

ReSkmer:

Modeling repeats improves population genomic distances between genome skims

by Eduardo Charvel, Isaac Thomas, Siavash Mirarab, Vineet Bafna



MIRARAB LAB



ALFRED P. SLOAN
FOUNDATION



Grainger
Bioinformatics
Center

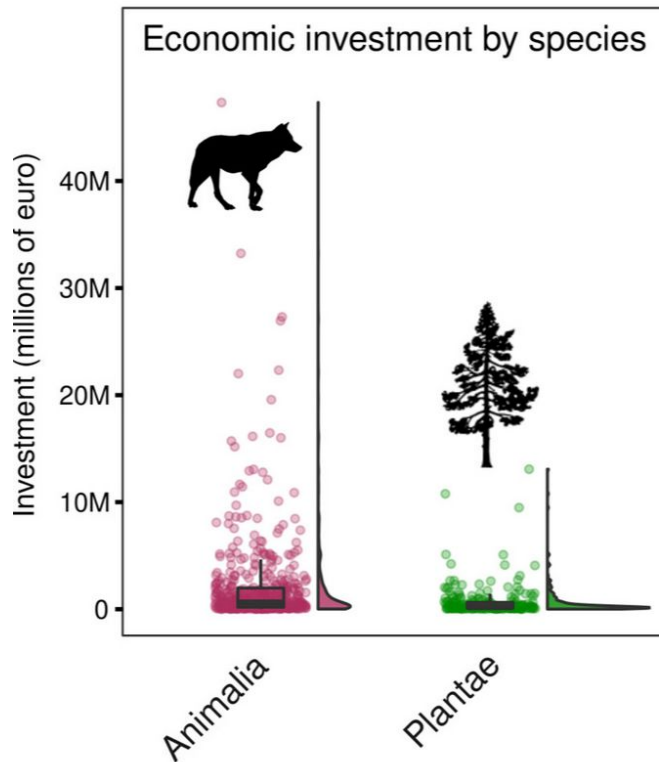
Accessible population genomics

Sequencing costs have
dropped.

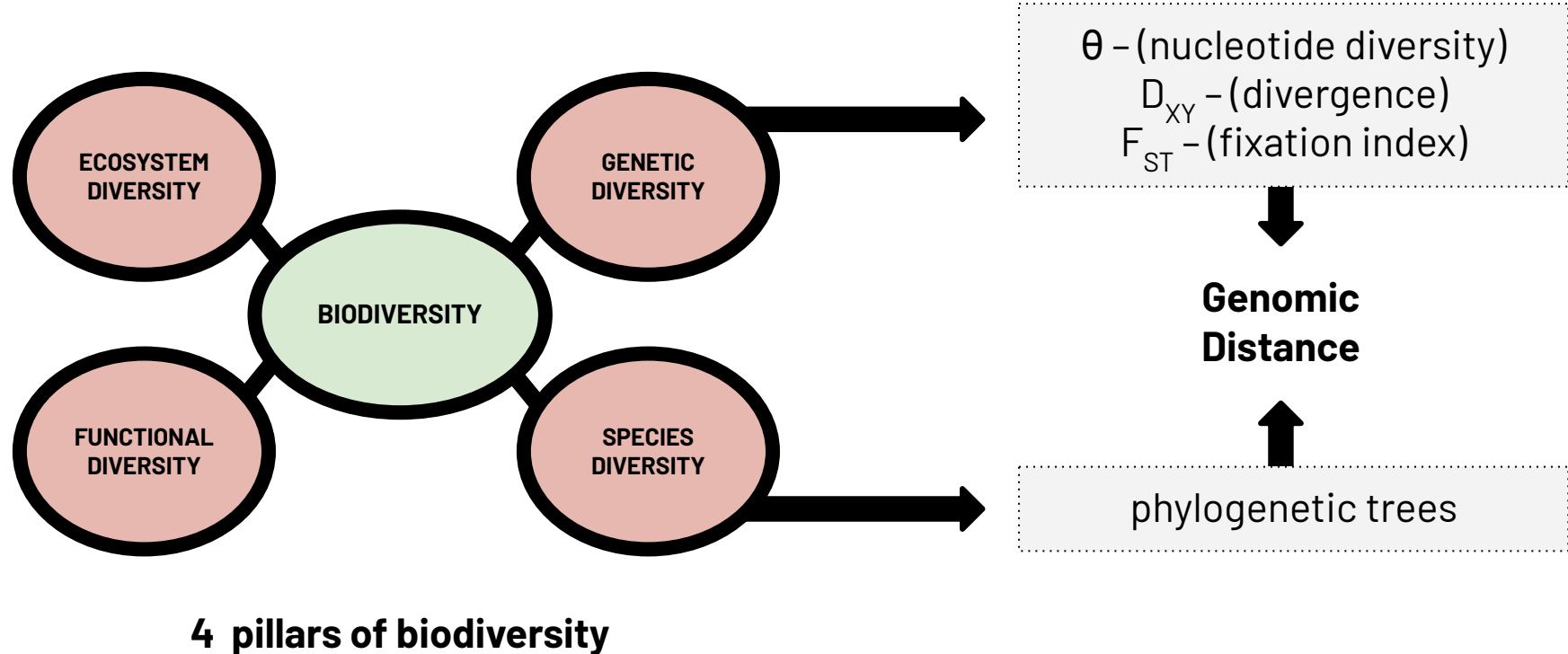
Conservation funding can be
biased.

Cost efficient methods for
measuring biodiversity are
crucial.

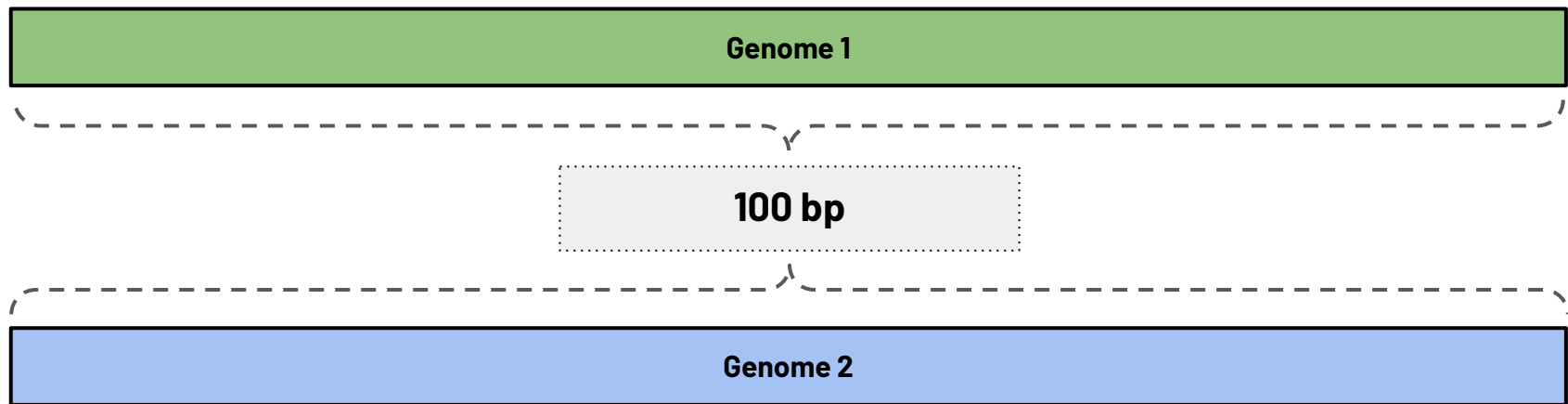
(Adamo et al. 2022)



How can we measure biodiversity efficiently?



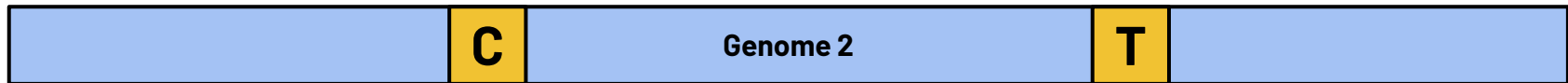
Genomic distance example



Genomic distance example

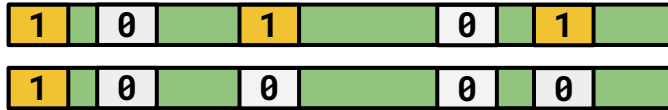


genomic distance
 $5/100 = 5\%$

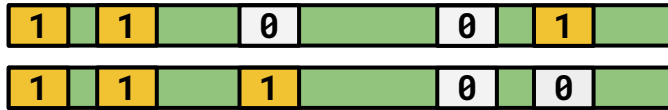


Genomic distances between populations

Population 1

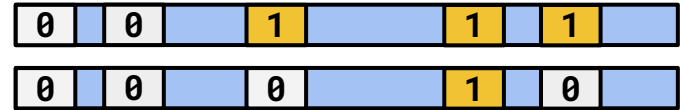


[...]

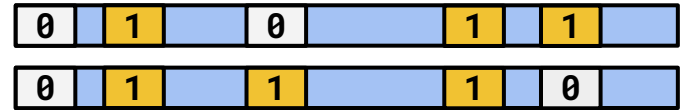


average genomic distance = 2%

Population 2



[...]



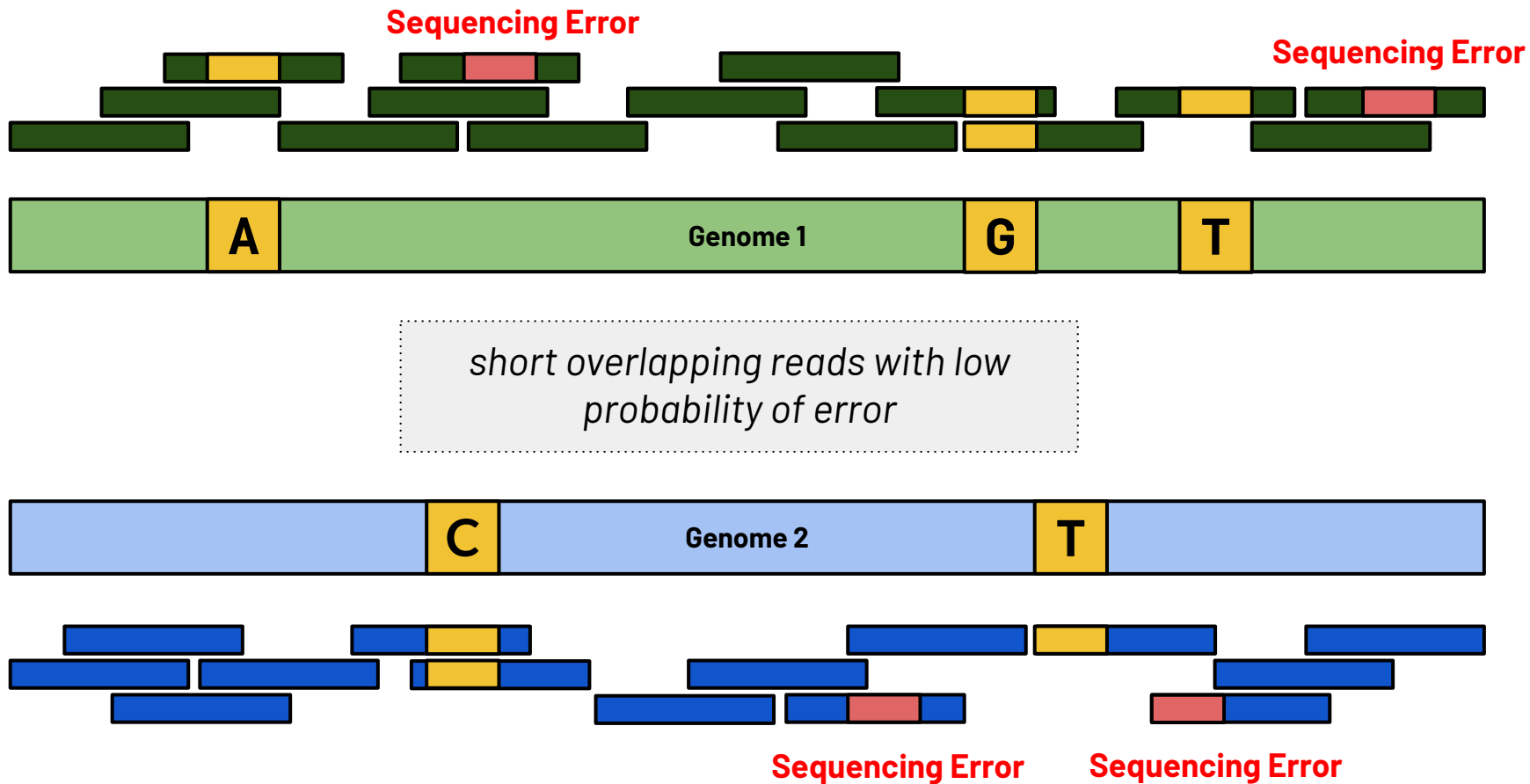
average genomic distance = 2%

average genomic distance = 3%

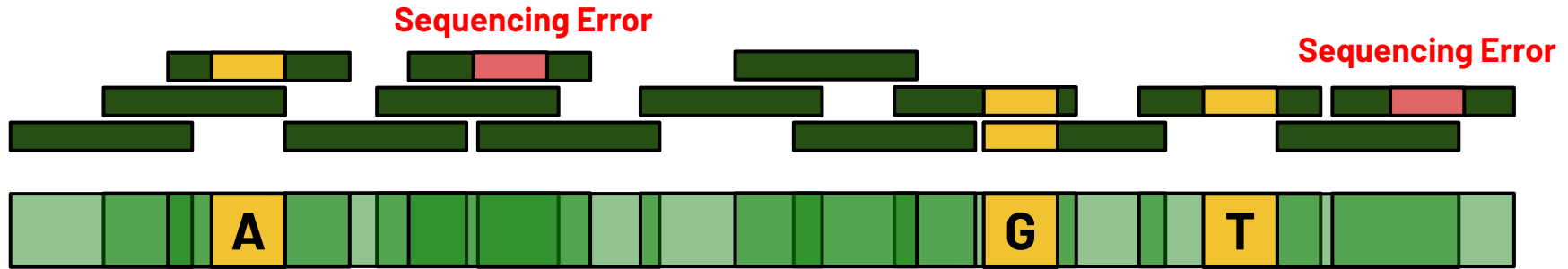
small values – population genomic distances

large values – phylogenetic distances

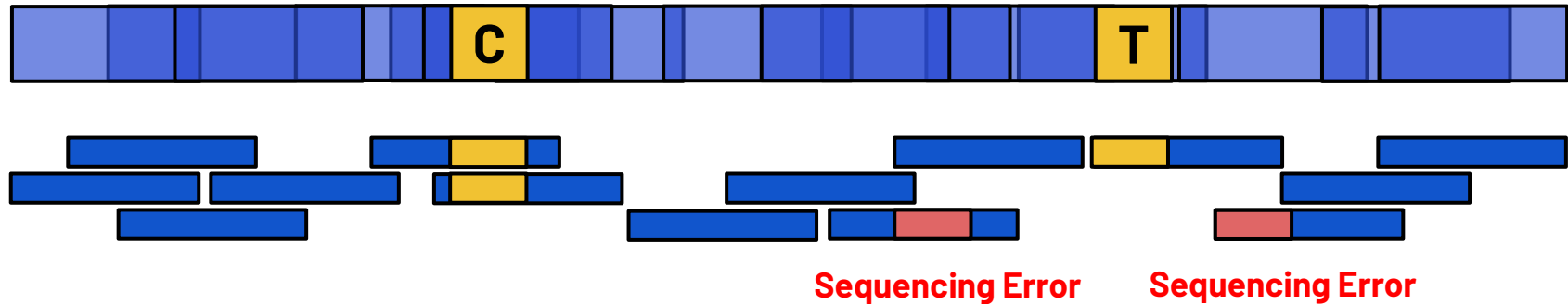
Short-read sequencing



Assembly / Alignment



sequences can then be **assembled/aligned**



Methods

Marker Based

*(single genes,
mitochondrial DNA,
microsatellites)*



Advantages

- Cost-effective Sanger Sequencing



Disadvantages

- Limited Resolution
- Requires Alignment

Methods

Marker Based

*(single genes,
mitochondrial DNA,
microsatellites)*



Advantages

- Cost-effective Sanger Sequencing



Disadvantages

- Limited Resolution
- Requires Alignment

RADseq

*(restriction enzyme
method, reduced
representation)*



- Cost-effective Method
- Reference Free



- Requires Alignment
- Reduced Resolution

Methods

Marker Based

(single genes,
mitochondrial DNA,
microsatellites)



Advantages

- Cost-effective Sanger Sequencing



Disadvantages

- Limited Resolution
- Requires Alignment

RADseq

(restriction enzyme
method, reduced
representation)



- Cost-effective Method
- Reference Free



- Requires Alignment
- Reduced Resolution

ANGSD

(SNP-based method)



- Chromosome-Level Resolution



- High Quality Assembly
- High(ish) Coverage
- Requires Alignment

Methods

Advantages

Disadvantages

Marker Based

(single genes,
mitochondrial DNA,
microsatellites)



- Cost-effective Sanger Sequencing



- Limited Resolution
- Requires Alignment

RADseq

(restriction enzyme
method, reduced
representation)



- Cost-effective Method
- Reference Free



- Requires Alignment
- Reduced Resolution

ANGSD

(SNP-based method)



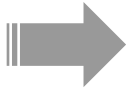
- Chromosome-Level Resolution



- High Quality Assembly
- High(ish) Coverage
- Requires Alignment

Mash

(k-mer based method,
short-read sequencing)



- Whole Genome Resolution
- Reference and Alignment Free



- High Coverage
- Assumes no repeats

Methods

Skmer

(*k*-mer based method,
short-read sequencing)



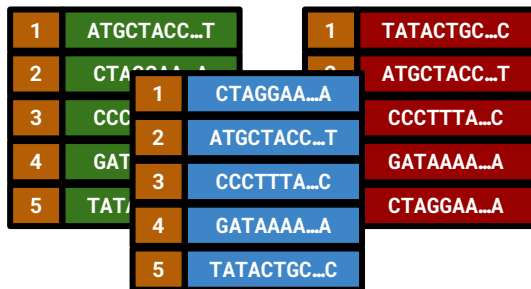
Advantages

- Whole Genome Resolution
- Reference and Alignment Free
- **Low Coverage**



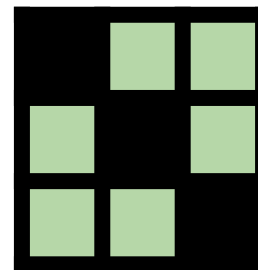
Disadvantages

- Assumes no repeats



fastq per sample

Input



distance matrix

Output

Skmer applications

Skmer approach improves species discrimination in taxonomically problematic genus *Schima* (Theaceae)

Han-Ning Duan ^{a,b}, Yin-Zi Jiang ^c, Jun-Bo Yang ^d, Jie Cai ^d, Jian-Li Zhao ^e, Lu Li ^b, Xiang-Qin Yu ^a

Chloroplast genome, nuclear ITS regions, mitogenome regions, and Skmer analysis resolved the genetic relationship among *Cinnamomum* species in Sri Lanka

Pradeepa C. G. Bandaranayake, Nathasha Naranpanawa, C. H. W. M. R. Bhagya Chandrasekara, Hiruna Samarakoon, S. Lokuge, S. Jayasundara, Asitha U. Bandaranayake, D. K. N. G. Pushpakumara, D. Siril A. Wijesundara

Published: September 20, 2023 • <https://doi.org/10.1371/journal.pone.0291763>

Method | [Open access](#) | Published: 13 February 2019

Skmer: assembly-free and alignment-free sample identification using genome skims

Shahab Sarmashghij, Kristine Bohmann, M. Thomas P. Gilbert, Vineet Bafna & Siavash Mirarab

Genome Biology 20, Article number: 34 (2019) | [Cite this article](#)

11k Accesses | 116 Citations | 10 Altmetric | [Metrics](#)

Genomic stability in the Galápagos *Scalesia* adaptive radiation: Consistent transposable element accumulation despite hybridization and ecological niche shifts

José Cerca, Patricia Jaramillo Díaz, Clément Goubert, Heidi Yang, Vanessa C. Bleker, Mario Fernández-Mazuecos, Pablo Vargas, Rowan Schley, Siyu Li, Juan Ernesto Guevara-Andino, Bent Petersen, Gitte Petersen, Neelima R. Sinha, Lene R. Nielsen, James H. Leebens-Mack, Gonzalo Rivas-Torres, Loren H. Rieseberg, Michael D. Martin
doi: <https://doi.org/10.1101/2024.09.30.614436>

Parallel introgression, not recurrent emergence, explains apparent elevational ecotypes of polyploid Himalayan snowtrout

Tyler K Chafin ^{1,2,*}, Binod Regmi ^{1,3}, Marlis R Douglas ¹, David R Edds ⁴, Karma Wangchuk ^{1,5}, Sonam Dorji ⁵, Pema Norbu ⁵, Sangay Norbu ⁵, Changlu Changlu ⁵, Gopal Prasad Khanal ⁵, Singye Tshering ⁵, Michael E Douglas ¹

► [Author information](#) ► [Article notes](#) ► [Copyright and License information](#)

PMCID: PMC8548808 PMID: [34729207](https://pubmed.ncbi.nlm.nih.gov/34729207/)

Phylogenetic Relationships and Next-Generation Barcodes in the Genus *Torreya* Reveal a High Proportion of Misidentified Cultivated Plants

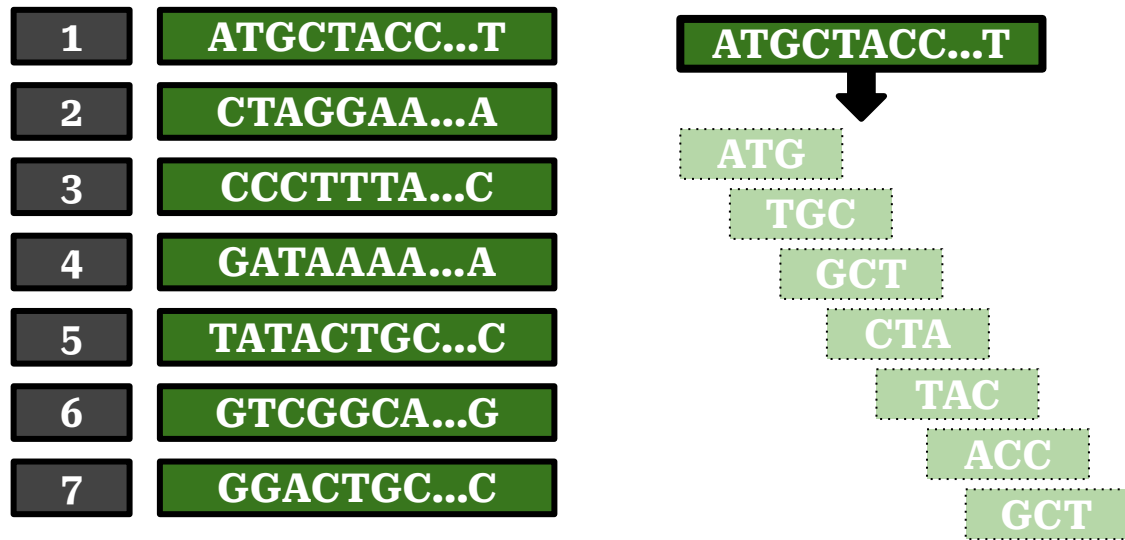
by Zhi-Qiong Mo ^{1,2,*}, Jie Wang ^{1,2,3,*}, Michael Möller ⁴, Jun-Bo Yang ³ and Lian-Ming Gao ^{1,5,*}

