

kWIP: The k-mer Weighted Inner Product

Kevin Murray

Borevitz Lab, CPEB, ANU

21 August 2015



Disclaimer

DRAFT



Disclaimer

DRAFT

Or as Norman says, kWIP is Kevin's Work In Progress



Collaboration

- ▶ This work is a collaborative effort
 - ▶ Norman Warthmann
 - ▶ Cheng Soon Ong
 - ▶ Chris Webers



Overview

- ▶ Motivation
- ▶ Technological overview
- ▶ Early results and plans
- ▶ Demonstration



Large-scale population genomics

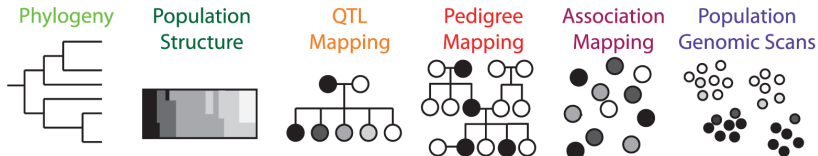
- ▶ Moving from 100s to 1000s and 10000s of samples *per PhD!*



Large-scale population genomics

- ▶ Moving from 100s to 1000s and 10000s 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*
 - ▶ Cross scale: *one tool to rule them all*

Fraction of genome



after Peterson *et al.* [1]



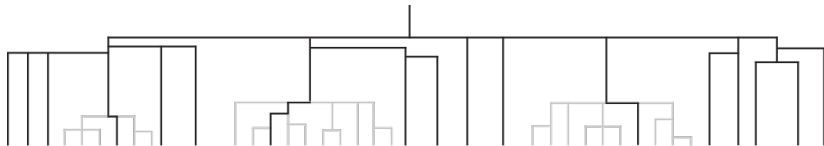
Rapid and Basic Clustering

- ▶ Rough approximation of sample relatedness required
- ▶ Enables a “zooming” approach to genetic analysis
- ▶ Required for population resampling



Rapid and Basic Clustering

- ▶ Rough approximation of sample relatedness required
- ▶ Enables a “zooming” approach to genetic analysis
- ▶ Required for population resampling

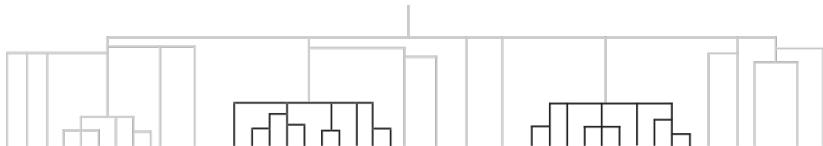


after Brachi *et al.* [2]



Rapid and Basic Clustering

- ▶ Rough approximation of sample relatedness required
- ▶ Enables a “zooming” approach to genetic analysis
- ▶ Required for population resampling

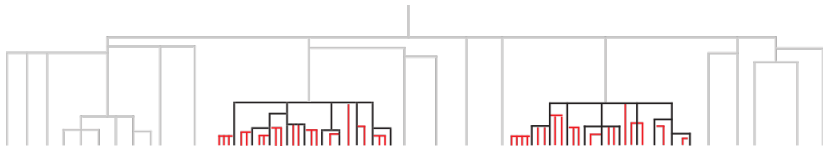


after Brachi *et al.* [2]



Rapid and Basic Clustering

- ▶ Rough approximation of sample relatedness required
- ▶ Enables a “zooming” approach to genetic analysis
- ▶ Required for population resampling



after Brachi *et al.* [2]



De novo technical control

- ▶ Sample DNA not very physically distinctive
 - ▶ Mix-ups and contamination occur
- ▶ Catching mix-ups early important
 - ▶ Not just for your own data: SRA not so perfect

Ath 80 FIGURE HERE



Technological Overview

- ▶ k -mer analysis
- ▶ Hashing and Probabilistic Data Structures
- ▶ Population and RunFrequency Hashes
- ▶ (Weighted) Inner Products



k -mer analysis

- ▶ Analyse k -length words of sequences

$k = 3$

ACGTGT

ACG

CGT

GTG

TGT



k -mer analysis

- ▶ Analyse k -length words of sequences
- ▶ Computationally and biologically appropriate
 - ▶ Fast
 - ▶ Constant-memory (using `khmer`)
 - ▶ Scalable and parallelisable
 - ▶ Cross-scale

$k = 3$

ACGTGT

ACG

CGT

GTG

TGT



Hashes and Hash Functions

- ▶ Hash function e.g.

