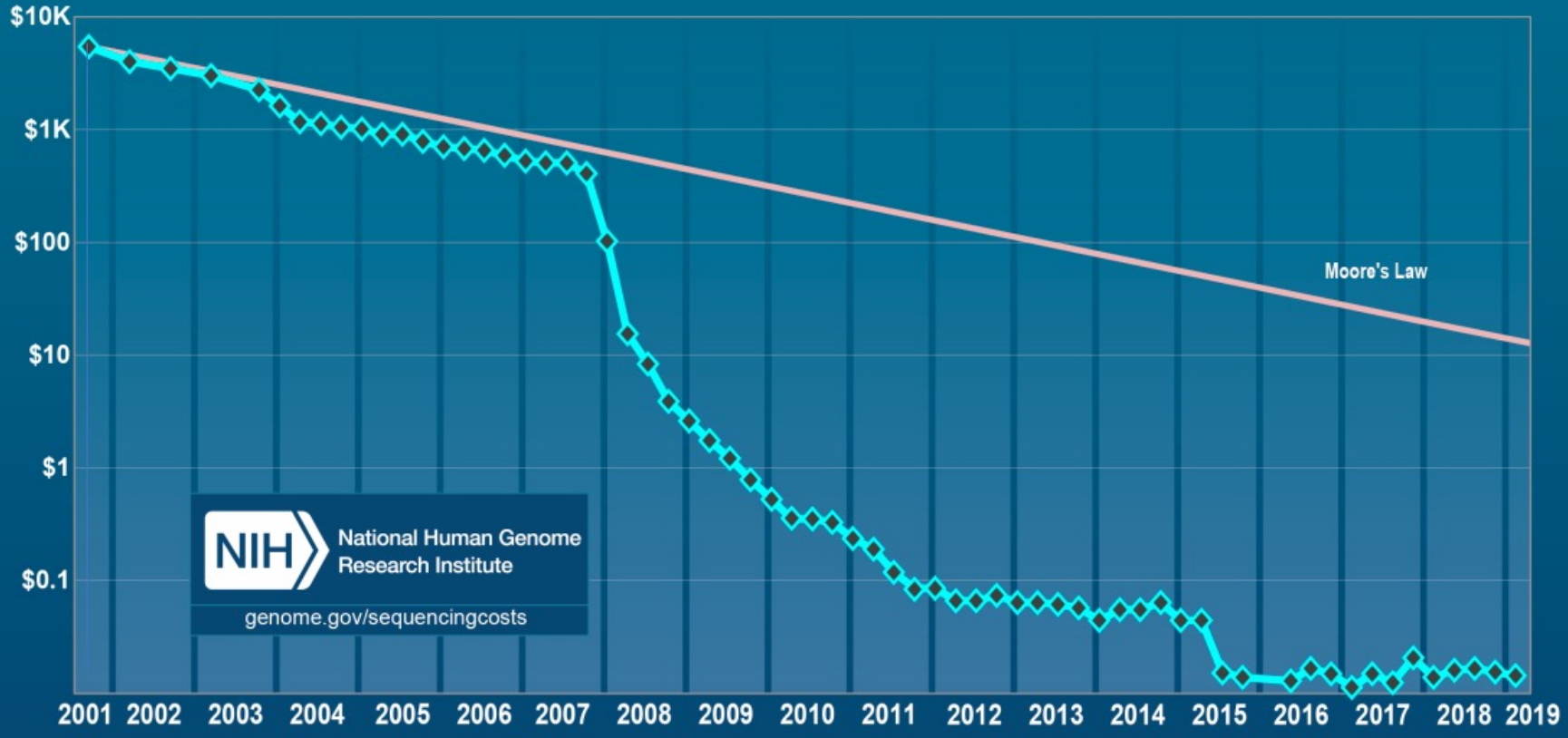
Introduction to low-coverage whole genome sequencing (lcWGS)

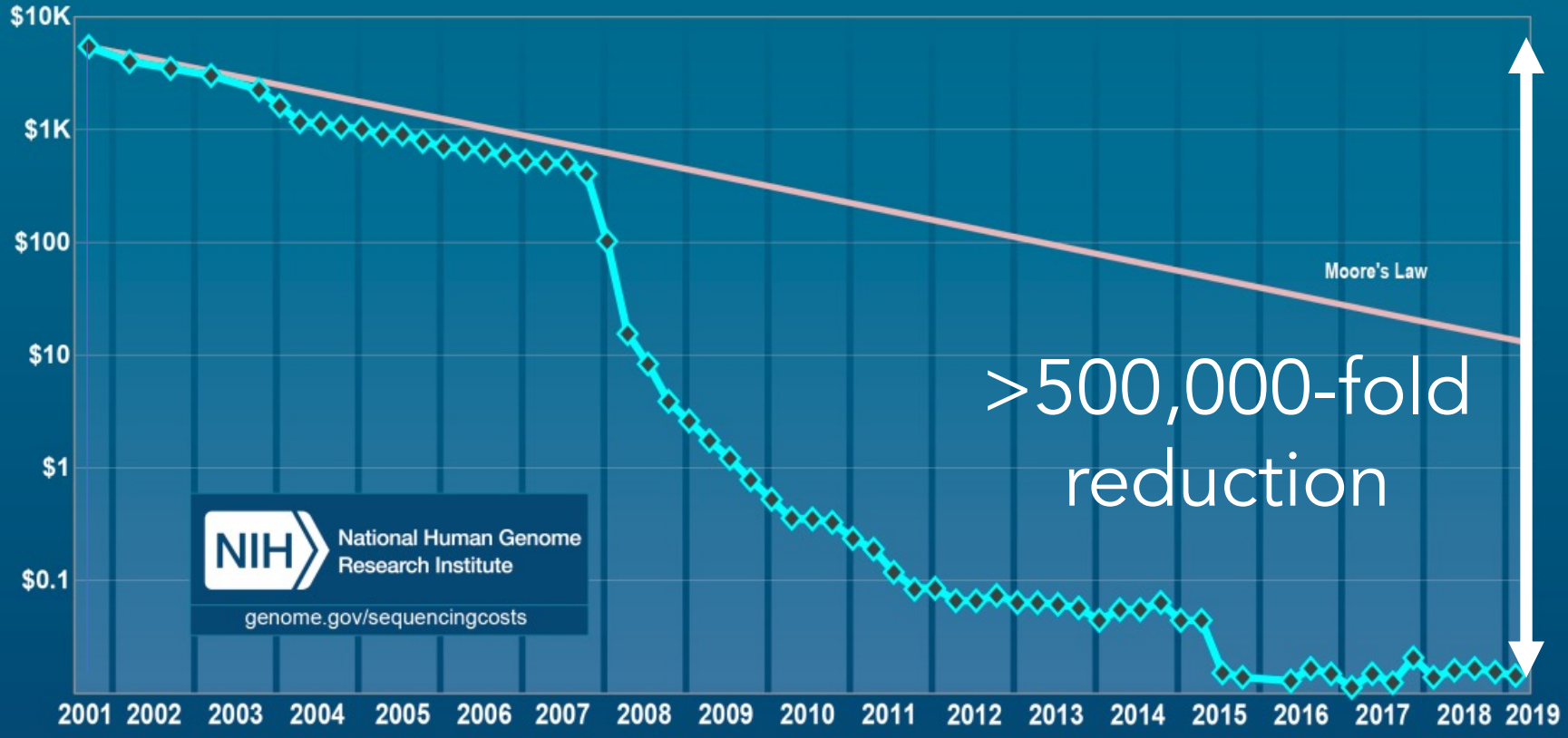
Plan for today

- Introduction to low-coverage whole genome sequencing (lecture)
- From sample to fastq (lecture + small exercise)
- From fastq to bam (lecture + practical)

Cost per Raw Megabase of DNA Sequence

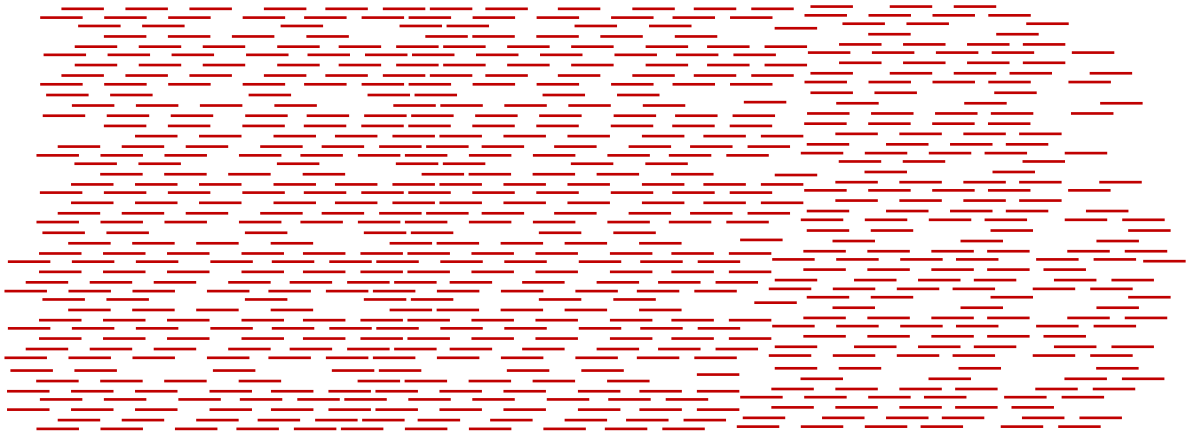


Cost per Raw Megabase of DNA Sequence

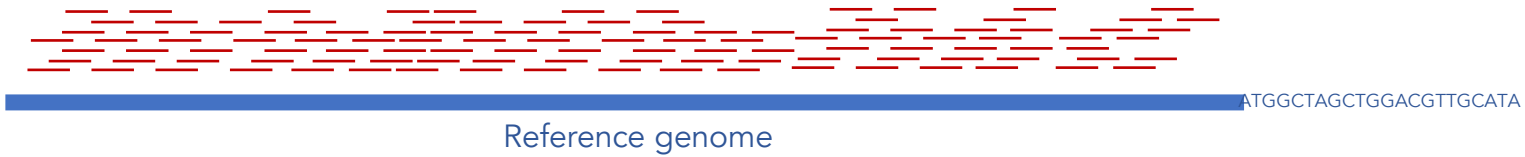


How to distribute sequencing effort?

Raw sequence reads

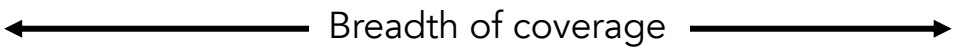


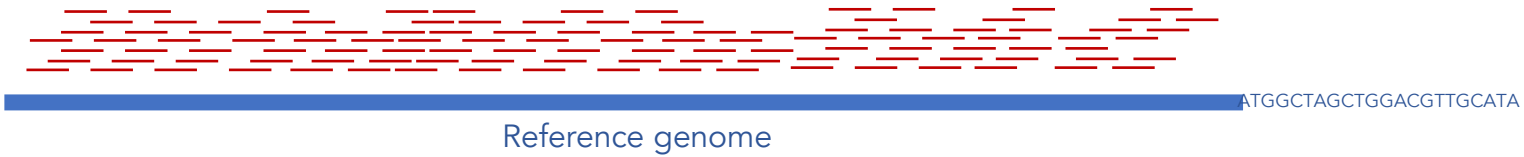
How to distribute sequencing effort?



Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

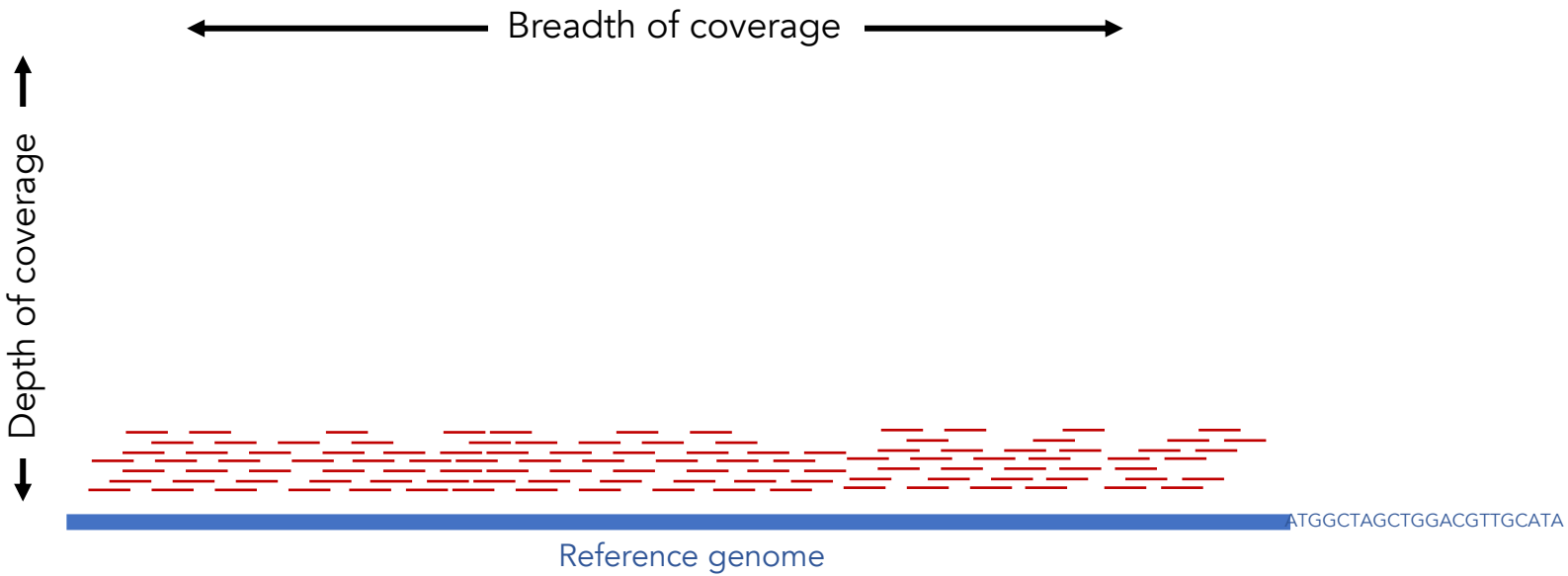
How to distribute sequencing effort?