Alignment-free distance estimation

1	ATGCTACC...T
2	CTAGGAA...A
3	CCCTTTA...C
4	GATAAAA...A
5	TATACTGC...C
6	GTCGGCA...G
7	GGACTGC...C

begin collection of
reads ~ **150 bp**

Alignment-free distance estimation

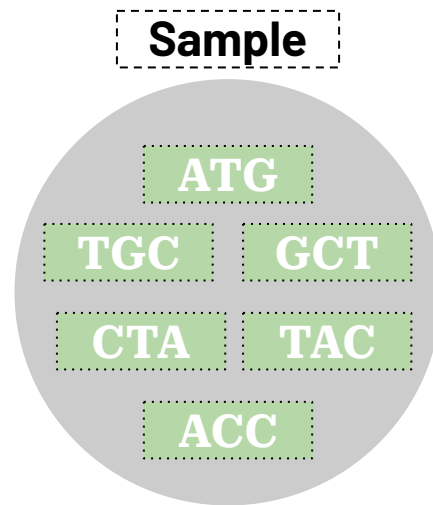
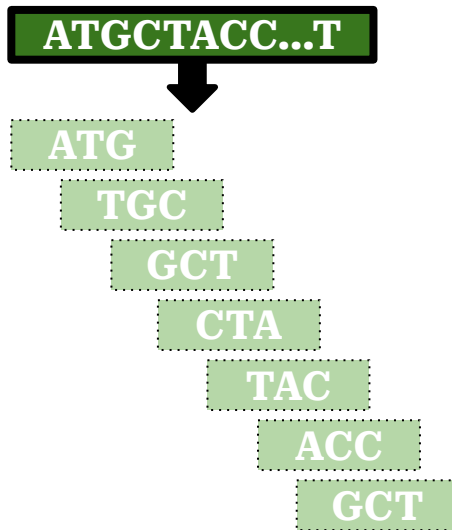


begin collection of
reads ~ **150 bp**

reads are decomposed
into **k-mers**

Alignment-free distance estimation

1	ATGCTACC...T
2	CTAGGAA...A
3	CCCTTTA...C
4	GATAAAA...A
5	TATACTGC...C
6	GTCGGCA...G
7	GGACTGC...C

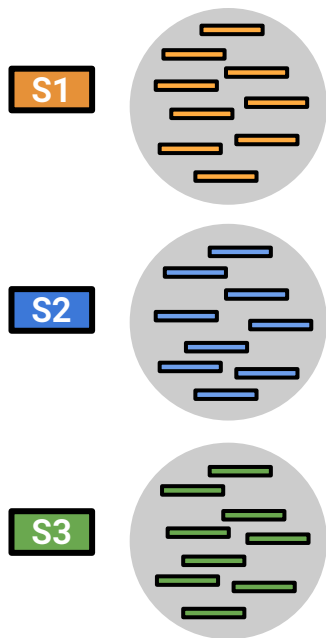


begin collection of
reads ~ **150 bp**

reads are decomposed
into **k-mers**

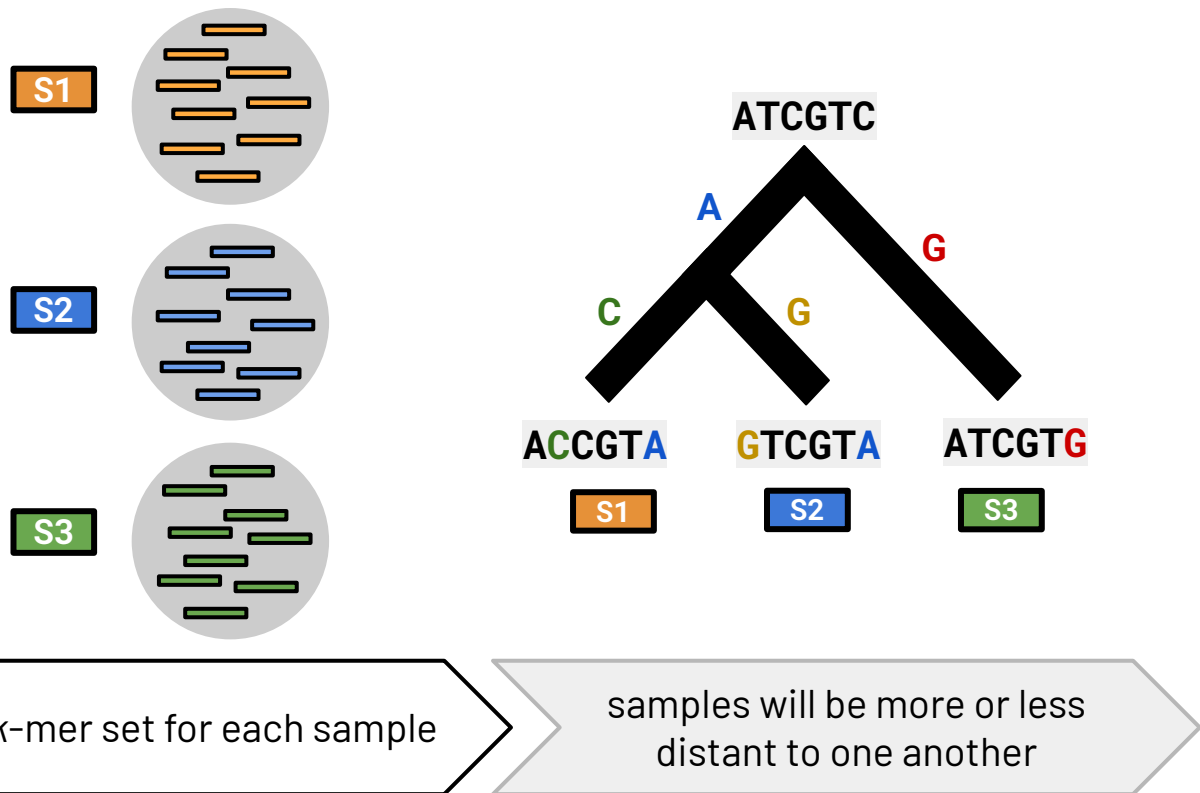
sets store **unique** k-mers

Alignment-free distance estimation

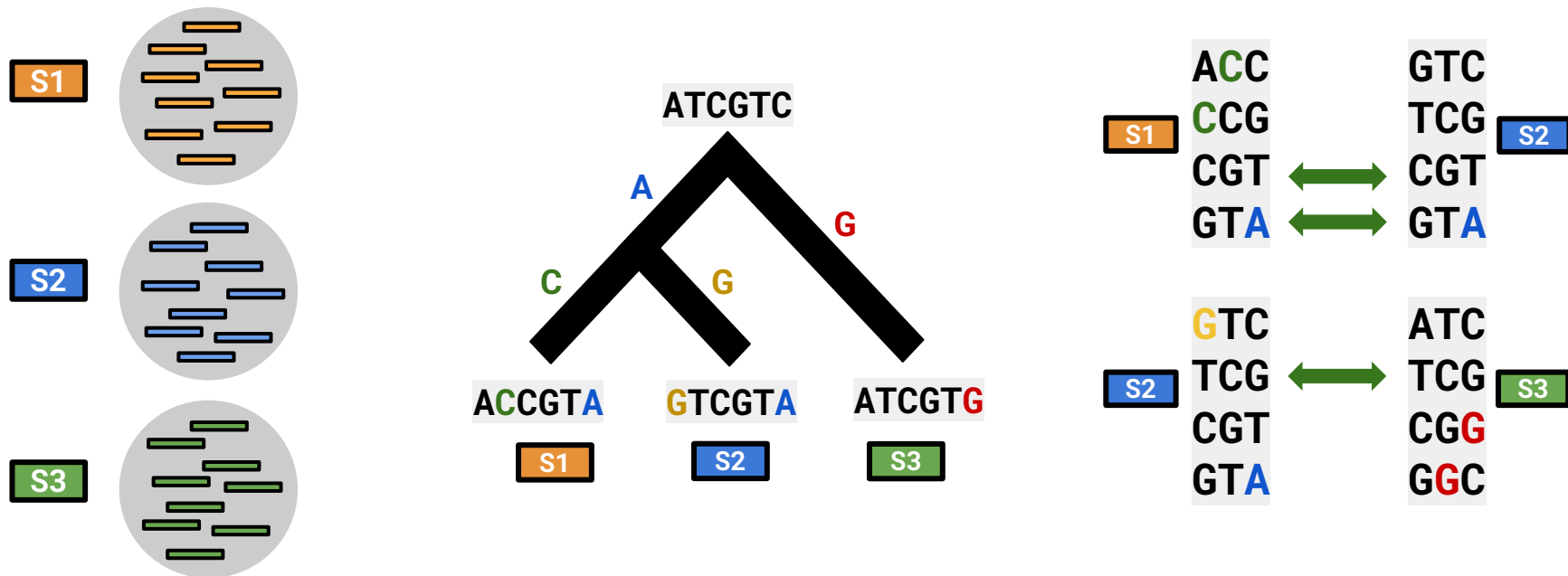


k -mer set for each sample

Alignment-free distance estimation



Alignment-free distance estimation

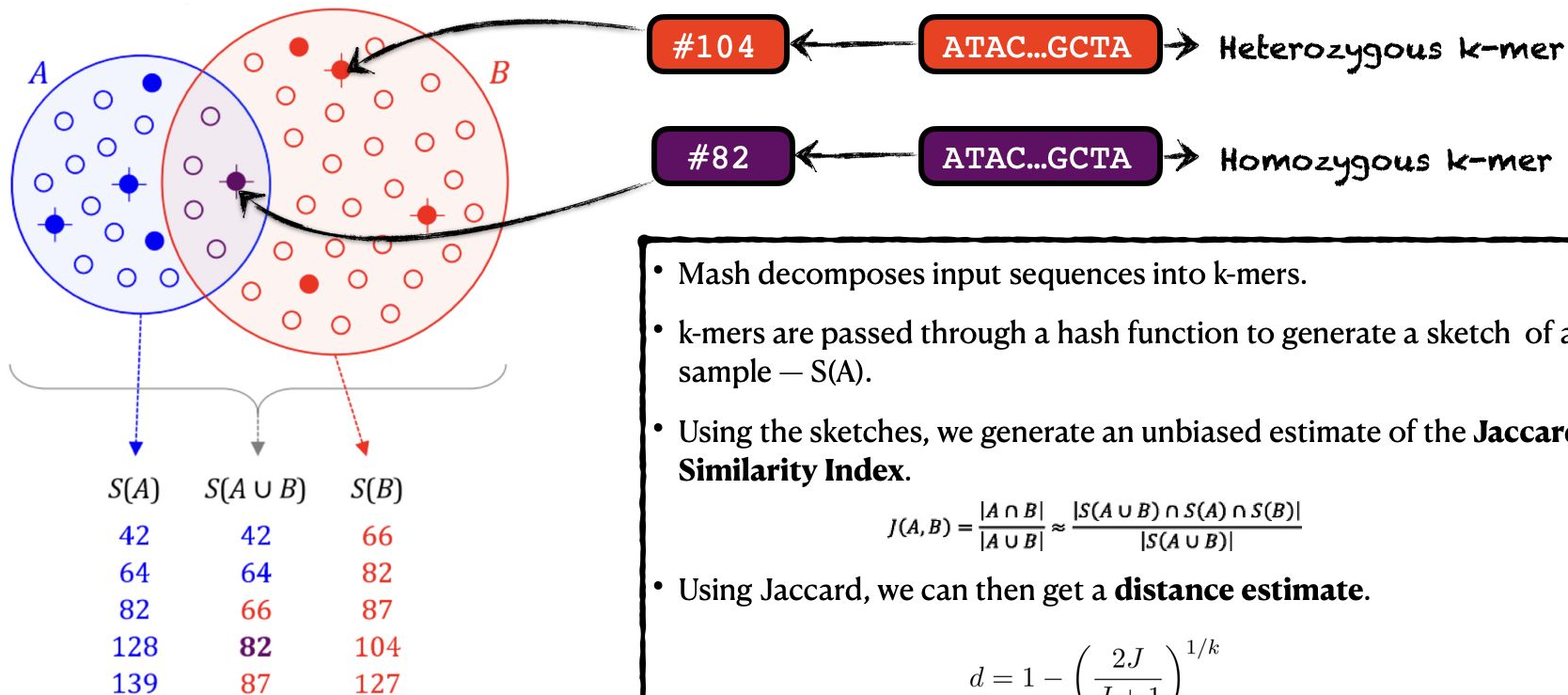


k-mer set for each sample

samples will be more or less
distant to one another

closely related samples will
have more k-mer matches

Mash for Alignment-Free Distances



- Mash decomposes input sequences into k-mers.
- k-mers are passed through a hash function to generate a sketch of a sample — $S(A)$.
- Using the sketches, we generate an unbiased estimate of the **Jaccard Similarity Index**.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

- Using Jaccard, we can then get a **distance estimate**.

$$d = 1 - \left(\frac{2J}{J + 1} \right)^{1/k}$$

Assembly-Free, Alignment-Free Distances

Decompose DNA
into k-mer sets

Jaccard Similarity
of sketches

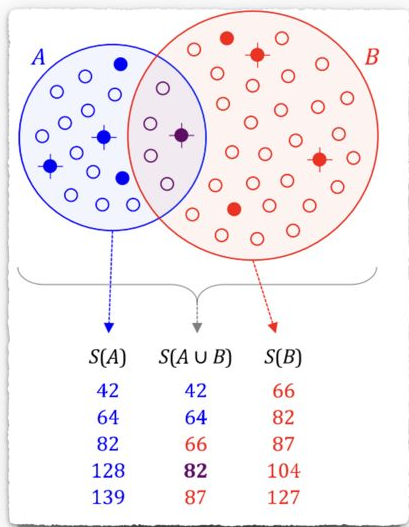
Results depend
on the data



● AGCT...GGCT
● GCTA...GCTA
● CTAT...CTAG

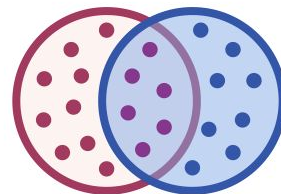


● AGCT...GGCT
● GCTA...GCTA
● CTAT...CTAG



(Ondov et al. 2016)

Complete
Genome



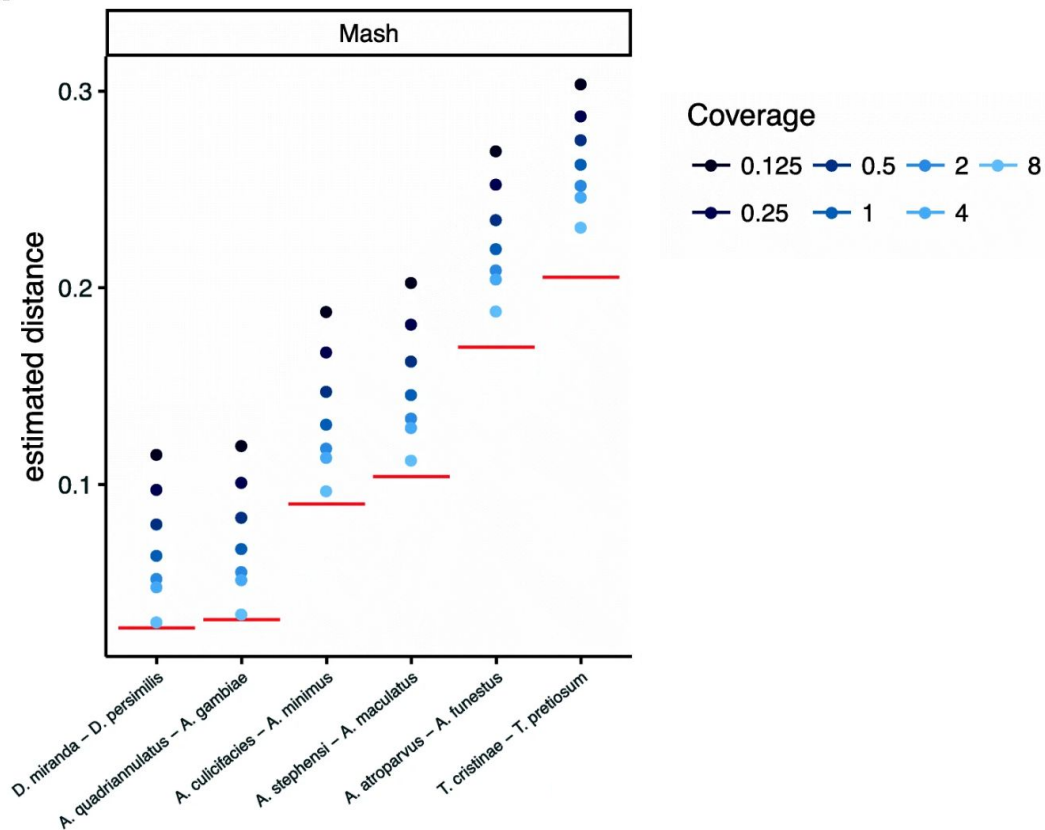
jaccard index (J)
k-mer size (k)



