kWIP: The k-mer Weighted Inner Product

Estimating genetic similarity of sequencing runs

Kevin Murray

PhD Candidate
Borevitz Lab, ANU

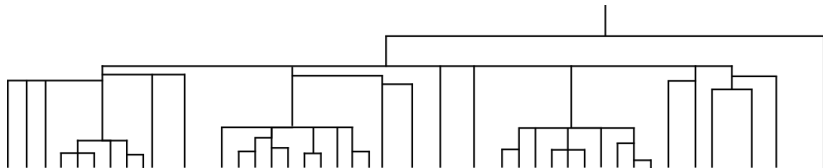2016-11-02

Collection Re-structuring

- Collect 100s or 1,000s of natural samples
- "First look" at genetic relatedness
  - Assert replicates cluster, detect mixups
  - Carry best samples to detailed analysis

after Brachi *et al.* [1]

Collection Re-structuring

- Collect 100s or 1,000s of natural samples
- "First look" at genetic relatedness
  - Assert replicates cluster, detect mixups
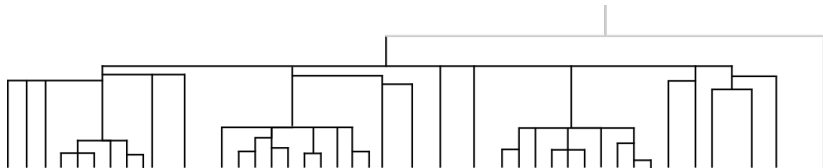 - Carry best samples to detailed analysis

after Brachi *et al.* [1]

Collection Re-structuring

- Collect 100s or 1,000s of natural samples
- "First look" at genetic relatedness
  - Assert replicates cluster, detect mixups
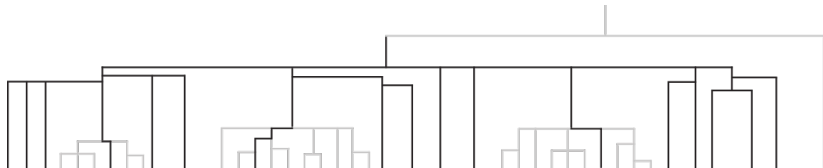 - Carry best samples to detailed analysis

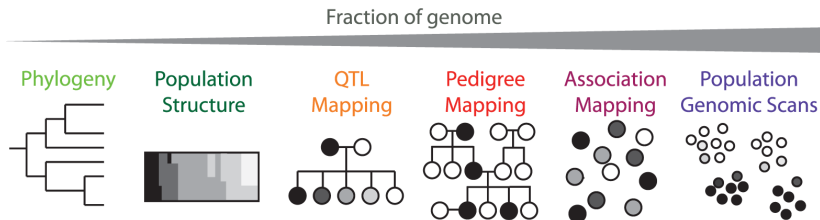

after Brachi *et al.* [1]

(Initial) Genetic Similarity Estimation

- Initial genetic analyses inspect
  - Outgroups (widest relationships)
  - Replicates
  - Mix-ups
  - Broad groupings
- Current genetic similarity, *not evolutionary history*

Fraction of genome

Phylogeny	Population Structure	QTL Mapping	Pedigree Mapping	Association Mapping	Population Genomic Scans

after Peterson *et al.* [2]

- 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*

Alignment-free Sequence Comparison

- Many existing metrics, and tools
  - $D2$ and related statistics
  - `spaced` and other spaced-word approaches[3,4]
  - `Cnidaria` and other Jaccard index approaches[5]
  - `mash` and other MinHash approaches[6]
- Most require assembled gene/genome sequence
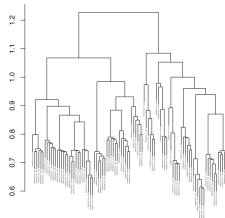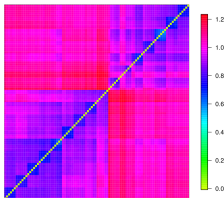- **Most assume evolutionary history between samples**

"You ever have one of those days where you just don't feel like aligning?"

