kWIP: The k-mer Weighted Inner Product, a *de novo* estimator of genetic similarity

Kevin D. Murray[1], Christfried Webers[2,3], Cheng Soon Ong[2,3], Justin O. Borevitz[1], and Norman Warthmann[1]

[1]Centre of Excellence in Plant Energy Biology, The Australian National University, Canberra, Australia
[2]Data61, CSIRO, Canberra, Australia
[3]Department of Computer Science, The Australian National University

September 16, 2016

Abstract

Modern techniques used in the elucidation of population genomic variation generate potentially overwhelming quantities of data. The nature and scale of this data demands the development of computationally efficient methods to determine genetic relatedness in an unbiased *de novo* manner. We present the k-mer Weighted Inner Product (kWIP), a novel assembly and alignment free estimator of genetic similarity. We show kWIP can recapitulate observed relationships among samples across diverse datasets and reconstruct the true relatedness between samples with simulated sequence. kWIP is licensed under the GNU GPL, and available from https://github.com/kdmurray91/kwip.

1 Introduction

Population genetics studies seek to determine the genetic relatedness of individuals of a species, within and between populations. Nowadays, the vast majority of studies in this field are performed using next generation sequencing (NGS) [31]. Many of the methods commonly employed to analyse NGS data rely on two overarching concepts: the assembly of a reference genome, and the comparison of a sample to this assembled reference genome (using re-sequencing and variant calling). These methods, while functional in model systems, are not ideal for use in a wide variety of studies. Generating a single reference genome is time consuming and costly if none is available [41, 34], and can lead to bias when samples differ dramatically from that reference [5, 18].

In many cases, a large volume of sequence data is first condensed to pairwise relatedness between individuals. This empirically provides a genetic who's who to

confirm population and family group assumptions. A recent review [3] found that sample misidentification occurs at an alarming rate. The rapid estimation of genetic relatedness between samples before any further analysis has the potential to detect such mix-ups before the investigator begins to draw conclusions using mis-labelled samples. Verifying that biological replicates belong to the same genetic lineage is another crucial initial step.

Estimating the genetic relatedness between a broad collection of natural accessions is needed to develop core sets for further detailed studies including breeding and conservation. When individuals from real world populations are collected, there is normally non-uniform genetic relatedness. Initially, one seeks to group samples into more closely related families or more distantly related populations. Genetic outliers can represent cryptic species and should be excluded. A process of population re-structuring [4] balances the genetic diversity among a subset of individuals for association studies. This process should avoid reference genome bias and have minimal per sample cost to maximise the number of samples.

Current methods to calculate genome-wide genetic relatedness from short read sequencing data commonly rely on alignment of short reads to a reference genome sequence, followed by variant calling. Practical limitations aside, calculation of genetic relatedness following read alignment exposes the investigator to a bias from the particular reference genome used for alignment. The field of alignment-free sequence comparison aims to combat these difficulties by avoiding the process of sequence alignment. Instead, approaches including decomposition to words (or k-mers), [32, 39, 12] sub-string or text processing algorithms [23, 24] and information theoretic measures of sequence similarity or complexity [42] are utilised as comparators. While avoiding issues inherent in sequence alignment, many alignment-free sequence comparison tools still require previous elucidation of each individual's genome sequence, which precludes their use in *de novo* analyses.

Recently, several algorithms enabling *de novo* comparison have been published. These extensions all attempt to reconstruct deeper phylogenetic relationships directly from sequencing reads. Spaced [32, 24] uses the Jensen-Shannon distance on spaced seeds (small k-mers a short distance from one another or with interspersed disregarded bases) to improve performance of phylogenetic reconstruction. Cnidaria [1] and AAF [10] use the Jaccard distance to reconstruct phylogenies, while `mash` [35] uses the MinHash approximation of Jaccard distance to the same effect. However, none of these methods report acceptable performance within species, meaning their utility in population genomics may be limited. Here we present `kWIP`, a tool which enables the estimation of genetic relatedness directly from sequencing reads. Using an inner product between k-mer counts, `kWIP` sensitively and accurately detects genetic relatedness between the individuals of a population, directly from sequencing reads.

2 Algorithms

`kWIP` operates on short sequencing reads generated by common modern sequencing platforms (e.g. Illumina). First, `kWIP` utilises `khmer` [44, 8] to count overlapping words of length k (k-mers) into a probabilistic data structure (sketch) for each

sample. `kWIP` then calculates a population-wide aggregate of k-mer frequency (population sketch) across all or a select set of samples. Using the population sketch, we calculate Shannon entropy of occurrence frequency for each sketch bin. We then calculate similarity as the inner product between each pair of sample sketches, weighted by Shannon entropy.

2.1 k-mer counting

For each sample, `kWIP` uses `khmer` to decompose sequencing reads into overlapping words of some fixed length, e.g. 20. The value of a reversible hash function is computed for each k-mer. k-mers are canonicalised by using the lexicographically smaller of a k-mer and its reverse complement. k-mers are counted using a sketch per sample. These sketches are vectors with prime number length, typically several billion elements in size (denoted S_i for sample i). The elements of these sketches are referred to as bins (indexed by b, e.g. S_{i_b}), and can store values between 0 and 255 (integer overflow is prevented). To count a k-mer, the b-th bin of the sketch (S_{i_b}) is incremented, where b is the hash value of the k-mer modulo the (prime) length of the sketch. For most use cases, we recommend counting k-mers between 19 and 21 bases long, as this balances the number of distinct k-mers with the uniqueness of each k-mer across samples.

