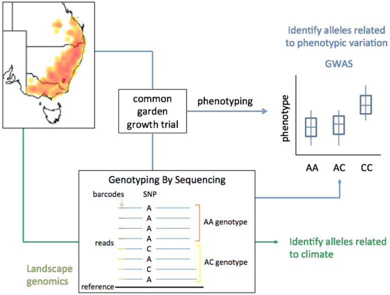
- Bioinformatics RA
  - TraitCapture Project Developer
  - Genomics (Low level sequence analysis)
  - Phenomics (Image analysis)
  - Sample tracking, data standards, HPC
- Starting PhD in Bioinformatics of Evolutionary Genomics next year
- *“Data Intensive Biologist”*

About Me

- 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)

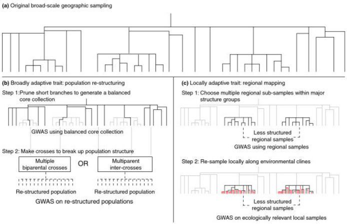
What do we use GBS for?

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



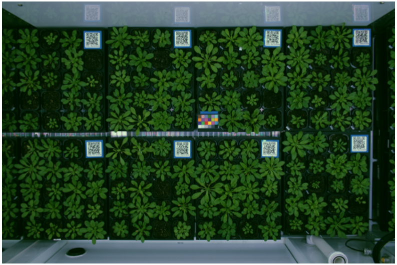
What do we use GBS for?

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


(a) Original broad-scale geographic sampling
(b) Broadly adaptive trait: population re-structuring
Step 1:Prune short branches to generate a balanced core collection
GWAS using balanced core collection
Step 2: Make crosses to break up population structure
Multiple biparental crosses
OR
Multiparent inter-crosses
Re-structured population
Re-structured population
GWAS on re-structured populations
(c) Locally adaptive trait: regional mapping
Step 1: Choose multiple regional sub-samples within major structure groups
Less structured regional samples
GWAS using regional samples
Step 2: Re-sample locally along environmental clines
Less structured regional samples
GWAS on ecologically relevant local samples

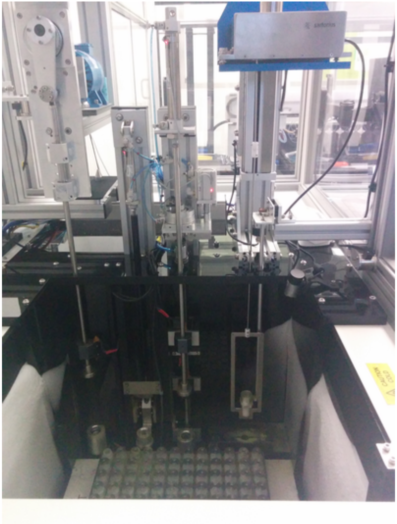
What do we use GBS for?

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



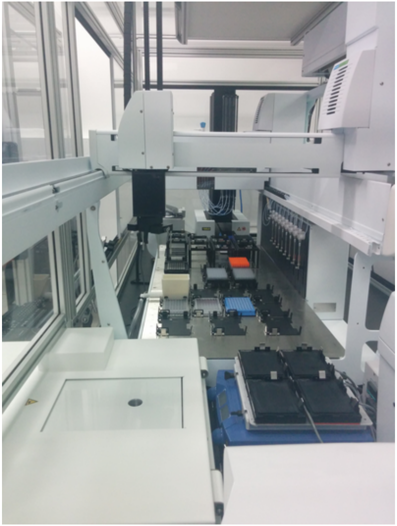
Library preparation

- We do all our library prep. in house
- Cost ≈ \$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

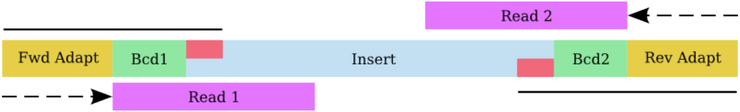


Library preparation

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

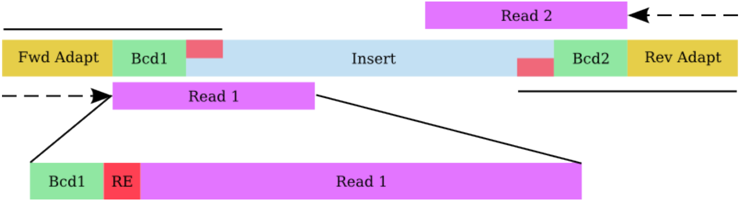
Library preparation protocol

- Modified protocol from Elshire et al. (2011)
- PE 101bp HiSeq 2500 reads
- We use in read barcodes
- Sequence 576-1152 samples/lane
- Need "combinatorial" barcoding (next slide)



Library preparation protocol

- Modified protocol from Elshire et al. (2011)
- PE 101bp HiSeq 2500 reads
- We use in read barcodes
- Sequence 576-1152 samples/lane
- Need "combinatorial" barcoding (next slide)



