ReSkmer:
Modeling repeats improves population genomic distances between genome skims

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

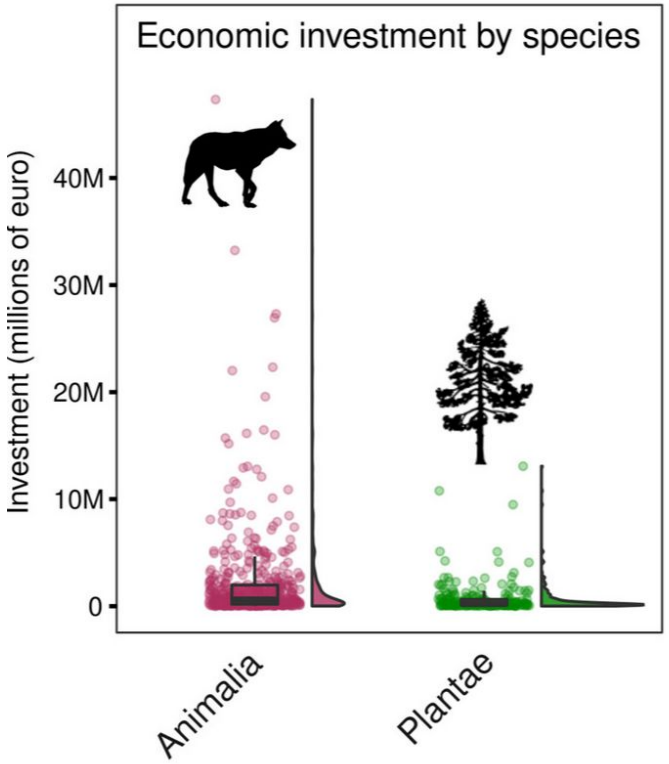
Accessible population genomics

Sequencing costs have dropped.

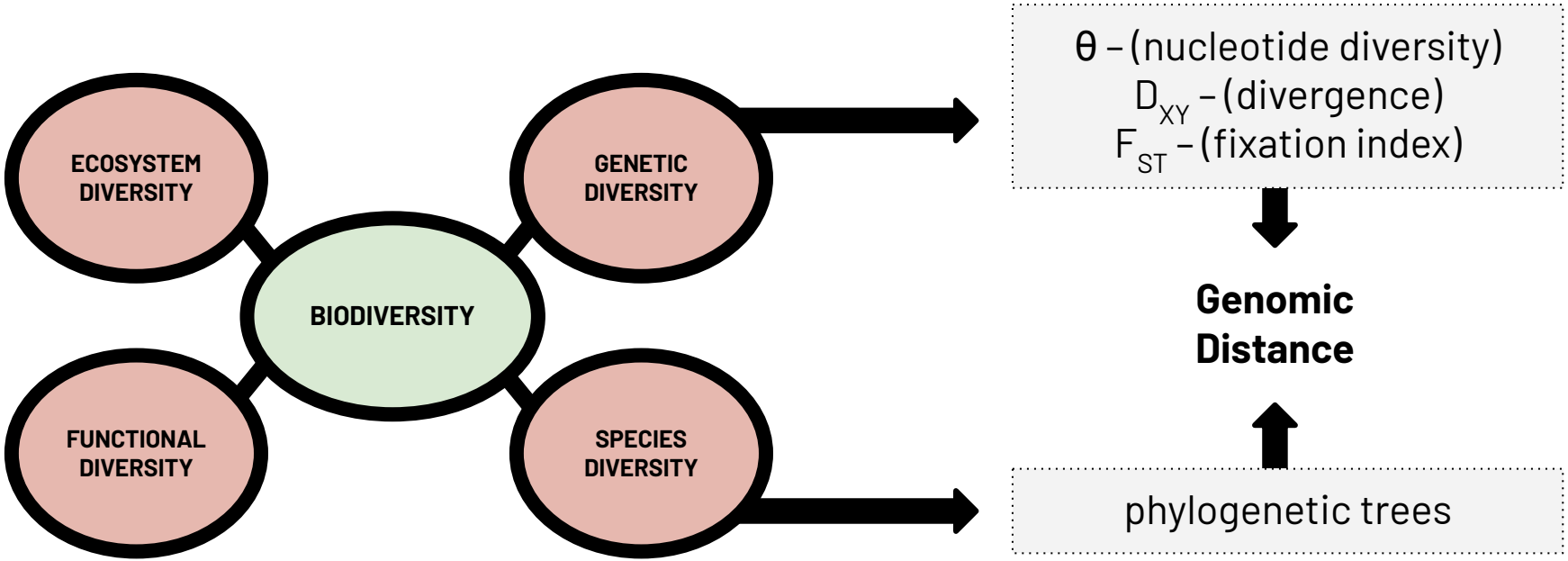
Conservation funding can be **biased**.

Cost efficient methods for measuring biodiversity are crucial.

(Adamo et al. 2022)

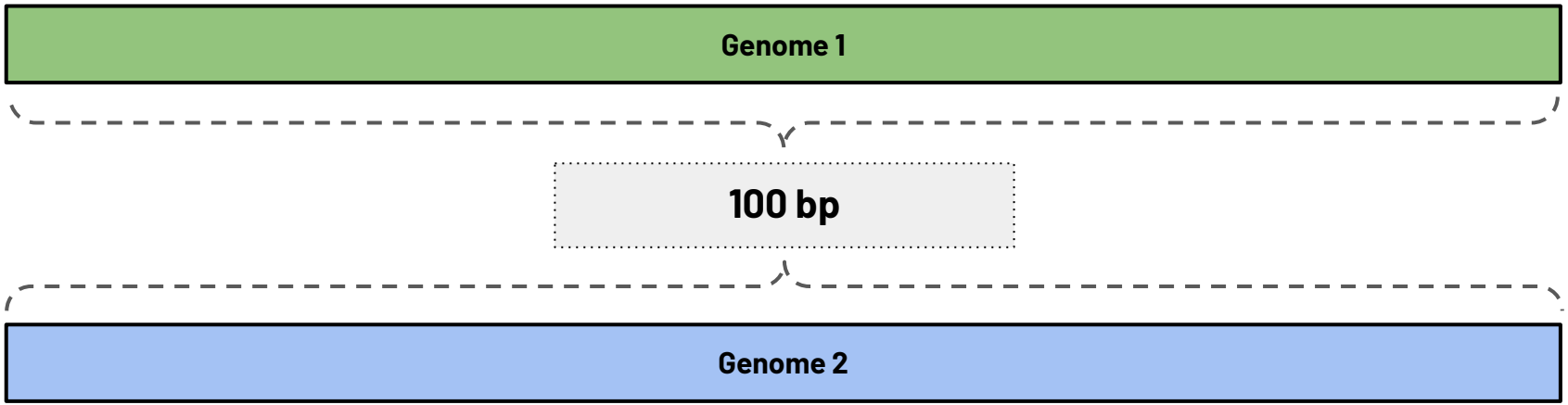


How can we measure biodiversity efficiently?

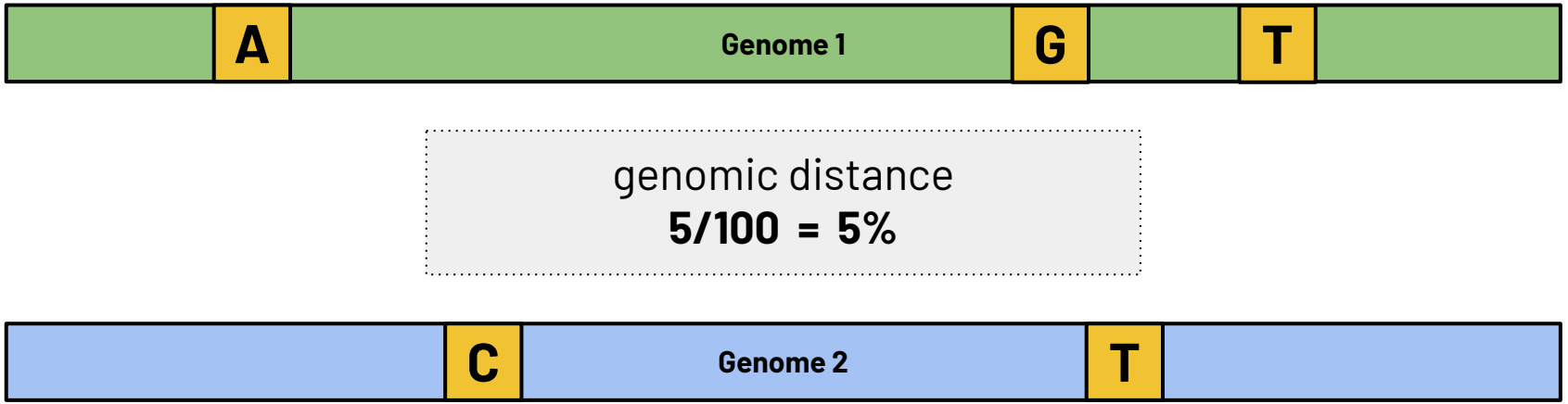


4 pillars of biodiversity

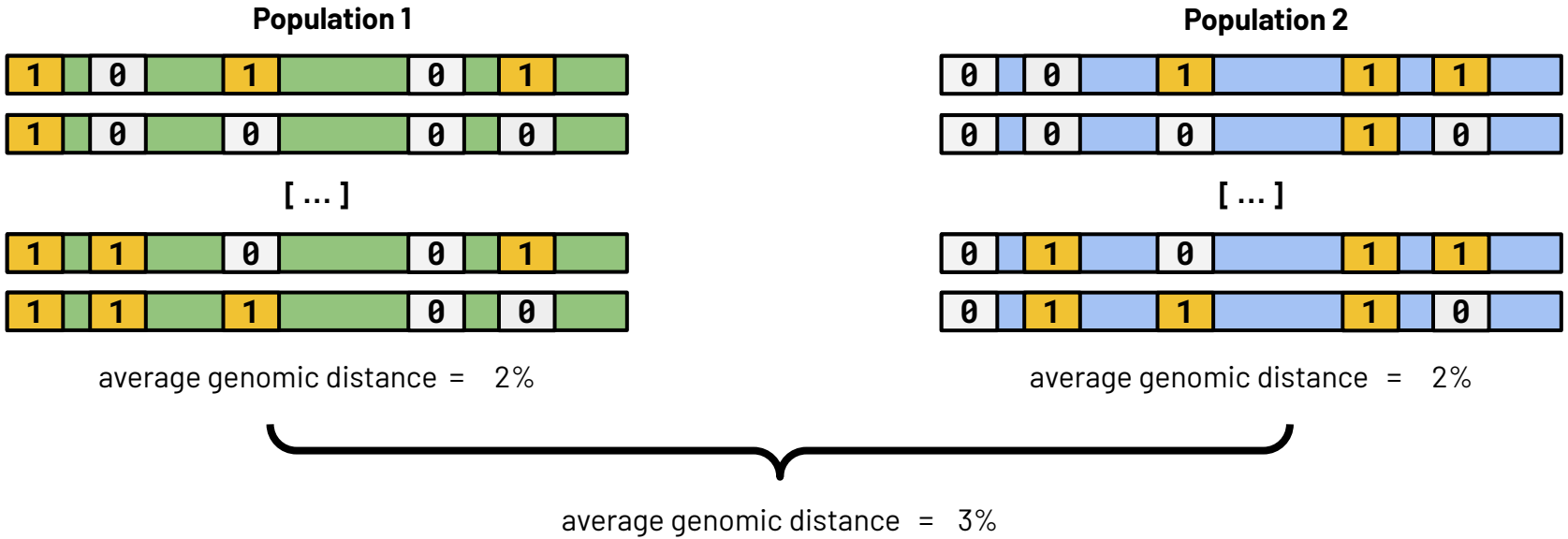
Genomic distance example



Genomic distance example



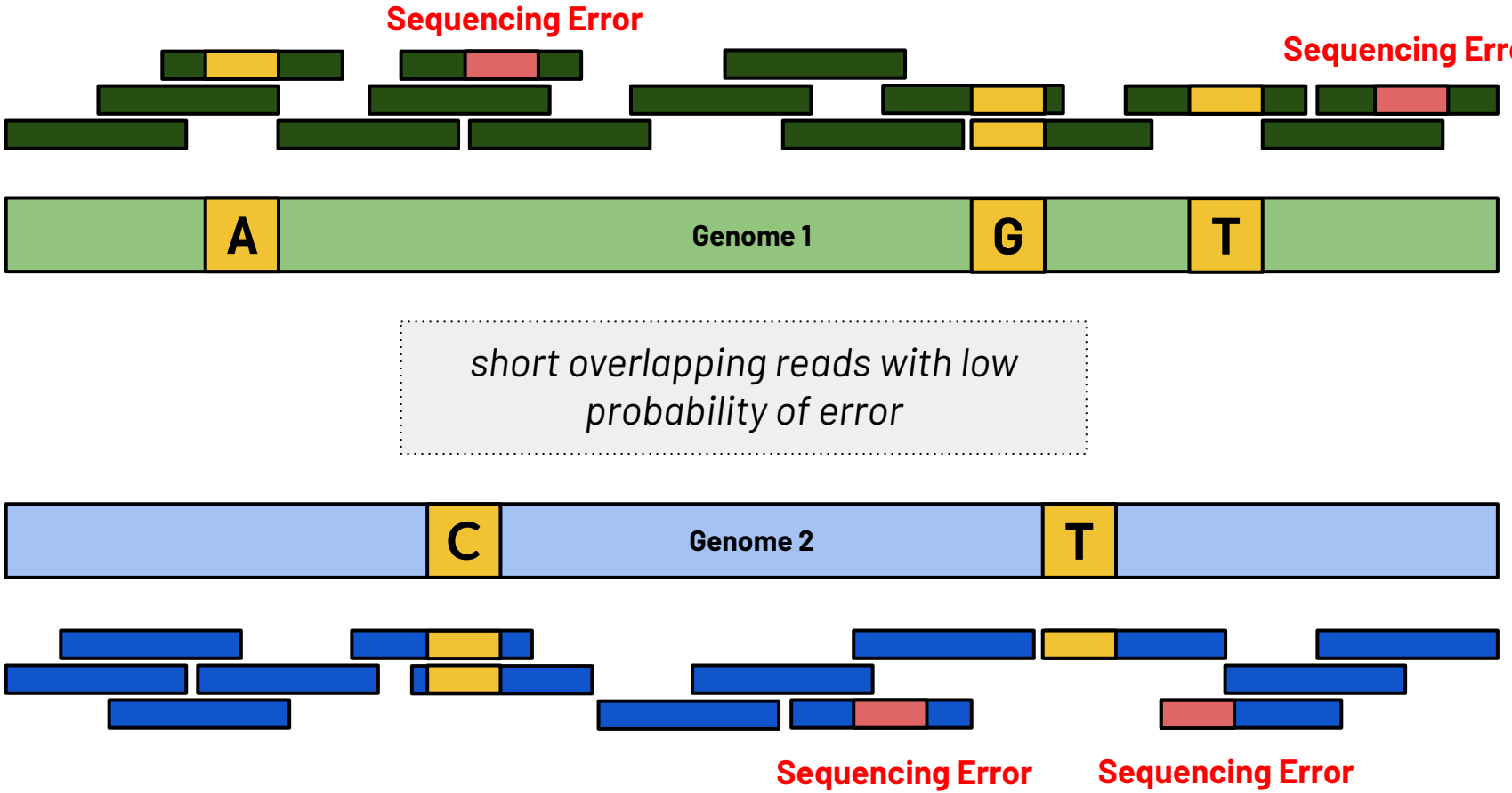
Genomic distances between populations



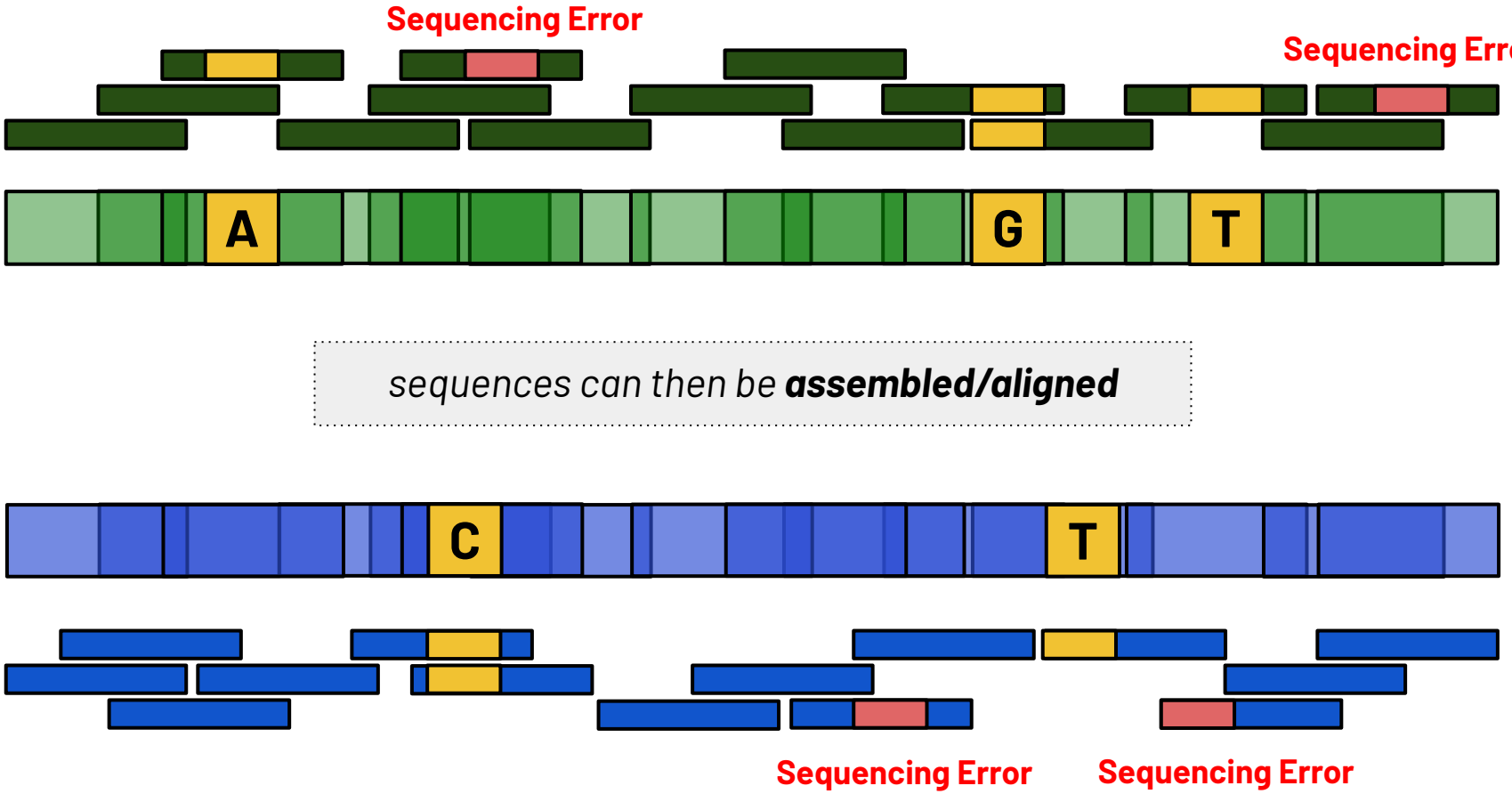
small values – population genomic distances

large values – phylogenetic distances

Short-read sequencing



Assembly / Alignment



Methods	Advantages	Disadvantages
Marker Based *(single genes, mitochondrial DNA, microsatellites)*	- Cost-effective Sanger Sequencing	- Limited Resolution - 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

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

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

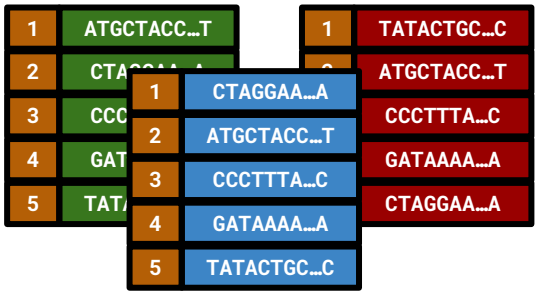
Skmer
(k-mer based method, short-read sequencing)