Note that the possible number of k-mers (4^k) is much larger than the length of a sketch. Therefore, aliasing or "collision" can occur between k-mers, but in practice is rare [44]. It is also worth noting that aliasing can only increase similarity between any two samples however this should occur uniformly across all sample pairs.

2.2 Weighting and similarity estimation

Genetic similarity is estimated as the inner product between each pair of sample sketches (S_i, S_j), weighted by the informational content of each bin. The population sketch (P) contains the frequency of occurrence of each bin, calculated as the number of samples with a non-zero count for each bin. We calculate a weight vector (H) of these occurrence frequencies using Shannon's entropy as per Equation (1). In the Weighted Inner Product (WIP) metric (or kernel), similarity is then calculated as the inner product over each sample's sketch, weighted by H as per Equation (2). The unweighted Inner Product (IP) metric is simply the inner product between to sample sketch vectors, $S_i \cdot S_j$. This produces a matrix of pairwise inner products K, commonly referred to as a kernel matrix. The kernel matrix is normalised using the Euclidean norm (3), and converted to distances using the "kernel trick" [16] as per Equation (4). Kernel matrices are confirmed to be positive semi-definite by checking that all eigenvalues are non-negative using the Eigen3 library [14].

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

$$K_{ij} = \sum_{b=1}^{n} S_{i_b} S_{j_b} H_b \tag{2}$$

$$K'_{ij} = \frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}} \tag{3}$$

$$D_{ij} = \sqrt{K'_{ii} + K'_{jj} - 2K'_{ij}} \tag{4}$$

3 Methods

We demonstrate `kWIP`'s performance using both real and simulated datasets. To quantify the performance of `kWIP` we use simulation studies. To show the utility of `kWIP` in low-coverage, large-scale population genomics datasets, we use data from the 3000 Rice Genomes Project [28, 40]. To show that `kWIP` estimates genetic similarity as well as SNP-based methods, we use a population genomics study of *Chlamydomonas* [11]. We show the applicability of `kWIP` to detecting microbiome structure from shotgun metagenomic data using a study of the rice root-associated microbiome [9].

The analyses performed here are implemented as Snakemake workflows [22], which describe all analysis steps including tool parameters. Where appropriate, random seeds have been fixed, and all analyses should be fully reproducible. Jupyter notebooks containing all post-analyses are available online. Precise program versions are noted in Supplementary Table 1. These snakemake workflows and Jupyter notebooks are available online at `https://github.com/kdmurray91/kwip-experiments`.

3.1 Simulated Dataset

We simulated several datasets to empirically determine the performance of `kWIP`. Populations of 12 individuals were simulated using `scrm` [38]. Branch lengths within each population were normalised such that the mean pairwise genetic distance (π) was equal. Branch lengths were then scaled over a range of π to test the effect of mean pairwise genetic distance. Genome sequences were simulated using DAWG2 [6]. Simulated data for three replicate sequencing runs were generated at various mean coverages using Mason2 [17] (with random variation in read number; coefficient of variation equal to 0.3 based on our experimental experience). We then used `khmer` to count k-mers in these simulated sequencing runs. Genetic similarity was estimated using `kWIP`, with both the weighted (WIP) and unweighted (IP) metrics.

The performance of each metric was measured relative to the true pairwise distances between simulated samples. The true distance matrix was calculated from the simulated aligned sample genomes using `scikit-bio`. Sample-wise entries were replicated to reflect the number of simulated runs; the true genetic distance between technical sequencing replicates was considered to be zero. Performance was calculated as Spearman's rank correlation (ρ) between all pairwise distances using `scipy` [20].

3.2 Datasets

We use several published datasets to demonstrate the performance and utility of `kWIP`. In all cases, data files were obtained from the NCBI Short Read Archive using

`sra-py` [33]. Reads were extracted using the SRA toolkit to interleaved FASTQ files. Low base quality regions were removed using sickle [21] in single-end mode. Counting of k-mers was performed using the `load-into-counting.py` script of `khmer`. Genetic similarity was estimated using `kWIP`, using both the WIP and IP metrics.

We use a selection of sequencing runs from the 3000 Rice Genomes project [28, 40]. We selected 100 sets of 96 replicate runs, which were chosen such that there were 16 samples with 6 replicate runs each. We ensured 8 samples each were described by Li, Wang, and Zeigler [28] as belonging to the Indica and Japonica families of *O. sativa*. To demonstrate the ability of `kWIP` to accurately cluster replicates at low coverage, we estimated similarity between runs in each of these 100 sets. We created a matrix describing the expected distances such that each run had a distance of zero to itself, a distance of 1 to each other run in its biological sample, a distance of 2 to each run from samples in the same rice group (Indica or Japonica), and a distance of 4 to each run from the other rice group. Spearman's rank correlation between this matrix and the distance matrices obtained using `kWIP` were calculated with `scipy`. A paired Student's t-test was performed between the estimates of relatedness from the WIP and IP metrics using R's `t.test` function.

