

# kWIP: The k-mer Weighted Inner Product

And the rest of my PhD...

Kevin Murray

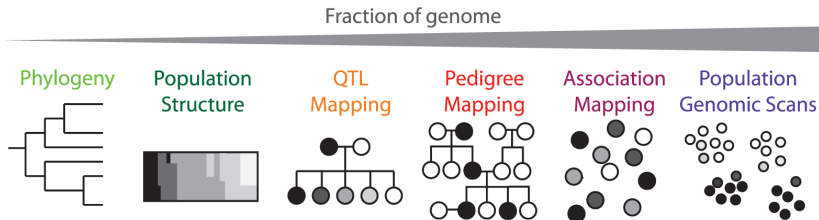
Borevitz Lab, CPEB, ANU

April 15, 2016



# Large-scale population genomics

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

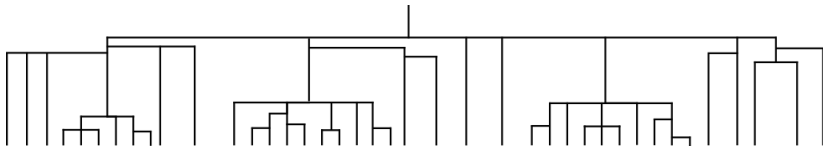


after Peterson *et al.* [1]



# Genetic Similarity Estimation

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

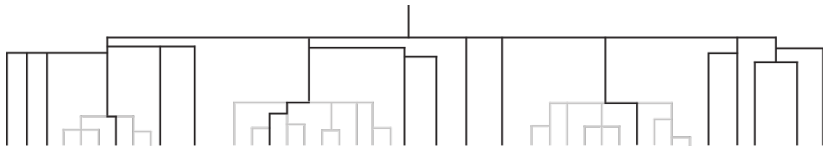


after Brachi *et al.* [2]



# Genetic Similarity Estimation

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

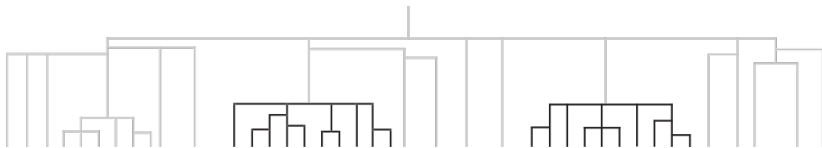


after Brachi *et al.* [2]



# Genetic Similarity Estimation

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

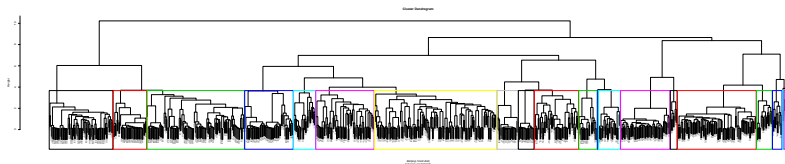


after Brachi *et al.* [2]



# Genetic Similarity Estimation

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







# Genetic Similarity Estimation

Initially, we care mostly about the deepest and shallowest branches of the tree.

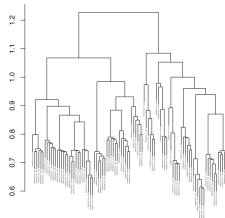
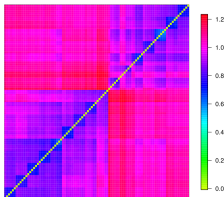




