kWIP: The k-mer Weighted Inner Product

Kevin Murray

PhD Candidate, 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

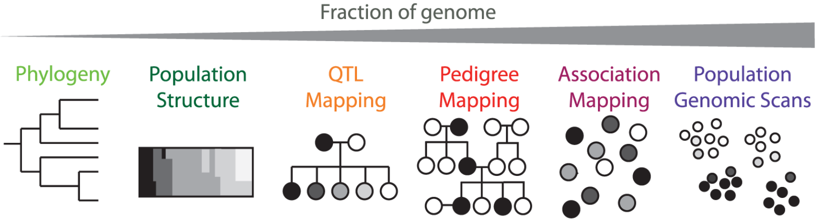
- Motivation
- Technological overview
- Early results and plans
- Demonstration

Large-scale population genomics

- Moving from 100s to 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*
  - Cross scale: *one tool to rule them all*



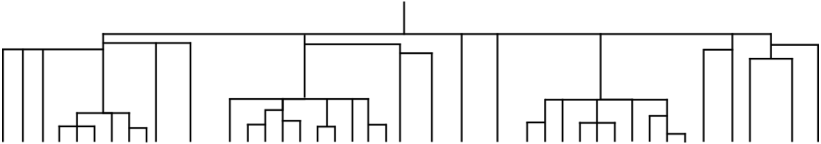
after Peterson *et al.* [1]

Rapid and Basic Clustering

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

Rapid and Basic Clustering

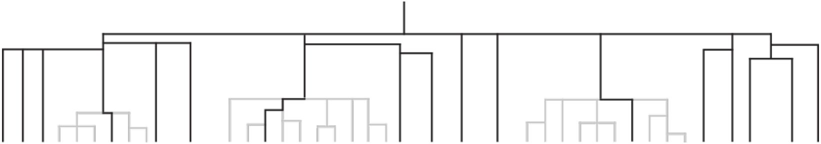
- Rough approximation of sample relatedness required
  - For natural collections
  - For association mapping
  - As a technical control



after Brachi *et al.* [2]

Rapid and Basic Clustering

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



after Brachi *et al.* [2]

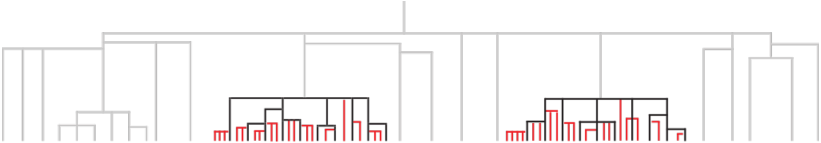
Rapid and Basic Clustering

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

after Brachi *et al.* [2]

Rapid and Basic Clustering

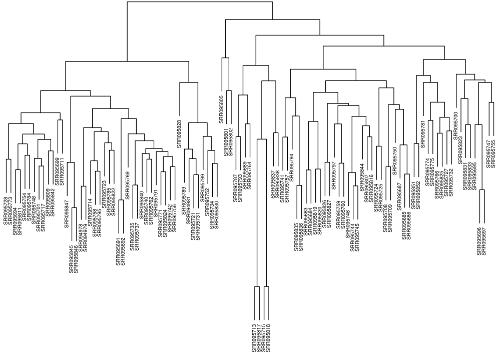
- Rough approximation of sample relatedness required
  - For natural collections
  - For association mapping
  - As a technical control



after Brachi *et al.* [2]

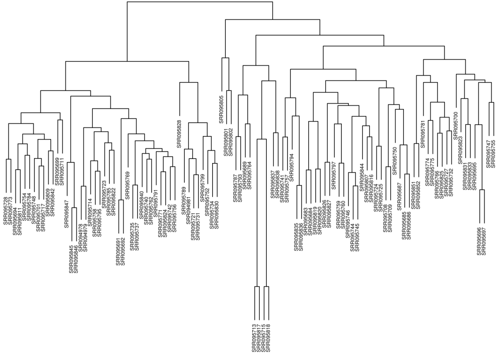
De novo technical control

- Sample DNA not very physically distictive
  - Mix-ups and contamination occur
- Not just for your own data: SRA not so perfect



