

Biological Problem: Global Biodiveristy Decline

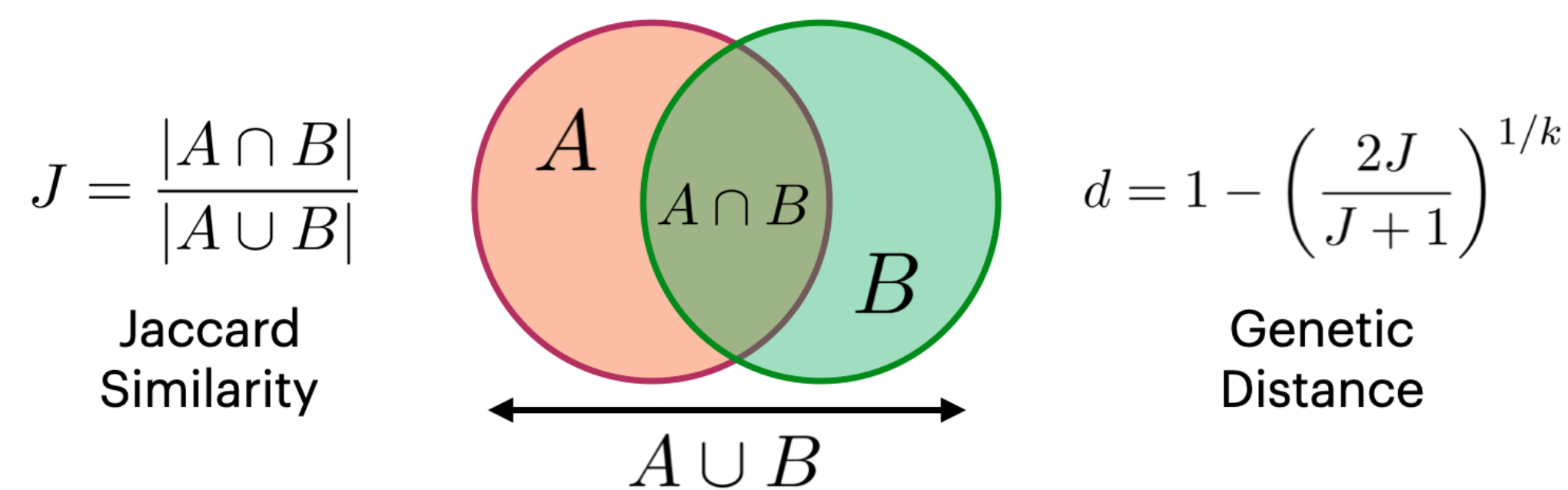
- World Wildlife Foundation: **73% decline in biodiversity (1970-2022)**
- Genetic diversity** is a powerful way to monitor biodiversity, but species must be sampled across time to be robust (**expensive**).



- Genome Skimming:** Low-coverage sequencing is an affordable way to get data, but only obtains **sparse** and **error-prone** reads.

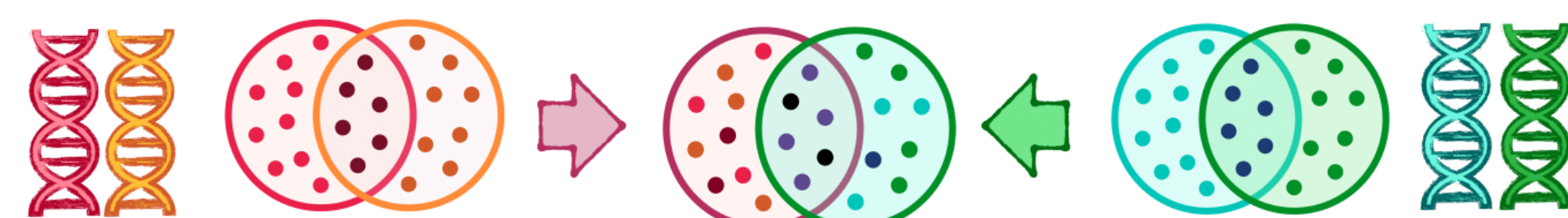


- Genomes can be represented by a collection of small sequences of length 'k' called **k-mers**.
- Methods (e.g., *Skmer*, *Mash*) exist that estimate genomic diversity between genome skims using **Jaccard Similarity** between sets of k-mers.

