The unrealized potential of genome skimming for sample identification

Siavash Mirarab

UC, San Diego (ECE Department)

Joint work with Shahab Sarmashghi, Metin Balaban, Nora Rachtman, Kristine Bohmann, Vineet Bafna, Tom Gilbert

Identifying a biological sample is essential to many applications: monitoring biodiversity and conservation to tracing food provenance

Sample identification is not easy!

The New York Times

Survey Finds That Fish Are Often Not What Label Says

SEATTLE — Many Europeans are fretting these days over horse meat, and whether it might have adulterated their shepherd's pie. Over here, it's all about the red snapper.

MENU nature International journal of science Subscribe Search Login

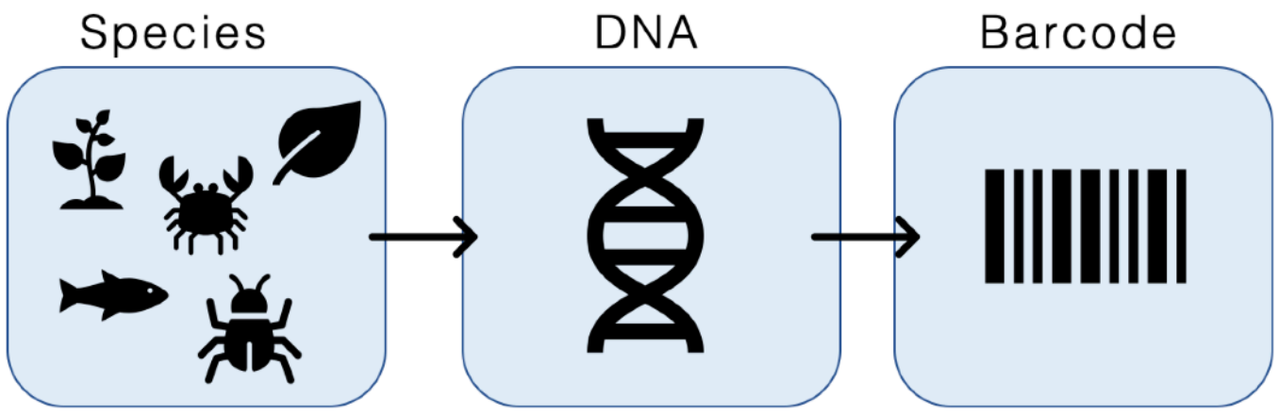
NEWS · 06 MAY 2019 · UPDATE 06 MAY 2019

Humans are driving one million species to extinction

Landmark United Nations-backed report finds that agriculture is one of the biggest threats to Earth's ecosystems.

Existing DNA barcoding

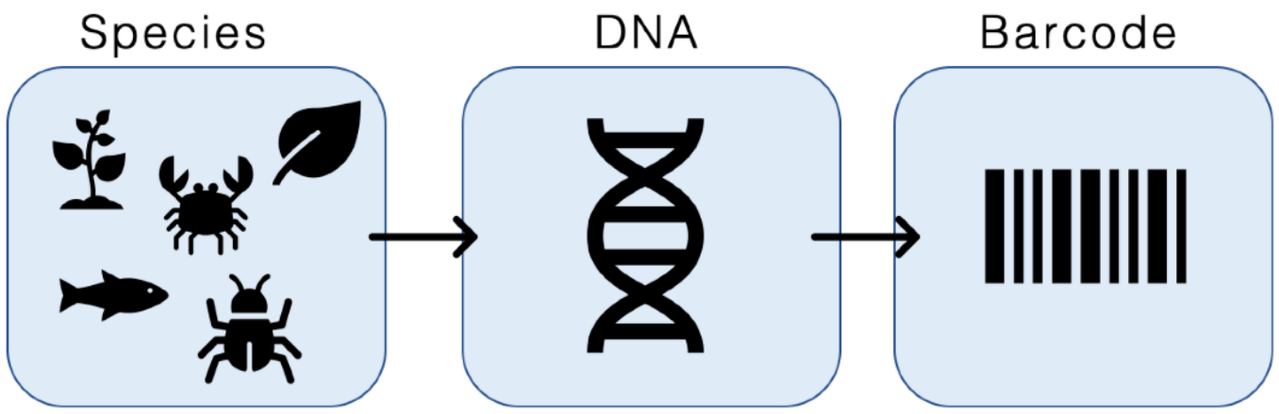
- Represent each species by its DNA sequence, like a barcode. Compare a query to a library of known barcodes (e.g., BOLD) to find matches
  - Uses the cheap PCR amplification of marker genes (e.g., COI)
  - Supposed to be varied enough to distinguish species



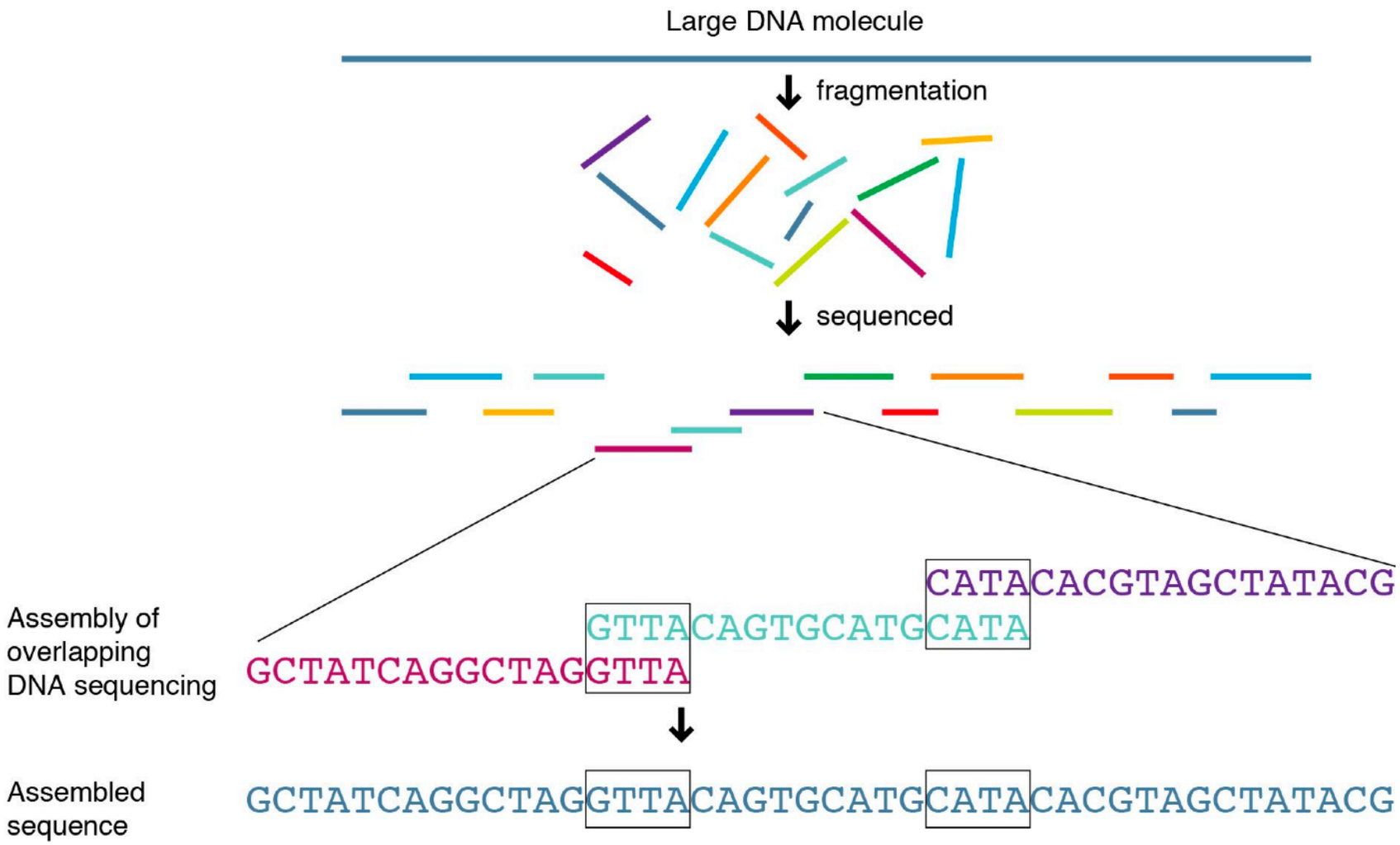
Existing DNA barcoding

- Represent each species by its DNA sequence, like a barcode. Compare a query to a library of known barcodes (e.g., BOLD) to find matches
  - Uses the cheap PCR amplification of marker genes (e.g., COI)
  - Supposed to be varied enough to distinguish species
- Barcodes are short, so have limited resolution
  - Often fail to distinguish species of the same genus

Hebert *et al.*, 2003, Proc. Royal Soc. B

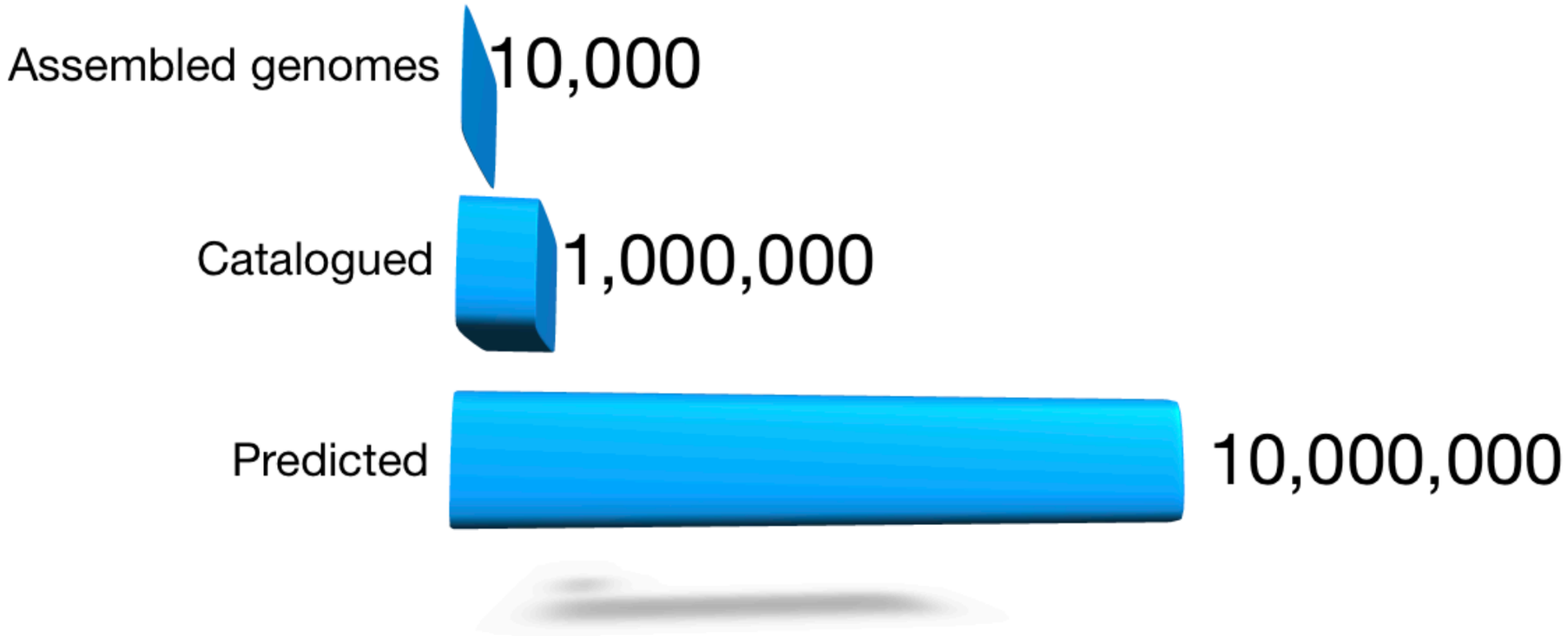


How about sequencing the whole genomes? Isn't that cheap these days?



Assembly is challenging and costly: it requires a high coverage of genome (>30X)

- Eukaryotic organisms



Mora *et al.*, 2011, Plos Biology

Genome skimming: low-coverage NGS

Quicke *et al.*, 2012, Mol. Eco. Resources

Coissac *et al.*, 2016, Molecular Ecology

Genome skimming: low-coverage NGS

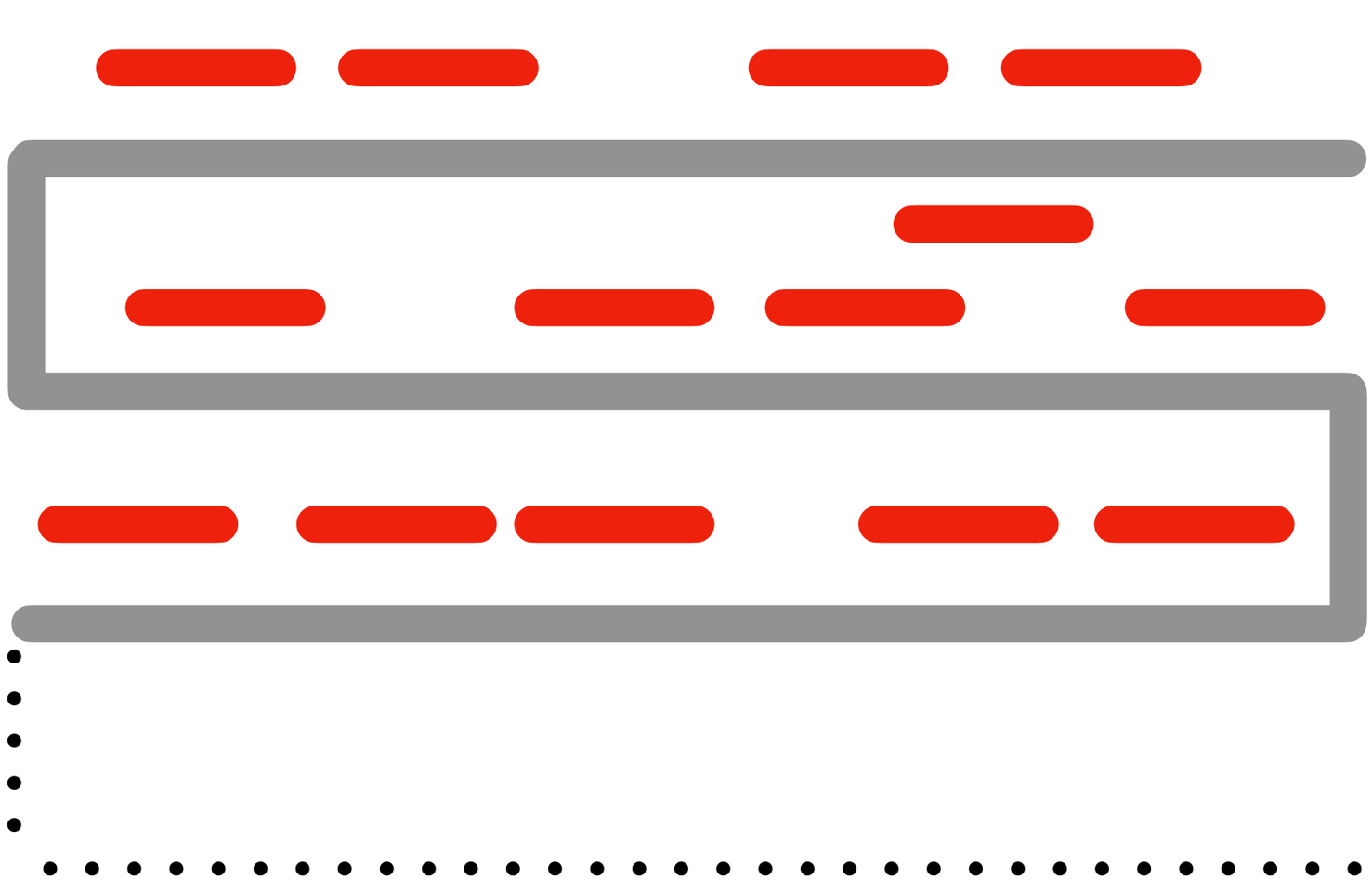
Quicke *et al.*, 2012, Mol. Eco. Resources

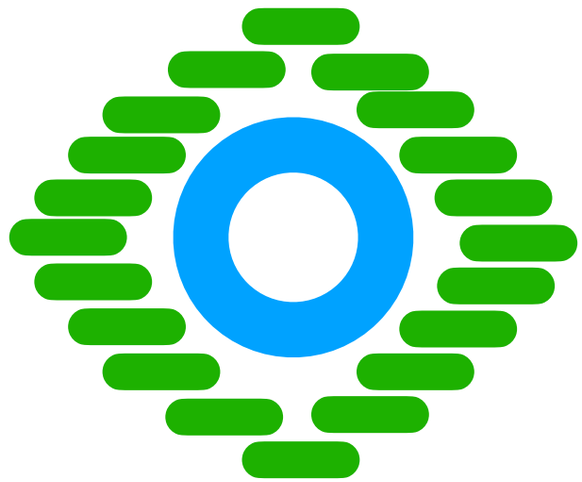
Coissac *et al.*, 2016, Molecular Ecology

Can we use all reads in a genome skim?