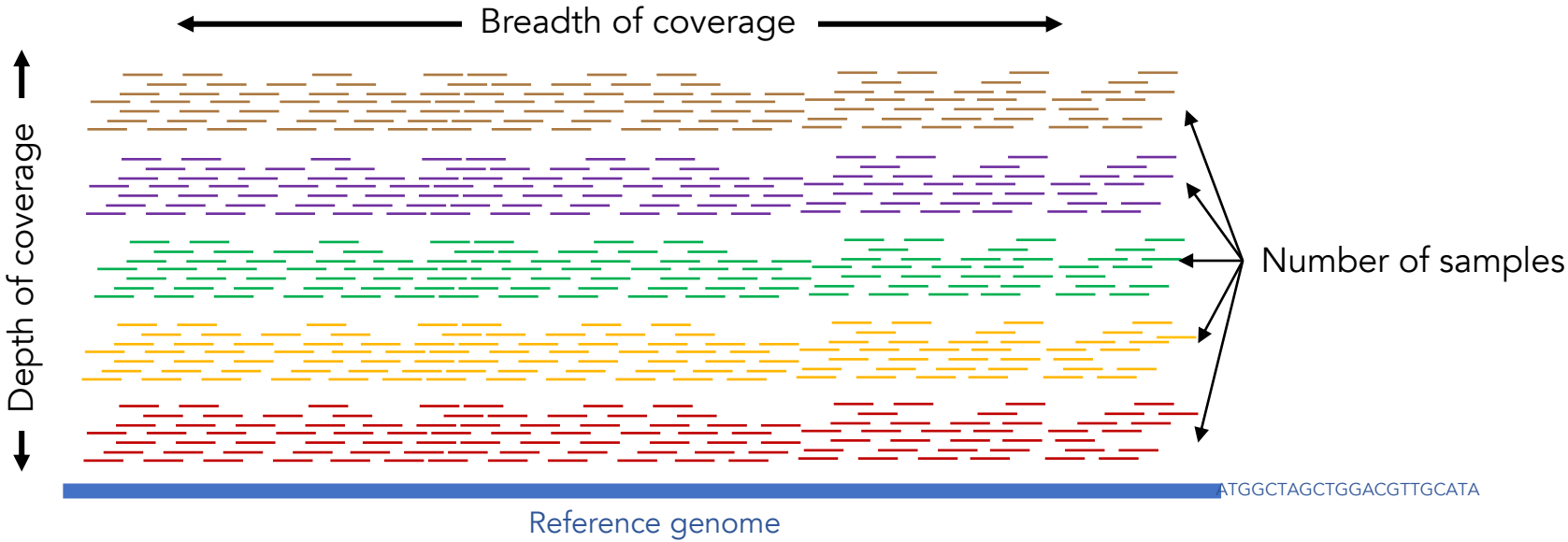
Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

How to distribute sequencing effort?



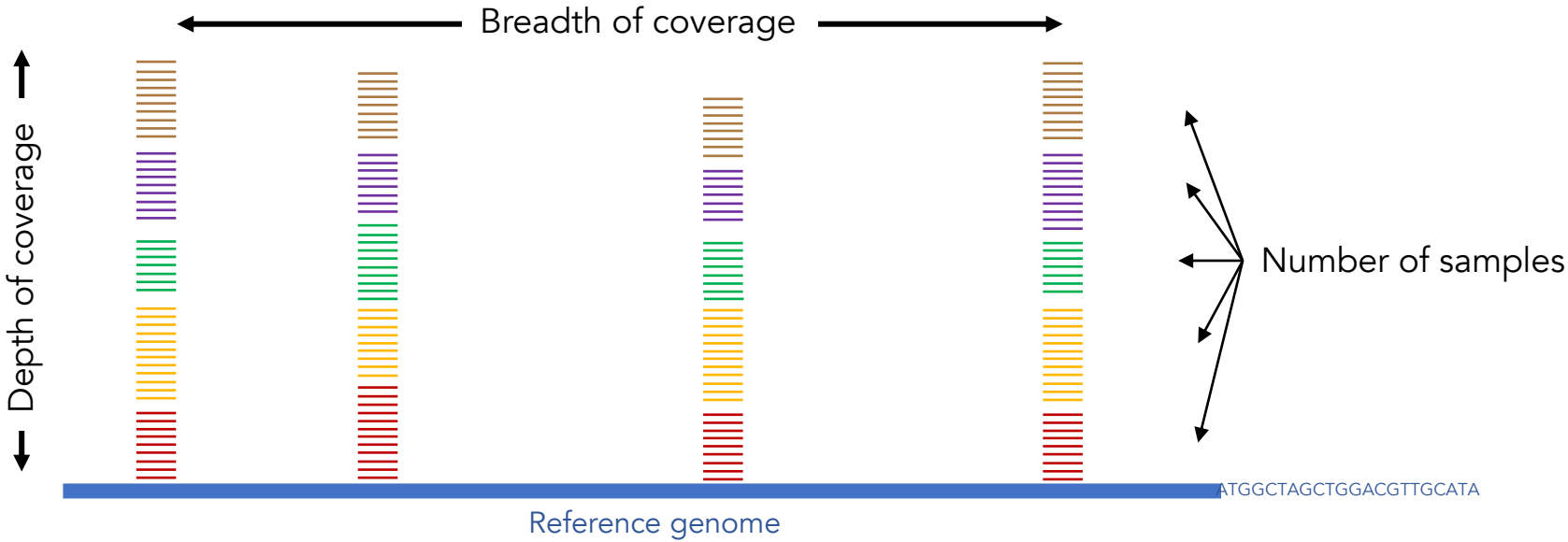
Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

How to distribute sequencing effort?



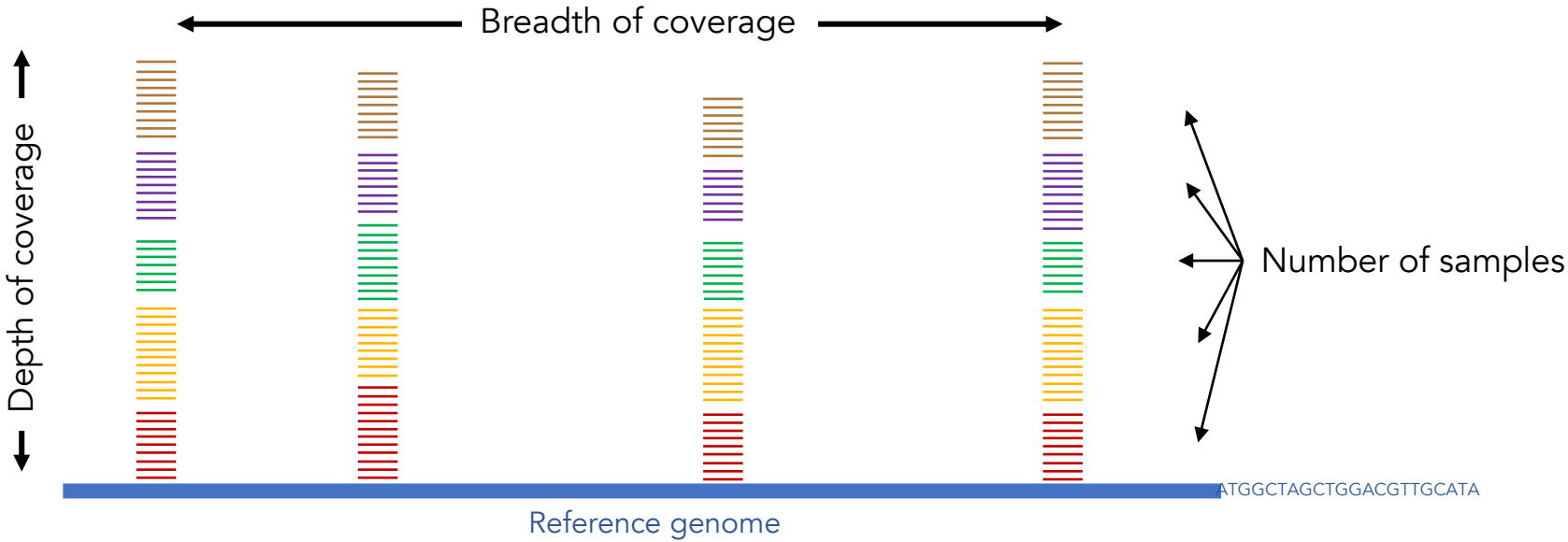
Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

RAD-seq approach



Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

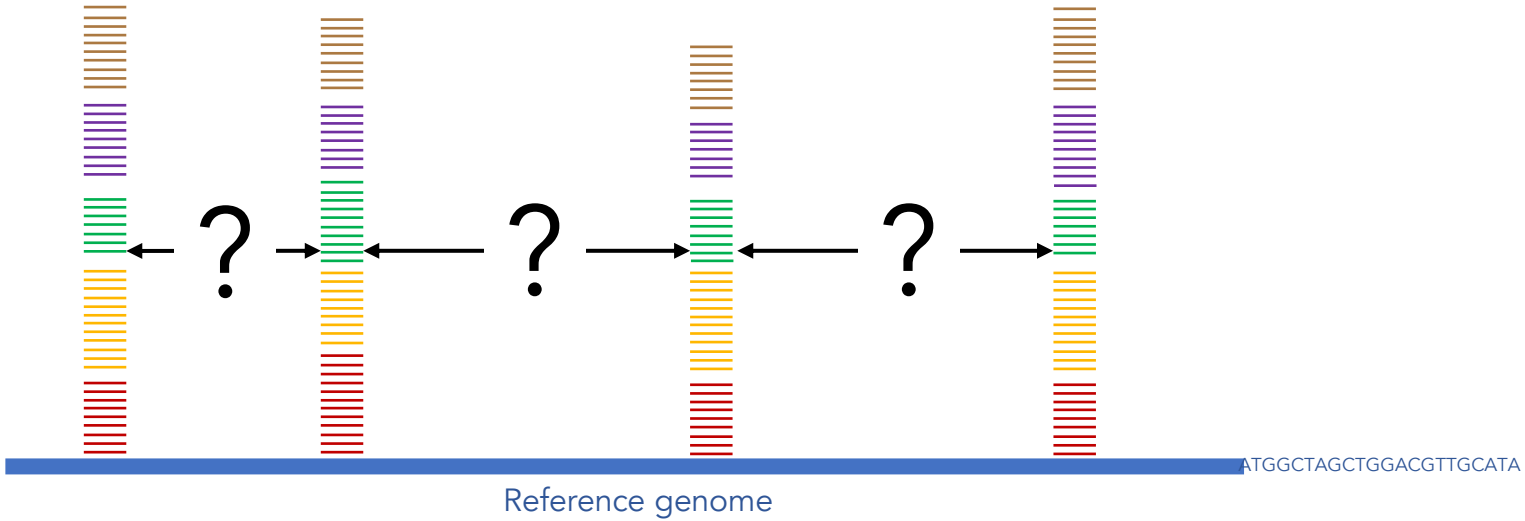
RAD-seq approach or SNP chip



Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

RAD-seq approach or SNP chip

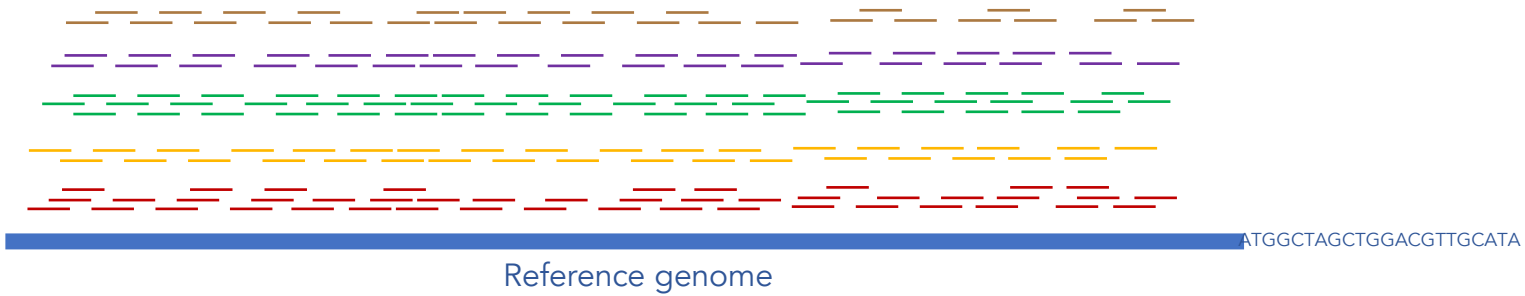
What are we missing?



Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

The value of trading depth for breadth

Maximizing the information content in our sequence data

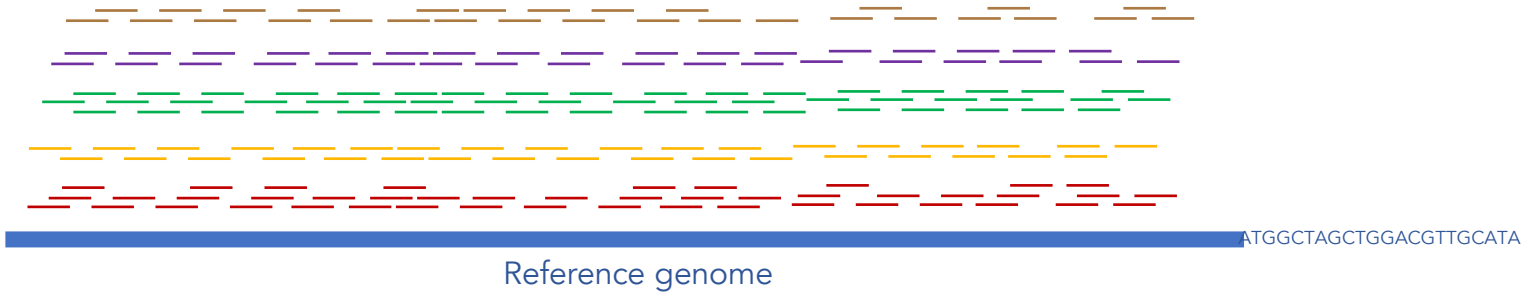


Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

The value of trading depth for breadth

Maximizing the information content in our sequence data

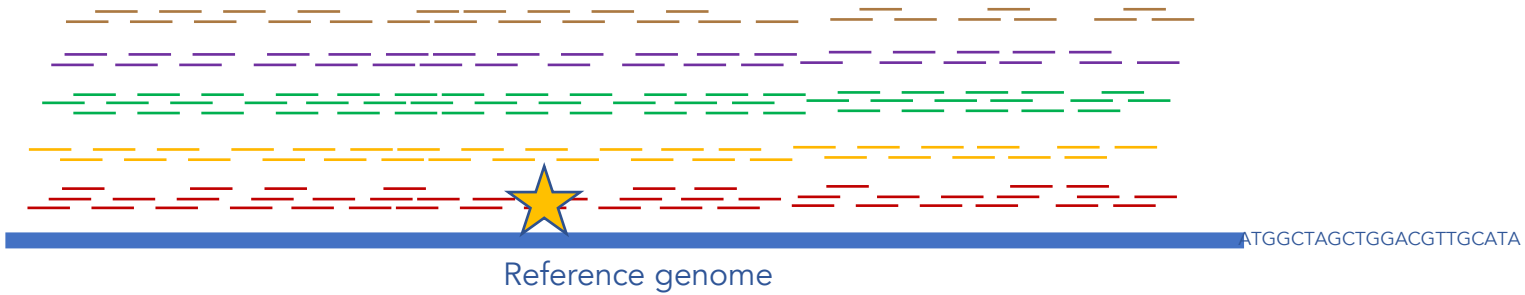
But low-coverage precludes robust genotype calls



Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

The value of trading depth for breadth

Do we care about the genotype at a particular SNP in a particular individual?