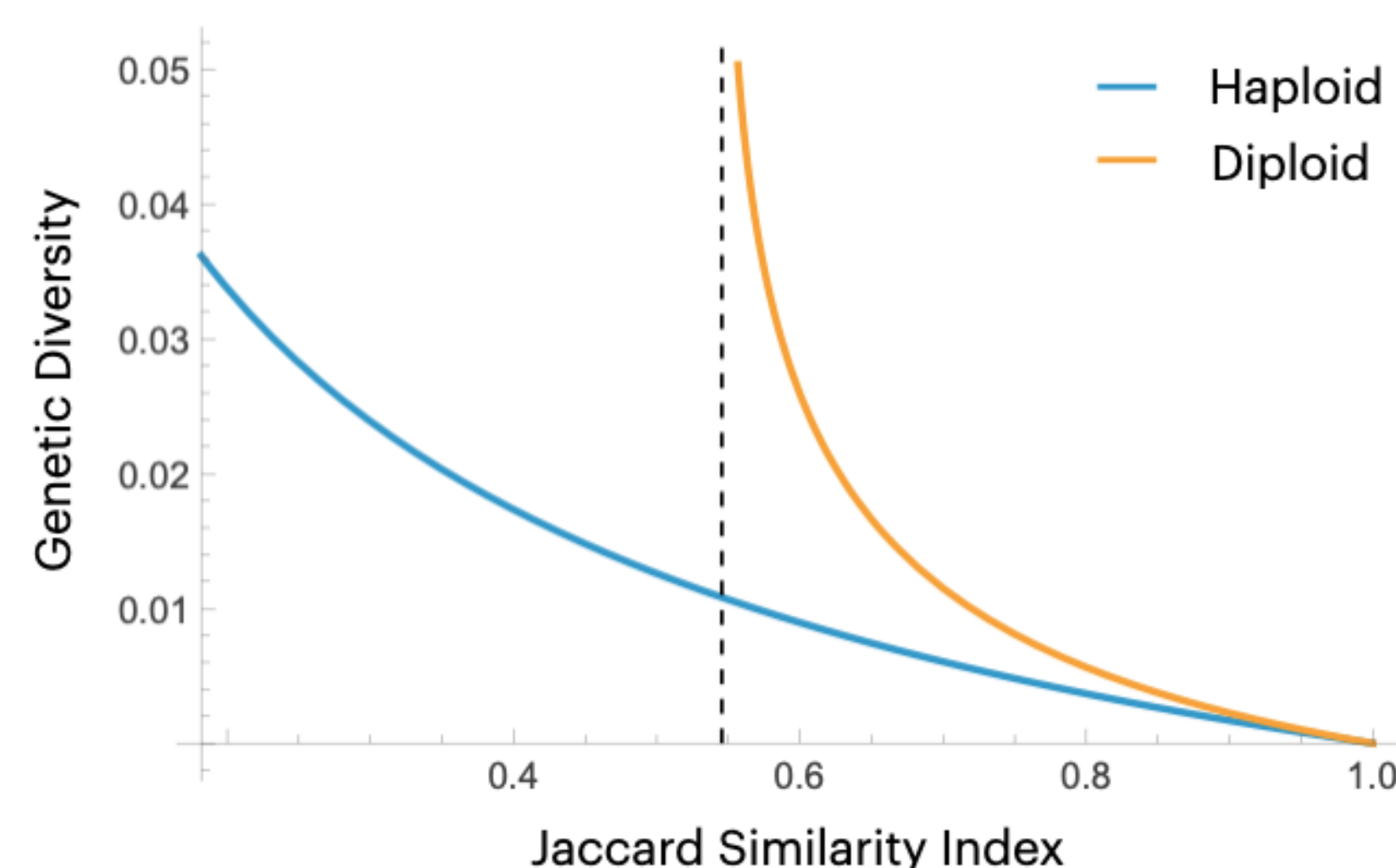


Challenge: Diploid Genetic Diversity via Genome Skims

- Many populations of conservation interest are **diploid** (i.e., have 2 sets of chromosomes), and existing methods do not model this.

