

# **Comparative genomics in conservation biology from an reference-free perspective**

**MATTHEW R. HELMUS, CHAI SHIAN KUA, CHARLES H. CANNON**

*Xishuangbanna Tropical Botanic Garden, Chinese Academy of Sciences*

**mrhelmus@gmail.com**

# Summary

Genomic data for any taxa can now be obtained on even modest research budgets, but translation of these data into meaningful results for conservation remains a challenge. Traditional analyses rely upon the construction of a physical map of a reference genome, which remains a costly and time-consuming endeavor. We propose that conservation biologists can best exploit short-read genomic sequence (SRS) data from nonmodel organisms by adopting a reference-free perspective—also termed assembly and alignment-free. Using reference-free techniques, sequence differences among samples (e.g., individuals, species) are directly discovered without any prior genomic knowledge.

One reference-free technique is to compare genomic sequences truncated to a short length  $k$ ,  $k$ -mers (**fig. 1**). Frequencies of  $k$ -mers shared among samples can be used for phylogenetic reconstruction with distance and parsimony methods (**fig. 2**), and to give estimates of genomic variation (**fig. 3**).  $k$ -mers that unify groups of samples can be identified (**see inset fig. 1**), and localized *de-novo* assembly of candidate markers can be performed (**fig. 5**). We recently tested this method on simulated SRS data from 104 assemble chloroplast genomes (**see Current Work**).

A second reference-free technique is to assemble *de novo* SRS data into contigs (**fig. 1**). The contigs are aligned among samples into clusters and polymorphisms are identified as candidate markers (**fig. 5**). Population genetic parameters can then be estimated (**fig. 4**).

Reference-free techniques have the potential to expand the taxonomic scope and utility of breakthroughs in DNA sequencing technology. Taking a reference-free perspective allows for rapid discovery of genetically based parameters and variation in the context of many questions in conservation biology.

# k-mers (fig. 1)

**extract DNA/RNA**

**preprocess samples**

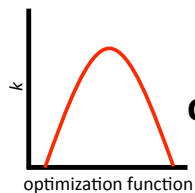
reduced complexity library  
chloroplast enrichment  
identifier tag addition, etc.

**high-throughput sequencing**



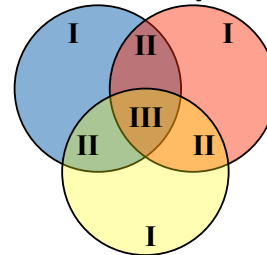
**short-read sequences (SRS)**

sample 1	sample 2...
1. GCTTCGCCCAACAAAATAGTACATTTAATGAAGATG	1. CACCAACATCATATATGTATTTAGTACATTTAATGAA
2. ACATTATATGTACATATATGTATTACCTGCCGAGCTT	2. CATATATGTATTACCTGCCGAGCTTTTAGCATCCATC
3. CCCCAGCAAGATCATATATGTATTATCCATTATCTAA	3. TCCCACCAACATCATATATGTATTATCCATTATCTAA
4. TAAACTATTCTTTGCCGAGCTTCCGCCCAAAAATA	4. ACAACTATTCTTTGCCGAGCTTTCATTTTATGTACA
5. TGTTCATTTTATGTACATATATGTATTATGCCCATTA	5. ATATTTTATACATCCAAATGTATTATCCTGTTCTTATG



**optimize length (i.e.,  $k$ )**  
(e.g., Sims *et al.* 2009)

**k-mer overlap in three samples**



k-mer types:  
I unique to a sample  
II exclusive to some samples  
III in all samples  
(Cannon *et al.* 2010)

**estimate k-mer frequencies**

```

1. CACCAACATCATATATGTATTTAGTACATTTAATGAA
1. CACCAACATCATATATGTATTTAGTACATTTAATGAA
1. CACCAACATCATATATGTATTTAGTACATTTAATGAA
...
1. CACCAACATCATATATGTATTTAGTACATTTAATGAA
2. CATATATGTATTACCTGCCGAGCTTTTAGCATCCATC
    
```

**k-mer filtering**

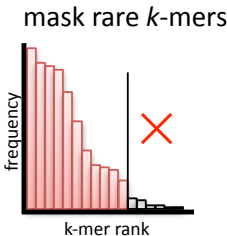
target a k-mer type

mask low-complexity k-mers

```

GCTTCGCCAC
TATATATATA
TACATTTAAT
GGGGGGGGGG
GATGCATATA
TGATTACCT
AAAAATTTT
TAGCATCCAT
    
```

etc.



