Lab chat on my PhD

"Novel informatic sequence analysis approaches to dissect complex plant genomes"

April 2, 2015

plant energy biology
ARC CENTRE OF EXCELLENCE

What?

- Novel algorithms to analyse large-scale genomics data

plant energy biology
ARC CENTRE OF EXCELLENCE

What?

- Novel algorithms to analyse large-scale genomics data
- Our wish-list:
  - Reference & alignment free
  - Tolerates any sequencing platform
  - Works with Borevitz-style “wide and shallow” expts: e.g. 1000 samples at 1x

plant energy biology
ARC CENTRE OF EXCELLENCE

What?

- Novel algorithms to analyse large-scale genomics data
- Our wish-list:
  - Reference & alignment free
  - Tolerates any sequencing platform
  - Works with Borevitz-style "wide and shallow" expts: e.g. 1000 samples at 1x
- *"...large data sets and high error rates combine to provide a ...challenge: it is now straightforward to generate data sets that cannot easily be analysed"* — C T Brown *et al.* (2012)

plant energy biology
ARC CENTRE OF EXCELLENCE

What?

- Novel algorithms to analyse large-scale genomics data
- Our wish-list:
  - Reference & alignment free
  - Tolerates any sequencing platform
  - Works with Borevitz-style "wide and shallow" expts: e.g. 1000 samples at 1x
- *"...large data sets and high error rates combine to provide a ...challenge: it is now straightforward to generate data sets that cannot easily be analysed"* — C T Brown *et al.* (2012)
- e.g: Norman & I analysing ≈20TB 3k rice genomes data. Our supercomputer allocation is 30TB.

plant energy biology
ARC CENTRE OF EXCELLENCE

How?

- *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

How?

- *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)
- Multi-layered "zooming" analysis
  - First pass basic clustering
  - Error correction
  - Population graph "alignment"
  - Variant calling

plant energy biology
ARC CENTRE OF EXCELLENCE

How?

- *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)
- Multi-layered "zooming" analysis
  - First pass basic clustering
  - Error correction
  - Population graph "alignment"
  - Variant calling
- In-silico experiment-driven development (new paradigm for me)

plant energy biology
ARC CENTRE OF EXCELLENCE

What have I been up to

- *k*-mer based clustering

What have I been up to

- k-mer based clustering
  - In collaboration w/ Sylvain Foret, Cheng-Soon Ong, Christfried Webers (NICTA)

plant energy biology
ARC CENTRE OF EXCELLENCE

What have I been up to

- k-mer based clustering
  - In collaboration w/ Sylvain Foret, Cheng-Soon Ong, Christfried Webers (NICTA)
  - Extending work in alignment-free sequence comparison (SF) and text/document clustering (C-SO, CW @ NICTA).

plant energy biology
ARC CENTRE OF EXCELLENCE

What have I been up to

- *k*-mer based clustering
  - In collaboration w/ Sylvain Foret, Cheng-Soon Ong, Christfried Webers (NICTA)
  - Extending work in alignment-free sequence comparison (SF) and text/document clustering (C-SO, CW @ NICTA).
  - Using Titus Brown's `khmer` (contributed a lot of code myself)

plant energy biology
ARC CENTRE OF EXCELLENCE

What have I been up to

- *k*-mer based clustering
  - In collaboration w/ Sylvain Foret, Cheng-Soon Ong, Christfried Webers (NICTA)
  - Extending work in alignment-free sequence comparison (SF) and text/document clustering (C-SO, CW @ NICTA).
  - Using Titus Brown's `khmer` (contributed a lot of code myself)
  - Initial results mixed: promising & and bit depressing

plant energy biology
ARC CENTRE OF EXCELLENCE

What have I been up to

- *k*-mer based clustering
  - In collaboration w/ Sylvain Foret, Cheng-Soon Ong, Christfried Webers (NICTA)
  - Extending work in alignment-free sequence comparison (SF) and text/document clustering (C-SO, CW @ NICTA).
  - Using Titus Brown's `khmer` (contributed a lot of code myself)
  - Initial results mixed: promising & and bit depressing
  - New C++ implementation: Can compute 100x100 matrix in 10 mins on 16 CPUs
  - Hashing takes $\approx$ 10mins, uses $\approx$ 1GB/sample.

plant energy biology
ARC CENTRE OF EXCELLENCE

What have I been up to

- *k*-mer based clustering
  - In collaboration w/ Sylvain Foret, Cheng-Soon Ong, Christfried Webers (NICTA)
  - Extending work in alignment-free sequence comparison (SF) and text/document clustering (C-SO, CW @ NICTA).
  - Using Titus Brown's `khmer` (contributed a lot of code myself)
  - Initial results mixed: promising & and bit depressing
  - New C++ implementation: Can compute 100x100 matrix in 10 mins on 16 CPUs
  - Hashing takes $\approx$ 10mins, uses $\approx$ 1GB/sample.
  - Norman, Sylvain, Cheng-Soon and Chris working on new metrics when SF & I are back