We use data from Flowers et al. [11] to show the ability of `kWIP` to detect subtle population structure. Genetic relatedness between 20 *Chlamydomonas* samples from this study was estimated with `kWIP` using the WIP metric. Classic Multidimensional Scaling (MDS) of the `kWIP` distance matrix was performed using the `cmdscale` function in R. We compare our MDS against the principal component analysis (PCA) of SNP genotypes as reported by Flowers et al. [11]. Note that for Euclidean distance matrices, MDS is equivalent to PCA [15].

We examined the effect of sample sequencing depth on the accuracy of `kWIP` in real datasets. To do so, we randomly sub-sampled each sample to coverages of between 0.01-fold and 200-fold average coverage across samples using the `sample` command of `seqtk`. The coefficient of variation in read number that existed in the original dataset (0.12) was preserved by sampling a random number of reads from an appropriate normal distribution. Spearman's rank correlation (ρ) was used to compare pairwise distances calculated at each average coverage to those from the original dataset.

To demonstrate the ability of `kWIP` to determine the relatedness of metagenomic datasets, we used data from a study of rice root microbiomes [9]. Illumina reads were downloaded from the short read archive as above, and extracted to FASTQ files. Low quality bases and adaptor sequences were removed using `skewer` [19]. k-mers were counted using `khmer` with a table size of 5×10^9 and relatedness estimated using `kWIP` with the WIP metric, and MDS was performed as above.

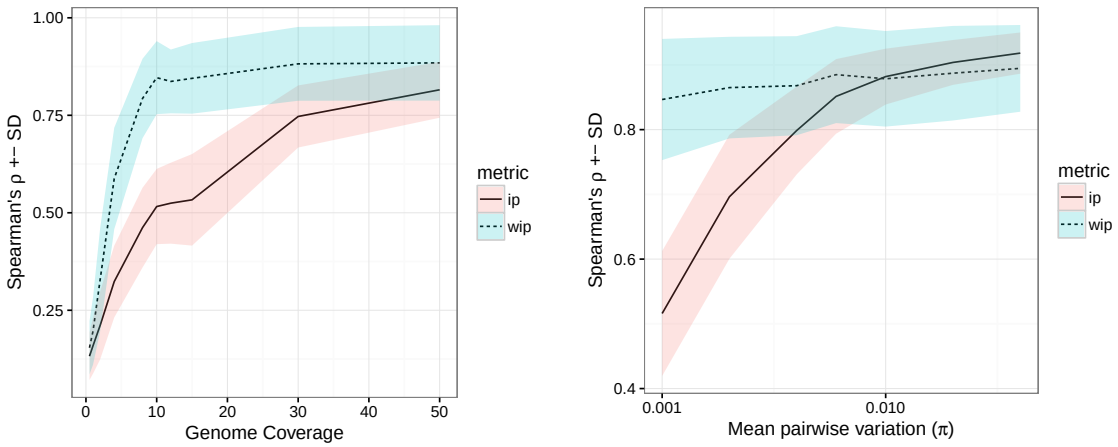
4 Results

We show that `kWIP` is able to determine genetic relatedness accurately in many scenarios. Using a simulated population re-sequencing experiment, we show that the weighting `kWIP` applied improves correlations with known truth (hereafter referred to as accuracy). We recover known technical and biological relationships between sequencing runs of the 3000 Rice Genomes project. We show that visualisation

of kWIP's estimation of genetic relationships between *Chlamydomonas* samples is nearly identical to that published by the dataset's authors who used traditional read mapping and variant calling to a reference genome.

4.1 Quantification of kWIP performance

Using simulated data we show that kWIP's correlation to known truth decreases with average sample sequencing depth (Figure 1a). Importantly, the WIP metric improves accuracy primarily at low coverage (below 30-fold). Above a certain coverage (in the case of the our simulations, approximately 50-fold), the performance of the WIP and IP metrics converges. As a general rule, if a sample has much lower coverage than the average, kWIP has difficulty accurately determining its relatedness to other samples. Therefore we advise excluding such poor quality samples. The improvement in accuracy of the WIP metric relative to the IP metric increases as mean pairwise genetic variation decreases, given constant coverage. While the accuracy of the IP metric decreases markedly below a mean pairwise distance of approximately 0.01, the WIP metric shows no such decrease (Figure 1b).



(a) (b)

Figure 1: The effect of average sample coverage and average pairwise genetic distance (π) on genetic similarity estimate accuracy. In (a) we show that at low to moderate coverage (2-30x) weighting increases accuracy, and the weighted metric obtains near-optimal accuracy at 10x coverage compared to above 30x for the unweighted metric. In (b) we show that the performance of the unweighted metric decreases rapidly as the mean pairwise distance between samples decreases, however this does not occur for the weighted metric. Lines indicate mean $\pm$ standard deviation of Spearman's ρ across 50 replicate runs.