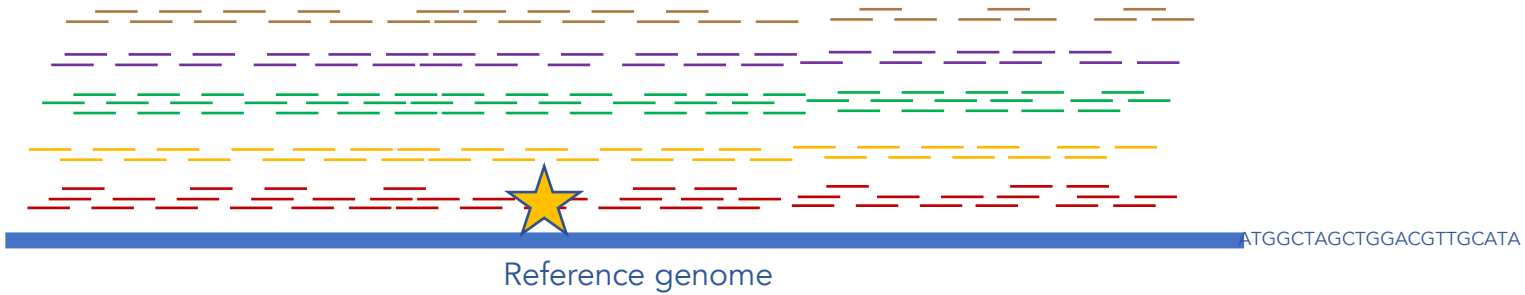


Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

The value of trading depth for breadth

Do we care about the genotype at a particular SNP in a particular individual?

Often not

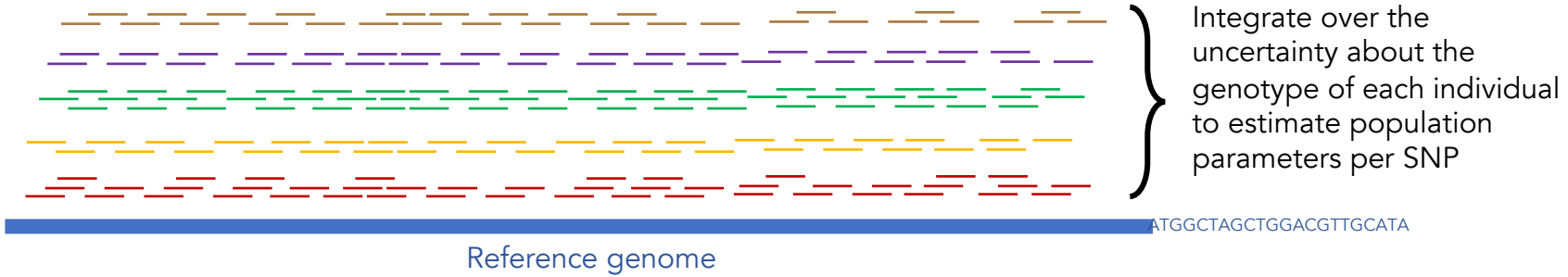


Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

The value of trading depth for breadth

Do we care about the genotype at a particular SNP in a particular individual?

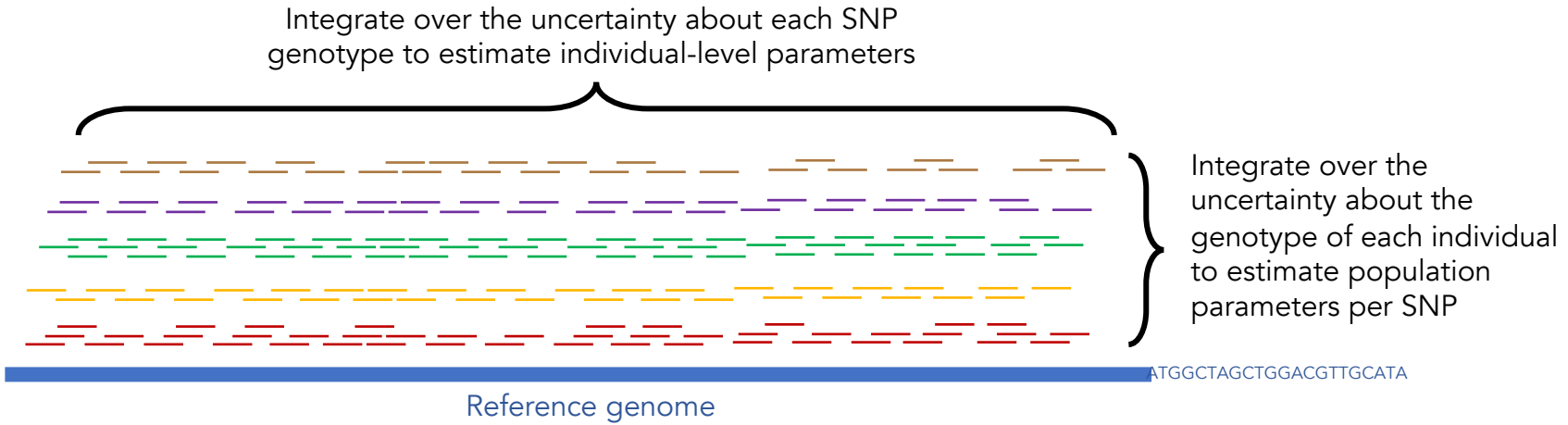
Often not



Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

The value of trading depth for breadth

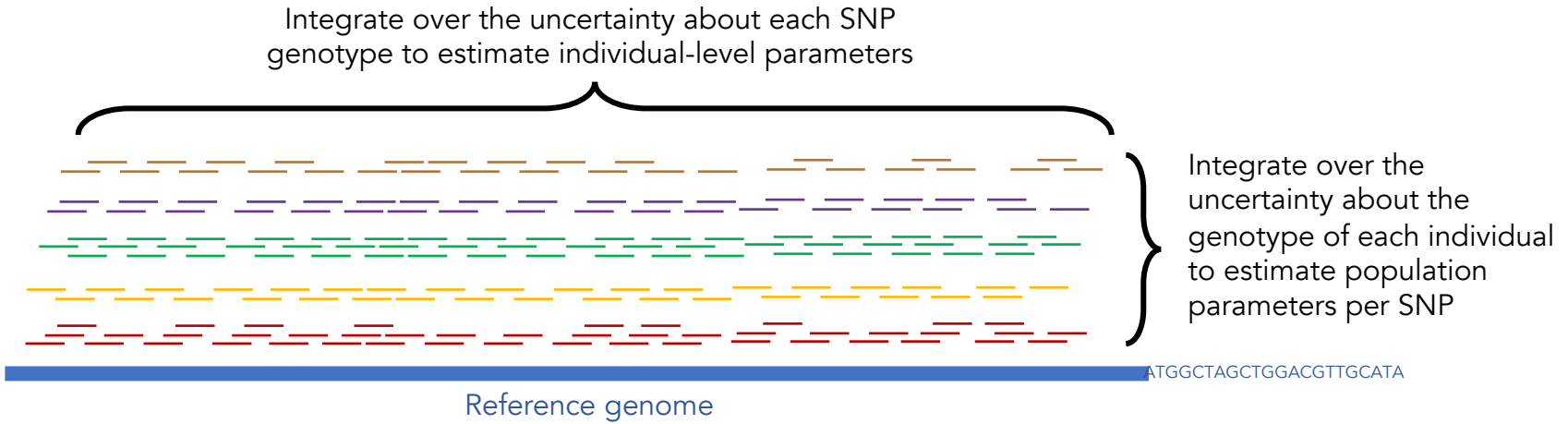
Do we care about the genotype at a particular SNP in a particular individual?



Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

The value of trading depth for breadth

Robust inference when downstream analysis is conducted in a probabilistic framework that takes uncertainty into account

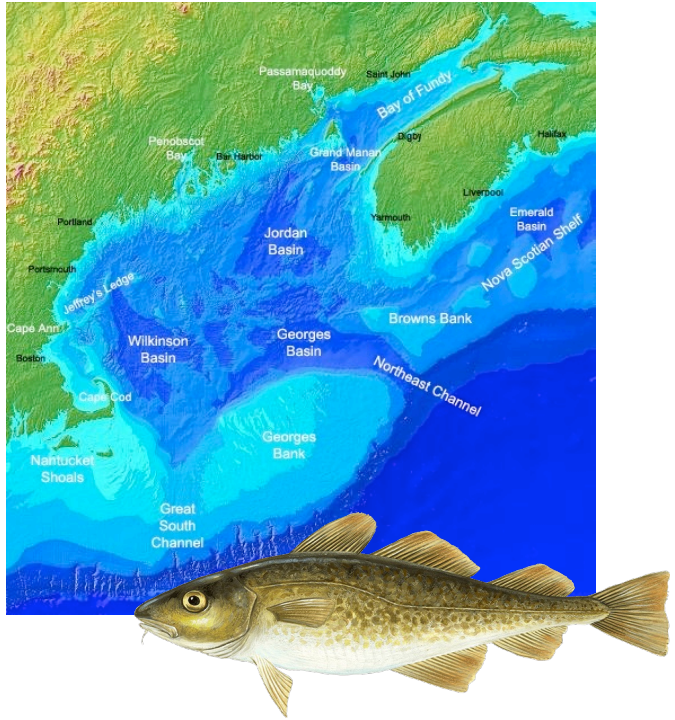


Adapted from https://sites.google.com/site/wiki4metagenomics/pdf/definition/coverage-read-depth