$$d = 1 - \left(\frac{2J}{J+1} \right)^{1/k}$$

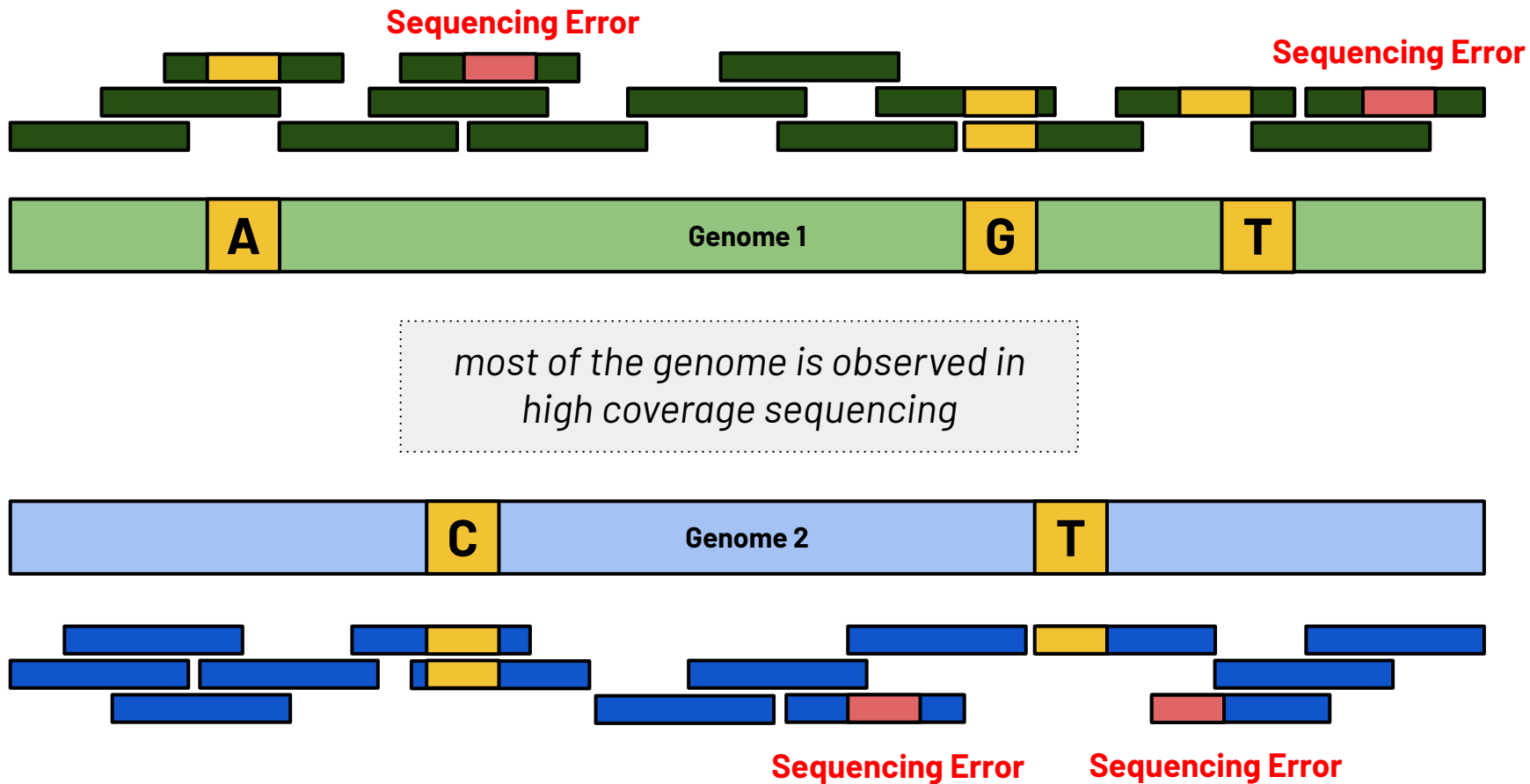
Mash fails with genome skimming data

a

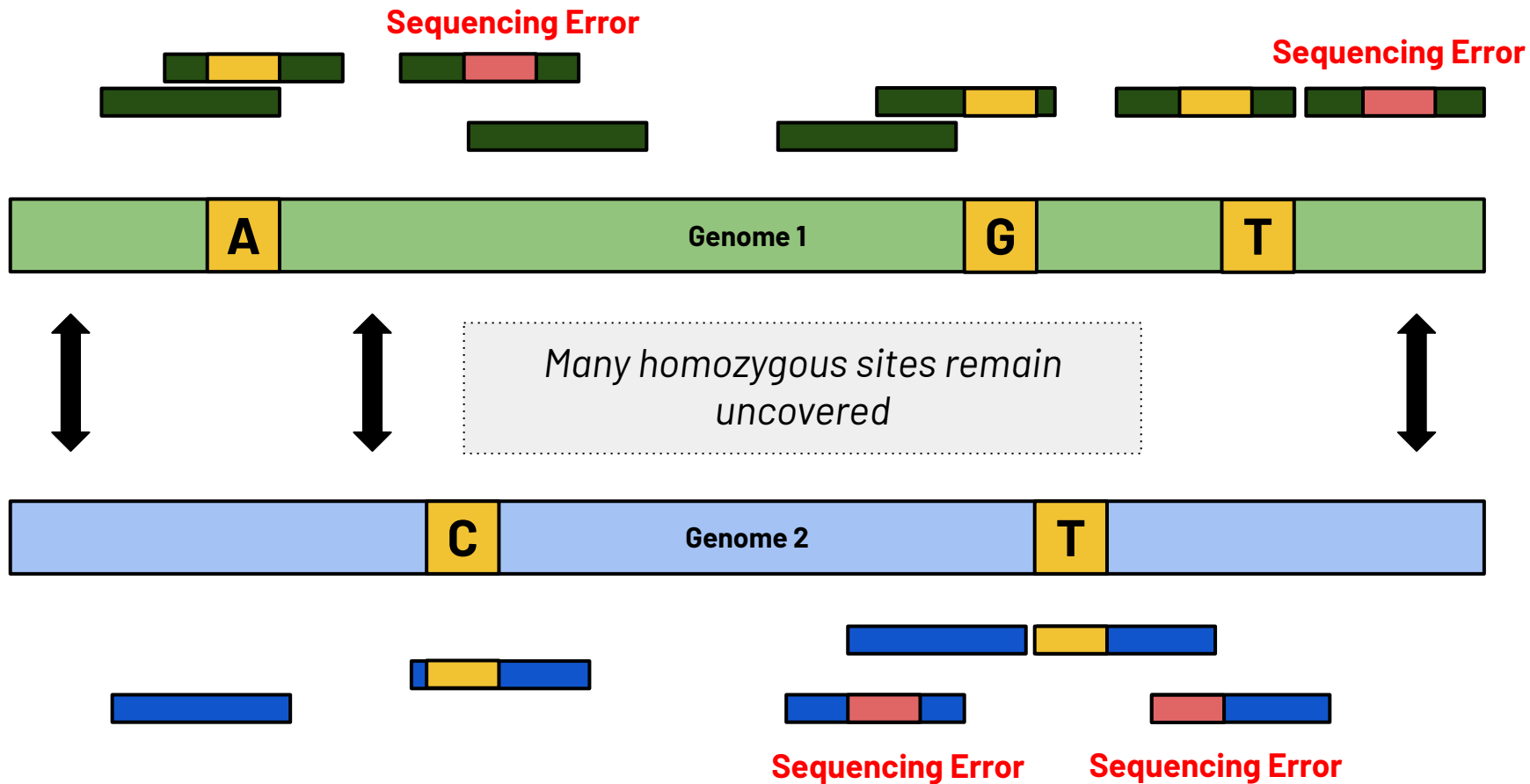


Low coverage
sequencing leads
to distance
overestimation
with mash

High coverage sequencing



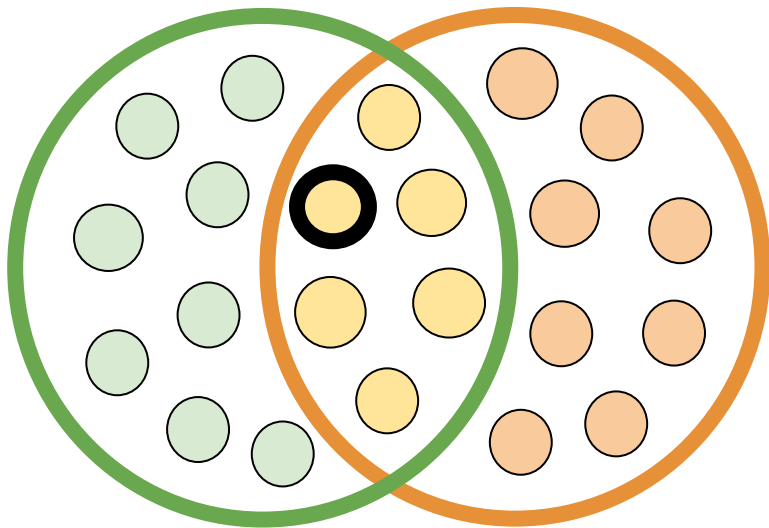
Low coverage sequencing



Jaccard Index example

High coverage
Jaccard Index

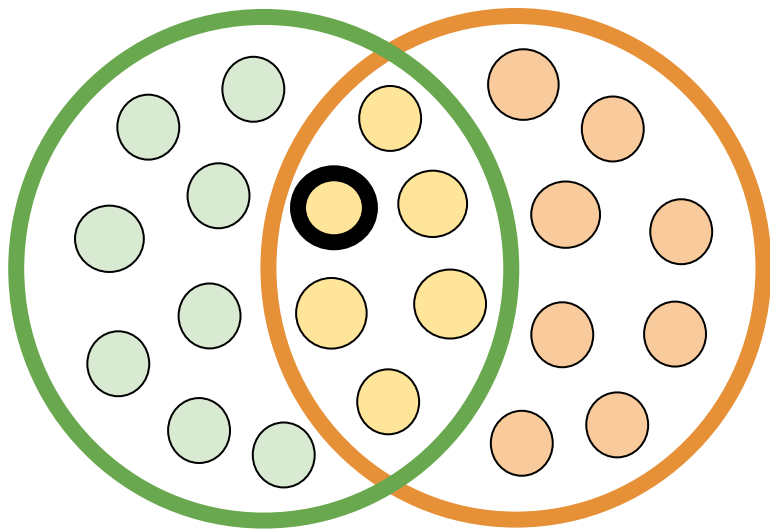
Low coverage
Jaccard Index



$$d = 1 - \left(\frac{2J}{J+1} \right)^{1/k}$$

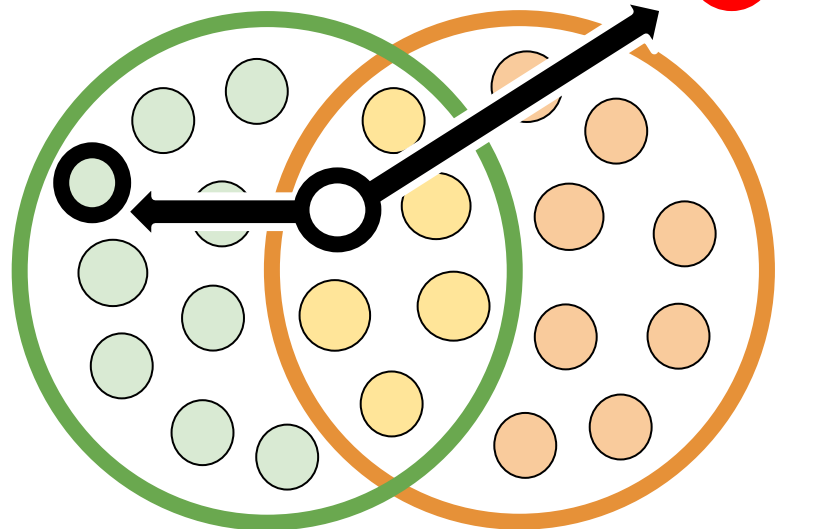
Jaccard Index example

High coverage
Jaccard Index



$$d = 1 - \left(\frac{2J}{J+1} \right)^{1/k}$$

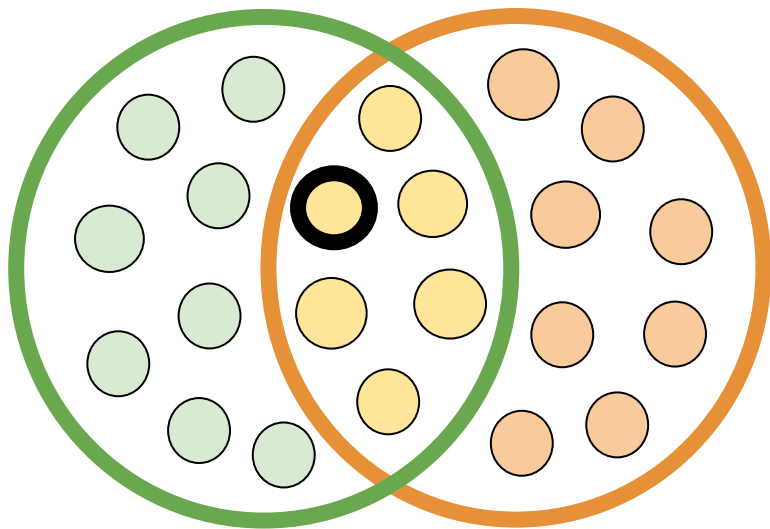
Low coverage
Jaccard Index



decreases similarity
(increases distance)

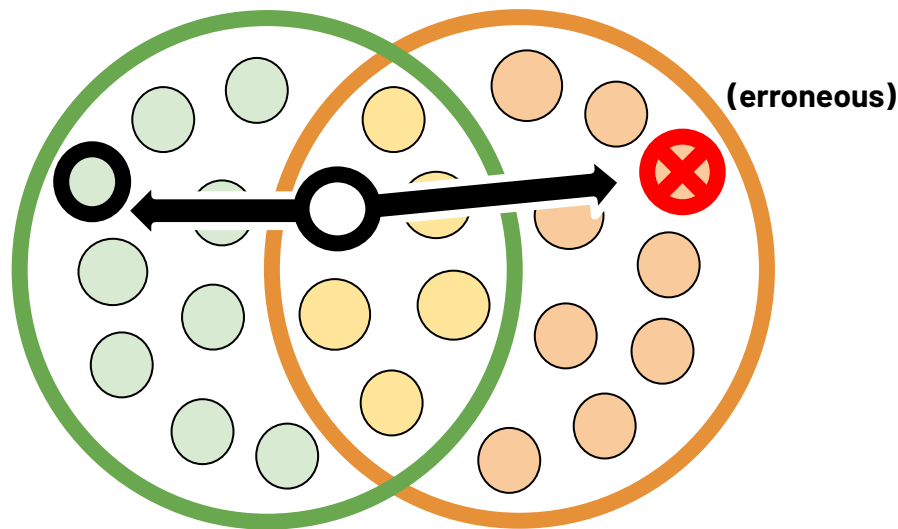
Jaccard Index example

High coverage
Jaccard Index



$$d = 1 - \left(\frac{2J}{J+1} \right)^{1/k}$$

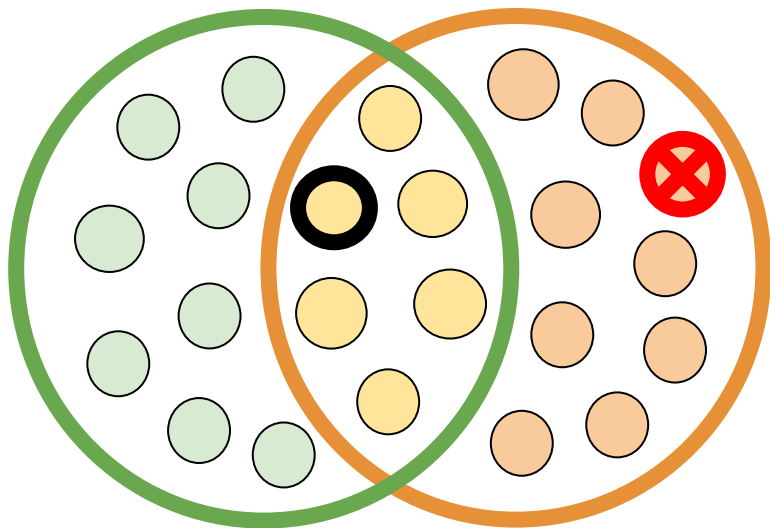
Low coverage
Jaccard Index



decreases similarity
(increases distance)

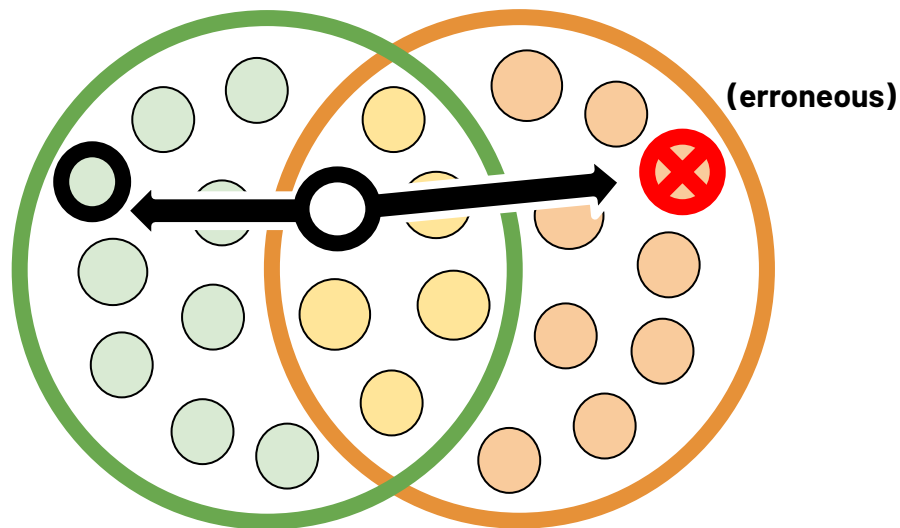
Jaccard Index example

High coverage
Jaccard Index



Doesn't high coverage also
have error?

Low coverage
Jaccard Index



decreases similarity
(increases distance)

Assembly-Free, Alignment-Free Distances

Decompose DNA
into k-mer sets

Jaccard Similarity
of sketches

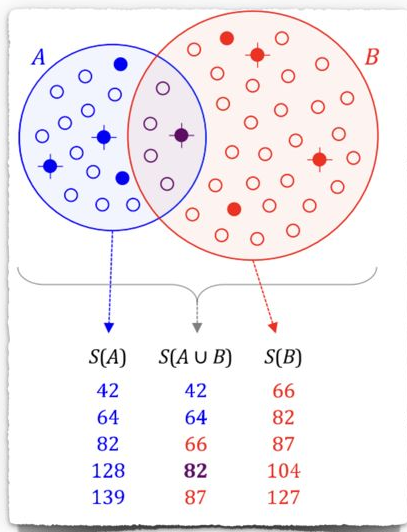
Results depend
on the data



● AGCT...GGCT
● GCTA...GCTA
● CTAT...CTAG

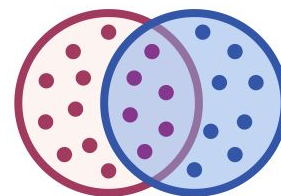


● AGCT...GGCT
● GCTA...GCTA
● CTAT...CTAG



(Ondov et al. 2016)

Complete
Genome

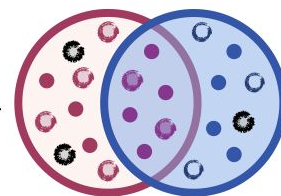


jaccard index (J)
k-mer size (k)



$$d = 1 - \left(\frac{2J}{J+1} \right)^{1/k}$$

Low
Coverage

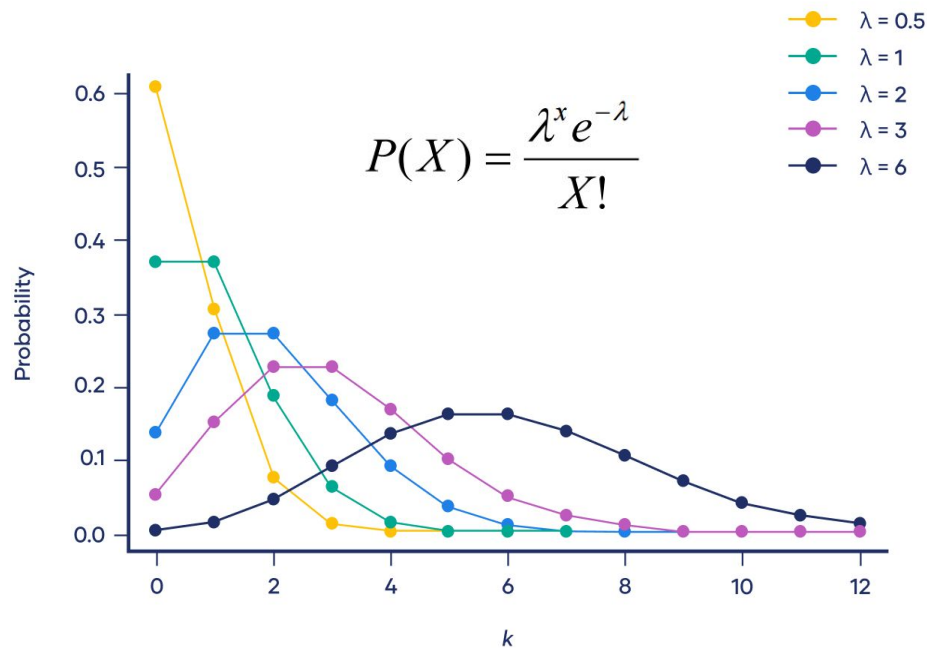


jaccard index (J)
k-mer size (k)
coverage ()
error rate ()



- erroneous k-mers
- missing k-mers
- observed k-mers

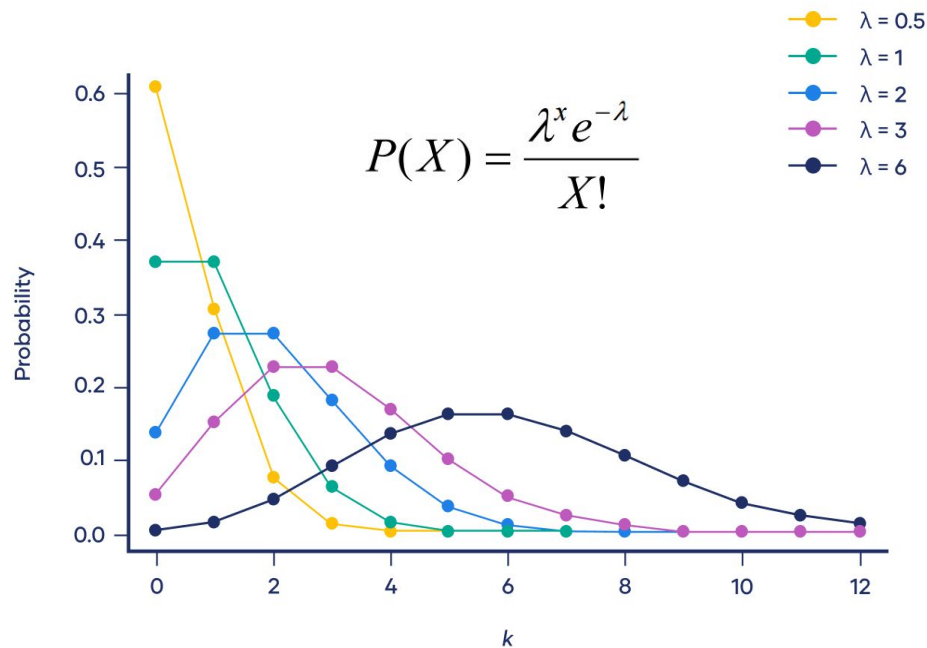
Modeling k-mer frequency



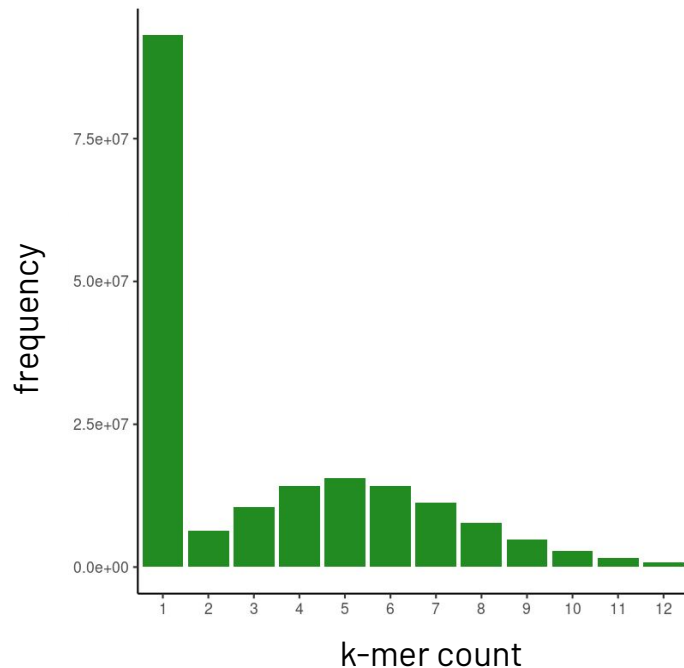
Poisson Distribution:

probability of a number of events occurring in a fixed interval, given an average rate of occurrence (λ)

Modeling k-mer frequency

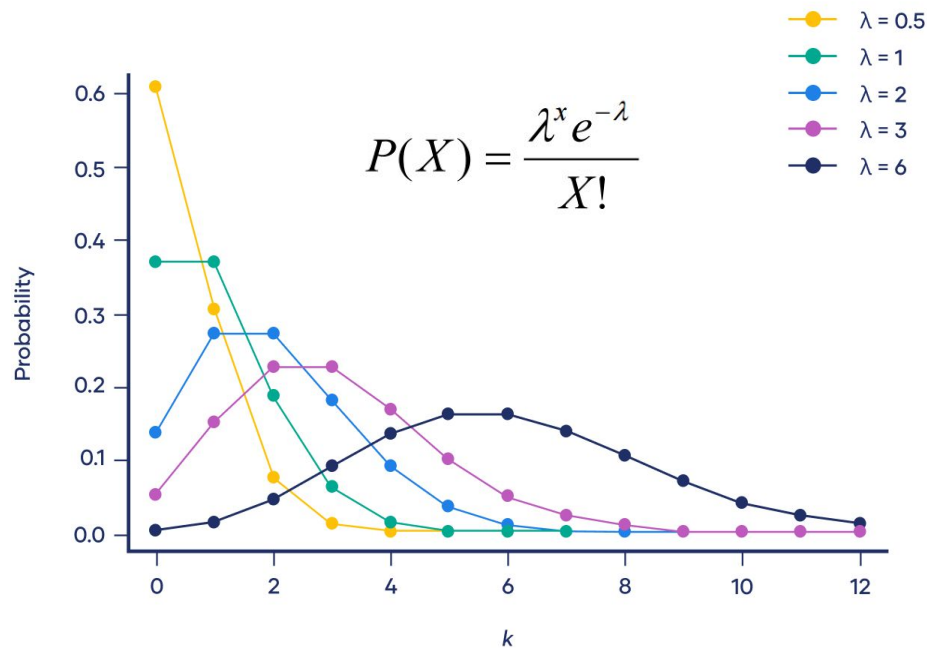


Poisson Distribution:
 probability of a number of events
 occurring in a fixed interval, given an
 average rate of occurrence (λ)

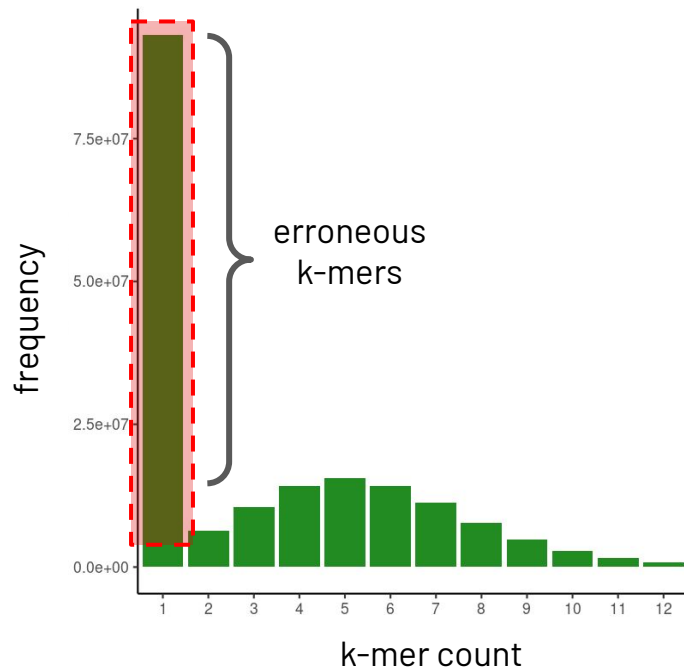


observed k-mer frequency
 histogram for a genome skim

Modeling k-mer frequency

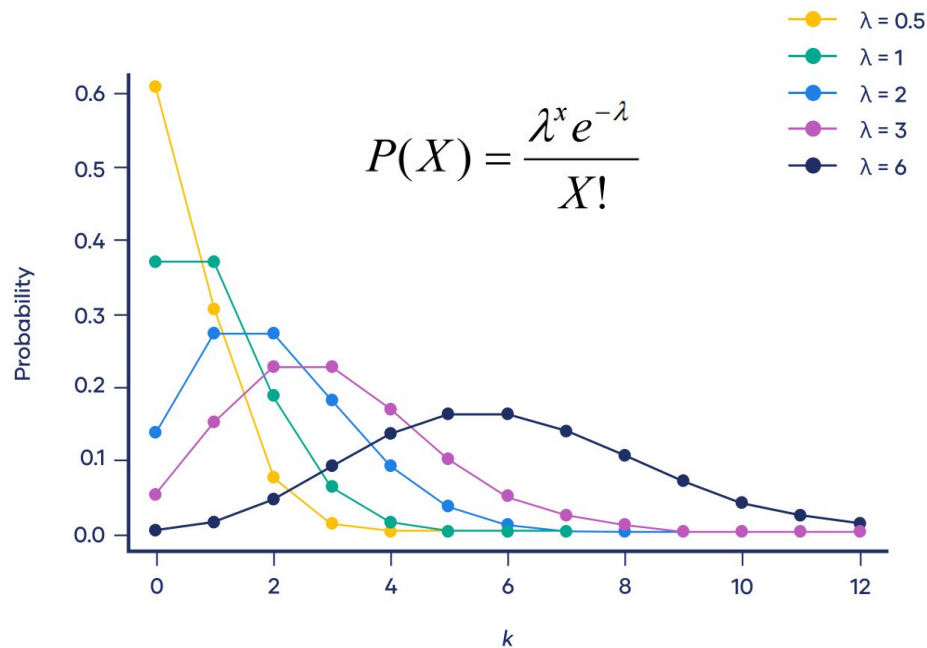


Poisson Distribution:
 probability of a number of events
 occurring in a fixed interval, given an
 average rate of occurrence (λ)