De novo technical control

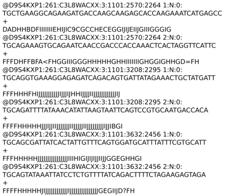
- Sample DNA not very physically distictive
  - Mix-ups and contamination occur
- Not just for your own data: SRA not so perfect



```
7d74f1b53174afc9fd3aec1dde1d996e SRR095715.fastq SRR095818.fastq
54da64f0343b69bcb959d33f127505d8 SRR095713.fastq SRR095817.fastq
```

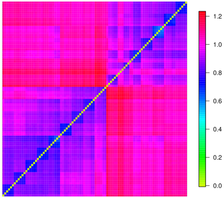
Presenting `kWIP`

- *k*-mer based *de novo* genetic clustering
- Weighted Inner Product between hashes to determine similarity
- Produces a distance matrix from raw NGS reads

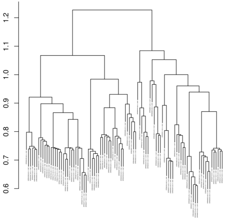


```
@D9S4KXP1:261:C3L8WACXX:3:1101:2570:2264 1:N:0:
TGCTGAAGGCAGAAGATGACCAAGCAAGAGCACCAAGAAATCATGAGCC
+
DADHHBDFIIIIIIIIEHIJIC9CGCCHECEGGIJIJEIIJGIIIGGGIG
@D9S4KXP1:261:C3L8WACXX:3:1101:2570:2264 2:N:0:
TGCAGAAAGTGCAGAATCAACCGACCCACCAAACTCACTAGGTTCATTC
+
FFFDHFFBFA<FHGGIIIGGGHHHHHGHHIIIIIIIIGHGGIGHHGD=FH
@D9S4KXP1:261:C3L8WACXX:3:1101:3208:2295 1:N:0:
TGCAGGTGAAAGGAGAGATCAGACAGTGATTATAGAAACTGCTATGATT
+
FFFHHHFHIJJJJJJJJJJJIJJJIJHHIJJJJIIJJJJJJJJJJJJJIJ
@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
+
FFFFHHHHHJIJJJJJJJJJJJJIJJJJJJJJJJJJJJJJGEGIIJD?FH
```

→



→



Technological Overview

- *k*-mer analysis
- Hashing and Probablistic Data Structures
- Population and Frequency 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
  - Alignment Free
  - Constant-memory (using `khmer`)
  - Fast: Scalable and parallelisable
  - Cross-scale

$k = 3$

```
ACGTGT
ACG
 CGT
  GTG
   TGT
```

Hashes and Hash Functions

- Hash function e.g.

  ```
  hash('ACG') => 5234315134
  ```

- For DNA, 2-bit encoding is used

Hashes and Hash Functions

- Hash function e.g.

  ```
  hash('ACG') => 5234315134
  ```

- For DNA, 2-bit encoding is used
- “Hash”: a probablistic data structure
  - Efficent way of counting k-mers
  - Constant memory
  - Easy set operations and inner product
  - Implicit de Bruijn graph
  - Implemented in C Titus Brown’s `khmer`

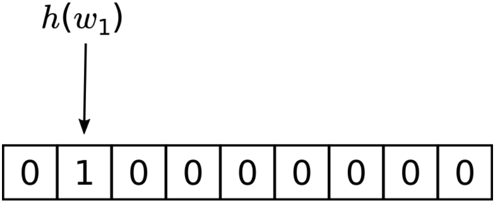
Hash

- Subset of a Count-Min Sketch/Counting Bloom Filter
- Vector of large prime length (e.g. 1e9 + 7)
- Indexed modulo length ($bin = h(w_i) \mod prime$)
- Aliasing can occur, hence probablistic