`hash('ACG')` => 5234315134



Hashes and Hash Functions

- ▶ Hash function e.g.

`hash('ACG')` => 5234315134

- ▶ “Hash”: a probabilistic data structure
 - ▶ Constant memory
 - ▶ Easy set operations and inner product
 - ▶ Implicit de Bruijn graph
 - ▶ Implemented in C Titus Brown's `khmer`
 - ▶ AKA Countgraph



Hash

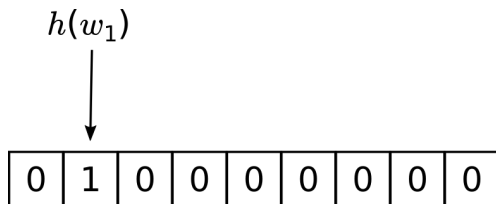
- ▶ Vector of large prime length (e.g. $1e9 + 7$)
- ▶ Indexed modulo length ($bin = h(w_i) \bmod prime$)

0	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---



Hash

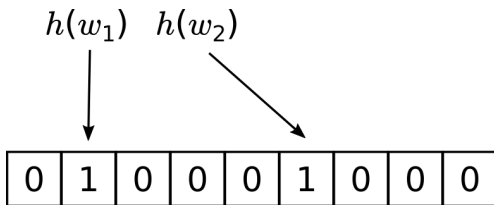
- ▶ Vector of large prime length (e.g. $1e9 + 7$)
- ▶ Indexed modulo length ($bin = h(w_i) \bmod prime$)





Hash

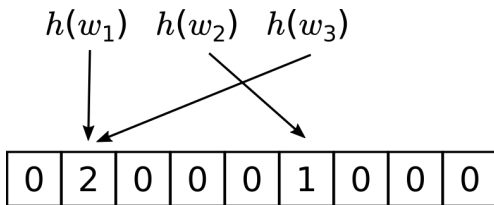
- ▶ Vector of large prime length (e.g. $1e9 + 7$)
- ▶ Indexed modulo length ($bin = h(w_i) \bmod prime$)





Hash

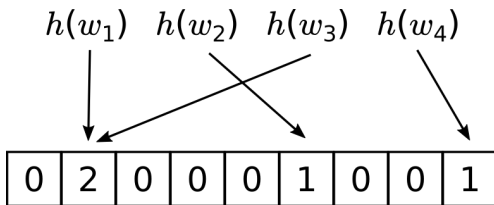
- ▶ Vector of large prime length (e.g. $1e9 + 7$)
- ▶ Indexed modulo length ($bin = h(w_i) \bmod prime$)





Hash

- ▶ Vector of large prime length (e.g. $1e9 + 7$)
- ▶ Indexed modulo length ($bin = h(w_i) \bmod prime$)





Hash Operations

- Population sum, or hash frequency

	2		2		1		1	
--	---	--	---	--	---	--	---	--

	2	1	7		1			
--	---	---	---	--	---	--	--	--

1	1				1	6		
---	---	--	--	--	---	---	--	--

1					2	3		
---	--	--	--	--	---	---	--	--

Population Sum

2	5	1	9	0	5	9	1	0
---	---	---	---	---	---	---	---	---



Hash Operations

► Population sum, or hash frequency

	2		2		1		1	
--	---	--	---	--	---	--	---	--

	2	1	7		1			
--	---	---	---	--	---	--	--	--

1	1				1	6		
---	---	--	--	--	---	---	--	--

1					2	3		
---	--	--	--	--	---	---	--	--

Population Sum

2	5	1	9	0	5	9	1	0
---	---	---	---	---	---	---	---	---

Frequency

2	3	1	2	0	4	2	1	0
---	---	---	---	---	---	---	---	---

 / 4



k -mer based clustering

- ▶ Alignment-free sequence clustering is a whole field
- ▶ $D2$ and friends
- ▶ Most require sequence gene/genome
- ▶ Many use inner product as similarity measure