Advantages

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

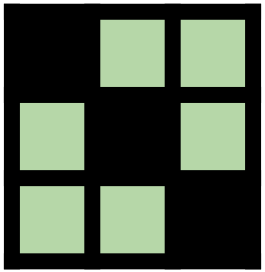
Disadvantages

- Assumes no repeats



fastq per sample

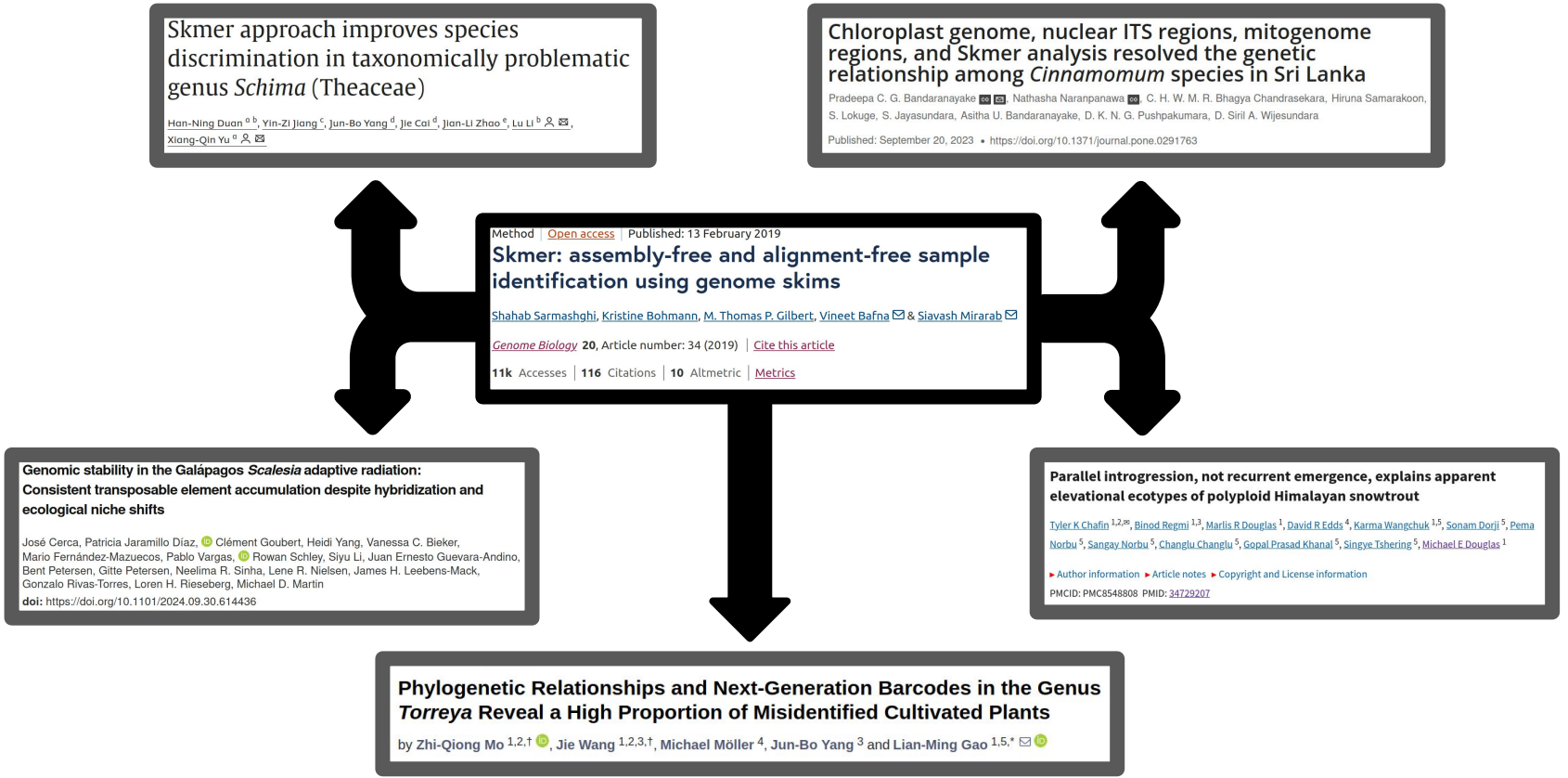
Input



distance matrix

Output

Skmer applications

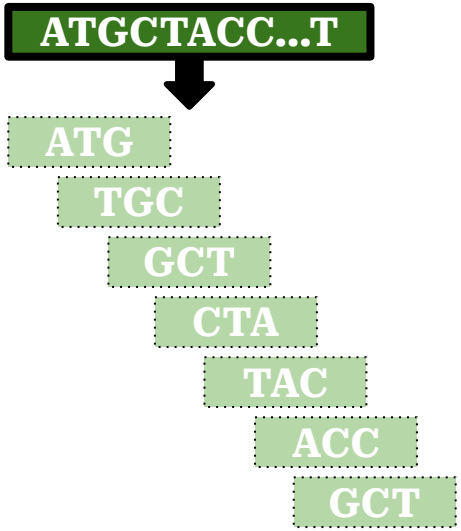


Alignment-free distance estimation

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**

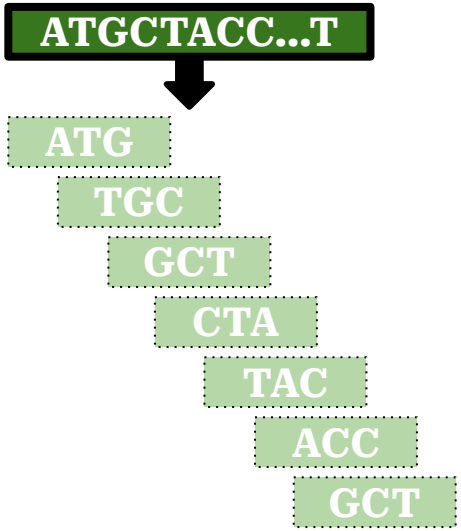
Alignment-free distance estimation

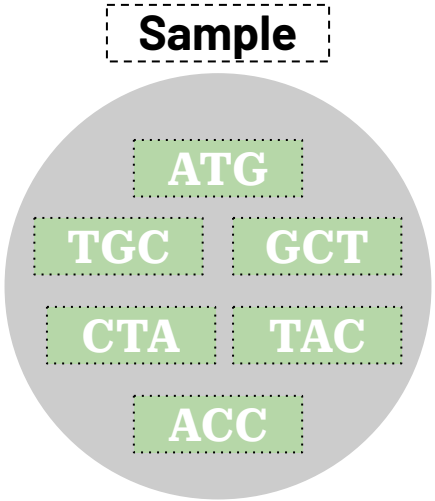


begin collection of reads **~ 150 bp**

reads are decomposed into ***k*-mers**

Alignment-free distance estimation



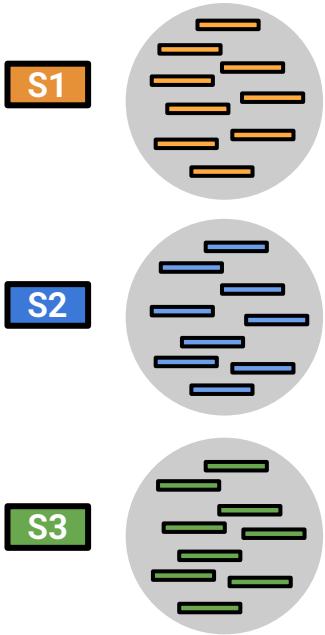


begin collection of reads **~ 150 bp**

reads are decomposed into ***k*-mers**

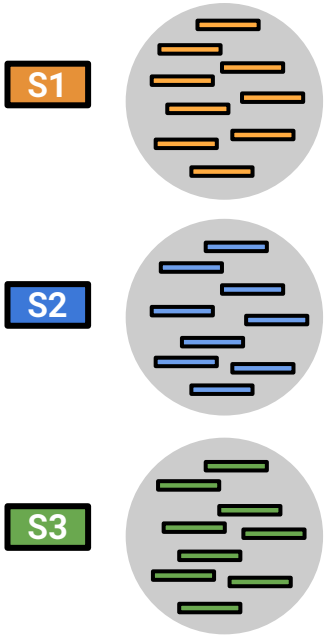
sets store **unique** *k*-mers

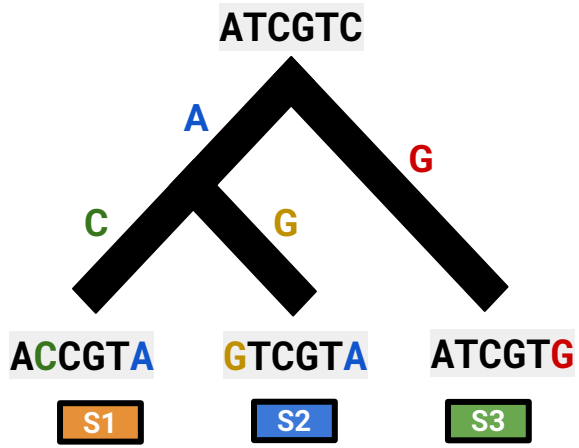
Alignment-free distance estimation



k-mer set for each sample

Alignment-free distance estimation

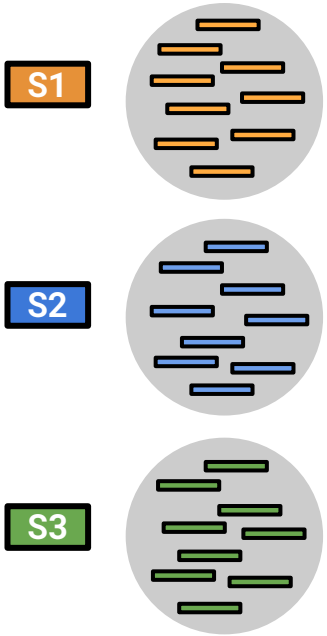


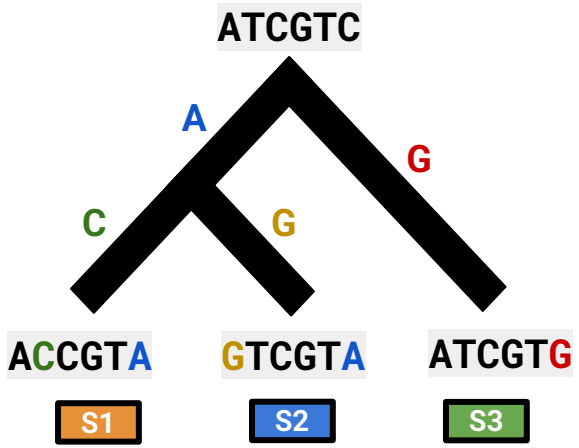


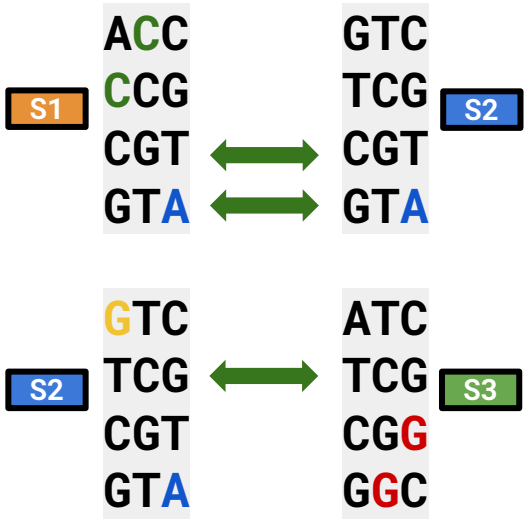
k-mer set for each sample

samples will be more or less distant to one another

Alignment-free distance estimation





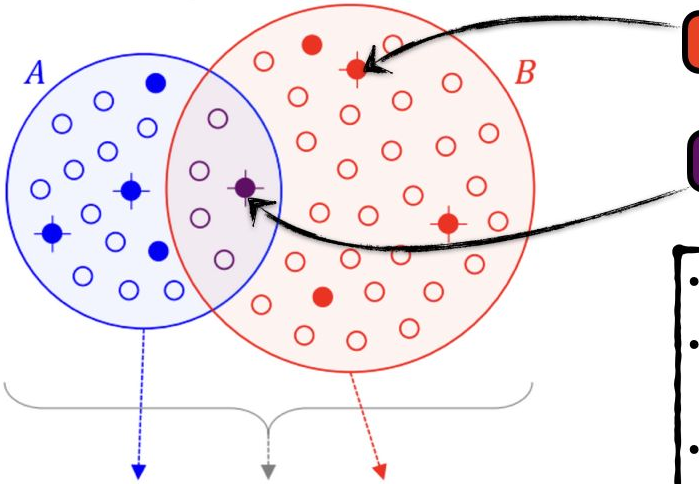


k-mer set for each sample

samples will be more or less distant to one another

closely related samples will have more *k*-mer matches

Mash for Alignment-Free Distances



$S(A)$	$S(A \cup B)$	$S(B)$
42	42	66
64	64	82
82	66	87
128	**82**	104
139	87	127

#104 ← ATAC…GCTA → Heterozygous k-mer

#82 ← ATAC…GCTA → Homozygous k-mer

- 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**.

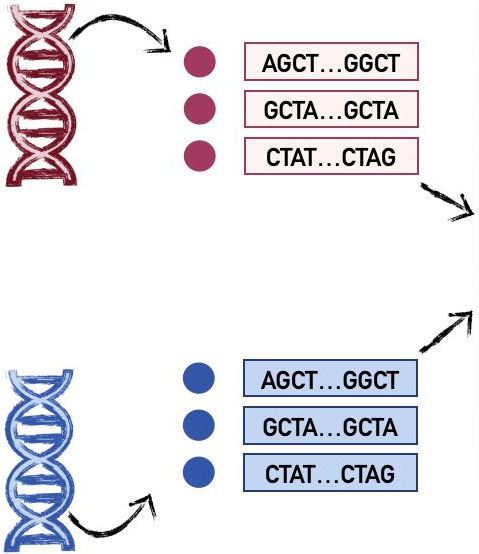
$$d = 1 - \left(\frac{2J}{J+1}\right)^{1/k}$$

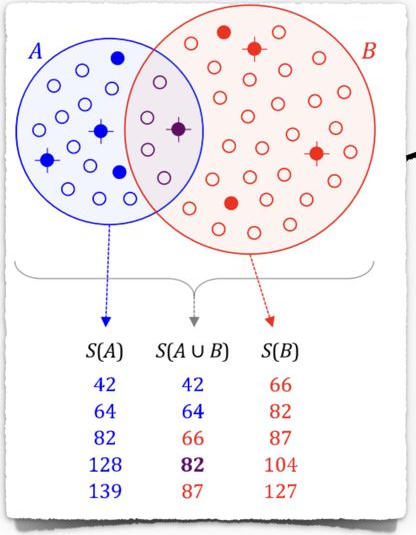
Assembly-Free, Alignment-Free Distances

Decompose DNA into k-mer sets

Jaccard Similarity of sketches

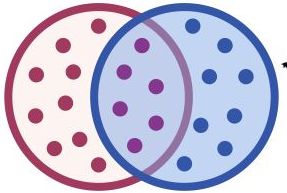
Results depend on the data





(Ondov et al. 2016)

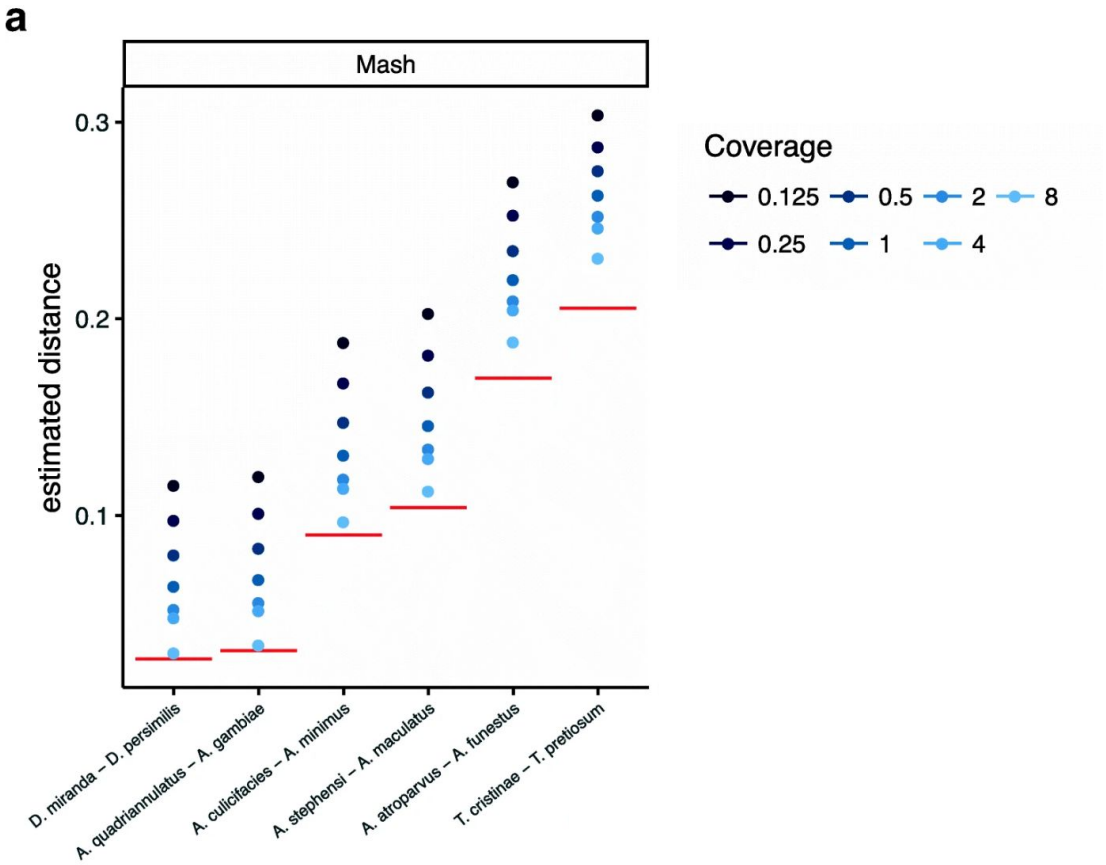
Complete Genome



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

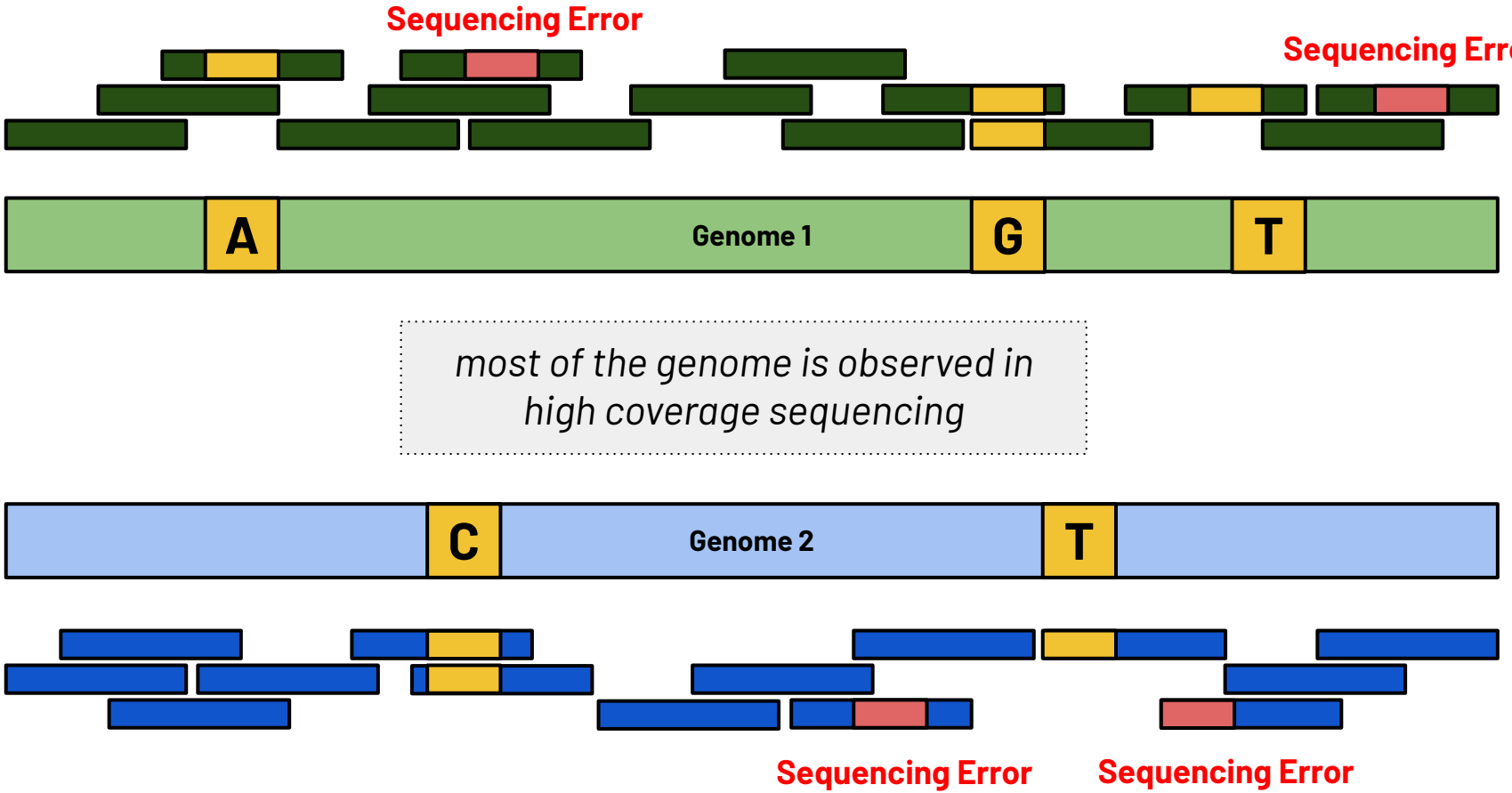
$$d = 1 - \left(\frac{2J}{J+1}\right)^{1/k}$$

