GBS QC and Pipelines

Kevin Murray

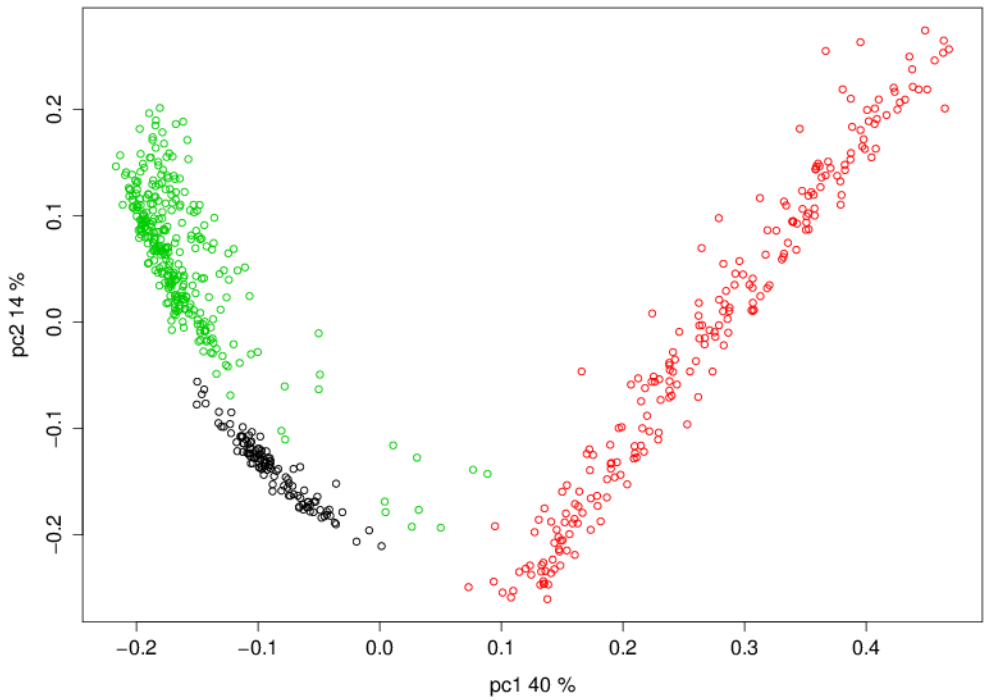
Borevitz Lab, CPEB, ANU

March 3, 2016

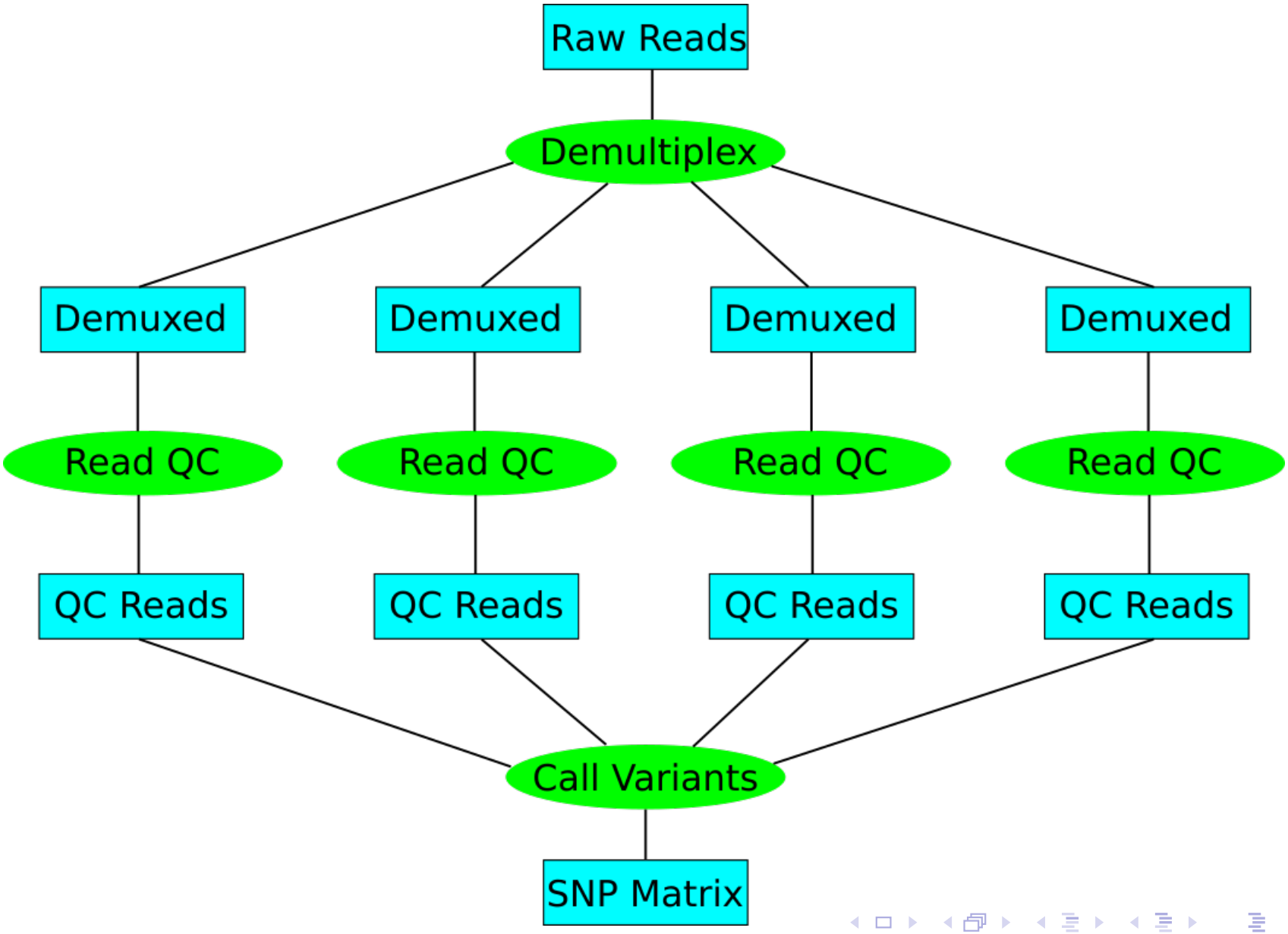
"The missing heritablility is in your shitty QC and assembly"

The GBS woes

- Reproducibility?
- Funky data
- Technical Noise
- Weird software requirements
- Obsolete software?
- Weird adaptors or barcodes?
- Care to add any?