0	0	0	0	0	0	0	0	0

Hash

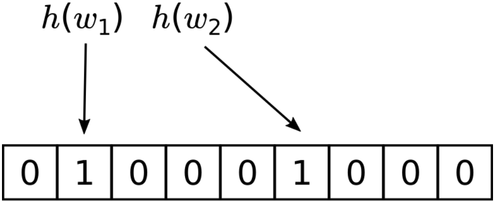
- Subset of a Count-Min Sketch/Counting Bloom Filter
- Vector of large prime length (e.g. 1e9 + 7)
- Indexed modulo length $(\mathit{bin} = h(w_i) \mod \mathit{prime})$
- Aliasing can occur, hence probablistic



0	1	0	0	0	0	0	0	0

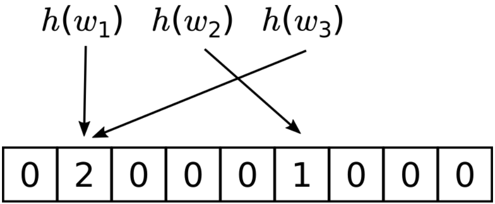
Hash

- Subset of a Count-Min Sketch/Counting Bloom Filter
- Vector of large prime length (e.g. 1e9 + 7)
- Indexed modulo length ($bin = h(w_i) \mod prime$)
- Aliasing can occur, hence probablistic



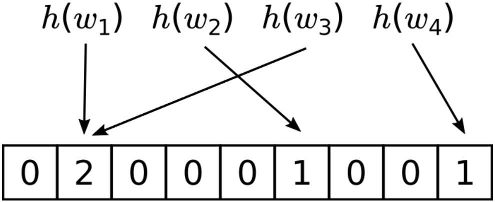
Hash

- Subset of a Count-Min Sketch/Counting Bloom Filter
- Vector of large prime length (e.g. 1e9 + 7)
- Indexed modulo length ($bin = h(w_i) \mod prime$)
- Aliasing can occur, hence probablistic



Hash

- Subset of a Count-Min Sketch/Counting Bloom Filter
- Vector of large prime length (e.g. 1e9 + 7)
- Indexed modulo length ($bin = h(w_i) \mod prime$)
- Aliasing can occur, hence probablistic



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

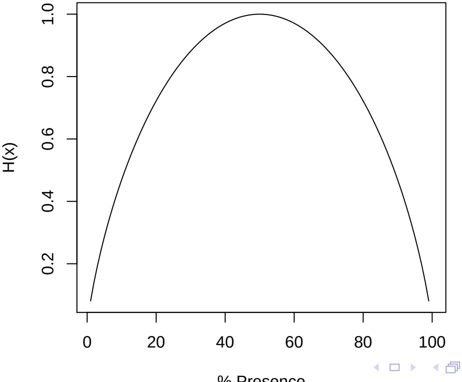
Hash Operations

- 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

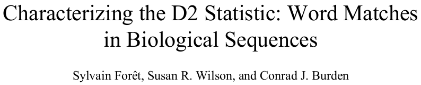
Shannon Entropy

- Measure of Information
- $-\sum_i p_i log(p_i)$
- `kWIP` weights by $H(frequency)$



k-mer based clustering

- Alignment-free sequence clustering is a whole field
- $D2$ and friends
- Most require assembled gene/genome sequence
- 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

kWIP

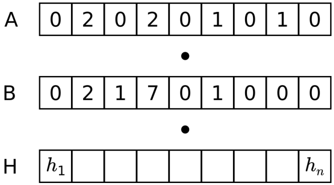
- The *k*-mer Weighted Inner Product
  - Extends alignment-free seq comparison to raw NGS data

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

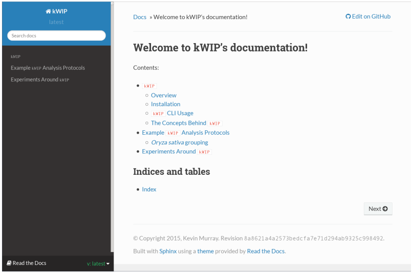
- The *k*-mer Weighted Inner Product
    - Extends alignment-free seq comparison to raw NGS data
- Algorithm:
    - For each sample: count all *k*-mers into a hash
    - For each analysis set, i.e “population”:
        - Calculate the entropy of population frequency (H)
        - For each pair of samples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

kWIP

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



Docs » Welcome to kWIP's documentation!