Nuclear genome

Organelle genome

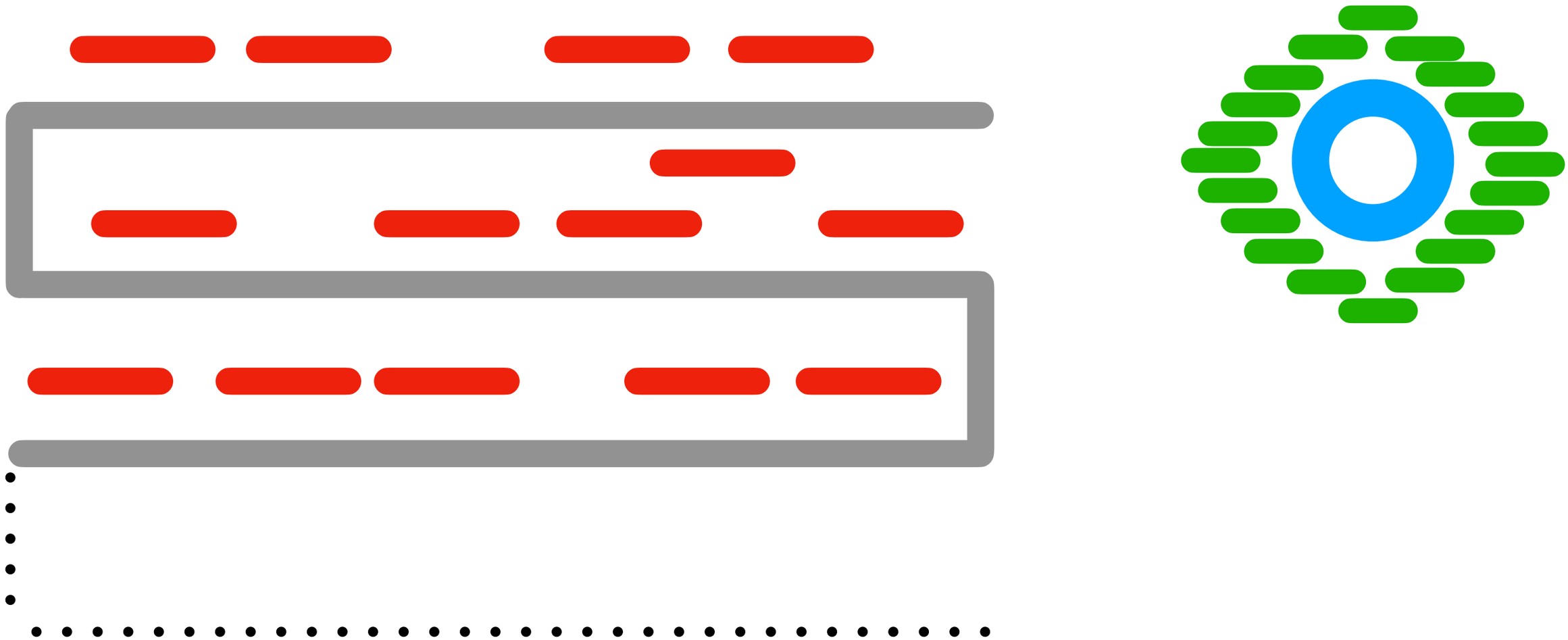




Can we use all reads in a genome skim?

Low read depth → Assembly is not possible

e genome

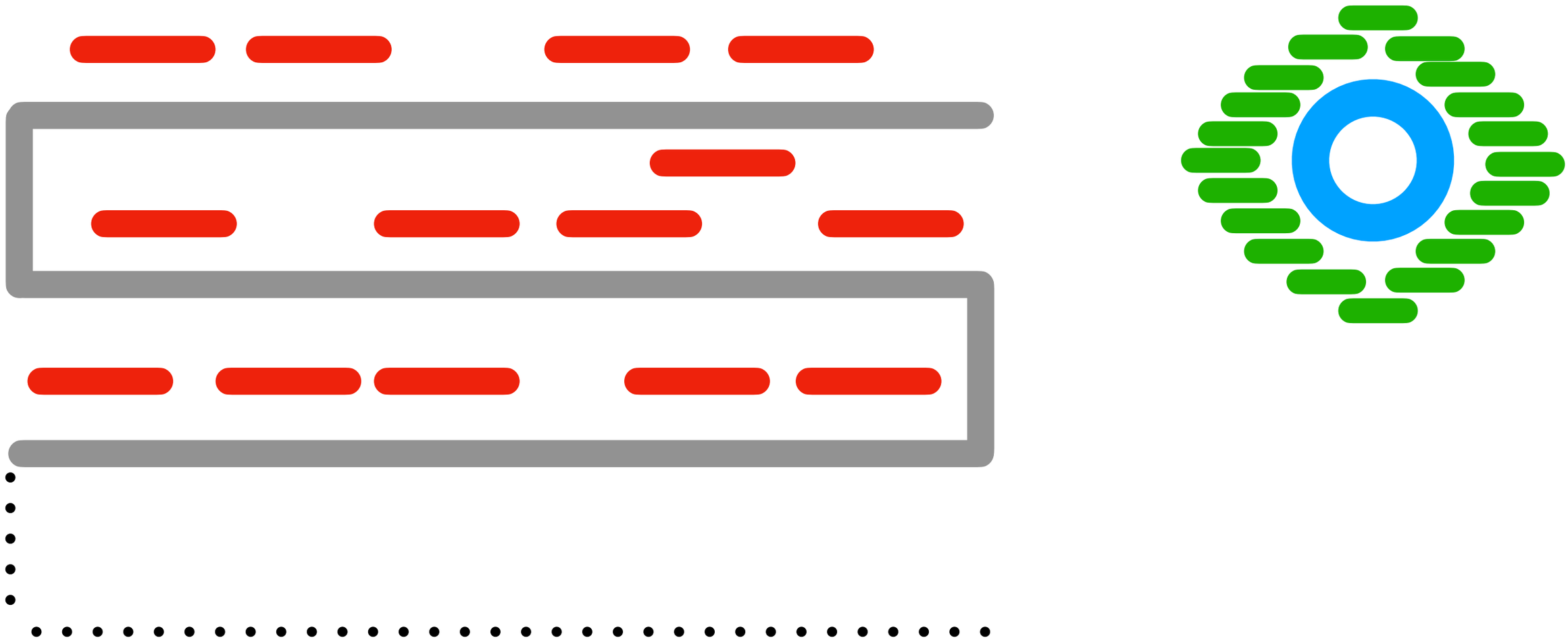


Can we use all reads in a

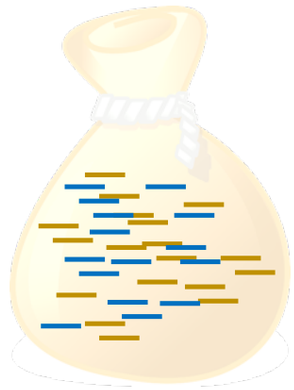
No reference → Alignment is not easy

Low read depth → Assembly is not possible

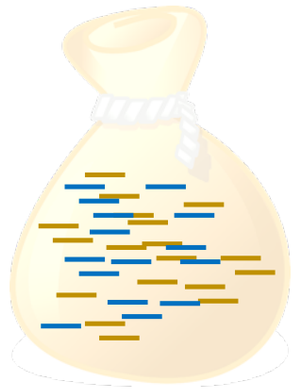
e genome



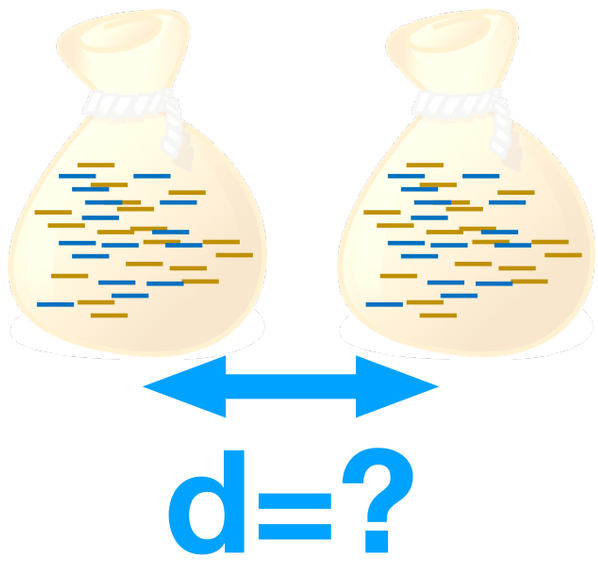
What if we barcode species as bags of reads?



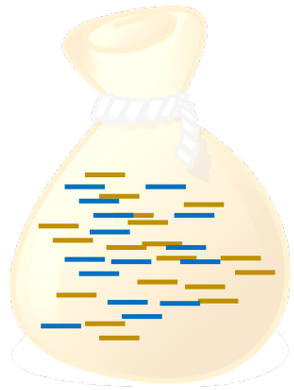
What if we barcode species as bags of reads?



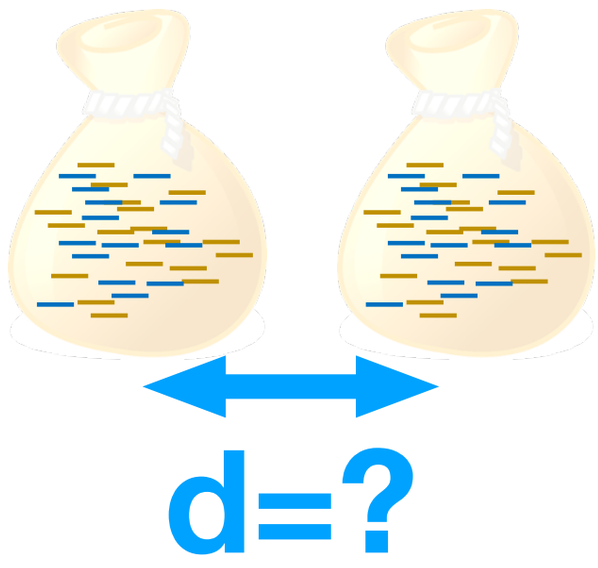
We need to quantify the distance between two genomes skims from these bags of reads



What if we barcode species as bags of reads?

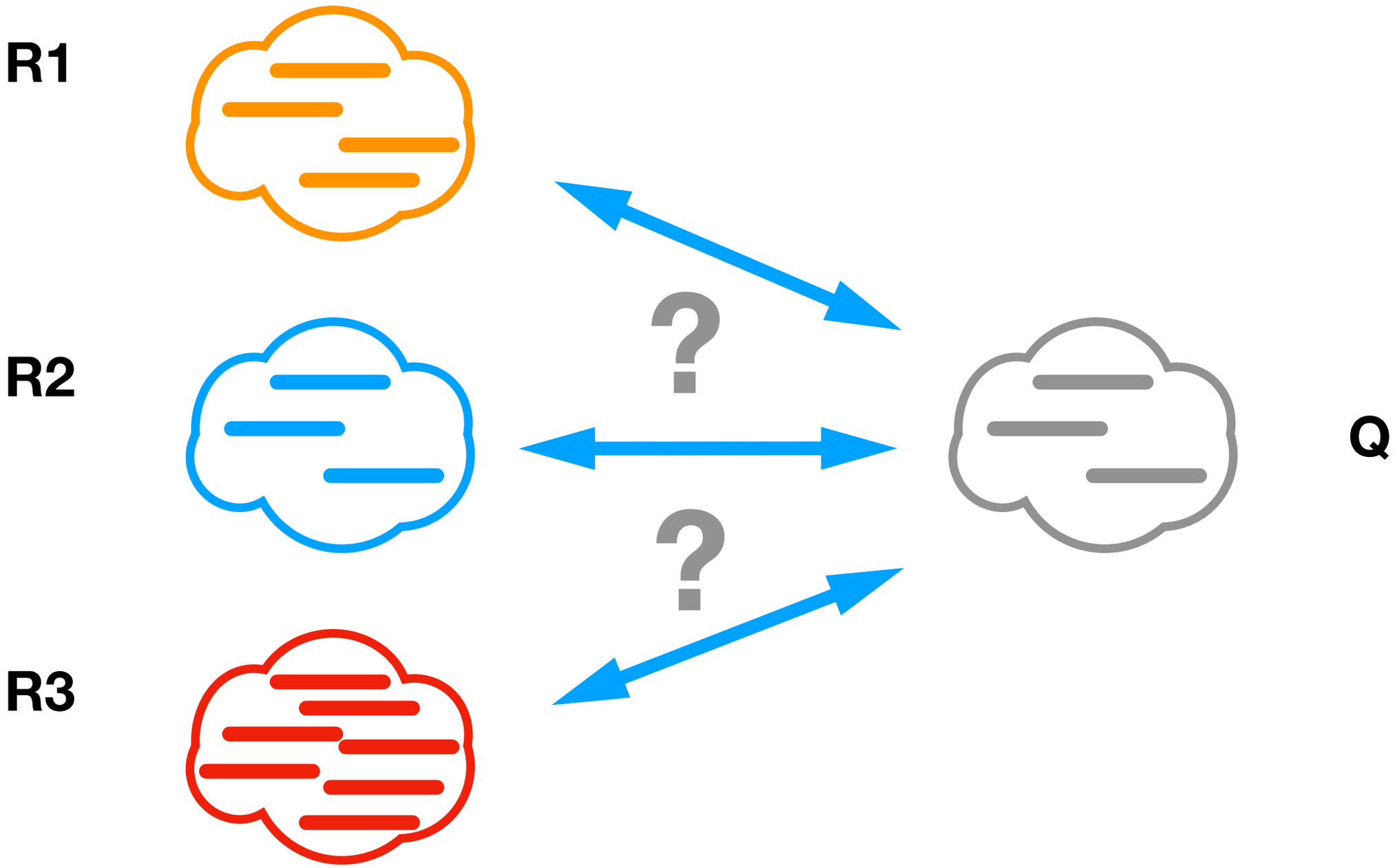


We need to quantify the distance between two genomes skims from these bags of reads



Can we do this when the bags of reads provides a low coverage of the genome (e.g., 1X)?

Compare to references



Genomic nucleotide distance

- Distance is defined as the fraction of sites that would not match if we had perfectly correct alignments of the fully assembled genomes (i.e., hamming distance)

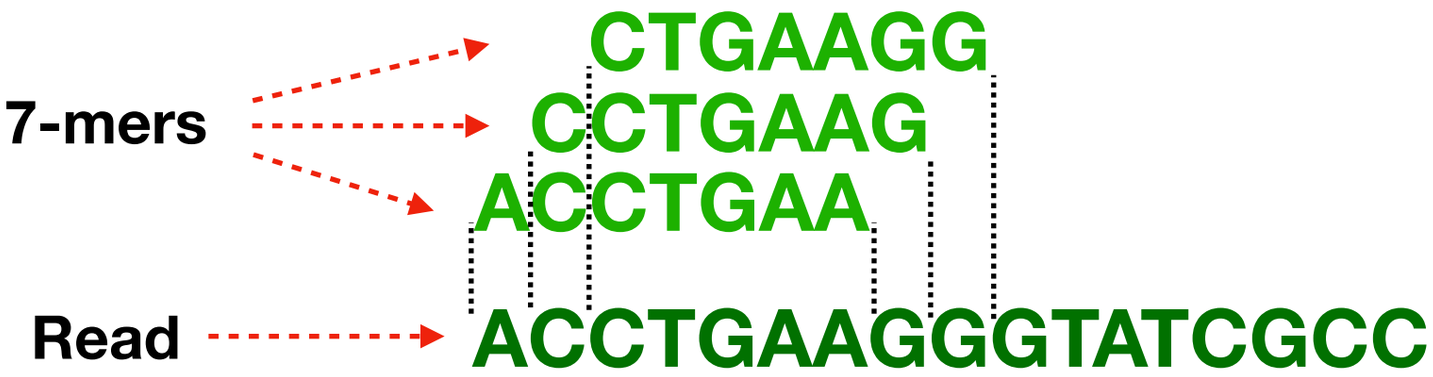
Genome1 …ACCTGAAGGGTATCGCC…

Genome2 …ACCTGAGGGGTATAGCC…

d=2/16

k-mer representation

- Reads are decomposed into consecutive k-mers



Alignment-free comparison of reads

- Sets of k-mers for two genome-skims are compared

Genome skim 1

CTGAA
CCTGA
ACCTG

ACCTGAAGGGTATC

TCGCC
ATCGC
TATCG

TATCGCCAAAAGCG

Genome skim 2

CTGAA
CCTGA
ACCTG

ACCTGAAGGGTATC

TCGCC
ATCGC
TATCG

TATCGCCAAAAGCG

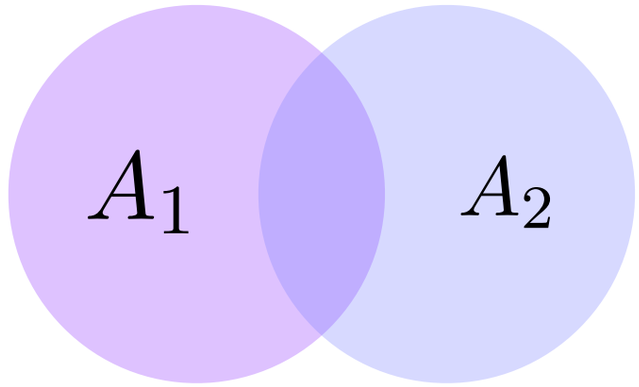
Alignment-free comparison of reads

- Sets of k-mers for two genome-skims are compared

Jaccard index

- Using Jaccard index to compute distances
- Measure of similarity between k-mer sets

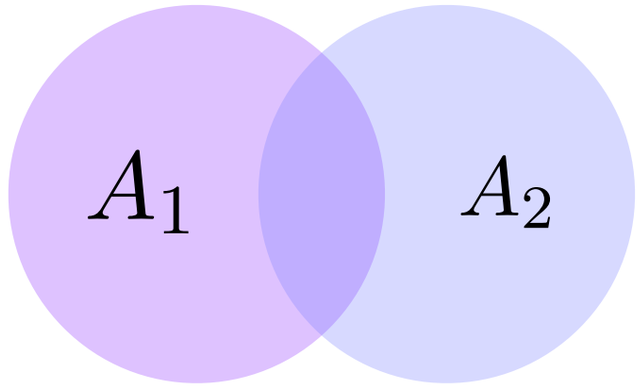
$$J = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}$$



Jaccard index

- Using Jaccard index to compute distances
- Measure of similarity between k-mer sets

$$J = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}$$



- Can be efficiently computed with a MinHash technique (Mash)

Ondov *et al.*, 2016, Genome Biology

Efficient Jaccard estimation

- Mash computes the Jaccard index efficiently via MinHash

Input size	Running time
100Mb	1min
500Mb	2min
1Gb	3min

Ondov *et al.*, 2016, Genome Biology

Jaccard to distance

- Distance (D) can be estimated from the Jaccard (J) using a simple Poisson model