Mash fails with genome skimming data

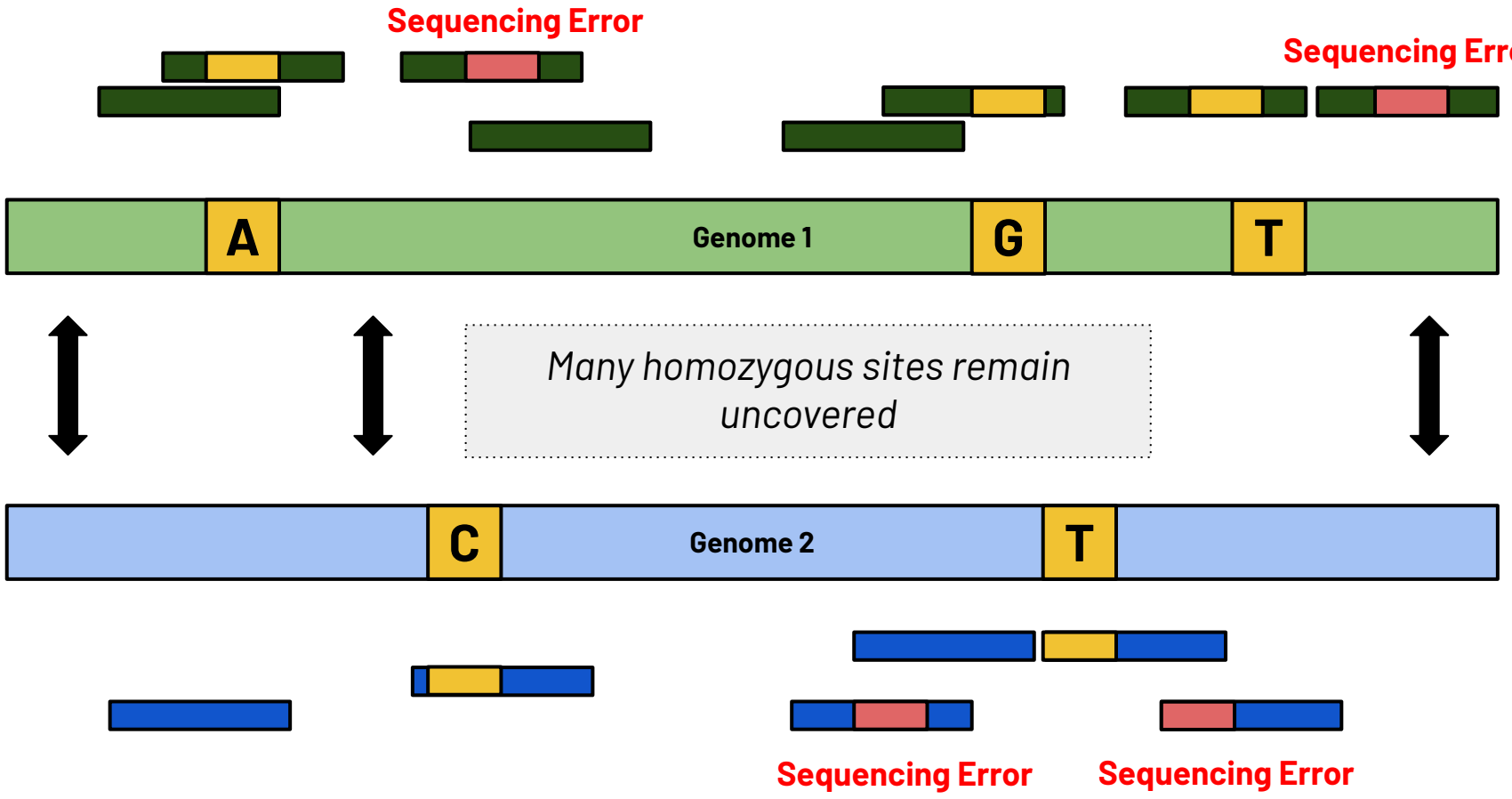


Low coverage sequencing leads to distance **overestimation** with mash

High coverage sequencing



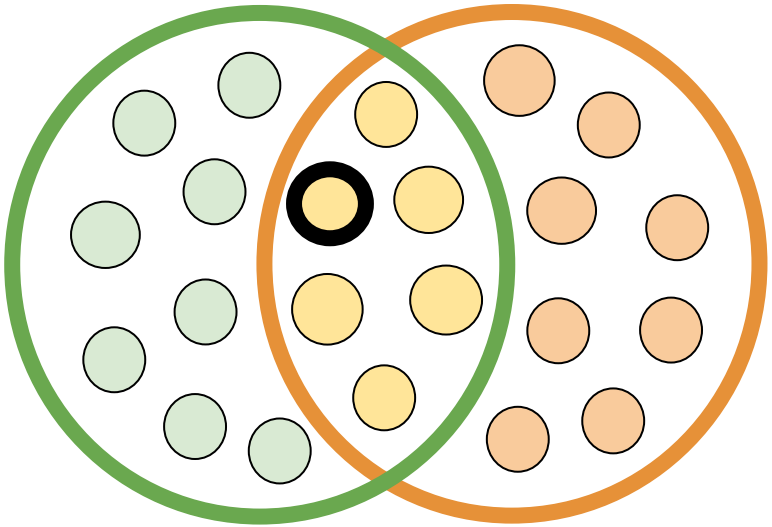
Low coverage sequencing



Jaccard Index example

High coverage
Jaccard Index

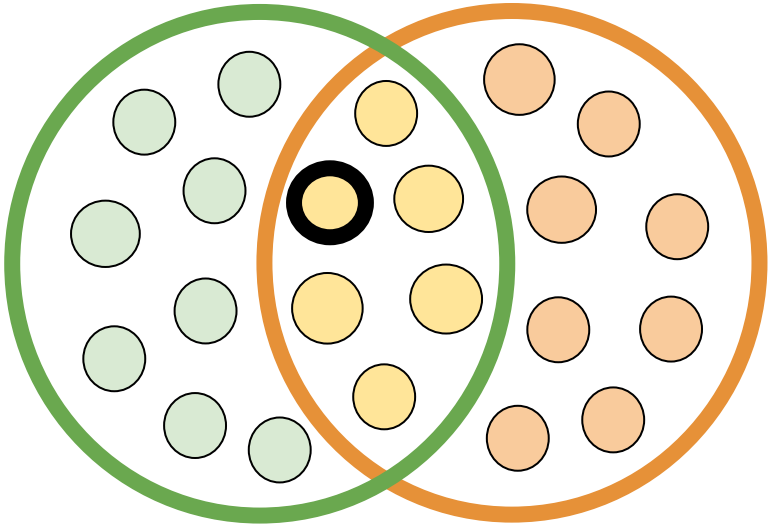
Low coverage
Jaccard Index



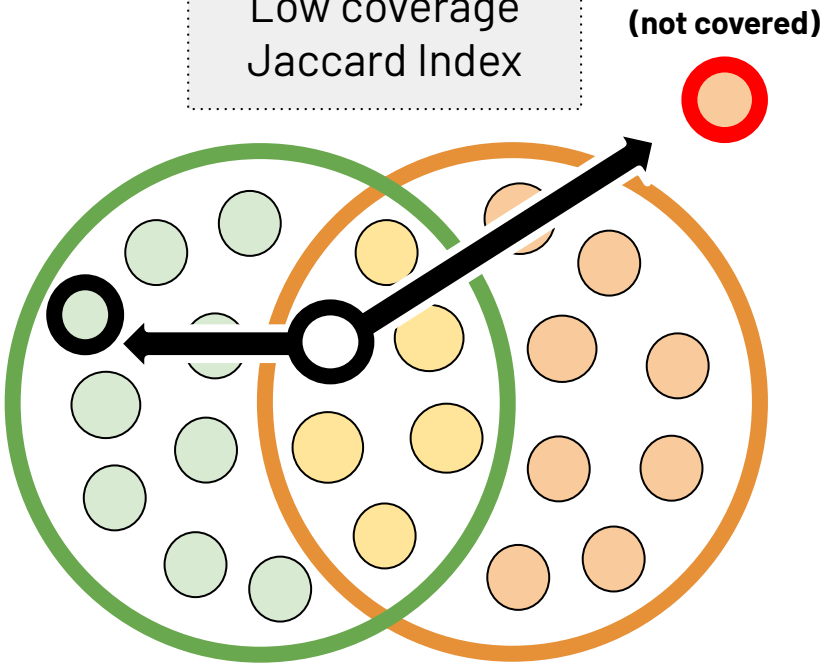
$$d = 1 - \left(\frac{2J}{J+1}\right)^{1/k}$$

Jaccard Index example

High coverage
Jaccard Index



Low coverage
Jaccard Index

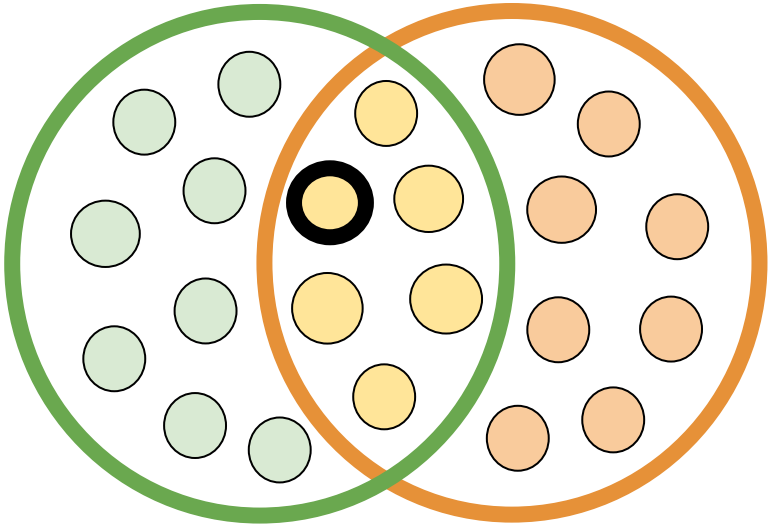


$$d = 1 - \left(\frac{2J}{J+1}\right)^{1/k}$$

decreases similarity
(increases distance)

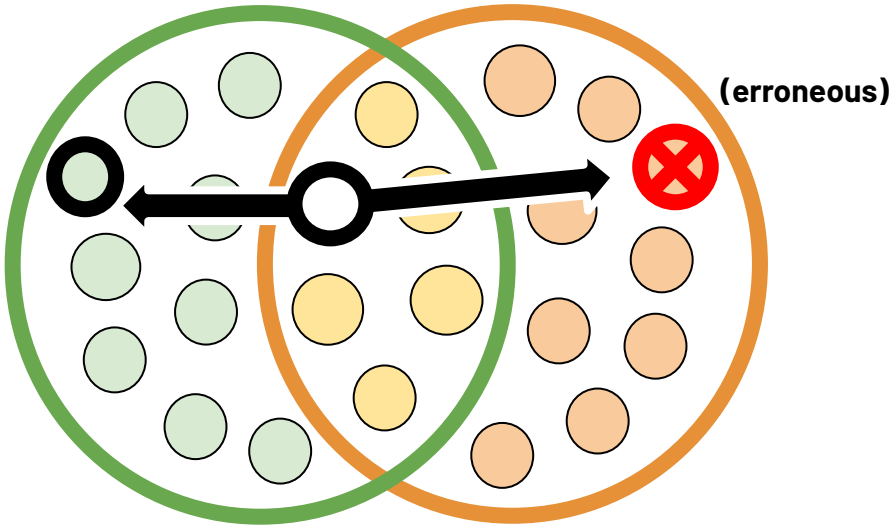
Jaccard Index example

High coverage
Jaccard Index



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

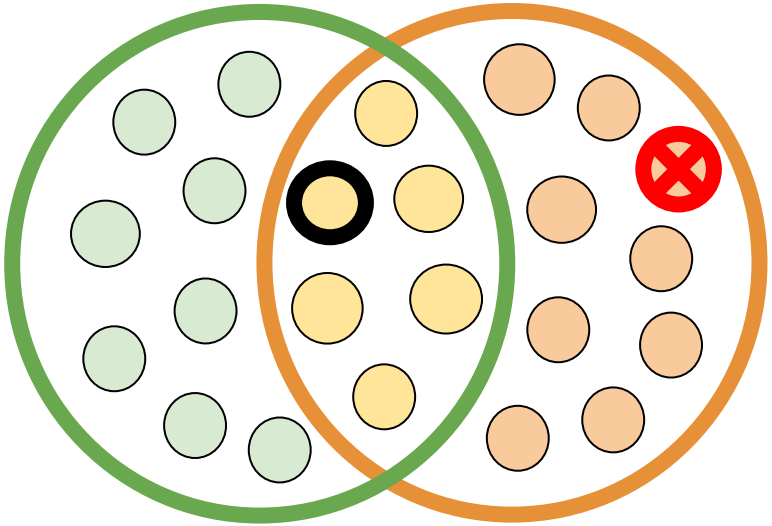
Low coverage
Jaccard Index



decreases similarity
(increases distance)

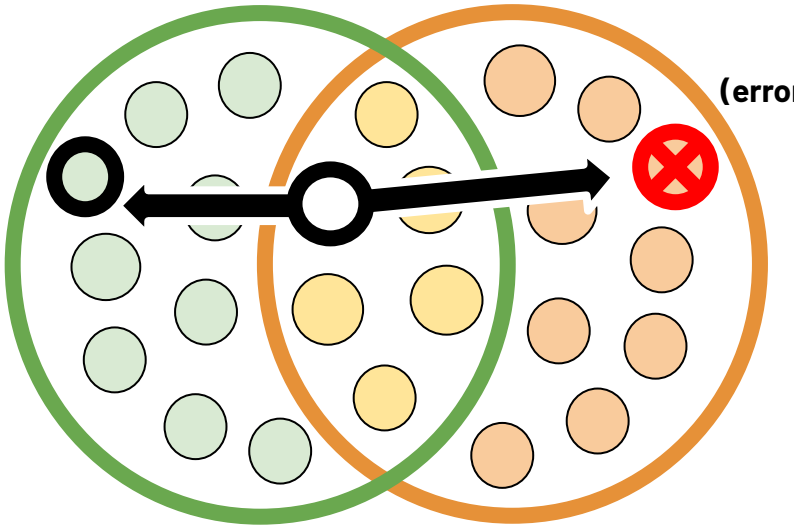
Jaccard Index example

High coverage
Jaccard Index



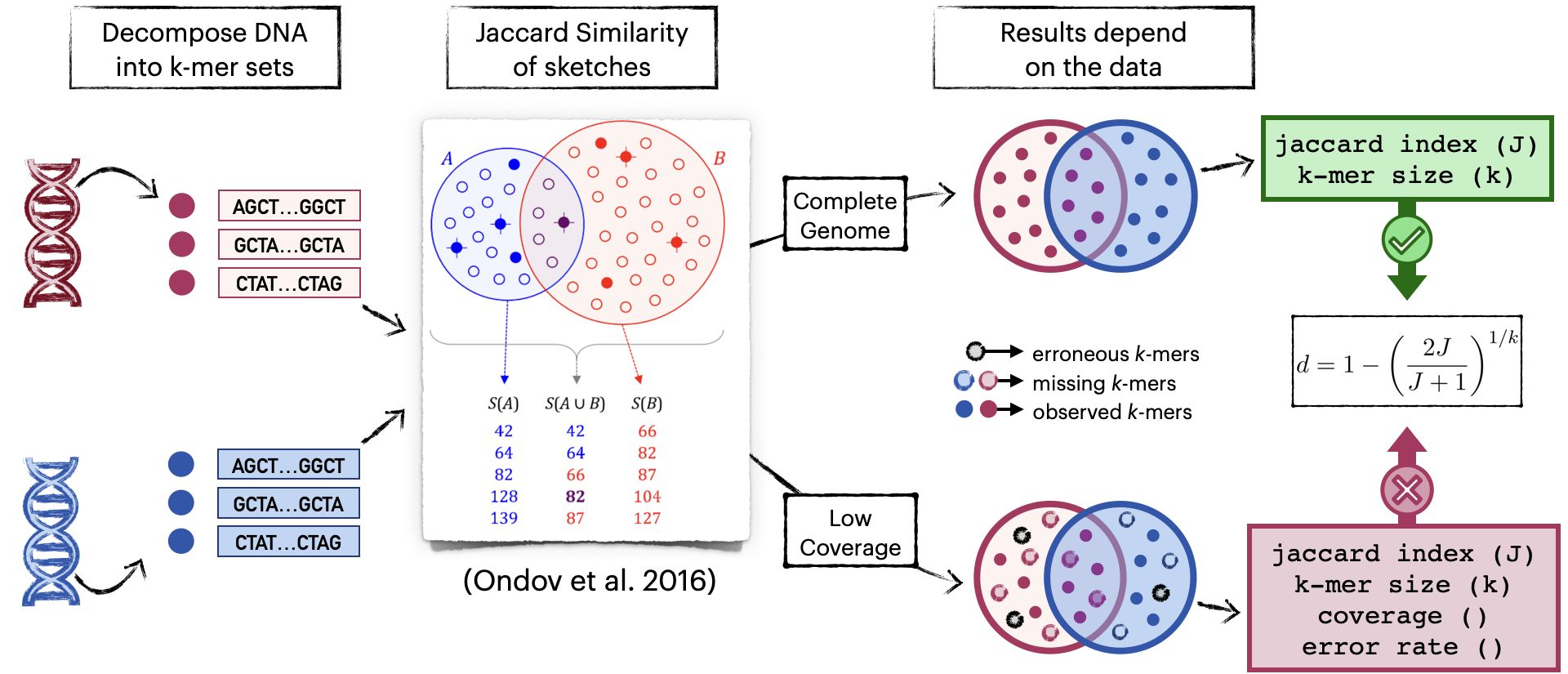
Doesn't high coverage also have error?

Low coverage
Jaccard Index

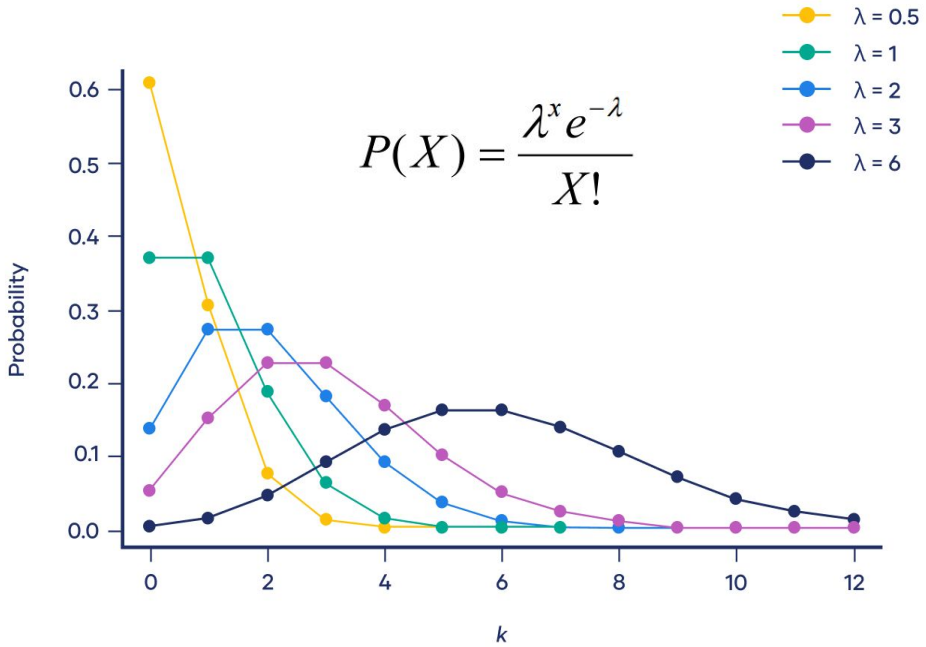


decreases similarity
(increases distance)

Assembly-Free, Alignment-Free Distances

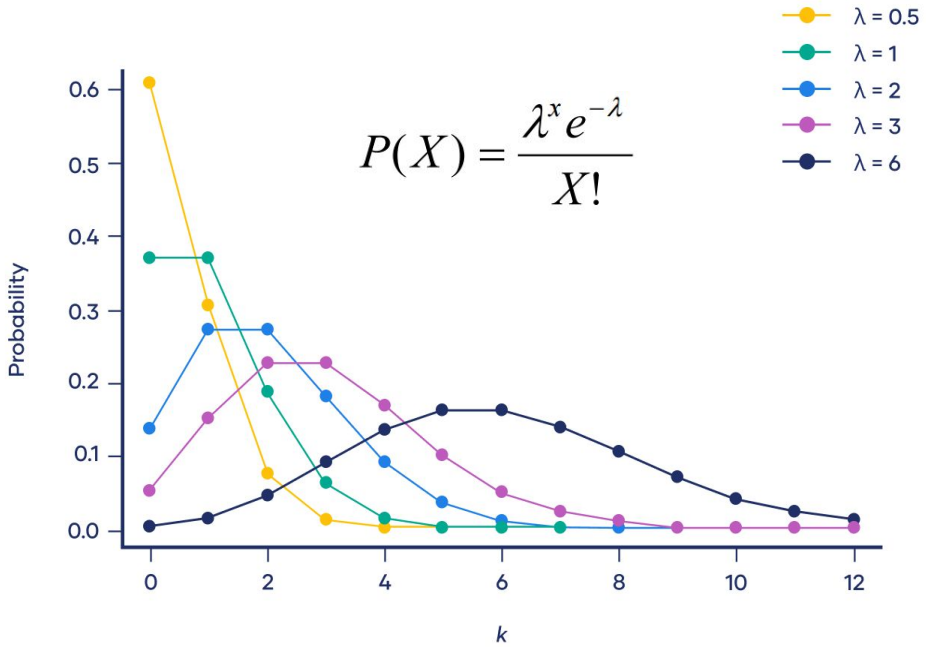


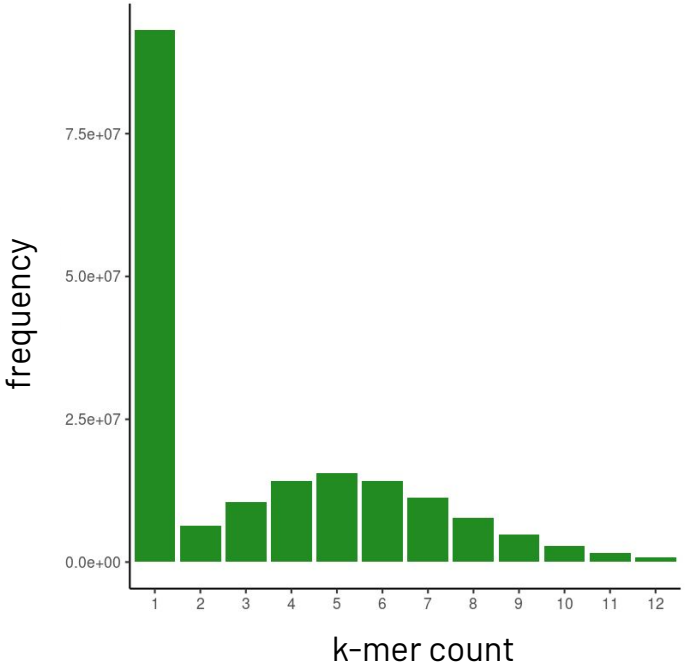
Modeling k-mer frequency



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

Modeling k-mer frequency

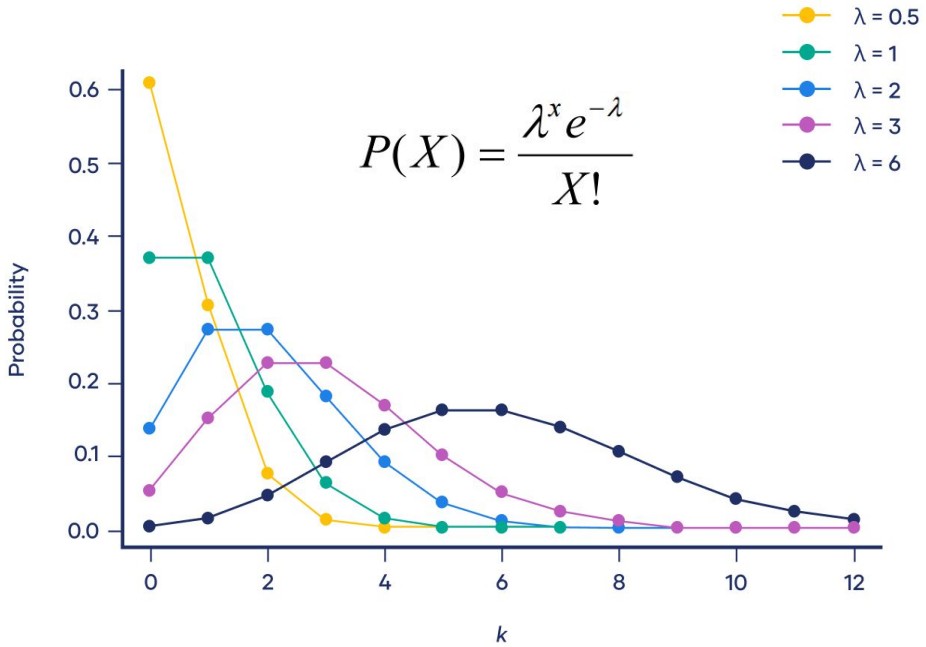




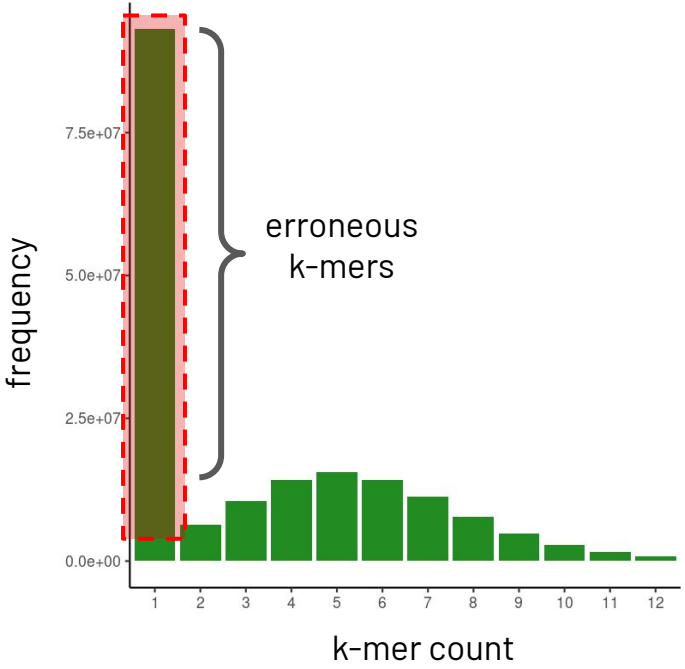
Poisson Distribution:
probability of a number of events occurring in a fixed interval, given an average rate of occurrence (λ)

observed k-mer frequency histogram for a genome skim

Modeling k-mer frequency

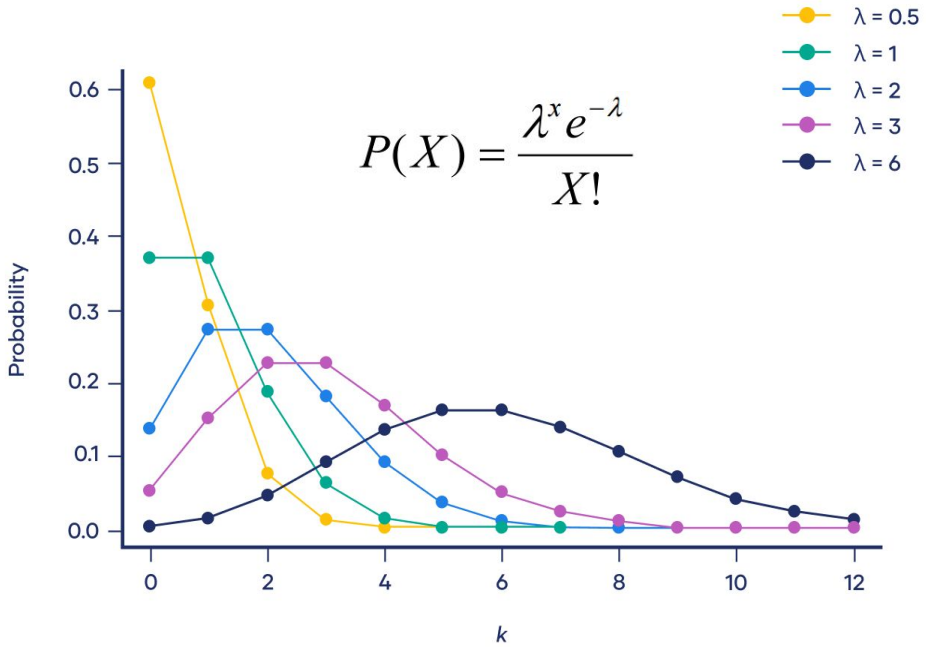


Poisson Distribution:
probability of a number of events occurring in a fixed interval, given an average rate of occurrence (λ)

