Bioinformatics for transcriptional and genome variation

Kevin Murray
@kdmurray91
kevin@kdmurray.id.au

Borevitz Lab, ANU

RNAseq bioinformatics session, 2015-04-29

Overview

- *Disclaimer: This is a whirlwind tour!*

plant energy biology
ARC CENTRE OF EXCELLENCE

Overview

- *Disclaimer: This is a whirlwind tour!*
- How we do RNAseq
- Experimental design: What we've done, how important it is
- RNAseq analysis pipelines
- DNA analysis pipelines

plant energy biology
ARC CENTRE OF EXCELLENCE

RNAseq in the capital

- We've done a fair bit ($\approx$ 15 lanes, mostly Pogson)
- Largely small-scale, $<$ 50 samples per experiment

plant energy biology
ARC CENTRE OF EXCELLENCE

RNAseq in the capital

- We've done a fair bit ($\approx$ 15 lanes, mostly Pogson)
- Largely small-scale, < 50 samples per experiment
- Developed our own sequence analysis pipelines

RNAseq in the capital

- We've done a fair bit ($\approx$ 15 lanes, mostly Pogson)
- Largely small-scale, < 50 samples per experiment
- Developed our own sequence analysis pipelines
- Wet lab: Mostly NEB/TruSeq kits; few attempts at custom library prep

plant energy biology
ARC CENTRE OF EXCELLENCE

Starting at the start. . .

- **Experimental design is key**

plant energy biology
ARC CENTRE OF EXCELLENCE

Starting at the start. . .

- **Experimental design is key**
- Design of sampling: randomisation at each step

plant energy biology
ARC CENTRE OF EXCELLENCE

Starting at the start. . .

- **Experimental design is key**
- Design of sampling: randomisation at each step
- Speed is essential during collection: responses can be fast

plant energy biology
ARC CENTRE OF EXCELLENCE

Starting at the start. . .

- **Experimental design is key**
- Design of sampling: randomisation at each step
- Speed is essential during collection: responses can be fast
- Replication vs coverage trade-off

plant energy biology
ARC CENTRE OF EXCELLENCE

Starting at the start. . .

- **Experimental design is key**
- Design of sampling: randomisation at each step
- Speed is essential during collection: responses can be fast
- Replication vs coverage trade-off
- Example experiment:
  - Arabidopsis (Col-0)
  - Three dynamic growth conditions
  - Before and after 1000μE light treatment
  - 4 biological reps each group
  - Tissue harvested within 60 seconds of end of stress
  - TruSeq RNASeq kits, 12 samples/lane

Deeper not always better

- Given the same amount of sequencing, go deep or wide?

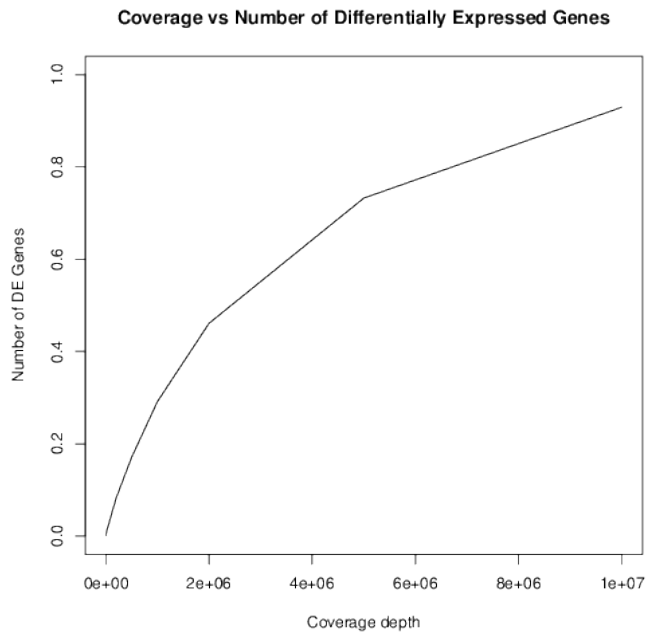
plant energy biology
ARC CENTRE OF EXCELLENCE