- W = The number of shared k-mers

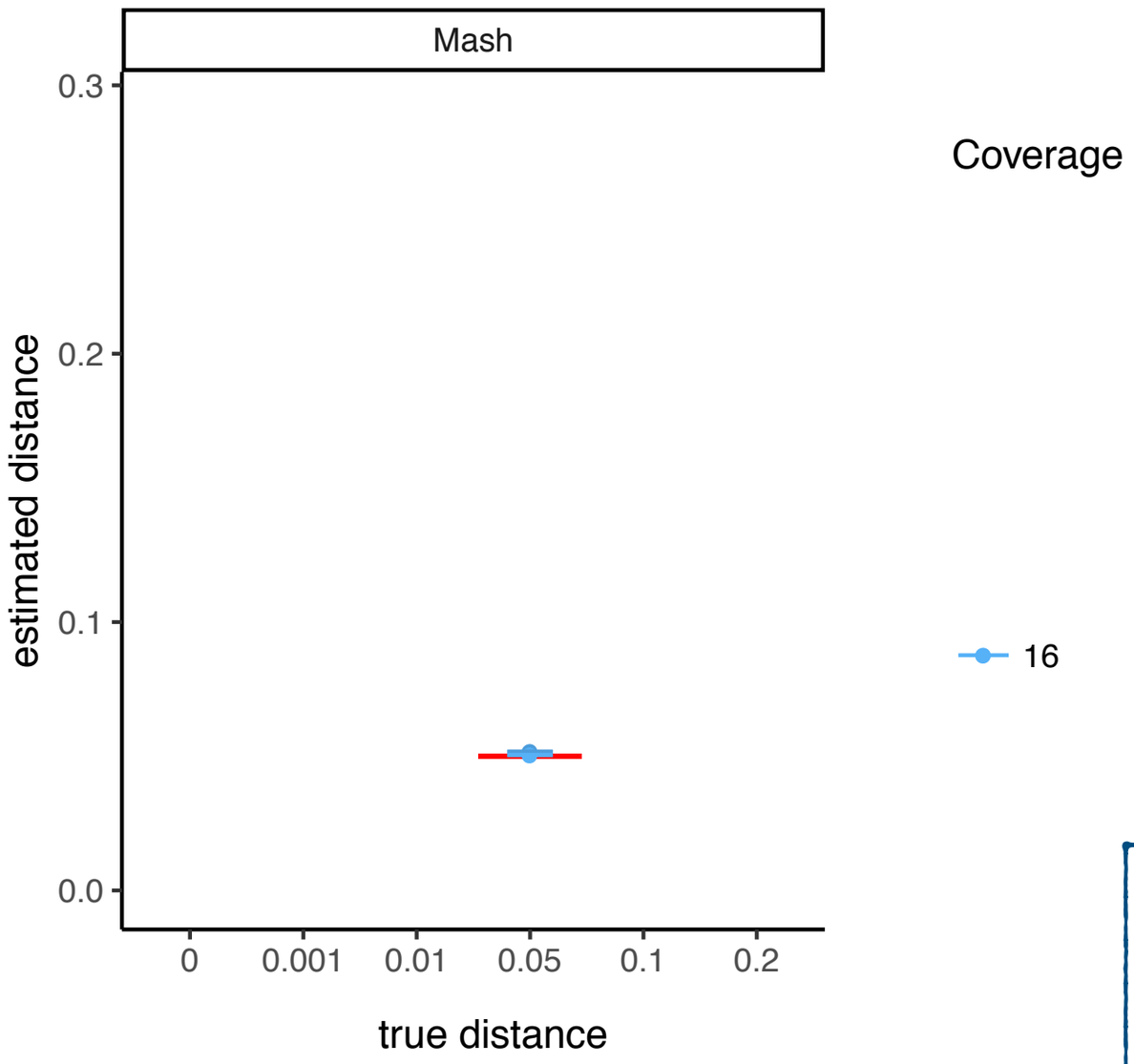
$$J = \frac{W}{L + L - W}$$

$$\rightarrow J = \frac{L(1-D)^k}{2L - L(1-D)^k}$$

$$\rightarrow \boxed{D = 1 - (\frac{2J}{1+J})^{1/k}}$$

Mash accuracy

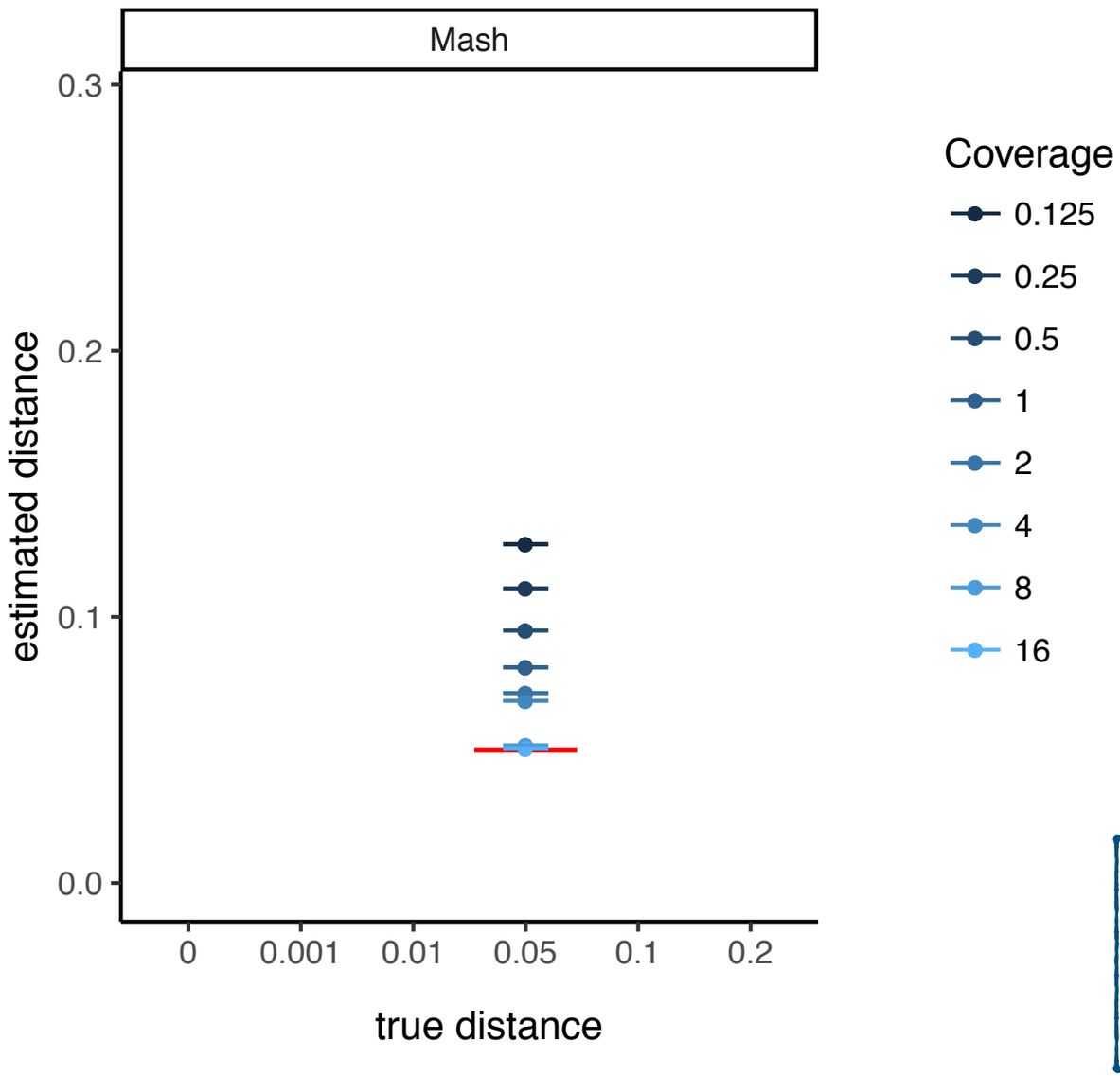
- Works well at high coverage



5% random mutations added to an insect genome

Mash accuracy

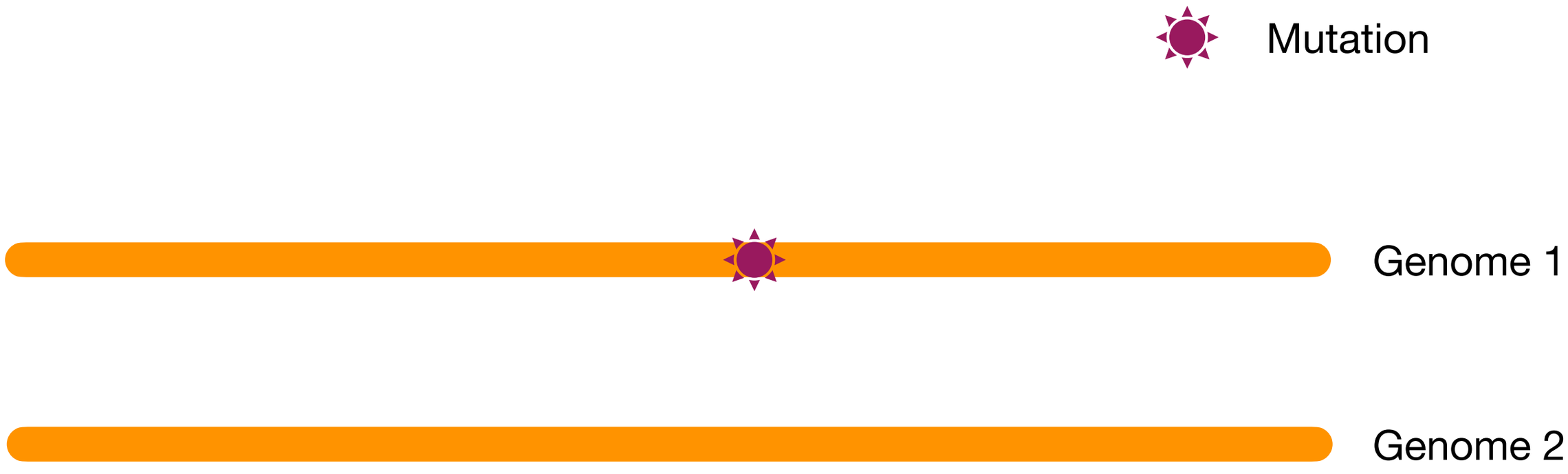
- Inaccurate when coverage is < 4X



5% random mutations added to an insect genome

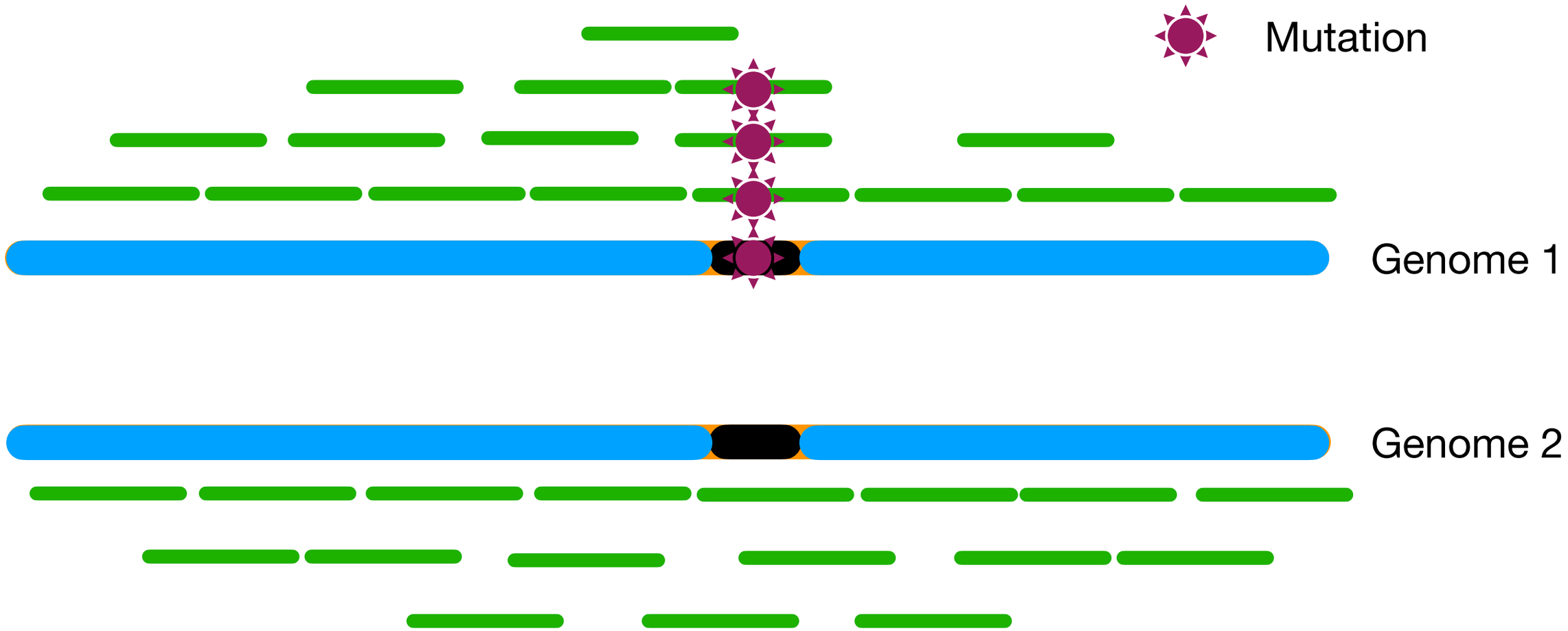
What happens when coverage is low?

- The Jaccard is reduced as some parts are not covered by any error-free read



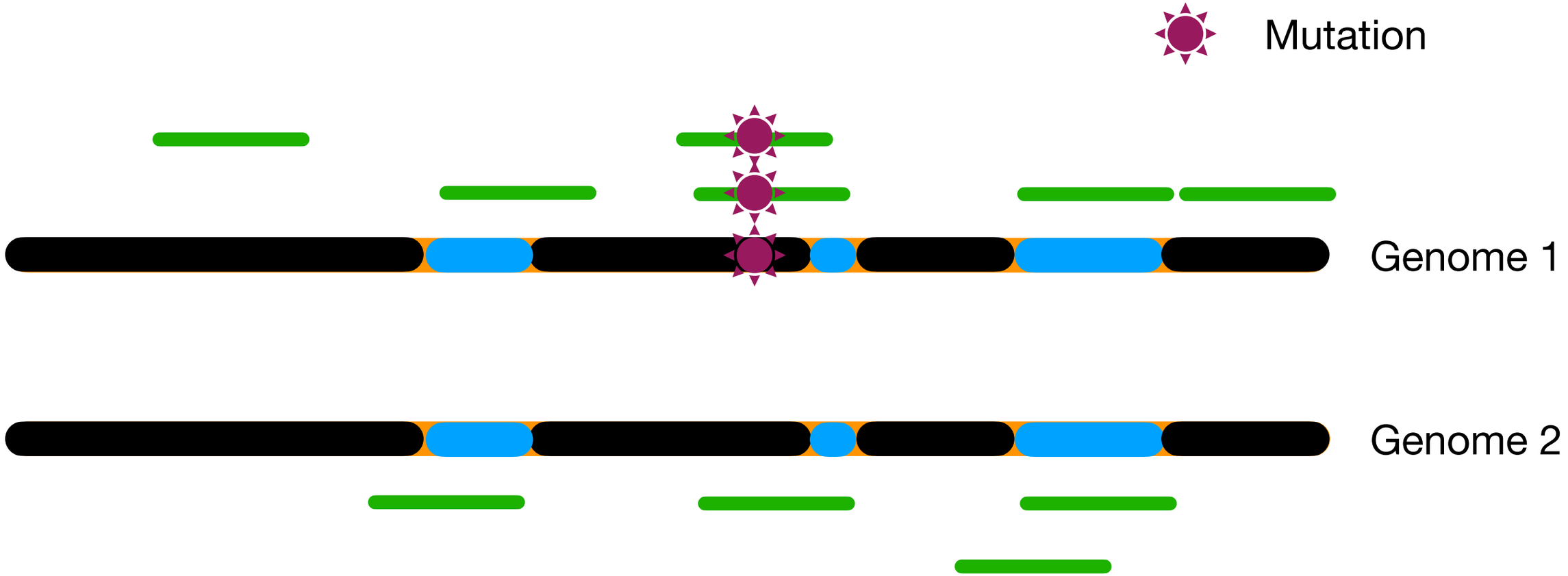
What happens when coverage is low?

- The Jaccard is reduced as some parts are not covered by any error-free read



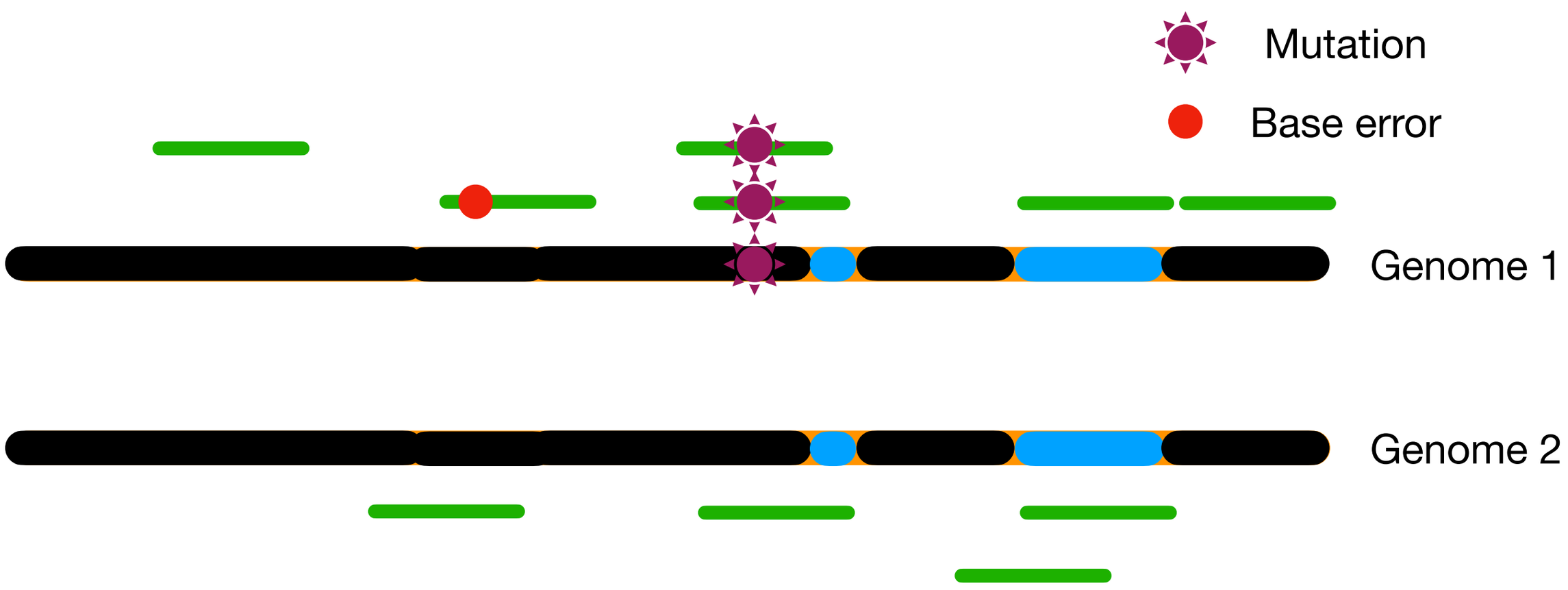
What happens when coverage is low?

- The Jaccard is reduced as some parts are not covered by any error-free read



What happens with sequencing error?