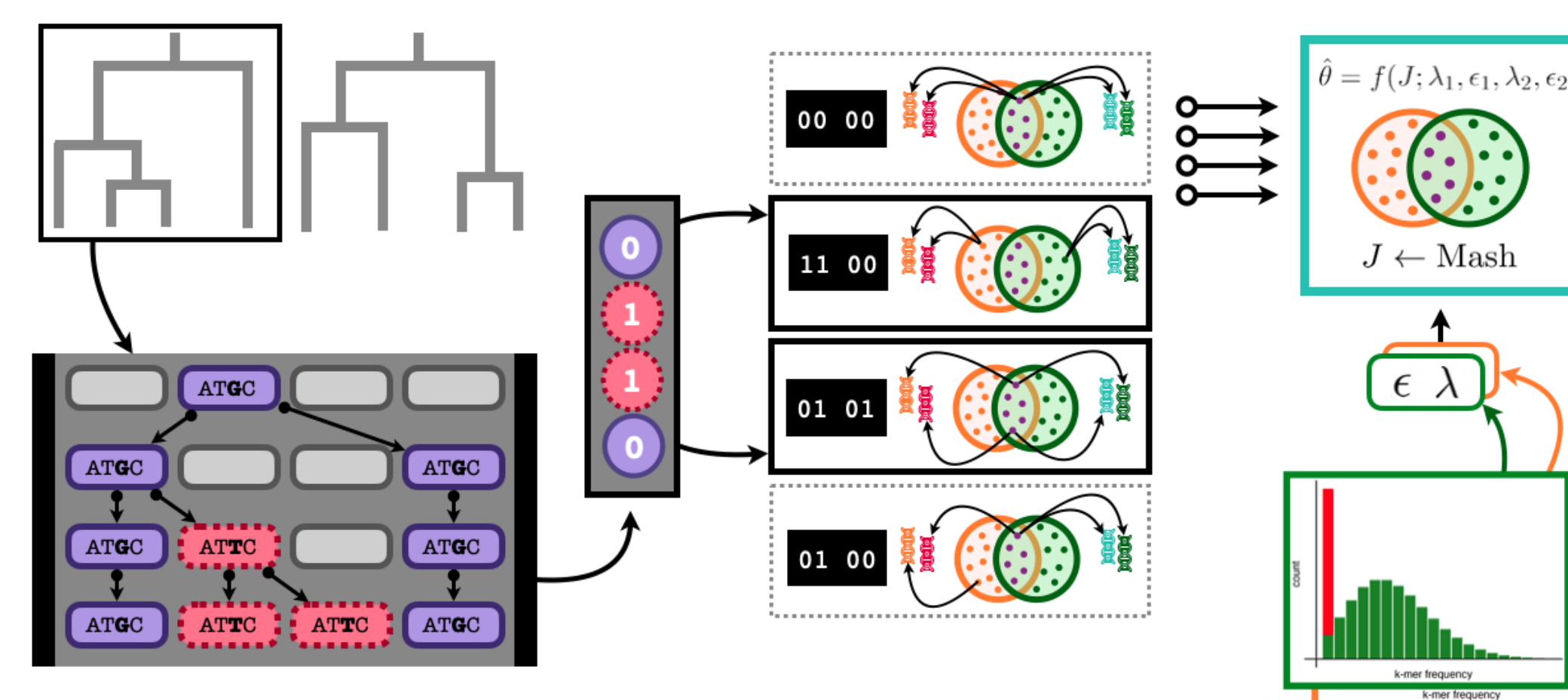


- Not modeling diploidy can lead to **incorrect estimates of biodiversity**.