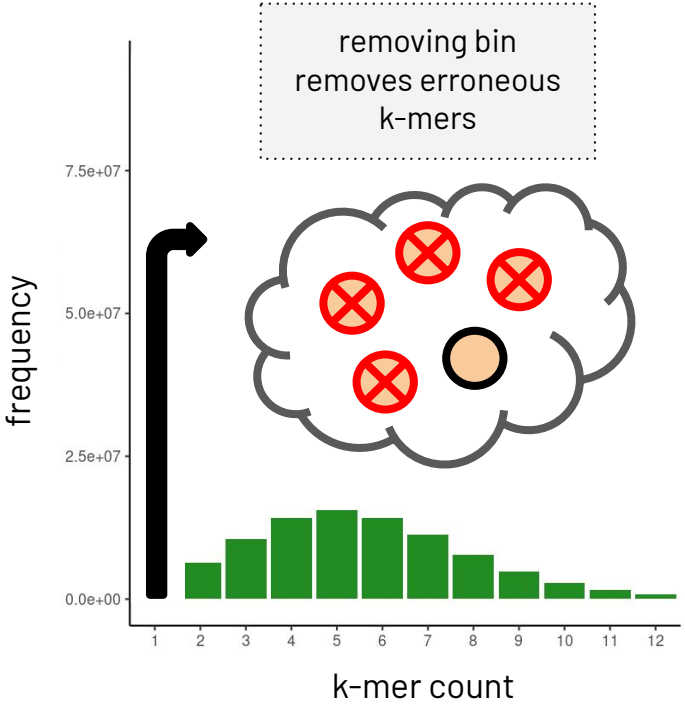


observed k-mer frequency histogram for a genome skim

Modeling k-mer frequency

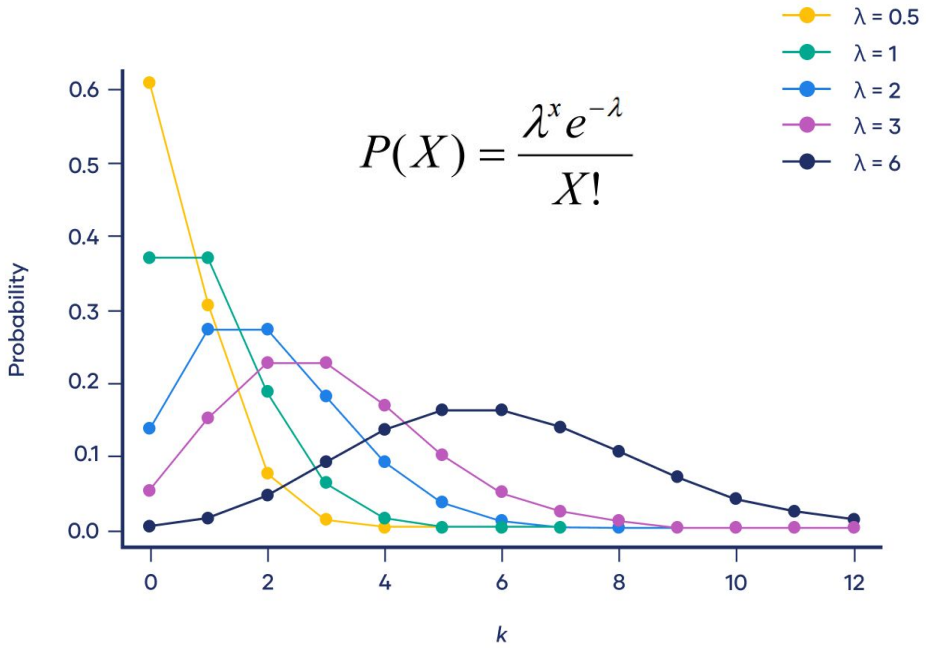


Poisson Distribution:
probability of a number of events occurring in a fixed interval, given an average rate of occurrence (λ)

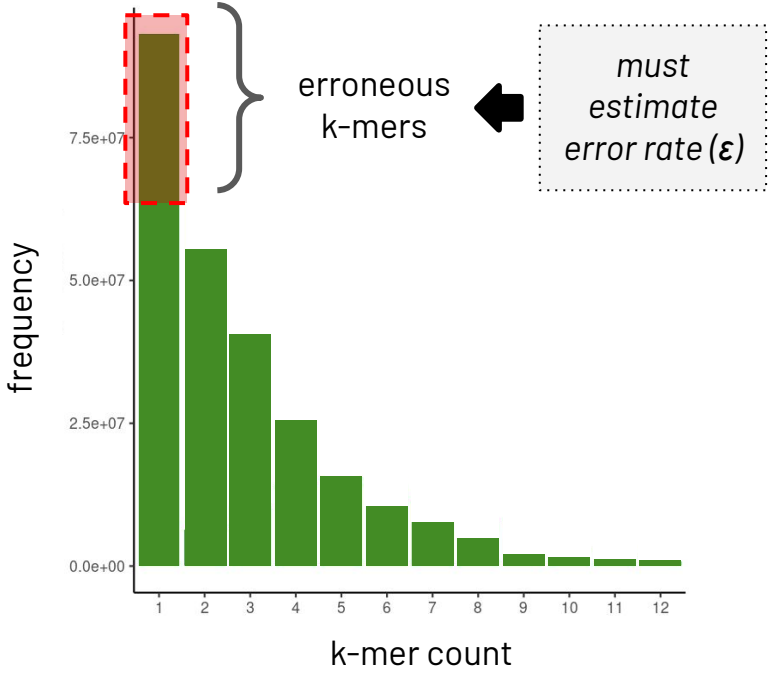


not possible for genome skims!

Modeling k-mer frequency

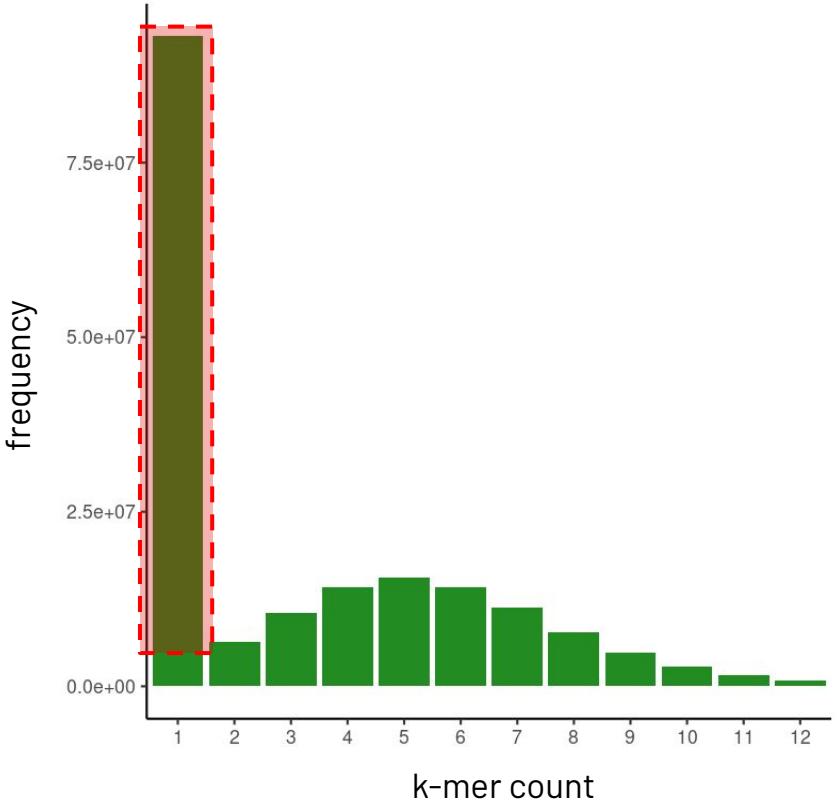


Poisson Distribution:
probability of a number of events occurring in a fixed interval, given an average rate of occurrence (λ)



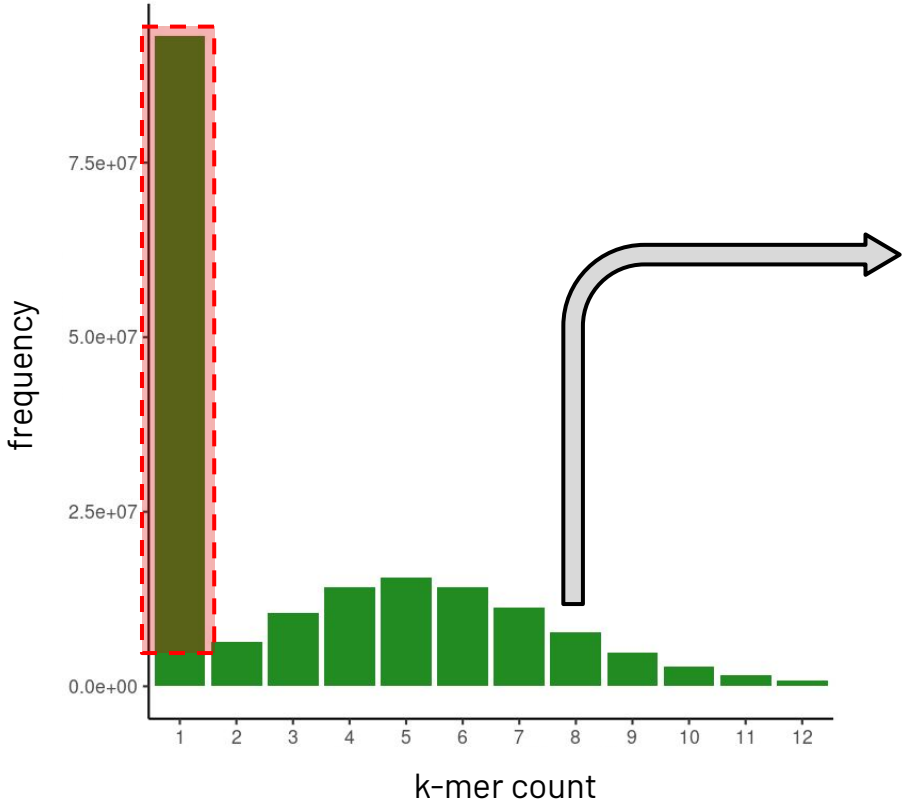
not possible for genome skims!

Skmer equations



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

Skmer equations

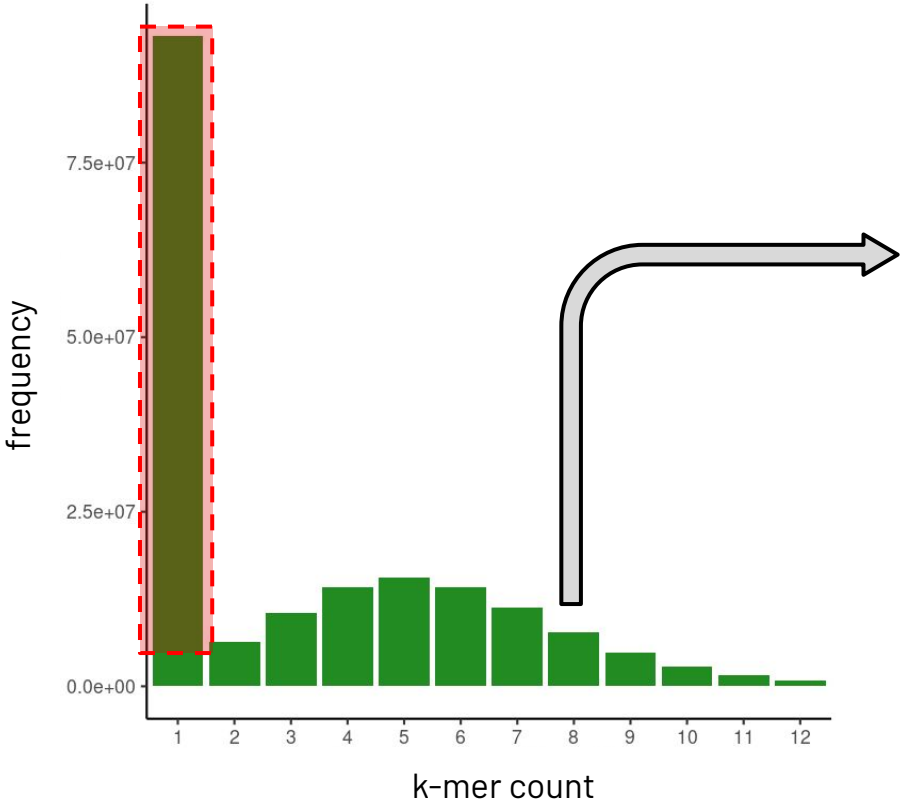


errors in large k-mers will create **novel k-mers**

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

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

Skmer equations



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

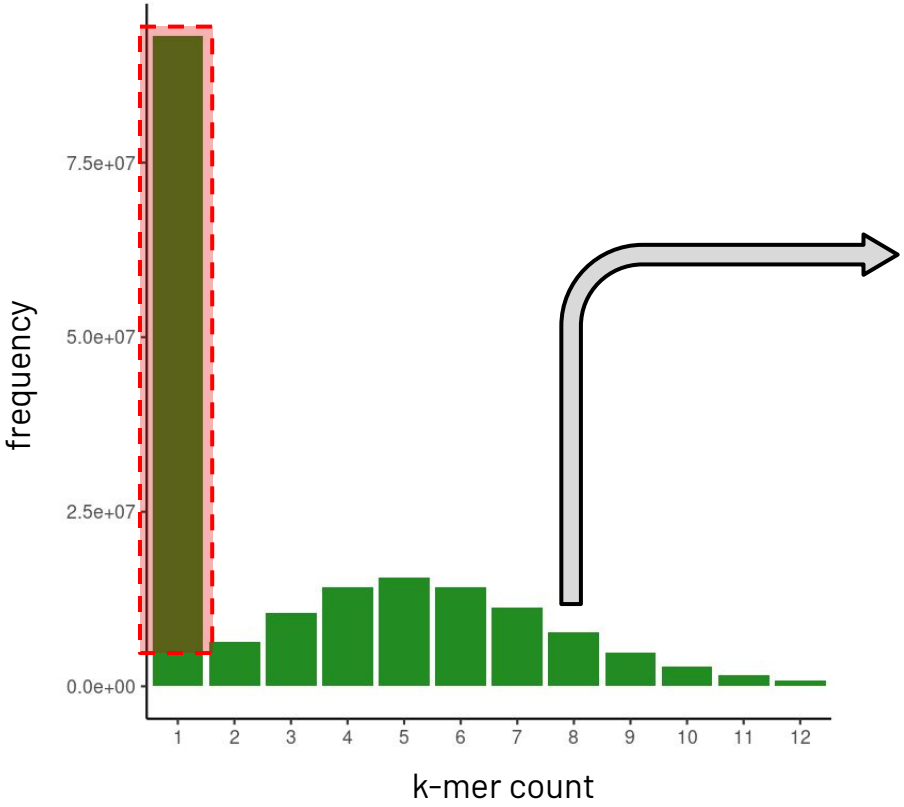
all other bins will follow an
error-free *k*-mer coverage poisson distribution (ξ)

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

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)

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

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** (Ɛ)

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

Distance equations

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

Mash

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

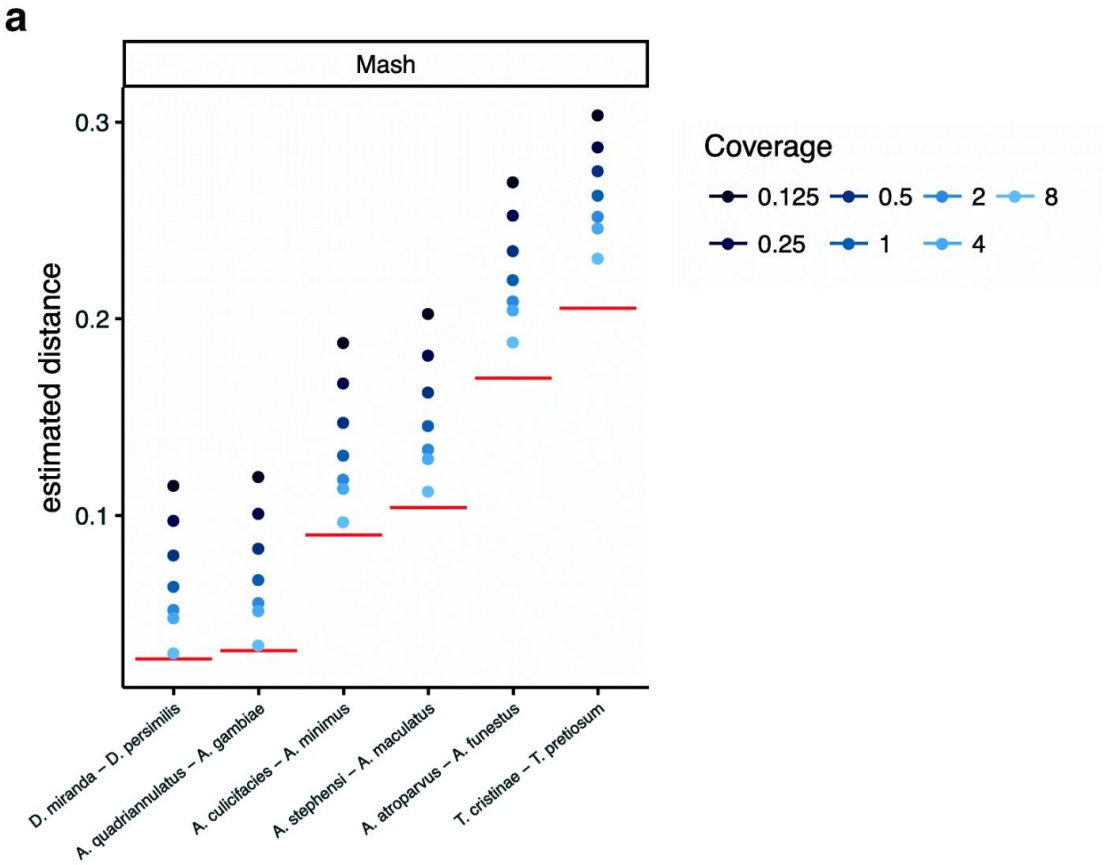
Skmer

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.

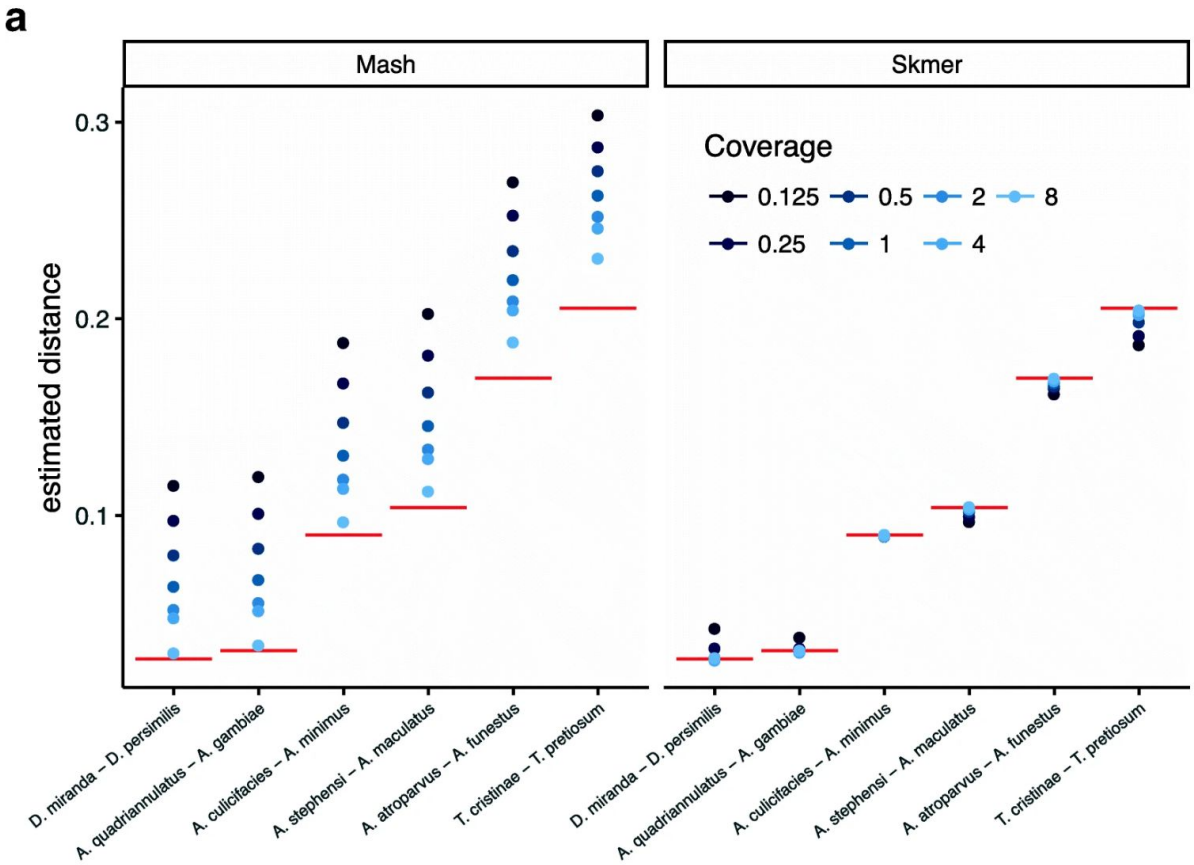
L_i = genome size.