- The Jaccard is reduced as some parts are not covered by any error-free read



Jaccard to distance with low coverage and error

- Write Jaccard as a function of genomic distance and sequencing parameters

- Number of reads covering a k-mer follows a Poisson distribution

Jaccard to distance with low coverage and error

- Write Jaccard as a function of genomic distance and sequencing parameters

- Number of reads covering a k-mer follows a Poisson distribution

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

Average number of k-mers per locus

Probability that a k-mer is covered

Jaccard to distance with low coverage and error

- Write Jaccard as a function of genomic distance and sequencing parameters

- Number of reads covering a k-mer follows a Poisson distribution

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

Average number of k-mers per locus

Probability that a k-mer is covered

η_i and ζ_i are simple funcions of coverage (λ_i) and sequencing error (ϵ_i)

Jaccard to distance with low coverage and error

- Write Jaccard as a function of genomic distance and sequencing parameters

- Number of reads covering a k-mer follows a Poisson distribution

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

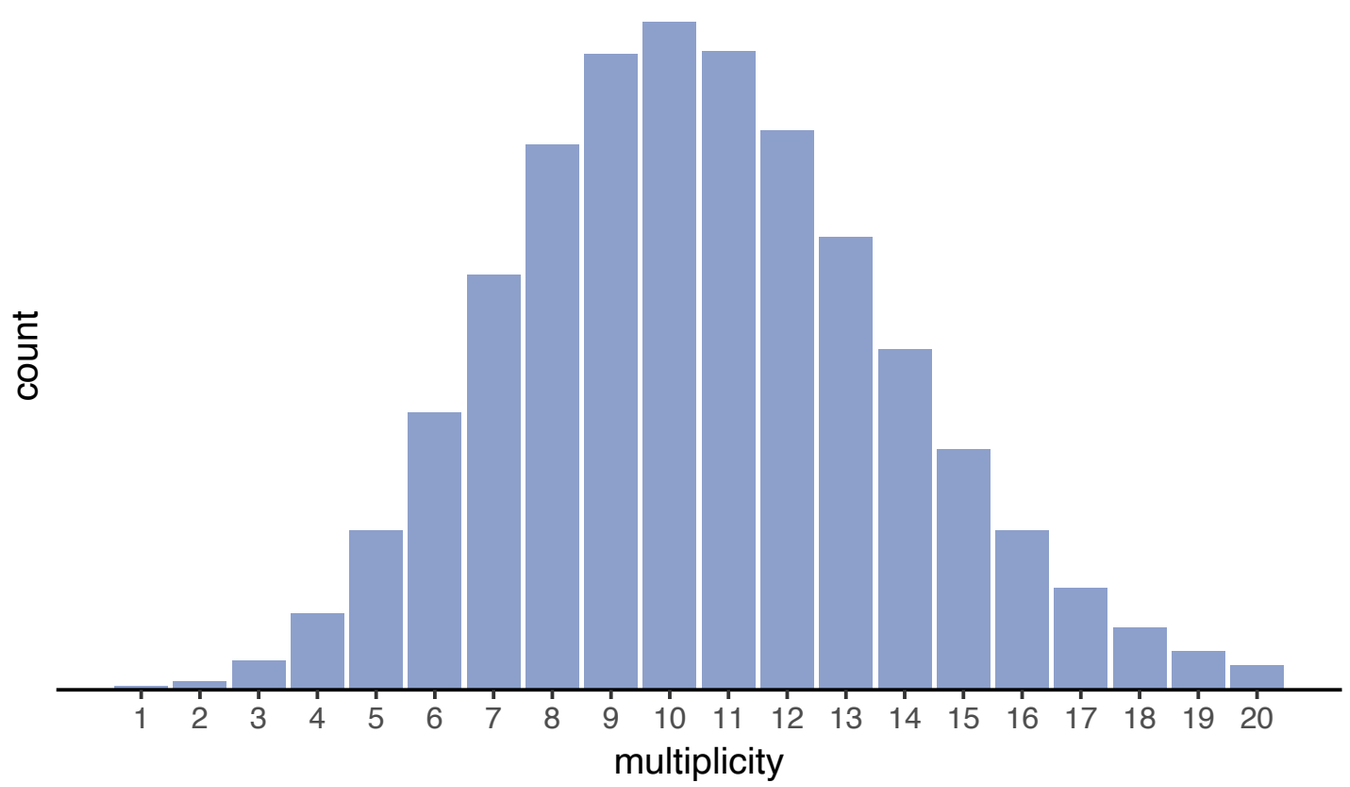
Average number of k-mers per locus

Probability that a k-mer is covered

η_i and ζ_i are simple funcions of coverage (λ_i) and sequencing error (ϵ_i)

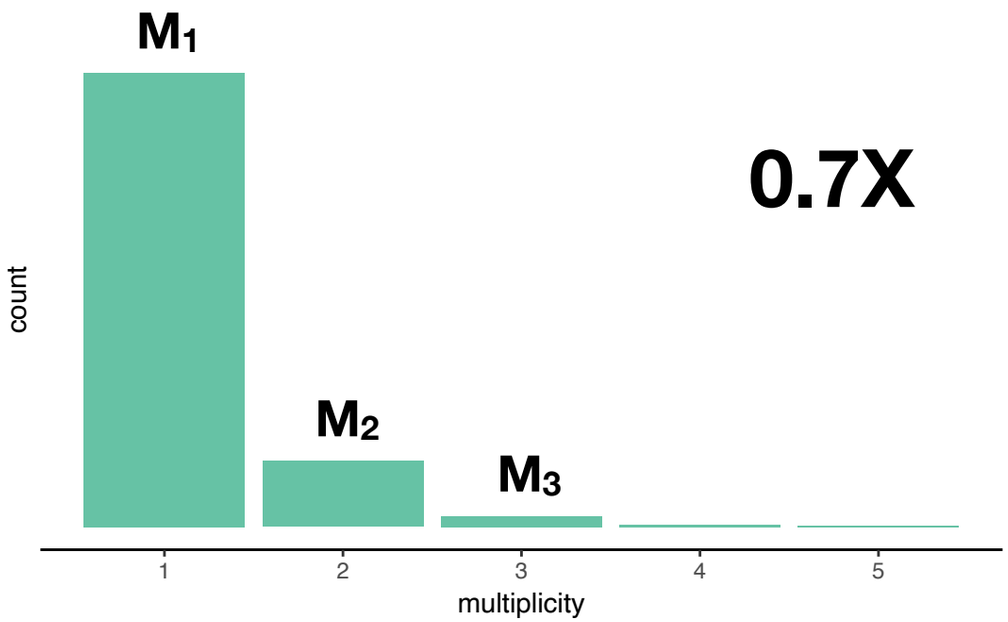
Estimating the sequencing parameters (λ_i and ϵ_i)

- Using k-mer profile (i.e., spectrum)
- For high coverage and zero error



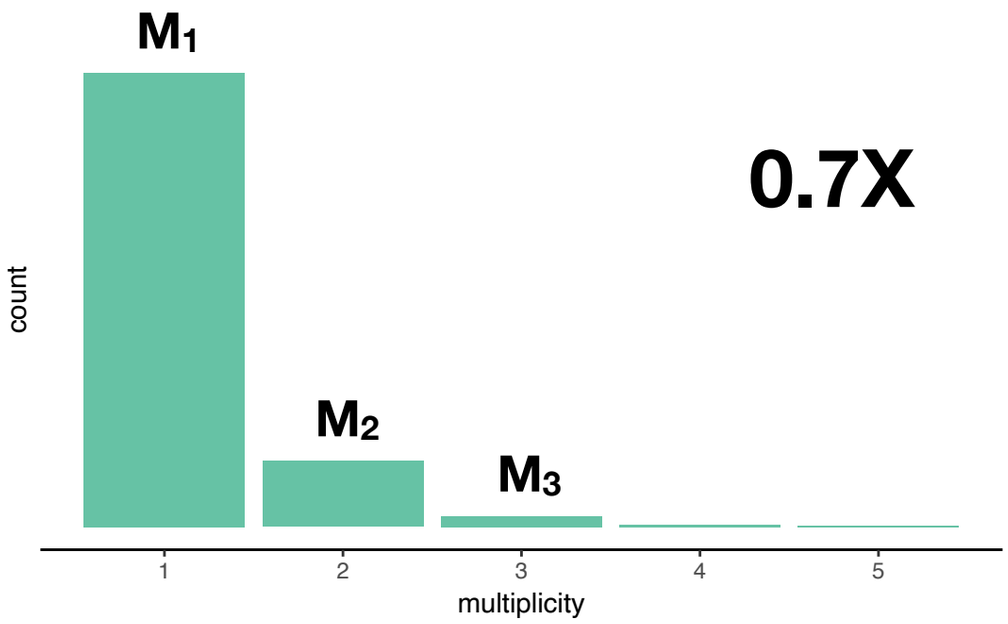
Interpreting k-mer counts at low coverage

- Just using the mode won't work
  - we don't see zeros, mode is always zero for low coverage, error adds to first bin



Interpreting k-mer counts at low coverage

- Just using the mode won't work
  - we don't see zeros, mode is always zero for low coverage, error adds to first bin



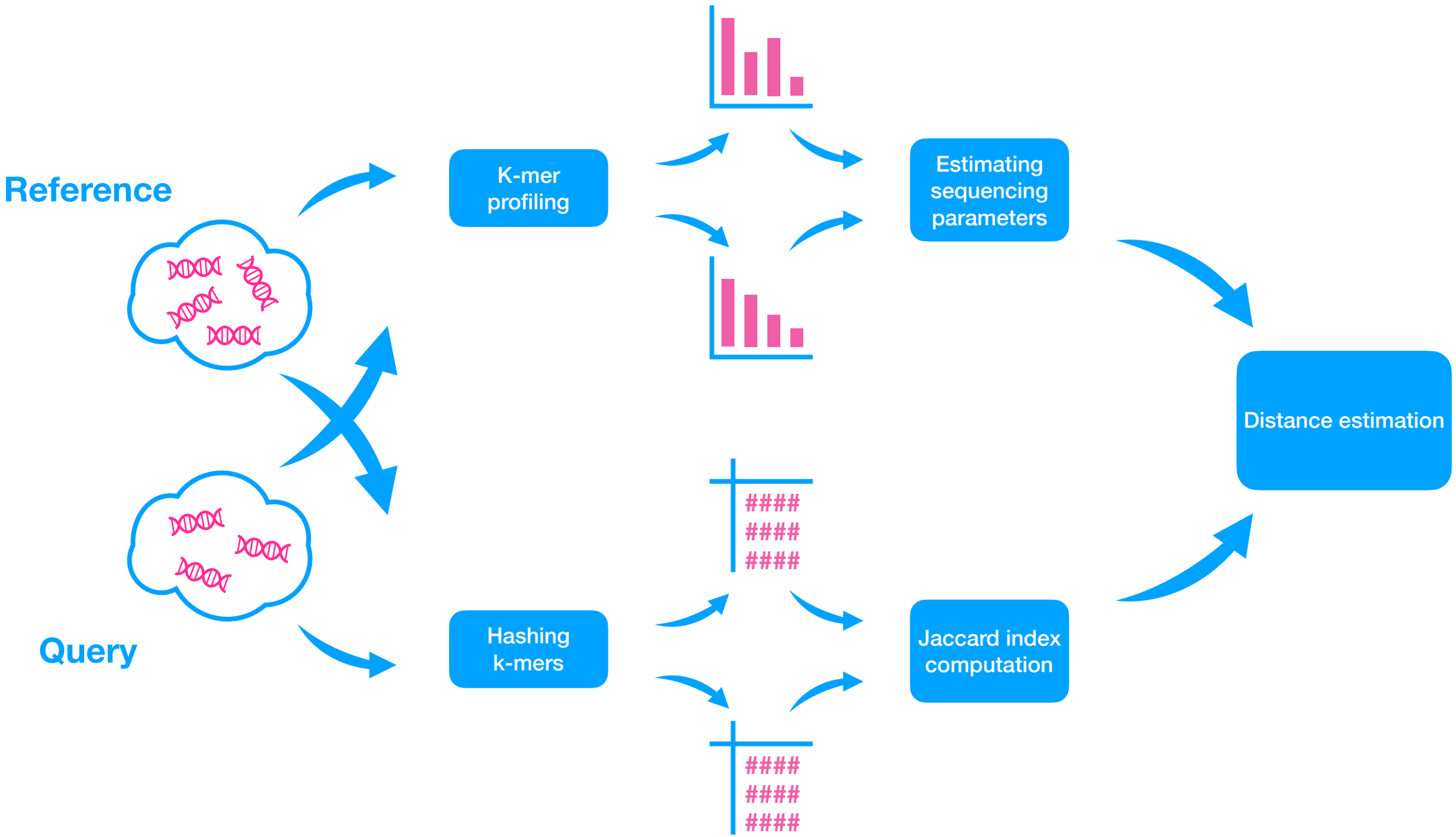
$$\mathbb{E}(M_i) = f(\lambda, \epsilon, L)$$

Coverage

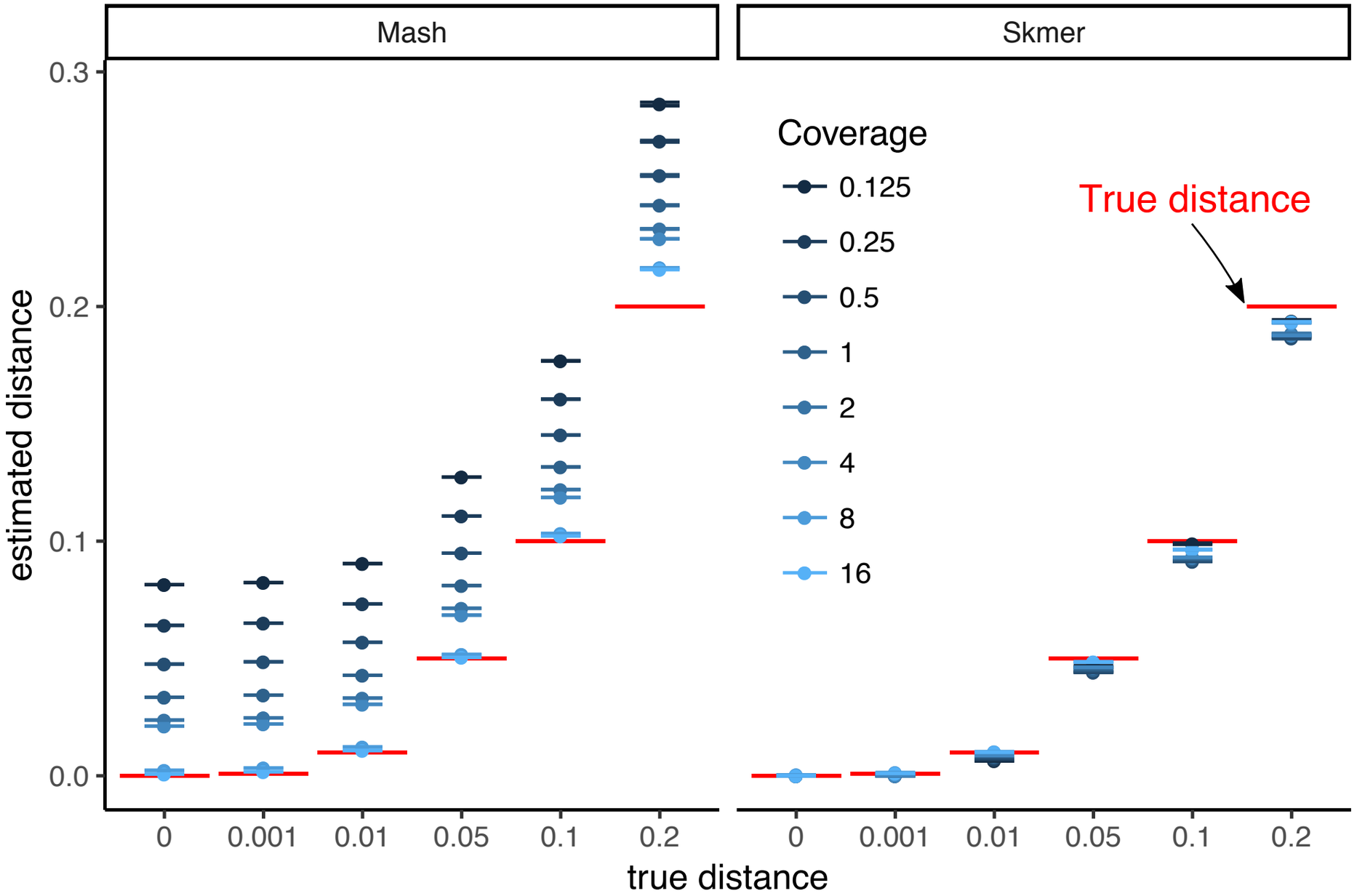
Error rate

Genome length

Skmer


Reference
Query
K-mer profiling
Hashing k-mers
Estimating sequencing parameters
Jaccard index computation
Distance estimation
####
####
####
####
####
####