The GBS pipeline

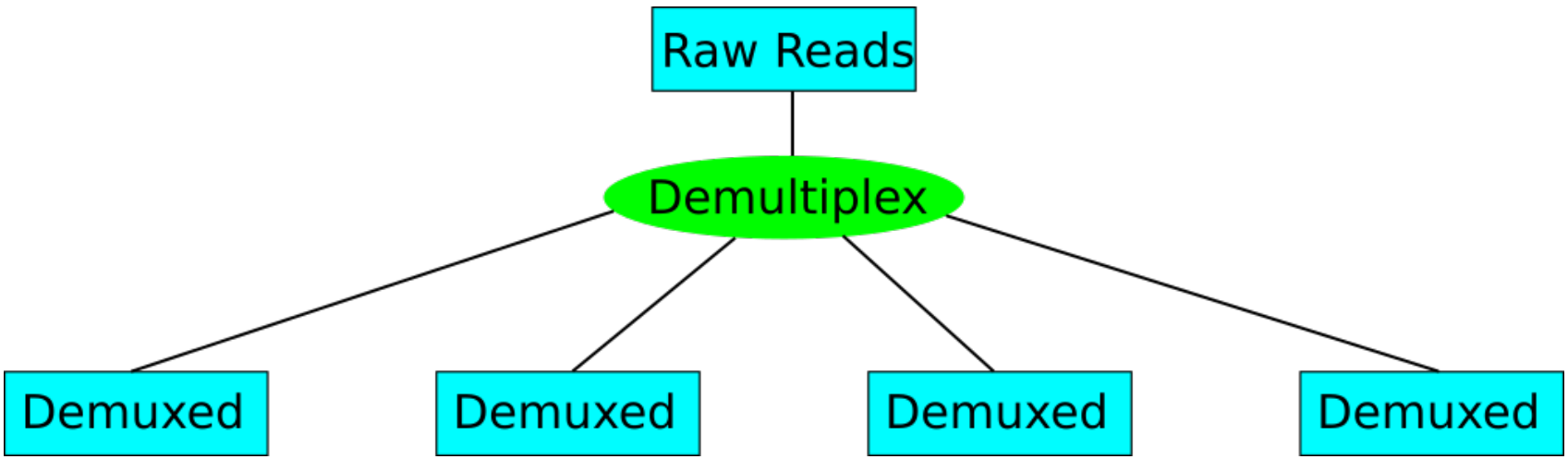


Demultiplex

- Use Axe!
- Written in 2014, pretty quick, stable software
- Have experiments that show it works (esp. for GBS)
- Planning to write it up this year
- Now part of the Debian and Ubuntu distributions!

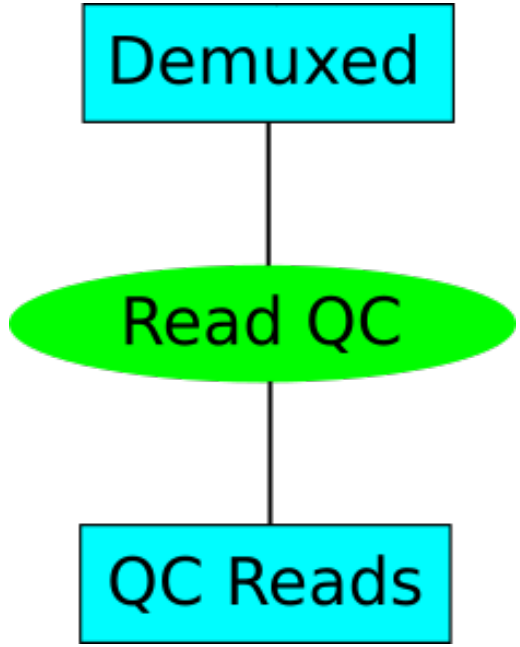
Demultiplex: Practicalities

- Axe takes a key file per lane
  - TSV: Barcode1, Barcode2, samplename
  - Sample names must be valid paths
  - Sample names must be unique within lane
- (normally) produces 1 interleaved Fastq per sample
- Reports stats, saves bad reads to file