Mash fails with genome skimming data



Low coverage sequencing leads to distance **overestimation** with mash

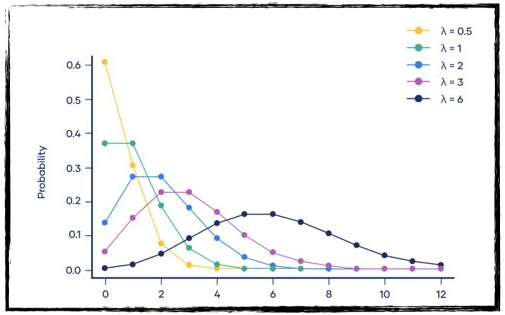
Mash fails with genome skimming data



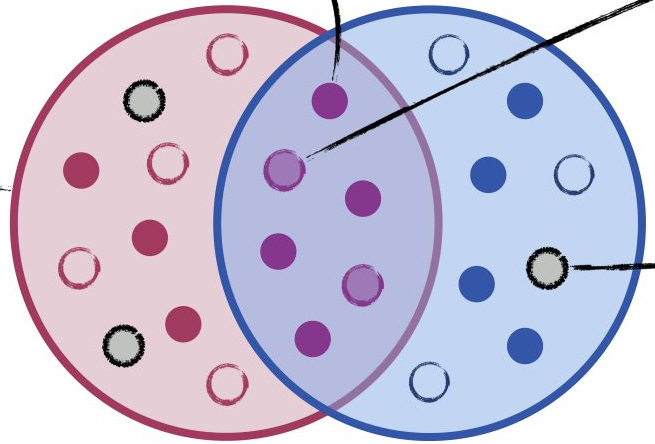
Skmer extends alignment-free, reference-free distance estimation to genome skims

Skmer Uses Simplifying Assumptions

1. Genome is mostly unique and haploid

2. 

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



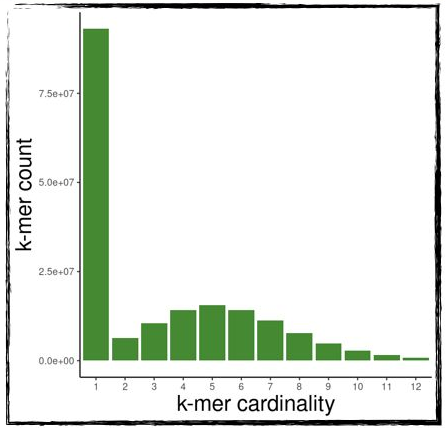
erroneous *k*-mers

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**



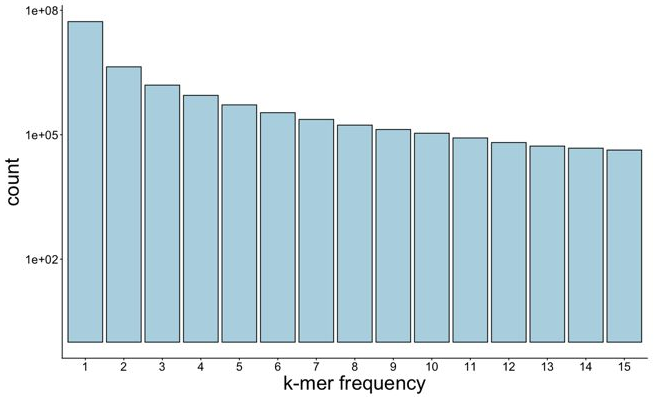
Genome Repeat Spectrum

K-mer based definition of repetitiveness

There are many kinds of repetitive DNA sequences

- Centromeres
- Gene Duplication
- Tandem Repeats
- Telomeres
- Transposons

Summarized with k-mer Repeat Spectrum



Uniqueness Ratio =

R1/ Genome Length

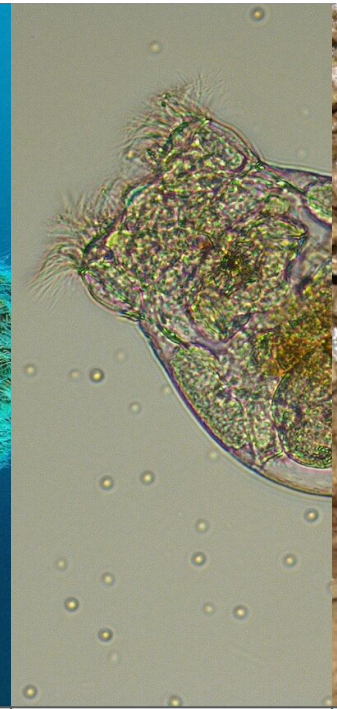
Genomes are not unique

S. moellendorffii

C. crispus

O. edulis

N. nomurai



B. plicatilis

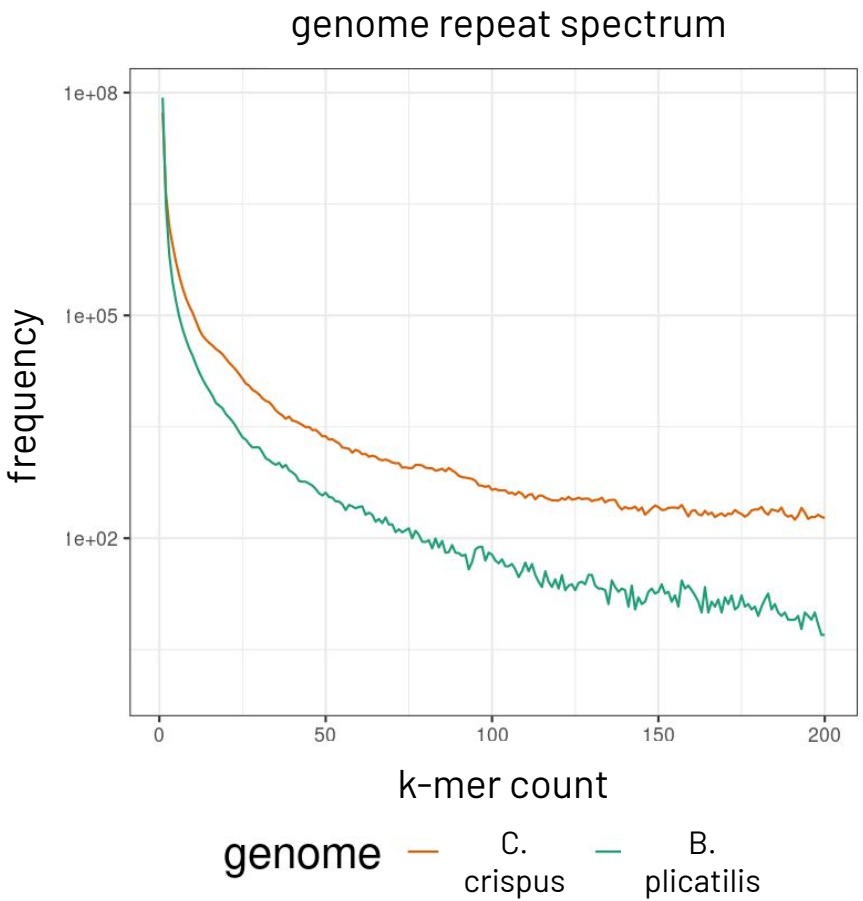
H. medicinalis

Low uniqueness

Medium uniqueness

High uniqueness

Genomes are not unique



Genomes come with a wide range of ***Uniqueness Ratios:***

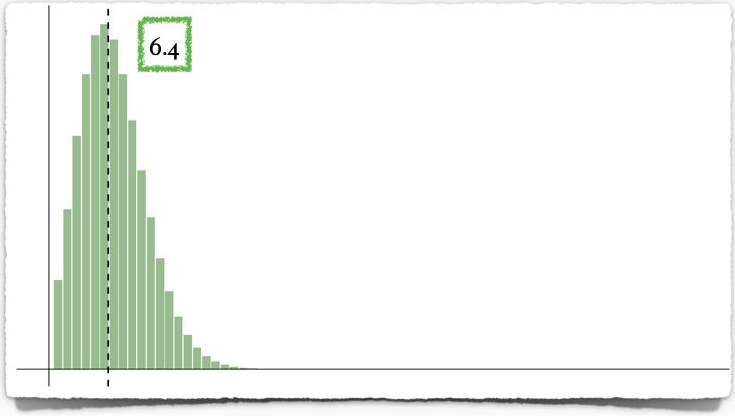
$$R_1 / L$$

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

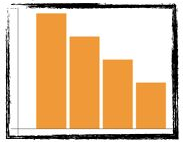
Repeats Violate Skmer Assumptions

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

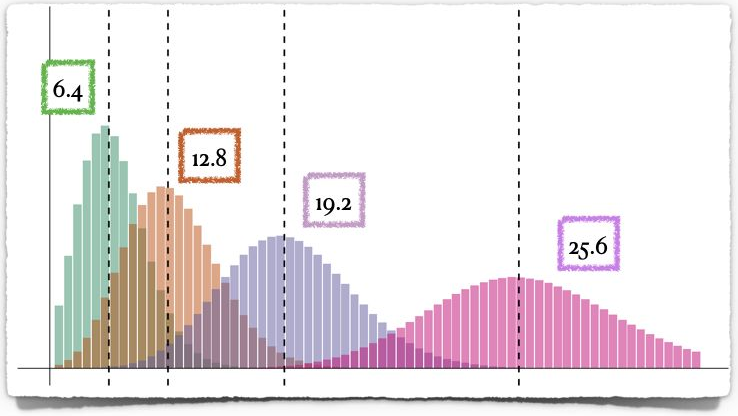
k-mer coverage = 6.4



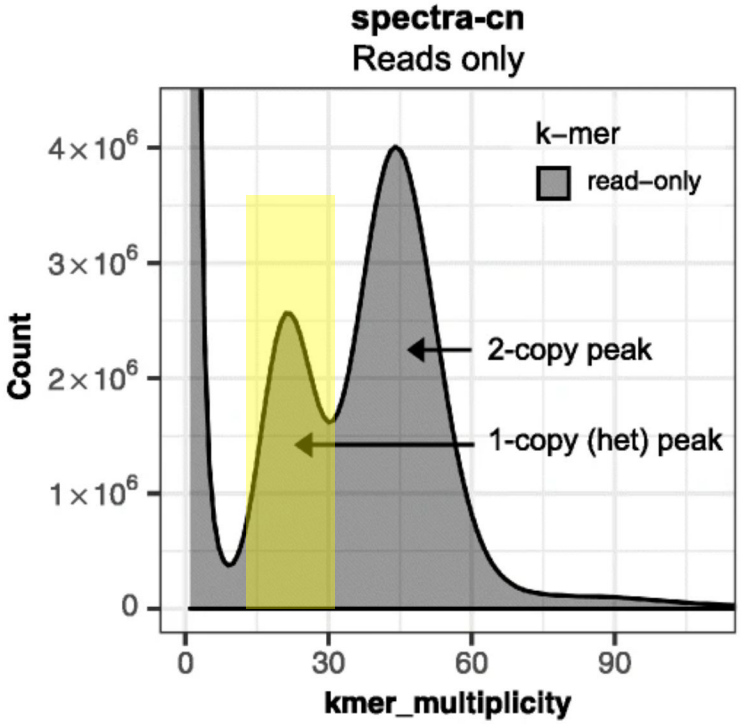
Skim k-mer frequency under genome with repeats



k-mer coverage = 6.4



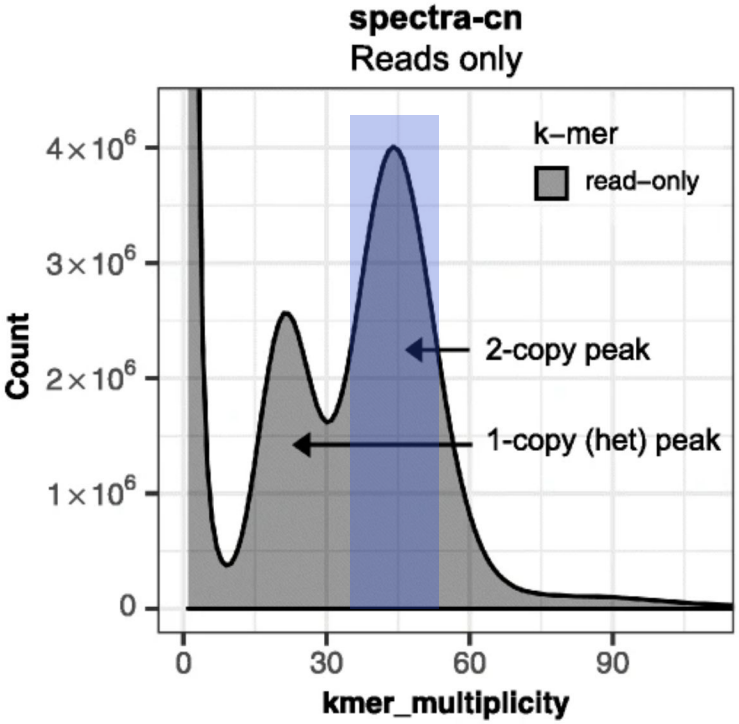
Repeats break Skmer assumptions



(Merqury, 2020)

collection of *k*-mers only occur once

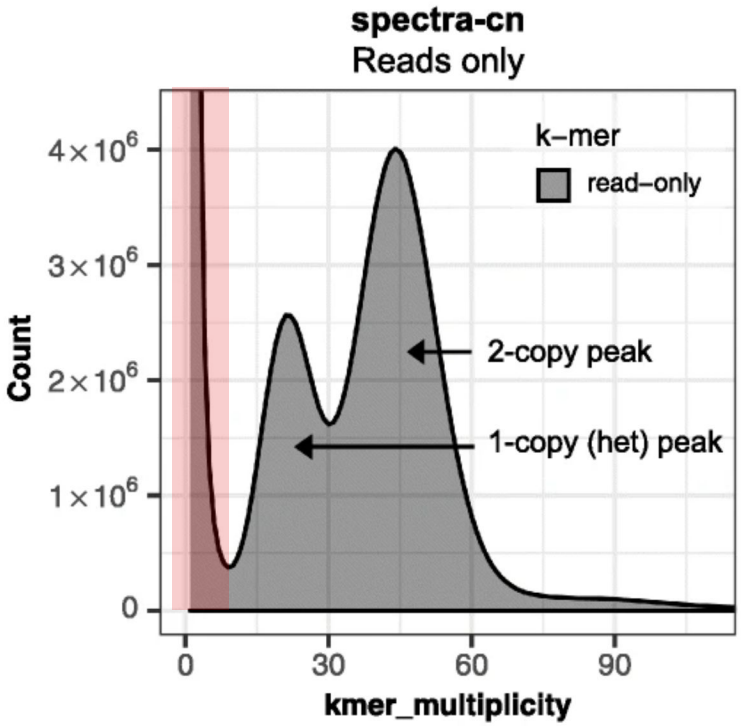
Repeats break Skmer assumptions



(Merqury, 2020)

collection of *k*-mers occur twice

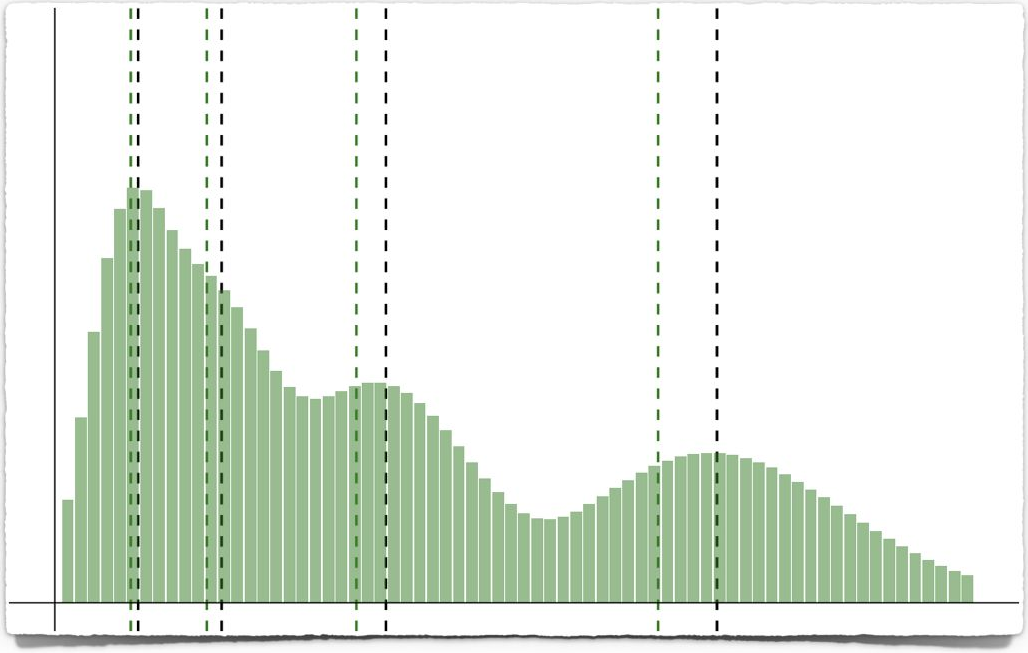
Repeats break Skmer assumptions



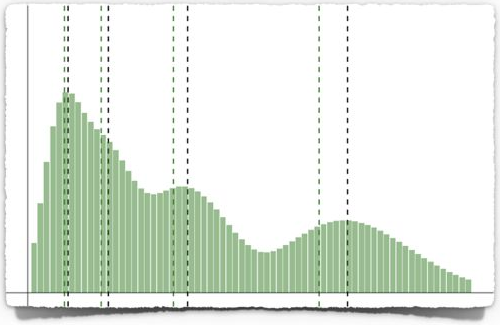
(Merqury, 2020)