Adaptive divergence in Atlantic cod in the Gulf of Maine

1. Use genomic analyses to characterize structure of active cod spawning populations in US waters

2. Identify adaptive genetic differences among populations



Collaborators

Adrienne Kovach

University of New Hampshire

Gemma Clucas

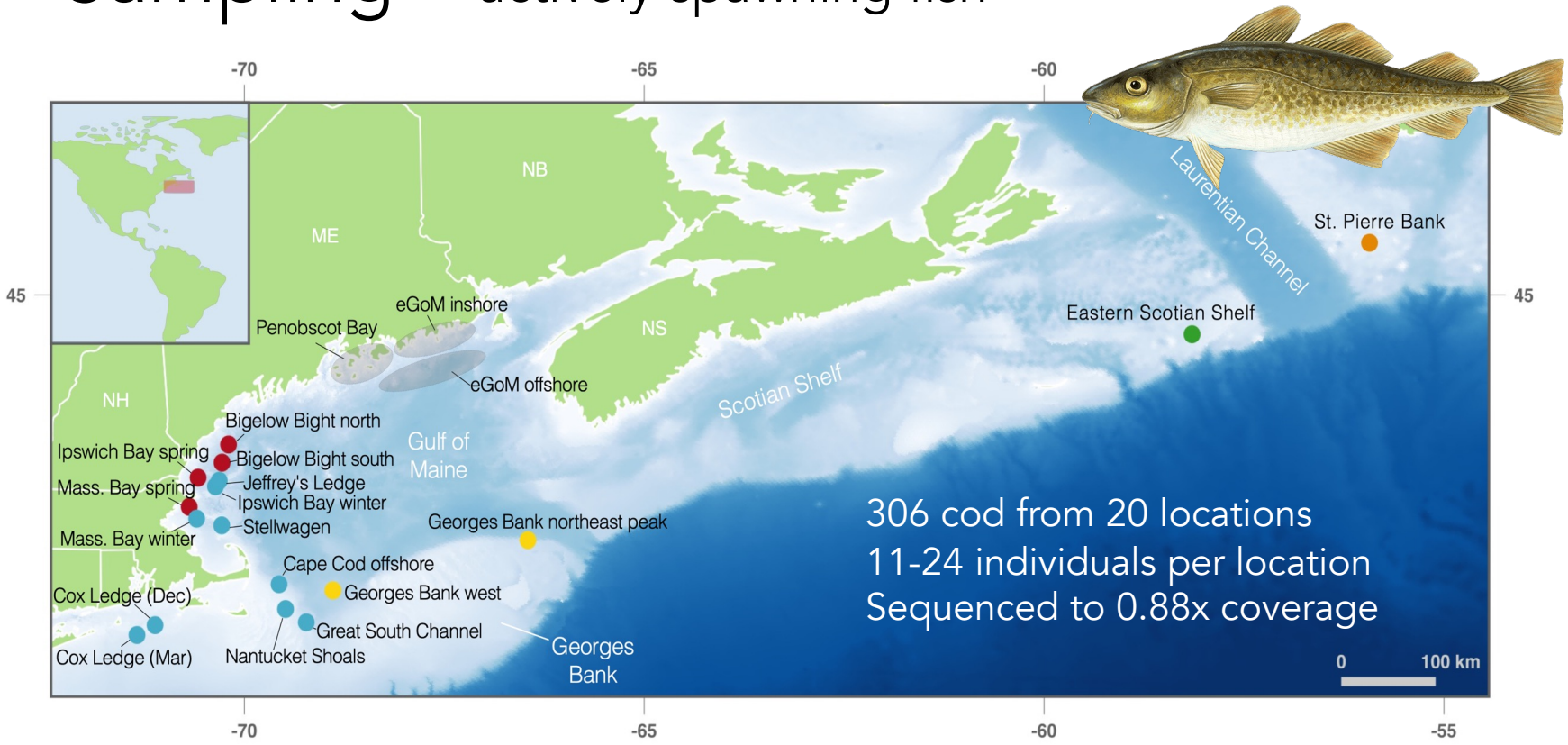
University of New Hampshire

Nicolas Lou

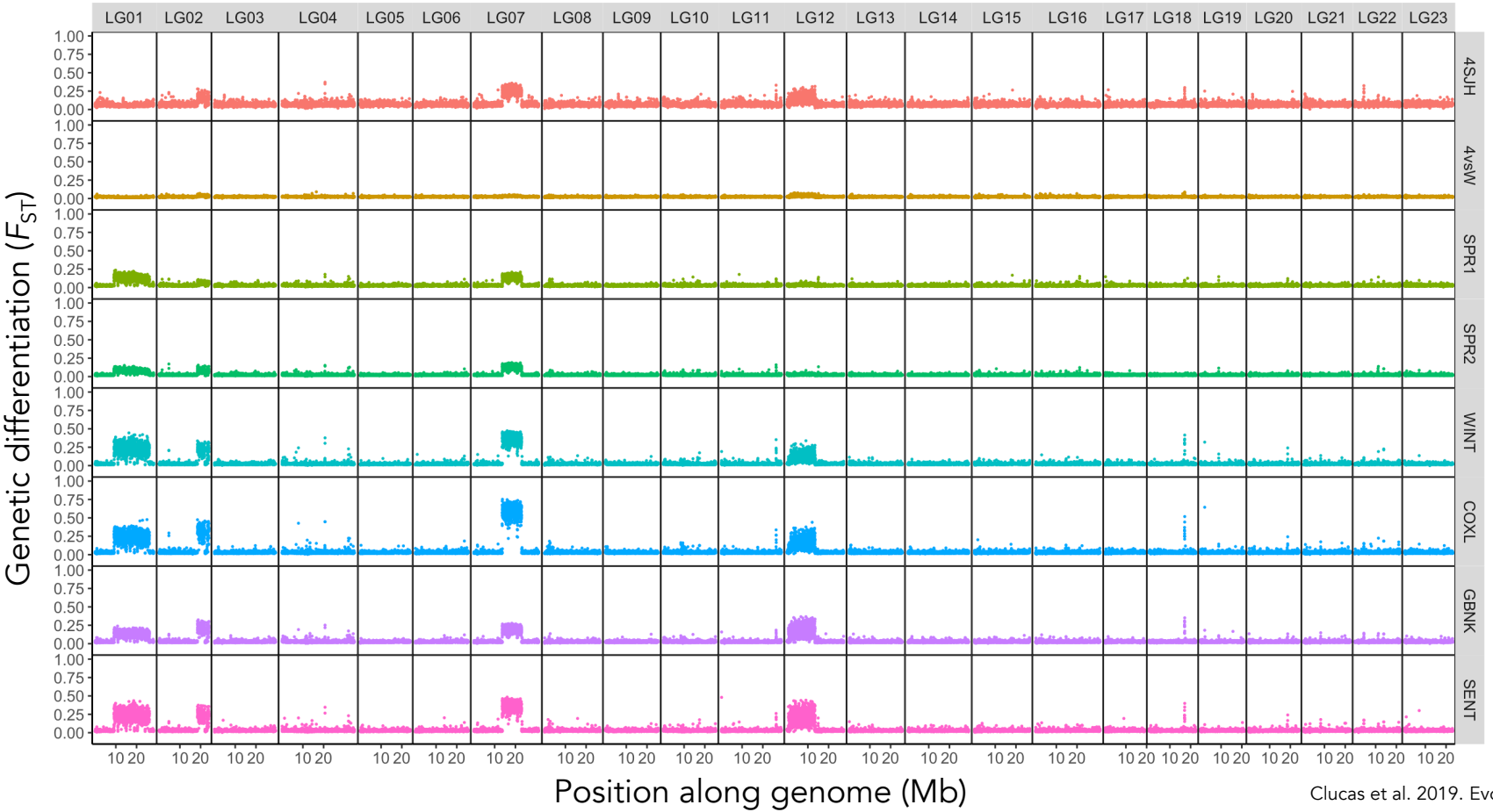
Cornell University

Sea Grant New Hampshire

Sampling – actively spawning fish

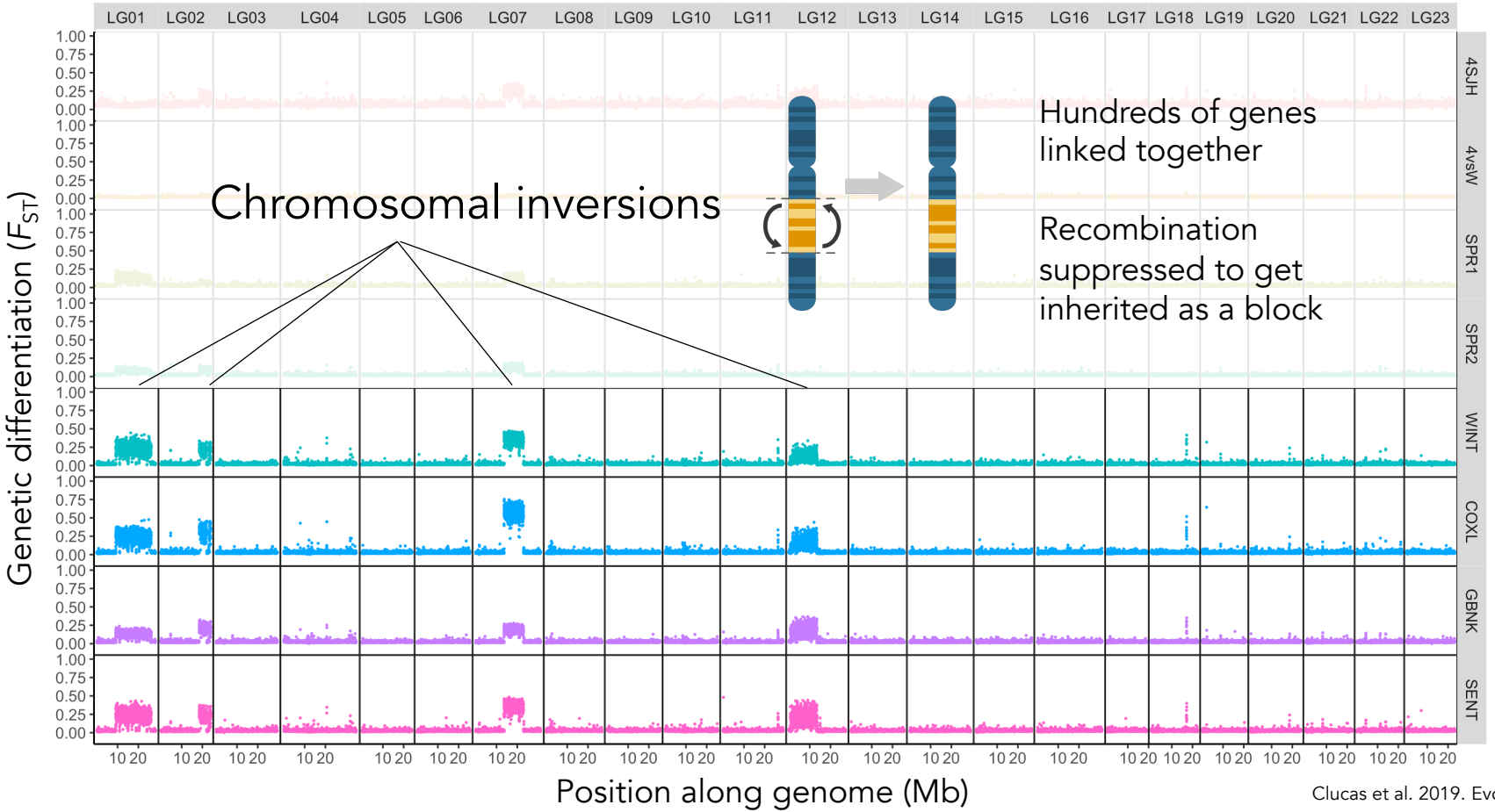


Pairwise comparisons to St. Pierre Bank, Canada (15 kb windows)



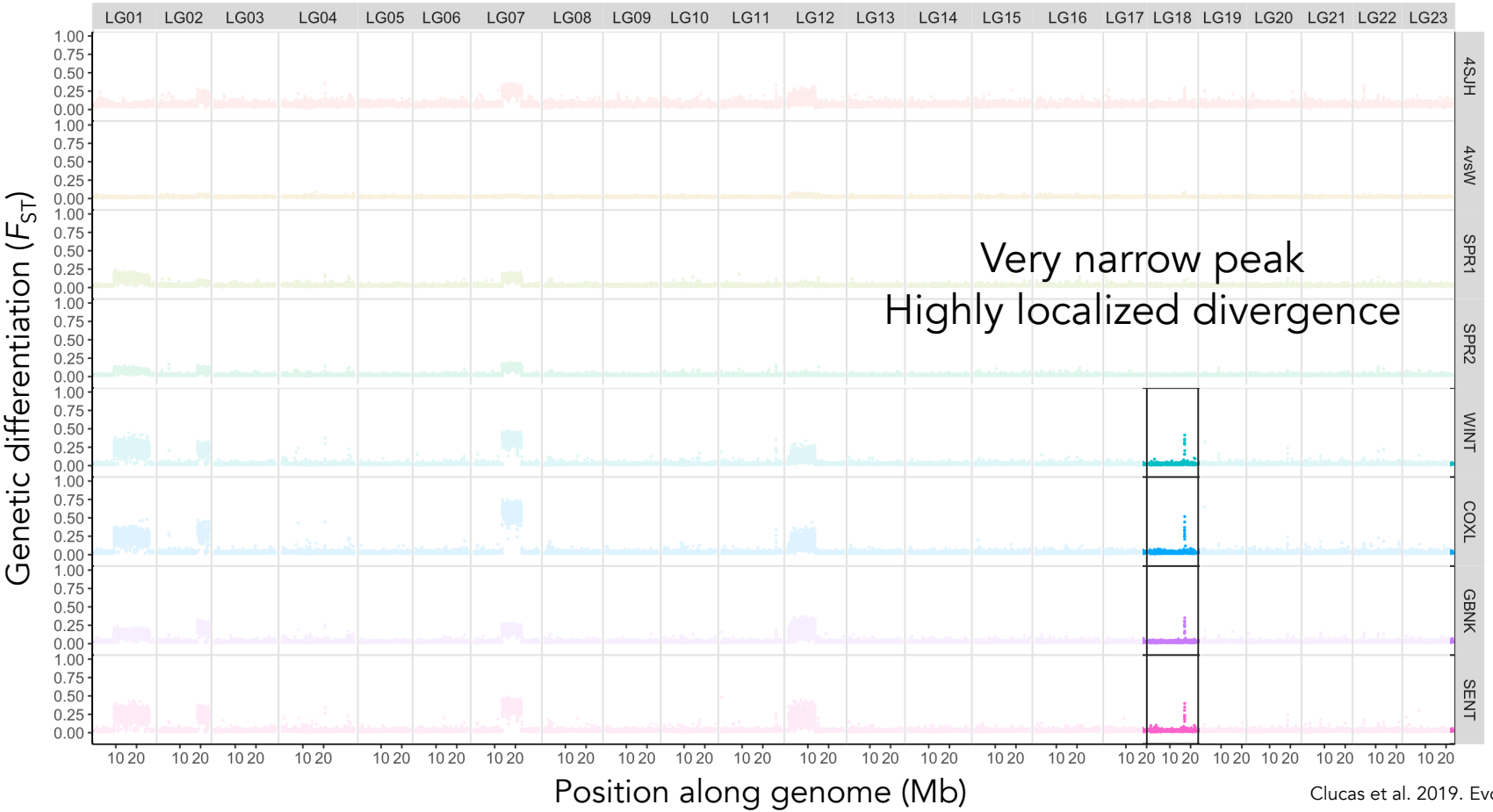
Clucas et al. 2019. Evol. Appl.

Pairwise comparisons to St. Pierre Bank, Canada (15 kb windows)



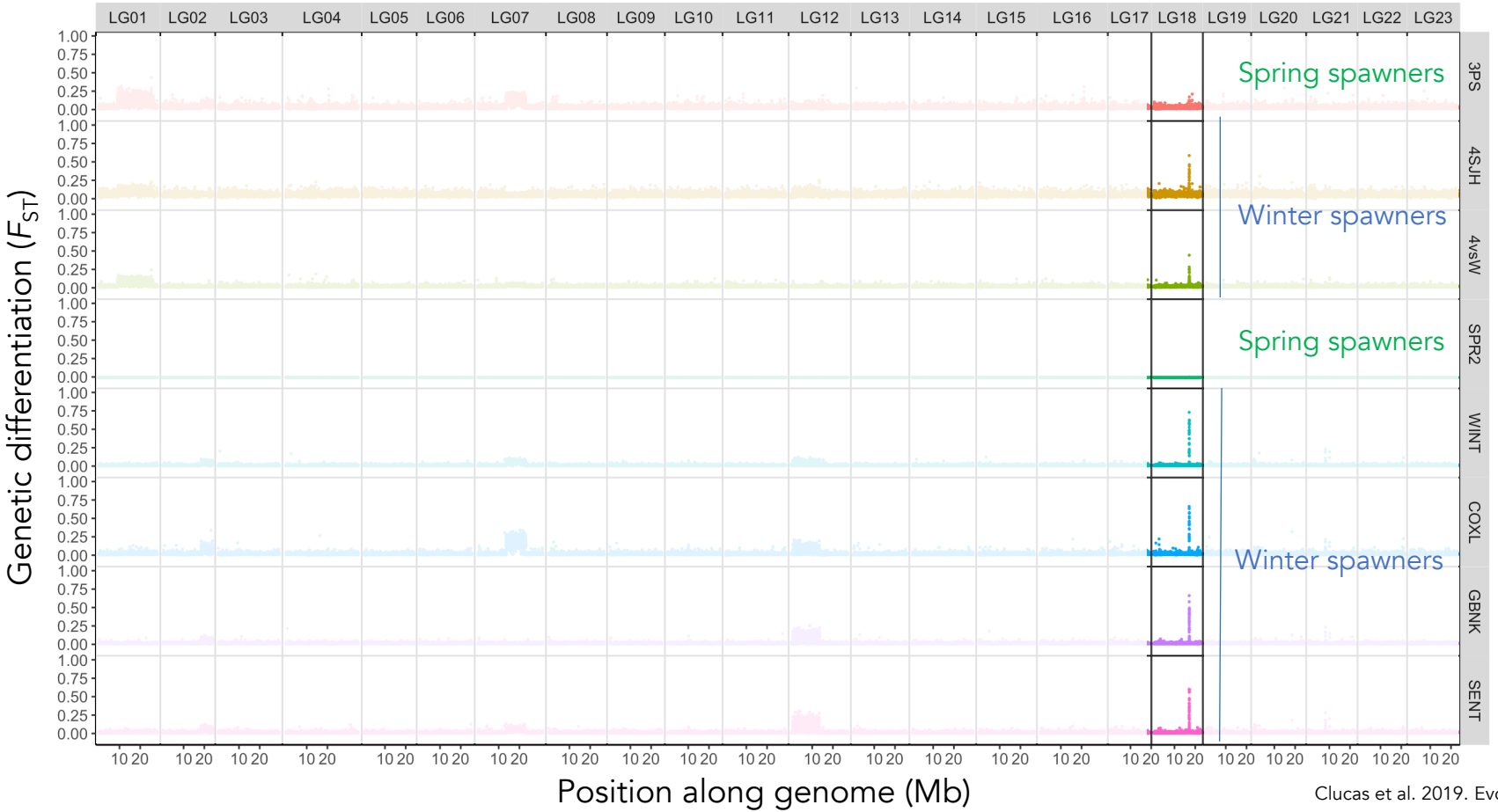
Clucas et al. 2019. Evol. Appl.