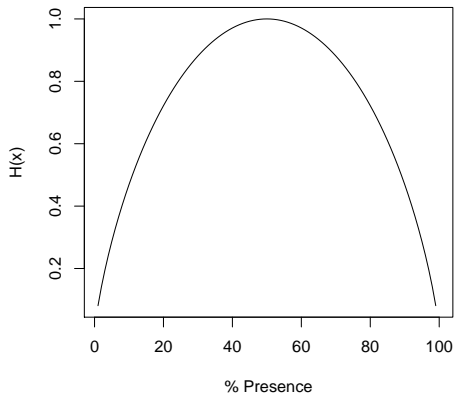
Characterizing the $D2$ Statistic: Word Matches
in Biological Sequences

Sylvain Forêt, Susan R. Wilson, and Conrad J. Burden



Shannon Entropy

- ▶ Measure of Information
- ▶ $-\sum_i p_i \log(p_i)$





kwIP

- ▶ The k -mer Weighted Inner Product
 - ▶ Extends alignment-free seq comparison to raw NGS data



kwIP

- ▶ The k -mer Weighted Inner Product
 - ▶ Extends alignment-free seq comparison to raw NGS data
- ▶ Algorithm:
 - ▶ For each sample: count all k -mers into a Hash



kWIP

- ▶ The k -mer Weighted Inner Product
 - ▶ Extends alignment-free seq comparison to raw NGS data
- ▶ Algorithm:
 - ▶ For each sample: count all k -mers into a Hash
 - ▶ For each analysis set, i.e “population”:
 - ▶ Calculate the informational entropy of hash frequency (H)
 - ▶ For each pair of samples 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



► The software:

- C++, >2000 lines of code
- Uses *khmer* for *k*-mer counting & hashing
- Parallelised, ≈ 10 hrs for 96 rice samples.
- GNU GPL licensed, source code on GitHub

kWIP
Docs

Welcome to kWIP's documentation!

Contents:

- [kWIP](#)
 - Overview
 - Installation
 - [kWIP CLI Usage](#)
 - [The Concepts Behind kWIP](#)
- [Example kWIP Analysis Protocols](#)
 - [Chrysa sativa grouping](#)
- [Experiments Around kWIP](#)

Indices and tables

- [Index](#)

Next

© Copyright 2015, Kevin Murray. Revision 8a8621a4a2573b6dc7a7e71d294ab9325c998492.
Built with [Sphinx](#) using a [theme](#) provided by [Read the Docs](#).



kWIP Experiments

- ▶ 3000 rice genomes:
 - ▶ 3000 rice lines from known families
 - ▶ Analysing in sets of ≈ 100 , from all major groups
 - ▶ Recover known grouping w/ kWIP, not w/ unweighted IP
 - ▶ Sensitive to read depth
- ▶ Simulation
 - ▶ Fake population genome sequencing studies

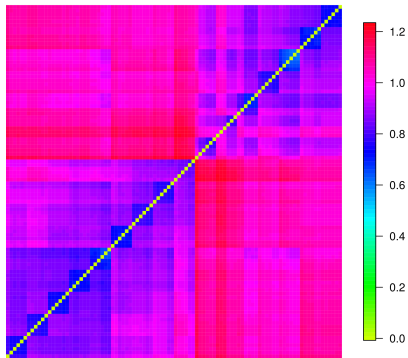


96 Rice Runs

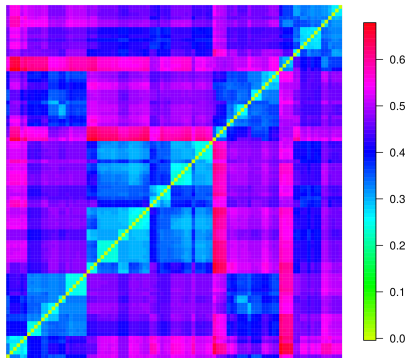
- ▶ Set of 96 rice runs from 12 samples (8 tech reps)
- ▶ 6 samples each from 2 major groups (Indica, Japonica)
- ▶ Expectations:
 - ▶ All runs cluster into groups of 6 reps (12 samples)
 - ▶ Big split between two groups: 6 samples each
- ▶ We see this with kWIP, not with Unweighted IP



