Lab chat on my PhD

“Novel algorithms for population-scale analysis of complex plant genomes”

July 2, 2015

What?

- Novel algorithms to analyse large-scale genomics data

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

How?

- *k*-mer analysis: analyse *k*-length words of sequence
  - Fast
  - Constant-memory (with `khmer`)
  - Scaleable and parallelisable

How?

- *k*-mer analysis: analyse *k*-length words of sequence
  - Fast
  - Constant-memory (with `khmer`)
  - Scaleable and parallelisable
- Multi-layered “zooming” analysis
  - First pass basic clustering
  - Error correction
  - Population graph “alignment”
  - Variant calling

How?

- *k*-mer analysis: analyse *k*-length words of sequence
  - Fast
  - Constant-memory (with `khmer`)
  - Scaleable and parallelisable
- Multi-layered "zooming" analysis
  - First pass basic clustering
  - Error correction
  - Population graph "alignment"
  - Variant calling
- In-silico experiment-driven development

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)

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).

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).
- Have functioning software package: `kWIP`

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).
- Have functioning software package: `kWIP`
- Using Titus Brown's `khmer` (contributed a lot of code myself)

kWIP

- The k-mer Weighted Inner Product

kWIP

- The k-mer Weighted Inner Product
- Algorithm:
  - For each sample: count all k-mers (k=20) into a hash

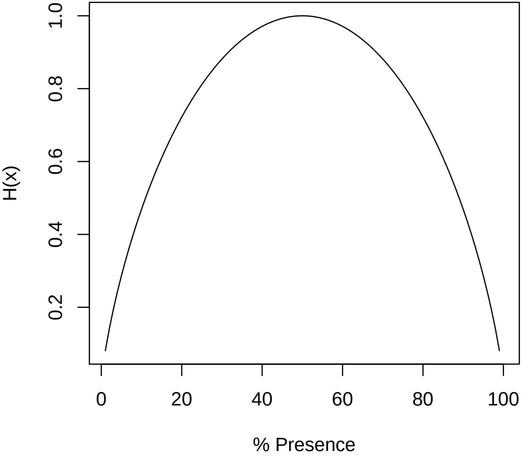
kWIP

- The k-mer Weighted Inner Product
- Algorithm:
  - For each sample: count all k-mers (k=20) into a hash
  - For each analysis set, a.k.a "population":
    - Calculate the informational entropy of hash bins (vector P)
    - For each pair of hashes A and B, calculate $A \cdot B \cdot P$

kWIP

- The k-mer Weighted Inner Product
- Algorithm:
    - For each sample: count all k-mers (k=20) into a hash
    - For each analysis set, a.k.a "population":
        - Calculate the informational entropy of hash bins (vector P)
        - For each pair of hashes A and B, calculate $A \cdot B \cdot P$

Shannon Entropy



kWIP

- The software:
  - C++, 2000 SLOC
  - Depends on `khmer`
  - Parallelised, ≈ 12 hrs for 96 rice samples.

kWIP

- The software:
  - C++, 2000 SLOC
  - Depends on `khmer`
  - Parallelised, $\approx$ 12 hrs for 96 rice samples.
- The paper:
  - Coming soon, planning to have it done by August
  - Involves many in-silico experiments

Rice Experiment

- 3000 rice lines, 25k sequence runs, 20TB data
- Analysing in sets of 96, from two major groups
- Looks very accurate, detect Basmatia as Jap, strange samples.

Drosophila

- Several read technologies
- Several species, population, reps
- Detect failed samples, repoduced known tree

Simulation??

- Need to think about simulation
- Time consuming to do well, but can make lots of data
- Can we do it somewhat dodgy?