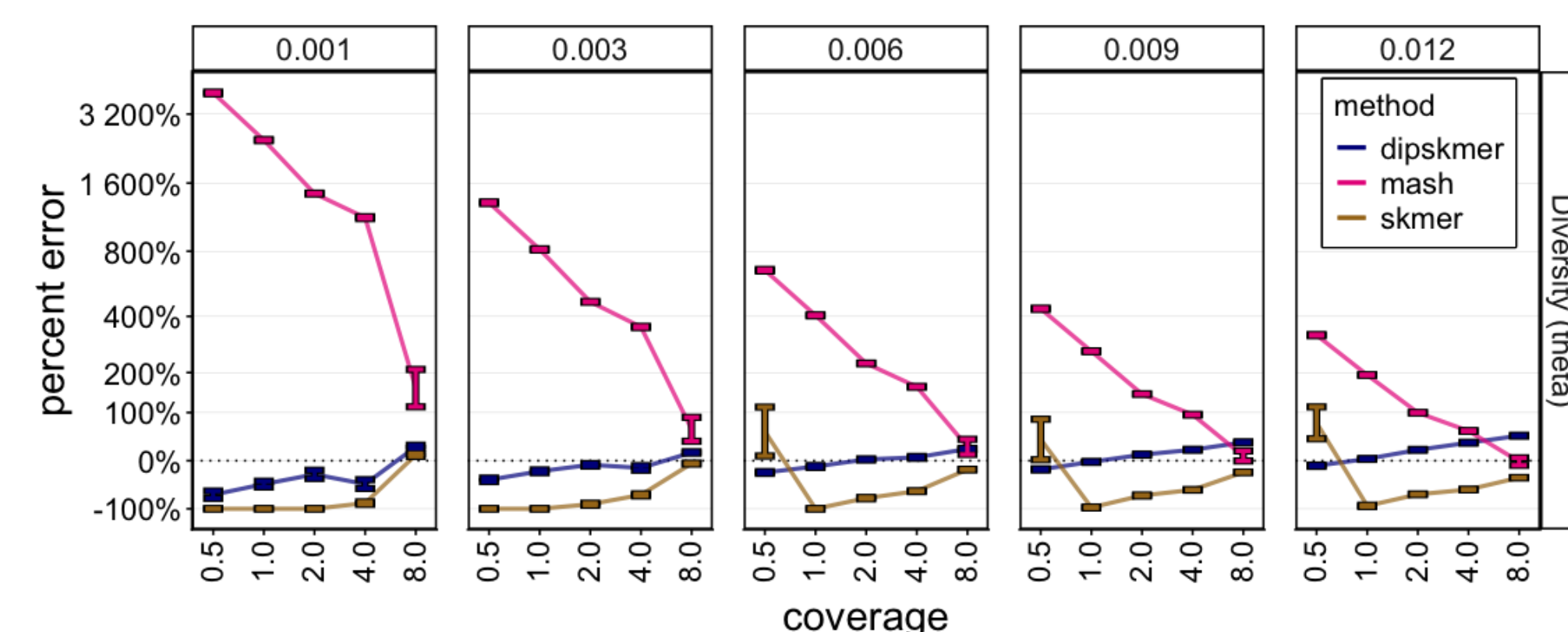


Modeling Diploid Evolution's Impact on k-mer Similarity

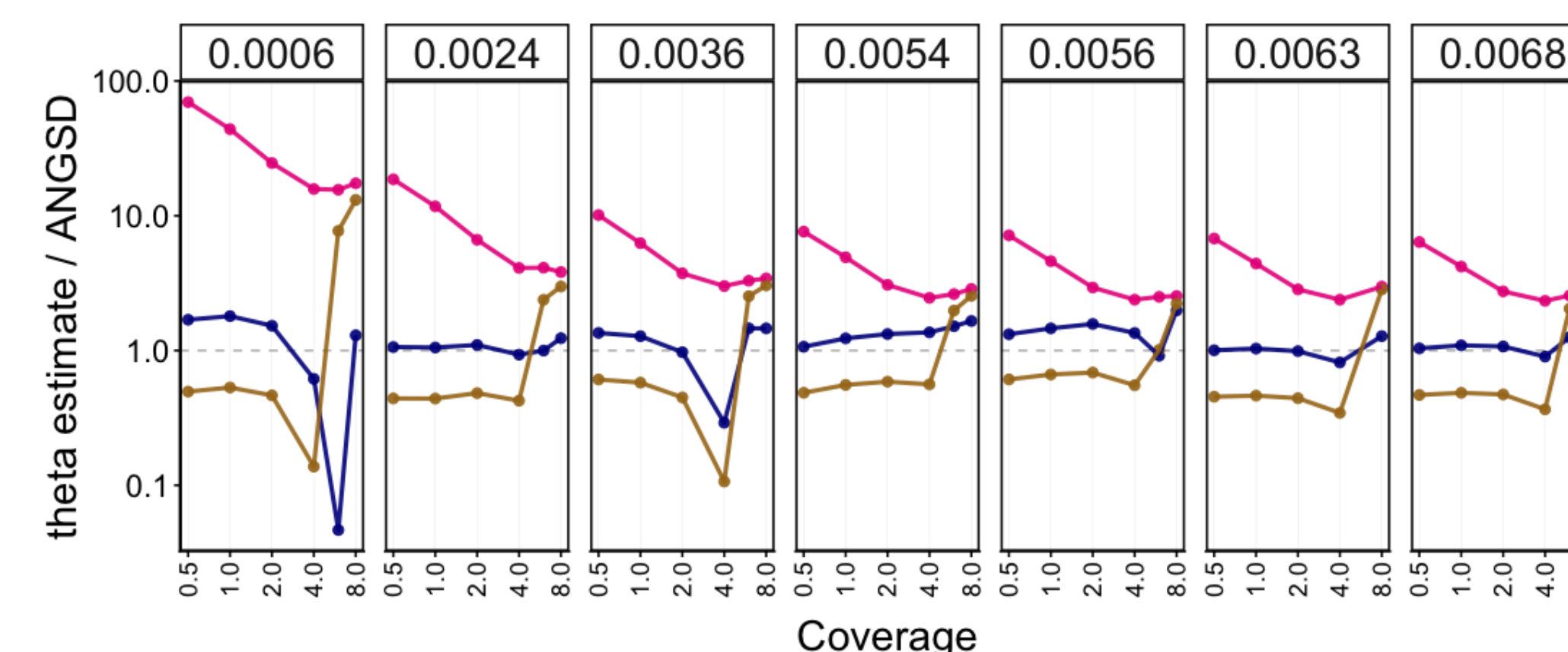
- Probability model for k-mer tree topologies across 4 chromosomes.
- Mutation probability model along a k-mer genealogy.
- Independent sorting probability of k-mers into samples.
- Coverage and sequencing error confound observed Jaccard Index.
- Full model integration to estimate genetic diversity ($\hat{\theta}$) from J , λ , and ϵ .