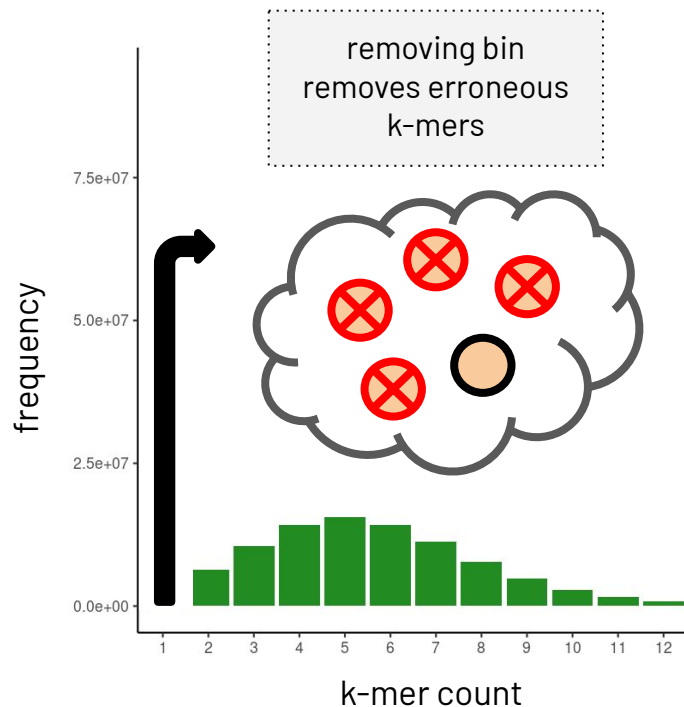


observed k-mer frequency
 histogram for a genome skim

Modeling k-mer frequency

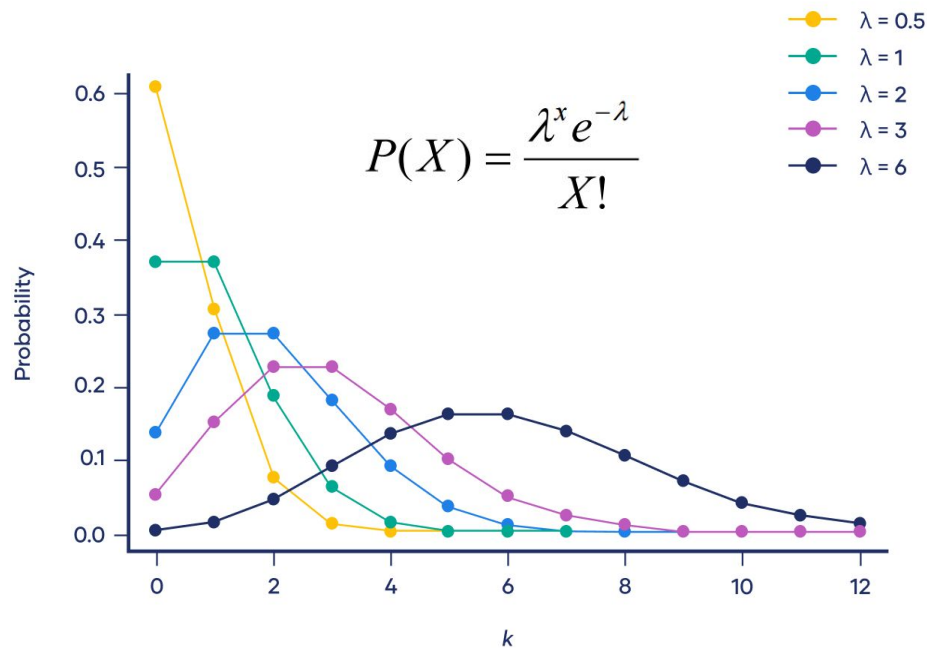


Poisson Distribution:
*probability of a number of events
 occurring in a fixed interval, given an
 average rate of occurrence (λ)*

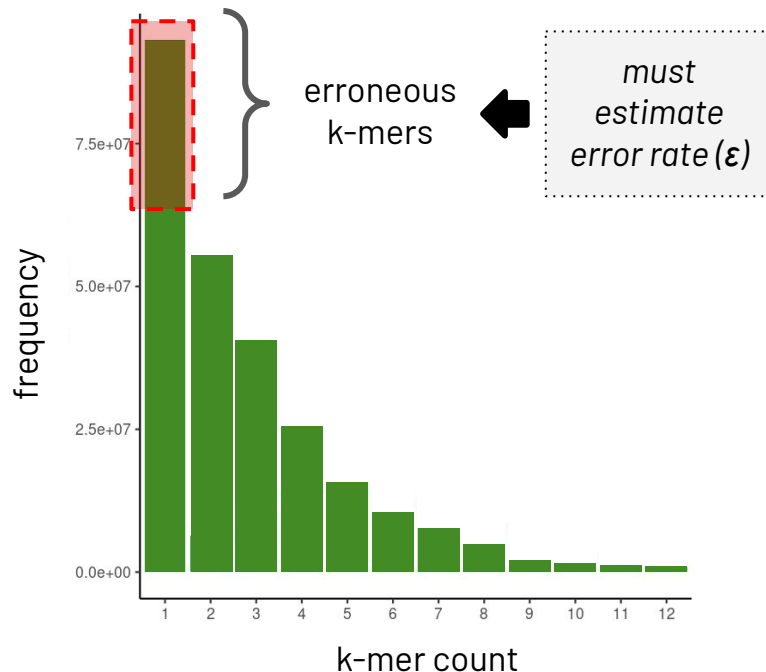


not possible for genome skims!

Modeling k-mer frequency

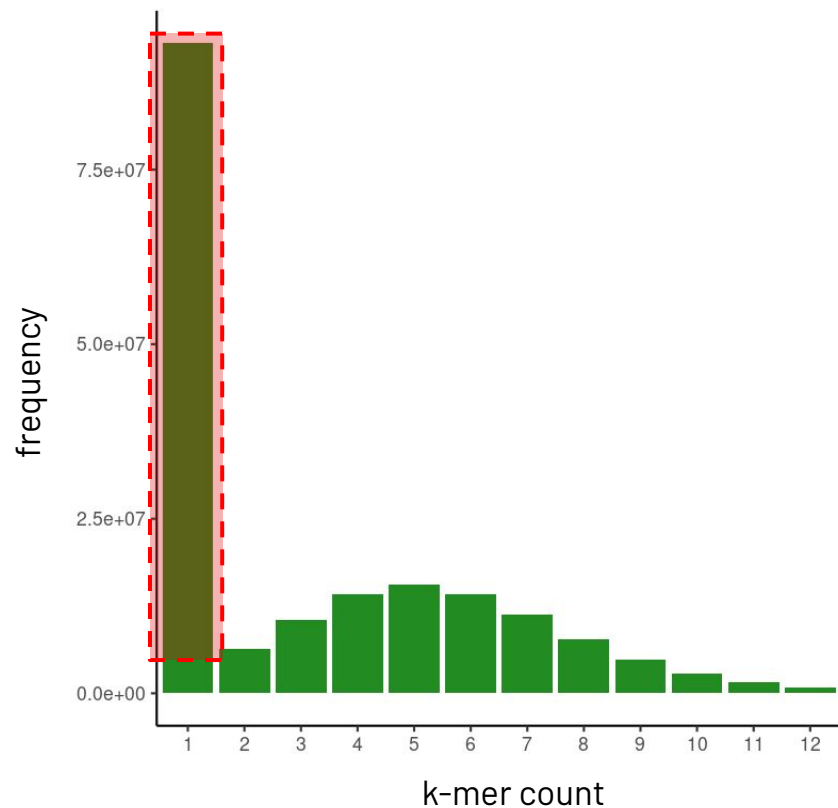


Poisson Distribution:
 probability of a number of events
 occurring in a fixed interval, given an
 average rate of occurrence (λ)



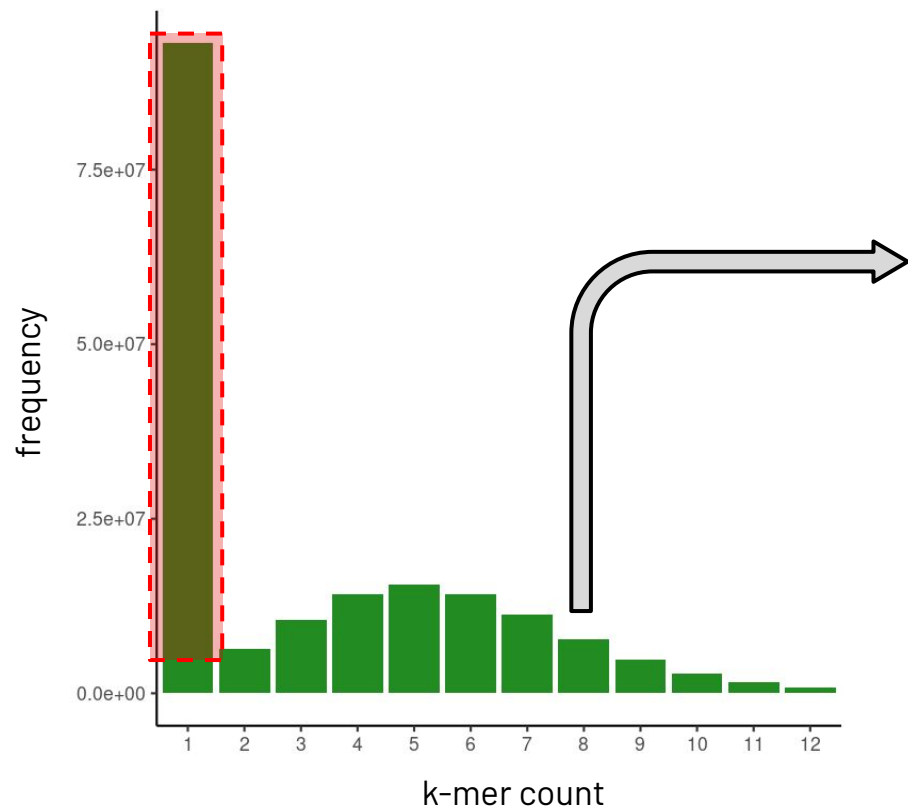
not possible for genome skims!

Skmer equations



errors in large k-mers will create **novel k-mers**

Skmer equations

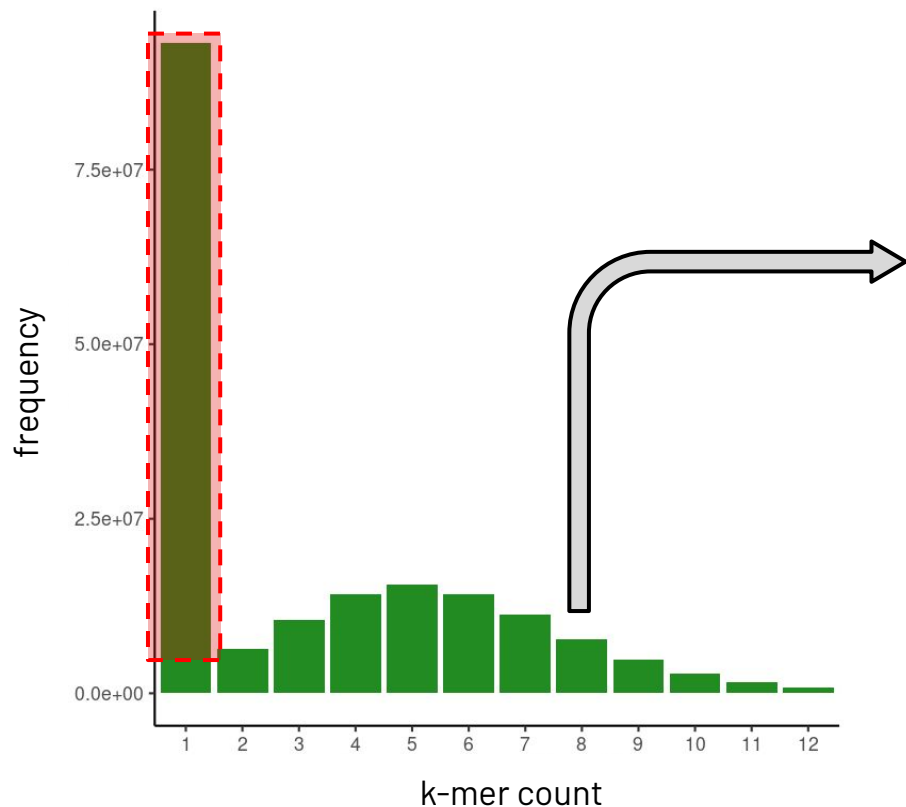


errors in large k-mers will create **novel k-mers**

all other bins will follow an **error-free k-mer coverage** poisson distribution (ξ)

$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!} \quad \Rightarrow \quad \text{expected size of bin} \quad \Rightarrow \quad \frac{\xi^i}{i!} e^{-\xi}$$

Skmer equations



errors in large k-mers will create **novel k-mers**

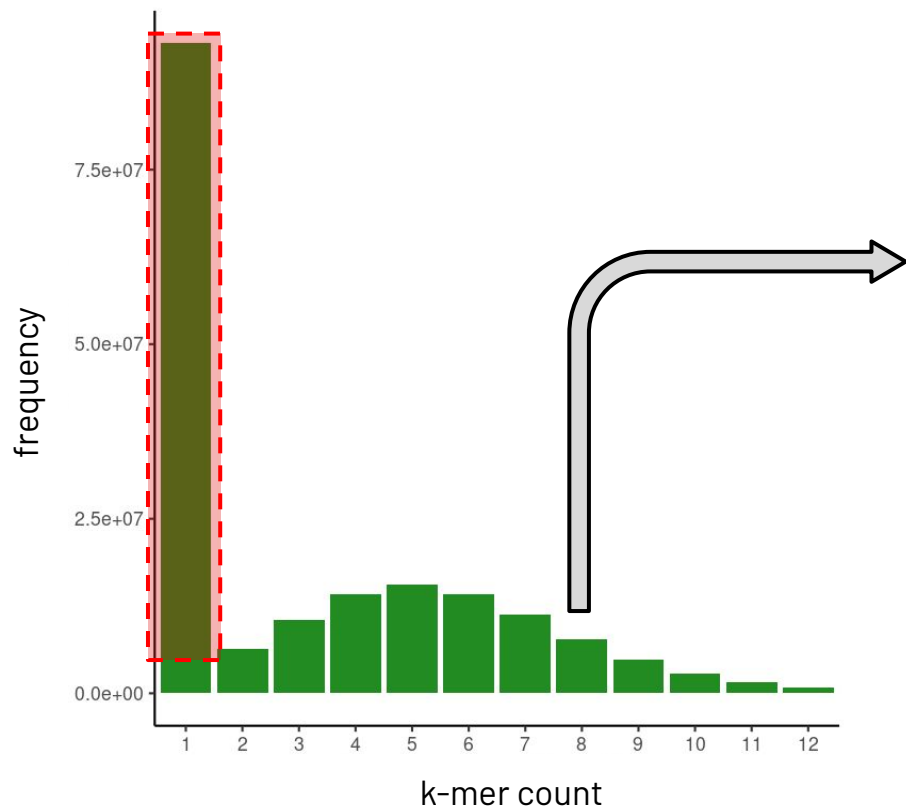
all other bins will follow an
error-free *k*-mer coverage poisson distribution (ξ)

$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!} \quad \Longrightarrow \quad \text{expected size of bin} \quad \Longrightarrow \quad \frac{\xi^i}{i!} e^{-\xi}$$

use the ratios of different bins to obtain estimate of
error-free *k*-mer coverage (ξ)

$$\xi = \frac{M_{h+1}}{M_h} (h + 1)$$

Skmer equations



errors in large k-mers will create **novel k-mers**

all other bins will follow an **error-free k-mer coverage** poisson distribution (ξ)

$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!} \implies \text{expected size of bin} \implies \frac{\xi^i}{i!} e^{-\xi}$$

use the ratios of different bins to obtain estimate of error-free k-mer coverage (ξ)

$$\xi = \frac{M_{h+1}}{M_h} (h + 1)$$

Do the same with the first bin to obtain **k-mer coverage** (λ) and then calculate **error** (ϵ)

$$\epsilon = 1 - (\xi/\lambda)^{1/k}$$

Distance equations

$$d = 1 - \left(\frac{2J}{J+1} \right)^{1/k}$$

Mash

$$d = 1 - \left(\frac{2(\zeta_1 L_1 + \zeta_2 L_2) J}{\eta_1 \eta_2 (L_1 + L_2)(1 + J)} \right)^{1/k}$$

Skmer

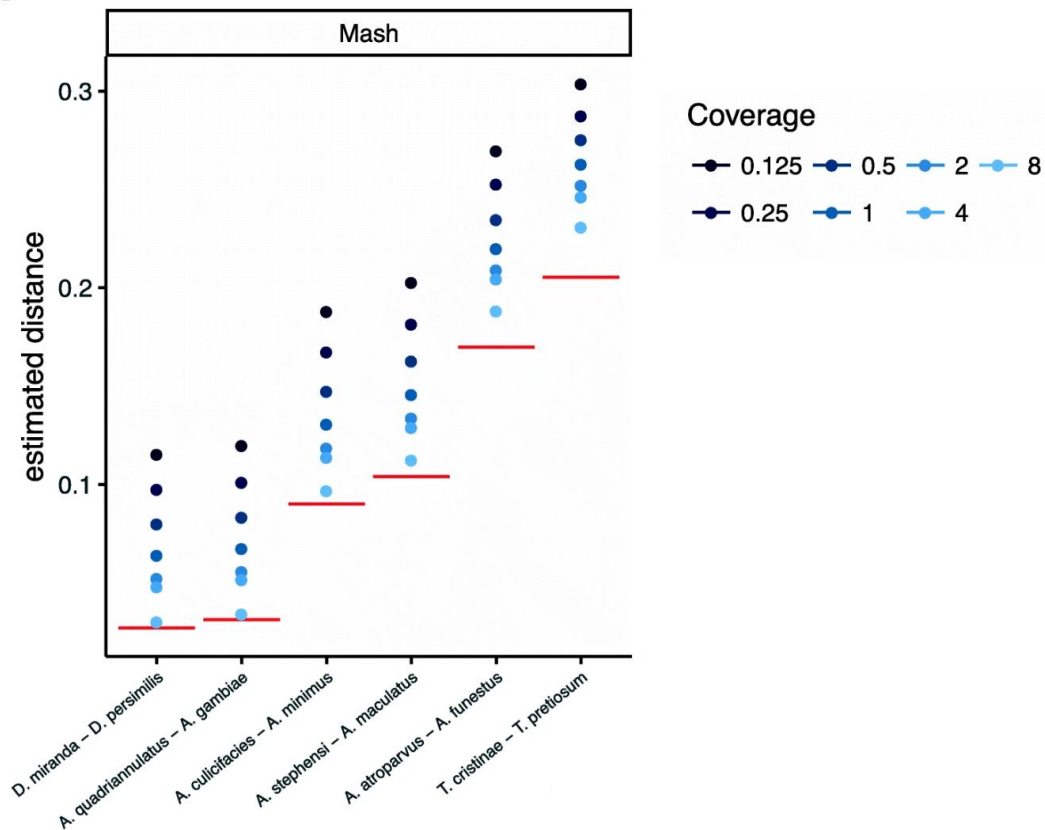
η_i (eta) = probability a k -mer is covered at least once without error.

ζ_i (zeta) = probability a k -mer is observed with or without error.

L_i = genome size.

Mash fails with genome skimming data

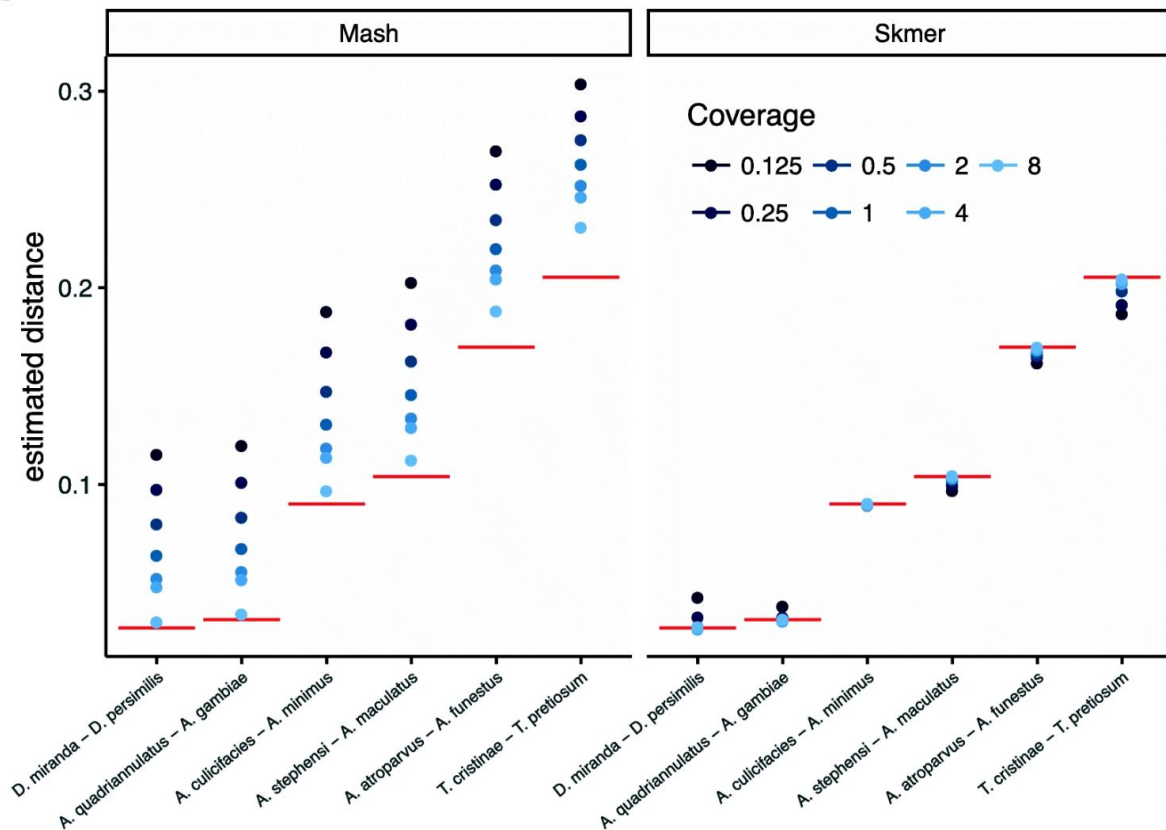
a



Low coverage
sequencing leads
to distance
overestimation
with mash

Mash fails with genome skimming data

a



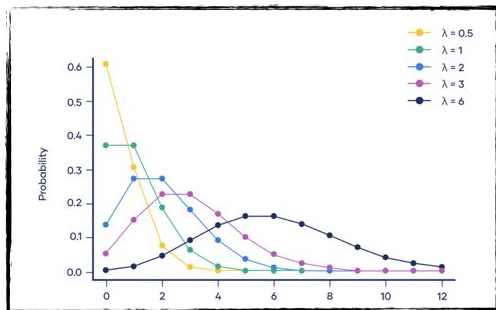
Skmer extends alignment-free, reference-free distance estimation to genome skims