4.2 Case Study: Replicate clustering

We show the utility of kWIP in verifying replicates. In Figures 2a and 2b we show a representative example of replicate clustering, showing the WIP metric is more able

to cluster replicates than the IP metric. In Figure 2c we show the distribution of rank correlation coefficients between distances obtained using the WIP and IP metrics and the expected clustering patterns for 100 subsets of 96 sequencing runs. The WIP metric outperforms the IP metric, having a higher mean correlation compared to the IP metric ($p << 0.001$, paired Student's T test, $n = 100$). To estimate relatedness between 96 rice samples took approximately 8 hours on a 64GB, 16 core (2.6 GHz Sandy Bridge) supercomputer node.

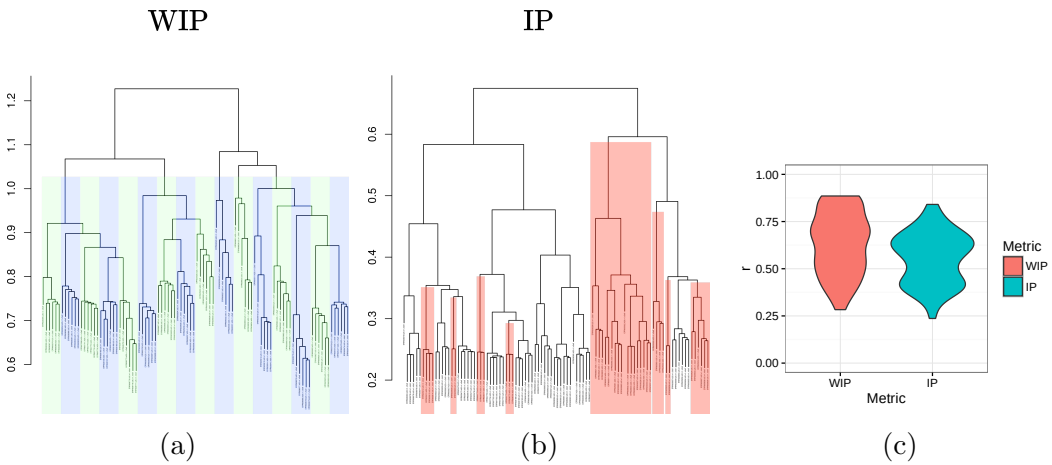
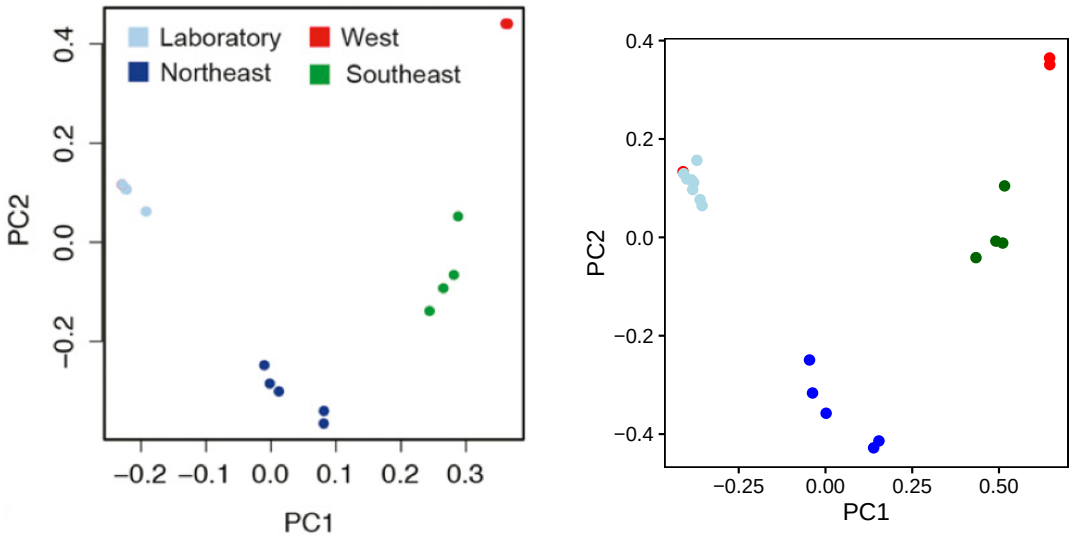


Figure 2: Weighting improves replicate clustering accuracy. In (a) and (b) we show a representative example demonstrating that the WIP metric (a) correctly clusters sets of 6 replicate runs into their respective samples (as indicated by blue and green bars) while the unweighted metric (b) cannot do so in all cases (indicated by red highlighting). In (c) we show rank correlation coefficients to expected relationships over 100 sets of 96 rice runs for the WIP and IP metrics.