# Presenting kWIP

- ▶  $k$ -mer based *de novo* genetic relatedness estimator
- ▶ Produces a distance matrix from raw NGS reads
- ▶ Uses Weighted Inner Product between  $k$ -mer counts

```
@D954KXP1.261.C3L8WACXX.3:1101:2570:2264 1:N:0:
TGCTGAAGGCAGAAGATGACCAAGCAAGAGCAAGAAATCATGAGCC
+
DADHHBDFIIIIIEHIIIC9CGCCECEGGIIIEIIGIIGGGGG
@D954KXP1.261.C3L8WACXX.3:1101:2570:2264 2:N:0:
TGCAGAAAGTGCAGAAATCAACCGACCCACCAAACTACTAGGTTCATTC
+
FFFDHFFBFA<FHGGIIIGGGHHHHHHHHHHHHHHGGGGHHHGD=FH
@D954KXP1.261.C3L8WACXX.3:1101:3208:2295 1:N:0:
TGCAGGTGAAGGAGAGATGACAGGTGATTATAGAACTGCTATGATT
+
FFFHFFHHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@D954KXP1.261.C3L8WACXX.3:1101:3208:2295 2:N:0:
TGCAGATTTTATAAAGATTAAGTAATTACGTCCTGCAATGACCACA
+
FFFFHHHHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@D954KXP1.261.C3L8WACXX.3:1101:3632:2456 1:N:0:
TGCAGCGATTATCATCTATTGTTTCAGTGAGTGATTATTCTTGTCATT
+
FFFFHHHHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@D954KXP1.261.C3L8WACXX.3:1101:3632:2456 2:N:0:
TGCAGTATAAATTCCTCTGTTTATCAGACTTTCTAGAAAGATAGA
+
FFFFHHHHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
GEGIIID7FH
```





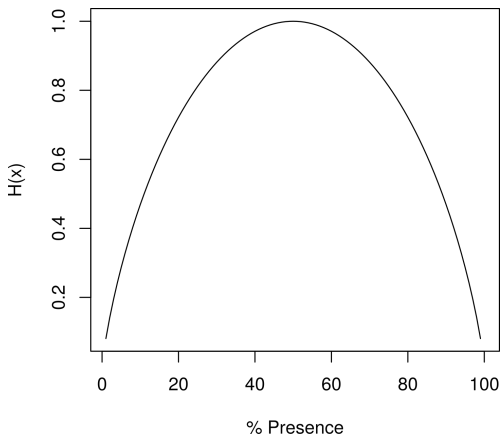
# $k$ -mer Sequence Comparison

- ▶ Many existing tools
  - ▶ *D2* and related statistics
  - ▶ Early steps in many sequence aligners
  - ▶ **spaced** and other spaced-word approaches<sup>3,4</sup>
  - ▶ **Cnidaria** and other Jaccard distance approaches<sup>5</sup>
  - ▶ **mash** and other MinHash approaches<sup>6</sup>
- ▶ Most require assembled gene/genome sequence
- ▶ Most target deeper relationships
- ▶ Many use inner product between  $k$ -mer counts
- ▶ kWIP extends these tools:
  - ▶ No assembly required
  - ▶ Weights inner product to improve accuracy



# Entropy Weighting

- ▶ kWIP weights by Shannon entropy:  $H(\text{frequency})$
- ▶ Shannon entropy: measure of information
- ▶  $-(P_i \log_2(P_i) + (1 - P_i) \log_2(1 - P_i))$





# kWIP Algorithm

- ▶ For each run: count all  $k$ -mers probabilistically
- ▶ For each analysis set:
  - ▶ Calculate the entropy of  $k$ -mer frequency ( $H$ )
  - ▶ For each pair of runs with  $k$ -mer counts  $A$  and  $B$ , calculate

$$\sum_{i=1}^n A_i \cdot B_i \cdot H_i$$

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++11,  $\approx 2000$  lines of code
  - ▶ Uses **khmer** for  $k$ -mer counting
  - ▶ GNU GPL licensed, source code on GitHub
  - ▶ Precompiled binaries provided

The screenshot shows the kWIP documentation website. The left sidebar is dark blue/black with a search bar and navigation links. The main content area is white and features the following elements:

- Header:** kWIP latest, Edit on GitHub
- Welcome message:** 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
- Footer:** © Copyright 2015, Kevin Murray. Revision 8a8621a4a2573bedcfa7e71d294ab9325c998492. Built with Sphinx using a theme provided by Read the Docs.