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.

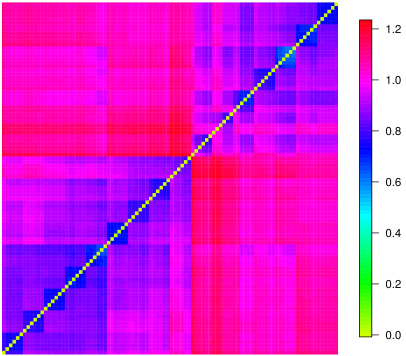
`kWIP` Experiments

- 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

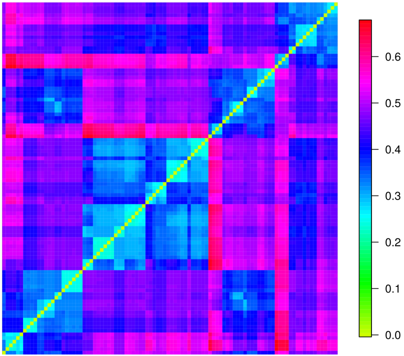
96 Rice Runs

- Set of 96 rice runs from 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)
- We see this with `kWIP`, not with Unweighted IP
- Took 10 hours on 16-core Raijin node, 60-80GB RAM

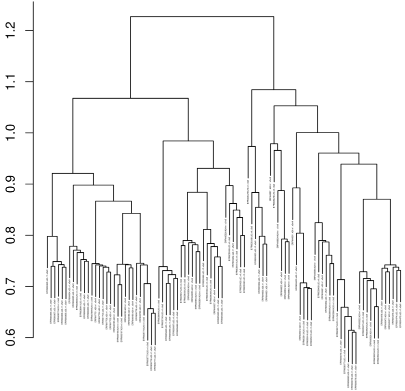
WIP



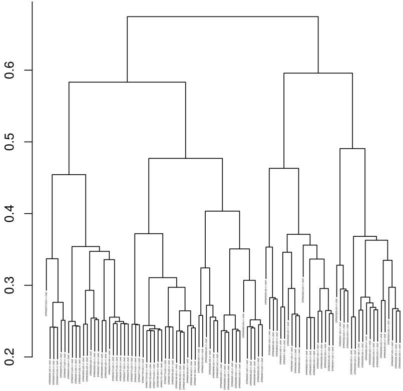
IP



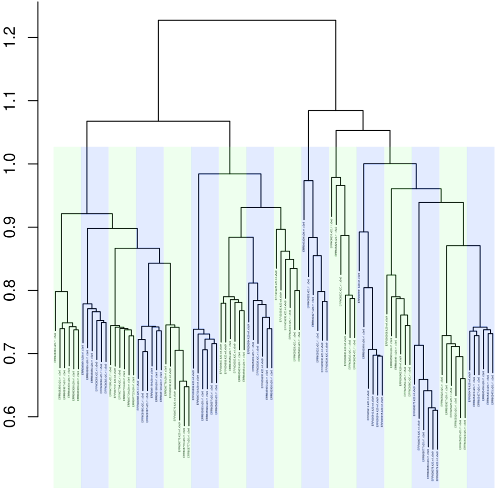
WIP



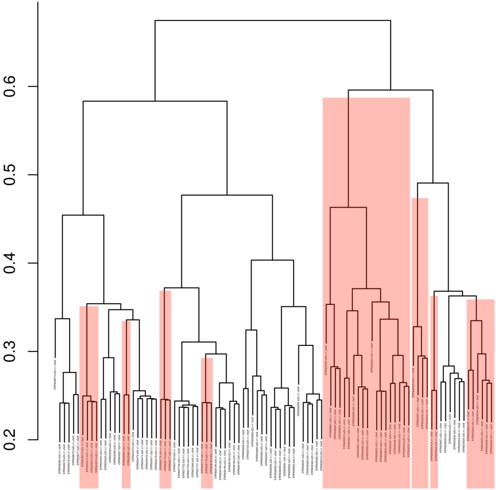
IP



WIP

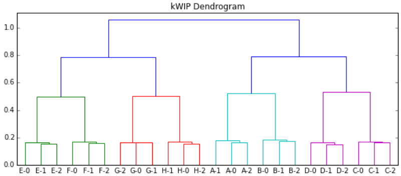


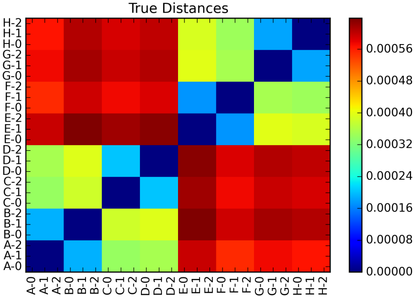
IP

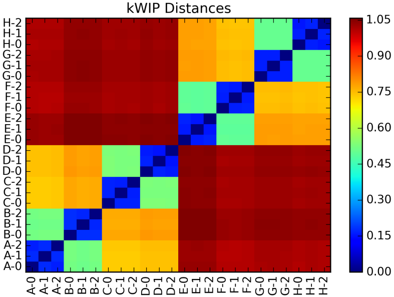


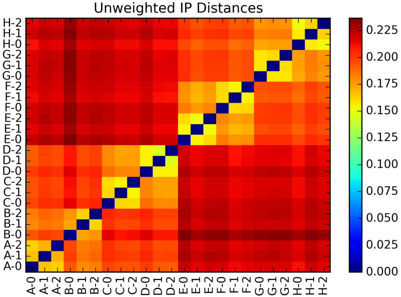
Simulation

- Aims to re-create a biforcating tree
- `kWIP` does so