4.3 Case Study: Population structure

Flowers et al. [11] sequenced 20 *Chlamydomonas reinhardtii* from continental USA, detecting significant and apparently geographic population structure. Figure 3 shows the genetic relatedness detected by Flowers et al. [11] as a principal component analysis (PCA) of SNP genotypes, alongside the highly similar relatedness detected using `kWIP`. In both analyses, we see the separation of laboratory strains (and one western sample) from both eastern and western samples (PC1). We see further structure among wild *Chlamydomonas* collected in western, southeastern and northeastern USA. In most cases, even the relationships between samples within each of these groups are highly similar between distances obtained using `kWIP` and those provided by Flowers et al. [11]. Additionally, we investigated the effect of reduced coverage on the fidelity of `kWIP`'s similarity estimation. Figure 4 shows that as coverage decreases, the accuracy of the estimated relationships decreases. We also show the visual effect of this decay on the visualisation of estimated genetic relatedness.



(a) (b)

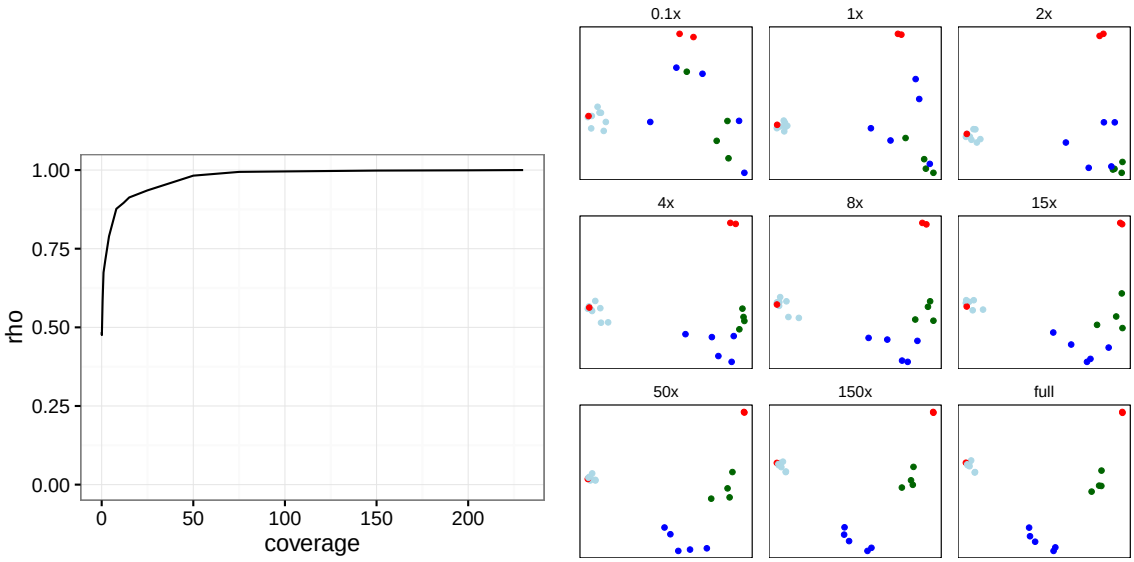
Figure 3: Genetic relatedness between *Chlamydomonas reinhardtii* (data from Flowers et al. [11]). (a) PCA of SNP genotypes reproduced from Flowers et al. [11]. (b) Relatedness obtained using `kWIP`. Flowers et al. [11] note that "Sample CC-4414 (red) is hidden behind the cluster of laboratory strains (light blue)", and this sample is almost obscured by these laboratory strains in the `kWIP` result.

4.4 Case Study: Metagenome relatedness

Edwards et al. [9] sequenced the rice root-associated microbiome, finding stratification of samples by biome and cultivation practice. Using `kWIP`, we detect highly similar microbial community structure. We observe a gradient of samples from within the root through the root-soil interface and into soil (Figure 5). We replicate the observed separation of samples by cultivation site. Additionally, we observe separation by cultivation practice, detecting the distinct microbiomes of rice in field plots and greenhouse pots, and detect the gradual change of sample microbiomes through time noted by Edwards et al. [9] (data not shown).

5 Discussion

The k-mer Weighted Inner Product (`kWIP`) estimates genetic distances between samples directly from NGS data, using a weighted inner product that aims to reduce the effect of the technical and biological noise between samples. To estimate genetic distance, `kWIP` uses the WIP metric, which weights k-mer counts by their informational entropy across an analysis set. Euclidean distances are then calculated from these inner products. This has the effect of down-weighting k-mers which are either highly abundant, or present in very few samples. These k-mers are typically either repetitive, invariable, or erroneous. By using Shannon entropy, the weights of infrequent or common k-mers are assigned lower, but non-zero weights, allow-



(a) (b)

Figure 4: The effect of average sequencing depth on estimation of genetic relatedness between *Chlamydomonas reinhardtii* (data from Flowers et al. [11]). (a) Spearman's rank correlation between sub-sampled datasets and the entire dataset across a range of subset average coverages. (b) Relatedness obtained using `kWIP` on selected sub-sampled datasets. "full" refers to the entire dataset, while "0.1x" refers to a sub-sampled dataset with average coverage of 0.1 over the *C. reinhardtii* genome (and likewise for 1x, 2x, and so on).