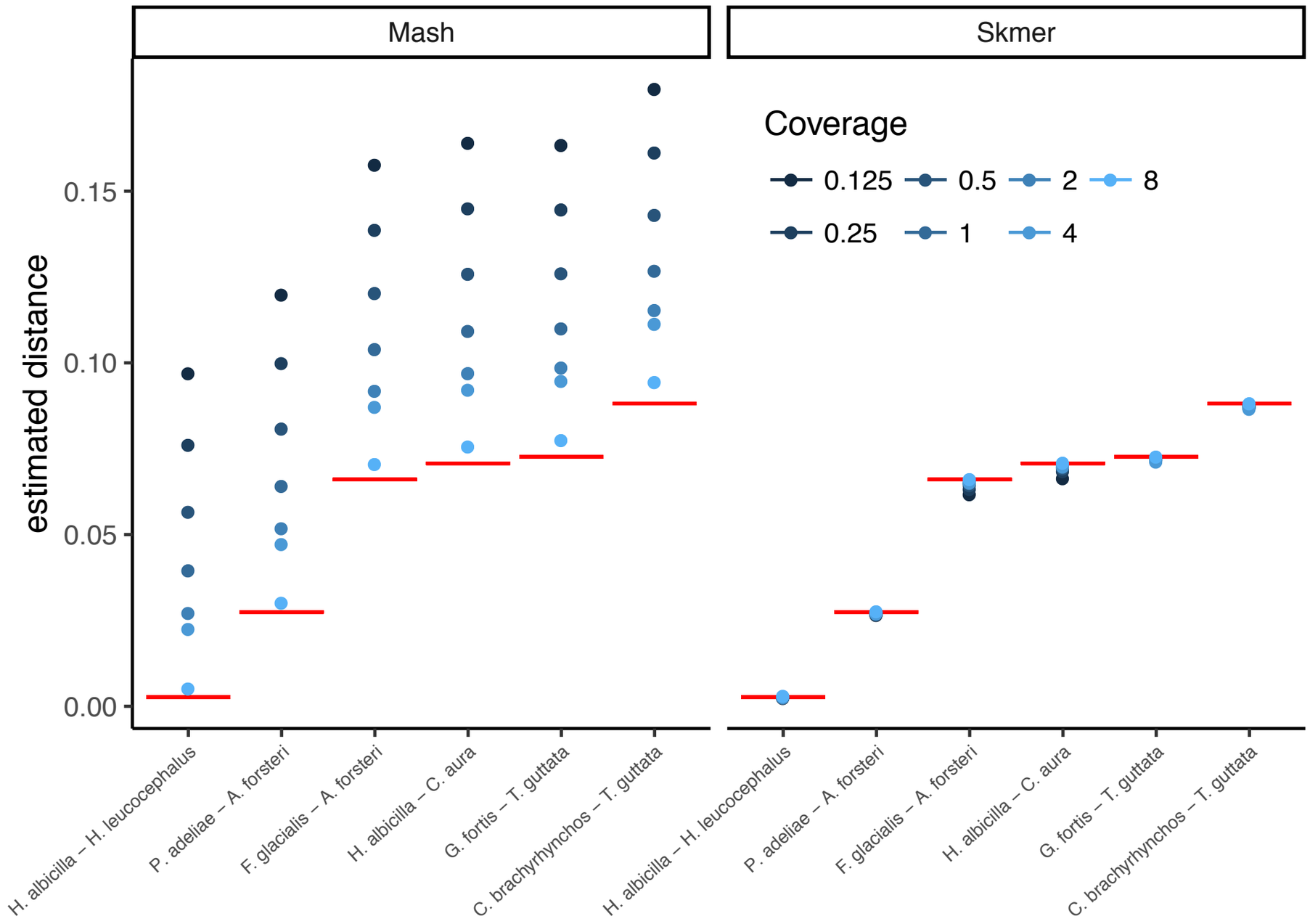
Skmer accuracy: simulations



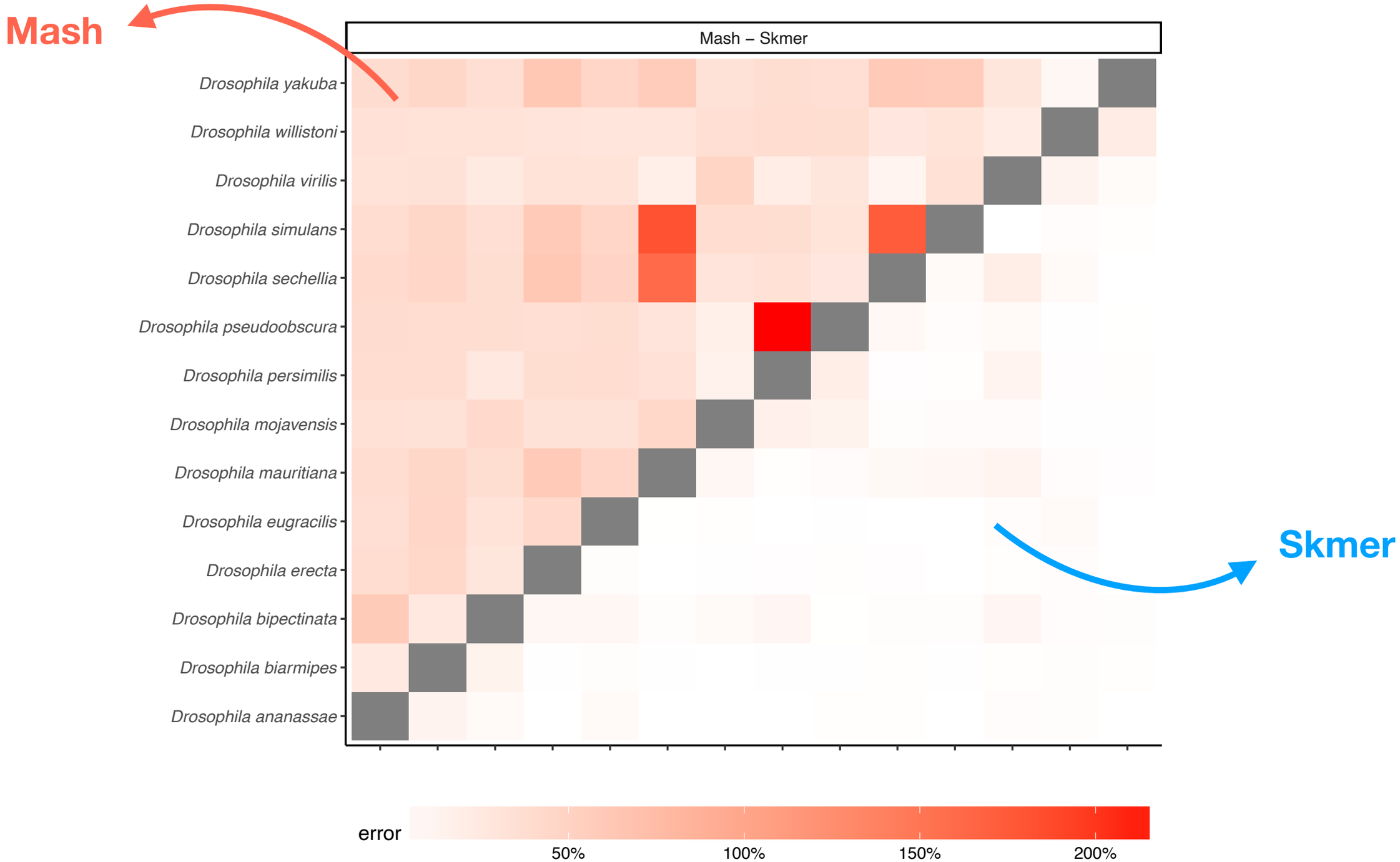
Random mutations added to an insect genome

Skmer accuracy: real data

Real pairs of bird genomes

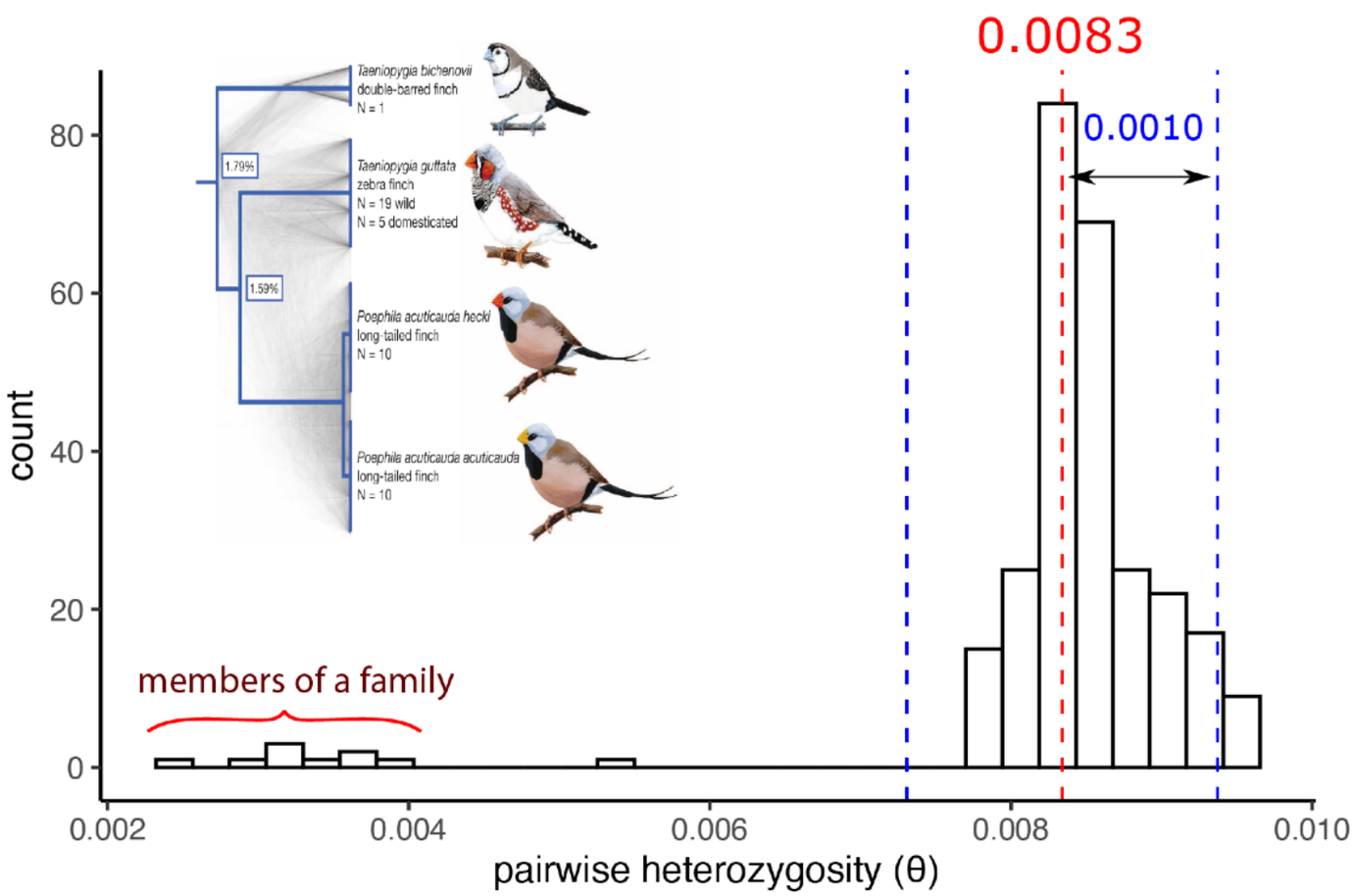


Results on real skims (~1/2X)



Stable recombination hotspots in birds

Sonal Singhal,[1,2]*† Ellen M. Leffler,[3,4]* Keerthi Sannareddy,[3] Isaac Turner,[4] Oliver Venn,[4] Daniel M. Hooper,[5] Alva I. Strand,[1] Qiye Li,[6] Brian Raney,[7] Christopher N. Balakrishnan,[8] Simon C. Griffith,[9] Gil McVean,[4] Molly Przeworski[1,2]†



Autosomal nucleotide diversity	Assembly	Skmer (2X)	Skmer+ RESPECT (2X)
Zebra finch	0.82%	0.69%	0.83%
Long-tailed finch	0.55%	0.43%	0.55%

[Unpublished work]

Research articles

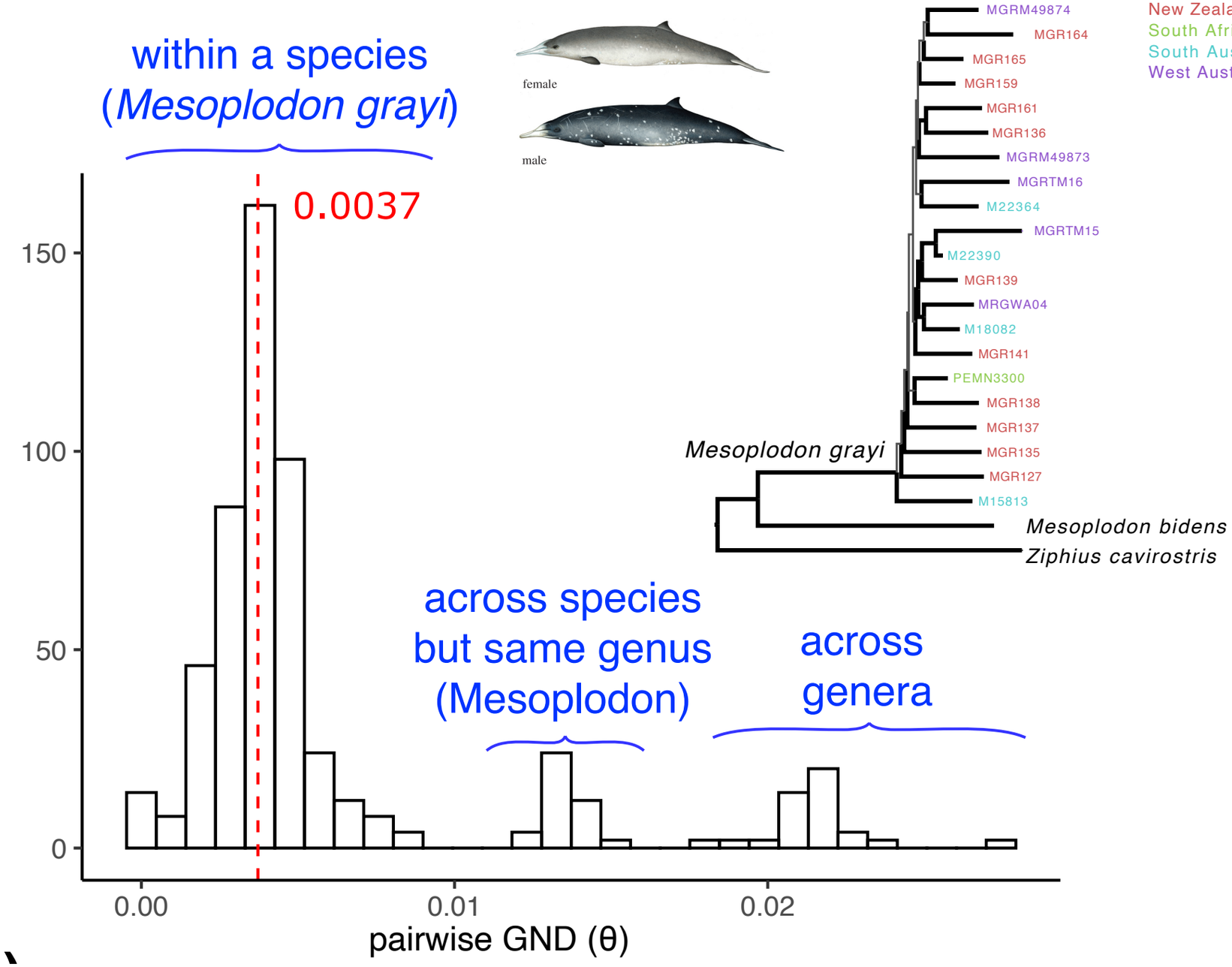
Ocean-wide genomic variation in Gray's beaked whales, *Mesoplodon grayi*

M. V. Westbury†, K. F. Thompson† ☐, M. Louis, A. A. Cabrera, M. Skovrind, J. A. S. Castruita, R. Constantine, J. R. Stevens and E. D. Lorenzen

Published: 24 March 2021 | https://doi.org/10.1098/rsos.201788

- Skims from 19 individuals of *Mesoplodon grayi* with coverage: 0.3X - 0.6X

- Two assemblies of closely-related beaked whale species (NCBI)

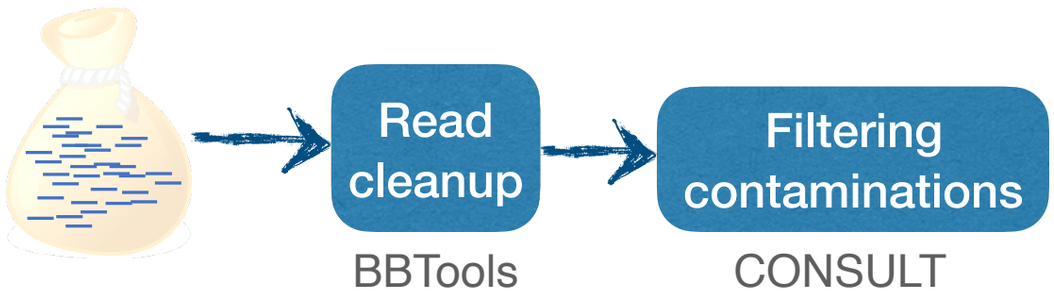


[Unpublished work]