Pairwise comparisons to St. Pierre Bank, Canada (15 kb windows)



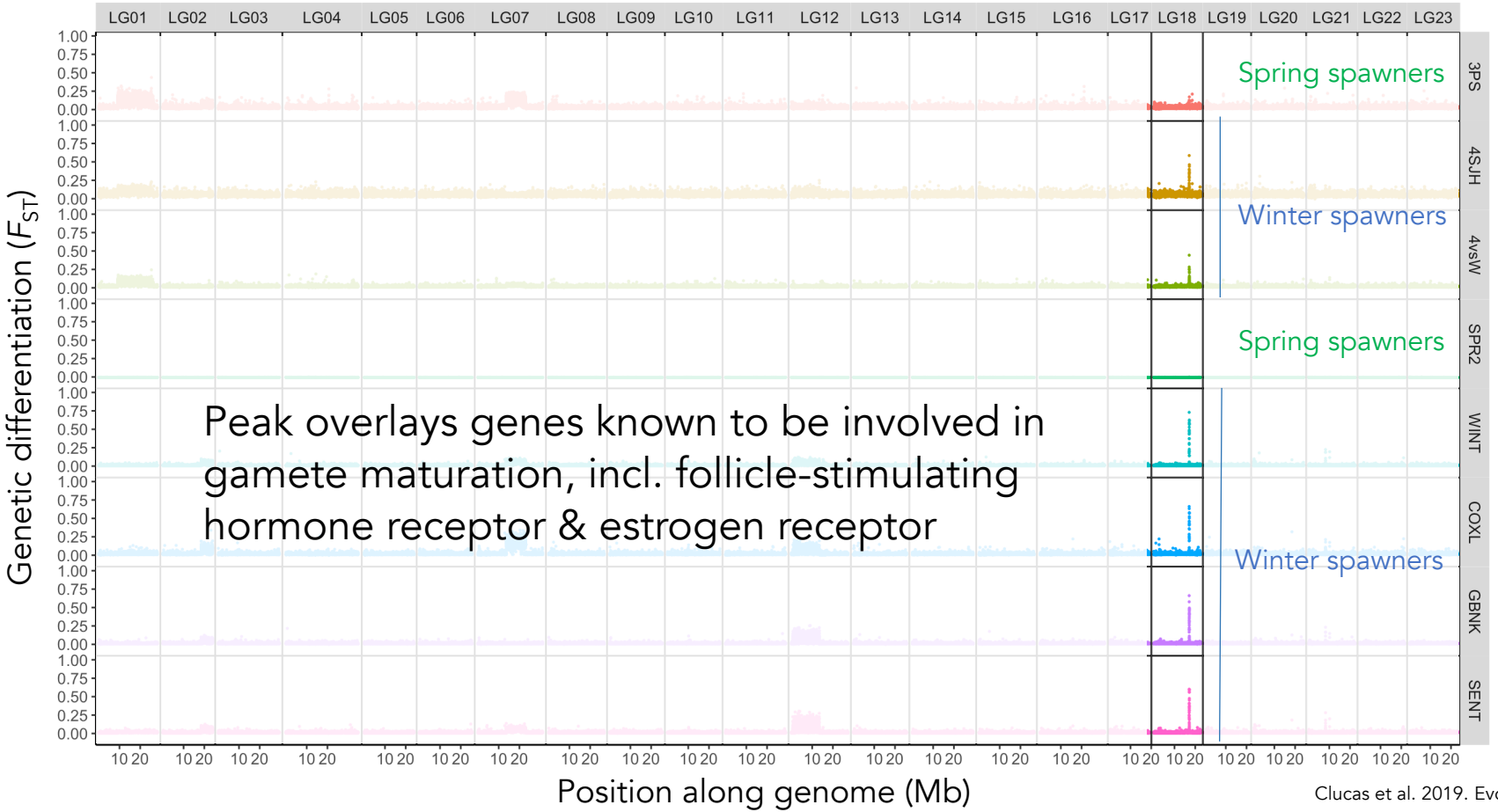
Clucas et al. 2019. Evol. Appl.

Pairwise comparisons to the wGoM spring spawners (5 kb windows)



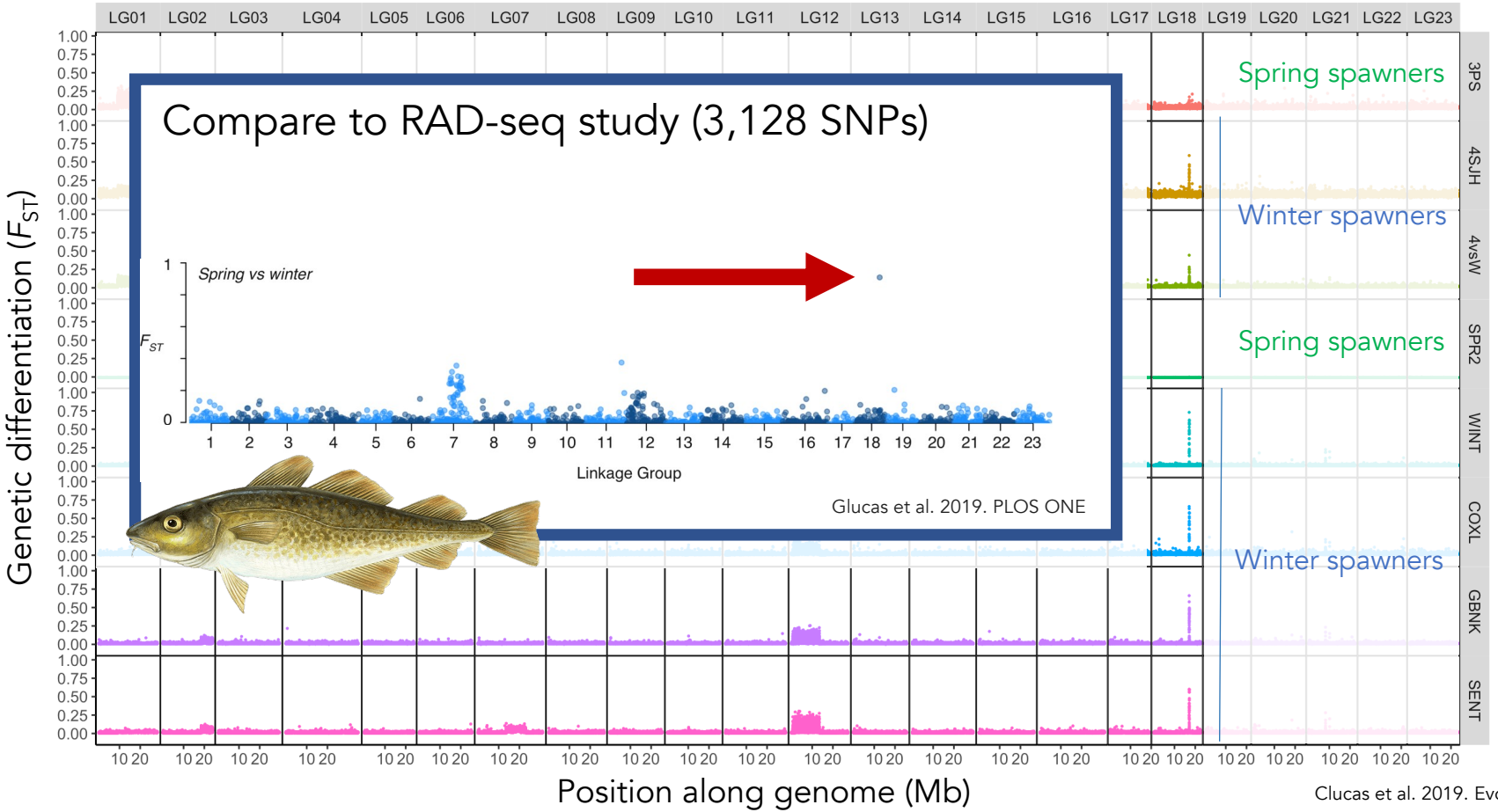
Clucas et al. 2019. Evol. Appl.

Pairwise comparisons to the wGoM spring spawners (5 kb windows)



Clucas et al. 2019. Evol. Appl.

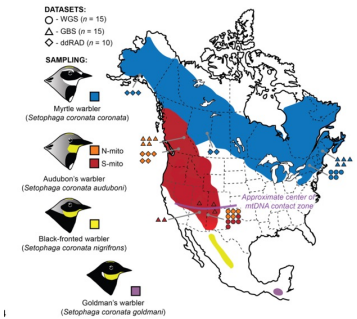
Pairwise comparisons to the wGoM spring spawners (5 kb windows)

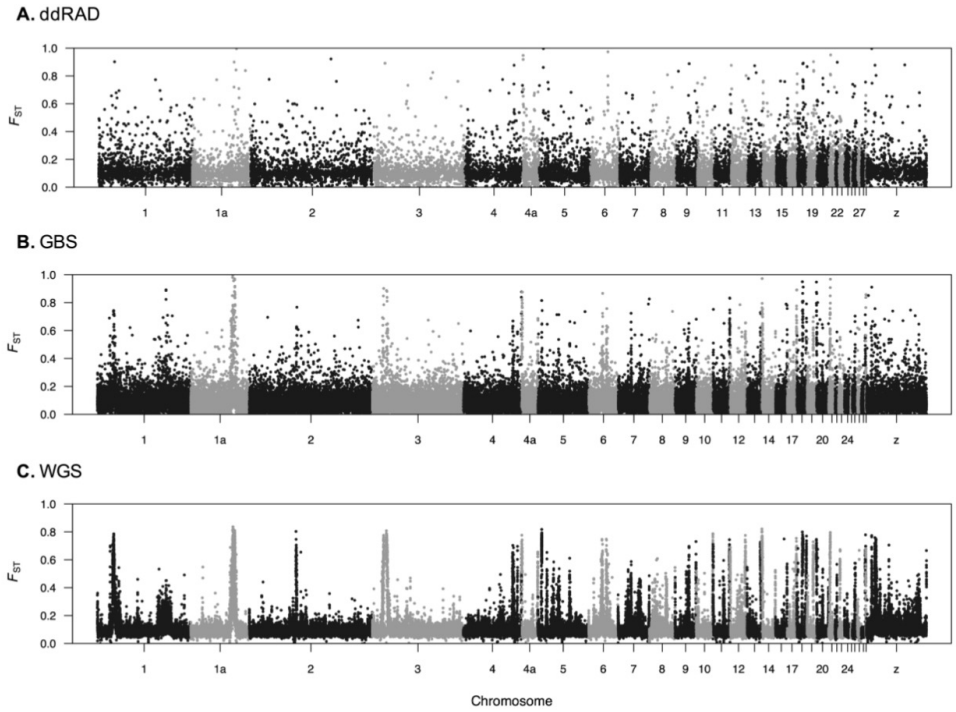


Clucas et al. 2019. Evol. Appl.