Skmer Uses Simplifying Assumptions

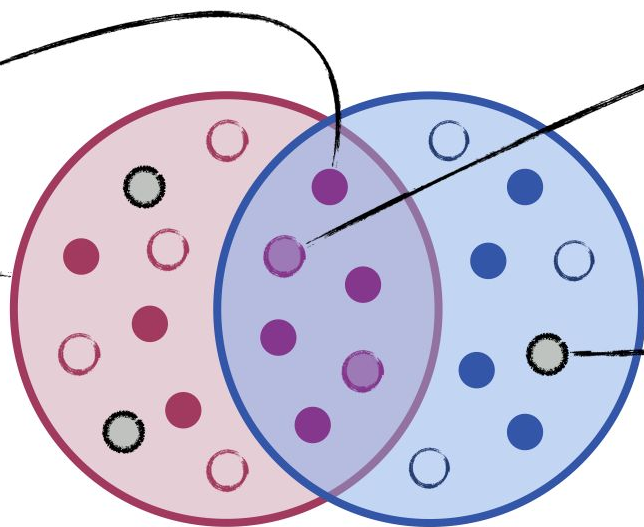
1

Genome is mostly unique and haploid

2



Error-free coverage can be modeled with a simple Poisson distribution



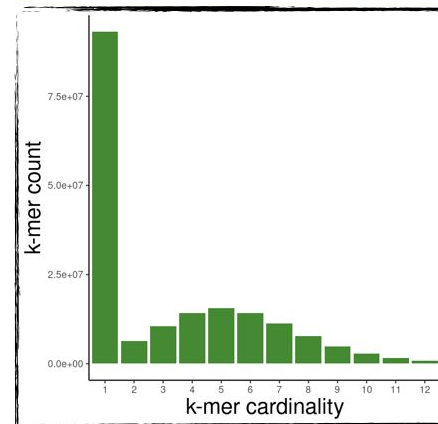
● → erroneous *k*-mers
● → missing *k*-mers
● → observed *k*-mers

3

Erroneous *k*-mers don't appear in the intersection

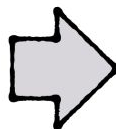
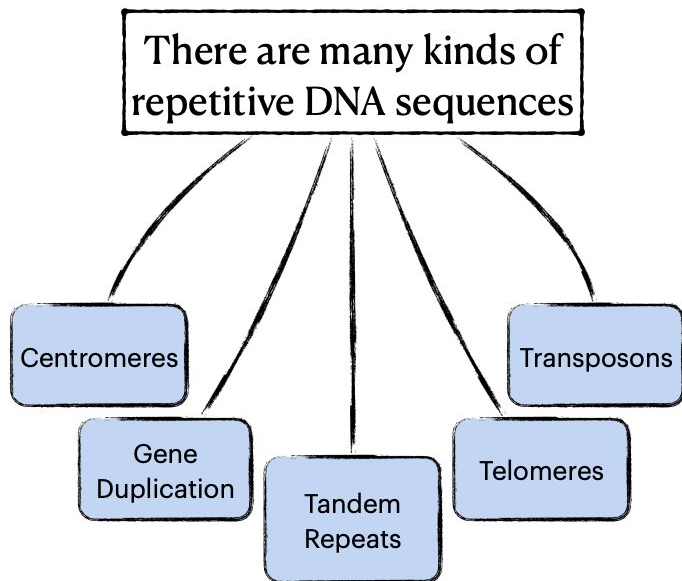
4

Erroneous *k*-mers appear only once at **low coverages**

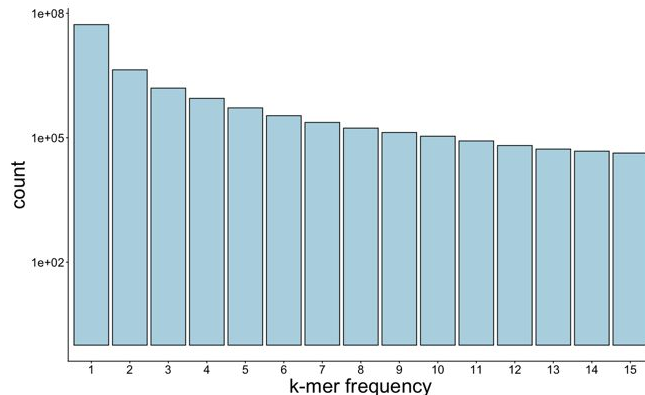


Genome Repeat Spectrum

K-mer based definition of repetitiveness



Summarized with k-mer Repeat Spectrum



Uniqueness Ratio =
 $R_1 / \text{Genome Length}$

Genomes are not unique



S. moellendorffii



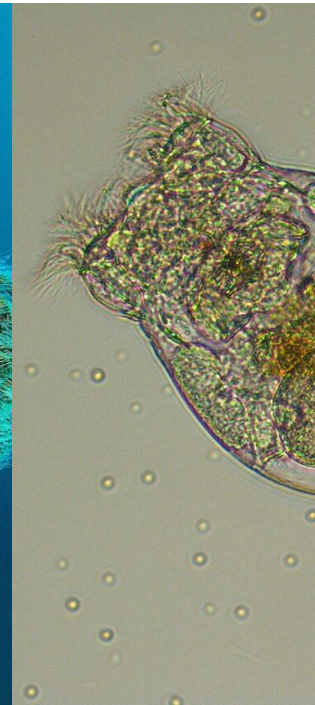
C. crispus



O. edulis



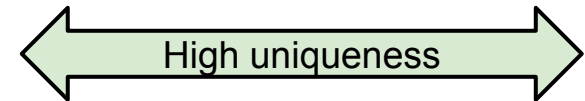
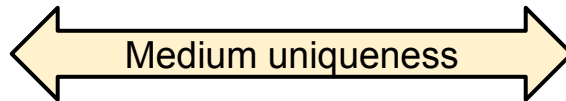
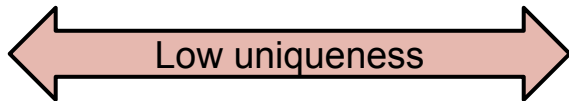
N. nomurai



B. plicatilis

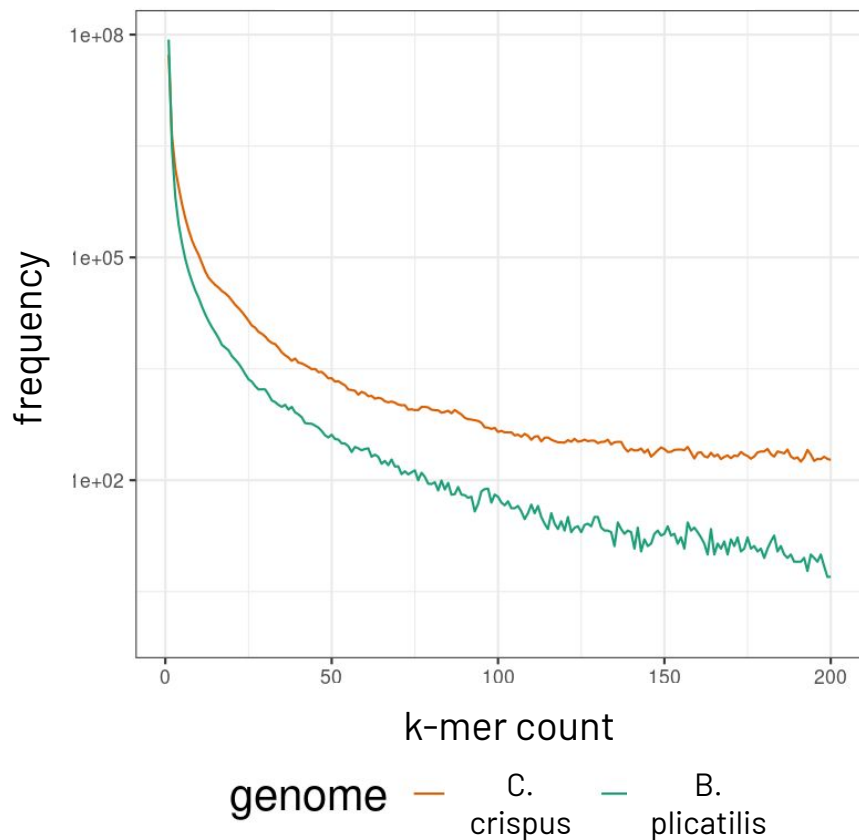


H. medicinalis



Genomes are not unique

genome repeat spectrum



Genomes come with a wide range of
Uniqueness Ratios:

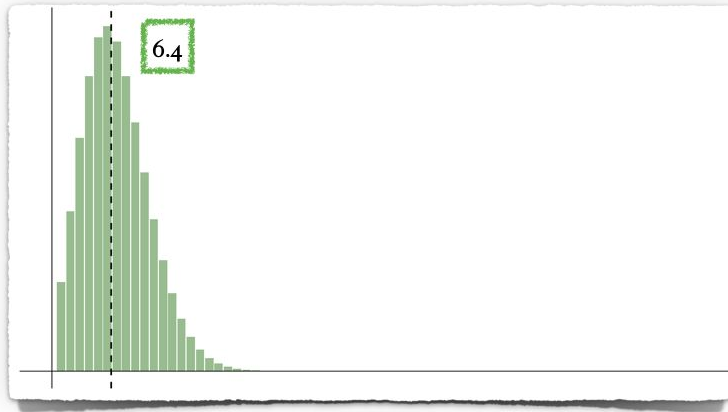
$$R_1 / L$$

<i>S. moellendorffi</i>	0.41
<i>C. crispus</i>	0.51
<i>O. edulis</i>	0.64
<i>N. nomurai</i>	0.71
<i>B. plicatilis</i>	0.85
<i>H. medicinalis</i>	0.90

Repeats Violate Skmer Assumptions

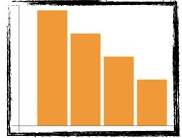
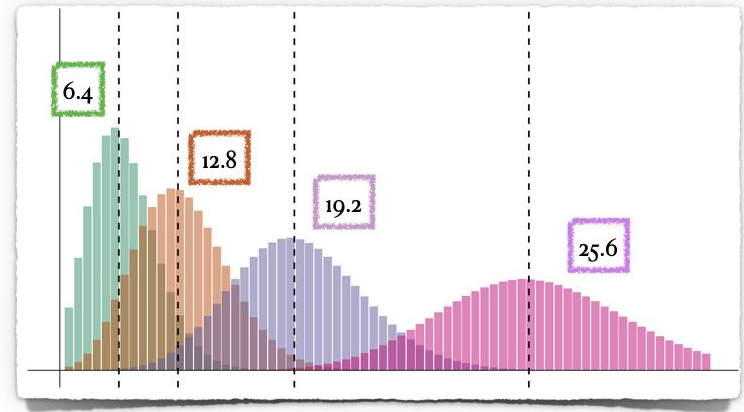
Skmer expectation of skim k-mer frequency under a unique genome

k-mer coverage = 6.4

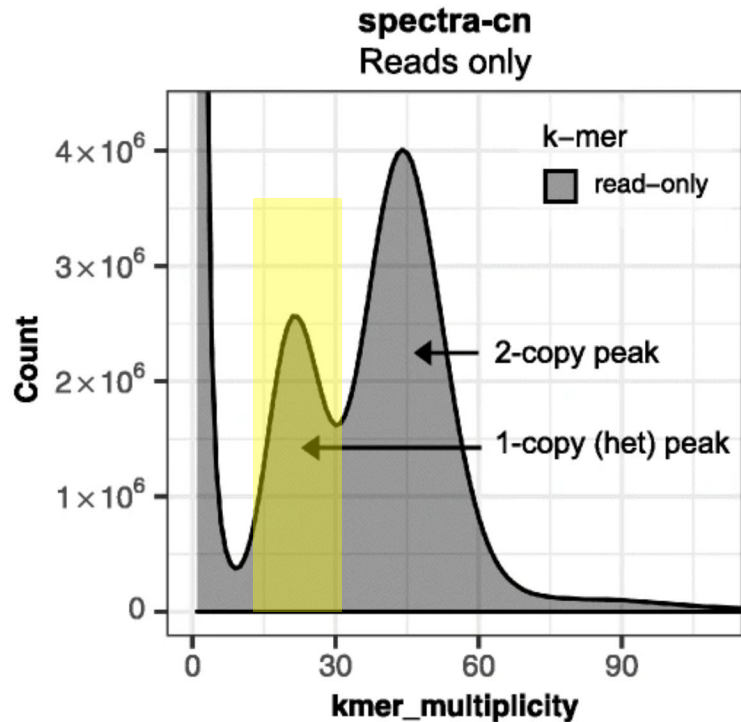


Skim k-mer frequency under genome with repeats

k-mer coverage = 6.4



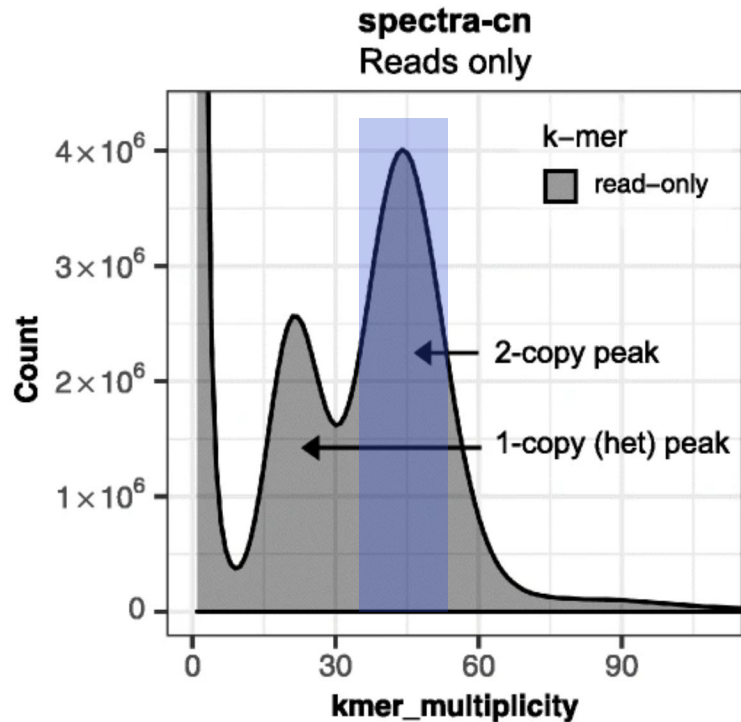
Repeats break Skmer assumptions



(Merqurey, 2020)

collection of k -mers only occur once

Repeats break Skmer assumptions

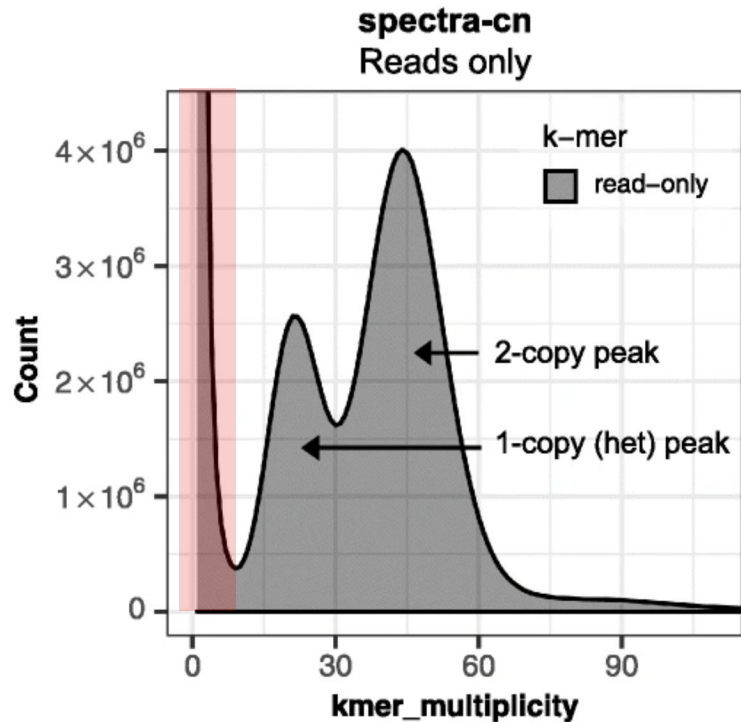


(Merqury, 2020)



collection of k -mers occur twice

Repeats break Skmer assumptions

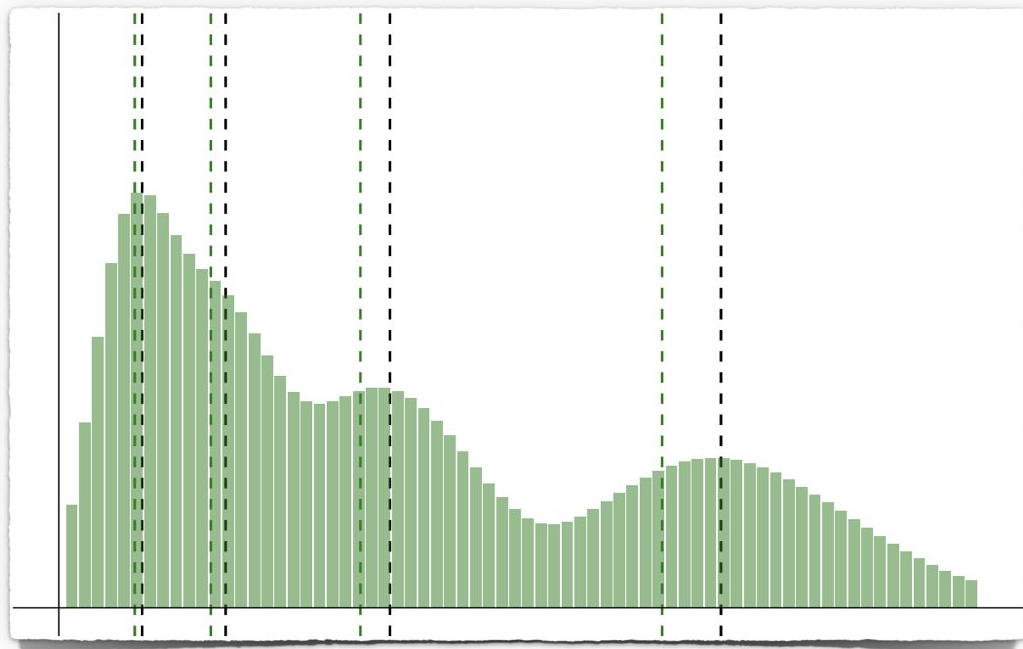


(Merqury, 2020)