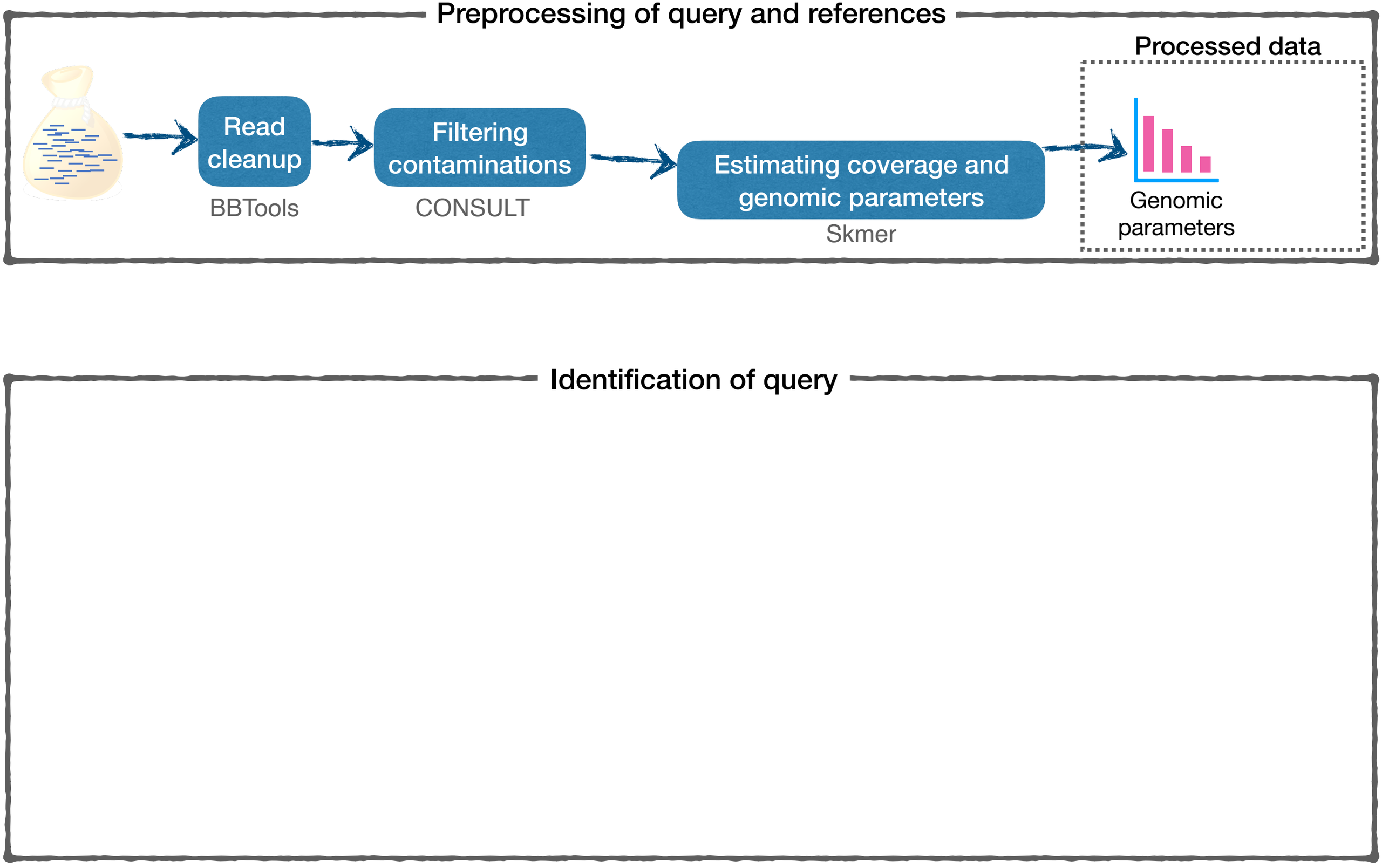
Preprocessing of query and references

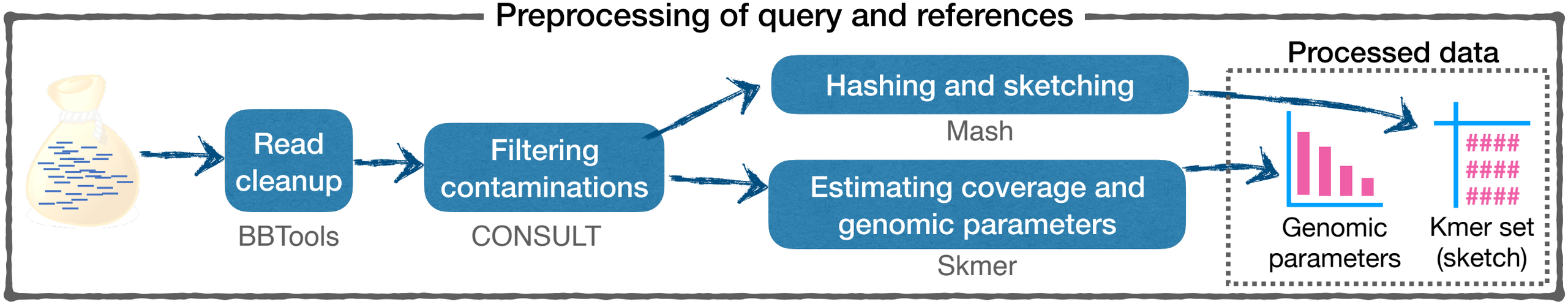
Identification of query

Preprocessing of query and references

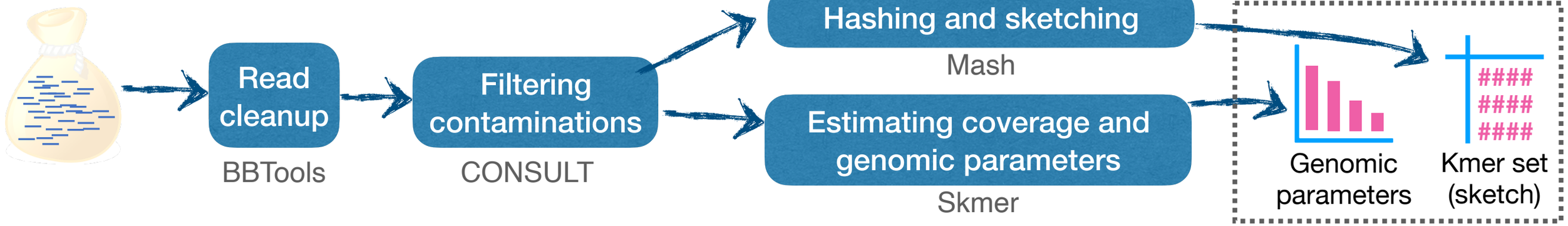


Identification of query

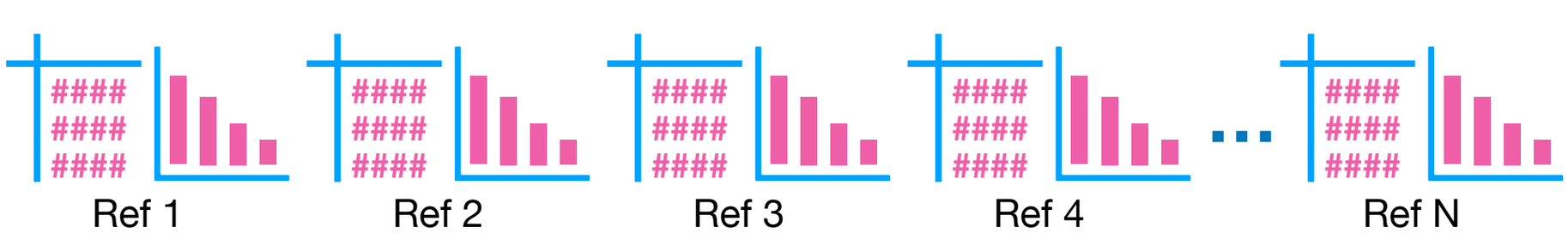

Preprocessing of query and references
Processed data
Read cleanup
BBTools
Filtering contaminations
CONSULT
Estimating coverage and genomic parameters
Skmer
Genomic parameters
Identification of query


Preprocessing of query and references
Processed data
Read cleanup
BBTools
Filtering contaminations
CONSULT
Hashing and sketching
Mash
Estimating coverage and genomic parameters
Skmer
Genomic parameters
Kmer set (sketch)
####
Identification of query

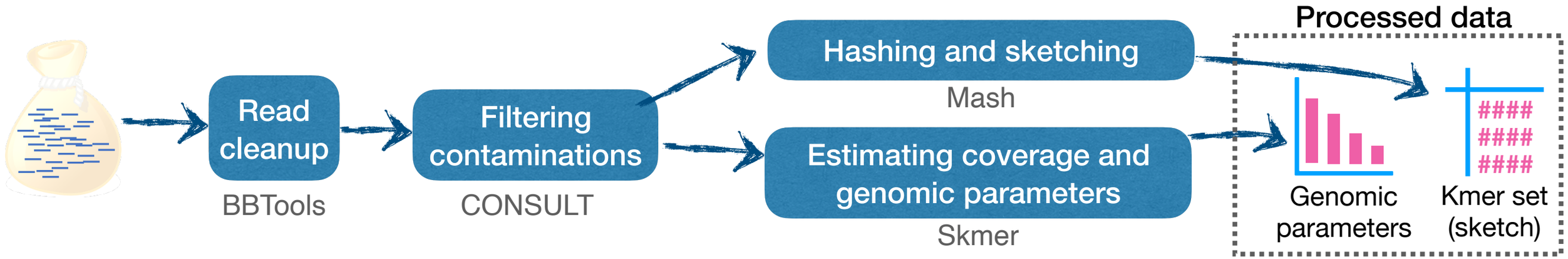
Preprocessing of query and references



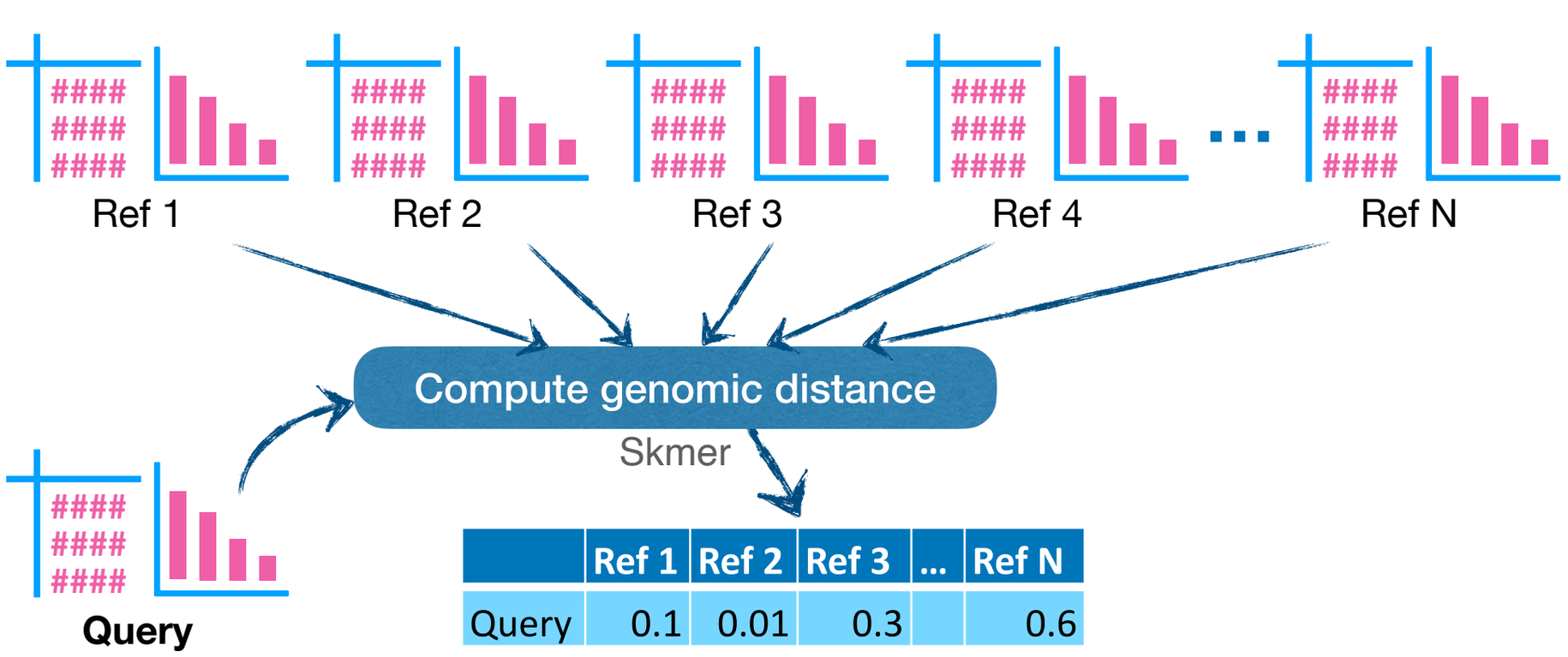
Identification of query



Preprocessing of query and references



Identification of query



	Ref 1	Ref 2	Ref 3	…	Ref N
Query	0.1	0.01	0.3		0.6

Skmer

https://github.com/shahab-sarmashghi/Skmer

```
$ conda install skmer
```

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1632-4

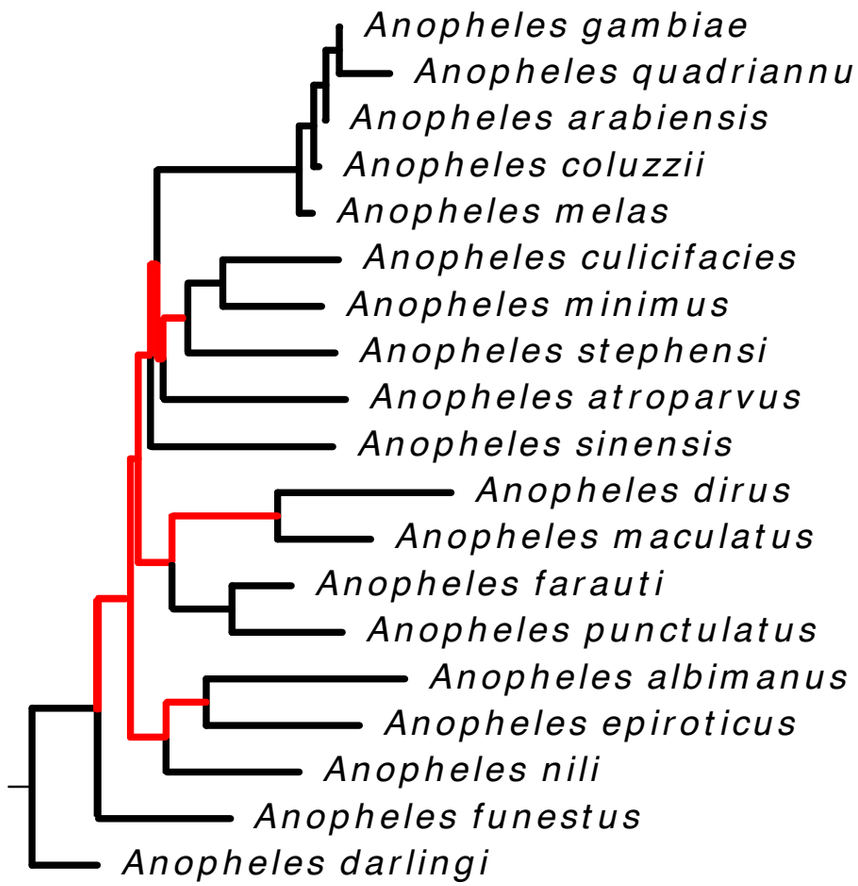
Distance-based phylogenetics

1. Compute all pairwise distance (e.g., GND)
   - $D_{i,j}$: distance of species i and j
2. Correct according to models of sequence evolution
   - Example: JC69 model:

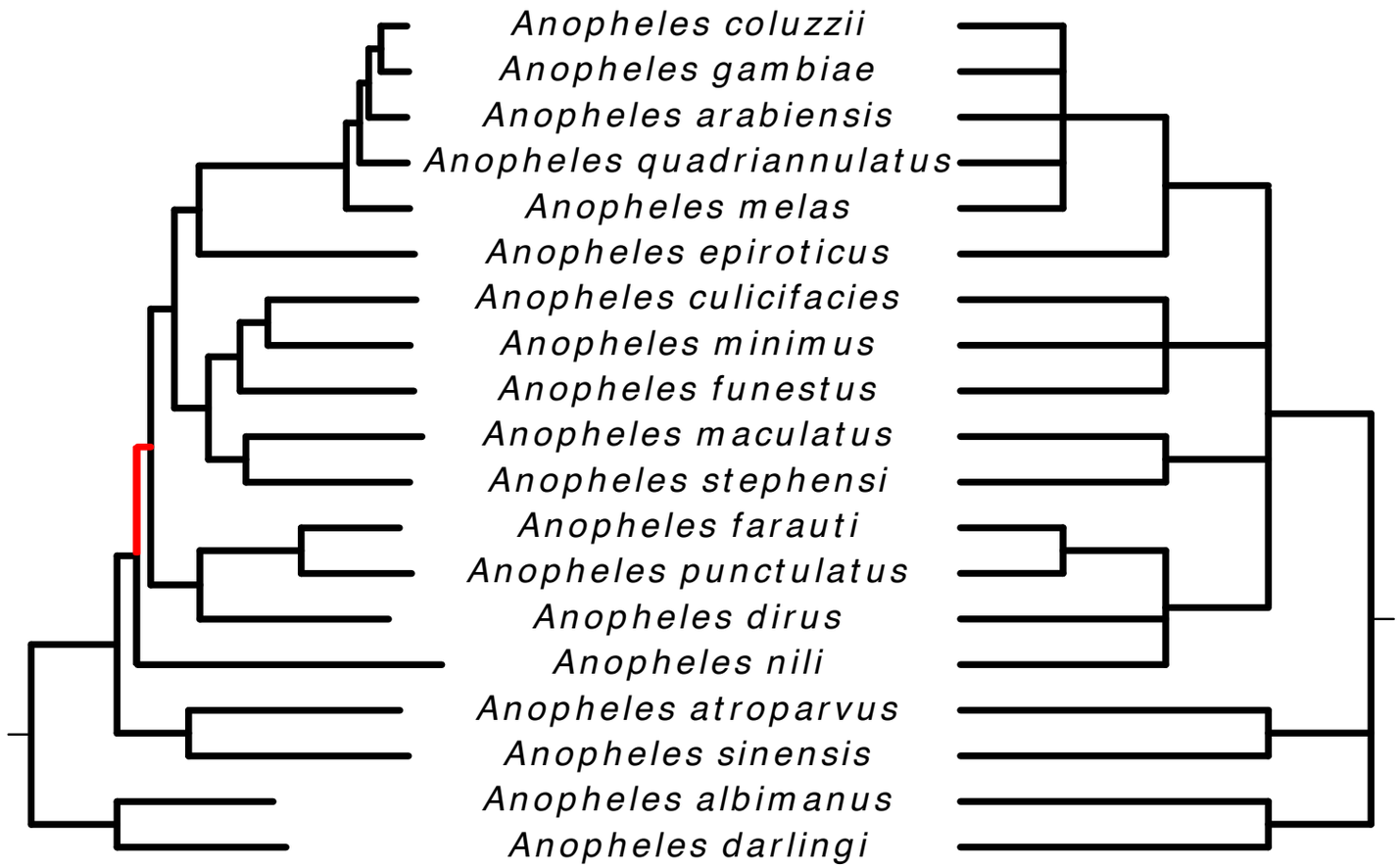
$$t_{i,j} = -\frac{3}{4}\ln(1 - \frac{4}{3}D_{i,j})$$