Direct comparison of ddRAD, GBS, and WGS data

Comparing subspecies of yellow-rumped warbler





41 SNPs/10 kb

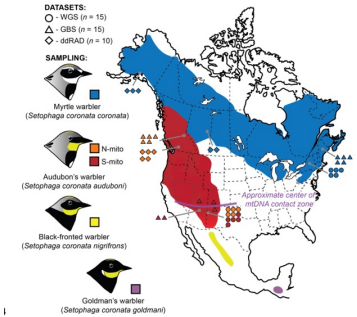
364 SNPs/10 kb

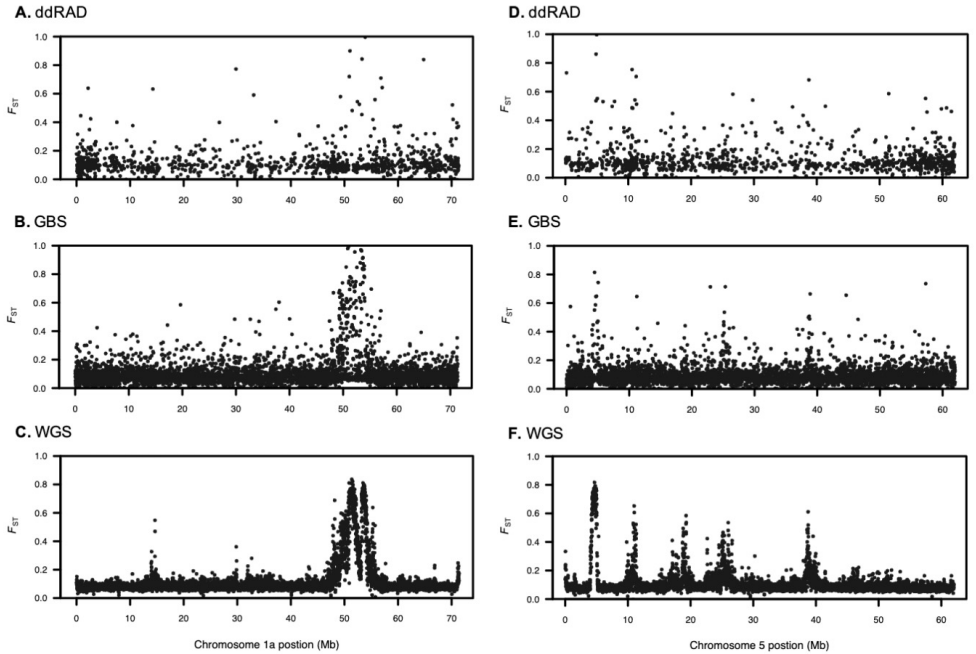
9361 SNPs/10 kb

Szarmach et al. 2021. Mol Ecol

Direct comparison of ddRAD, GBS, and WGS data

Comparing subspecies of yellow-rumped warbler





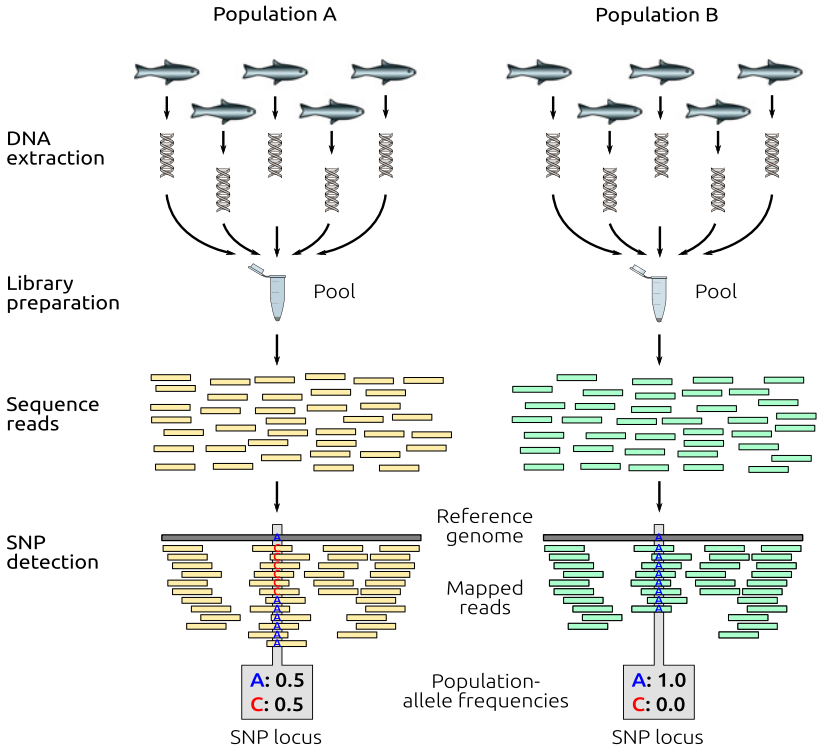
41 SNPs/10 kb

364 SNPs/10 kb

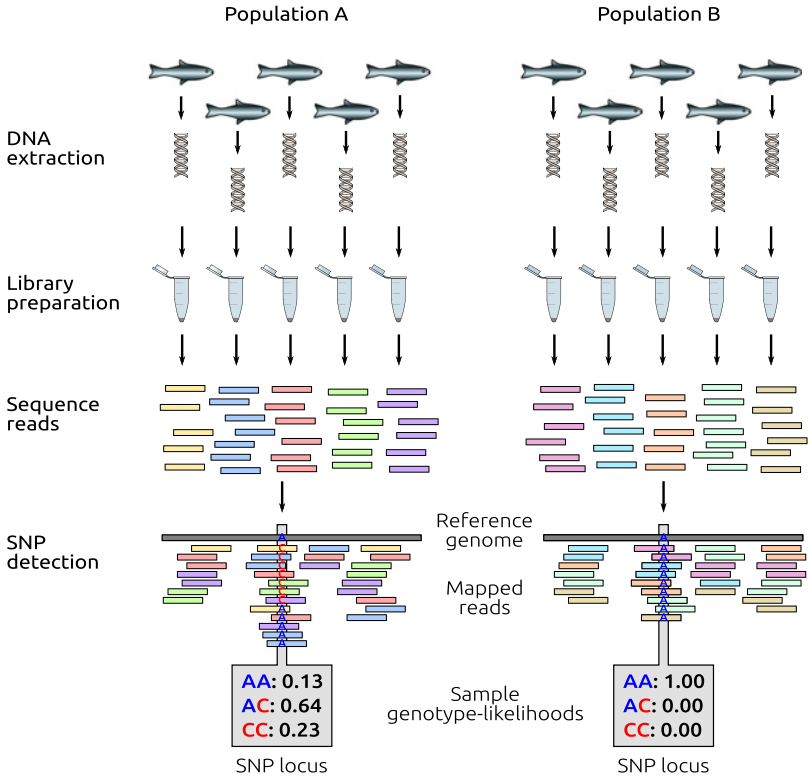
9361 SNPs/10 kb

Szarmach et al. 2021. Mol Ecol

Pool-seq



Low-coverage WGS



From Fuentes-Pardo and Ruzzante. 2017. Mol Ecol

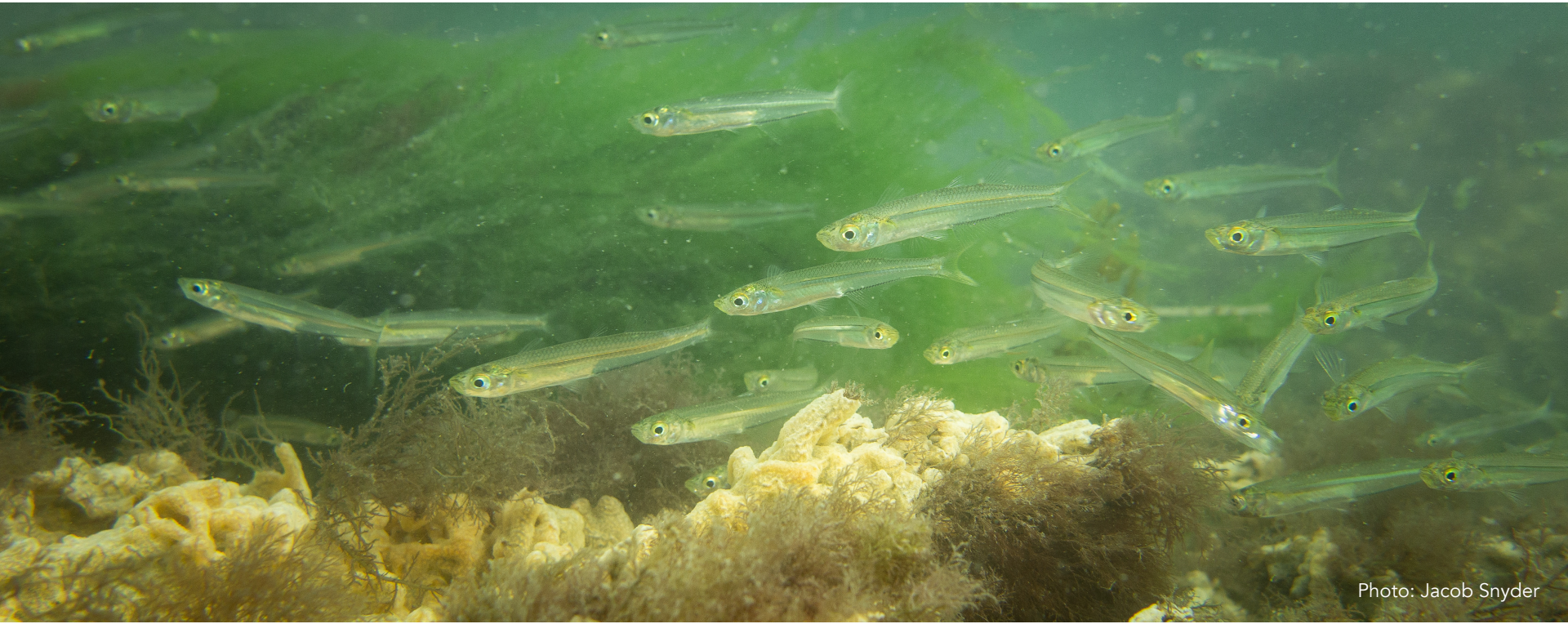
Advantages of lcWGS over Poolseq

- Can account for uneven contribution of samples in pool

- Retain individual information about e.g.

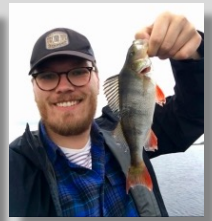
    - Cryptic population mixing
    - Relatedness
    - Estimates of individual heterozygosity or inbreeding
    - Linkage disquilibrium

Atlantic silverside *Menidia menidia*



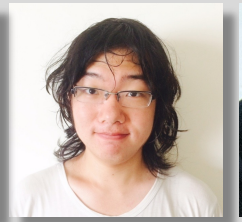
Collaborators:

Aryn Wilder
San Diego Zoo



Arne Jacobs
U. of Glasgow

Anna Tigano
U. of British Columbia



Nicolas Lou
Cornell University

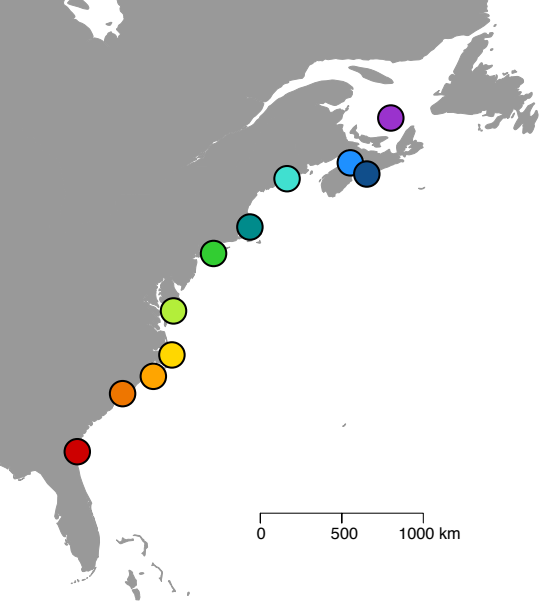
Hannes Baumann
U. of Connecticut

Steve Palumbi
Stanford University

David Conover
U. of Oregon

Funding:

BIO-OCE 1434325
BIO-OCE 1756316

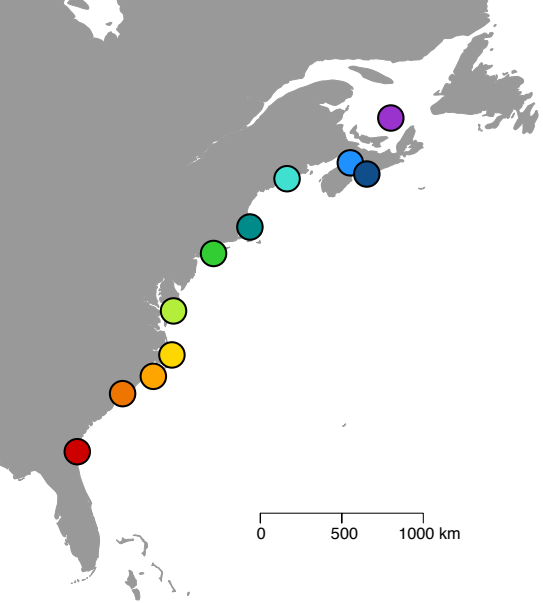


Individual-level analysis

Enables analysis of the genetic distance among individuals

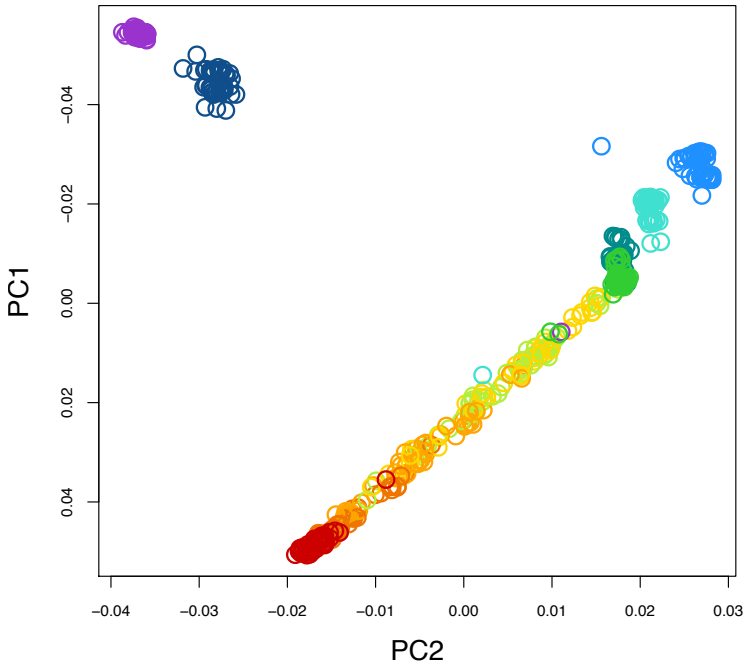
n=50 silversides per site (sampled during spring spawning)

Low-coverage ‘in silico’ exome capture



Individual-level analysis

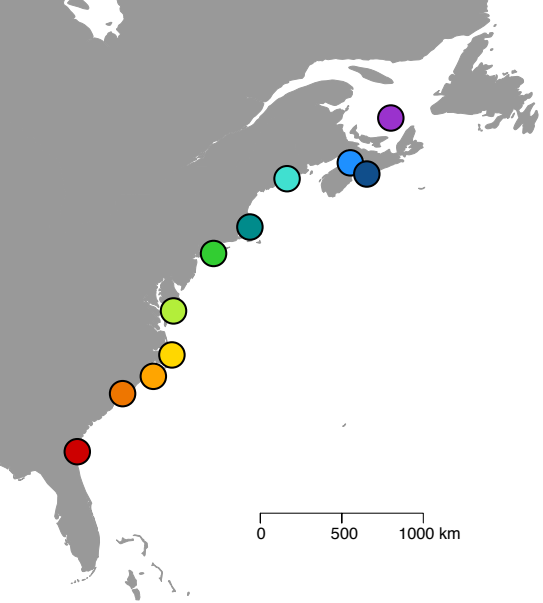
Enables analysis of the genetic distance among individuals



Multi-dimensional scaling of genetic distance (across 2.3 million SNPs)

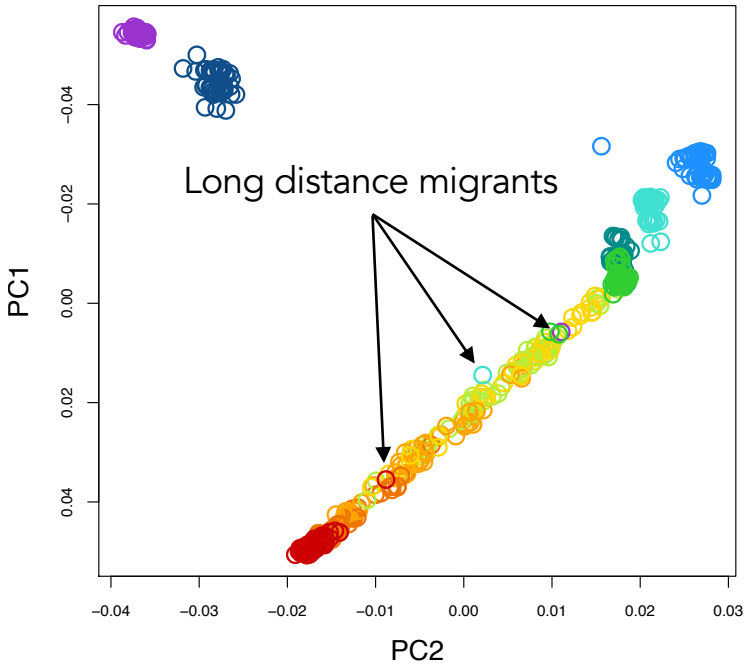
n=50 silversides per site (sampled during spring spawning)

Low-coverage ‘in silico’ exome capture (~1.3x)



Individual-level analysis

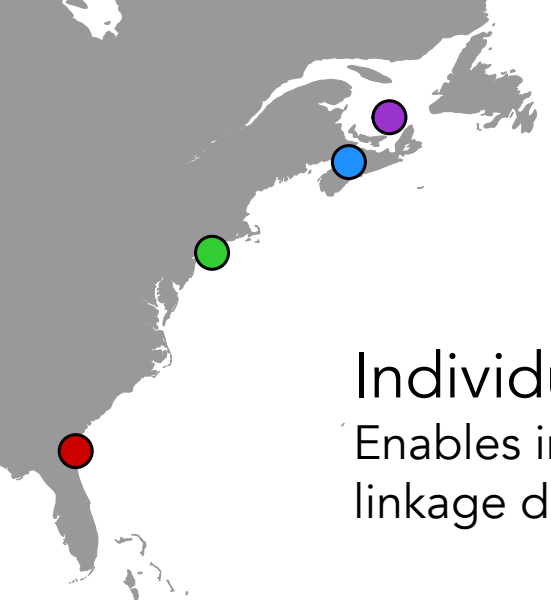
Enables identification of divergent individuals and mixed samples