Deeper not always better

- Given the same amount of sequencing, go deep or wide?
- For our experiments: wide, more samples, almost always better
- **Your mileage may vary!**

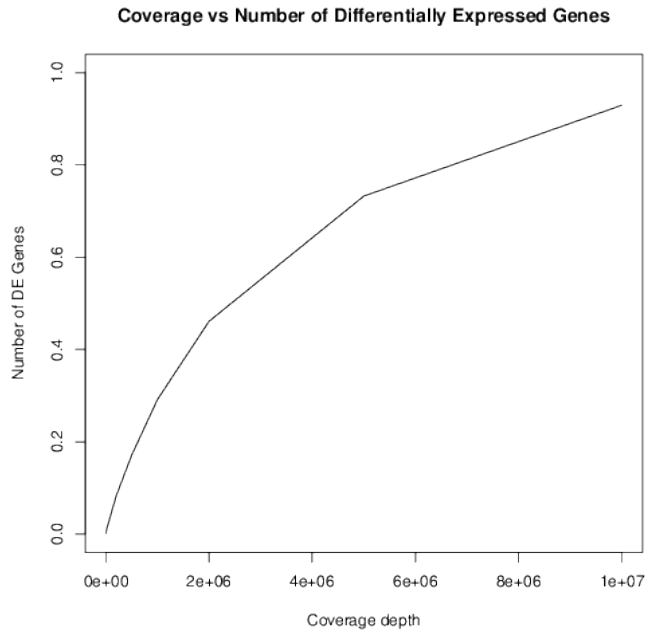
Deeper not always better

- Given the same amount of sequencing, go deep or wide?
- For our experiments: wide, more samples, almost always better
- **Your mileage may vary!**



Deeper not always better

- Given the same amount of sequencing, go deep or wide?
- For our experiments: wide, more samples, almost always better
- **Your mileage may vary!**



- See Kliebenstein, (2012) FIPS: Exploring the Shallow End; Estimating Information Content in Transcriptomics Studies.

plant energy biology
ARC CENTRE OF EXCELLENCE

Sequence analysis

- Data is rawer than with microarrays
  - Needs summarisation to counts: *"sequence analysis"*
  - At large scale, can be a bottleneck

plant energy biology
ARC CENTRE OF EXCELLENCE

Sequence analysis

- Data is rawer than with microarrays
  - Needs summarisation to counts: *"sequence analysis"*
  - At large scale, can be a bottleneck
- Have developed pipelines to do this efficiently

plant energy biology
ARC CENTRE OF EXCELLENCE

Sequence analysis

- Data is rawer than with microarrays
  - Needs summarisation to counts: *"sequence analysis"*
  - At large scale, can be a bottleneck
- Have developed pipelines to do this efficiently
- Many re-analysis runs, not on supercomputer

Sequence analysis

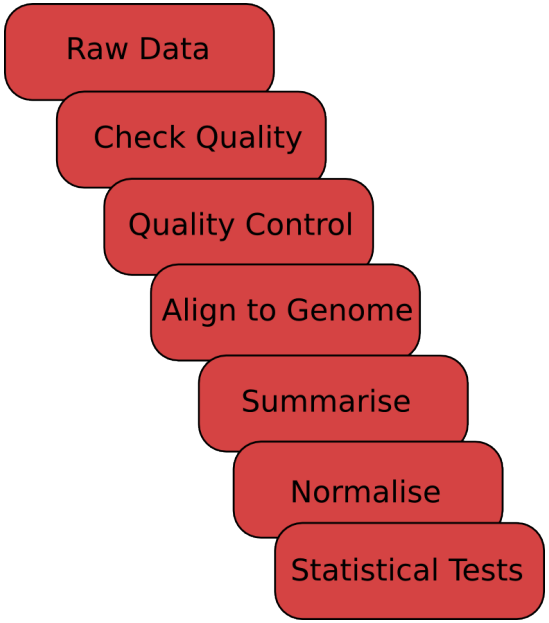
- Data is rawer than with microarrays
  - Needs summarisation to counts: *"sequence analysis"*
  - At large scale, can be a bottleneck
- Have developed pipelines to do this efficiently
- Many re-analysis runs, not on supercomputer
- Faster than others by $> 2 - 3$x, no loss in accuracy

