kWIP: The k-mer Weighted Inner Product

Kevin Murray

PhD Candidate, Borevitz Lab, CPEB, ANU

26 October 2015

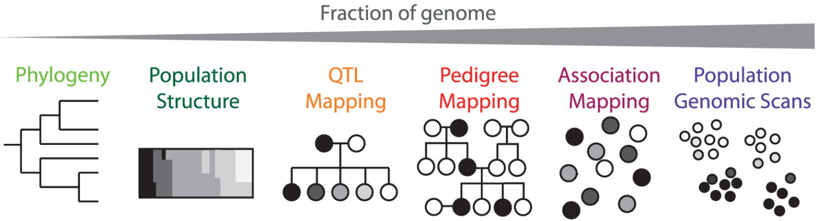
Overview

Large-scale population genomics

- Moving from 100s to 1,000s or 10,000s of samples *per PhD!*

Large-scale population genomics

- Moving from 100s to 1,000s or 10,000s of samples *per PhD!*
- Efficient algorithms to analyse large-scale genomic data
  - Reference & alignment free: *less bias, de novo*
  - Platform/protocol agnostic: *future proof*
  - Computationally efficient: *not the bottleneck*



Fraction of genome

Phylogeny

Population Structure

QTL Mapping

Pedigree Mapping

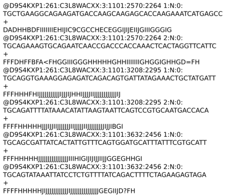
Association Mapping

Population Genomic Scans

after Peterson *et al.* [1]

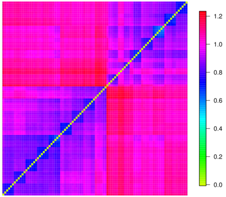
Presenting `kWIP`

- *k*-mer based *de novo* estimate of genetic similarity
- Produces a distance matrix from raw NGS reads

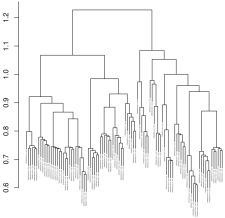


```
@D9S4KXP1:261:C3L8WACXX:3:1101:2570:2264 1:N:0:
TGCTGAAGGCAGAAGATGACCAAGCAAGAGCACCAAGAAATCATGAGCC
+
DADHHBDFIIIIIIIIEHIJIC9CGCCHECEGGIJIJEIIJGIIIGGGIG
@D9S4KXP1:261:C3L8WACXX:3:1101:2570:2264 2:N:0:
TGCAGAAAGTGCAGAATCAACCGACCCACCAAACTCACTAGGTTCATTC
+
FFFDHFFBFA<FHGGIIIGGGHHHHHGHHIIIIIIIIGHGGIGHHGD=FH
@D9S4KXP1:261:C3L8WACXX:3:1101:3208:2295 1:N:0:
TGCAGGTGAAAGGAGAGATCAGACAGTGATTATAGAAACTGCTATGATT
+
FFFHHHFHIJJJJJJJJJJJIJJJJIJHHIJJJJIIJJJJJJJJJJJJIJ
@D9S4KXP1:261:C3L8WACXX:3:1101:3208:2295 2:N:0:
TGCAGATTTTATAAACATATTAAGTAATTCAGTCCGTGCAATGACCACA
+
FFFFHHHHHJJJIJJIIJJJJJJJIIJJJJJIIJJJJJJJJJJIJJIBGI
@D9S4KXP1:261:C3L8WACXX:3:1101:3632:2456 1:N:0:
TGCAGCGATTATCACTATTGTTTCAGTGGATGCATTTATTTCGTGCATT
+
FFFHHHHHJJJJJJJJJJJJJJJJJIIIHIGIJIJIJJIIJJGGEGHHGI
@D9S4KXP1:261:C3L8WACXX:3:1101:3632:2456 2:N:0:
TGCAGTATAAATTATCCTCTGTTTTATCAGACTTTTCTAGAAGAGTAGA
+
FFFFHHHHHJIJJJJJJJJJJJIJJJJJJJJJJJJJJJJJGEGIIJD?FH
```

→



→

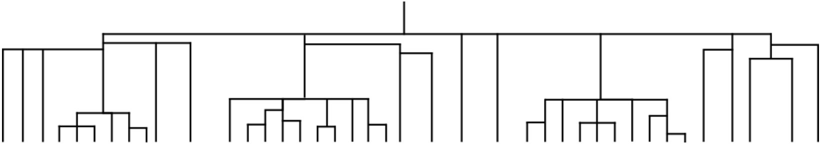


Why Estimate Similarity?