erroneous k-mers

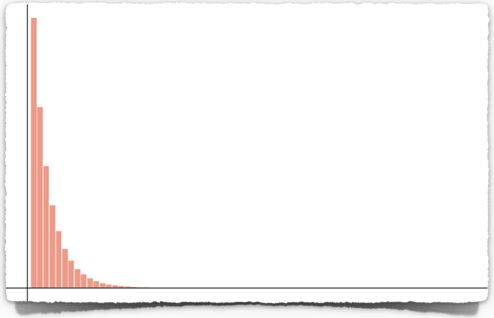
K-mer Frequency Histogram with Error



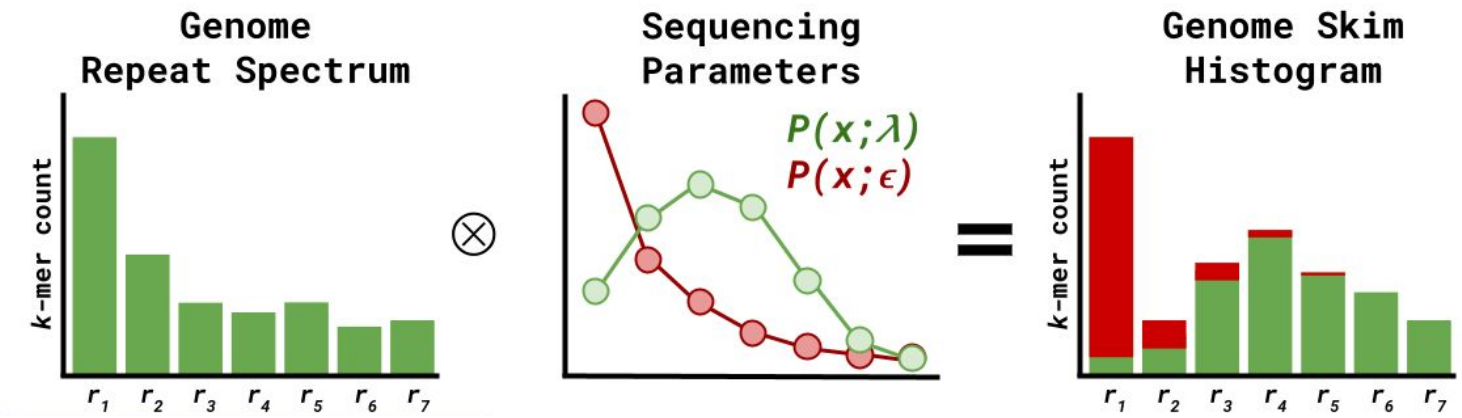
xi



eps

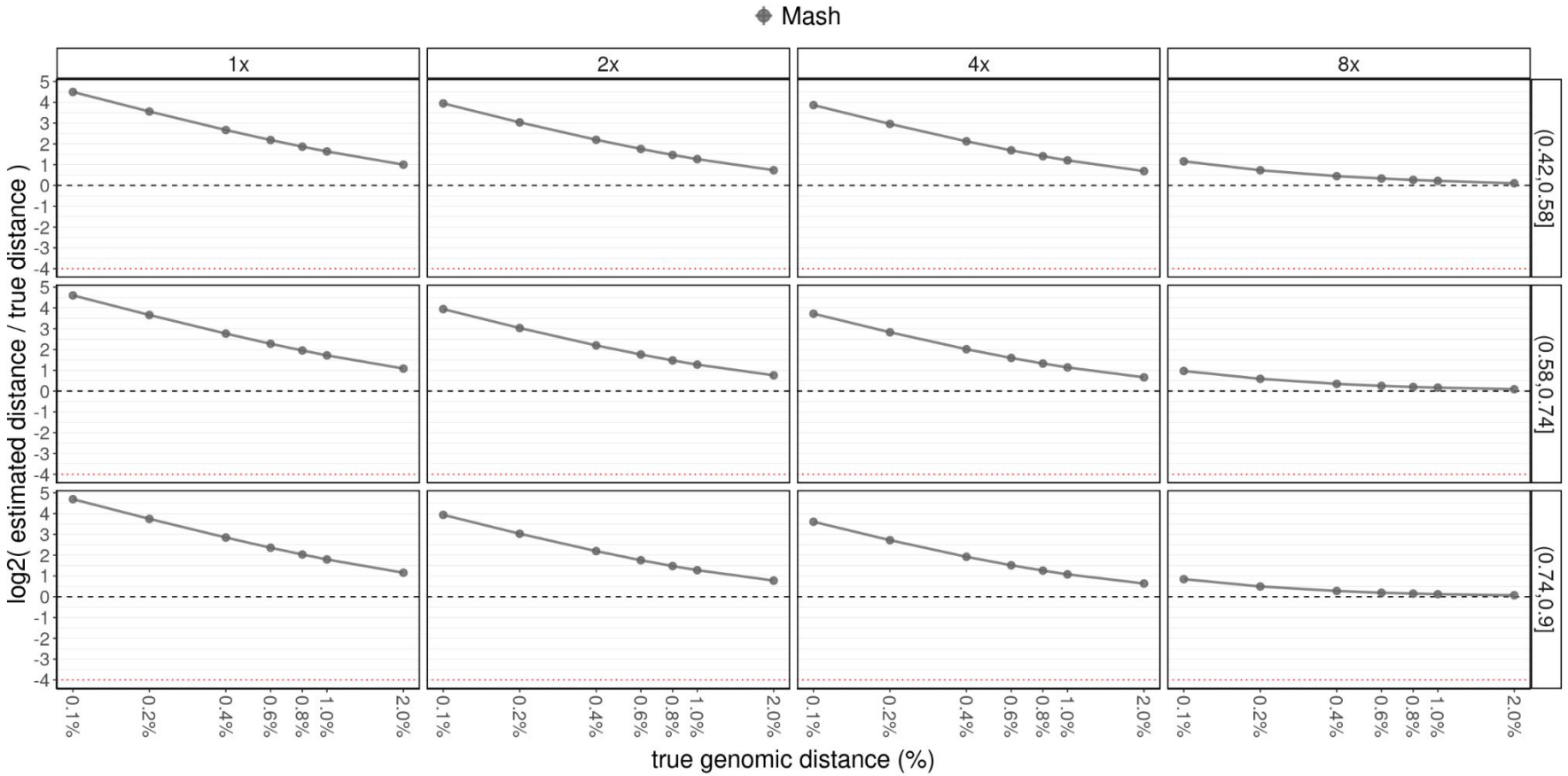


ReSkmer: estimating sequencing parameters

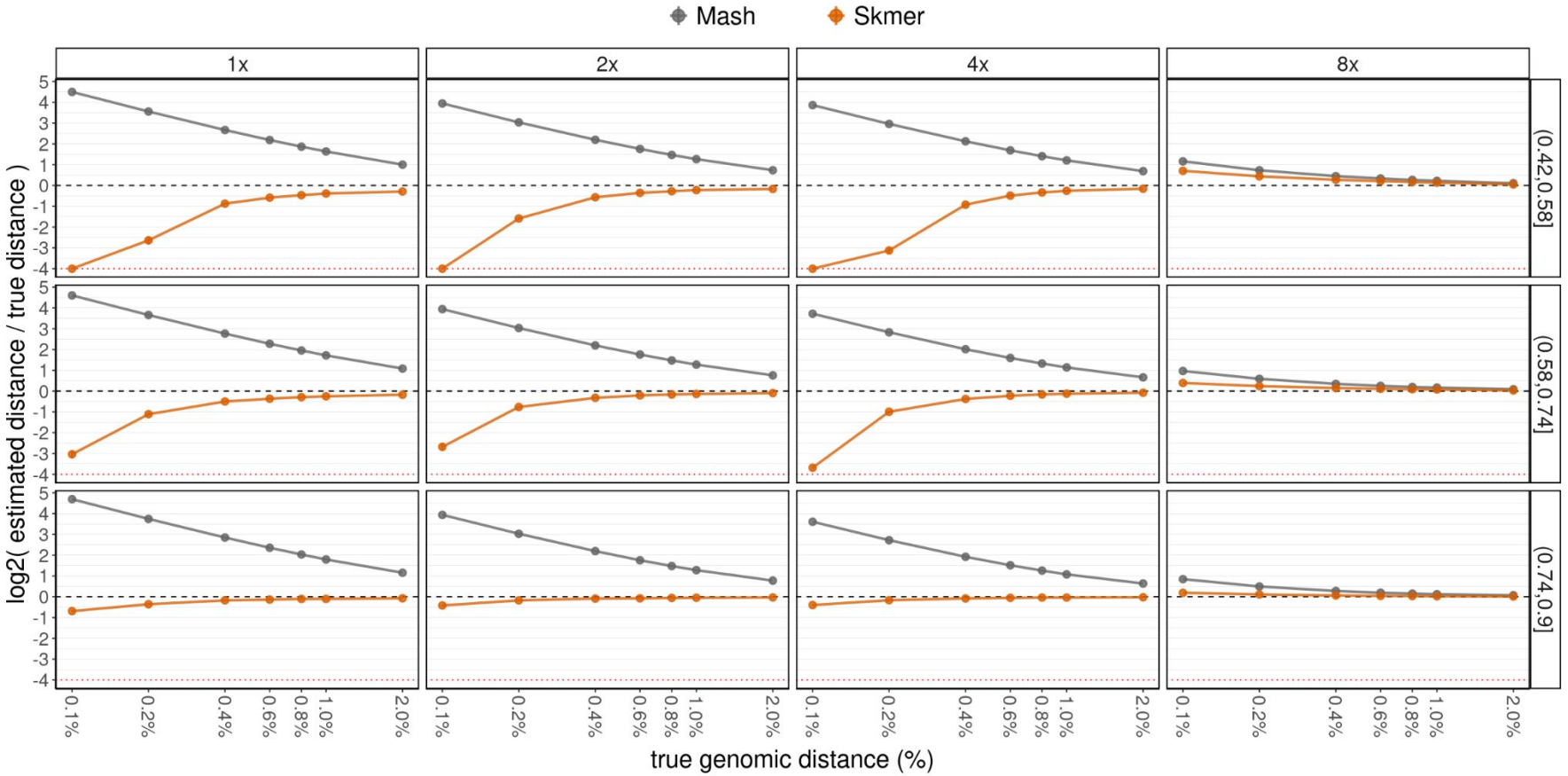


Each genome skim is a **convolution** between **sequencing parameter** distributions and it's **repeat spectrum**

Mash Results



Skmer results



Methods

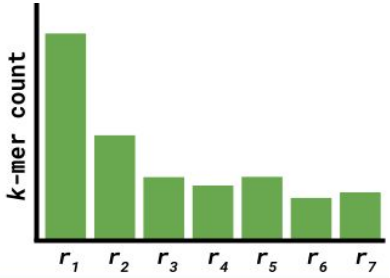
ReSkmer
(k-mer based method, short-read sequencing)

Advantages

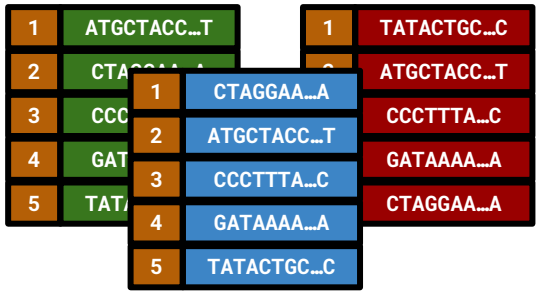
- Whole Genome Resolution
- Reference and Alignment Free

Disadvantages

- **Low Coverage**
- **Models repeats**

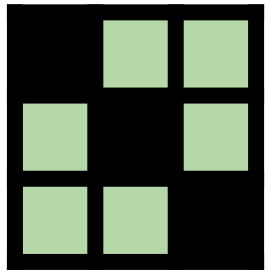


genome repeat spectrum



fastq per sample

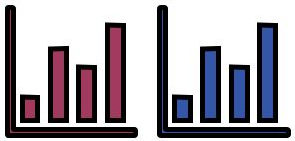
Input



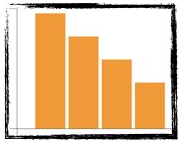
distance matrix

Output

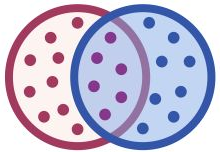
Correcting for Repetitiveness



Skimming k-mer Histograms

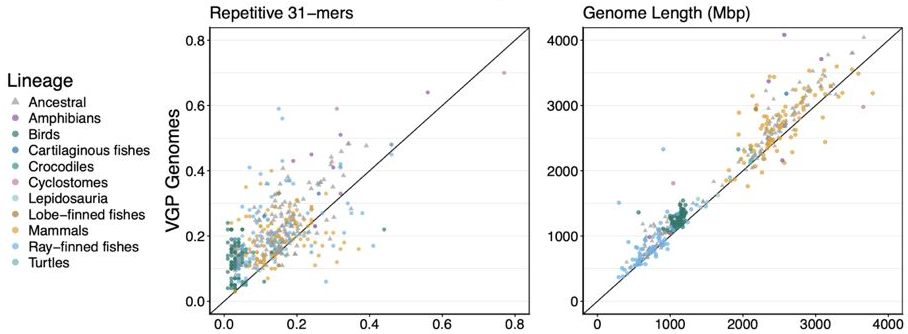


Genome Repeat Spectrum



Jaccard Index

Distance



Newer assemblies can capture repetitiveness

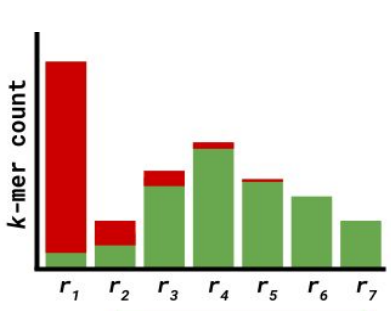
Estimating repeat spectra and genome length from low-coverage genome skims with RESPECT

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

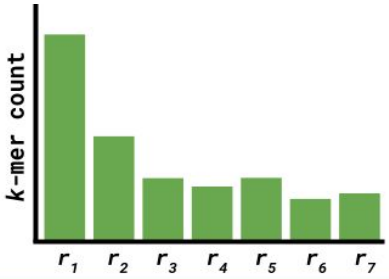
If no available genome, spectrum can be estimated.

Using repeat spectrum to estimate distance

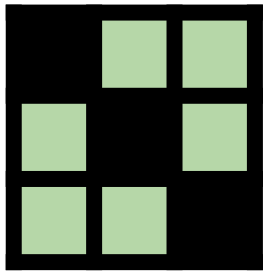
For 'short' distances **(0.001 ~ 0.05)**



genome skim



genome repeat spectra



Accurate Distance Matrix

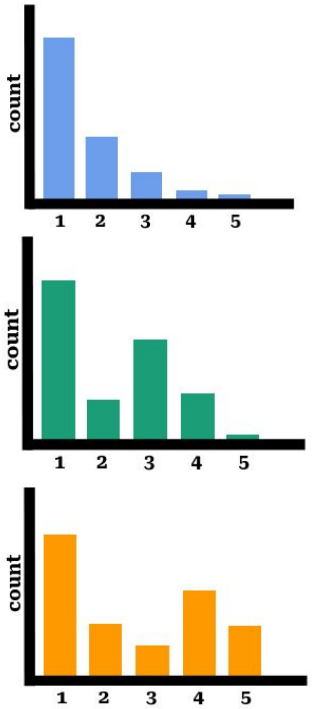
get an idea of the k-mer counts of the original genome

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

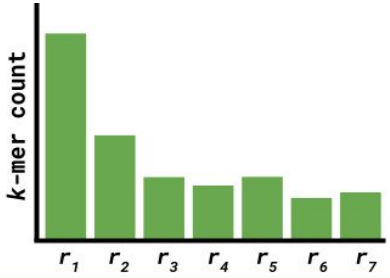
1. **High Quality Assembly** 2. **Repeat Spectra Estimate** (via Respect)



RESPECT

Repeat Spectra and **Genome Length** Estimation

RESPECT

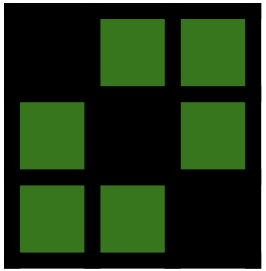


genome repeat spectrum

ReSkmer

Use corresponding spectra to **estimate sequencing parameters** and **distance matrix**

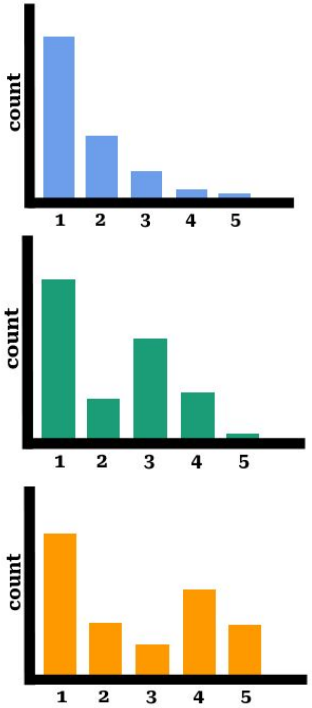
ReSkmer



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

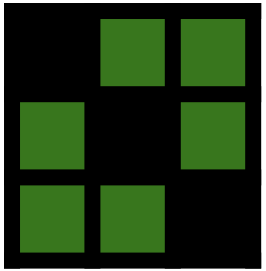
sample	r1	r2	...
skim_1	5016990	902682	...
skim_2	4906813	873642	...
skim_3	5206770	892112	...

set of repeat spectra estimates

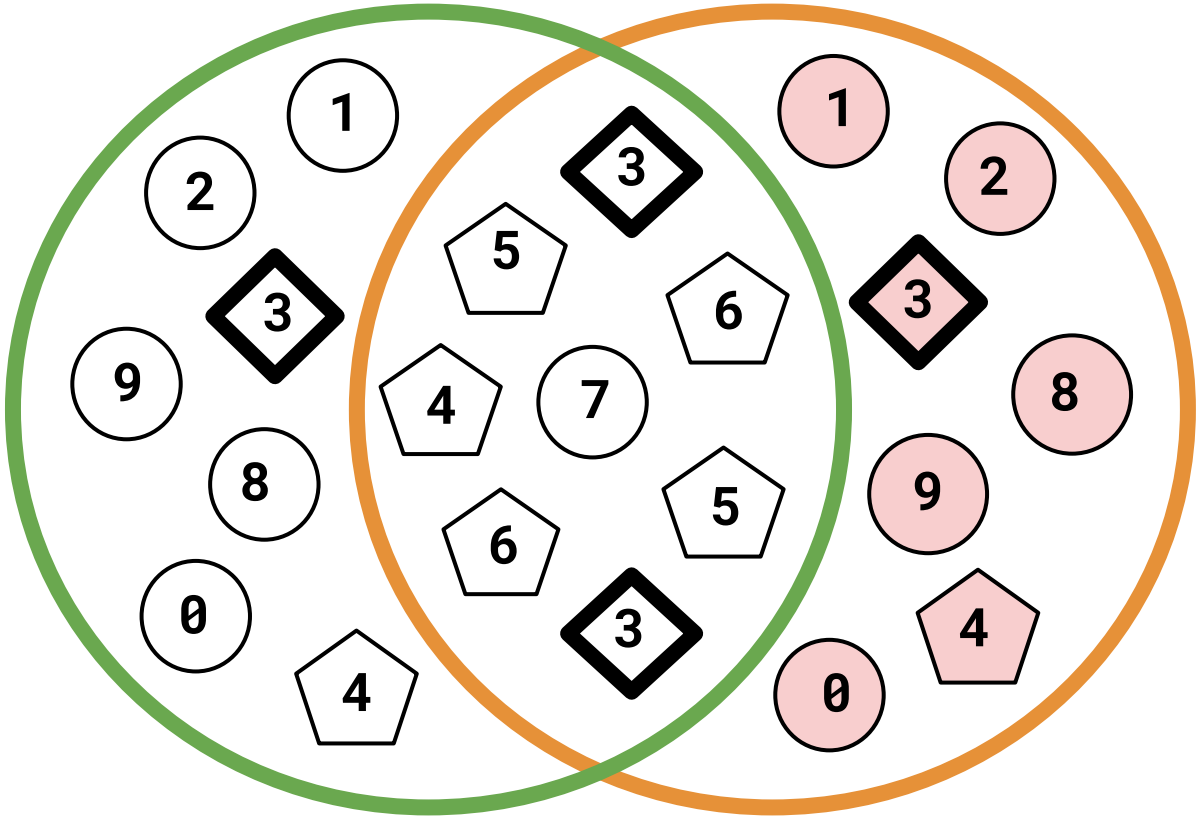
ReSkmer

Use corresponding spectra to **estimate sequencing parameters** and **distance matrix**

ReSkmer

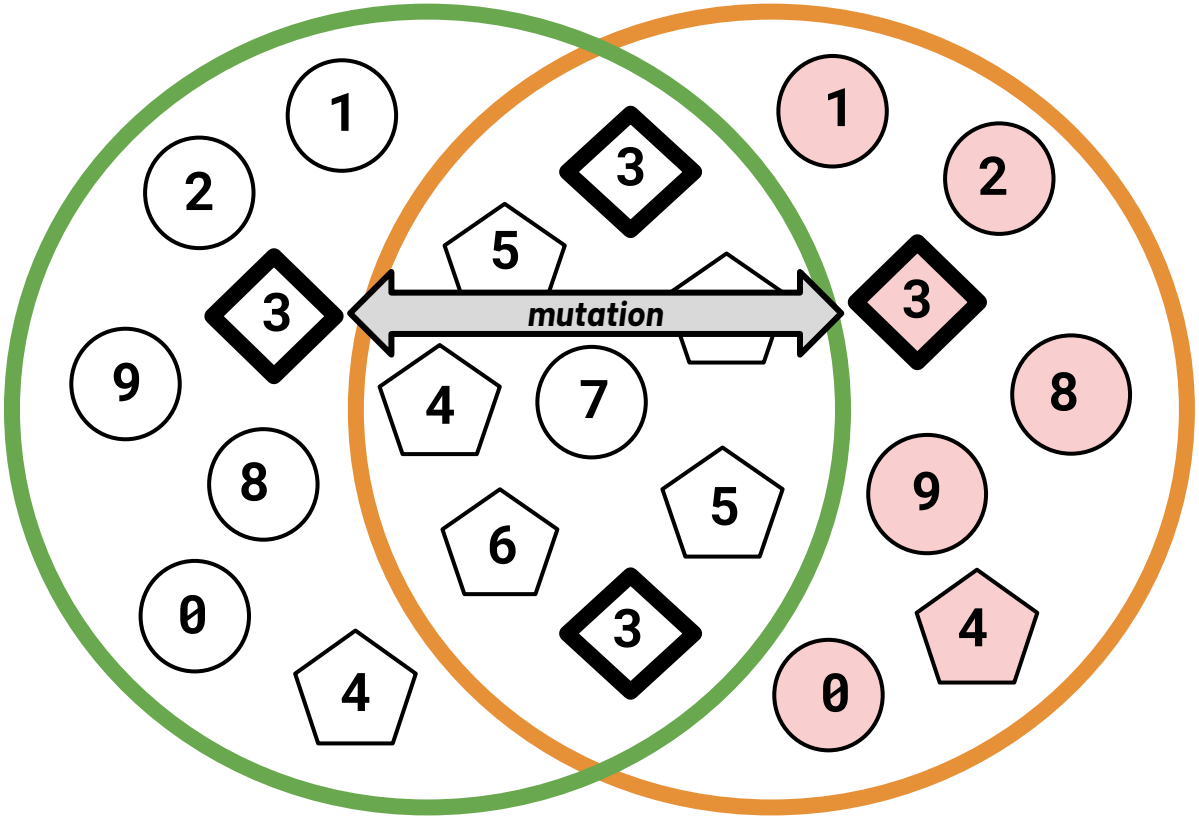


How repeats affect the intersection

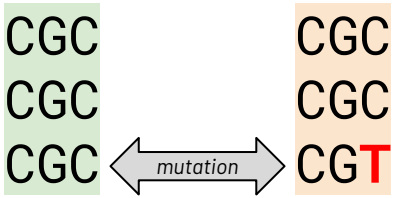


intersection between the sets of *k*-mers

How repeats affect the intersection

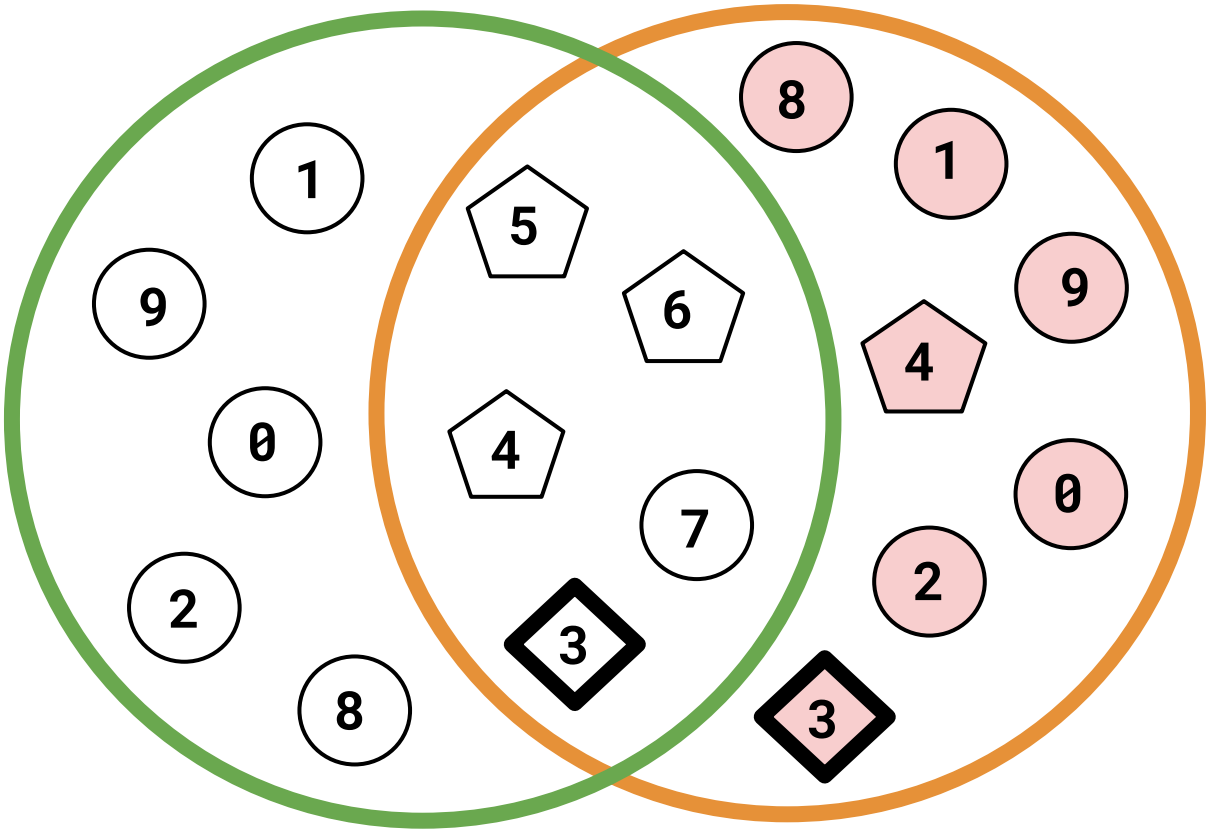


intersection between the sets of *k*-mers

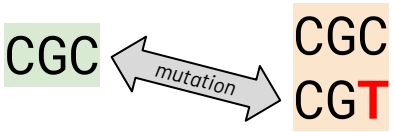


$$\mathbb{E}[I_{\text{multiset}}] = f(d)$$
$$I_{\text{multiset}} = 8$$

How repeats affect the intersection

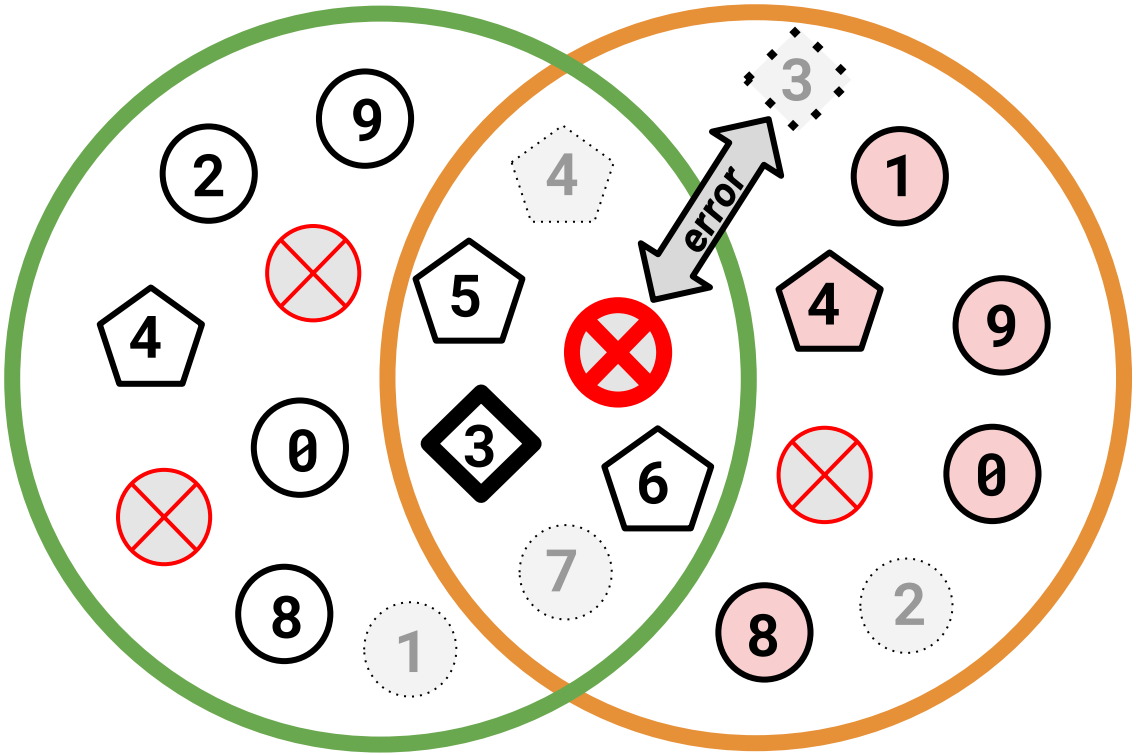


constraints of **sets** force us to remove information about occurrence

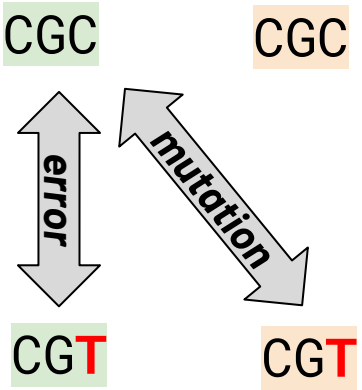


$$\mathbb{E}[I_{\text{set}}] = f(d, R)$$
$$I_{\text{set}} = 5$$

How repeats affect the intersection



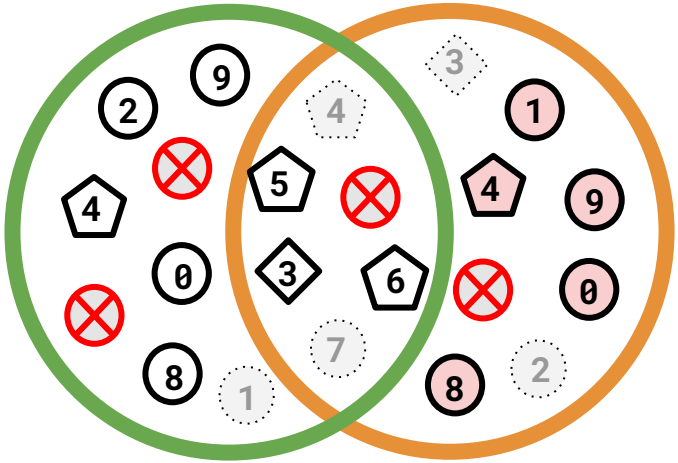
repeats allow for collisions between mutations and errors