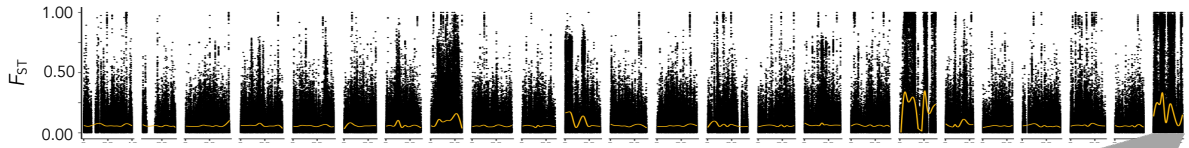


Multi-dimensional scaling of genetic distance (across 2.3 million SNPs)

n=50 silversides per site (sampled during spring spawning)

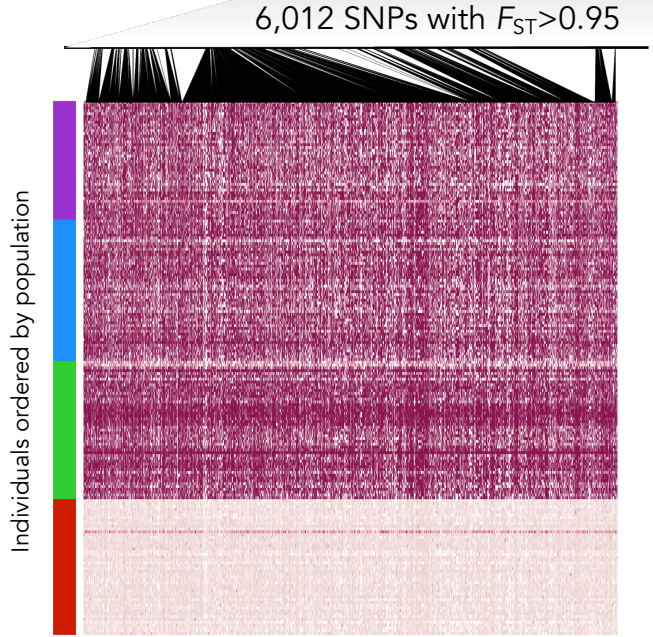
Low-coverage ‘in silico’ exome capture (~1.3x)





Individual analysis

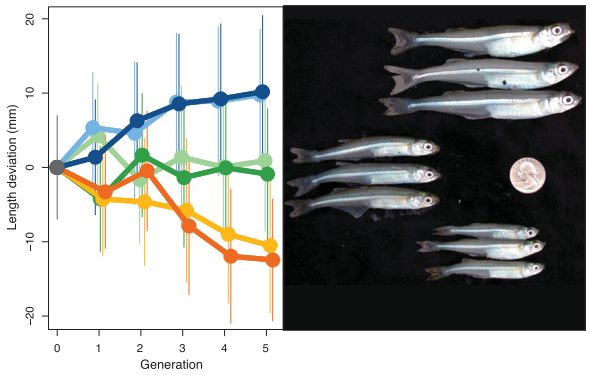
Enables inference of linkage disequilibrium



Most likely genotype for each individual

- Northern homozygote
- Heterozygote
- Southern homozygote

Wilder et al. 2020. Evolution Letters



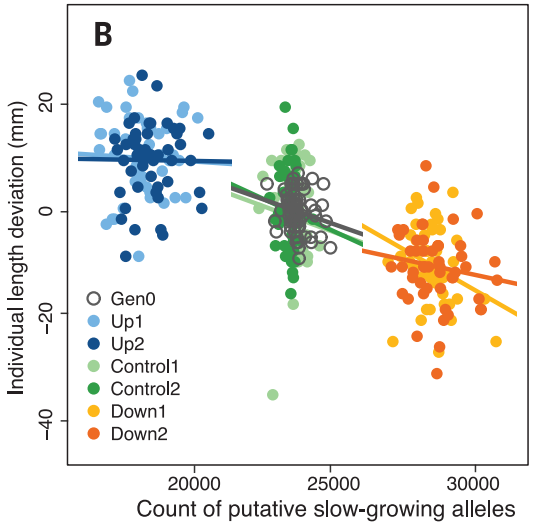
Individual-level analysis

Enables individual genotype-phenotype association analysis

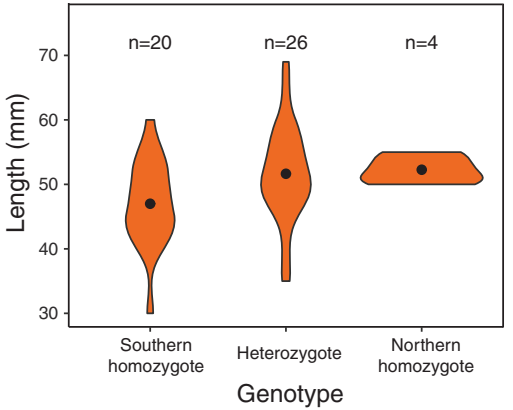
~50 silversides per population in generations 0 and 5

Low-coverage ‘in silico’ exome capture (~1.3x)

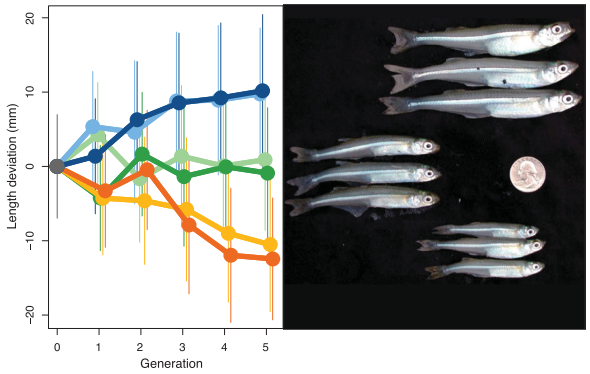
Polygenic score



Haplotyping inversions



Therkildsen et al. 2019. Science



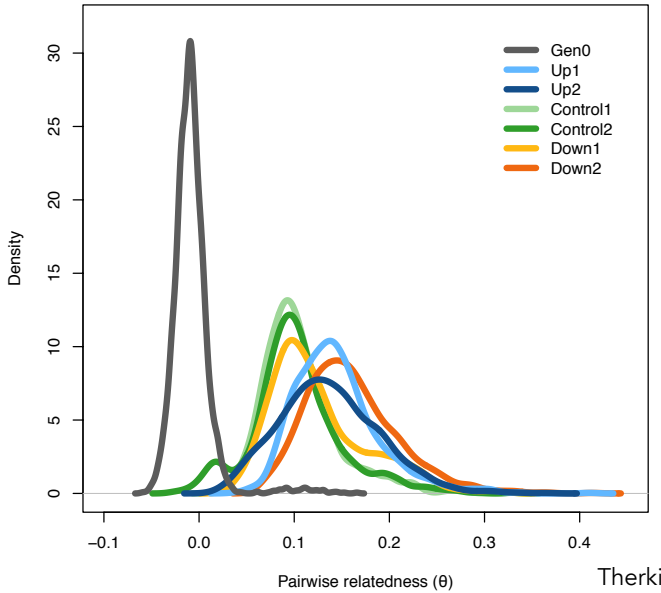
Individual-level analysis

Enables computation of relatedness and individual diversity statistics

~50 silversides per population in generations 0 and 5

Low-coverage ‘in silico’ exome capture (~1.3x)

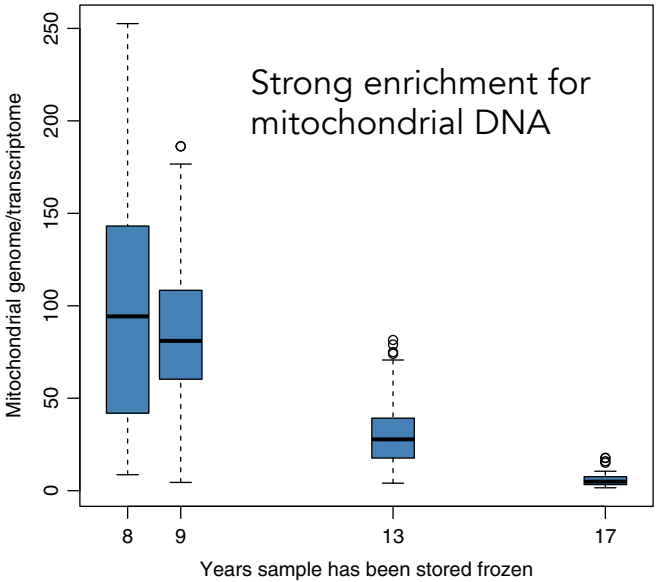
Pairwise relatedness



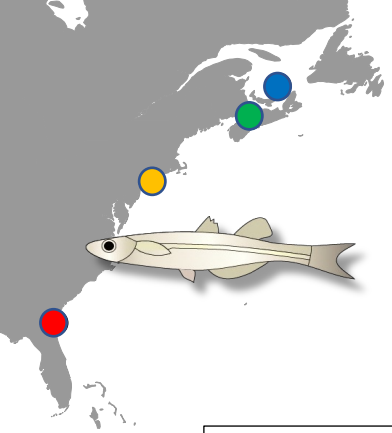
Therkildsen et al. 2019. Science

Individual-level analysis

Enables recovery of full mitochondrial genome sequences

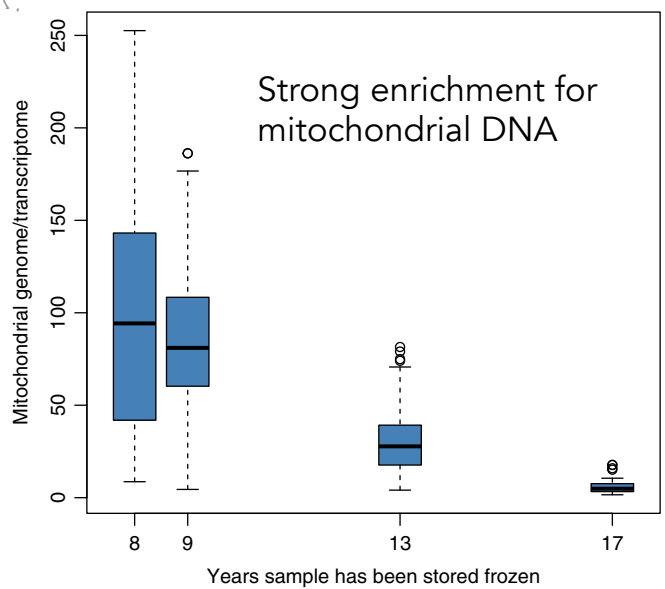


Therkildsen and Palumbi 2017. Mol. Ecol.

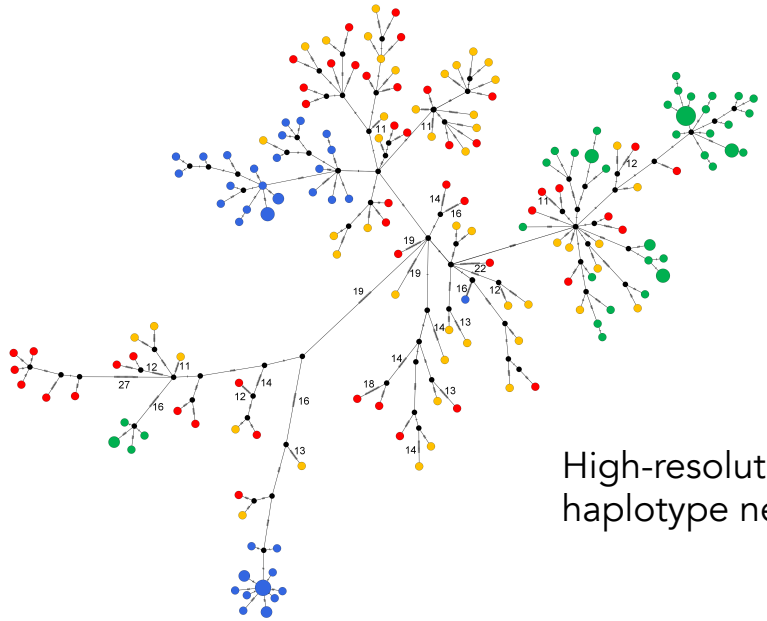


Individual-level analysis

Enables recovery of full mitochondrial genome sequences



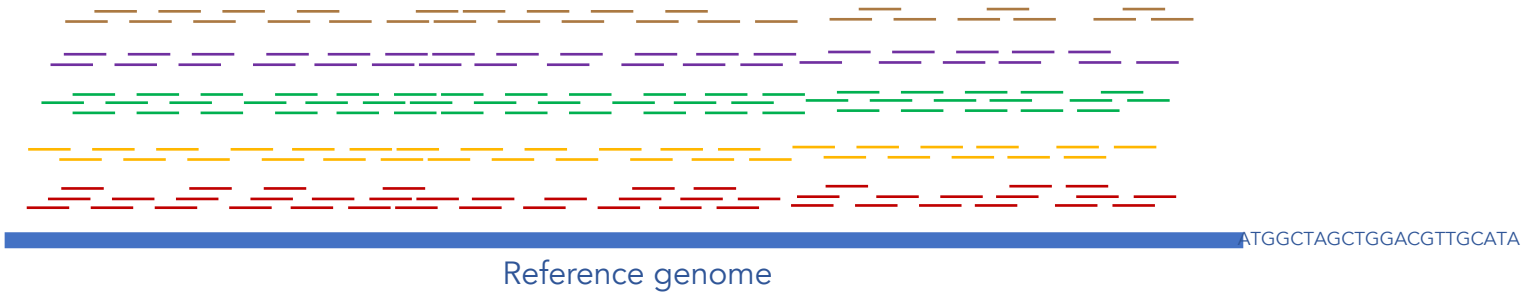
Therkildsen and Palumbi 2017. Mol. Ecol.



High-resolution haplotype network

Lou et al. 2018. Mar. Biol.

Cheap reads are short –
We need a reference sequence for mapping



Cheap reads are short –
We need a reference sequence for mapping

- *Do novo* genome assembly has become much more accessible even on modest budgets

- The literature is getting flooded with new genome assemblies
  - But the quality of assemblies varies widely
    - Contiguity
    - Accuracy

From the pre-course survey

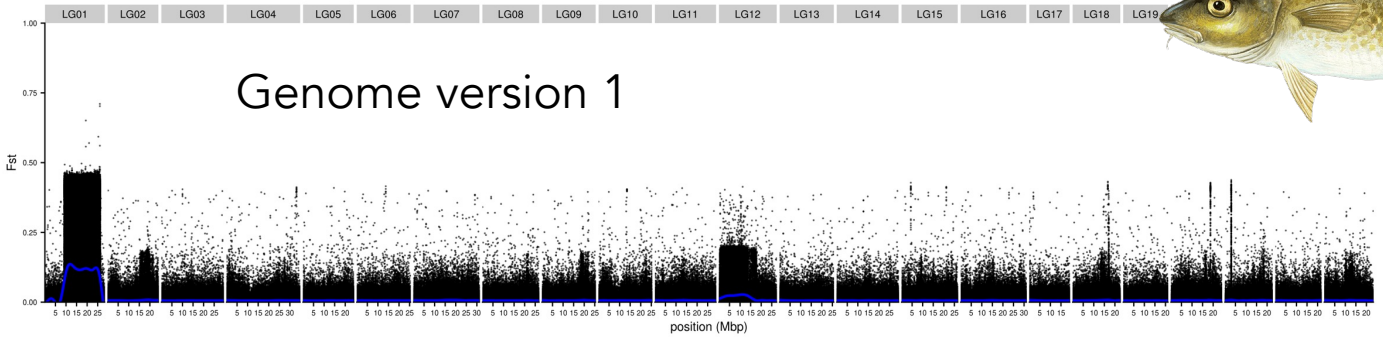
Is there a reference genome available for your primary study species?