- Rough approximation of sample relatedness required
  - For natural collections
  - As a technical control

Why Estimate Similarity?

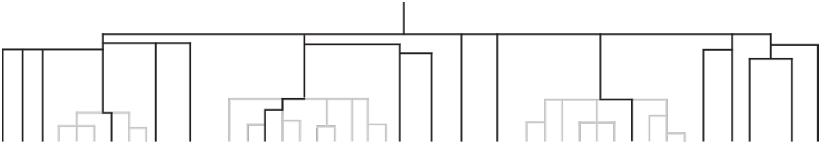
- Rough approximation of sample relatedness required
  - For natural collections
  - As a technical control



after Brachi *et al.* [2]

Why Estimate Similarity?

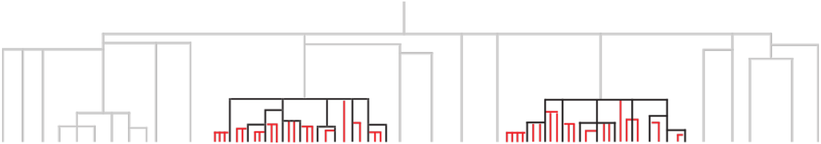
- Rough approximation of sample relatedness required
  - For natural collections
  - As a technical control



after Brachi *et al.* [2]

Why Estimate Similarity?

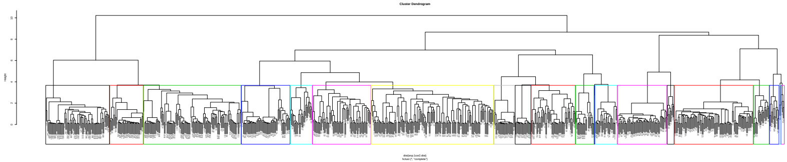
- Rough approximation of sample relatedness required
  - For natural collections
  - As a technical control



after Brachi *et al.* [2]

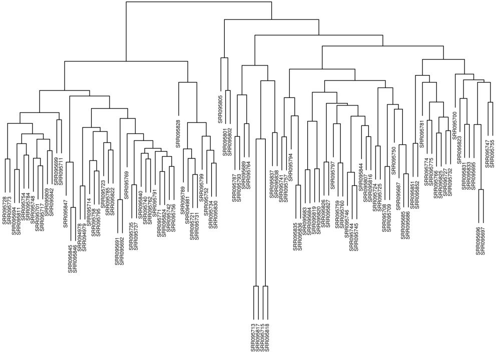
Why Estimate Similarity?

- Rough approximation of sample relatedness required
  - For natural collections
  - As a technical control



Why Estimate Similarity?

- Rough approximation of sample relatedness required
  - For natural collections
  - As a technical control



Technological Overview

- k-mer counting (bag-of-words)
    - Decompose sequences to overlapping words of k

Technological Overview

- k-mer counting (bag-of-words)
  - Decompose sequences to overlapping words of k
- Hashing and Probabilistic Data Structures
  - Efficient storage & compute of “bag of words”

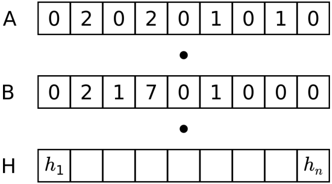
Technological Overview

- k-mer counting (bag-of-words)
  - Decompose sequences to overlapping words of k
- Hashing and Probabilistic Data Structures
  - Efficient storage & compute of “bag of words”
- (Weighted) Inner Products
  - Similarity metric weighted by Shannon entropy

kWIP Algorithm

- For each run: count all k-mers into a Hash
- For each analysis:
  - Calculate the entropy of population frequency (H)
  - For each pair of runs A and B, calculate