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

$$I = 4$$

How repeats affect the intersection



I = I_0 + I_1

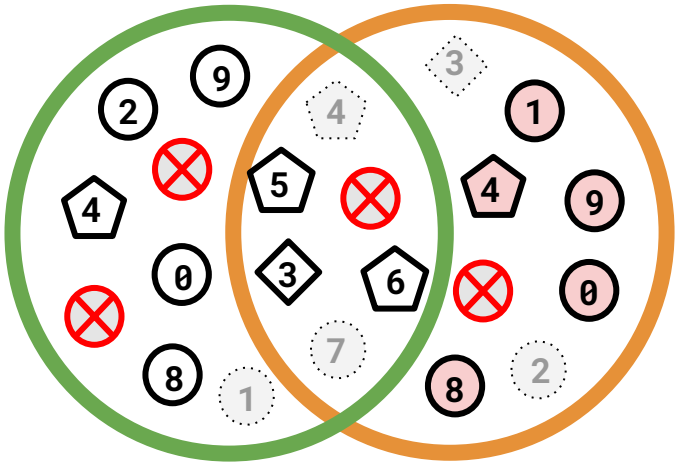
intersection is comprised of two different sets of kmers

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

are not mutated in the evolution from parent to child

in intersection due to error

How repeats affect the intersection



1. coverage
2. genome length
3. error rate

replace **ratio-based** estimators with **spectrum-based** estimators

I = I_0 + I_1

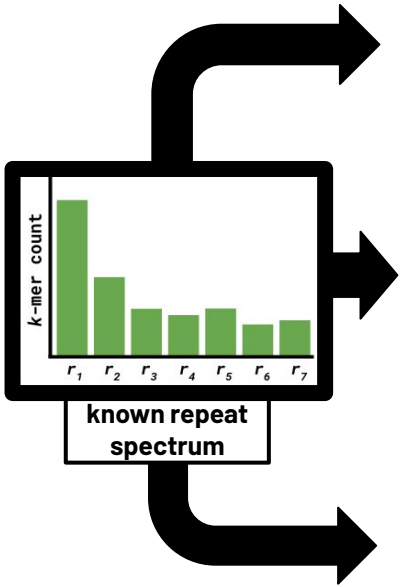
intersection is comprised of two different sets of kmers

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

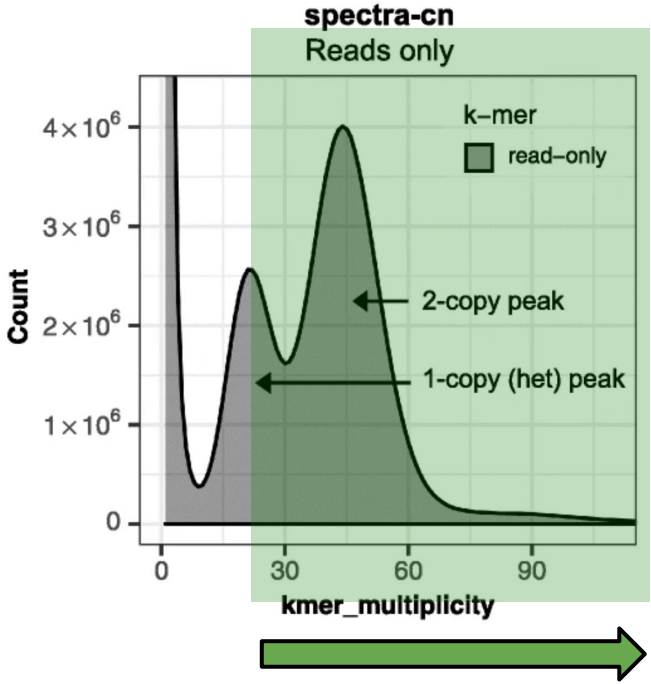
are not mutated in the evolution from parent to child

in intersection due to error

ReSkmer: estimating sequencing parameters

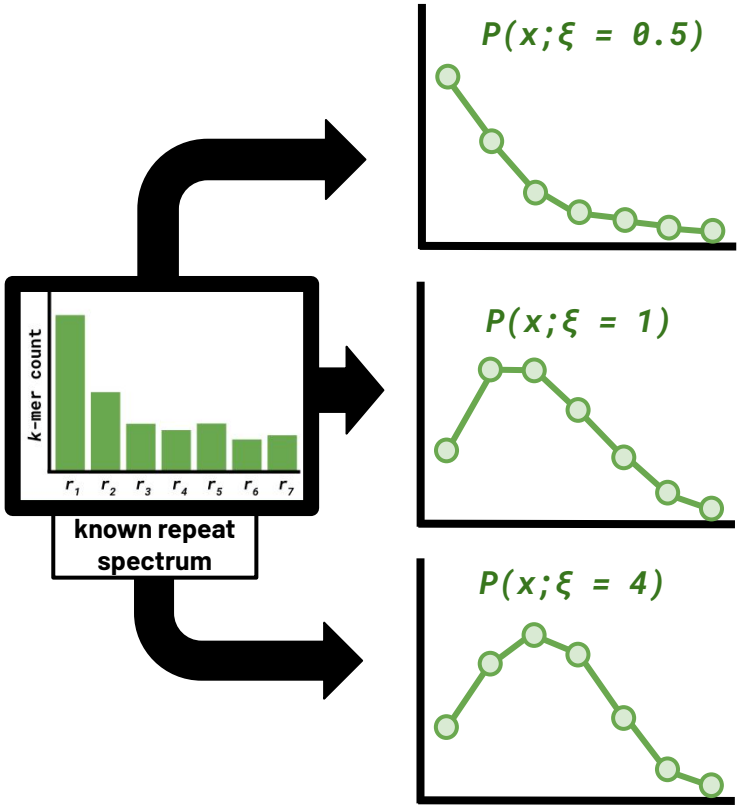


(ξ) Error Free k-mer coverage

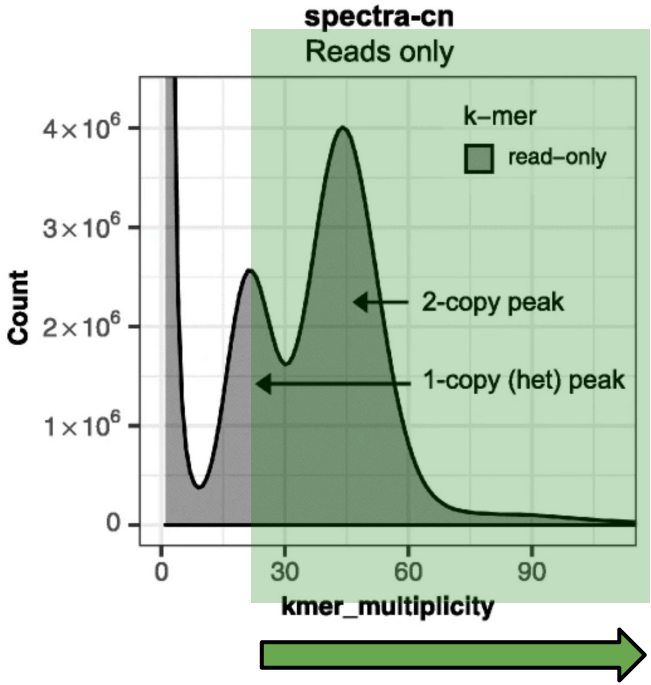


choose range of bins with no error

ReSkmer: estimating sequencing parameters

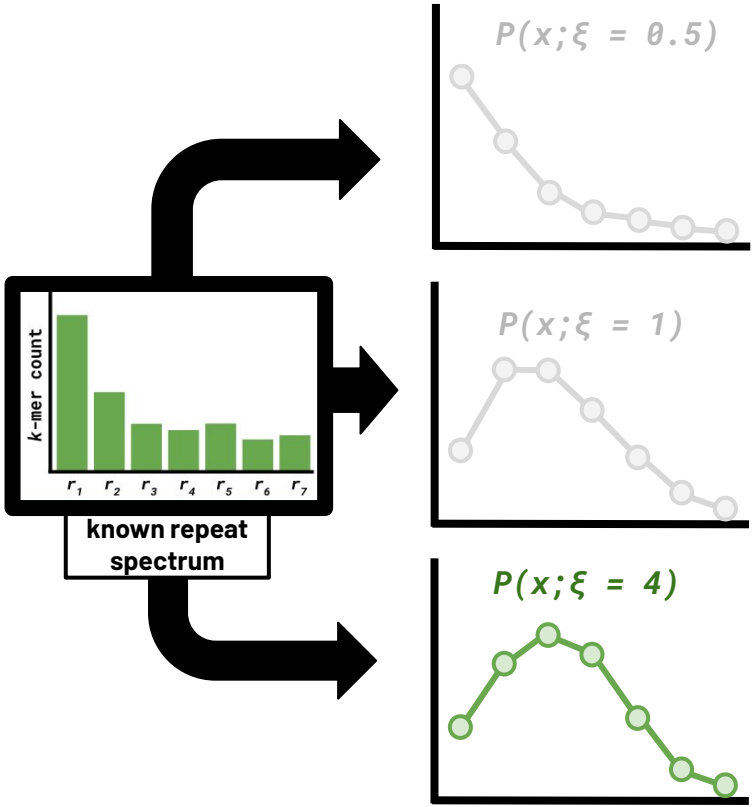


Iterate over potential coverage until match



choose range of bins with no error

ReSkmer: estimating sequencing parameters



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

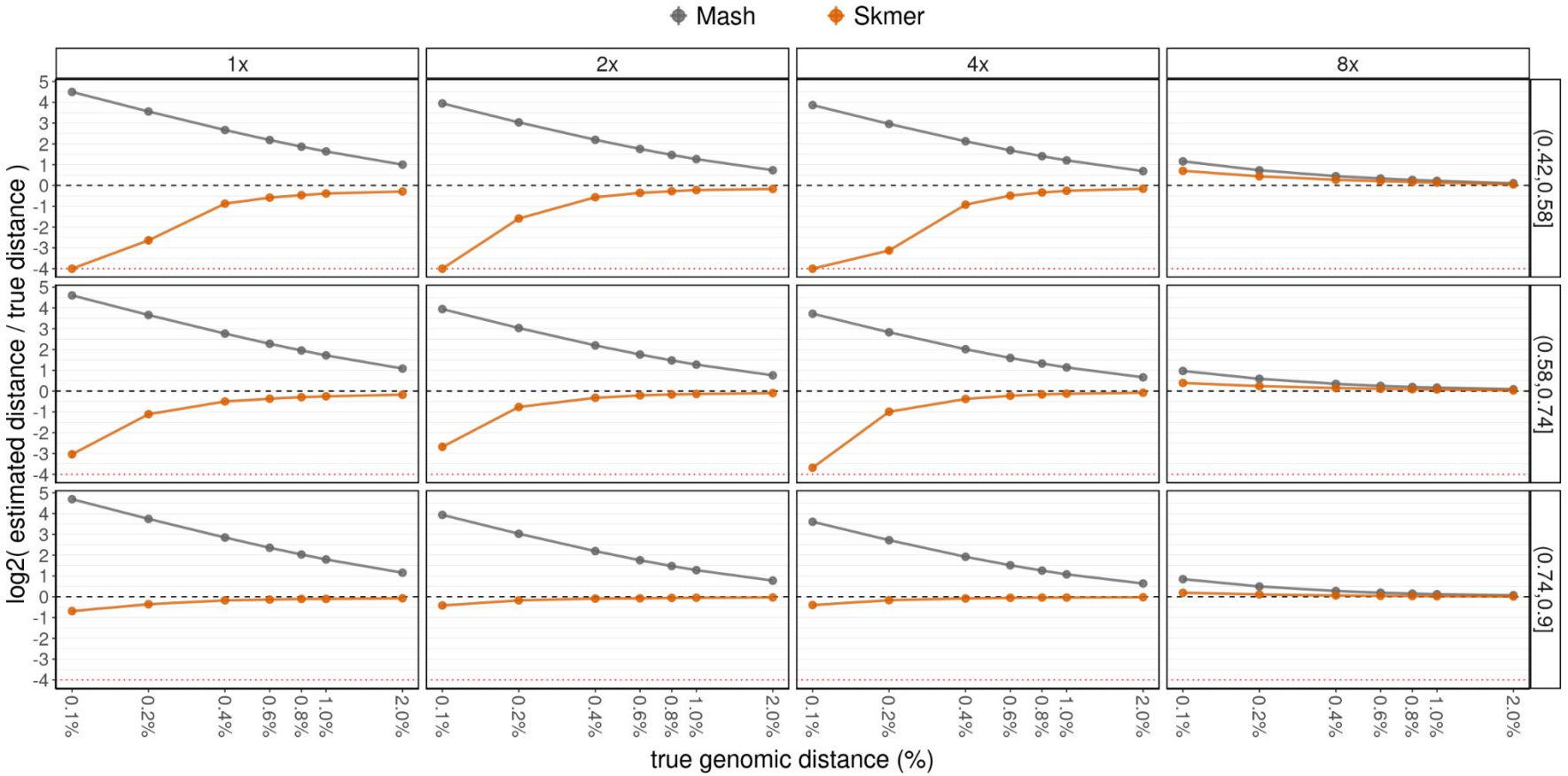
$$\lambda = \frac{\sum_h h o_h}{L}$$

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

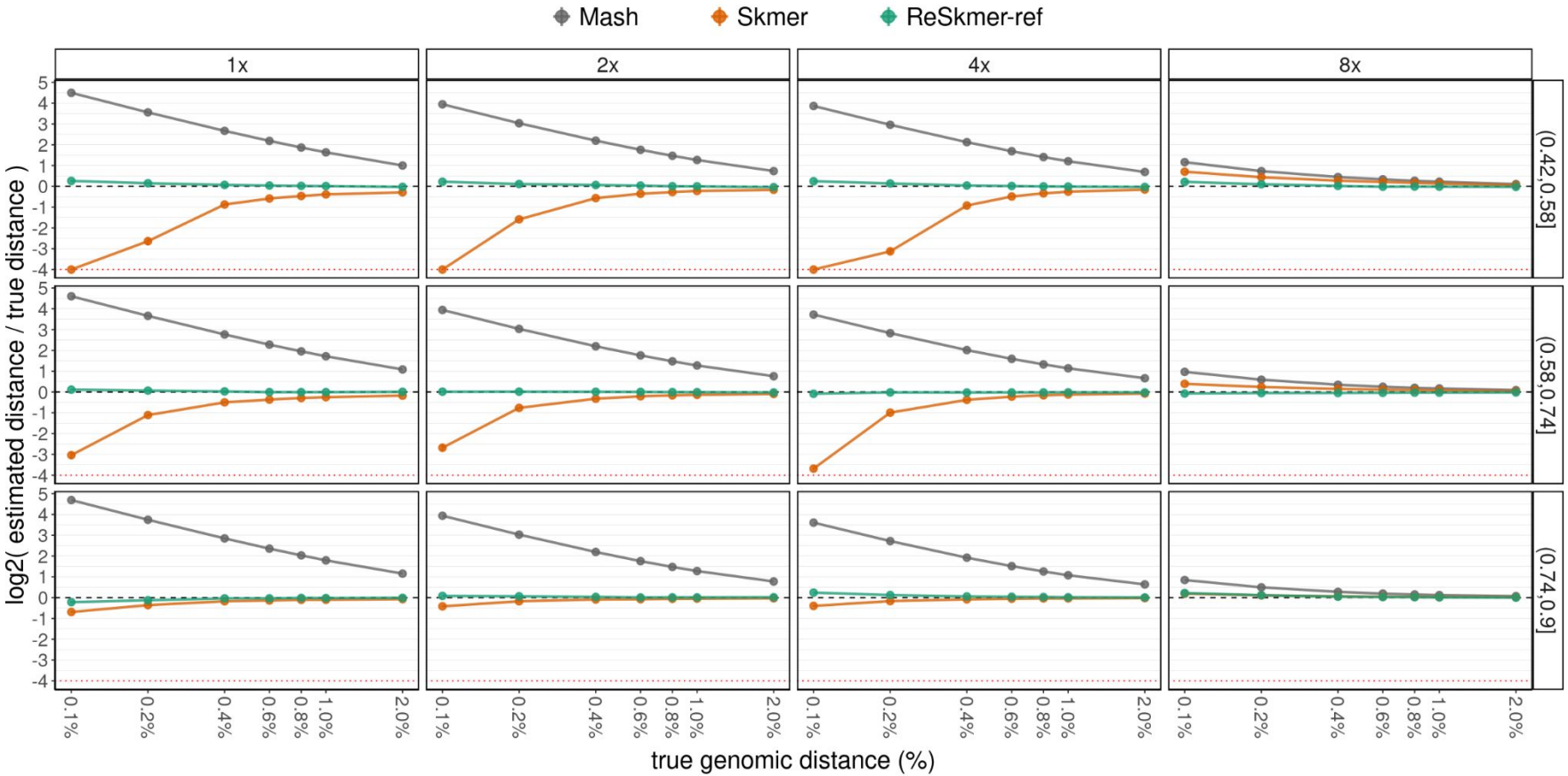
$$\epsilon = 1 - \left(\frac{\xi}{\lambda}\right)^{1/k}$$