Presenting `kWIP`

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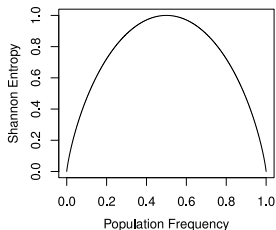- Uses Weighted Inner Product between k-mer counts

kWIP Algorithm

- Count each sample's k-mers probabilistically (`khmer`)
- Calculate information content of each k-mer
- Compute each pairwise distance using weighed inner product (WIP)

kWIP – Software

- Uses `khmer` for k-mer counting
- Parallelised with OpenMP
- GNU GPL licensed, `C++11` source code on GitHub
- Precompiled binaries provided
- Documentation & tutorials online

kWIP Case Studies

- Rice genomes project[7]
  - 3000 rice varieties (25k runs)
  - $\approx$ 2-fold sequencing per run
- Chlamydomonas[8]
  - High-coverage sequencing of 20 wild & lab strains
- Rice root-associated microbiome metagenomics
  - Shotgun sequencing of root-soil interface
- Simulation
  - Fake population genome sequencing studies

Replicate and Subspecies Clustering

- Set of 96 rice runs from 16 samples, 6 tech reps.
- $\approx$ 3-fold sequencing per run
- Expectations:
  - All runs cluster into samples of 6 reps
  - Big split between 2 major groups (Indica, Japonica)
- Recover known grouping w/ `kWIP`, not w/ unweighted IP
- Took 6 hours on 16 CPU, 64GB RAM supercomputer node

WIP
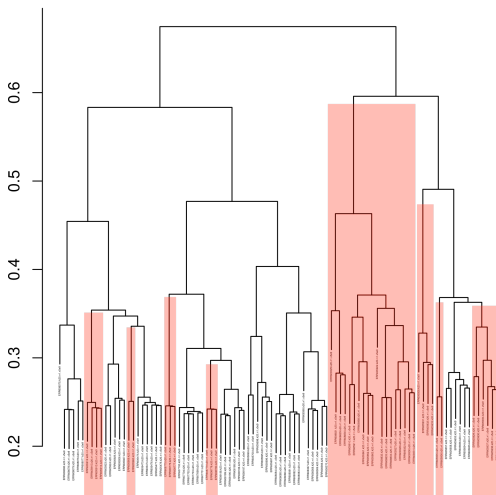
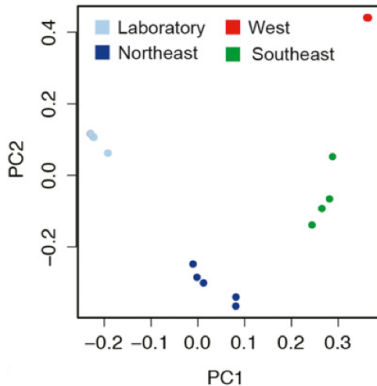# IP
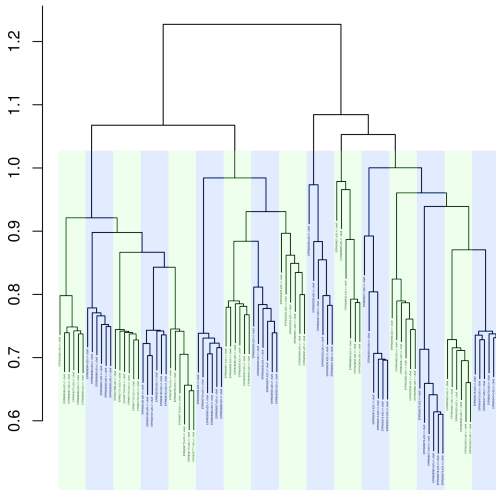
- Avoid reference bias with "leftover assembly"[8]
  - Sequence *very* deep (> 200x)
  - Map to reference
  - Assemble umapped reads
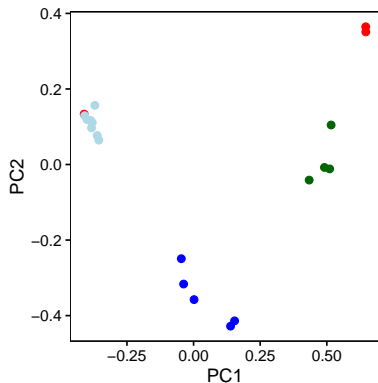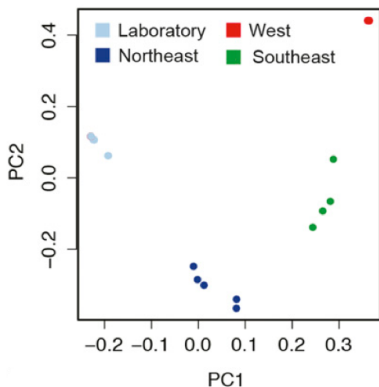 - Map to reference + leftovers
  - Call variants
  - `SNPrelate` + PCA