WIP



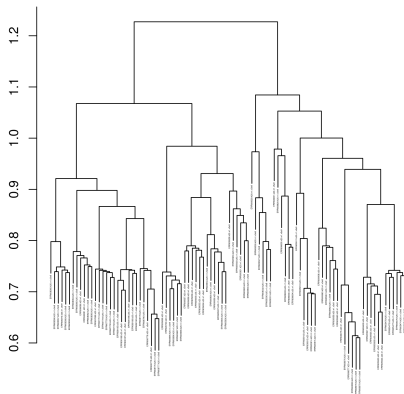
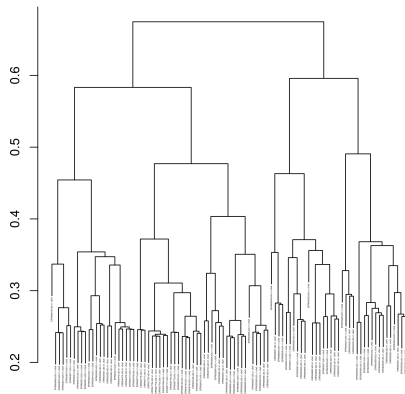
IP





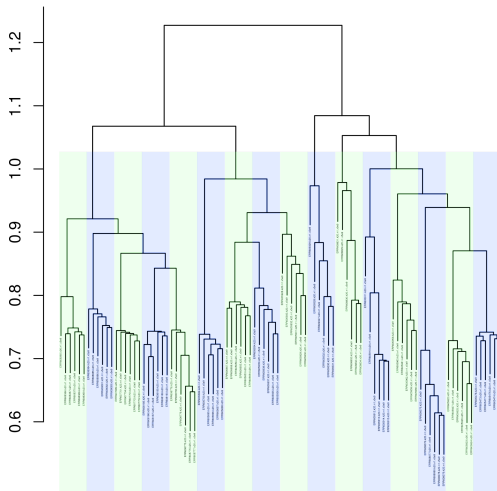
WIP

IP



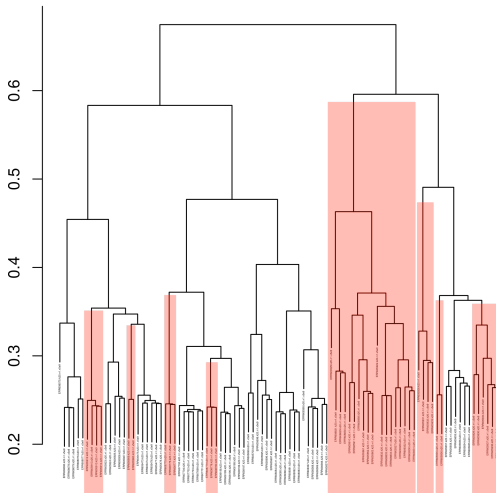


WIP





IP





Simulation

- ▶ Now for an Jupyter notebook





Thanks

- ▶ My collaborators: Cheng Soon Ong, Christfried Webers, Norman Warthmann
- ▶ {Super,Ad}visors: Justin, Sylvain, Gavin and Barry
- ▶ ~~khmer~~ folks: C. Titus Brown, Michael Crusoe, Camille Scott (DIB-lab) @ UC Davis
- ▶ Yourselfes



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).



Missing heritability in the field?

- ▶ Find more diversity in the field!
- ▶ Sample natural populations
 - ▶ Ecological hypotheses of trait selection, adaptation
 - ▶ Sample widely as possible across non-uniform genetic diversity



Missing heritability in the field?

- ▶ Find more diversity in the field!
- ▶ Sample natural populations
 - ▶ Ecological hypotheses of trait selection, adaptation
 - ▶ Sample widely as possible across non-uniform genetic diversity
- ▶ Now **complexity limited**: complex kinship & population structure
- ▶ Mandates development of economic, accurate large scale population genomics

