kWIP: The k-mer Weighted Inner Product

K Murray[1], C Webers[2], C S Ong[2], S Forêt[3], J Borevitz[1], N Warthmann[1]

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

[2]: NICTA, Canberra

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

kevin.murray@anu.edu.au Poster URL: http://git.io/vcxYF

Abstract

Modern techniques in population genomics generate unprecedented quantities of data within which complex genetic histories reside. The scale and complexity of these data require the development of new approaches to the analysis of genetic data. 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

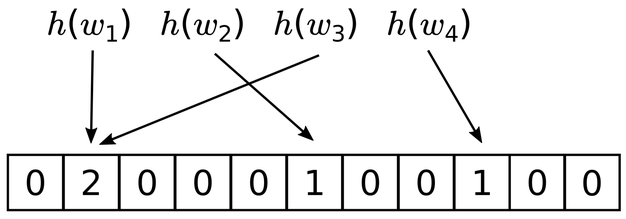
Modern population genomics requires sequencing many thousands of samples. To distil knowledge from this data requires analysis by sequence comparison. To compare such datasets, algorithmic improvement is required. Alignment-free sequence comparison is a powerful field of algorithms which overcome some 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.

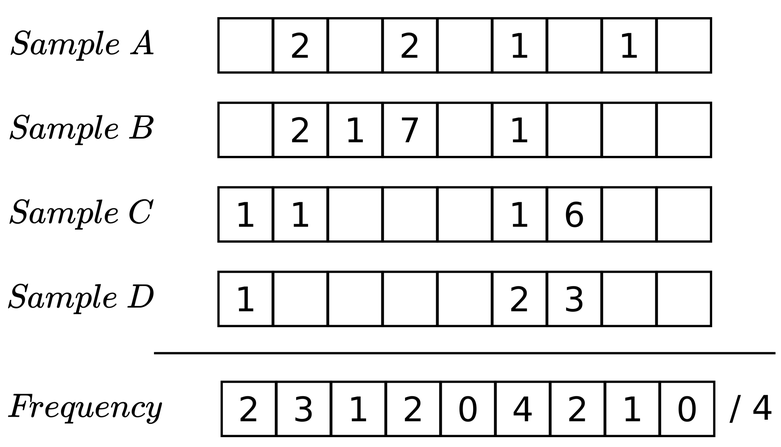
Hashing

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).

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

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

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 in a thread-safe manner.

References

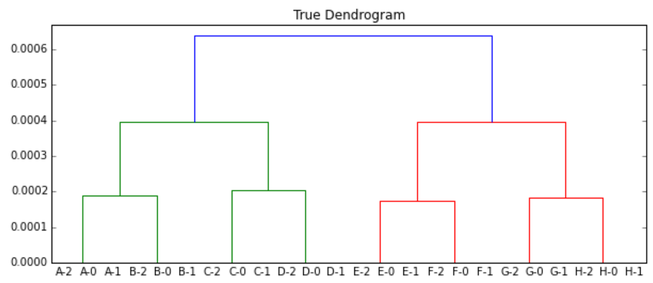
[1] Michael R. Crusoe et al. "The khmer software package: enabling efficient nucleotide sequence analysis". en. In: *F1000Research* (2015).

Experimental Validation

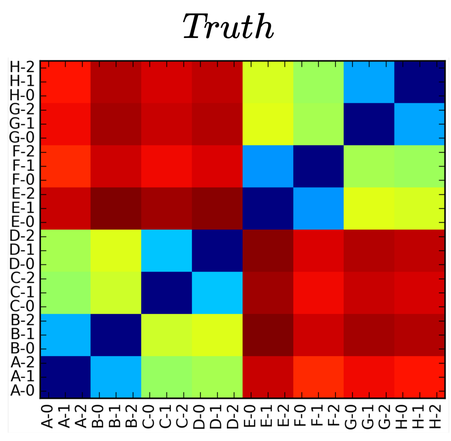
We present an initial experimental validation of kWIP and show increased performance of the weighted inner product metric compared to the unweighted metric.

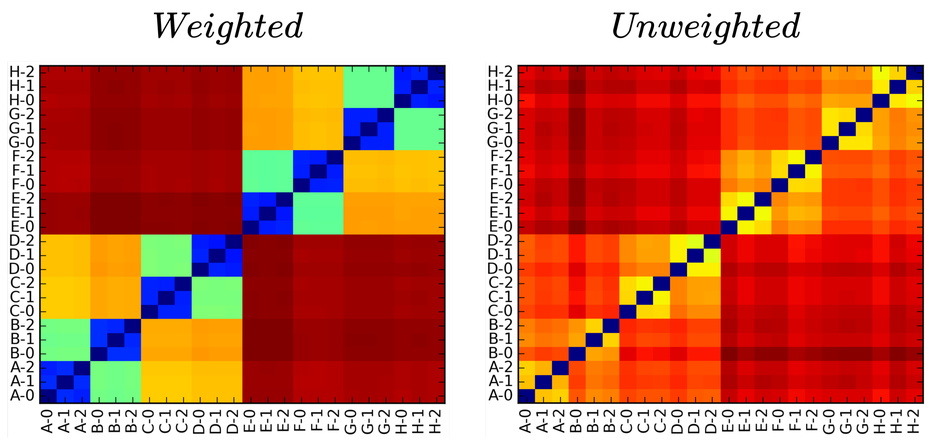
Methods

Random genomes with equal base composition, and genetic variation from these genomes, were simulated using the mason genome simulation suite [2] and seq-gen[3]. Reads were hashed using khmer version 2.0, and kWIP was run with default parameters in both weighted and unweighed mode.



Simulation Results





Above we see kWIP's estimation of the genetic relatedness between simulated samples. Note the increased resolution and fidelity of the WIP when compared to the IP distance matrix.

kWIP can calculate pairwise relatedness between 96 samples of 3Gbp in approximately 10 hours on 16 CPUs.

Summary

kWIP: a fast tool to determine approximate genetic relatedness

- Entirely *de novo* and alignment free
- Hashes k-mers into probabilistic data structures
- Calculates inner products between sample hashes
- Available from https://github.com/kdmurray91/kwip under the GNU GPL v3+.

Forthcoming Research

A paper describing kWIP in more detail is in preparation. We plan to deploy kWIP for use as a quality assurance measure across several large-scale plant population genome sequencing projects. We plan to implement MPI-based parallelism, and to extend kwip to full counting Bloom filters or counting de Bruijn graphs.

Acknowledgements

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

[2] M. Holtgrewe. "Mason A Read Simulator for Second Generation Sequencing Data". In: *Technical Report FU Berlin* (2010).

[3] A. Rambaut and N. C. Grassly. "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees". eng. In: *Computer applications in the biosciences: CABIOS* 13.3 (1997), pp. 235–238.