- Neighbor joining trees have Robinson-Foulds distance of 0




True Distances


kWIP Distances
H-2
H-1
H-0
G-2
G-1
G-0
F-2
F-1
F-0
E-2
E-1
E-0
D-2
D-1
D-0
C-2
C-1
C-0
B-2
B-1
B-0
A-2
A-1
A-0
A-0
A-1
A-2
B-0
B-1
B-2
C-0
C-1
C-2
D-0
D-1
D-2
E-0
E-1
E-2
F-0
F-1
F-2
G-0
G-1
G-2
H-0
H-1
H-2
1.05
0.90
0.75
0.60
0.45
0.30
0.15
0.00


Unweighted IP Distances
H-2
H-1
H-0
G-2
G-1
G-0
F-2
F-1
F-0
E-2
E-1
E-0
D-2
D-1
D-0
C-2
C-1
C-0
B-2
B-1
B-0
A-2
A-1
A-0
A-0
A-1
A-2
B-0
B-1
B-2
C-0
C-1
C-2
D-0
D-1
D-2
E-0
E-1
E-2
F-0
F-1
F-2
G-0
G-1
G-2
H-0
H-1
H-2
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000

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
- 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).

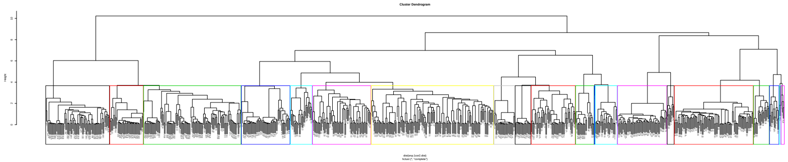
.

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


Cluster Dendrogram
Height
dist(snp.1cor2.dist)
hclust (*, "complete")