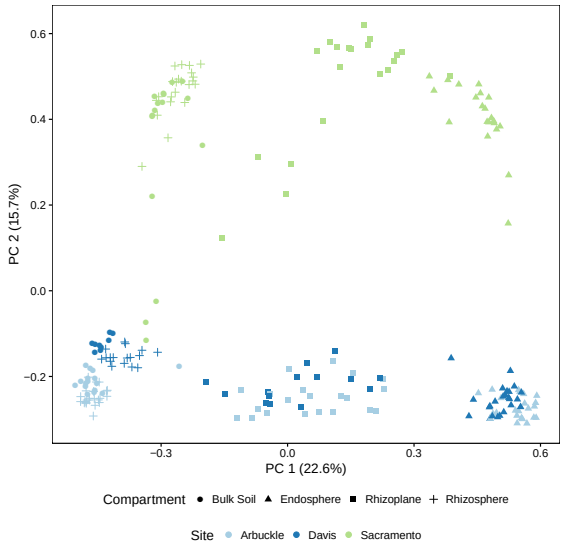


Figure 5: Estimation of relatedness of metagenomes. Here, we use `kWIP` to examine the data of Edwards et al. [9]. We replicate these authors' observations of stratification of root-associated microbiomes by biome compartment (PC1) and collection site (PC2).

ing some contribution of their signal. kWIP is able to estimate the genetic distance between samples with less data than is typically used to call SNPs against a reference, without requiring a reference sequence. It is agnostic of sequencing protocol and platform allowing use into the future. kWIP outputs a distance matrix between samples, which may be used to cluster or classify samples into unbiased groups even where differentiation is quite low. The distance matrices kWIP produces can also be used to quantitatively compare genetic distance to geographic or environmental distances via mantel tests or generalised dissimilarity modelling.

We have demonstrated the applicability and effectiveness of kWIP using simulations and several published datasets. Through simulations, we show that the novel weighting kWIP uses improves accuracy specifically in cases where there is little genetic differentiation. Using the 3000 rice genome dataset [40], we reconstruct known relationships between and memberships to all major genetic groups of *Oryza sativa*, and observe the correct clustering of technical and biological replicates (see Figure 2). Using a population re-sequencing experiment in *Chlamydomonas* [11] we precisely recreate a visualisation of population relatedness, arguably improving resolution compared to a reference-genome based variant calling approach (Figure 3).

The central innovations of kWIP are the use of a fixed-sized, probabilistic data structure for counting k-mers, and the use of a weighted inner product as a measure of similarity between samples. Such data structures (e.g. Count-Min Sketches and Bloom Filters) have been used extensively for k-mer counting by many previous tools (e.g. [8, 30]), however these data structures have not been widely adopted by alignment free sequence comparison algorithms. Weighting of the inner product addresses both the noise from rare variants and sequencing errors. It also addresses noise due to variable coverage across the genome that leads to fixed sites not being observed in all samples.

An inner product between k-mer counts has long been used in the field of alignment free phylogenetics, where it is referred to as the D_2 statistic. There have been many derivatives of the D_2 statistic that seek to enhance its accuracy where evolutionary distance is large and sites may have mutated multiple times (e.g. D_2^S and D_2^* [36, 43, 2]). These enhanced models of sequence evolution are not necessary where mutation events occur at independent sites as is usually the case within populations (see Figure 1b). In addition, inner products between sequences have been used as a kernel for protein sequence classification algorithms [26, 25].

Further improvements in kWIP's performance could be made in a variety of ways. By applying smoothing to sample k-mer counts, one can differentiate between k-mers that are genuinely not in the genomes of a sample and those that were not observed due to low coverage or stochastic sampling of the genome. This smoothing approach has been used to account for unseen n-grams in natural language modelling [7], a conceptually similar problem. It is possible that an alternative distance metric such as the Jaccard distance or Manhattan distance could improve the performance of kWIP, especially if combined with entropy weighting as such metrics have been used previously for alignment free sequence comparison [29, 35]. Furthermore, alternative kernel functions including those considering inexact matches could be investigated [27, 25]. Finally, kWIP currently scales quadratically ($\mathcal{O}(n^2)$) with regards to the

number of samples, making the analysis of very large data sets (with 10,000s of samples) computationally expensive. To resolve this limitation, a classification algorithm utilising the WIP metric as a kernel function could be used to compare samples against a core set without computing all pair-wise distances. Sequence Bloom Trees use a similar approach to classify unassembled datasets of sketched k-mers [13, 37].

The next frontier is characterisation of complex metagenome samples. Recently, several researchers have applied methods conceptually similar to `kWIP` for metagenome similarity calculation [35]. Here, we show that `kWIP` is able to recreate microbial community structure. In addition to simply calculating the similarities between metagenomes, the calculation of intra- and inter-sample diversity is often of primary interest. The ability to calculate these measures *de novo* would be highly useful due to the difficulty of assembling metagenomes accurately.

In most population based experiments, technical sources of error are dwarfed by the error from insufficient sampling. This is especially true when rare cryptic lineages are present, and in conditions of non-random mating where population structure is substantial. Such population level noise can only be overcome with large, broad studies informed by unbiased preliminary genomic analysis. The use of `kWIP` to process low coverage whole genome sequence data makes feasible these large, more balanced study designs which aim to assess the major groupings among individuals and their genome wide diversity. Focused genetic mapping study designs split major groups and select a single member from each family, creating subsets with similar relatedness [4]. As sample, library, and sequencing costs decline and tools such as `kWIP` become utilised, more large data sets that allow joint analysis of diversity among samples and across the genome will be generated.