Simulation Results

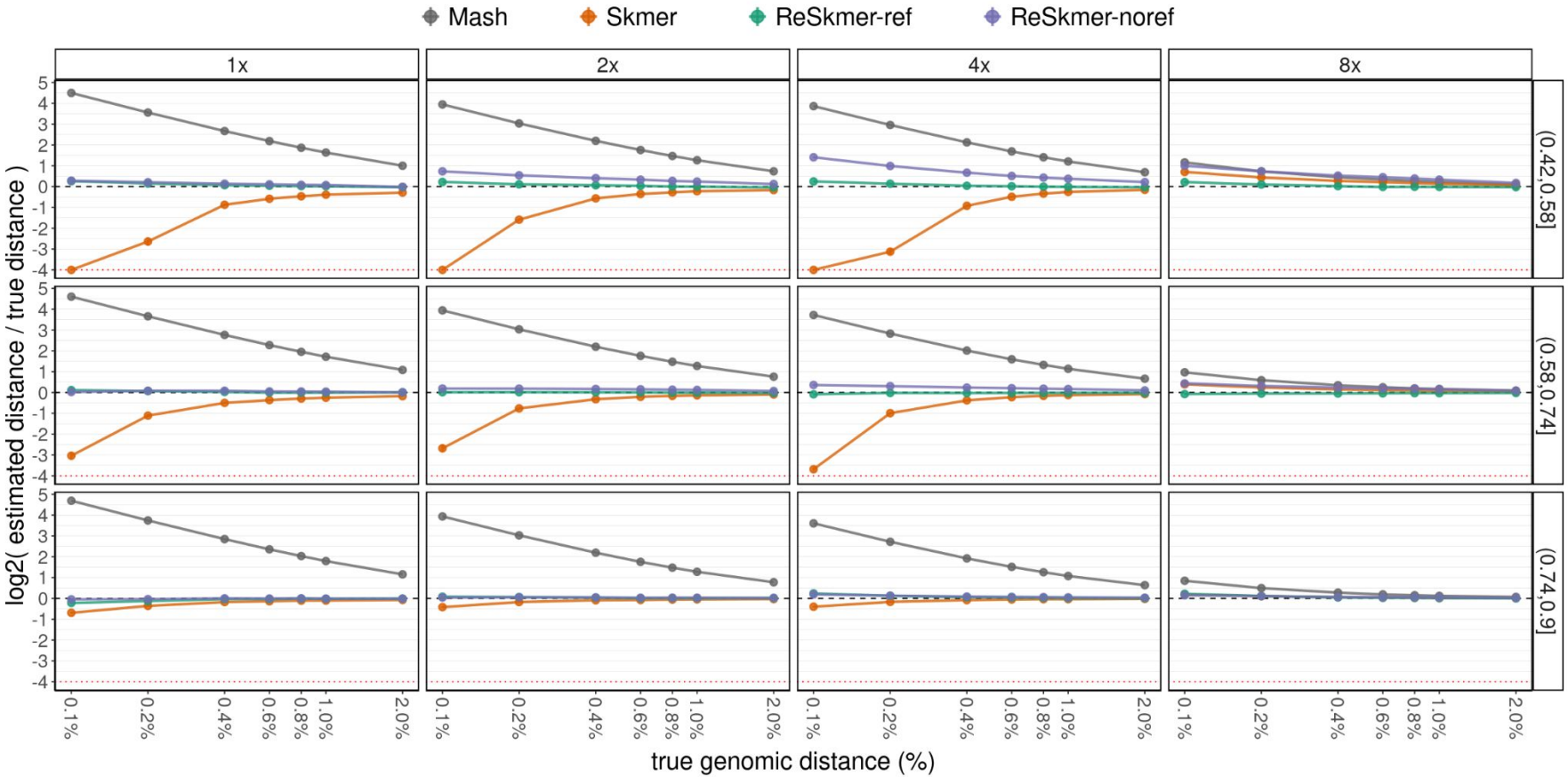
ReSkmer results



ReSkmer results



ReSkmer results



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

1X

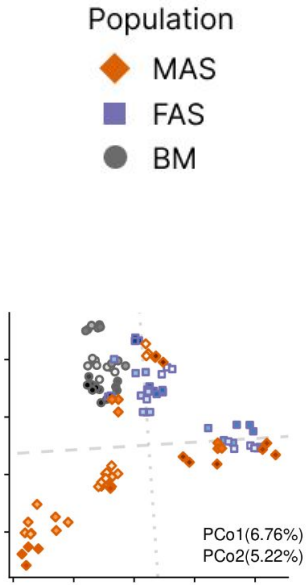
4X

Mash

Skmer

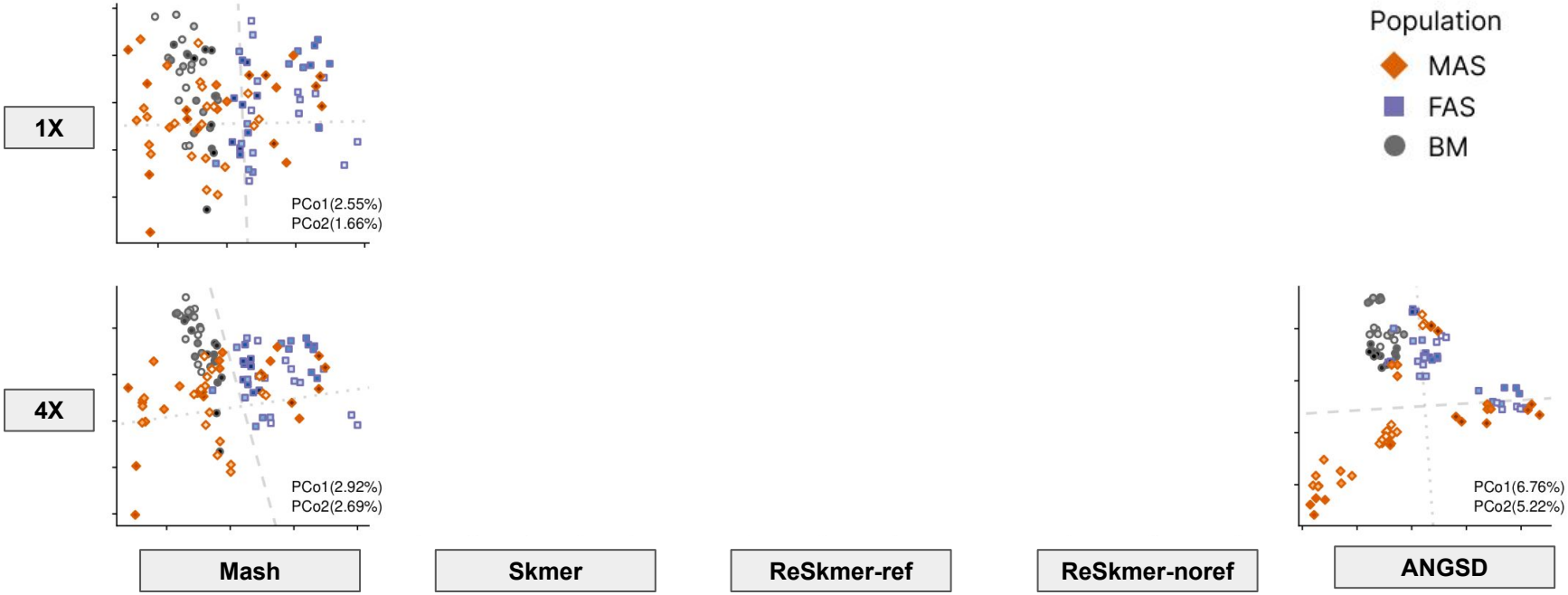
ReSkmer-ref

ReSkmer-noref

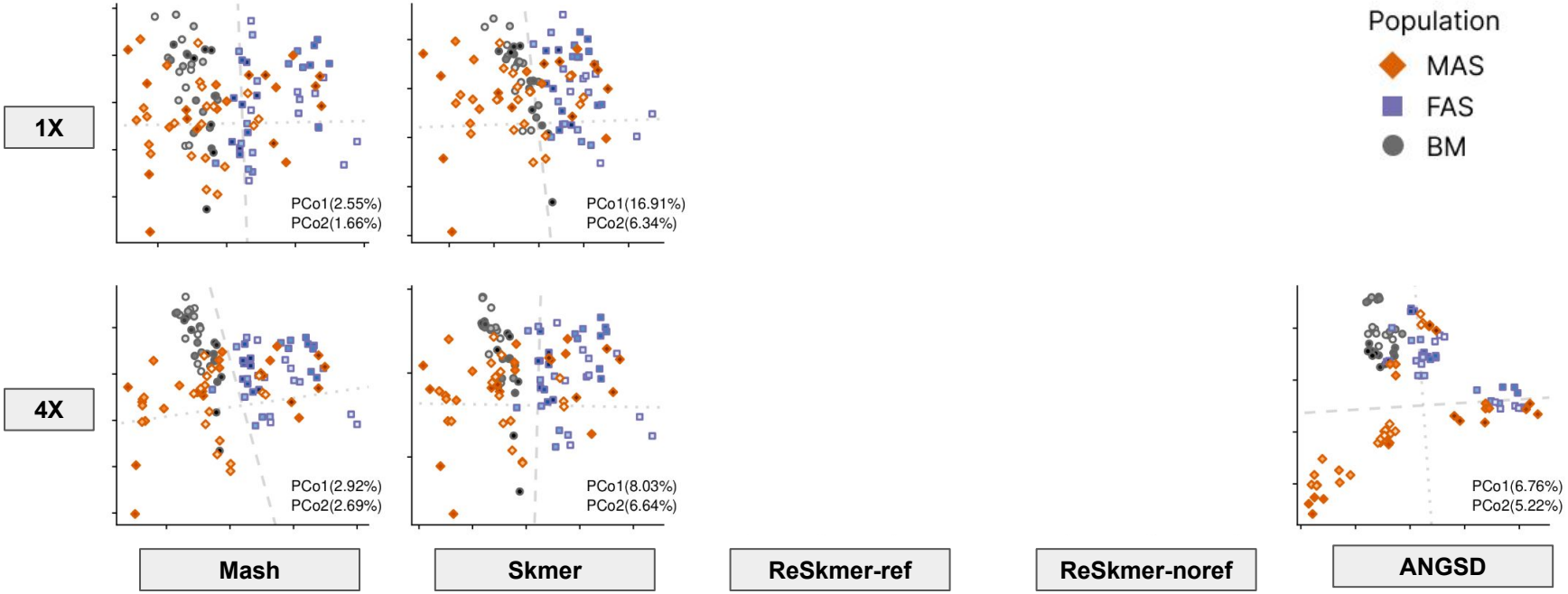


ANGSD

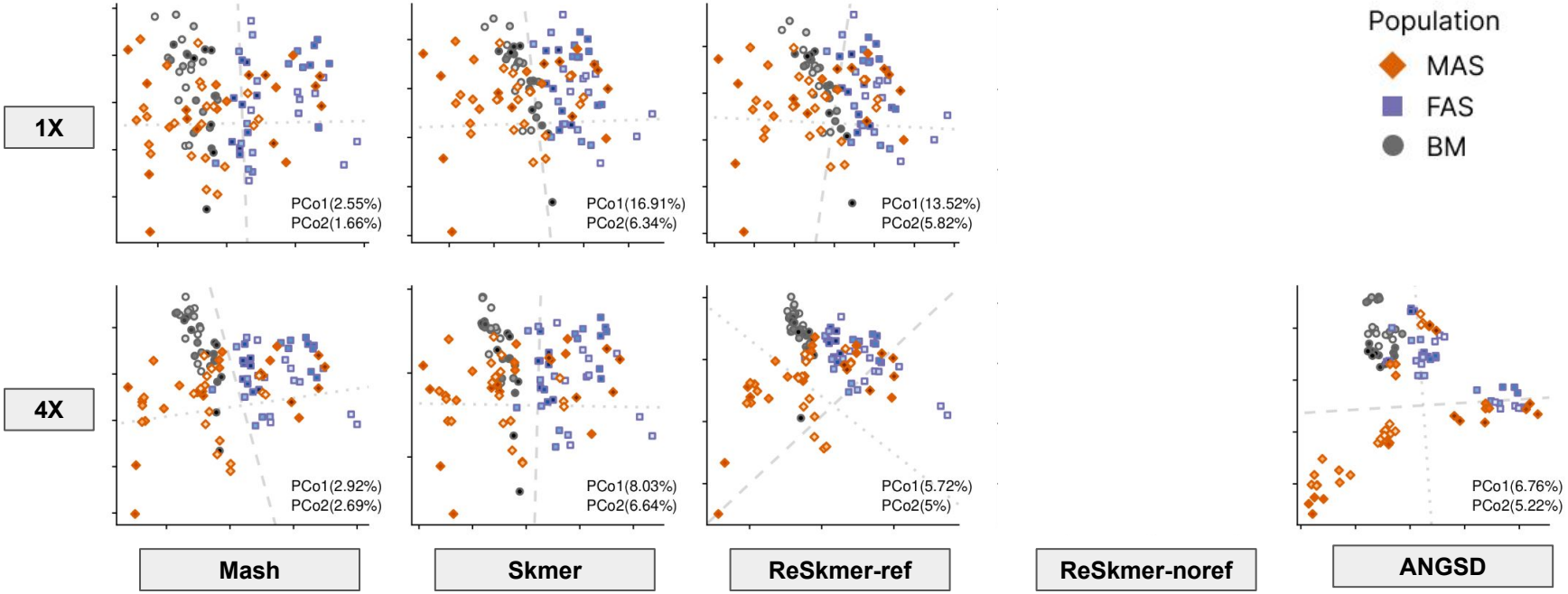
Honeybee clustering data



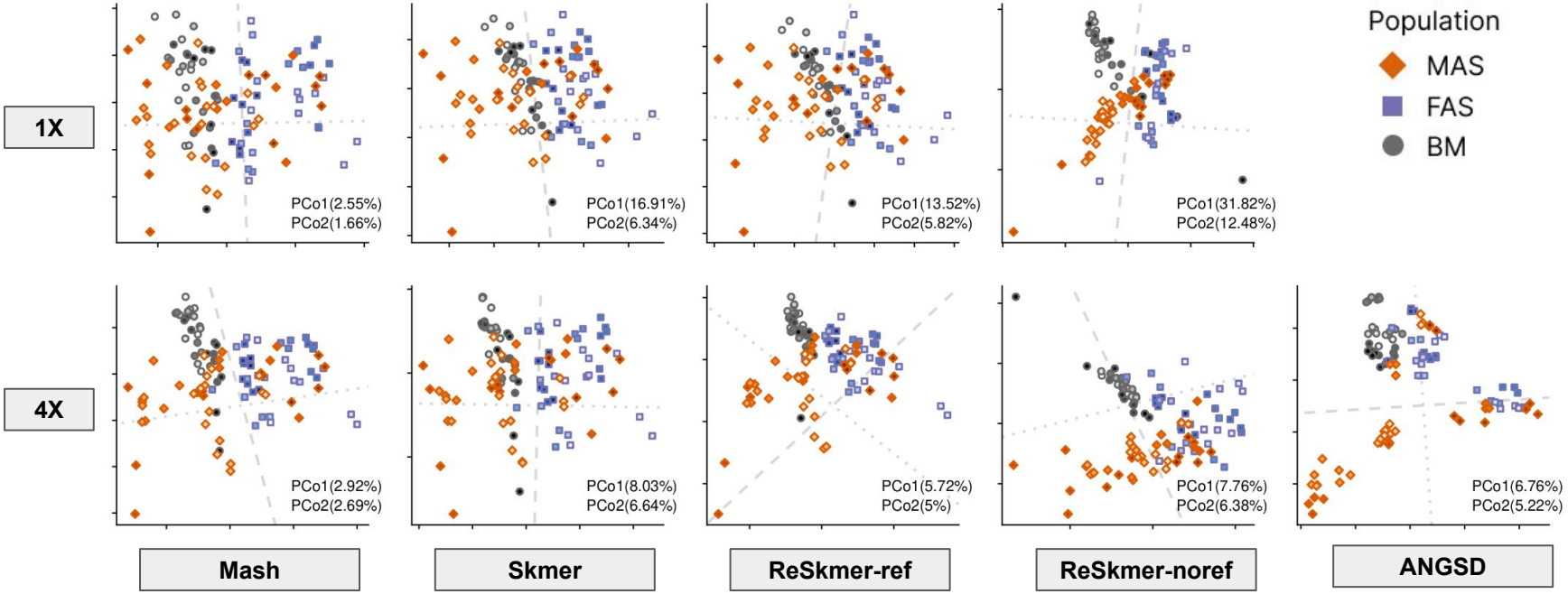
Honeybee clustering data



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 FREE

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?

	Apis mellifera				*Drosophila melanogaster*			
	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?

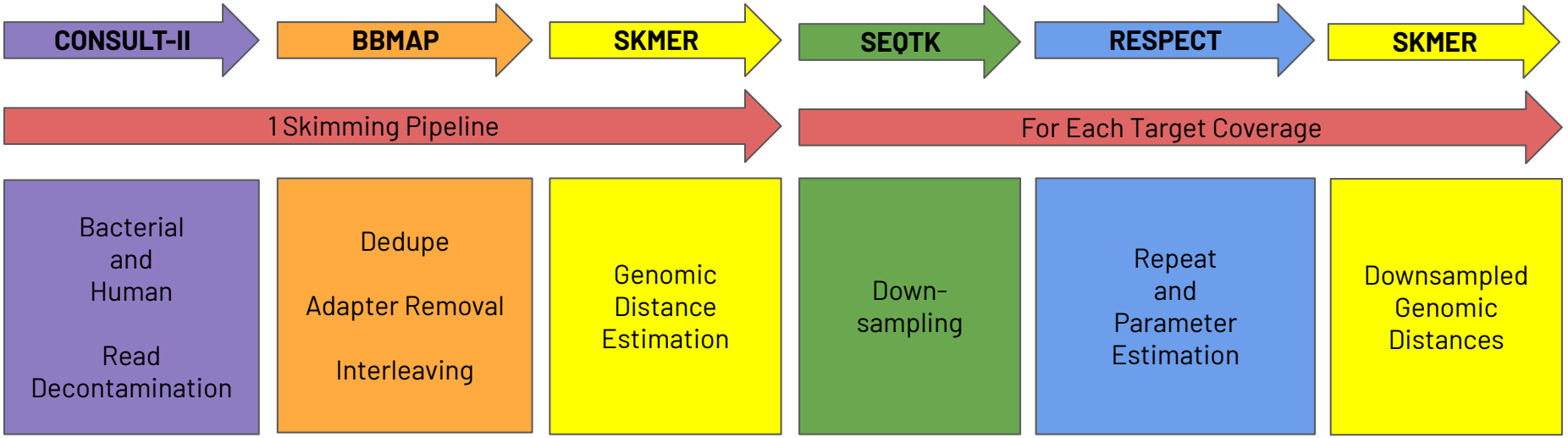
	Apis mellifera				*Drosophila melanogaster*			
	Mash	Skmer	ReSkmer (ref)	ReSkmer (noref)	Mash	Skmer	ReSkmer (ref)	ReSkmer (noref)
1×	0.12	0.61	0.60	**0.65**	0.12	0.58	**0.88**	0.76
2×	0.40	0.69	0.62	**0.71**	0.12	0.61	**0.90**	0.86
4×	0.63	**0.77**	0.58	0.71	0.40	0.69	0.88	**0.90**
6×	0.67	**0.72**	0.57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

Correlation results

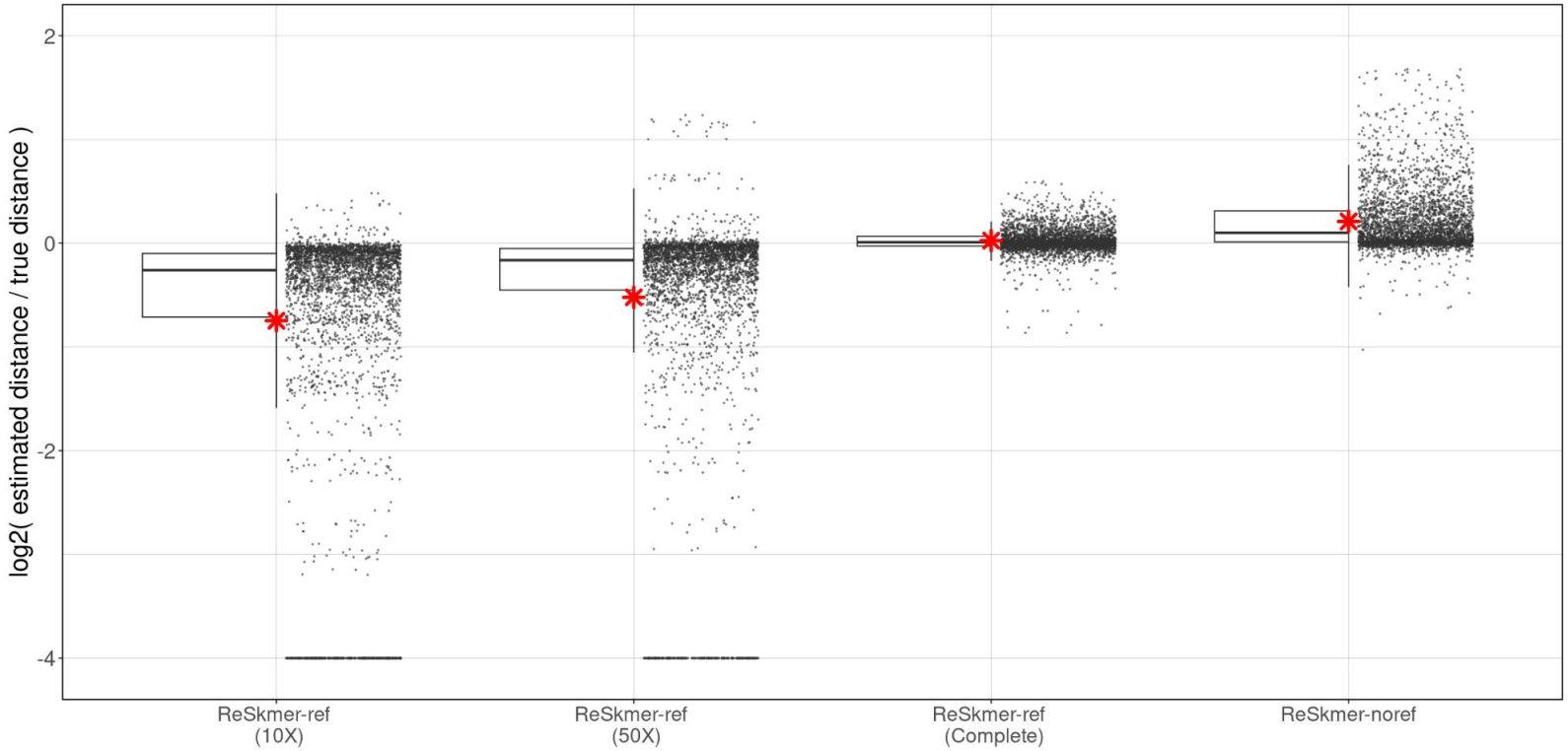
How well do **alignment-free** methods correlate with **alignment-based** methods?

	Apis mellifera				*Drosophila melanogaster*			
	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?



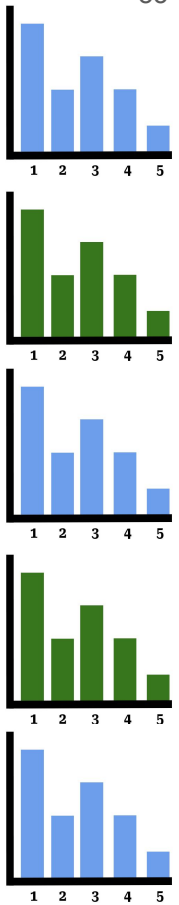
Conclusions

- 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 Monteiro[4], Shahab Sarmashghi[5], Glenn Dunshea[6], Vineet Bafna[3], and Siavash Mirarab[1]

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

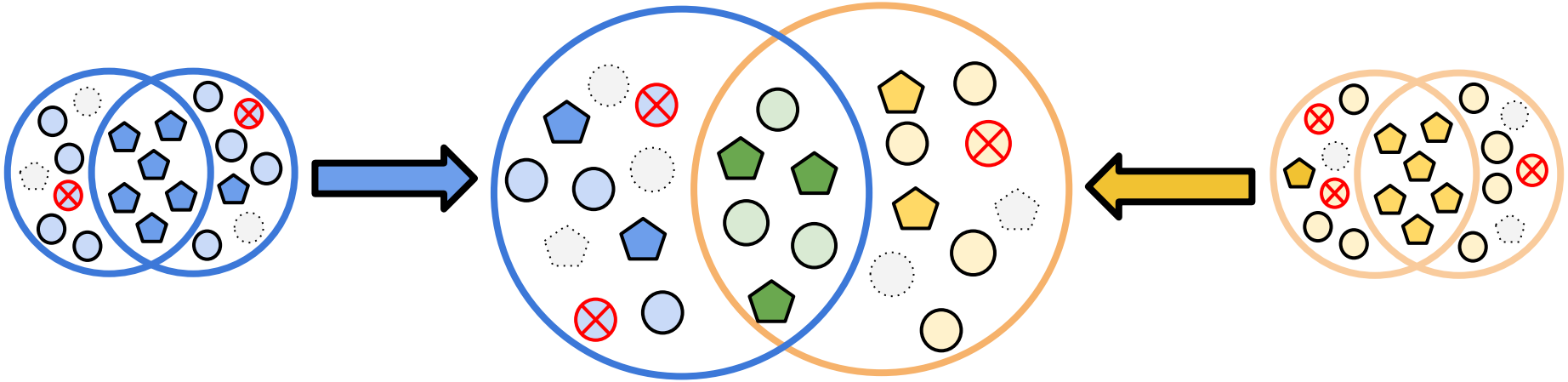


Future Work

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

- Whole Genome Resolution
- Reference and Alignment Free

- **<u>Low Coverage</u>**
- Models diploidy



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