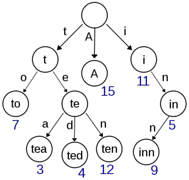
Combinatorial Barcoding

- We use dual in read barcodes
- Independent barcodes for R1 & R2
- Set of 96x12 = 1152/lane
- Must be balanced & staggered
- Hamming distance >2

TGCG		TGCG		CTCG	ATGAAAG	1_A1
AGGAT		AGGAT		TGCA	ATGAAAG	1_A2
TTCAGA		TTCAGA		ACTA	ATGAAAG	1_A3
CGCGGT		CGCGGT		AACT	ATGAAAG	1_A5
GAATTCA	X	GAATTCA	=	......		
CTACGGA		CTACGGA		GCTGTGGA	CTTGCTT	12_H8
....		....		GTGAGGGT	CTTGCTT	12_H10
CCGGATAT		CCGGATAT		TATCGGGA	CTTGCTT	12_H11
TTCCTGGA		TTCCTGGA		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 Management

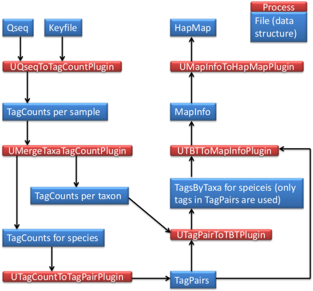
- Sample-tracking system manages FASTQs after demux:
  - QC reads (scythe, sickle, seqqs)
  - After QC, 16-mer globally unique ID re-added to reads
  - Auto-creates TASSEL KeyFile & working dir for new analysis
  - Auto-remove 16-mer unique ID from reads if not using TASSEL.

Sample Management

- 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

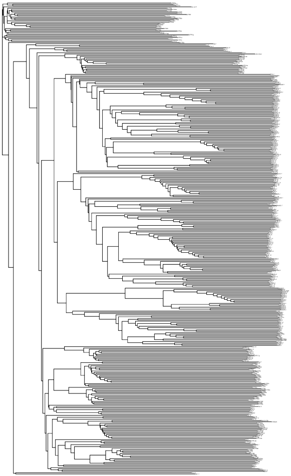
Downstream Analysis

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



Downstream Analysis

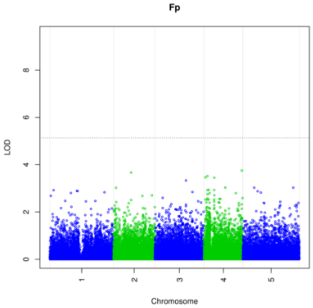
- Variant calling:
  - TASSEL!!
  - We use UNEAK mostly
  - Ref-based pipeline w/ BWA-MEM
- Post analysis in R
  - Custom filtering for missing data/paralogy
  - H-Clust



plant energy biology
ARC CENTRE OF EXCELLENCE

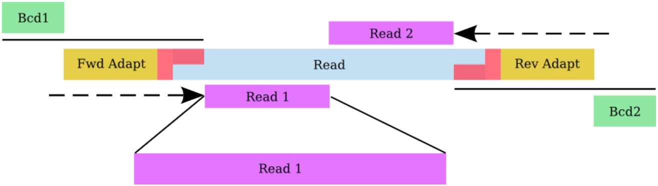
Downstream Analysis

- 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



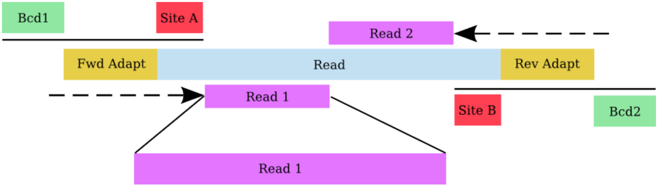
The Future!

- Better adaptors



The Future!

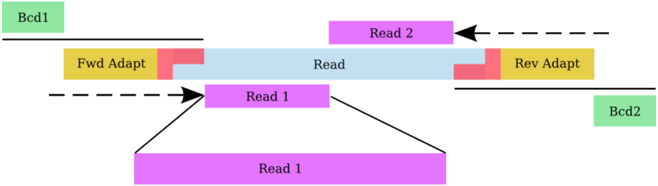
- Better adaptors
- Home-brew NexTera
  - NextRAD
  - Low coverage WGS
  - Long Pseudo-reads



The Future!

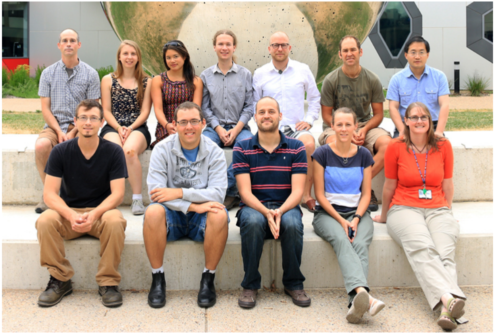
- Better adaptors
- Home-brew NexTera
  - NextRAD
  - Low coverage WGS
  - Long Pseudo-reads

- 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