DipSkmer Estimates Genetic Diversity Accurately



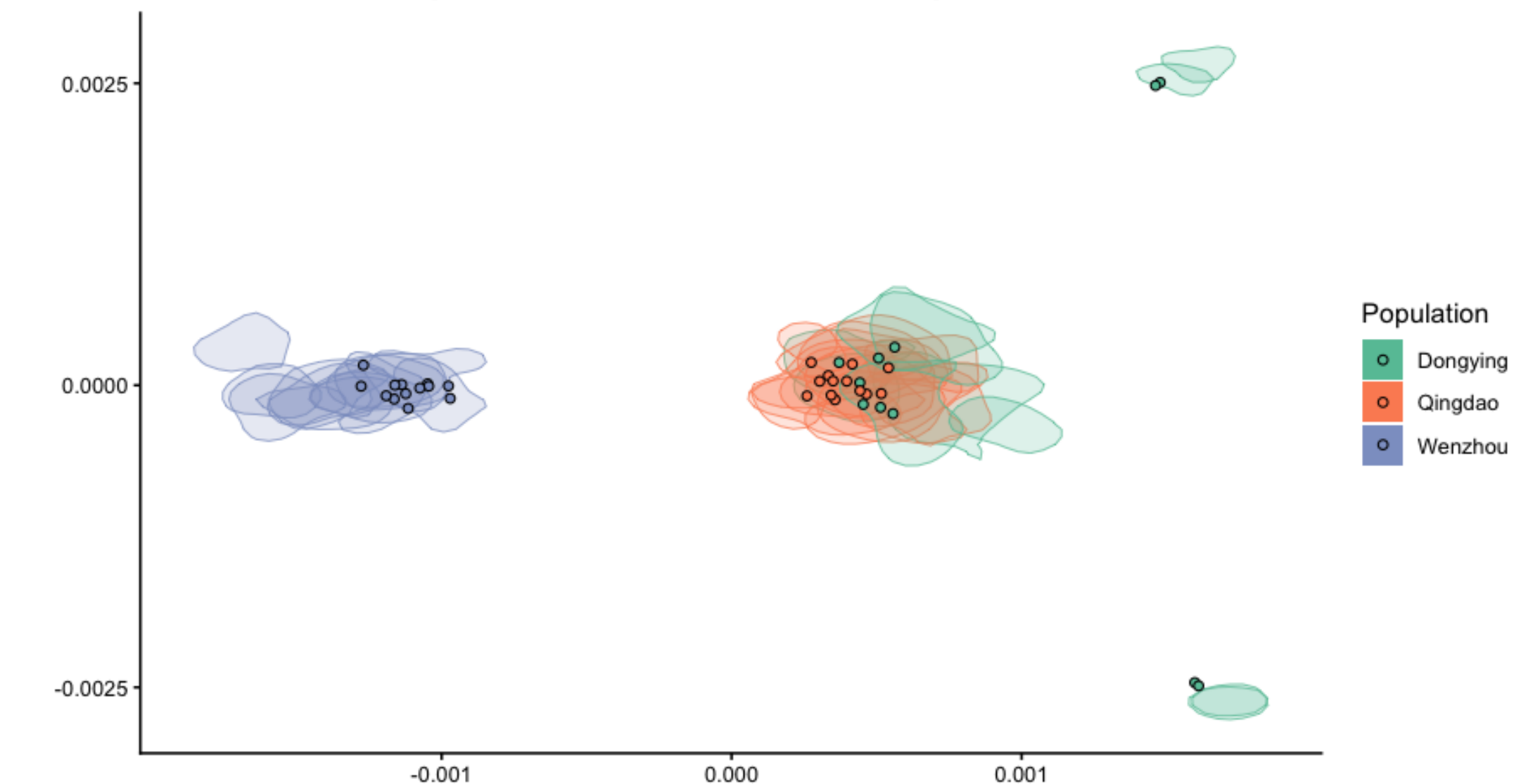
DipSkmer is More Accurate in Simulations: Panels represent different true θ values we obtain by modulating population size. Coverage is shown on the x-axis, and absolute error is shown on the y-axis.



DipSkmer Best Matches High-Coverage Method in Biological Datasets: We estimate genetic diversity from real populations (panels) with varying θ values (top strip). The y-axis shows the ratio between our θ estimates and those obtained by a high-coverage analysis (ANGSD).