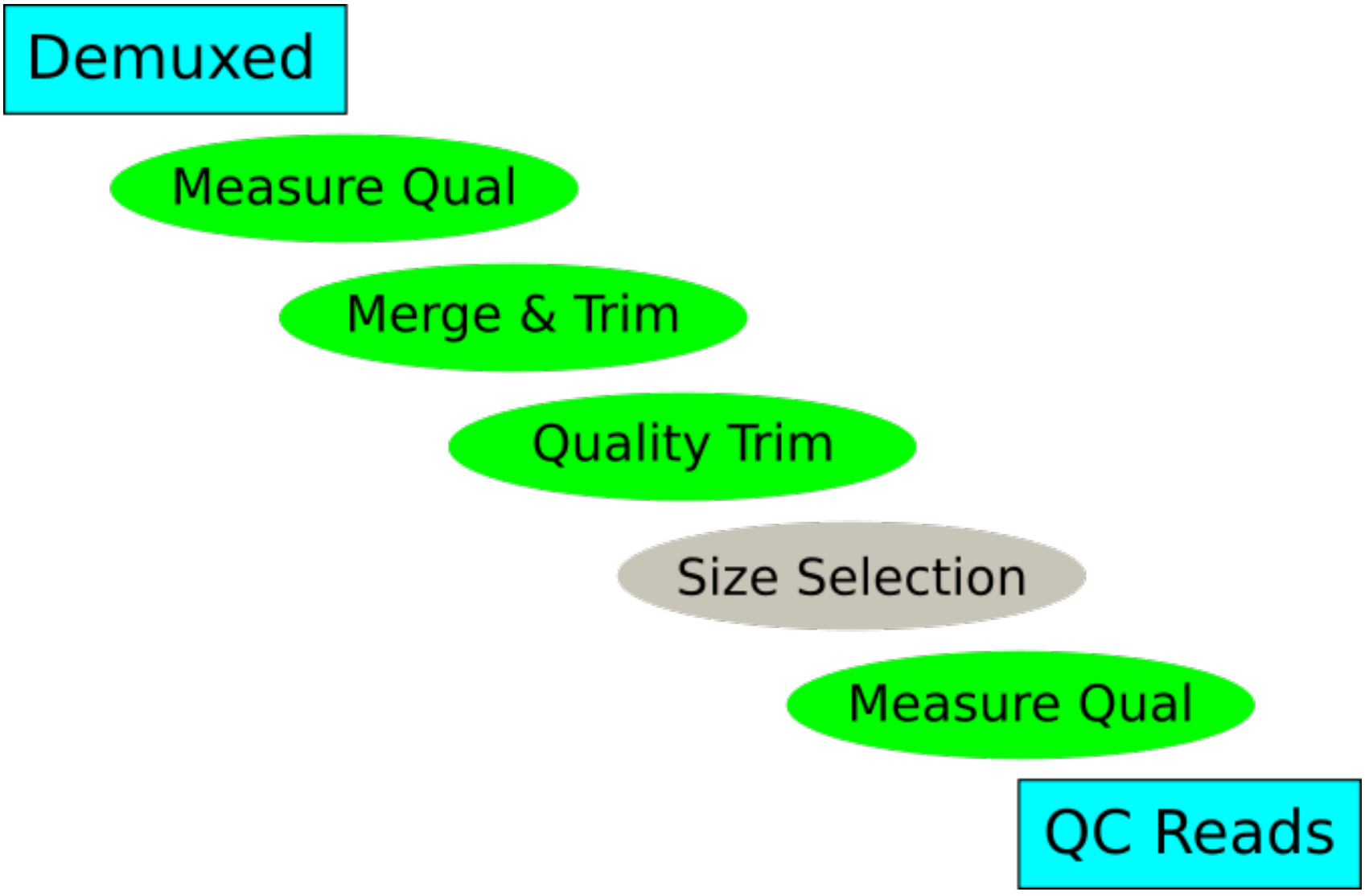


GBSQC

- New tool
- Written last year, to help Megan & Jared's data
- C++ library and CLI tool
- One-stop-shop QC tool

