kWIP: The k-mer Weighted Inner Product

Kevin Murray[1,3], **C Webers**[2], **C S Ong**[2], **J Borevitz**[1,3], **N Warthmann**[1,3]

[1]: ARC CoE in Plant Energy Biology, ANU, Canberra

[2]: National ICT Australia (NICTA), and The Australian National University, Canberra, Australia

[3]: Research School of Biology, ANU, Canberra

kevin.murray@anu.edu.au

Poster URL: http://git.io/vcxYF

Abstract

We present the k-mer Weighted Inner Product, a *de novo*, alignment free measure of genetic similarity between samples in a population. kWIP, is an efficient tool implementing this metric that can determine the genetic relatedness between samples without alignment or assembly. We show kWIP can reconstruct the true relatedness between samples directly from sequencing reads generated with various modern sequencing platforms, as well as from simulated data.

Introduction

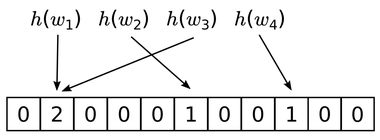
Modern population genomics requires sequencing many thousands of samples. To distil knowledge from this data requires analysis by sequence comparison. To compare such datasets, algorithmic improvement is required. Alignment-free sequence comparison promises to overcome some shortcomings of sequence alignment. However, few alignment free algorithms can process raw data from modern sequencing platforms, which sequence genomes as millions of short fragments. kWIP extends alignment-free sequence comparison algorithms to accept sequencing data directly.

Algorithms

kWIP works by decomposing sequencing reads to short k-mers, hashing these k-mers using a constant-memory data structure, and performing pairwise distance calculation between these sample k-mer hashes. One can calculate the inner product between hashes as a similarity measure. However, this treats all k-mers as of equal importance and accuracy. Therefore, kWIP applies a weight to each k-mer to reduce the contribution of technical noise to the overall signal, and focus on k-mers which provide maximal information about relatedness within a population.

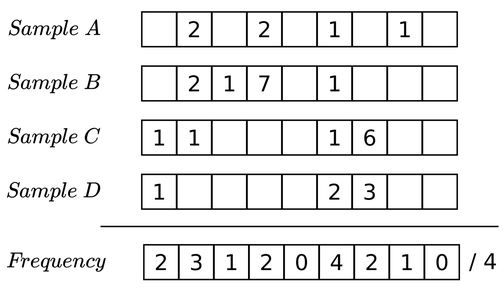
Hashing

Sequence reads are decomposed into k-mers and counted in a probabilistic data structure (Hash). This hashing is performed using the khmer C++ library [1].



Entropy vector weighting

To calculate the weighting applied to each k-mer, we first calculate the frequency of occurrence of the k-mer in the population. This is simply the proportion of samples with non-zero counts of a given k-mer.



The Shannon entropy of this frequency is used as the weights of each k-mer, calculated per (1).

$$H = -\sum_i P(x_i) log_2(P(x_i)) \tag{1}$$

Inner Product Calculation

Sample similarity is calculated pairwise between all samples as the inner product of hashes. The inner product between two hashes alone is calculated as (2). The weighted inner product calculation is calculated per (3).

$$\langle A, B \rangle = \sum_i A_i \cdot B_i \tag{2}$$

$$\langle A, B | H \rangle = \sum_i A_i \cdot B_i \cdot H_i \tag{3}$$

Implementation

kWIP is implemented in C++11, utilising the khmer C++ library. Weighted and unweighted inner products have been implemented. kWIP uses OpenMP to parallelise distance matrix calculation.

Experimental Validation

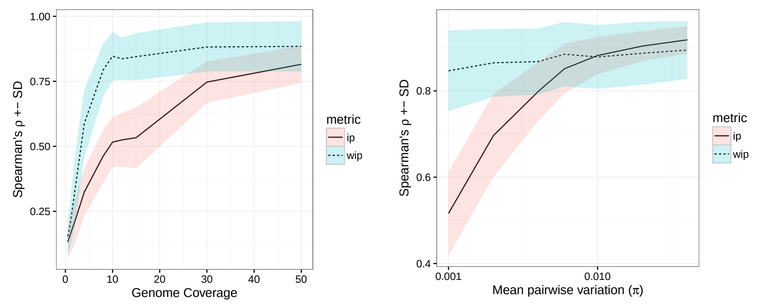
We present an initial experimental validation of kWIP and show increased performance of the weighted inner product metric compared to the unweighted metric.