# kWIP Case Studies

- ▶ 3000 rice genomes
  - ▶ 3000 rice samples (25k runs)
  - ▶ The 3,000 rice genomes project [7]
- ▶ Chlamydomonas
  - ▶  $\approx 20$  lines from USA
  - ▶ Flowers *et al.* [8]
- ▶ Simulation
  - ▶ Fake population genome sequencing studies

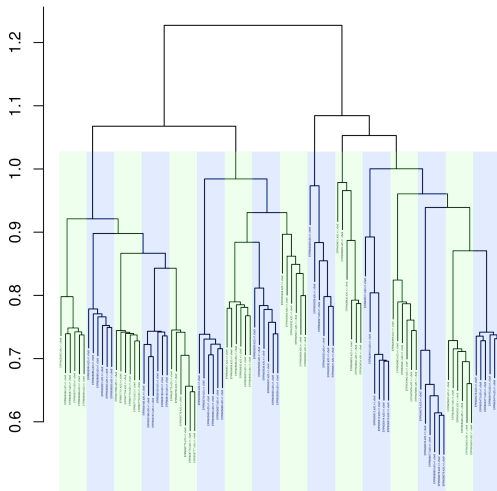


## 96 Rice Runs

- ▶ Set of 96 rice runs from 16 samples (6 tech reps ea)
- ▶ About half/half from 2 major groups (Indica, Japonica)
- ▶ Expectations:
  - ▶ All runs cluster into groups of 6 reps (16 samples)
  - ▶ Big split between two groups: (7 and 9 respectively here)
- ▶ Recover known grouping w/ kWIP, not w/ unweighted IP
- ▶ Sensitive to read depth
- ▶ Took 6 hours on 16 CPU, 64GB RAM supercomputer node



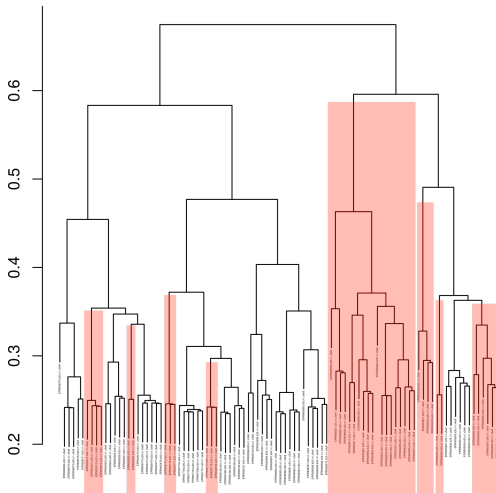
# WIP







## IP



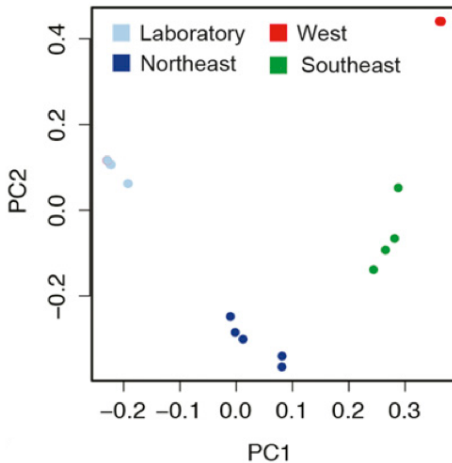


# Chlamydomonas

- ▶ High coverage re-sequencing with leftover assembly
  - ▶ Map to reference
  - ▶ Assemble missing sample genome from leftover reads
  - ▶ Map again to reference + leftovers
  - ▶ Call variants
  - ▶ Calculate distance
- ▶ Compare kWIP to SNP-based distance calculation
  - ▶ Compare PCA visualisation of each



# Chlamydomonas



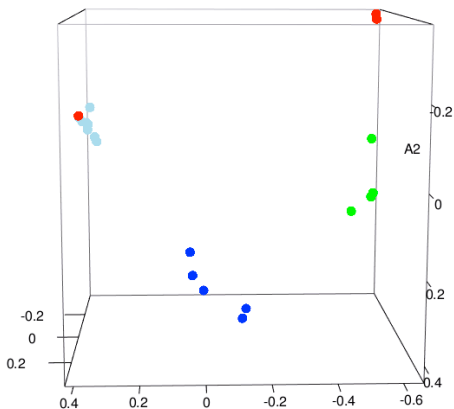
“Sample CC-4414 (red) is hidden behind the cluster of laboratory strains (light blue)”

## Population structure of *Chlamydomonas* in USA

Data from Flowers *et al.* [8]



# Chlamydomonas



Population structure of *Chlamydomonas* in USA

Data from Flowers *et al.* [8]