3. Build the phylogeny using an algorithm of choice
   - Neighbour joining, Minimum Evolution, etc.

Building Phylogeny



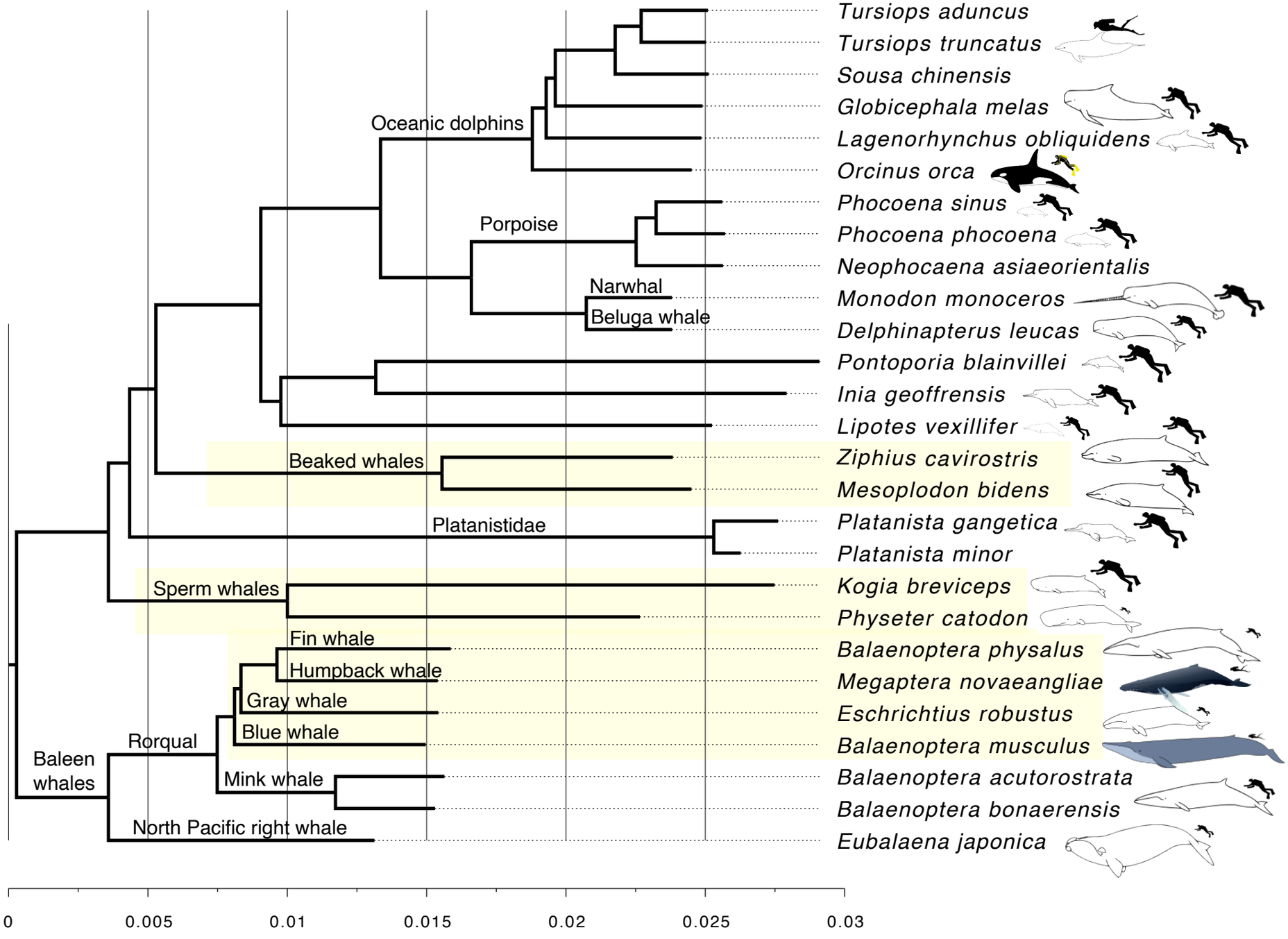
Marker Phylogeny



Skmer distances+ JC69 correction

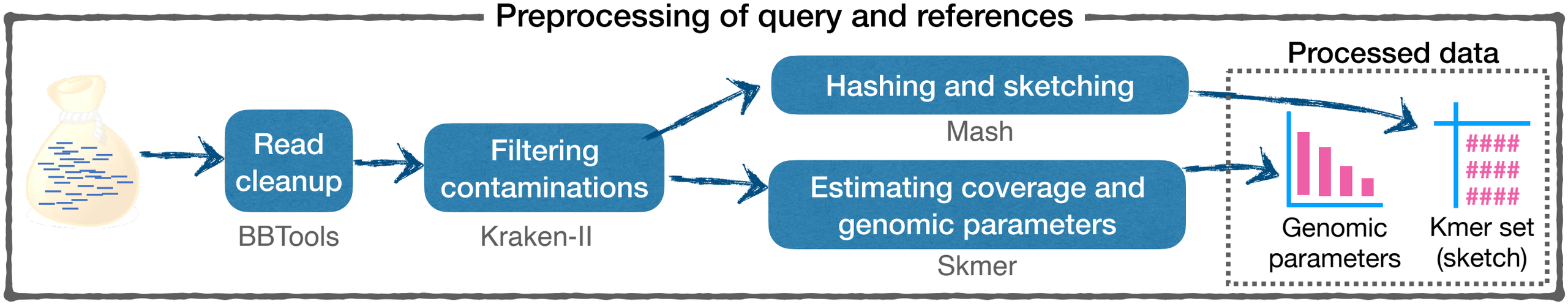
oTL Reference

Cetacean phylogeny

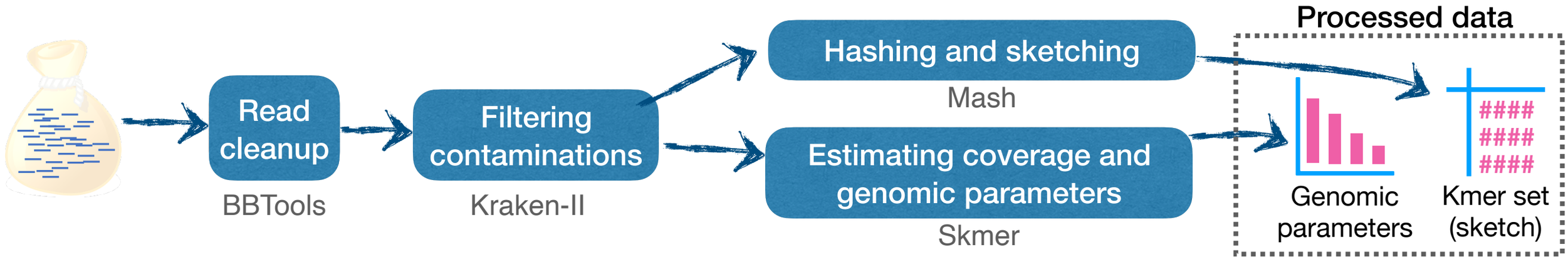


Good tools to use

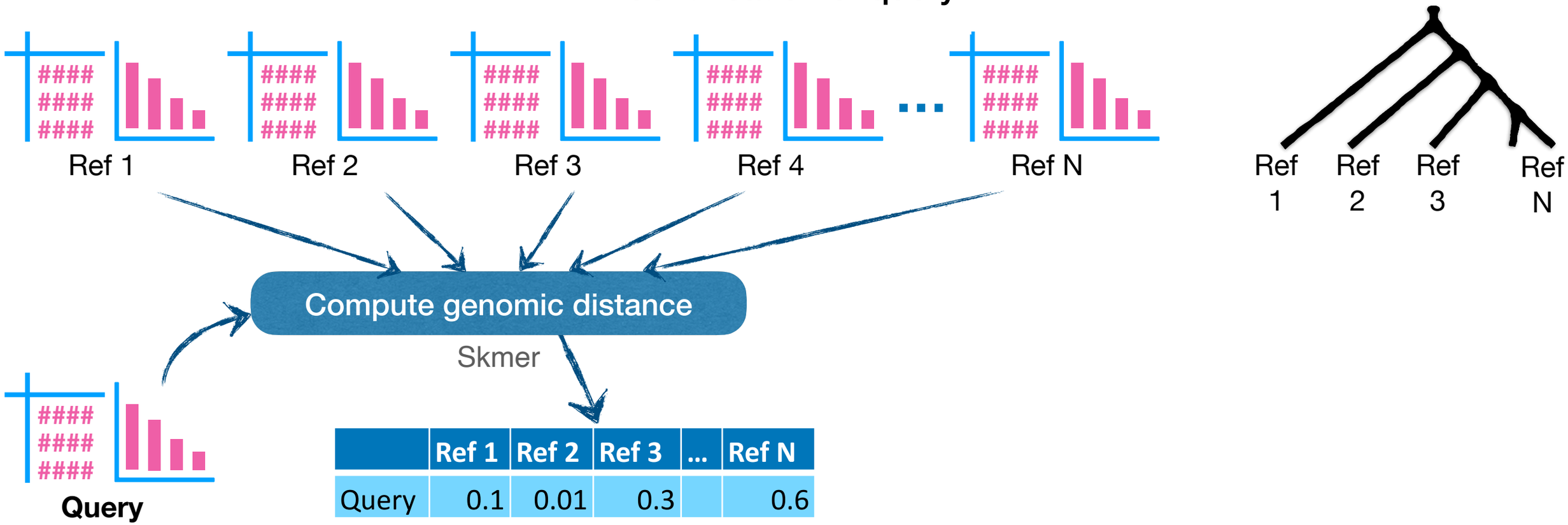
- FastME is my favorite


Preprocessing of query and references
Read cleanup
BBTools
Filtering contaminations
Kraken-II
Hashing and sketching
Mash
Estimating coverage and genomic parameters
Skmer
Processed data
Genomic parameters
Kmer set (sketch)
Identification of query
Ref 1
Ref 2
Ref 3
Ref 4
Ref N
Compute genomic distance
Skmer
Query
Ref 1 Ref 2 Ref 3 ... Ref N
Query 0.1 0.01 0.3 0.6

Preprocessing of query and references

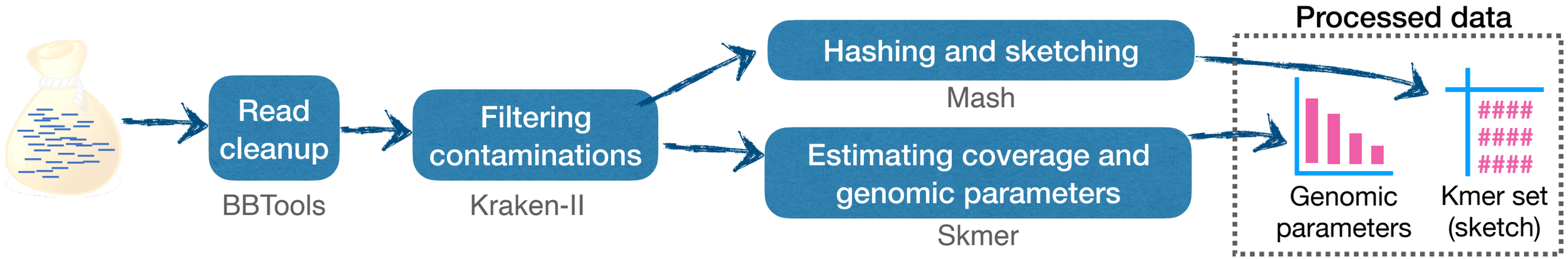


Identification of query

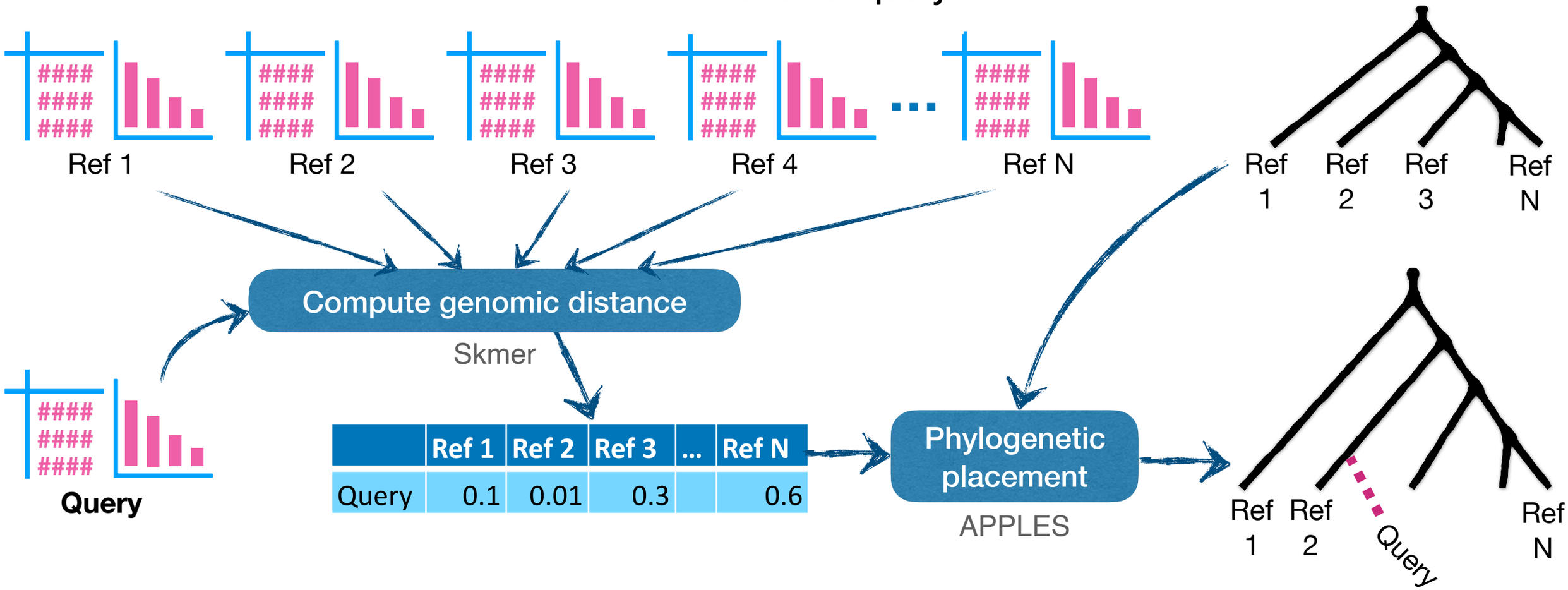


	Ref 1	Ref 2	Ref 3	...	Ref N
Query	0.1	0.01	0.3		0.6

Preprocessing of query and references

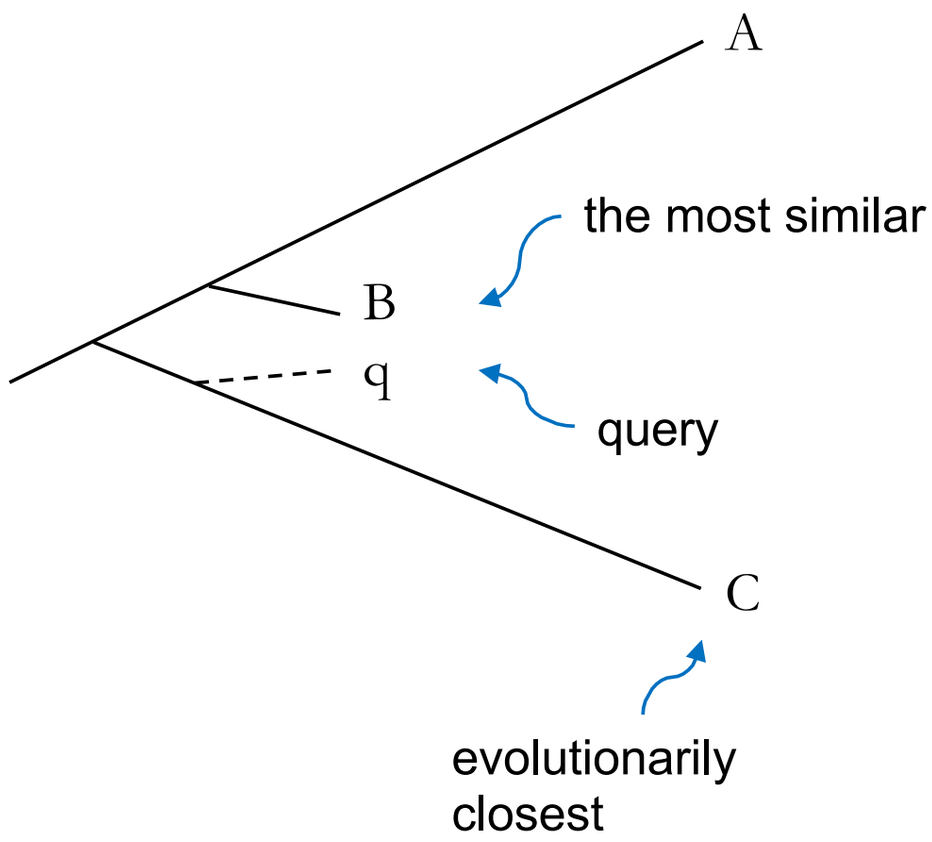


Identification of query

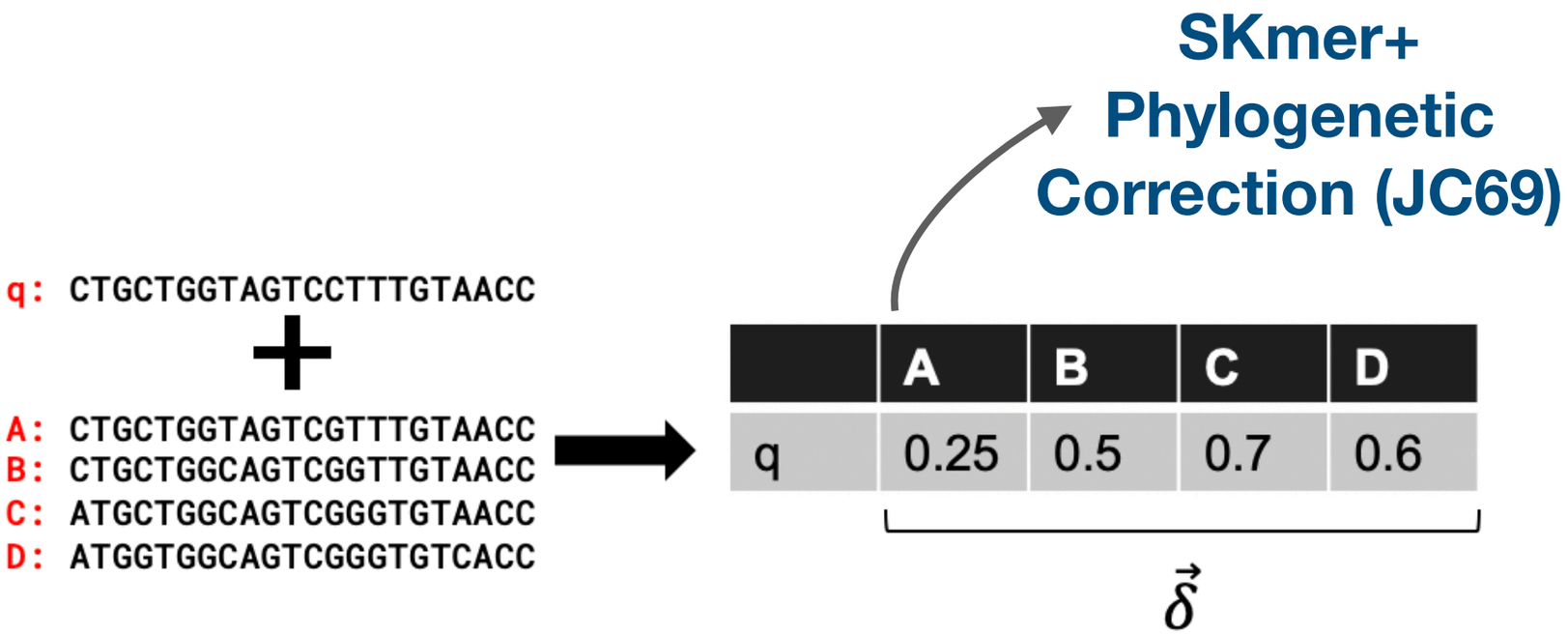


Closest match not always enough

- Alternative to placement: closest match
- For incomplete reference sets, assignment to closest is misleading
- When rate of evolution changes best match can mislead.