mask rare k-mers

**de novo assembly of SRS**  
(with filtered k-mers masked)

```

GCTTAATGAAGATG
AATGAAGATGCCCAACCAAC
AAGATGCCCAACCAACGCCAA
ATGCCCAACCAACGCCAACATCATAT
CCACCAACGCCAACATCATATATGTATT
CGCACCAACATCATATATGTATTATCCATTATCTAAAAATTTTT
contig 1: GCTTAATGAAGATGCCCAACCAACGCCAACATCATATATGTATTATCCATTATCT
contig 2: AAGAGCTGAAGATGCCCTTCTTCCGCCAACCTTAAACGCCAACCTGAAGATGCCCA
    
```

(Ratan *et al.* 2010)

**comparative analysis**

# Assembly/Alignment-Free Phylogenetics (fig. 2)

## k-mer frequency table

k-mer	sample							
	A	B	C	D	E	F	G	H
GCTTCGCCCC	10	20	23	0	4	0	3	0
ACCAAAATAG	2	5	0	0	10	34	20	20
TACATTTAAT	0	0	0	0	0	3	0	0
GAAGATGTCT	32	45	67	10	13	12	10	9
...								
AAGTTAAACT	0	0	0	0	15	14	12	14
ATTCTTTGCC	4	0	6	5	0	0	0	0
GAGCTTCGCC	12	19	13	4	4	0	27	28

## pairwise distance metric

e.g., Euclidean distance, Jensen-Shannon divergence

## distance matrix

	A	B	C	D	E	F	G	H
A	1	0.08	0.08	1	1	1	1	1
B	0.08	1	0.03	1	1	1	1	1
C	0.08	0.03	1	0	1	1	1	1
D	1	1	1	1	0.65	0.65	0.65	0.65
E	1	1	1	0.65	1	0.02	0.02	0.05
F	1	1	1	0.65	0.02	1	0.02	0.05
G	1	1	1	0.65	0.02	0.02	1	0.05
H	1	1	1	0.65	0.05	0.05	0.05	1

## phylogenetic reconstruction

e.g., neighbor joining, UPGMA

(e.g., Sims et al. 2009)

## k-mer character matrix

k-mer	A	B	C	D	E	F	G	H
GCTTCGCCCC	1	1	1	0	1	0	1	0
ACCAAAATAG	1	1	0	0	1	1	1	1
TACATTTAAT	0	0	0	0	0	1	0	0
GAAGATGTCT	1	1	1	1	1	1	1	1
...								
AAGTTAAACT	0	0	0	0	1	1	1	1
ATTCTTTGCC	1	0	1	1	0	0	0	0
GAGCTTCGCC	1	1	1	1	1	0	1	1

## phylogenetic reconstruction

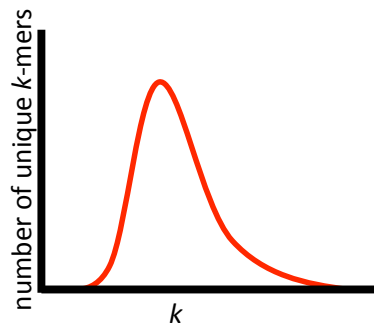
e.g., parsimony, but evolutionary models need to be tested/made for maximum likelihood, Bayesian.

(Cannon et al. in prep)

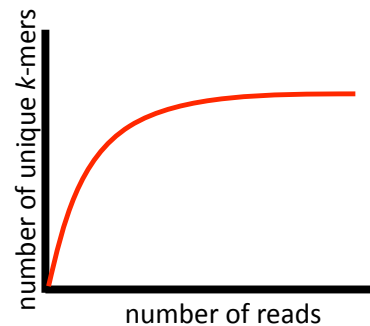
# Genomic Variation (fig. 3)

estimates based on variation in  $k$ -mer frequencies (fig. 2)

unique  $k$ -mer distribution



accumulation curve for a given  $k$

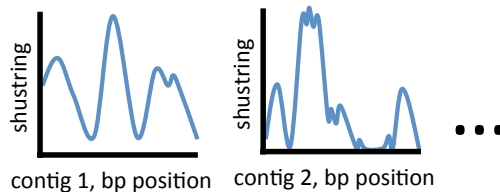


nucleotide diversity based on  $k$ -mers as restriction sites  
(Nei and Li 1979, Pavoine and Bailly 2007)

$$\pi = \sum_{i=1}^n \sum_{j=1}^i x_i x_j \pi_{ij}$$

estimates based on polymorphisms in contigs (fig. 4, 5)

shortest unique substrings  
(Haubold B. & Wiehe 2006)



estimate  $\vartheta$  from contigs and SNPs (fig. 5)  
(e.g Hellmann et al. 2008)

$$\theta = 4N_e\mu$$

estimate heterozygosity from contigs and SNPs (fig. 5)  
(Lynch 2008)

contig 1: GCTTAATGAAGATGCCCCACCAACGCACCAACATCATATATGTATTATCCATTATCTAA  
contig 2: AAGAGCTGAAGATGCCCTTCTTCGGCACCAACTTAAACGCACCAACTGAAGATGCCCC