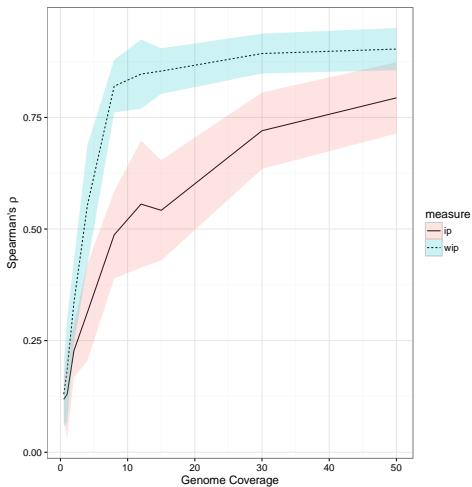


# Simulation

- ▶ Perform simulated sequencing experiment:
  - ▶ Simulate natural population structure
  - ▶ Simulate sample genomes, sequencing runs
  - ▶ Hash reads, kWIP
  - ▶ Compare known truth to kWIP results  
(with Spearman's Rank Correlation,  $\rho$ )
- ▶ kWIP quantitatively outperforms unweighted equivalent
  - ▶ Effect of coverage on accuracy
  - ▶ Accuracy across scale of variation



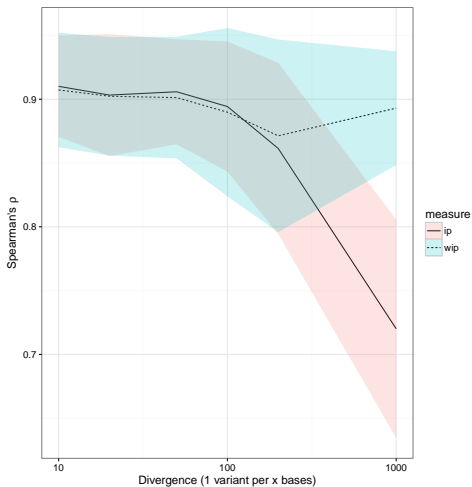
# Simulation Results



Coverage vs Accuracy



# Simulation Results



Average variation vs Accuracy



## kWIP Summary

- ▶ kWIP is implemented, no known bugs
- ▶ Publicly available at [github.com/kdmurray91/kwip](https://github.com/kdmurray91/kwip)
- ▶ We show the utility of kWIP
- ▶ Publication coming soon





## What's next?

- ▶ Finish kWIP paper
- ▶ Complete & Publish GBS analysis tools
  - ▶ GBS is a quick & cheap sequencing method
  - ▶ Have written improved analysis tools
  - ▶ These need publication



# What's next?

- ▶ Finish kWIP paper
- ▶ Complete & Publish GBS analysis tools
  - ▶ GBS is a quick & cheap sequencing method
  - ▶ Have written improved analysis tools
  - ▶ These need publication
- ▶ Further  $k$ -mer analysis methods
  - ▶ Marrying assembly and  $k$ -mer comparison concepts
  - ▶ Pan-genome representations
- ▶ *Eucalyptus* population genomics
  - ▶ Collaboration with Rose Andrew @ UNE
  - ▶ Extending a pilot study in a hybrid population of *E. sideroxylon* and *E. albens*.



# Thanks

- ▶ My kWIP collaborators
  - ▶ Norman Warthmann, Christfried Webers, Cheng Soon Ong, Sylvain Forêt
- ▶ Supervisors:
  - ▶ Justin Borevitz, Gavin Huttley, Barry Pogson, Sylvain Forêt
- ▶ **khmer** folks (DIB-lab, UC Davis)
- ▶ Yourselfes



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).



Morgenstern, B. *et al.* Estimating evolutionary distances between genomic sequences from spaced-word matches. *Algorithms for Molecular Biology* **10**, 5 (2015).



Leimeister, C.-A. *et al.* Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, btu177 (2014).



Aflitos, S. A. *et al.* Cnidaria: fast, reference-free clustering of raw and assembled genome and transcriptome NGS data. *BMC Bioinformatics* **16**, 352 (2015).



Ondov, B. D. *et al.* Fast genome and metagenome distance estimation using MinHash. *bioRxiv*, 029827 (2015).



The 3,000 rice genomes project. The 3,000 rice genomes project. *GigaScience* **3**, 7 (2014).



Flowers, J. M. *et al.* Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga *Chlamydomonas reinhardtii*. *The Plant Cell* **27**, 2353–2369 (2015).