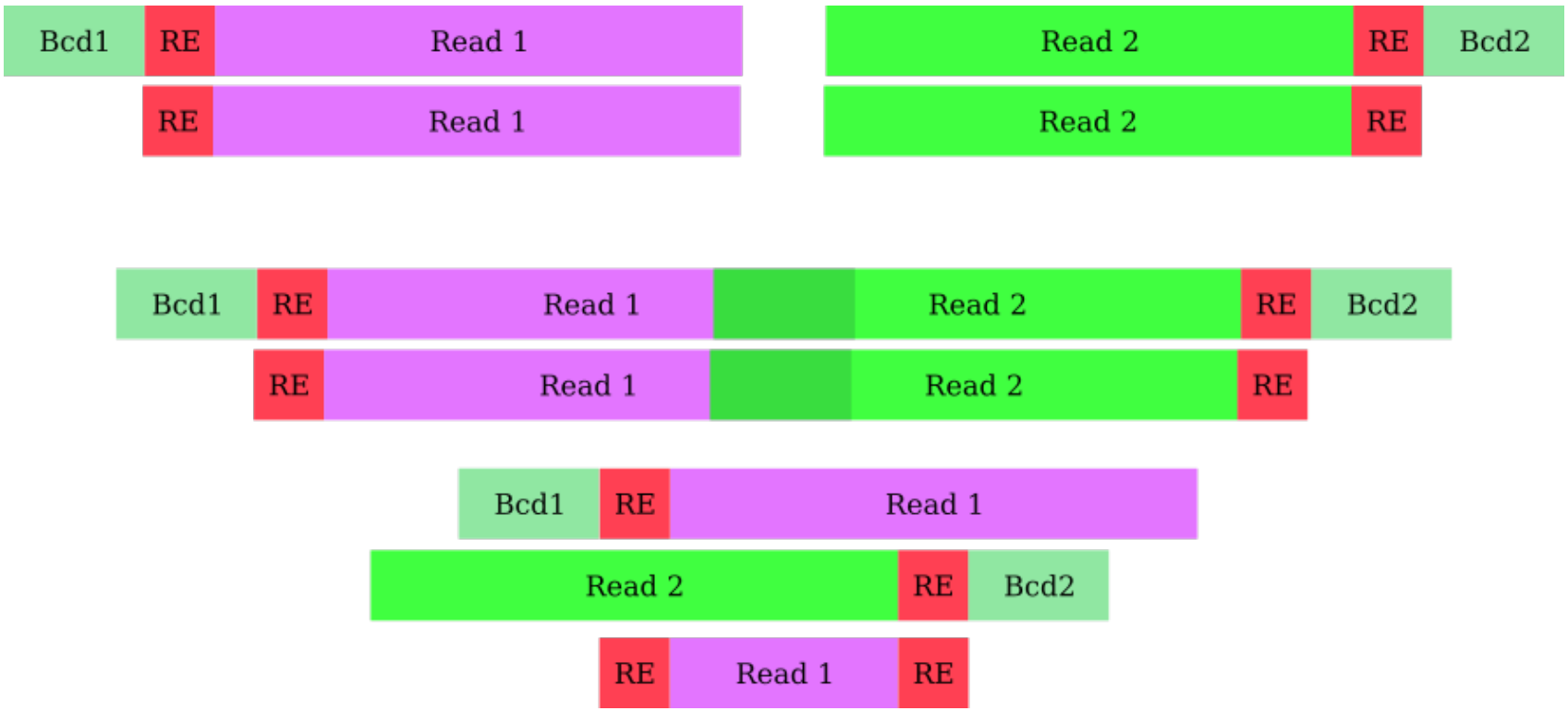


GBSQC

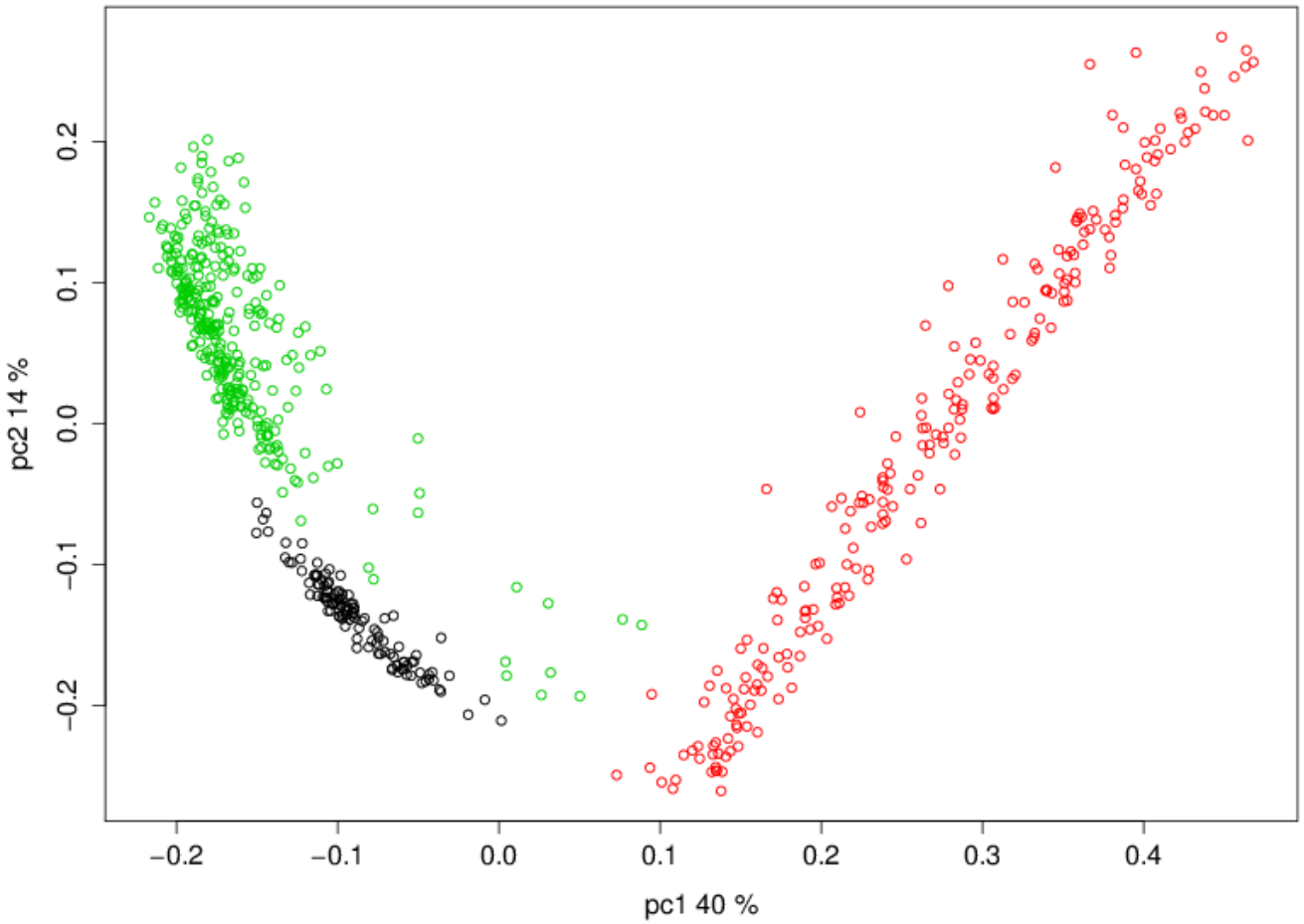

Demuxed
Measure Qual
Merge & Trim
Quality Trim
Size Selection
Measure Qual
QC Reads