Simulation Methods

Simulated population genome sequencing was used to test the performance of kWIP. Populations were generated at random[2] (with fixed π) , and genomes simulated using evolutionary models[3]. Sequencing reads were simulated at various read coverages[4], before k-mer counting[1] and analysis with kWIP. Accuracy is calculated using rank order correlation (Spearman's rho) of true pairwise genomic distance and kWIP's estimate of genetic relatedness.

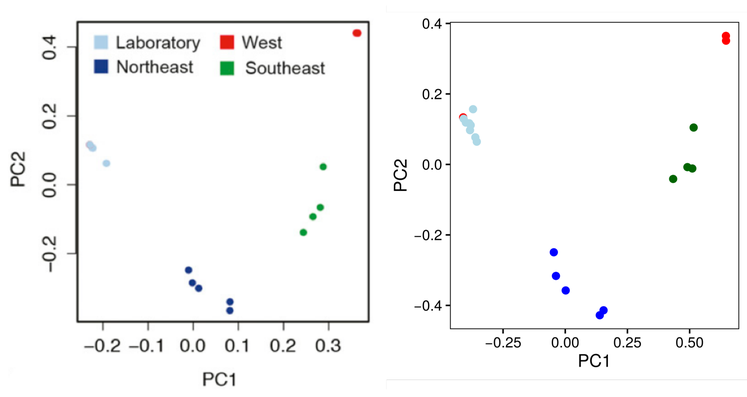
Coverage and Divergence affects performance

kWIP more accurately estimates genetic distance at higher average sample coverage and average pairwise genetic distance. At coverages common in population genomics (1-30x coverage) kWIP outperforms the unweighted equivalent; the performance of these measures eventually converge. Similarly, kWIP's performance outperforms the unweighted metric at low average pairwise genetic distance (π).



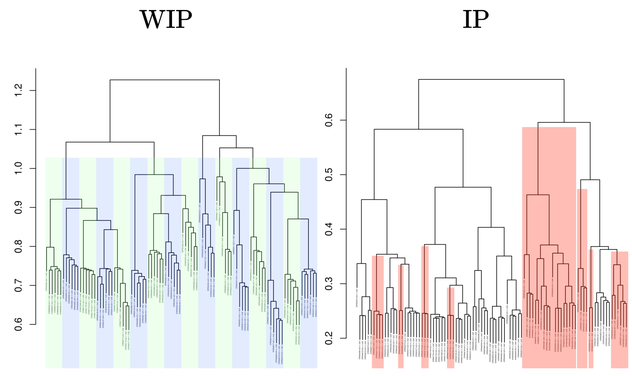
Population structure detected

Using a *Chlamydomonas* population re-sequencing experiment[5], we show kWIP can detect population structure approximately as well as a state-of-the-art reference-based variant calling pipeline.



Replicates are accurately clustered

Using data from the 3000 Rice genomes sequencing project[6] we show weighting improves replicate clustering accuracy. Representative example show below (erroneous clustering indicated by red highlighting).



Summary

kWIP: a fast tool to determine approximate genetic relatedness

- Entirely *de novo* and alignment free
- Efficient k-mer counting into probabilistic data structures (using khmer)
- Uses entropy weighting to amplify signal above technical noise
- Available from https://github.com/kdmurray91/kwip under the GNU GPL

Forthcoming Research

A paper describing kWIP in more detail is in preparation. We plan to deploy kWIP across several large-scale plant population genome sequencing projects. An MPI-parallelised implementation is in preparation.

Acknowledgements

We thank Sylvain Forêt, Conrad Burden, Terry Neeman, Ben Kahler, Gavin Huttley and Cameron Jack for comments or advice on algorithms and experiments presented here.

References

1. Crusoe, M. R. *et al.* The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research* (2015).
2. Staab, P. R., Zhu, S., Metzler, D. & Lunter, G. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* **31,** 1680–1682 (2015).
3. Cartwright, R. A. DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics* **21,** iii31–iii38 (Suppl 3 2005).
4. Holtgrewe, M. Mason – A Read Simulator for Second Generation Sequencing Data. *Technical Report FU Berlin* (2010).
5. 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).
6. The 3,000 rice genomes project. The 3,000 rice genomes project. *GigaScience* **3,** 7 (2014).