6 Acknowledgements

We thank Sylvain Fort, Terry Neeman, Conrad Burden, Gavin Huttley, Ben Kaehler and Cameron Jack for comments and advice on the metrics, algorithms, and experiments reported here. We thank Luisa Teasdale for comments on earlier drafts of this manuscript. This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

References

[1] Saulo Alves Aflitos et al. "Cnidaria: Fast, Reference-Free Clustering of Raw and Assembled Genome and Transcriptome NGS Data". In: *BMC Bioinformatics* 16 (2015), p. 352.

[2] Elizabeth S. Allman, John A. Rhodes, and Seth Sullivant. "Statistically-Consistent K-Mer Methods for Phylogenetic Tree Reconstruction". In: (Nov. 5, 2015).

[3] Joy Bergelson et al. "A Proposal Regarding Best Practices for Validating the Identity of Genetic Stocks and the Effects of Genetic Variants". In: *The Plant Cell* 28.3 (Mar. 1, 2016), pp. 606–609.

[4] Benjamin Brachi, Geoffrey P Morris, and Justin O Borevitz. "Genome-Wide Association Studies in Plants: The Missing Heritability Is in the Field". In: *Genome biology* 12.10 (2011), p. 232.

[5] Dbora Y. C. Brandt et al. "Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Project Phase I Data". In: *G3: Genes—Genomes—Genetics* 5.5 (May 1, 2015), pp. 931–941.

[6] Reed A. Cartwright. "DNA Assembly with Gaps (Dawg): Simulating Sequence Evolution". In: *Bioinformatics* 21 (Suppl 3 Jan. 11, 2005), pp. iii31–iii38.

[7] Stanley Chen and Joshua Goodman. "An Empirical Study of Smoothing Techniques for Language Modeling". In: *PROCEEDINGS OF THE 34TH ANNUAL MEETING OF THE ACL*. 1996, pp. 310–318.

[8] Michael R. Crusoe et al. "The Khmer Software Package: Enabling Efficient Nucleotide Sequence Analysis". In: *F1000Research* (Sept. 25, 2015).

[9] Joseph Edwards et al. "Structure, Variation, and Assembly of the Root-Associated Microbiomes of Rice". In: *Proceedings of the National Academy of Sciences* 112.8 (Feb. 24, 2015), E911–E920.

[10] Huan Fan et al. "An Assembly and Alignment-Free Method of Phylogeny Reconstruction from next-Generation Sequencing Data". In: *BMC Genomics* 16.1 (July 14, 2015), p. 522.

[11] Jonathan M. Flowers et al. "Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga Chlamydomonas Reinhardtii". In: *The Plant Cell* 27.9 (Sept. 1, 2015), pp. 2353–2369.

[12] Sylvain Fort, Susan R Wilson, and Conrad J Burden. "Characterizing the D2 Statistic: Word Matches in Biological Sequences". In: *Statistical Applications in Genetics and Molecular Biology* 8.1 (2009), pp. 1–21.

[13] Simon Gog et al. "From Theory to Practice: Plug and Play with Succinct Data Structures". In: *Experimental Algorithms: 13th International Symposium, SEA 2014, Copenhagen, Denmark, June 29 July 1, 2014. Proceedings.* Ed. by Joachim Gudmundsson and Jyrki Katajainen. Cham: Springer International Publishing, 2014, pp. 326–337.

[14] Gal Guennebaud, Benot Jacob, et al. *Eigen v3*. 2010.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York, 2009.

[16] Thomas Hofmann, Bernhard Schlkopf, and Alexander J. Smola. "Kernel Methods in Machine Learning". In: *The Annals of Statistics* 36.3 (June 2008), pp. 1171–1220.

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

[18] Zamin Iqbal et al. "De Novo Assembly and Genotyping of Variants Using Colored de Bruijn Graphs". In: *Nature Genetics* 44.2 (Jan. 8, 2012), pp. 226–232.

[19] Hongshan Jiang et al. "Skewer: A Fast and Accurate Adapter Trimmer for next-Generation Sequencing Paired-End Reads". In: *BMC Bioinformatics* 15.1 (June 12, 2014), p. 182.

[20] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open Source Scientific Tools for Python.* [Online; accessed 2016-09-15]. 2001.

[21] N A Joshi and J N Fass. *Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files.* Version 1.33. 2011.

[22] Johannes Kster and Sven Rahmann. "Snakemake — a Scalable Bioinformatics Workflow Engine". In: *Bioinformatics* 28.19 (Jan. 10, 2012), pp. 2520–2522.

[23] Chris-Andre Leimeister and Burkhard Morgenstern. "Kmacs: The K-Mismatch Average Common Substring Approach to Alignment-Free Sequence Comparison". In: *Bioinformatics* (May 13, 2014), btu331.

[24] Chris-Andre Leimeister et al. "Fast Alignment-Free Sequence Comparison Using Spaced-Word Frequencies". In: *Bioinformatics* (Apr. 3, 2014), btu177.