TrimMerge

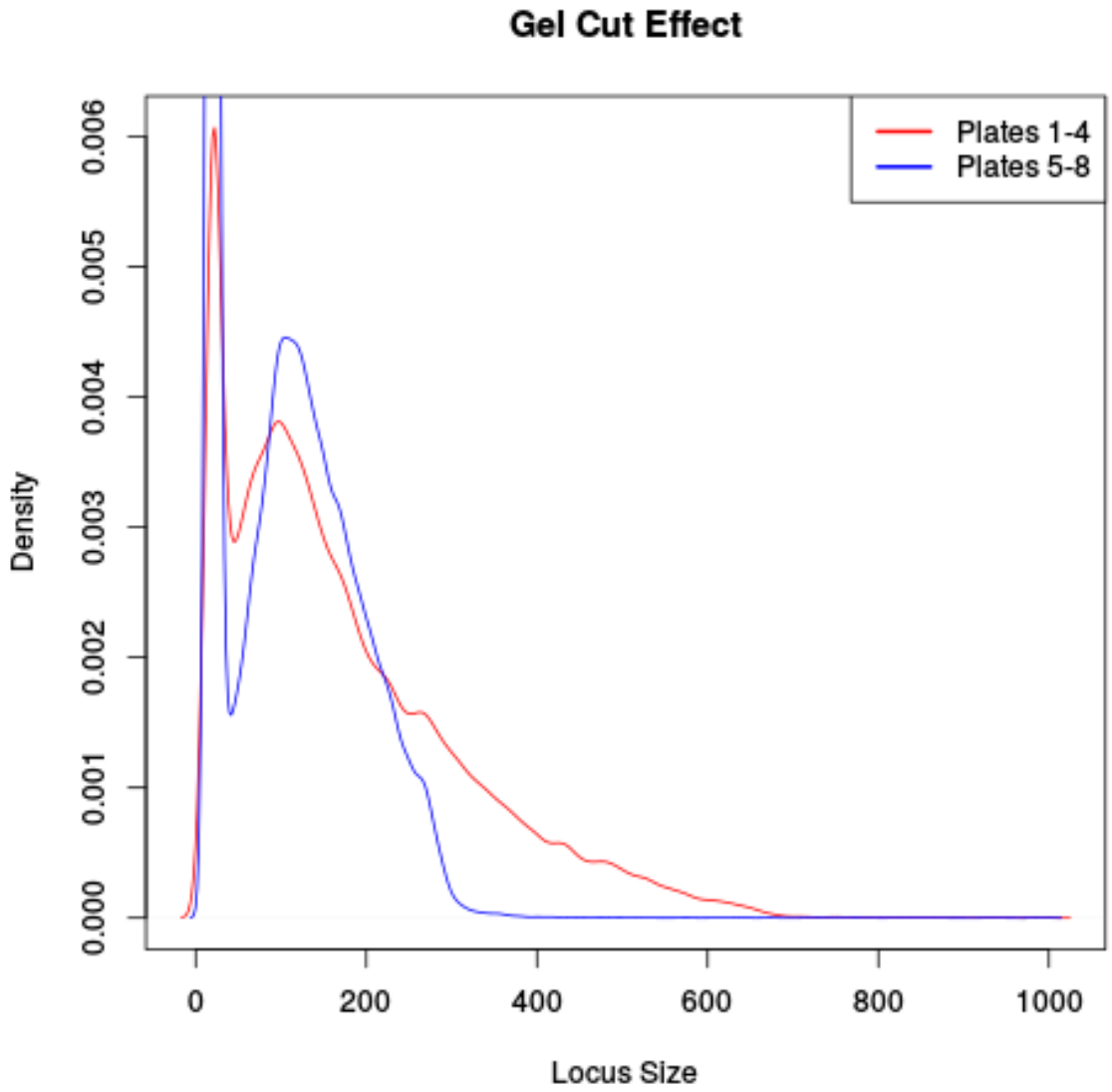
- Full Needleman-Wunsch Global alignment
- Outputs the fragment once if < 190bp



in silico Gel Cut



in silico Gel Cut



in silico Gel Cut

- Thanks to TrimMerge, we know fragment size
- Can digitally select a size range (e.g. 60-150)
- Hopefully will help reproducibility

Drinking the Stacks Kool-Aid

- TASSEL UNEAK
  - is deprecated
  - Seems to to silly things at times
- Stacks
  - Maintained
  - New version handels our read types (allegedly)
  - Seems more efficient computationally
- Lets agree to trial stacks for all *de novo* work

Database-automated Demux

- We have all info need to demux in a DB
- We have a somewhat automated QC pipeline
- Let's join the two:
  - Poll DB for new lanes
  - Automatically backup to archive
  - One command to run demux & QC
  - Automatically generate Stacks script
  - Allow independent analysis directory

Slides I didn't get time to write

But we can talk about now

- Contamination
- Database automated demux
- Common workflow language pipeline