plant energy biology
ARC CENTRE OF EXCELLENCE

Sequence analysis

- Data is rawer than with microarrays
  - Needs summarisation to counts: *"sequence analysis"*
  - At large scale, can be a bottleneck
- Have developed pipelines to do this efficiently
- Many re-analysis runs, not on supercomputer
- Faster than others by $> 2 - 3$x, no loss in accuracy
- Open source, public:
- `https://github.com/kdmurray91/RNAseqPipeline`
  `https://github.com/pedrocrisp/NGS-pipelines`

Sequence analysis pipeline



- `fastqc`
- `scythe`
- `sickle`
- `subread/subjunc`
- `featurecounts`
- `edgeR`
  - `TMM` normalisation
  - `exactTest` or `glmFit`
  - Also using `limma`'s `voom`
- `R` scripts for post-analysis
  - `GOseq`
- Diagnostic plots **highly recommended!**

Robinson & Oshlack (2010); Liao *et al.* (2013a; 2013b); Robinson *et al.* (2013); Young *et al.* (2010)

plant energy biology
ARC CENTRE OF EXCELLENCE

A Change of Pace

- Back to the slow, simple world of DNA for a moment...

Existing DNA variation pipelines

- Genotyping-by-sequencing
  - Reference & *de-novo* analysis
  - Porting to NCI NF
  - Currently manual, working to automate
  - Processed $>$ 5000 samples
- Reference-based genotype calling
  - Pipelines exist
  - Not used a lot, requires deeper coverage
  - See Norman's talk yesterday

Novel algorithms

- My PhD topic

plant energy biology
ARC CENTRE OF EXCELLENCE

Novel algorithms

- My PhD topic
- Our wish-list:
  - Works with "wide and shallow" experiments: e.g. 1000 samples at 1x
  - Reference & alignment free
  - Tolerates any sequencing kind
  - *Free, Open source, Open Data*

plant energy biology
ARC CENTRE OF EXCELLENCE

Novel algorithms

- My PhD topic
- Our wish-list:
  - Works with "wide and shallow" experiments: e.g. 1000 samples at 1x
  - Reference & alignment free
  - Tolerates any sequencing kind
  - *Free, Open source, Open Data*
- e.g: gearing up to sequence 7500+ Eucalyptus, generating over 10 TB **raw** sequence data.

plant energy biology
ARC CENTRE OF EXCELLENCE

Novel algorithms

- My PhD topic
- Our wish-list:
  - Works with "wide and shallow" experiments: e.g. 1000 samples at 1x
  - Reference & alignment free
  - Tolerates any sequencing kind
  - *Free, Open source, Open Data*
- e.g: gearing up to sequence 7500+ Eucalyptus, generating over 10 TB **raw** sequence data.
- *k*-mer analysis: analyse *k*-length words of sequence
  - Fast
  - Constant-memory (with `khmer`)
  - Scaleable (linear time w/ number of samples)
  - Parallelisable (within & across nodes)

plant energy biology
ARC CENTRE OF EXCELLENCE

Thanks

- Borevitz lab (Norman, Justin, Megan, Steve)
- Pogson lab (Pete Crisp)
- Genome Discovery Unit