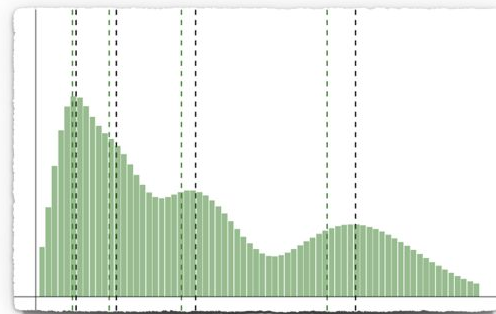


erroneous k-mers

K-mer Frequency Histogram with Error



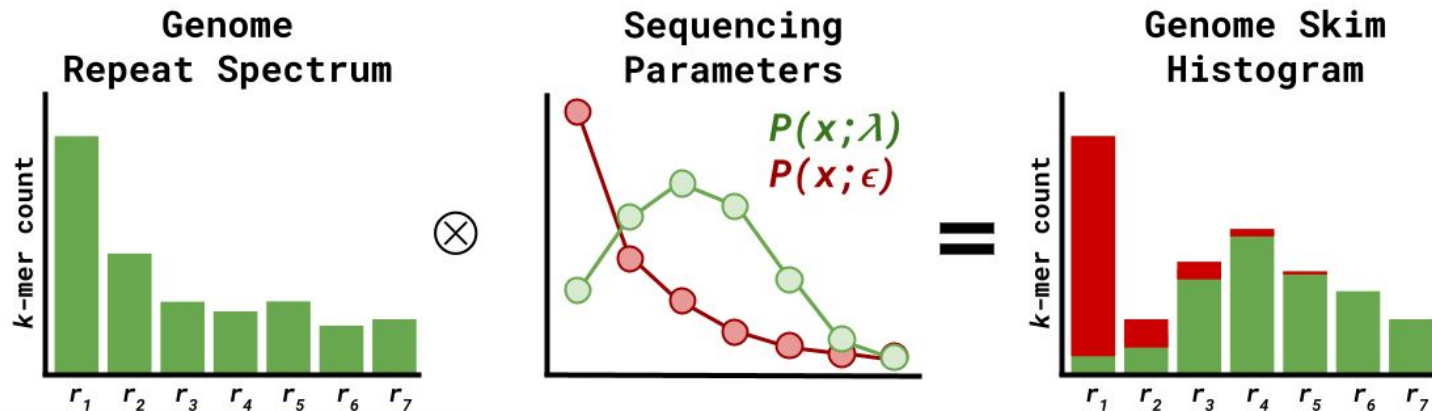
$\mathbf{xi}$



$\mathbf{eps}$

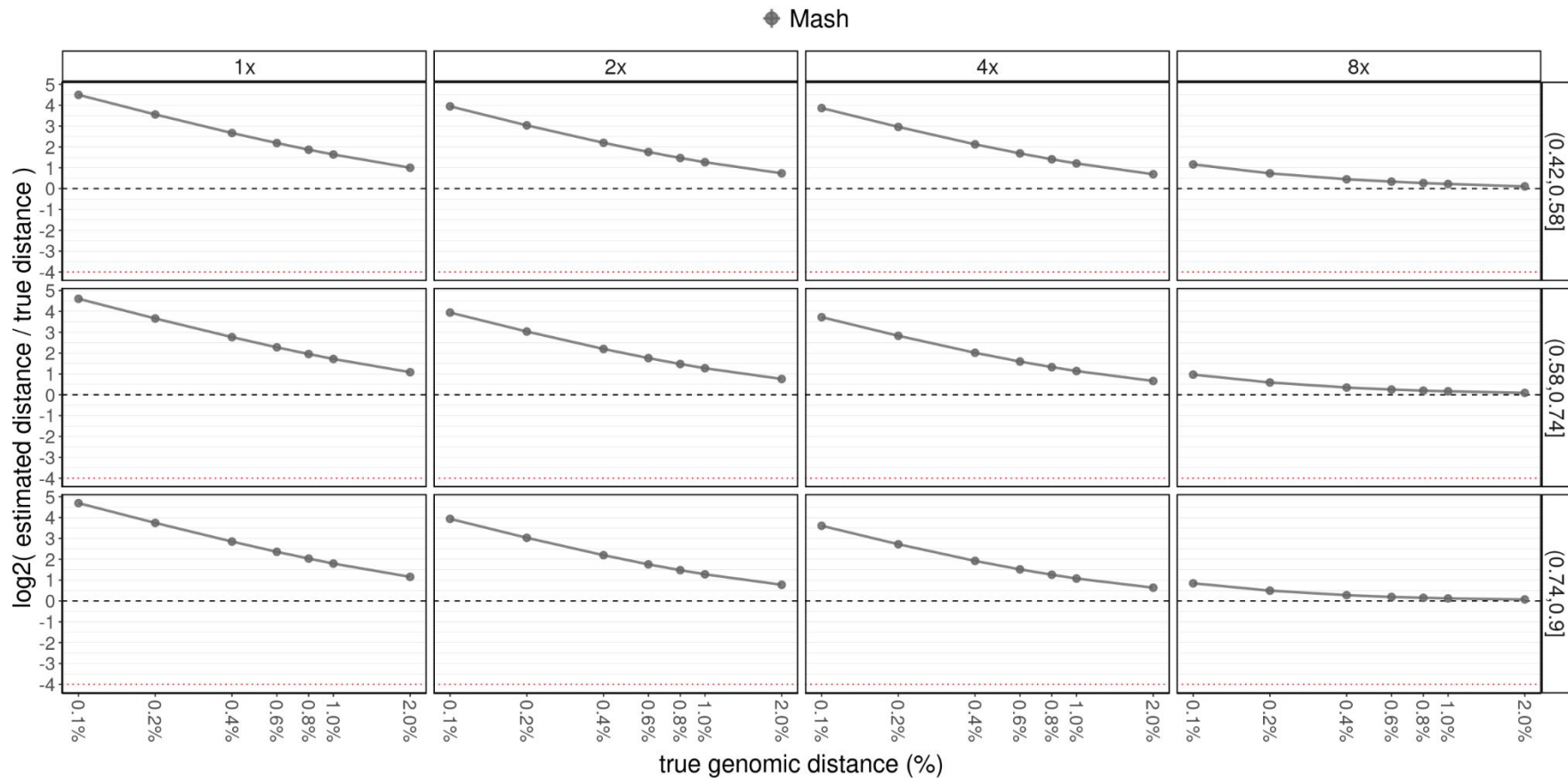


ReSkmer: estimating sequencing parameters

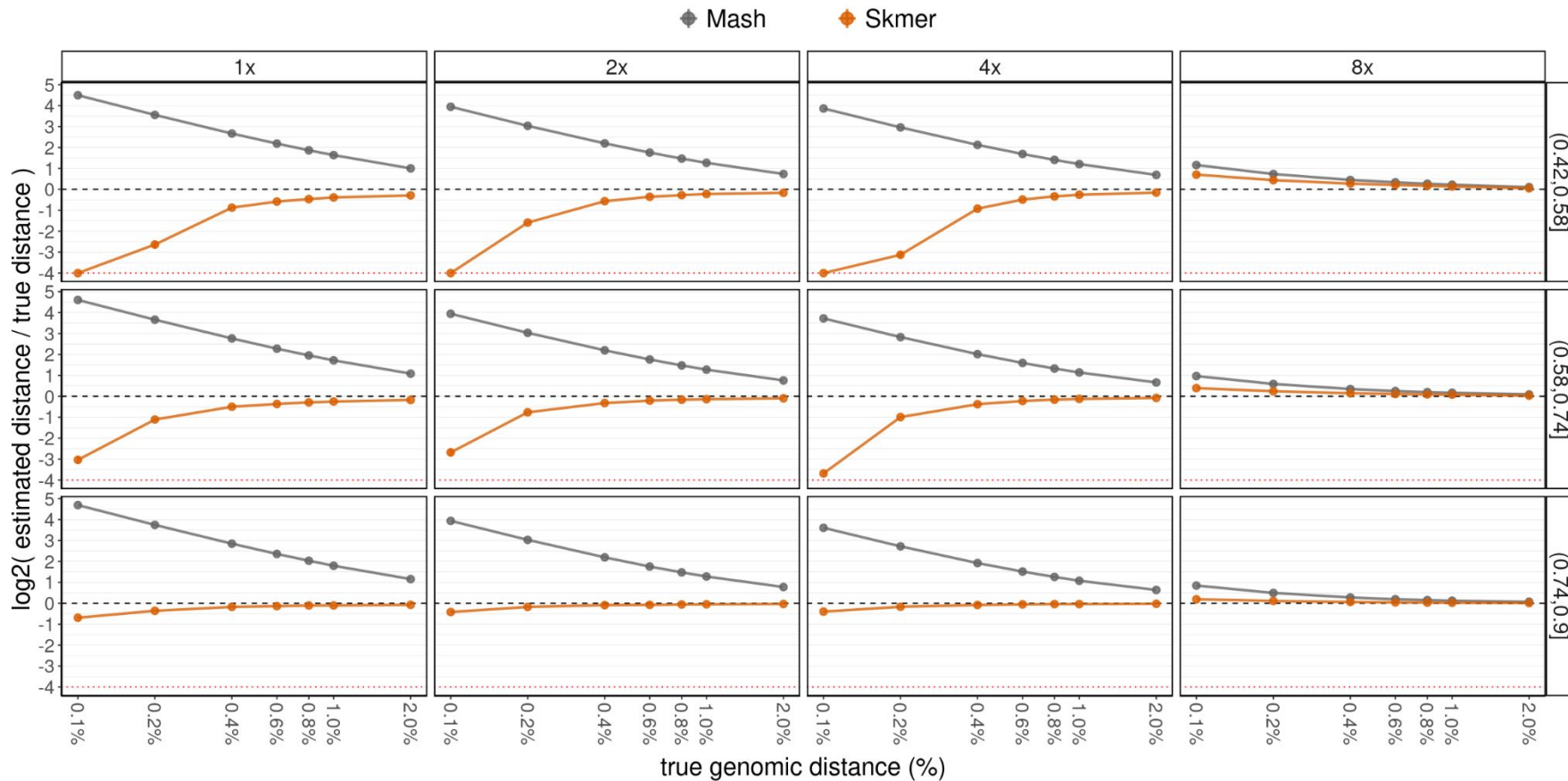


Each genome skim is a **convolution** between **sequencing parameter** distributions and its **repeat spectrum**

Mash Results



Skmer results



Methods

ReSkmer

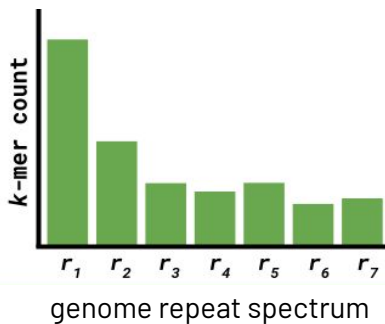
(*k*-mer based method,
short-read sequencing)

Advantages

- Whole Genome Resolution
- Reference and Alignment Free

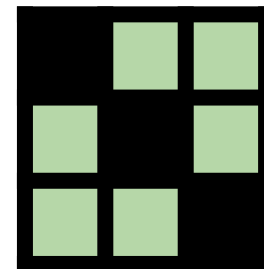
Disadvantages

- Low Coverage
- Models repeats



1	ATGCTACC...T	1	TATACTGC...C
2	CTAGGAA...A	2	ATGCTACC...T
3	CCCTTTA...C	3	CCCTTTA...C
4	GATAAAA...A	4	GATAAAA...A
5	TATACTGC...C	5	CTAGGAA...A

fastq per sample

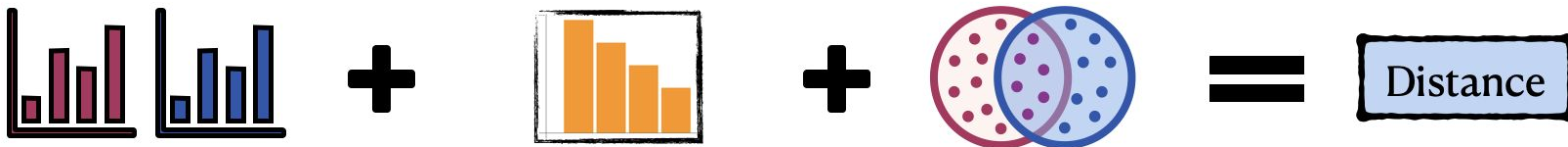


distance matrix

Input

Output

Correcting for Repetitiveness



Skimming k-mer
Histograms

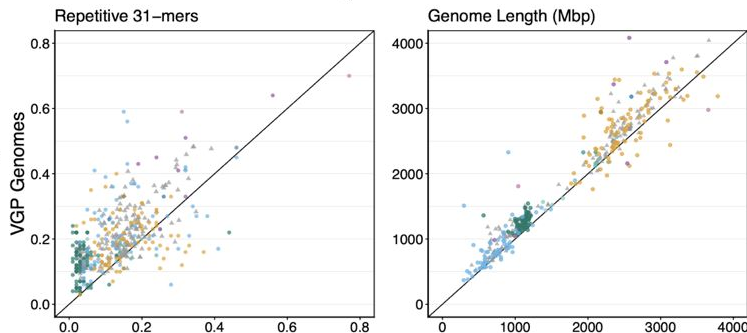
Genome Repeat
Spectrum

Jaccard Index

Distance

Lineage

- ▲ Ancestral
- Amphibians
- Birds
- Cartilaginous fishes
- Crocodiles
- Cyclostomes
- Lepidosauria
- Lobe-finned fishes
- Mammals
- Ray-finned fishes
- Turtles



Estimating repeat spectra and genome length from low-coverage genome skims with RESPECT

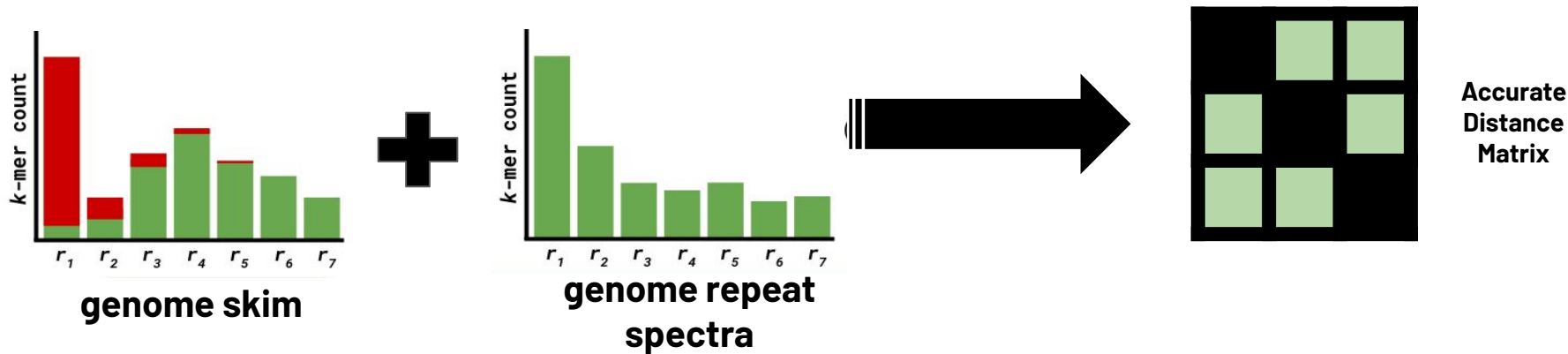
Shahab Sarmashghi, Melin Balaban, Eleonora Rachman, Behrouz Touri, Siavash Mirarab, Vineet Bafna

If no available
genome, spectrum
can be estimated.

Newer assemblies can capture repetitiveness

Using repeat spectrum to estimate distance

For 'short' distances (0.001 ~ 0.05)



get an idea of the k-mer counts of the original genome

1. **High Quality Assembly**

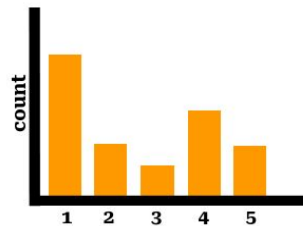
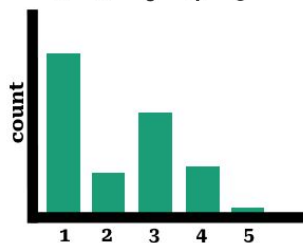
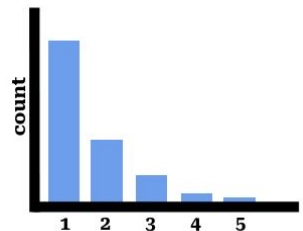
2. **Repeat Spectra Estimate** (via Respect)

Using repeat spectrum to estimate distance

get an idea of the k-mer counts of the original genome

1. **High Quality Assembly**

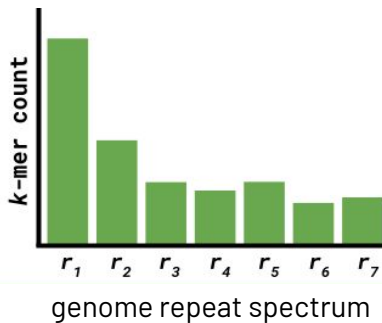
2. **Repeat Spectra Estimate** (via Respect)



RESPECT

Repeat
Spectra
and
Genome
Length
Estimation

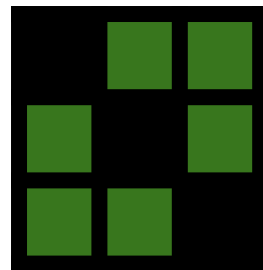
RESPECT



ReSkmer

Use
corresponding
spectra to
estimate
sequencing
parameters
and
distance matrix

ReSkmer

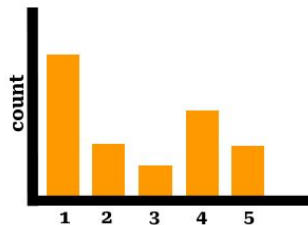
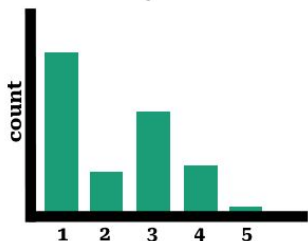
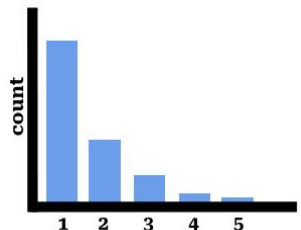


Using repeat spectrum to estimate distance

get an idea of the k-mer counts of the original genome

1. **High Quality Assembly**

2. **Repeat Spectra Estimate** (via Respect)



RESPECT

Repeat Spectra and Genome Length Estimation

RESPECT

Estimating repeat spectra and genome length from low-coverage genome skims with RESPECT

Shahab Sarmashghi, Metin Balaban, Eleonora Rachtman, Behrouz Touri, Siavash Mirarab, Vineet Bafna

sample	r1	r2	...
skim_1	5016990	902682	...
skim_2	4906813	873642	...
skim_3	5206770	892112	...

set of repeat spectra estimates

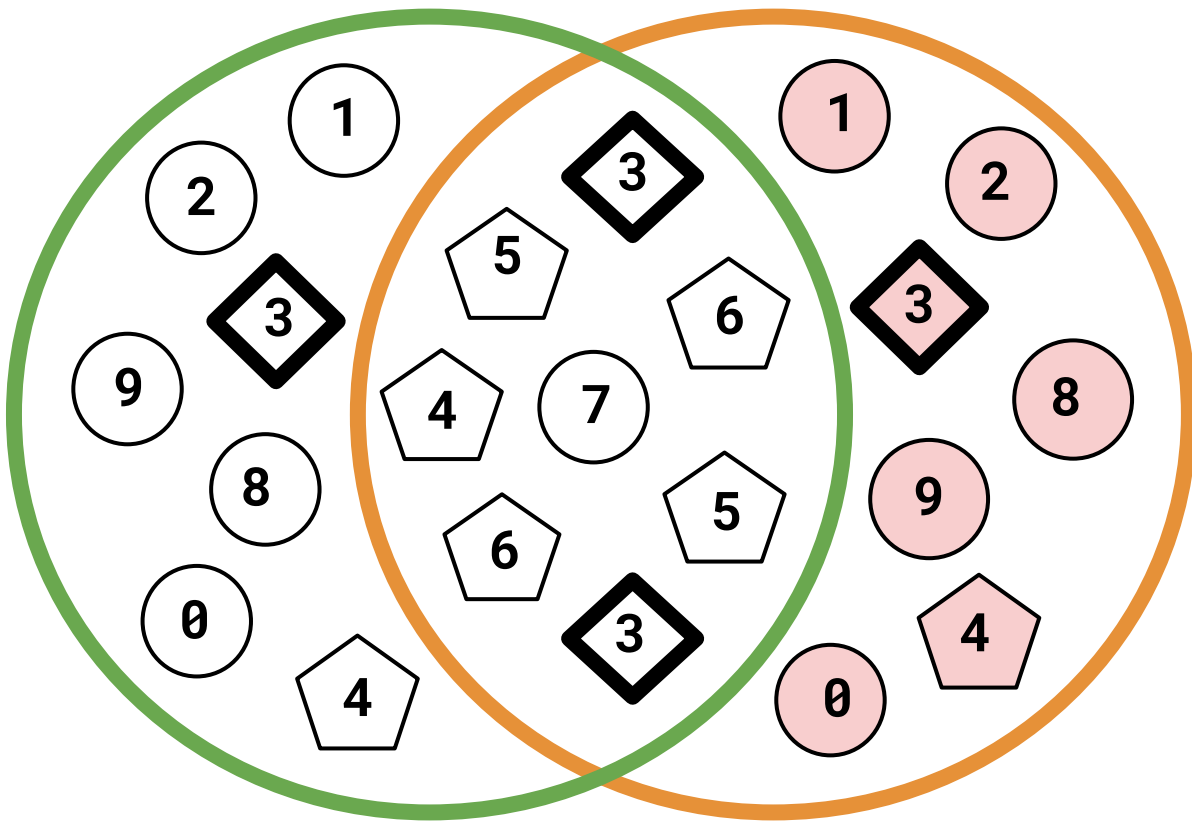
ReSkmer

Use corresponding spectra to estimate sequencing parameters and distance matrix

ReSkmer

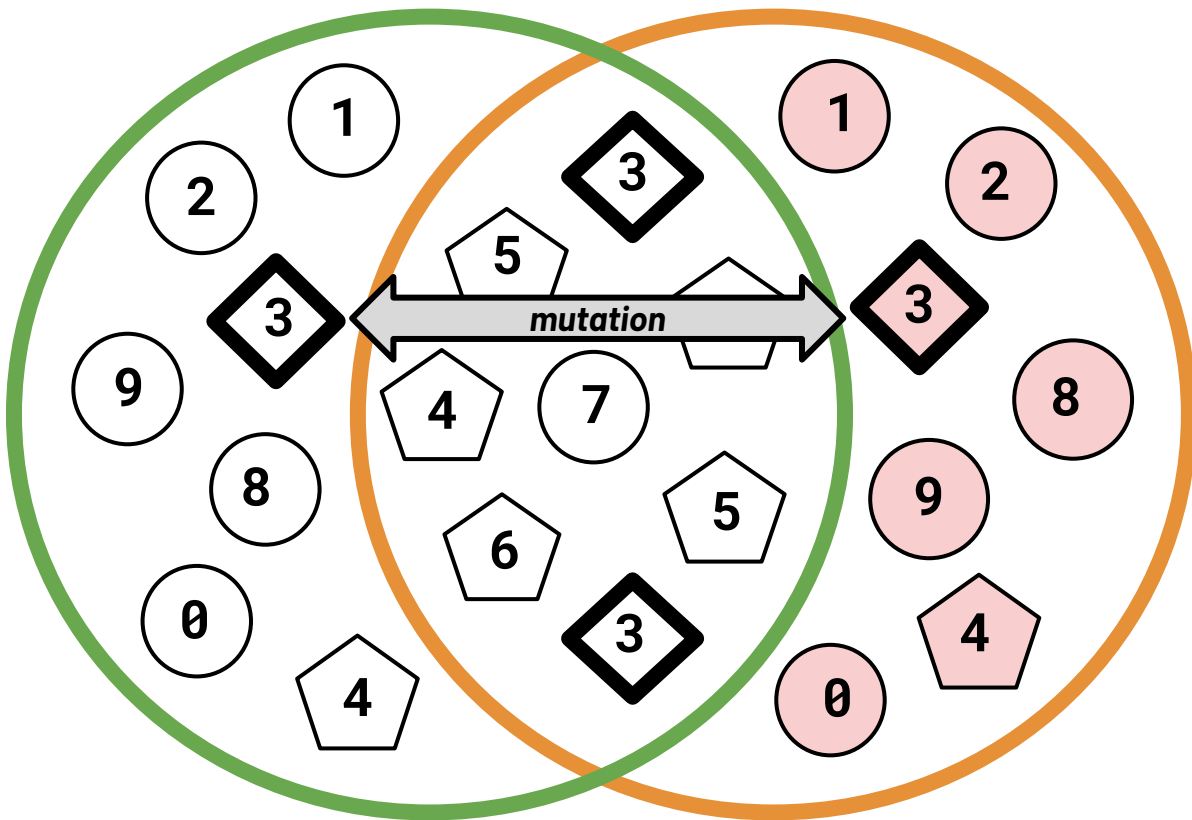


How repeats affect the intersection

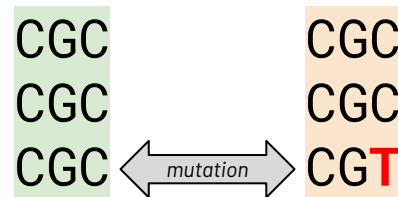


intersection between the
sets of k -mers

How repeats affect the intersection



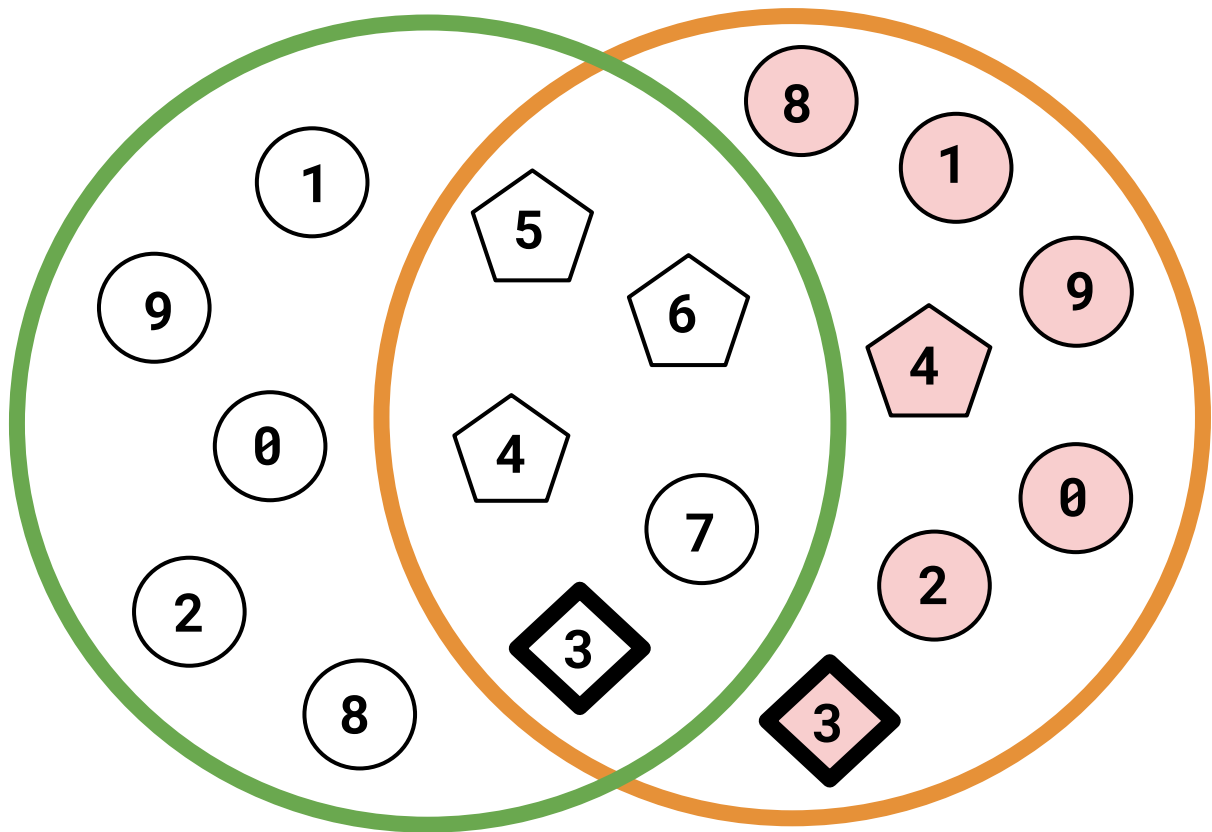
intersection between the
sets of k -mers



$$\mathbb{E}[I_{\text{multiset}}] = f(d)$$

$$I_{\text{multiset}} = 8$$

How repeats affect the intersection



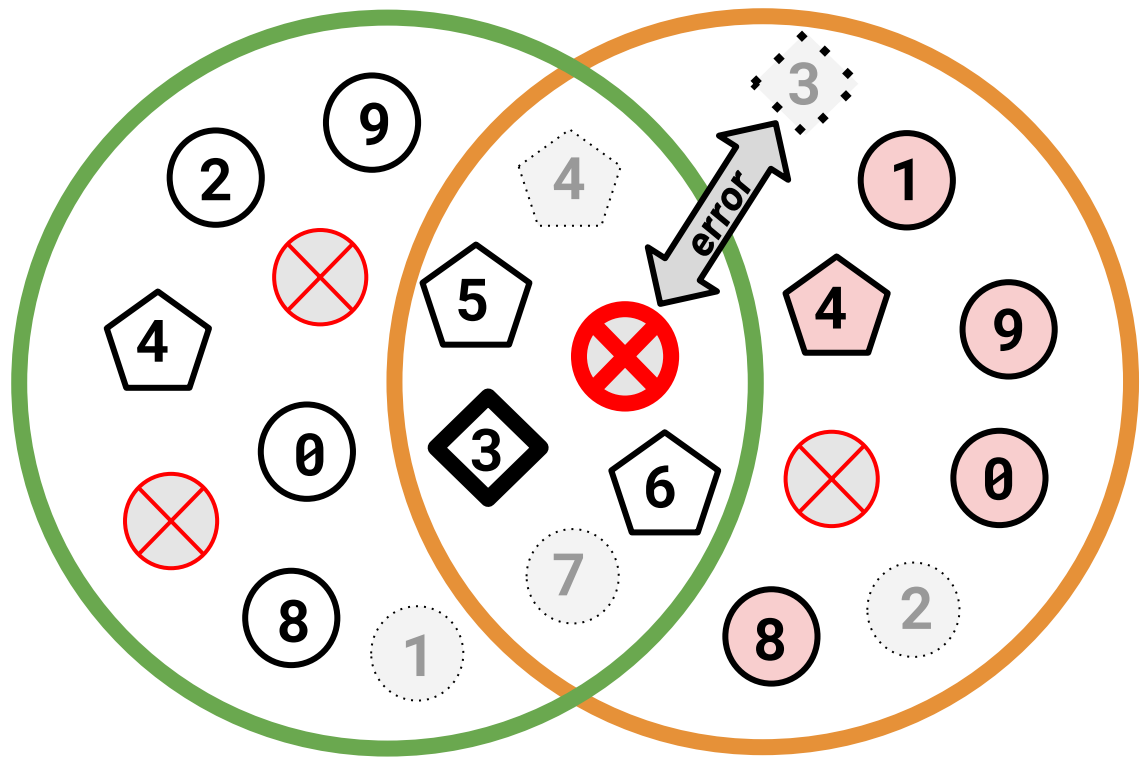
constraints of **sets** force us to remove information about occurrence