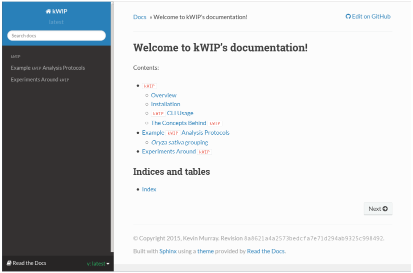
    $$\sum_{i=0}^{n} A_i \cdot B_i \cdot H_i$$



A	0	2	0	2	0	1	0	1	0
					•				
B	0	2	1	7	0	1	0	0	0
					•				
H	h_1								h_n

kWIP

- The software:
  - `C++`, >2000 lines of code
  - Uses **`khmer`** for k-mer counting & hashing
  - Parallelised, fast
  - GNU GPL licensed, source code on GitHub



Welcome to kWIP's documentation!

Contents:

- `kWIP`
  - Overview
  - Installation
  - `kWIP` CLI Usage
  - The Concepts Behind `kWIP`
- Example `kWIP` Analysis Protocols
  - *Oryza sativa* grouping
- Experiments Around `kWIP`

Indices and tables

- Index

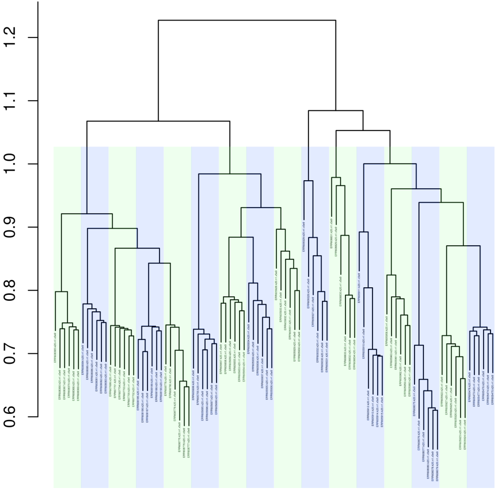
© Copyright 2015, Kevin Murray. Revision `8a8621a4a2573bedcfa7e71d294ab9325c998492`.

Built with Sphinx using a theme provided by Read the Docs.

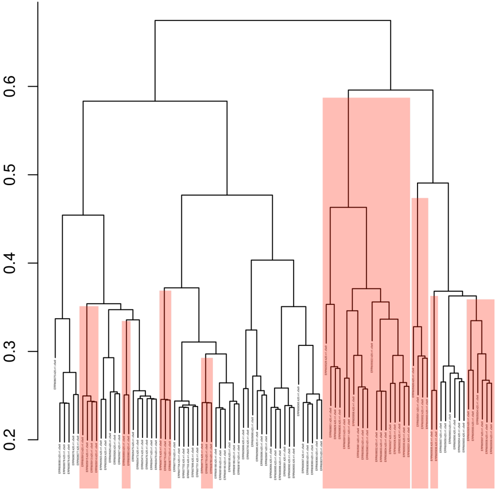
Demonstration

- Set of 96 rice runs from 16 samples (6 tech reps ea)
- Expectations:
  - All runs cluster into groups of 6 reps (16 samples)
  - Big split between Indica & Japonica: (7 and 9 respectively here)
- We see this with `kWIP`, not with Unweighted IP
- Took 8 hours on 16 CPU Raijin node, 60-80GB RAM

WIP



IP



Conclusions

- `kWIP` is implemented, β software
- A few users in the audience (thanks!)
- Further simulations and experiments required
- Publication in preparation

Thanks

- Supervisors: J Borevitz, S Forêt, G Huttley and B Pogson
- Christfried Webers, Cheng Soon Ong, Norman Warthmann
- `khmer` folks: C. Titus Brown, Michael Crusoe, Camille Scott (DIB-lab) @ UC Davis
- Beta testers
- Yourselves

References

Peterson, B. K. *et al.* Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE* **7,** e37135 (2012).

Brachi, B., Morris, G. P. & Borevitz, J. O. Genome-wide association studies in plants: the missing heritability is in the field. *Genome biology* **12,** 232 (2011).

.