[25] Christina S. Leslie et al. "Mismatch String Kernels for Discriminative Protein Classification". In: *Bioinformatics (Oxford, England)* 20.4 (Mar. 1, 2004), pp. 467–476.

[26] Christina Leslie, Eleazar Eskin, and William Stafford Noble. "The Spectrum Kernel: A String Kernel for SVM Protein Classification". In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2002). key: leslie_spectrum_2002, pp. 564–575.

[27] Christina Leslie and Rui Kuang. "Fast String Kernels Using Inexact Matching for Protein Sequences". In: *J. Mach. Learn. Res.* 5 (Dec. 2004), pp. 1435–1455.

[28] Jia-Yang Li, Jun Wang, and Robert S. Zeigler. "The 3,000 Rice Genomes Project: New Opportunities and Challenges for Future Rice Research". In: *GigaScience* 3.1 (May 28, 2014), p. 8.

[29] Yang Young Lu et al. "COCACOLA: Binning Metagenomic Contigs Using Sequence COmposition, Read CoverAge, CO-Alignment, and Paired-End Read LinkAge". In: *Bioinformatics* (June 2, 2016), btw290.

[30] Pll Melsted and Jonathan K. Pritchard. "Efficient Counting of K-Mers in DNA Sequences Using a Bloom Filter". In: *BMC bioinformatics* 12 (2011), p. 333.

[31] Michael L. Metzker. "Sequencing Technologies the next Generation". In: *Nature Reviews Genetics* 11.1 (Jan. 2010), pp. 31–46.

[32] Burkhard Morgenstern et al. "Estimating Evolutionary Distances between Genomic Sequences from Spaced-Word Matches". In: *Algorithms for Molecular Biology* 10.1 (Feb. 11, 2015), p. 5.

[33] Kevin Murray. "SRApy: Pythonic Tools for Accessing the Short Read Archive". In: *Zenodo* (May 2016).

[34] Bjrn Nystedt et al. "The Norway Spruce Genome Sequence and Conifer Genome Evolution". In: *Nature* 497.7451 (May 30, 2013), pp. 579–584.

[35] Brian D. Ondov et al. "Mash: Fast Genome and Metagenome Distance Estimation Using MinHash". In: *Genome Biology* 17 (2016), p. 132.

[36] Gesine Reinert et al. "Alignment-Free Sequence Comparison (I): Statistics and Power". In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 16.12 (Dec. 2009), pp. 1615–1634.

[37] Brad Solomon and Carl Kingsford. "Fast Search of Thousands of Short-Read Sequencing Experiments". In: *Nature Biotechnology* 34.3 (Mar. 2016), pp. 300–302.

[38] Paul R. Staab et al. "Scrm: Efficiently Simulating Long Sequences Using the Approximated Coalescent with Recombination". In: *Bioinformatics* 31.10 (May 15, 2015), pp. 1680–1682.

[39] Jie Tang et al. "A Novel K-Word Relative Measure for Sequence Comparison". In: *Computational Biology and Chemistry* 53, Part B (Dec. 2014), pp. 331–338.

[40] The 3,000 rice genomes project. "The 3,000 Rice Genomes Project". In: *GigaScience* 3.1 (May 28, 2014), p. 7.

[41] The Arabidopsis Genome Iniative. "Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana". In: *Nature* 408.6814 (Dec. 14, 2000), pp. 796–815.

[42] Susana Vinga. "Information Theory Applications for Biological Sequence Analysis". In: *Briefings in Bioinformatics* 15.3 (Jan. 5, 2014), pp. 376–389.

[43] Lin Wan et al. "Alignment-Free Sequence Comparison (II): Theoretical Power of Comparison Statistics". In: *Journal of Computational Biology* 17.11 (Oct. 25, 2010), pp. 1467–1490.

[44] Qingpeng Zhang et al. "These Are Not the K-Mers You Are Looking For: Efficient Online K-Mer Counting Using a Probabilistic Data Structure". In: *PLoS ONE* 9.7 (July 25, 2014), e101271.

A Supplimentary Methods

Software	Version	URL/Citation
`khmer`	2.0 (c2cc671)	"Zhang et al. [44] and Crusoe et al. [8]"
`kwip`	0.2.0 (ce9b1db)	https://github.com/kdmurray91/kwip
`seqtk`	r93 (c282ed8)	https://github.com/lh3/seqtk
`dawg2`	commit f9ebcd8	https://github.com/reedacartwright/dawg
`sickle`	1.33 (f3d6ae3)	https://github.com/najoshi/sickle
`mason2`	2.0.3 (88e3a0b)	Holtgrewe [17]
`scikit-bio`	0.5.0	https://github.com/biocore/scikit-bio
`scipy`	0.18.0	Jones, Oliphant, Peterson, et al. [20]
`skewer`	0.2.2 (978e8e4)	Jiang et al. [19]
`sra-py`	0.1.15 (fb35481)	Murray [33]
`scrm`	1.7.0	Staab et al. [38]

Table 1: Summary of Software Versions. Versions refer to the precice installed version, and where available, git commit identifiers are given in parentheses.