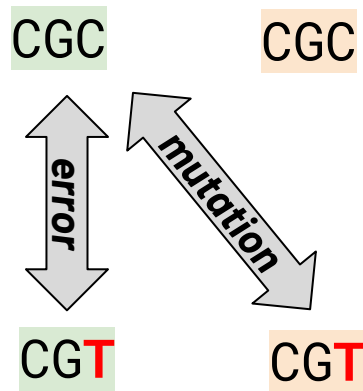
$$\mathbb{E}[I_{\text{set}}] = f(d, R)$$

$$I_{\text{set}} = 5$$

How repeats affect the intersection



repeats allow for collisions between mutations and errors



$$\mathbb{E}[I] = f(d, R, \lambda, \epsilon)$$

$$I = 4$$

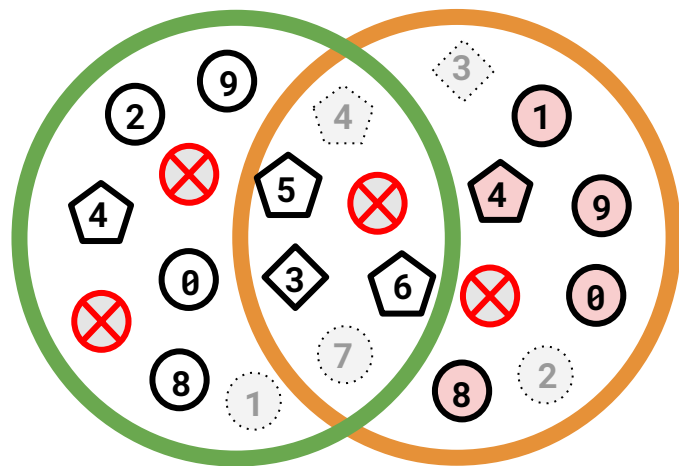
 I_0 +

are not mutated in the evolution from parent to child

$$I_1$$

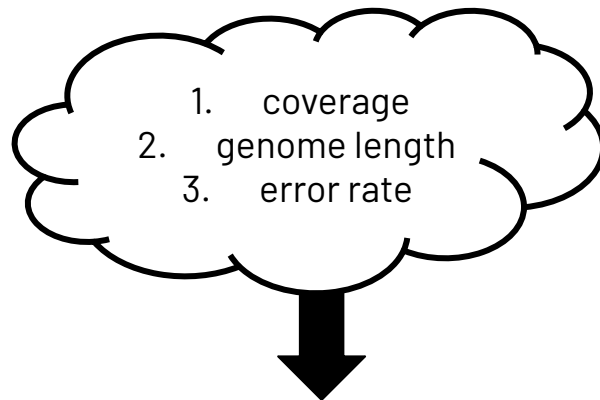
in intersection
due to error

How repeats affect the intersection



I =

intersection is comprised of
two different sets of kmers



replace
ratio-based
estimators with
spectrum-based
estimators

I_0

+

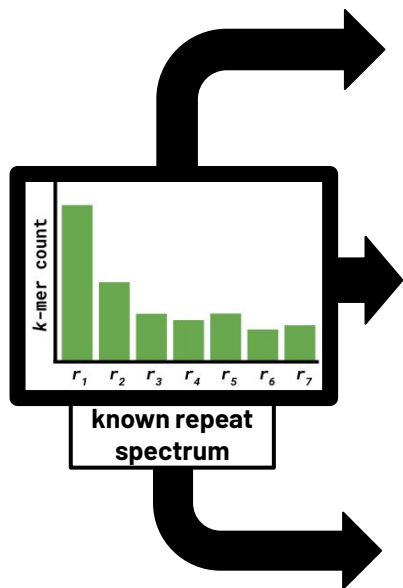
I_1

sampled **without error** in skim
1 and skim 2

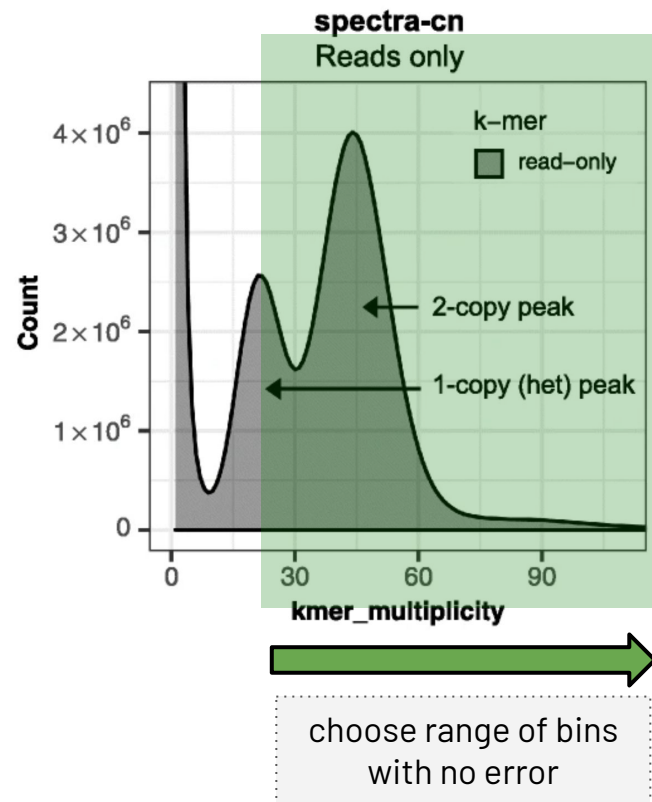
are not mutated in the
evolution from parent to child

in intersection
due to error

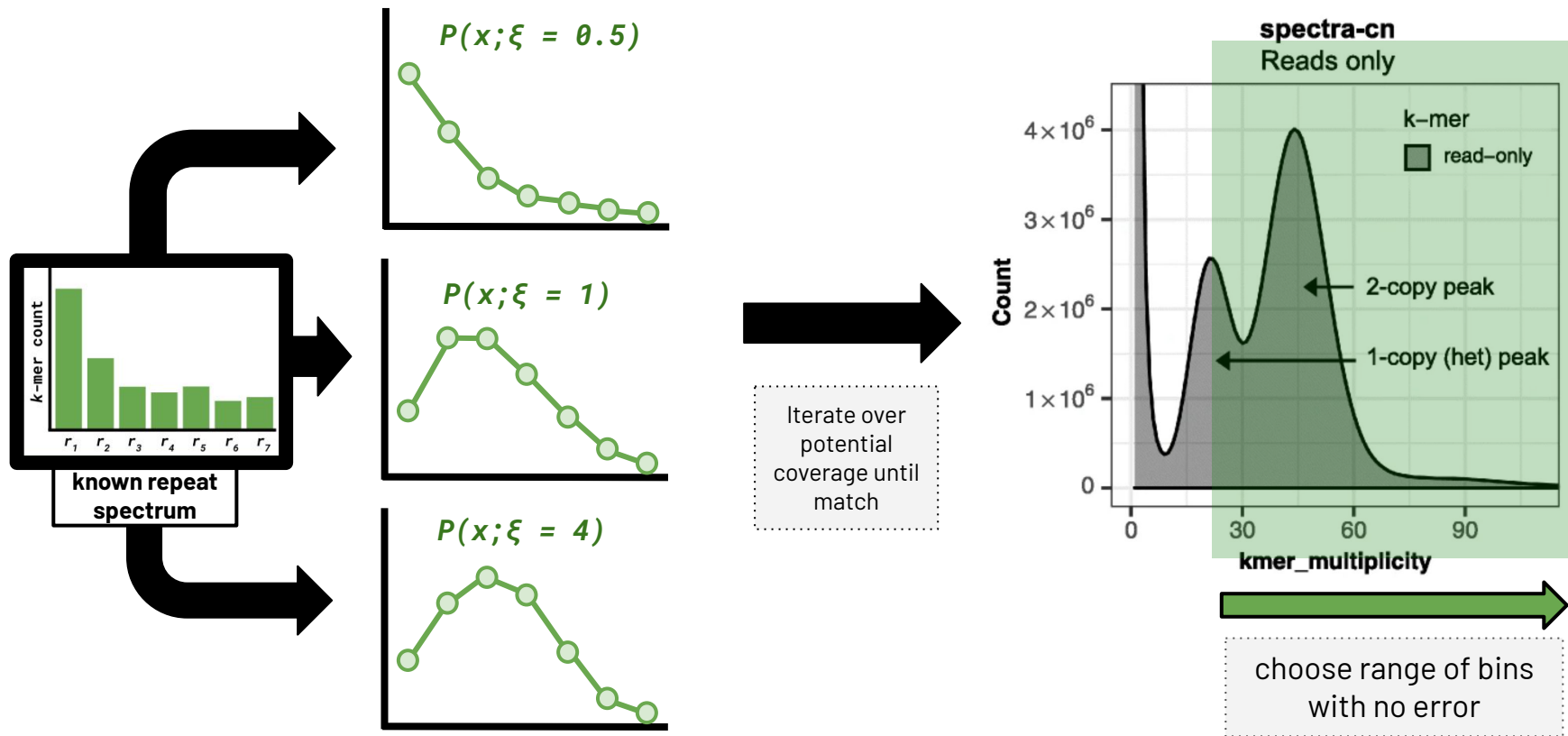
ReSkmer: estimating sequencing parameters



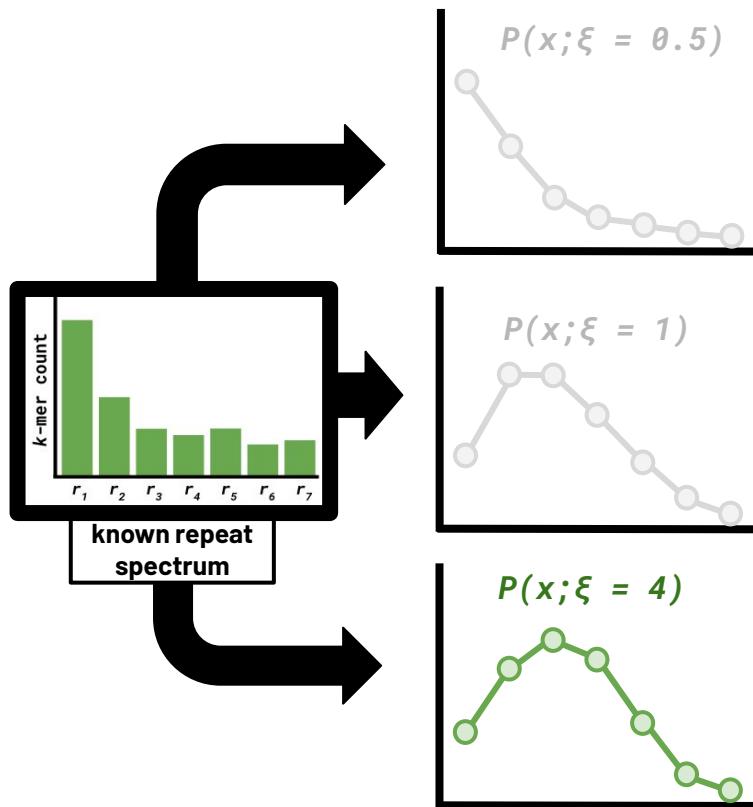
*(ξ) Error Free
k-mer coverage*



ReSkmer: estimating sequencing parameters



ReSkmer: estimating sequencing parameters



lambda (k-mer coverage):
all observed k-mers / genome length

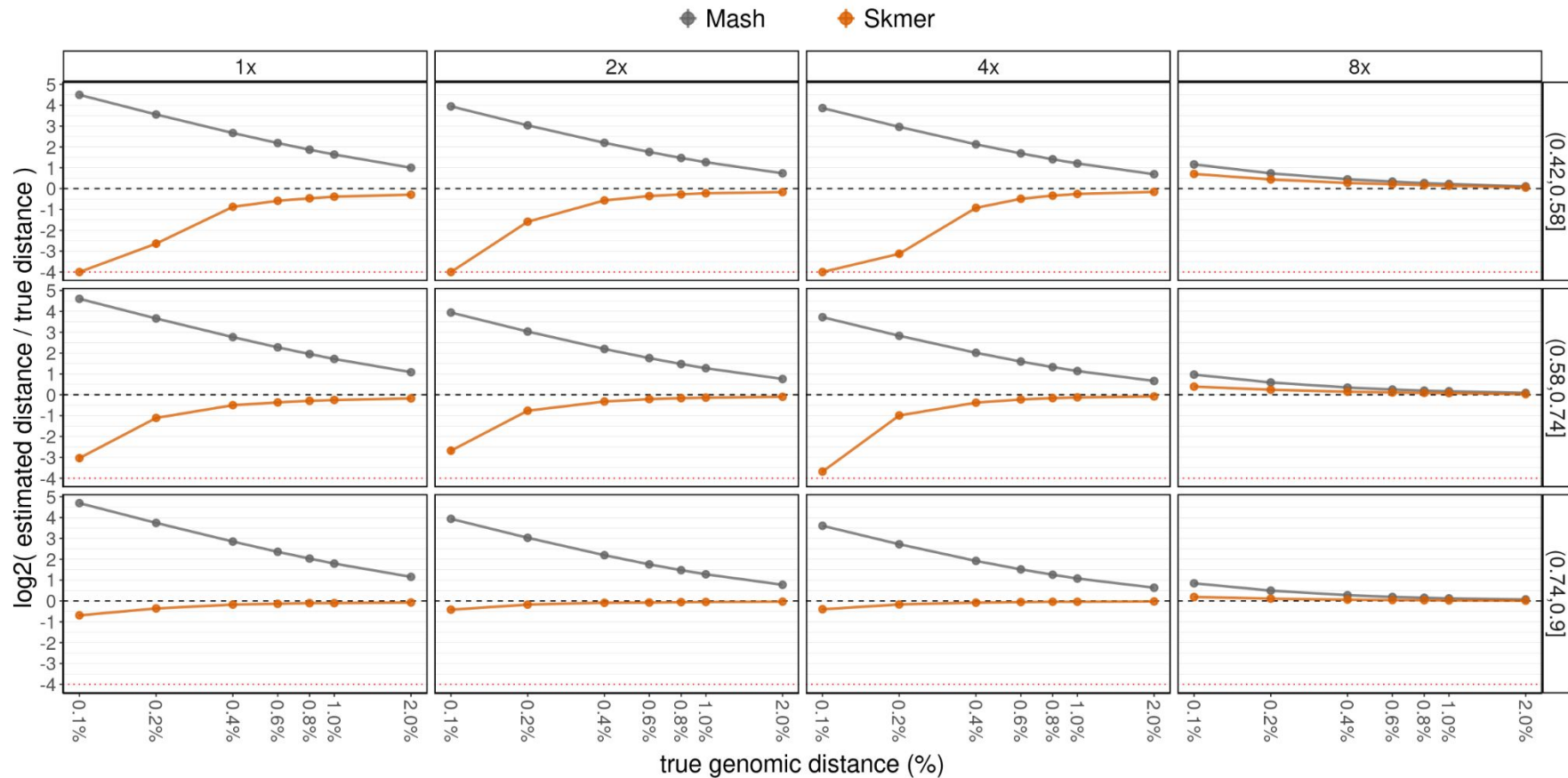
$$\lambda = \frac{\sum_h h o_h}{L}$$

Use **xi** and **lambda** to get **epsilon** .

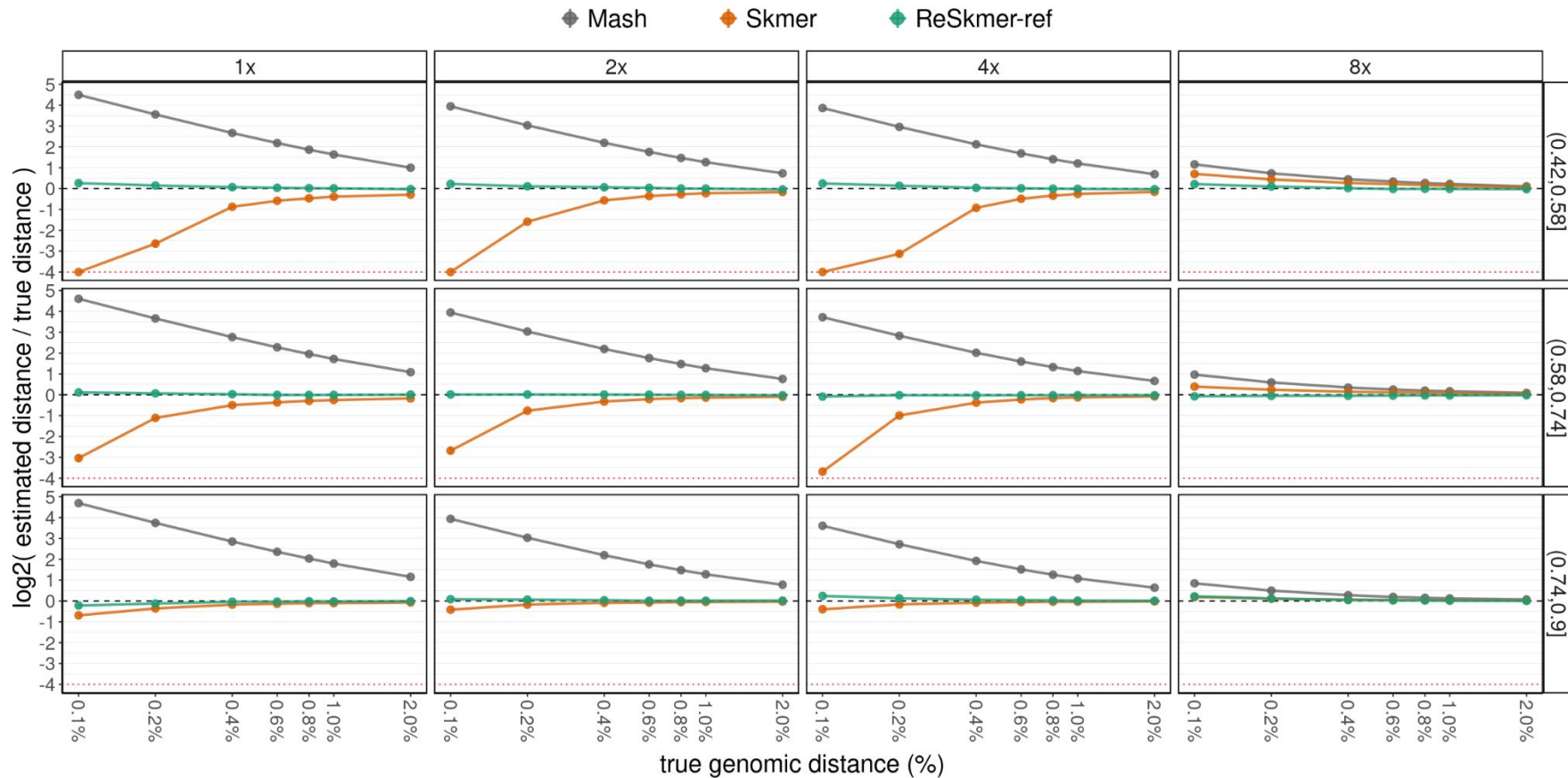
$$\epsilon = 1 - \left(\frac{\xi}{\lambda} \right)^{1/k}$$

Simulation Results

ReSkmer results

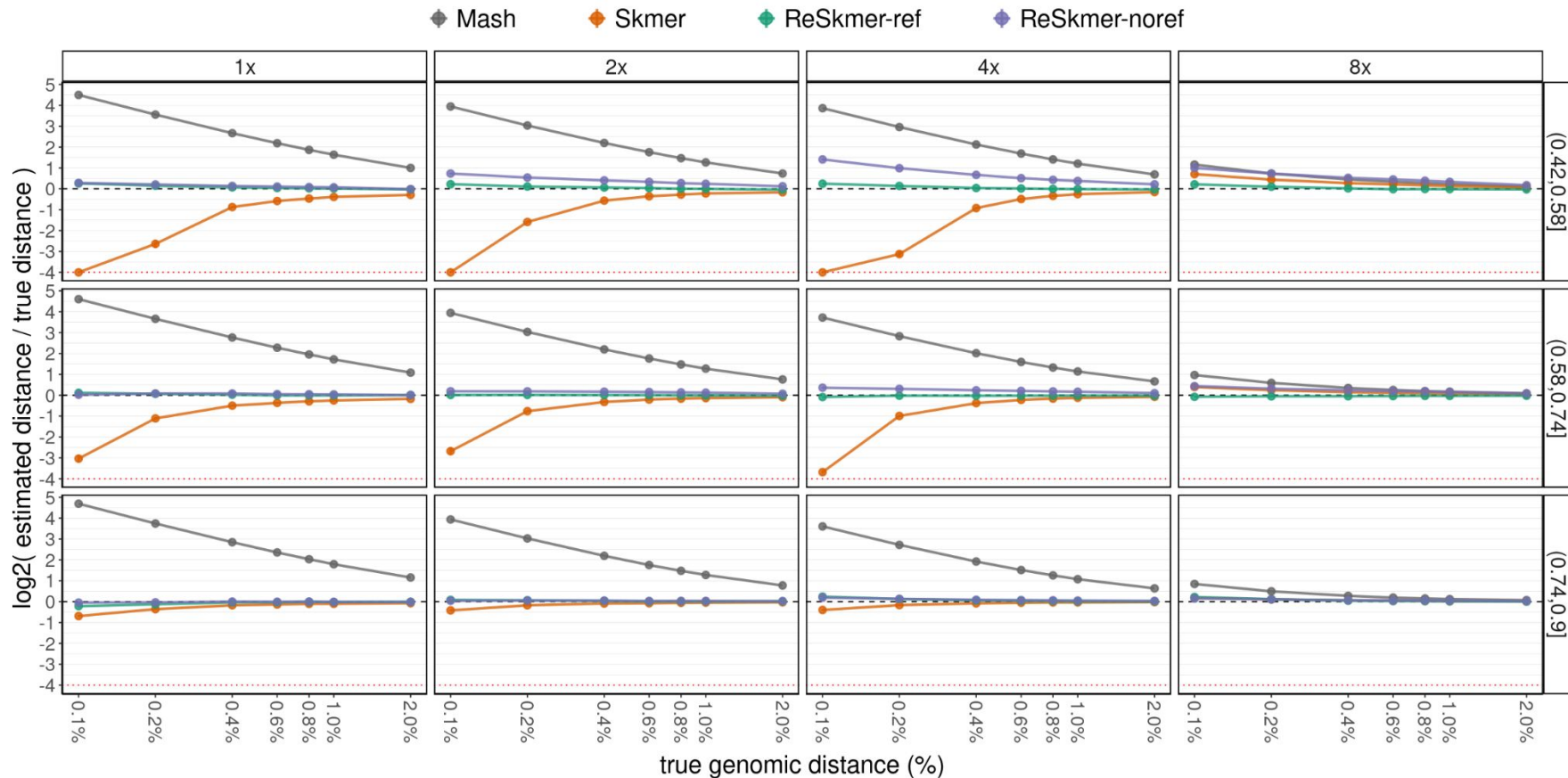


ReSkmer results



ReSkmer results

72



Apis mellifera Results

JOURNAL ARTICLE

Integrative Genomics Reveals the Genetics and Evolution of the Honey Bee's Social Immune System