Population Structure in Smelt Populations

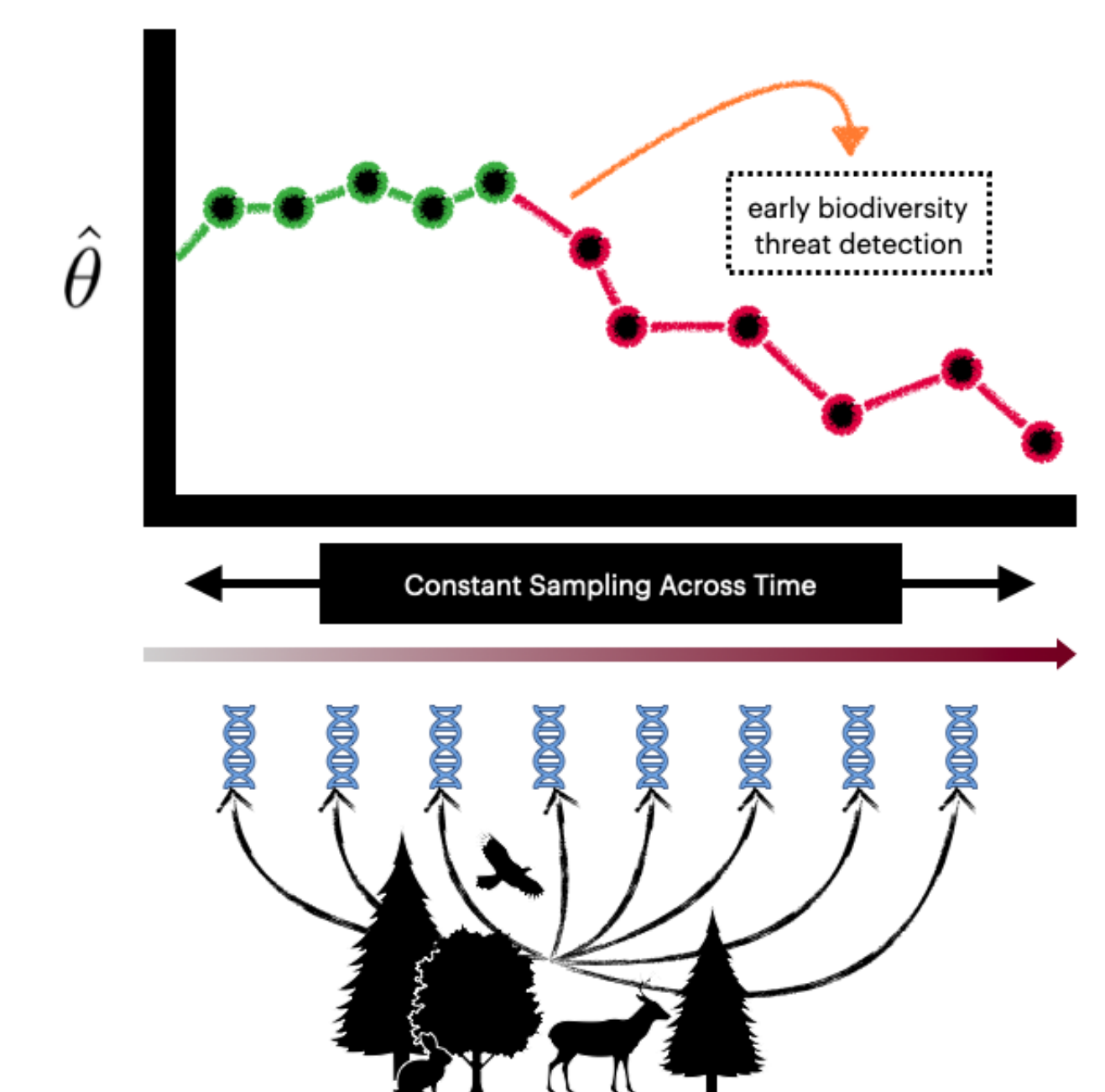
Population Structure Comparison



DipSkmer PCoA Analysis of Smelt Populations: We estimate θ across samples in a biological dataset comprised of three smelt populations (*Silago sinica*). Points show the structure inferred using ANGSD with high coverage samples. Polygons represent the structure inferred using DipSkmer.

Conclusion, Applications, and Future Work:

- Conclusions:** DipSkmer can use affordable low-coverage sequencing to estimate *genetic diversity* and infer *population structure* with the same accuracy as high-coverage methods.
- Applications:** Opens the door to frequent and scalable biodiversity monitoring.
- Future Work:** Currently DipSkmer can only be used for population parameter estimation, future work should focus on Diploid phylogenomics or reference-free GWAS approaches.



References

- [1] Brian D and et al Ondov. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, December 2016.
- [2] Shahab Sarmashghi and et al. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology*, 20(1):34, December 2019.