Distance-based Phylogenetic Placement Framework



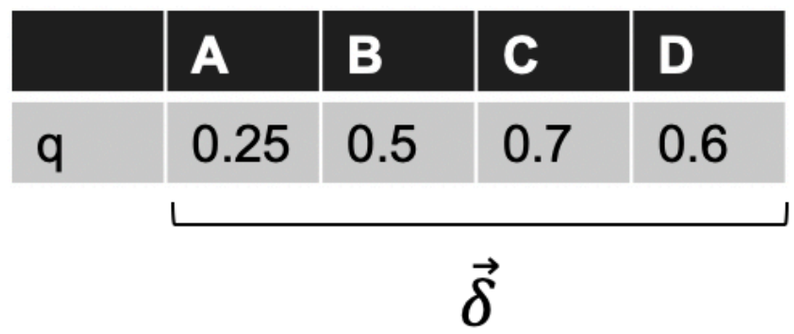
Distance-based Phylogenetic Placement Framework

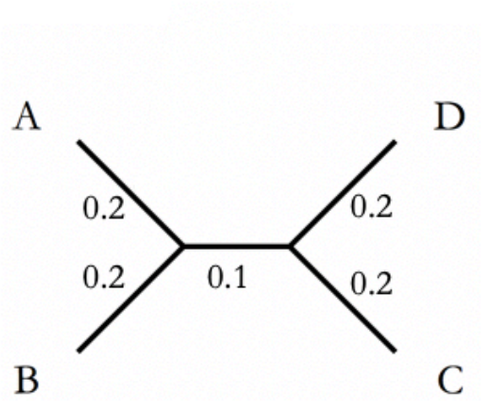
SKmer+ Phylogenetic Correction (JC69)

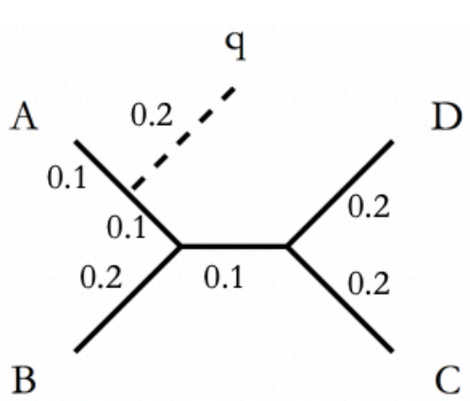
q: CTGCTGGTAGTCCTTTGTAACC

\+

A: CTGCTGGTAGTCGTTTGTAACC
B: CTGCTGGCAGTCGGTTGTAACC
C: ATGCTGGCAGTCGGGTGTAACC
D: ATGGTGGCAGTCGGGTGTCACC



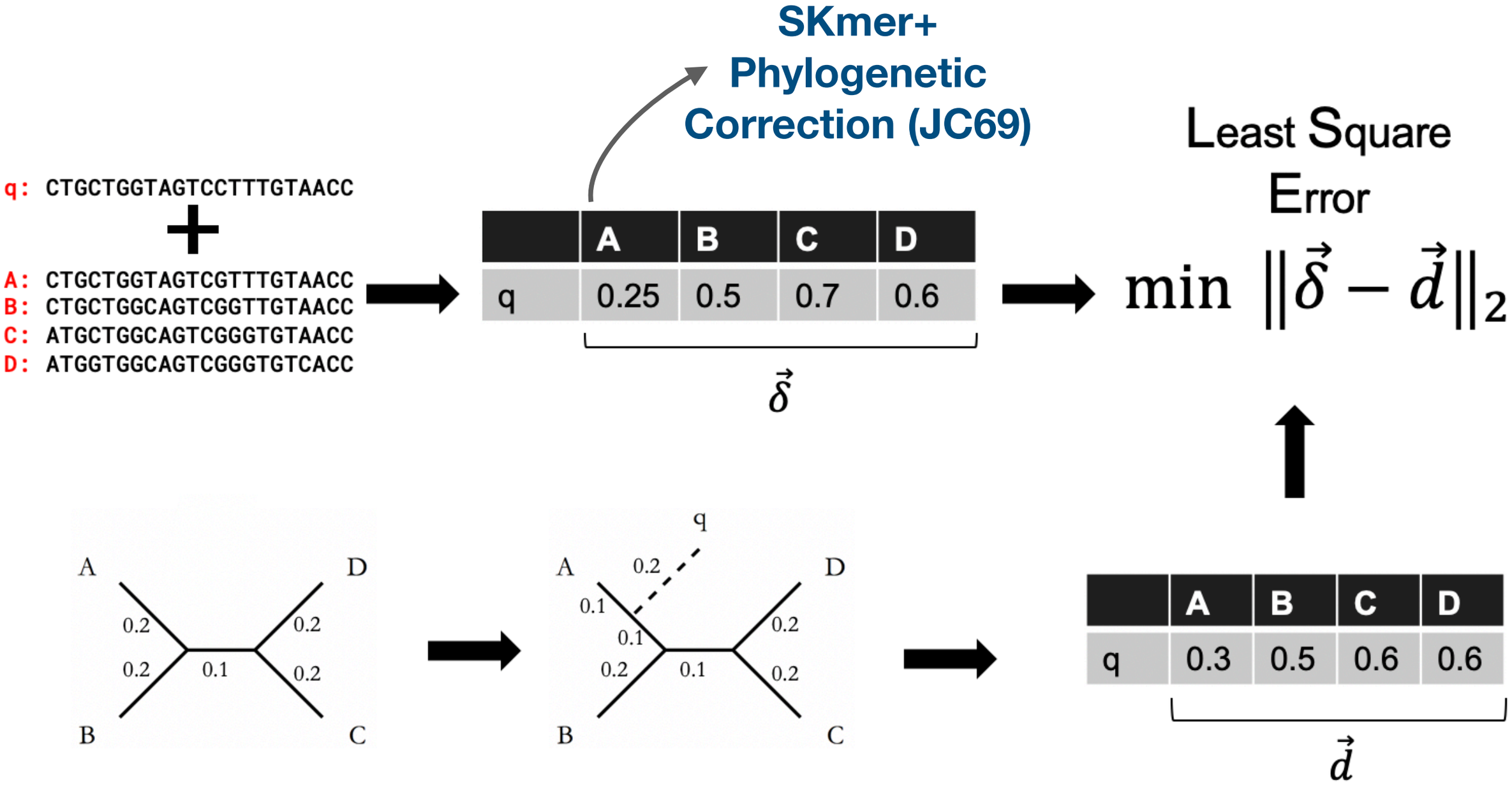




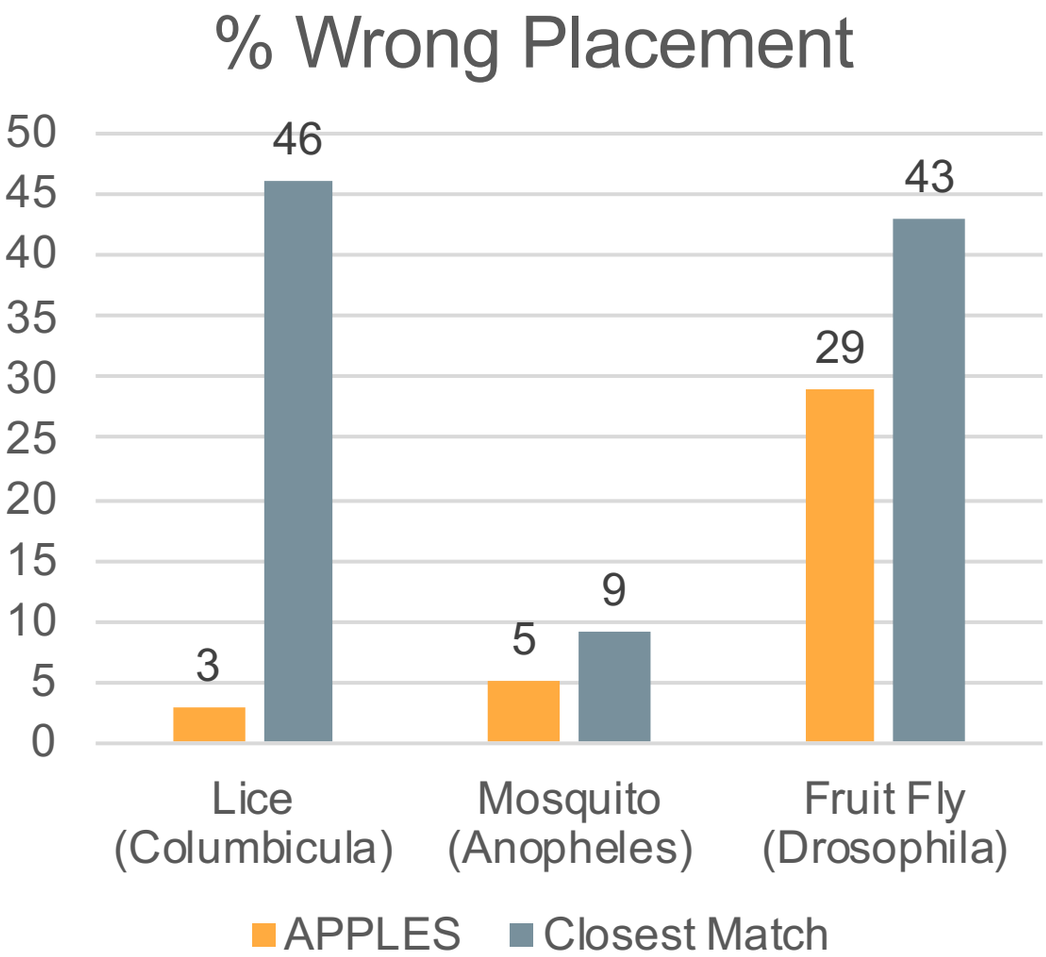
	A	B	C	D
q	0.3	0.5	0.6	0.6

$\vec{d}$

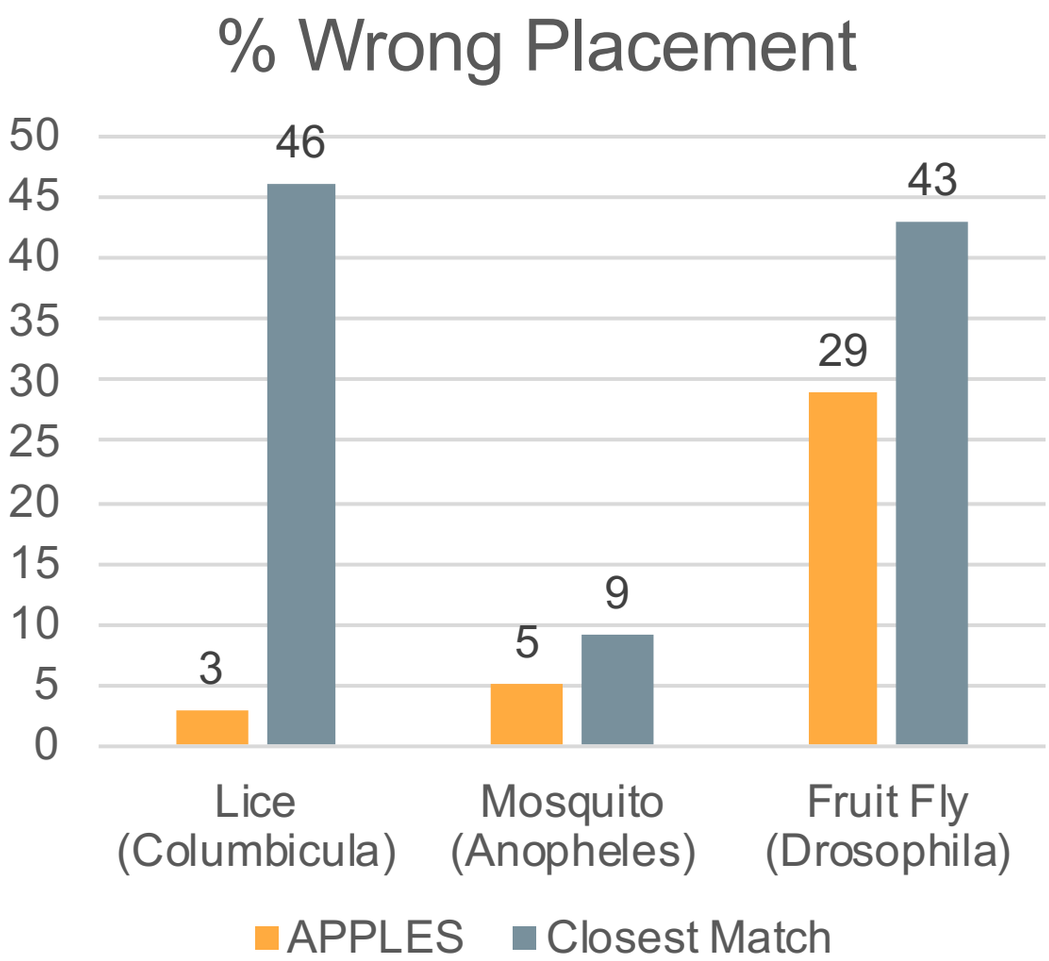
Distance-based Phylogenetic Placement Framework

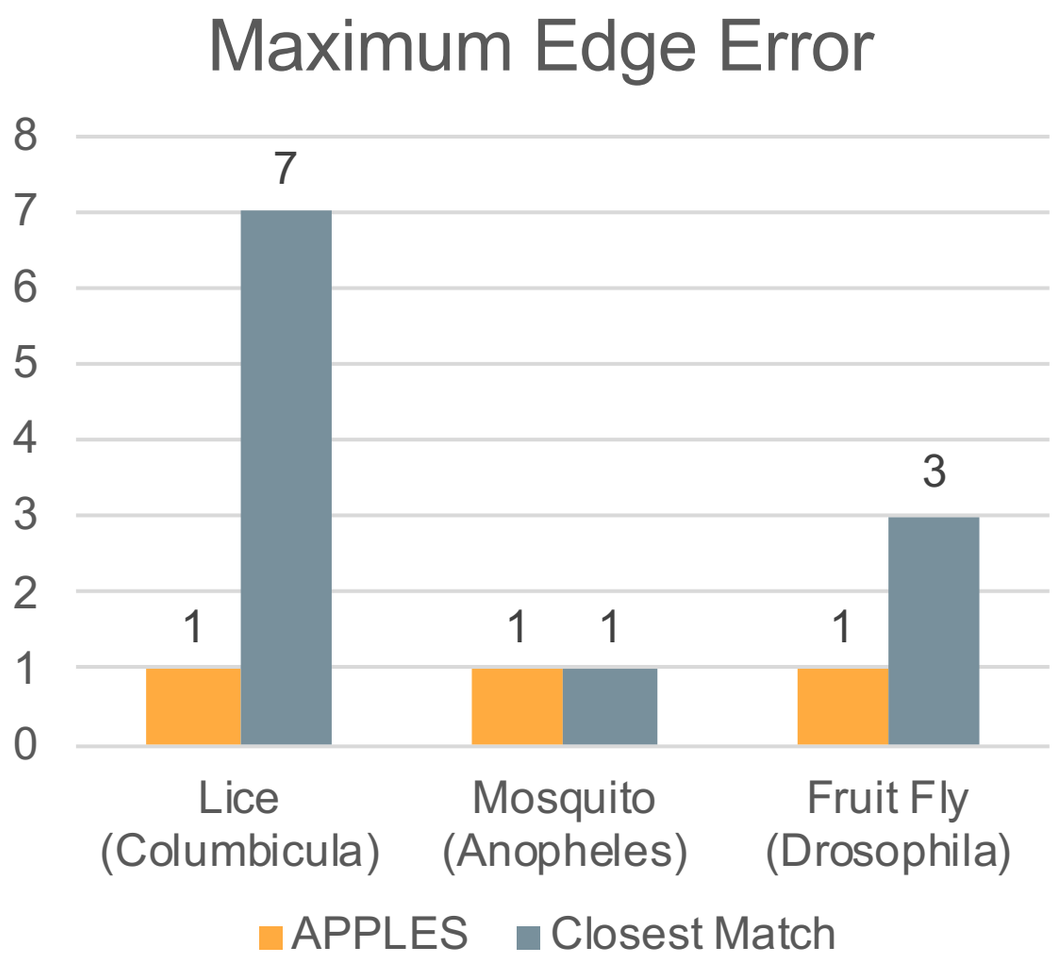


Placement accuracy results



Placement accuracy results





APPLES

https://github.com/balabanmetin/apples

```
$ pip install apples
```

Systematic Biology

https://doi.org/10.1093/sysbio/syz063

MOLECULAR ECOLOGY

NEWS AND VIEWS | Open Access

Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification

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

First published: 16 June 2020 | https://doi.org/10.1111/mec.15507

Publications

- Bohmann, Kristine, Siavash Mirarab, Vineet Bafna, and M. Thomas P. Gilbert. “Beyond DNA Barcoding: The Unrealized Potential of Genome Skim Data in Sample Identification.” Molecular Ecology 29, no. 14 (2020): 2521–34. https://doi.org/10.1111/mec.15507.

- **Skmer:** Sarmashghi, Shahab, Kristine Bohmann, M Thomas P Gilbert, Vineet Bafna, and Siavash Mirarab. “Skmer: Assembly-Free and Alignment-Free Sample Identification Using Genome Skims.” Genome Biology 20, no. 1 (2019): 34. https://doi.org/10.1186/s13059-019-1632-4.

- **Contamination:** Rachtman, Eleonora, Metin Balaban, Vineet Bafna, and Siavash Mirarab. “The Impact of Contaminants on the Accuracy of Genome Skimming and the Effectiveness of Exclusion Read Filters.” Molecular Ecology Resources 20, no. 3 (2020): 1755-0998.13135. https://doi.org/10.1111/1755-0998.13135.

- **APPLES:** Balaban, Metin, Shahab Sarmashghi, and Siavash Mirarab. “APPLES: Scalable Distance-Based Phylogenetic Placement with or without Alignments.” Edited by David Posada. Systematic Biology 69, no. 3 (2020): 566–78. https://doi.org/10.1093/sysbio/syz063

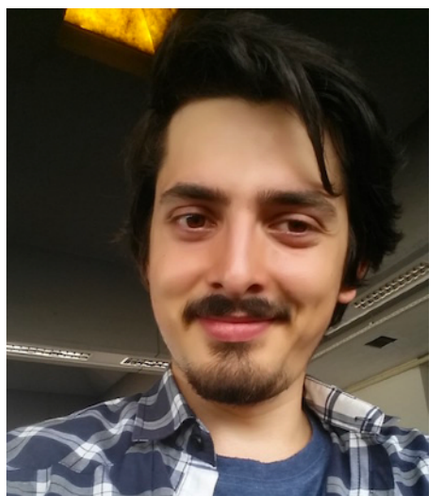
- **Mixed samples:** Balaban, Metin, and Siavash Mirarab. “Phylogenetic Double Placement of Mixed Samples.” Bioinformatics 36, no. Supplement_1 (2020): i335–43. https://doi.org/10.1093/bioinformatics/btaa489.

Shahab Sarmashghi

Vineet Bafna

Kristine Bohmann

Tom Gilbert



Metin Balaban

Eleonora (Nora) Rachtman

1815485 and 1565862