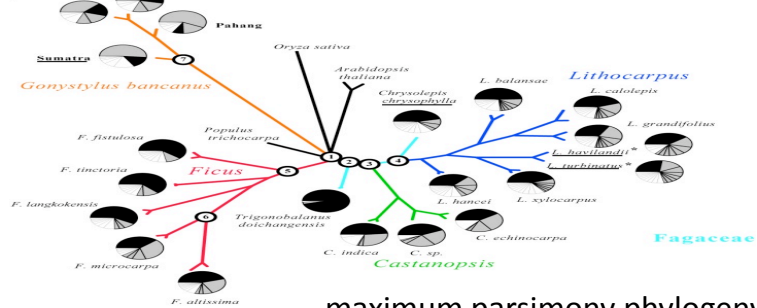
...

# Current Work

## Methods

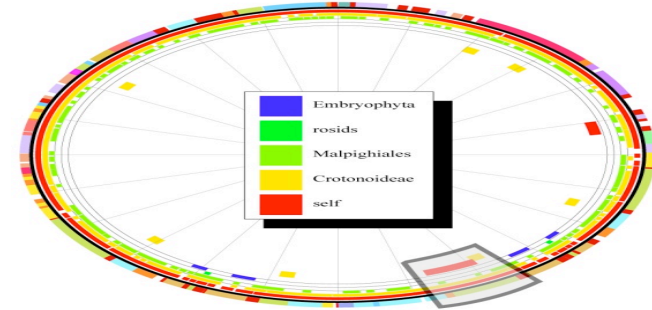
contact (Chuck Cannon: [chuck.cannon@gmail.com](mailto:chuck.cannon@gmail.com))

### reference-free phylogenetics



maximum parsimony phylogeny  
of South-East Asian tropical tree  
clades

### reference-free marker identification



proof of concept: markers that unify assembled  
chloroplast genomes at various taxonomic levels

## Applications

genetically identify tropical timber products

contact (Chai Shian Kua: [cskua1@gmail.com](mailto:cskua1@gmail.com))



transcriptomics of phenotypic plasticity

contact (Jocelyn Behm: [jebehm@wisc.edu](mailto:jebehm@wisc.edu))



# Population Genetics (fig. 4)

align/cluster contigs across samples

```
ind 1: GCTTAATGAAGATGCCCCACCAACGCTCCAAAAACATCATATATGTATTA
ind 2:      ATGCCCCACCAACGCACCAAAAAACATGATATATGTATTATC
ind 3:      GAAGATGCCCCACCAACGCACCAAAAAACATCATATATGTAT
...
```

align/cluster SRS data across samples

```
ind 1: GCTTAATGAAGATGC
ind 1:  AATGAAGATGCCCCACCAAC
ind 2:  AAGATGCCCCACCA
ind 3:  AATGAAGATGCCCCACCAACGCA
ind 2:  TGAAGATGCCCCACCAACGCACCAACAT
ind 1:  CCCACCAACGCACCAA
```

OR

find variant positions in each cluster

cluster 1

```
ind 1: GCTTAATGAAGATGCCCCACCAACGCTCCAAAAACATCATATATGTATTA
ind 2:      ATGCCCCACCAACGCACCAAAAAACATGATATATGTATTATC
ind 3:      GAAGATGCCCCACCAACGCACCAAAAAACATCATATATGTAT
...
```

cluster 2

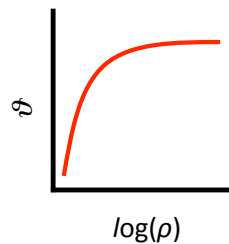
```
ind 1:      GCCCCTTCTCCGCACCAACTTAACGCACCAACTGAAGATGCCCC
ind 2: AAGAGCTGAAGATGCCCTTTTCCGCACCAACTTAACGCACCAACTGAAGA
ind 3:      AGATGCCCTTTTCCGCACCAACTTAACGCACCAACTGAGGATGCC
...
```

estimate population-genetic parameters

$$\theta = 4N_e\mu \quad \rho = 4N_e c \quad \epsilon_{sequencing} \quad LD = r_{XY}^2$$

composite likelihood estimators  
(Hellmann *et al.* 2008; Jiang *et al.* 2009)

$$L(\rho) = \prod p(X | \rho)$$

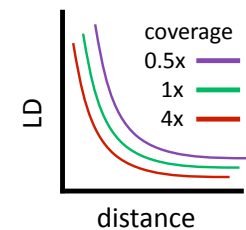


$\vartheta$  based on modified Watterson estimator

$$\hat{\theta}_w = \frac{K}{a_n}$$

(Johnson and Slatkin 2006, 2008; Lynch 2008; Haubold *et al.* 2010)

linkage disequilibrium distance decay  
(Jiang *et al.* 2009)



# Definitions

**sliding window** – a bioinformatic technique where a fixed length is slid along a DNA sequence at a fixed steps in order to tabulate  $k$ -mers or calculate summary statistics. To tabulate  $k$ -mers the window length is  $k$ .