Brock A Harpur , Maria Marta Guarna , Elizabeth Huxter , Heather Higo ,
Kyung-Mee Moon , Shelley E Hoover , Abdullah Ibrahim , Andony P Melathopoulos ,
Suresh Desai , Robert W Currie ... [Show more](#)

Genome Biology and Evolution, Volume 11, Issue 3, March 2019, Pages 937–948, <https://doi.org/10.1093/gbe/evz018>

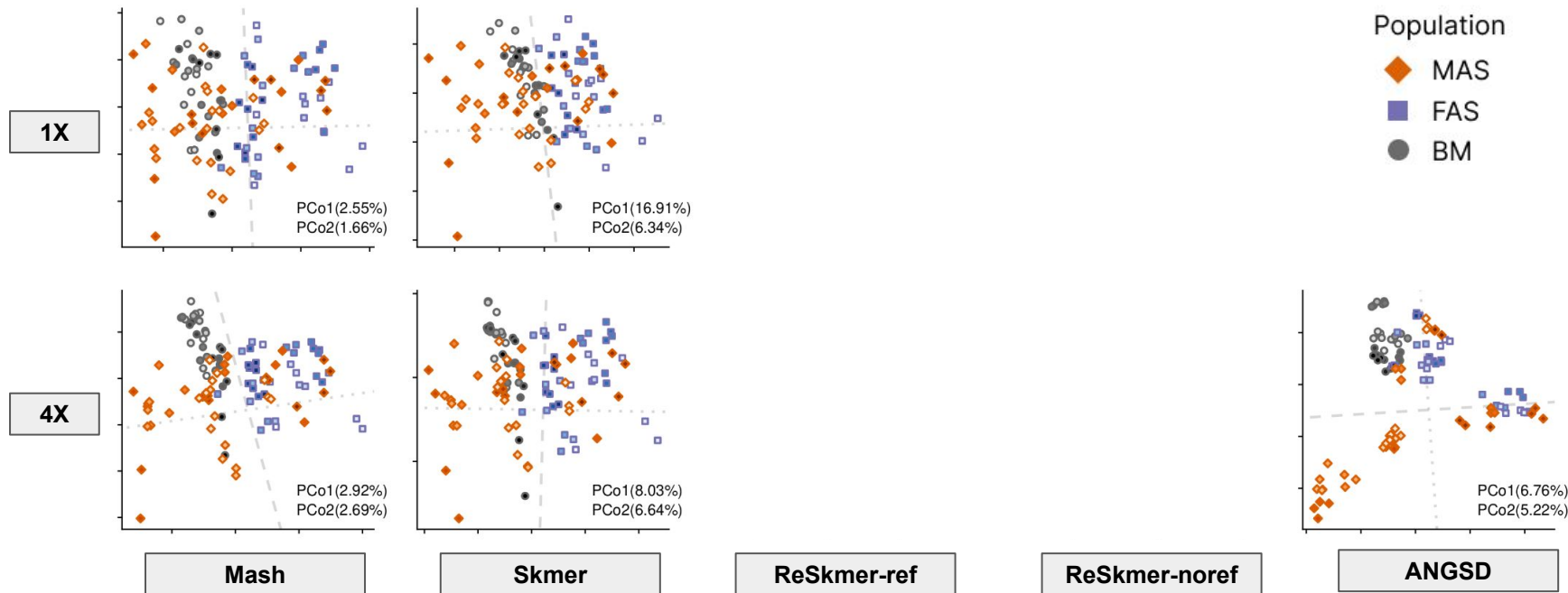
Honeybee clustering data



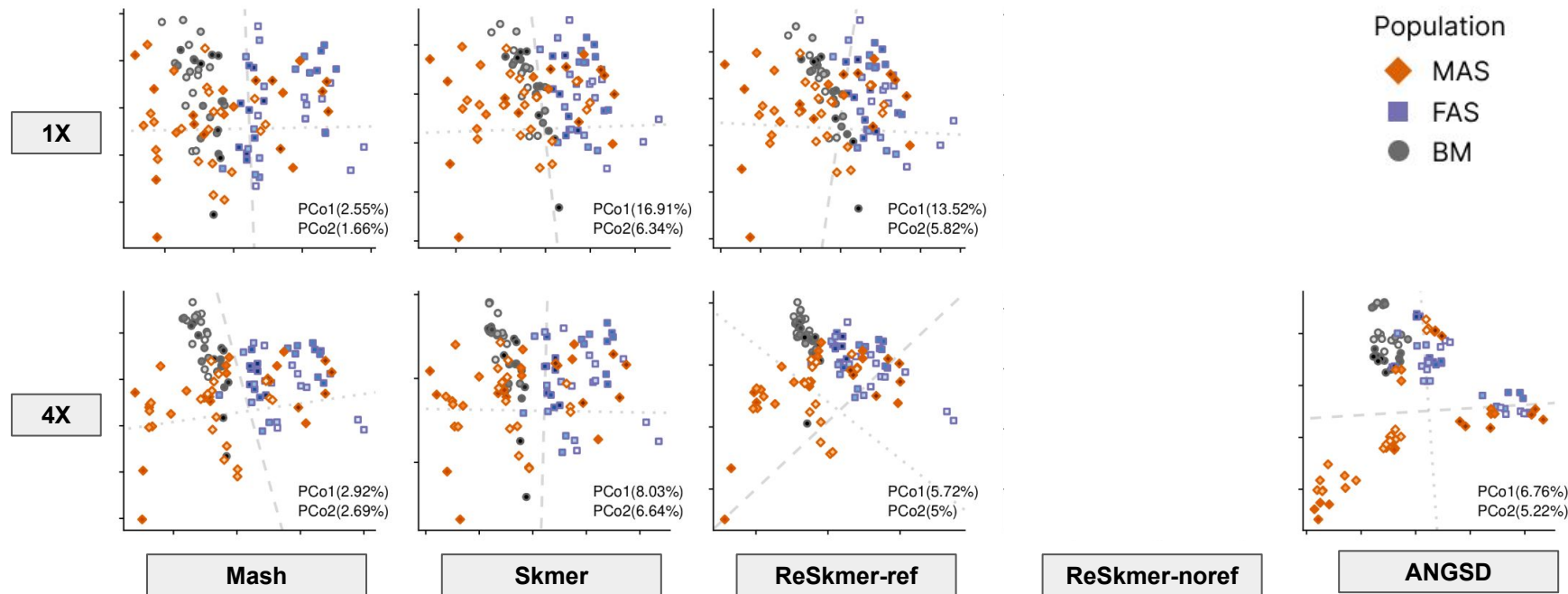
Honeybee clustering data



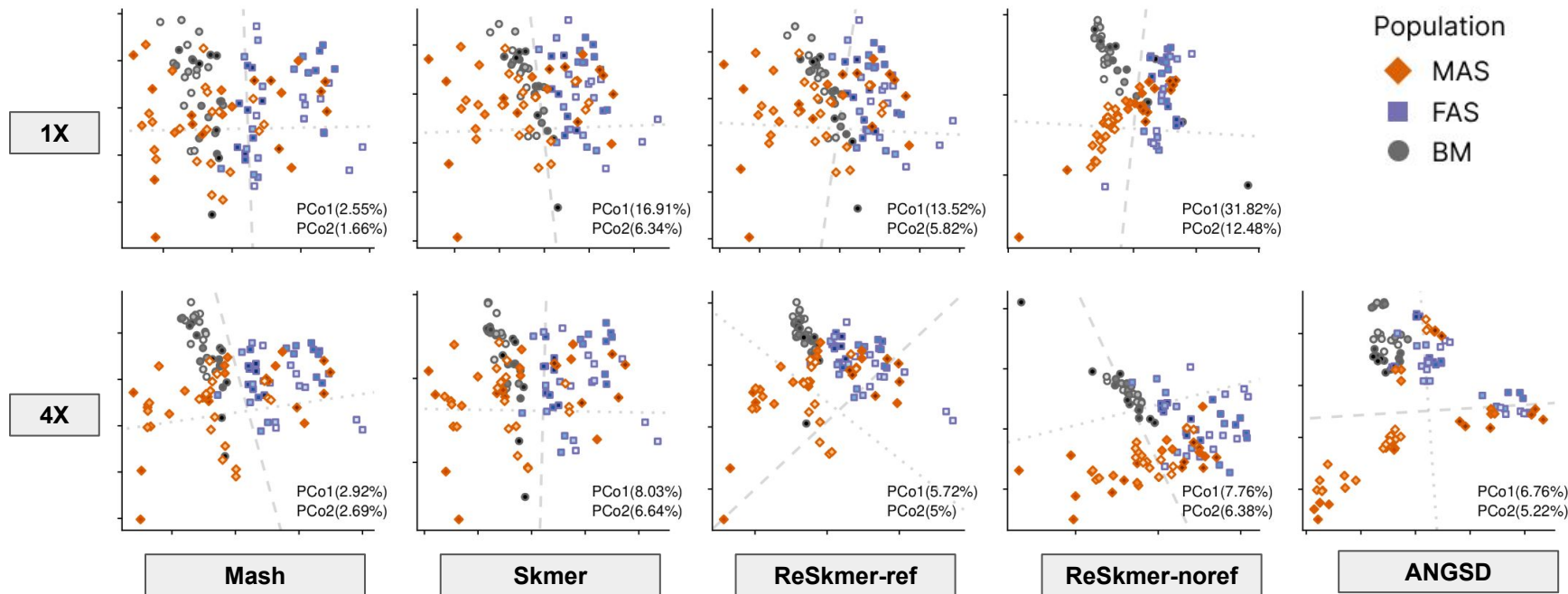
Honeybee clustering data



Honeybee clustering data



Honeybee clustering data



Correlation results

JOURNAL ARTICLE

The *Drosophila* Genome Nexus: A Population Genomic Resource of 623 *Drosophila melanogaster* Genomes, Including 197 from a Single Ancestral Range Population

Justin B Lack , Charis M Cardeno, Marc W Crepeau, William Taylor, Russell B Corbett-Detig, Kristian A Stevens, Charles H Langley , John E Pool 

Genetics, Volume 199, Issue 4, 1 April 2015, Pages 1229–1241, <https://doi.org/10.1534/genetics.115.174664>

Correlation results

How well do **alignment-free** methods correlate with **alignment-based** methods?

	<i>Apis mellifera</i>				<i>Drosophila melanogaster</i>			
	Mash	Skmer	ReSkmer (ref)	ReSkmer (noref)	Mash	Skmer	ReSkmer (ref)	ReSkmer (noref)
1×	0.12	0.61	0.60	0.65	0.12	0.58	0.88	0.76
2×	0.40	0.69	0.62	0.71	0.12	0.61	0.90	0.86
4×	0.63	0.77	0.58	0.71	0.40	0.69	0.88	0.90
6×	0.67	0.72	0.57	0.53	0.31	0.40	0.83	0.90

ANGSD

Unified Genotyper

Correlation results

How well do **alignment-free** methods correlate with **alignment-based** methods?

	<i>Apis mellifera</i>				<i>Drosophila melanogaster</i>			
	Mash	Skmer	ReSkmer (ref)	ReSkmer (noref)	Mash	Skmer	ReSkmer (ref)	ReSkmer (noref)
1×	0.12	0.61	0.60	0.65	0.12	0.58	0.88	0.76
2×	0.40	0.69	0.62	0.71	0.12	0.61	0.90	0.86
4×	0.63	0.77	0.58	0.71	0.40	0.69	0.88	0.90
6×	0.67	0.72	0.57	0.53	0.31	0.40	0.83	0.90

ANGSD

Unified Genotyper

Correlation results

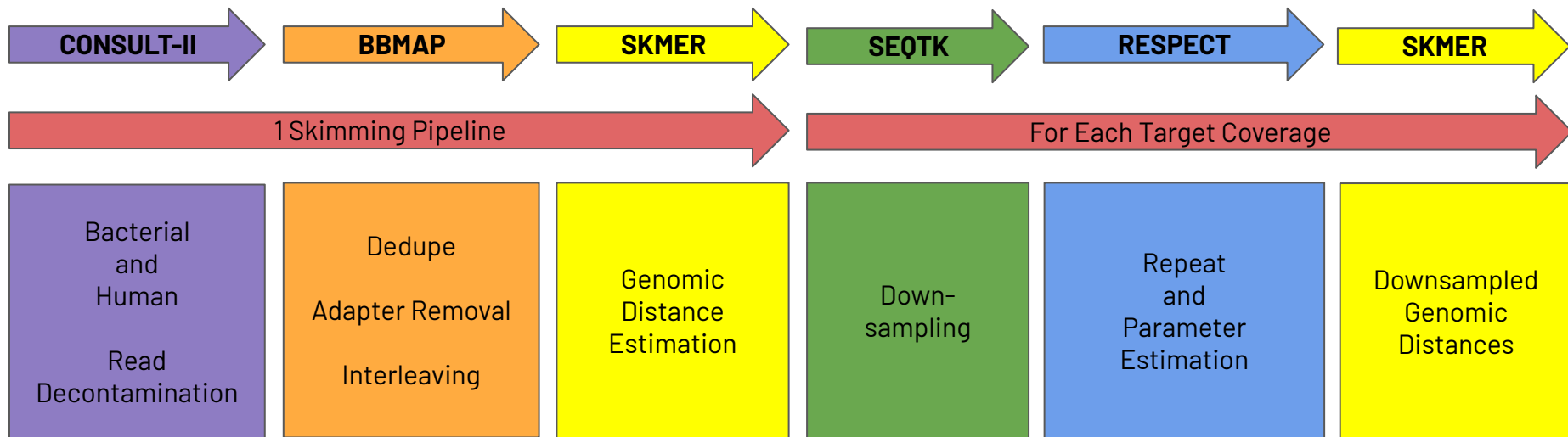
How well do **alignment-free** methods correlate with **alignment-based** methods?

	<i>Apis mellifera</i>				<i>Drosophila melanogaster</i>			
	Mash	Skmer	ReSkmer (ref)	ReSkmer (noref)	Mash	Skmer	ReSkmer (ref)	ReSkmer (noref)
1×	0.12	0.61	0.60	0.65	0.12	0.58	0.88	0.76
2×	0.40	0.69	0.62	0.71	0.12	0.61	0.90	0.86
4×	0.63	0.77	0.58	0.71	0.40	0.69	0.88	0.90
6×	0.67	0.72	0.57	0.53	0.31	0.40	0.83	0.90

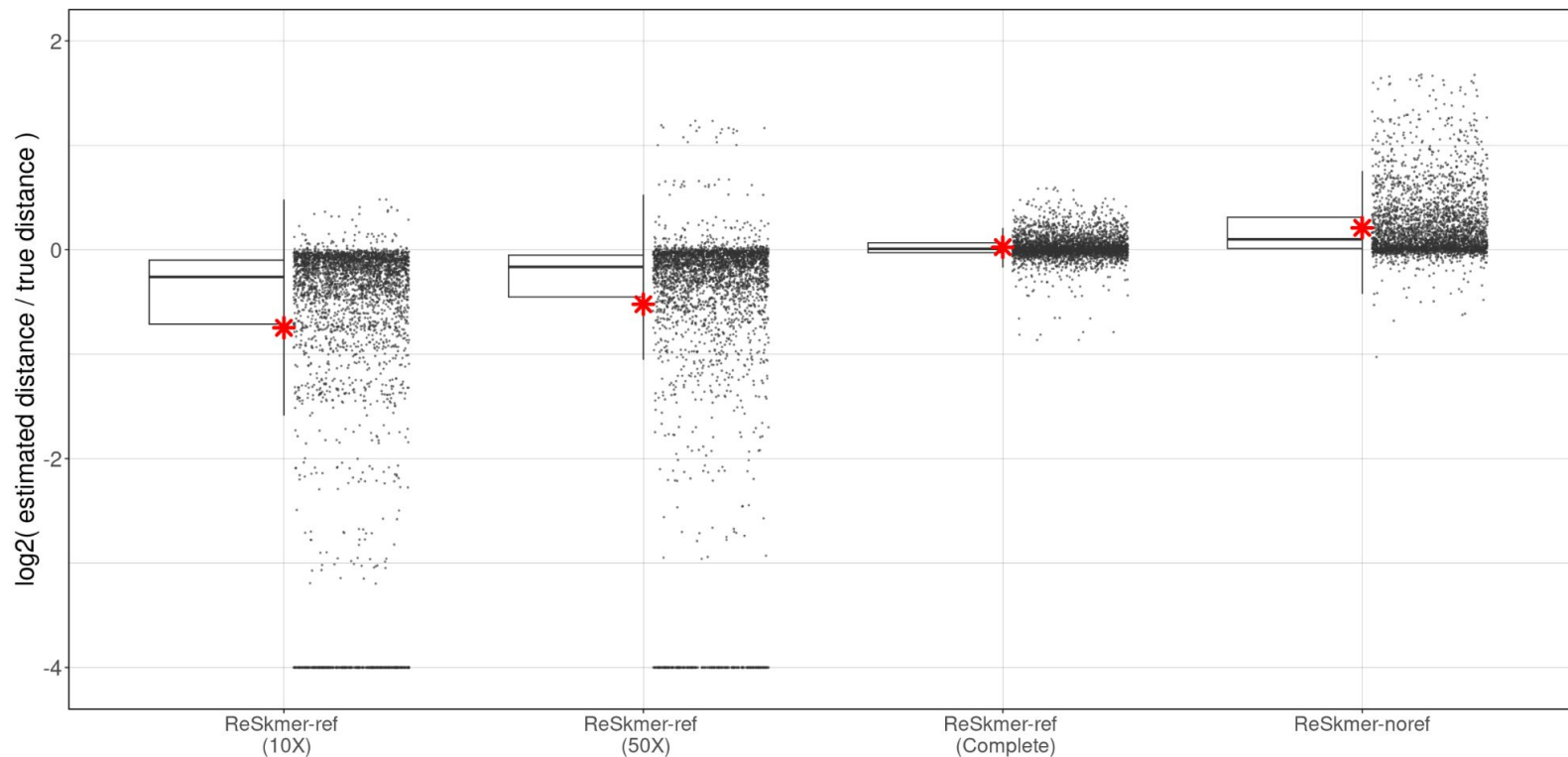
ANGSD

Unified Genotyper

Skimming Pipeline Overview



What Repeat Spectrum should I use?



Conclusions

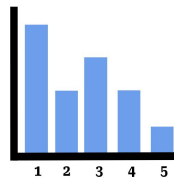
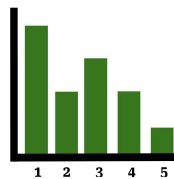
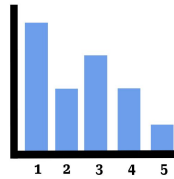
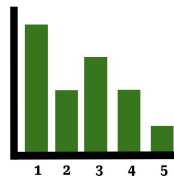
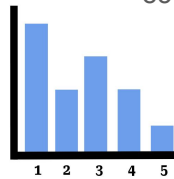
- ReSkmer models how **repeats affect the Jaccard Index** between k -mer sets
- ReSkmer builds on the groundwork set up by Mash and Skmer
 - Fast and scalable, whole genome resolution, reference and alignment-free, low coverage
- ReSkmer allows us to calculate **very small distances very accurately** (pop-gen)
- ReSkmer submitted to be published and is under review

ReSkmer: Modeling repeats allows k -mer-based alignment-free

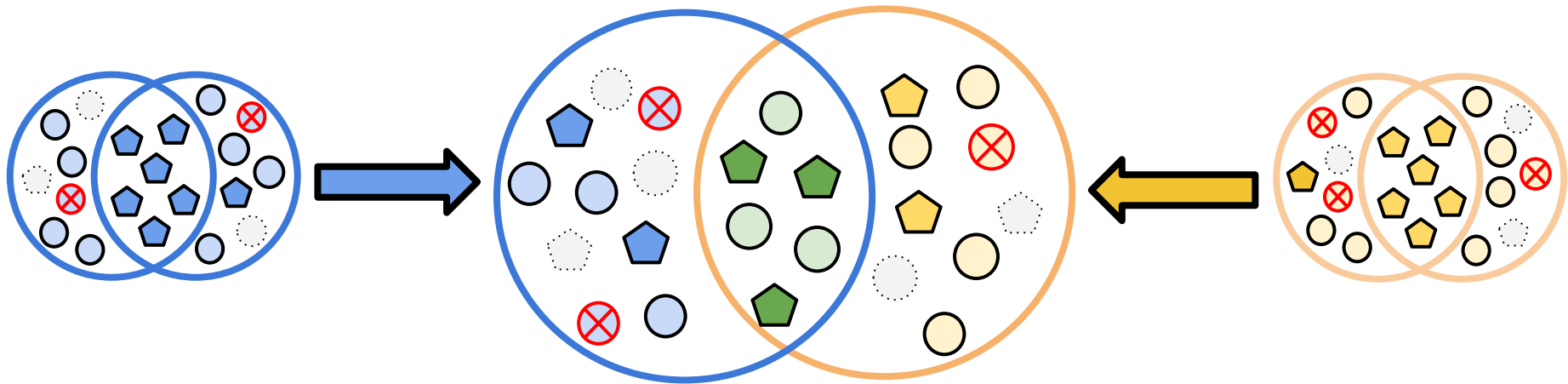
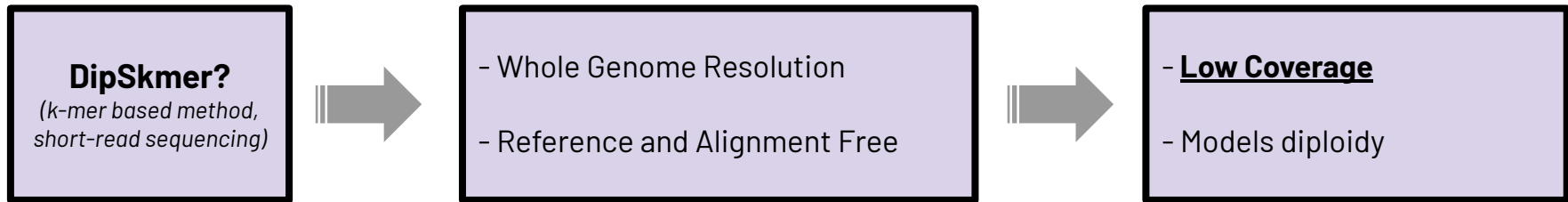
methods to calculate population genomic distances

Eduardo Charvel^{*2}, Isaac Thomas^{*3}, Homère J. Alves Monteiro⁴, Shahab Sarmashghi⁵,
Glenn Dunshea⁶, Vineet Bafna³, and Siavash Mirarab¹

- **Code is here!** <https://github.com/echarvel3/ReSkmer>



Future Work





ALFRED P. SLOAN
FOUNDATION



FIELD. Grainger
Bioinformatics
Center

Thank you!

Mirarab Lab:

Ali, Daira, Shayasteh, Eleanora, Alan, Puoya, Kaushik, and Yueyu

Funding:

Field Museum: Grainger Bioinformatics Center
Minderoo Foundation
Alfred P. Sloan Foundation

edcharvel@ucsd.edu