- Slides at `git.io/vfAof`

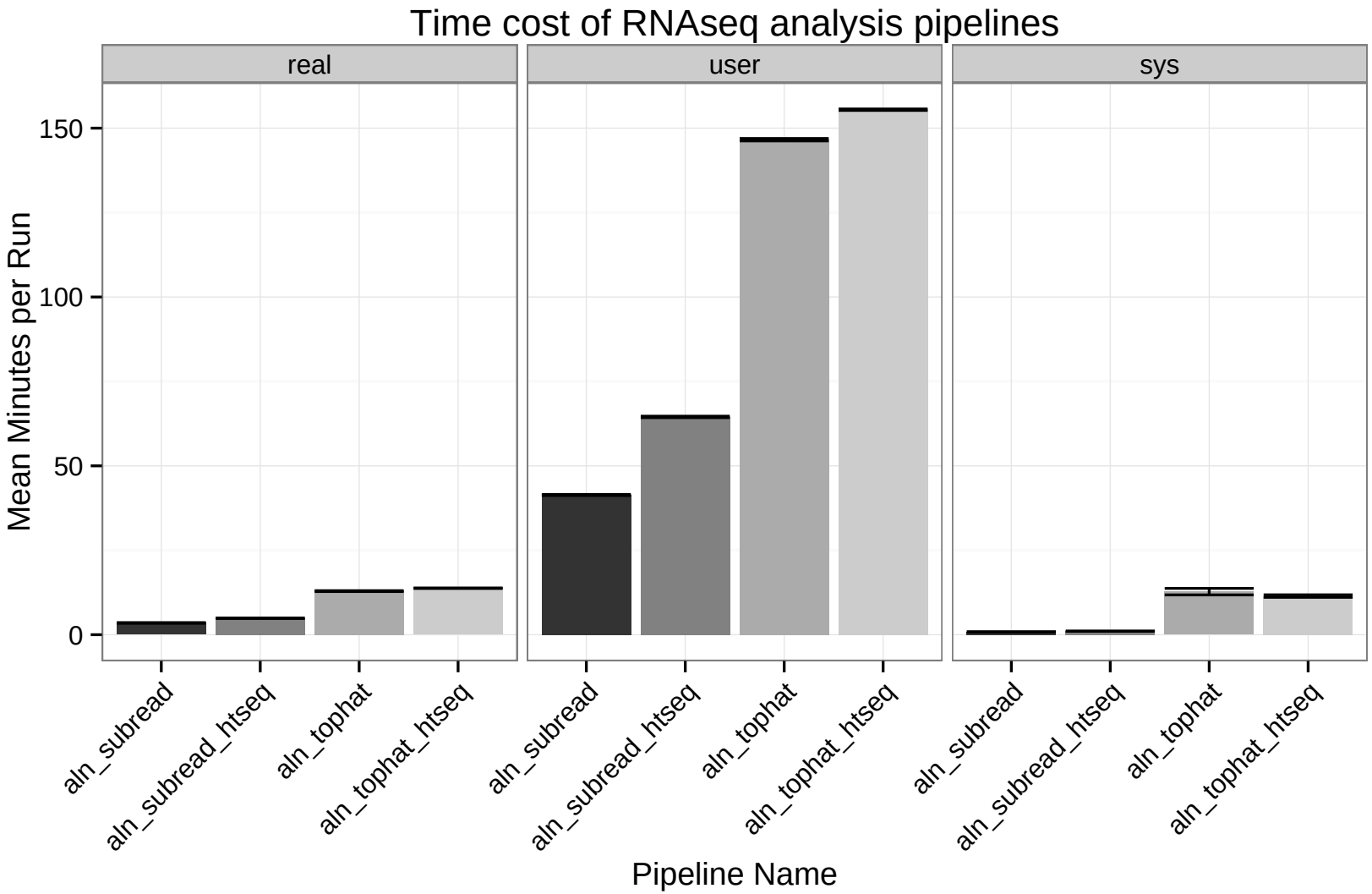
plant energy biology
ARC CENTRE OF EXCELLENCE

Grab-bag of capabilities

- Confirm genotype using RNAseq reads
- Check technical reps are true

Pipeline performance

- Faster than others by $> 2 - 3$x



MDS Plots save time!

- If your reps don't cluster, time to cry into beer.