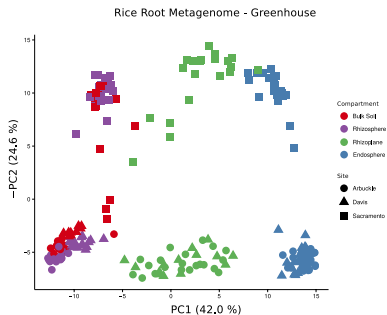*"Sample CC-4414 (red) is hidden behind the cluster of laboratory strains (light blue)"*

Chlamydomonas – with `kWIP`

- Download SRA files
- Count k-mers
- Run `kWIP`

Metagenomes
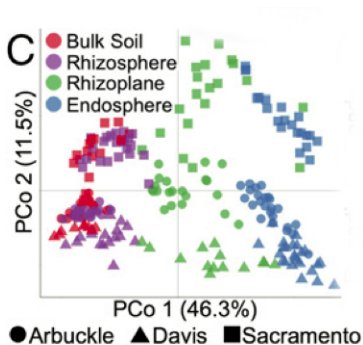
Simulation

- Perform simulated sequencing experiment (50 times):
  - Simulate natural population structure
  - Simulate sample genomes
  - Simulate sequencing runs (with random variation)
  - Sketch reads, `kWIP`
  - Compare `kWIP` results to known truth
  - Spearmans Rank Correlation (ρ), "Performance"
- kWIP quantitatively outperforms unweighted equivalent
  - Performs reasonably at low-moderate coverage
  - Performance stable across scale of variation
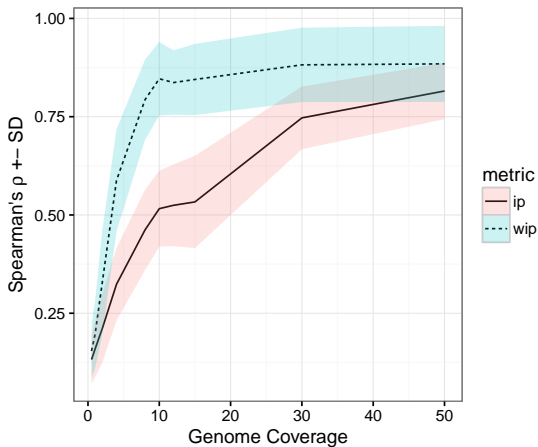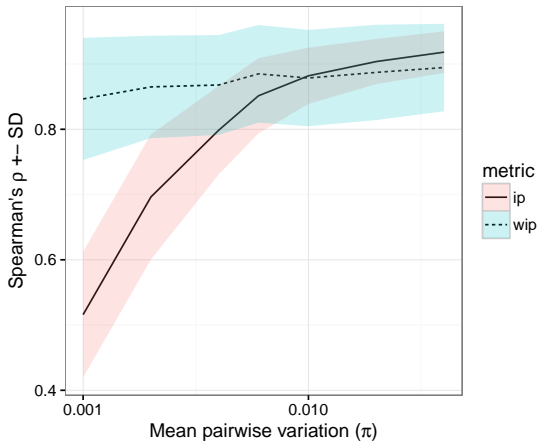
kWIP Summary

- `kWIP` is implemented, no known bugs
- Publicly available at `github.com/kdmurray91/kwip`
- We show the utility of `kWIP`
- Publication in review at PLoS Comp. Biol. (`bit.do/kwip`)
- Version 2 on the way
  - MPI parallel
  - More metrics
  - Even faster

Thanks

- Norman Warthmann, Justin Borevitz
- Christfried Webers, Cheng Soon Ong
- Sylvain Forêt
- AB^3ACBS Organisers and Yourselves

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).

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).

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).

.