**$k$ -mer frequencies** – the relative abundances of unique  $k$ -mers in a genomic data set (e.g., a SRS data set). Also called a feature frequency profile.

**sequence cluster** – a set of overlapping sequences (e.g.,  $k$ -mers, SRS) from multiple genomes or multiple reads of a genome that are from homologous genomic regions and are thus alignable. Also called a super island.

**de-novo assembly** – the assembly of reads from the sequence clusters of one chromosome without the use of a reference sequence.

**contig** – a completely sequenced fragment from one chromosome assembled from a sequence cluster. The term is also sometimes used to mean sequence cluster.

**masked sequence** – a sequence (e.g., a  $k$ -mer) within a larger sequence (e.g., a SRS) that is not used when performing bioinformatic analyses such as assembly and alignment of contigs.

**genome coverage** – the average number of reads per nucleotide in a genome. Coverage affects  $k$ -mer frequencies, the ability to cluster sequences, and estimates of population genetic parameters based on SRS data.

**genome enabled taxa** – taxa that are closely related to a species whose genome has been well studied, assembled and mapped.

**genomic/genetic markers** – sequences of DNA that differentiate and group genomes (e.g., single nucleotide polymorphism). From a reference-free perspective, the chromosomal location of markers does not, at least initially, need to be identified.

**sequence homoplasy** – the convergence of DNA to similar sequences through mutation either with or without selection. Homoplasy may cause errors in reference-free analyses when the analyses rely on the assumption of sequence orthology (i.e., sequences derived from a common ancestor and not due to duplication events).

**sequence paralogy** – similar sequences either within or among genomes derived from a duplication event. Relative  $k$ -mer frequencies can be strongly dependent on duplications.

## Reference-Free Software

(working list)

### Phylogenetics

*KR* (Haubold et al. 2008)

### Genomic Variation

*kmer\_count* (Cannon et al. in prep)  
*ade4 R package* (Pavoine 2007)

### Population genetics

*mIRho* (Haubold et al. 2010)

### Marker Identification

*DIAL* (Ratan et al. 2010)  
*kmer\_count* (Cannon et al. in prep)  
*shustring* (Haubold et al. 2005)



# References

- Cannon C.H., Kua C.-S., Zhang D. & Harting J.R. (2010). Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack. *Molecular Ecology*, 19 146–160.
- Cannon C.H., Ruan J., Helmus M.R., Zhang D. & Kua C.-S. (in prep). Assembly-free phylogenomic analysis of next-gen sequence data.
- Haubold B., Pierstorff N., Moller F. & Wiehe T. (2005). Genome comparison without alignment using shortest unique substrings. *BMC Bioinformatics*, 6, 123.
- Haubold B. & Wiehe T. (2006). How repetitive are genomes? *BMC Bioinformatics*, 7, 541.
- Haubold B., Pfaffelhuber P. & Lynch M. (2010). mlRho--A program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Molecular Ecology*, 19, 277-284.
- Hellmann I., Mang Y., Gu Z., Li P., de la Vega F.M., Clark A.G. & Nielsen R. (2008). Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome research*, 18, 1020-1029.
- Jiang R., Tavaré S. & Marjoram P. (2009). Population Genetic Inference From Resequencing Data. *Genetics*, 181, 187-197.
- Johnson P.L.F. & Slatkin M. (2006). Inference of population genetic parameters in metagenomics: A clean look at messy data. *Genome research*, 16, 1320-1327.
- Johnson P.L.F. & Slatkin M. (2008). Accounting for Bias from Sequencing Error in Population Genetic Estimates. *Mol Biol Evol*, 25, 199-206.
- Lynch M. (2008). Estimation of Nucleotide Diversity, Disequilibrium Coefficients, and Mutation Rates from High-Coverage Genome-Sequencing Projects. *Mol Biol Evol*, 25, 2409-2419.
- Nei M. & Li W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76, 5269-5273.
- Pavoine S. & Bailly X. (2007). New analysis for consistency among markers in the study of genetic diversity: development and application to the description of bacterial diversity. *Bmc Evolutionary Biology*, 7, 156.
- Ratan A., Zhang Y., Hayes V., Schuster S. & Miller W. (2010). Calling SNPs without a reference sequence. *BMC Bioinformatics*, 11, 130.
- Sims G.E., Jun S.R., Wua G.A. & Kim S.H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 2677-2682.