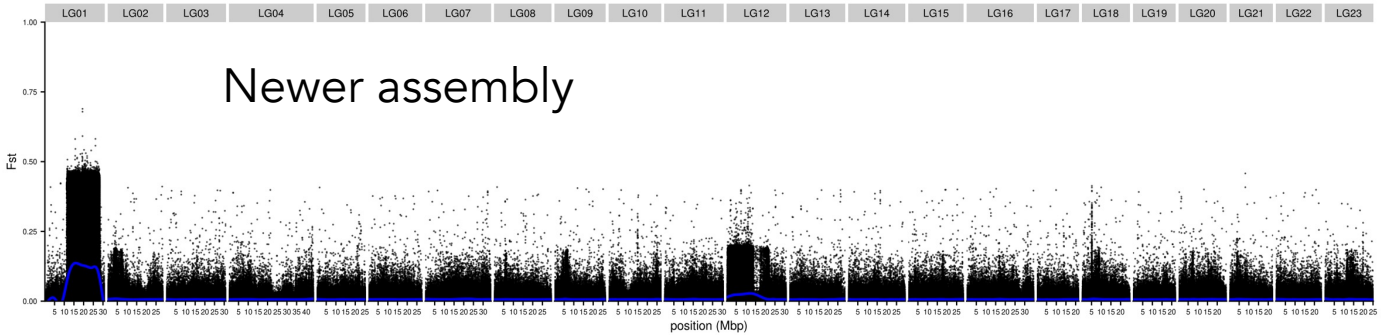
25 responses

Discussion in breakout rooms

- How important is it to have a high-quality genome assembly for successful analysis of low-coverage whole genome re-sequencing data?

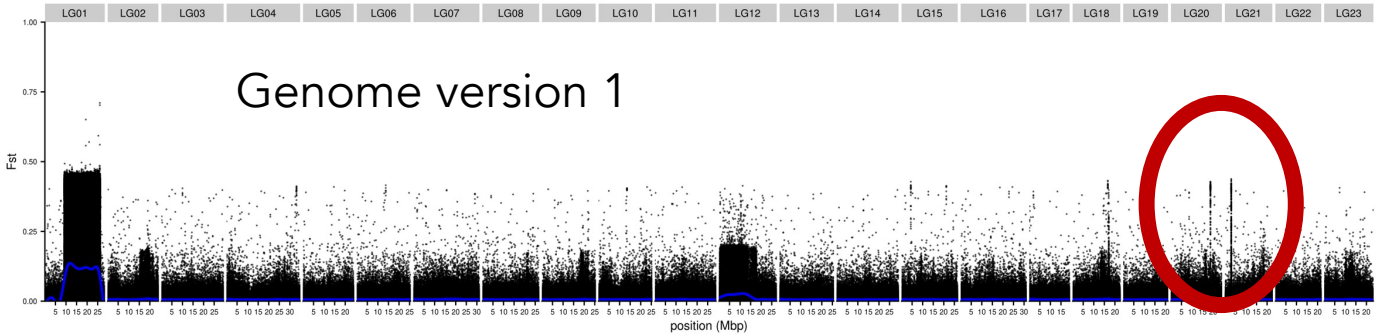
Examples from Atlantic cod

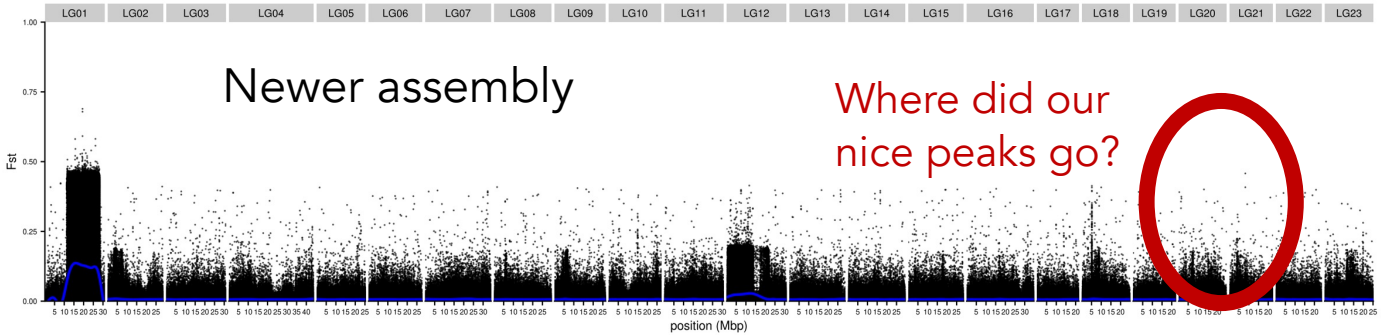




Nicolas Lou, unpubl.

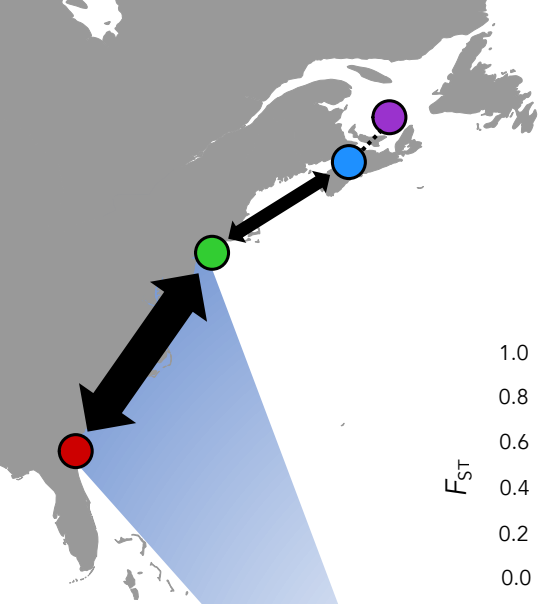
Examples from Atlantic cod



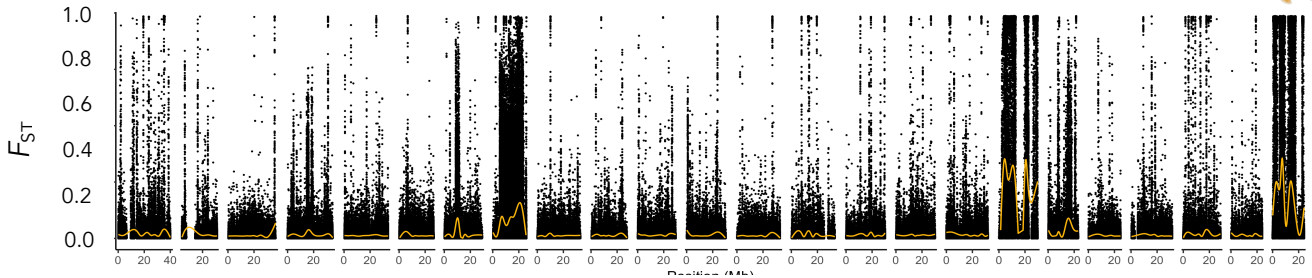


Nicolas Lou, unpubl.

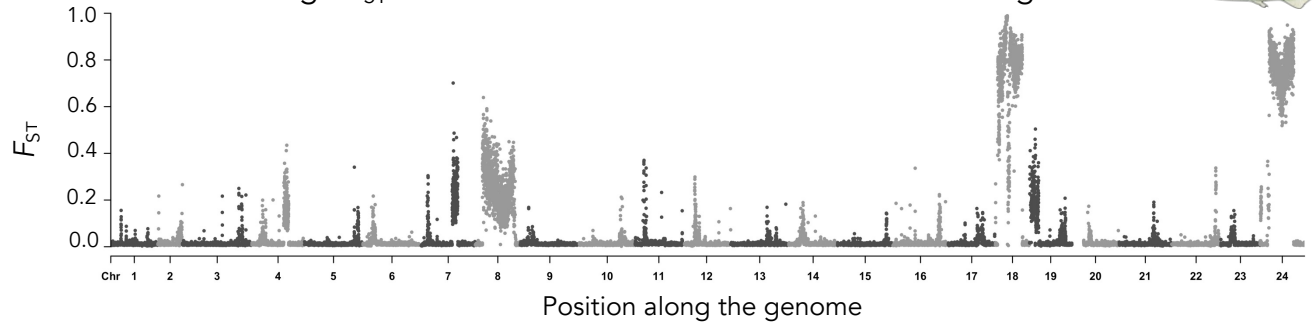
Heterogeneous gene flow and divergence patterns across latitudes



F_{ST} for transcriptome SNPs anchored to the medaka genome



Average F_{ST} for SNPs in 15kb windows on the silverside genome



Highly fragmented genome assembly

Comparison of white and summer suckers

Nicolas Lou, unpubl.

Highly fragmented genome assembly

How do we even begin to make sense of this?

Comparison of white and summer suckers

Nicolas Lou, unpubl.

New technologies facilitate highly contiguous genome assemblies

- Long-read (PacBio and Nanopore) and linked-read (10X and haplotagging) sequencing
- Hi-C
- Optical mapping

VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

PHASE I | TECHNOLOGY | WHO WE ARE | PUBLICATIONS | NEWS | EVENTS | WAYS TO HELP | CONTACT | ERGA | DONATE

The Vertebrate Genomes Project introduces a new era of genome sequencing

The VGP reports its first discoveries towards generating near error-free...all living vertebrate species in a special issue of Nature (reported April 28, 2021)

LIST OF~70,000 EXTANT VERTEBRATE SPECIES

GENOME ARK — All VGP data, raw reads, mito assemblies

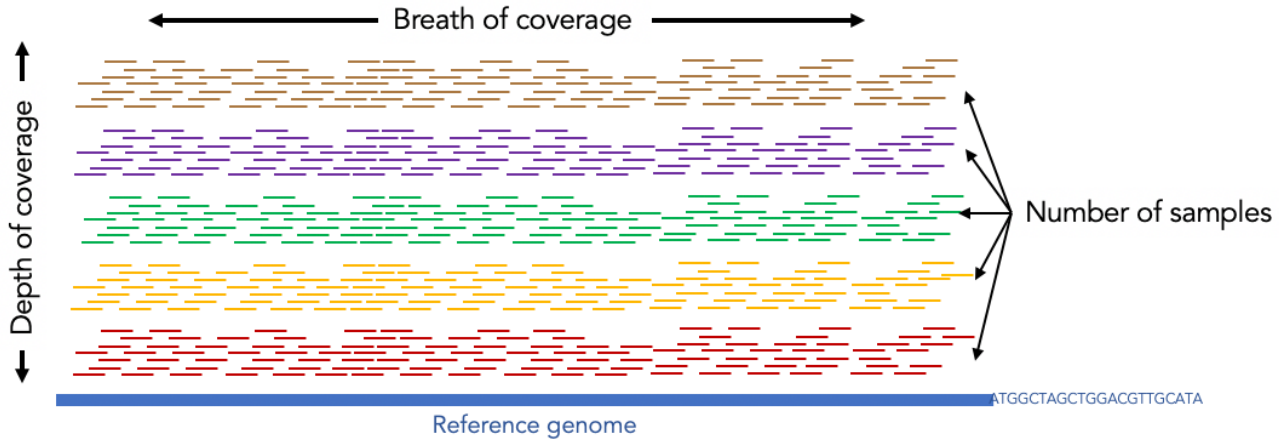
Set of primary assemblies — UCSC VGP GB

VGP ENSEMBL — Annotations

Annotations — VGP GENBANK

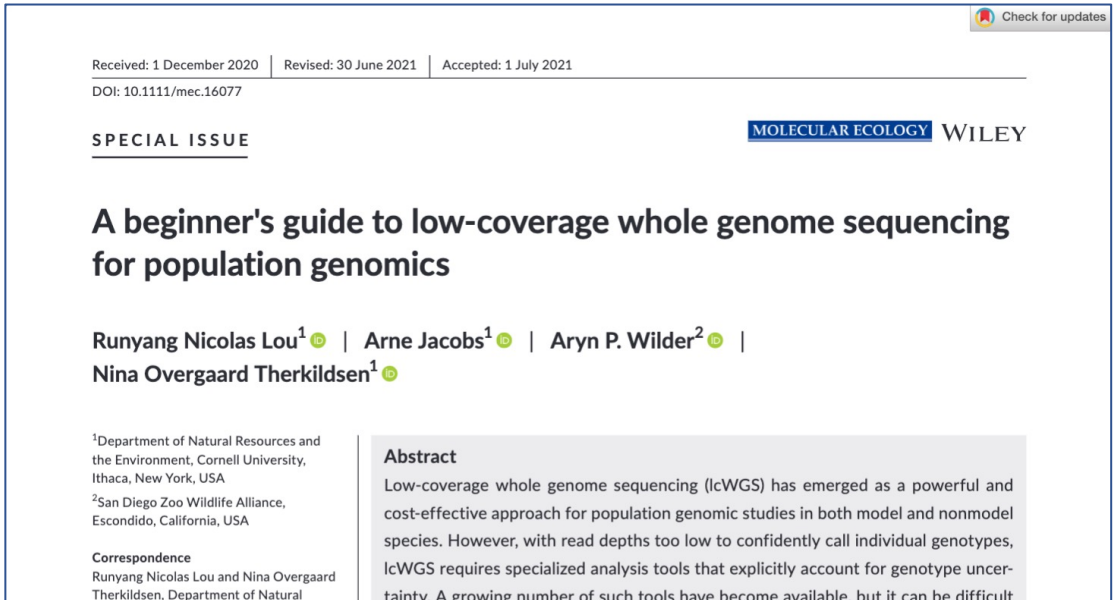
Experimental design for lcWGS

- Trade-off between depth per individual and number of individuals
- How low should we go?



Experimental design for lcWGS

- Trade-off between depth per individual and number of individuals
- How low should we go?



Check for updates

Received: 1 December 2020 | Revised: 30 June 2021 | Accepted: 1 July 2021

DOI: 10.1111/mec.16077

SPECIAL ISSUE

MOLECULAR ECOLOGY WILEY

A beginner's guide to low-coverage whole genome sequencing for population genomics

Runyang Nicolas Lou[1] | Arne Jacobs[1] | Aryn P. Wilder[2] | Nina Overgaard Therkildsen[1]

[1]Department of Natural Resources and the Environment, Cornell University, Ithaca, New York, USA

[2]San Diego Zoo Wildlife Alliance, Escondido, California, USA

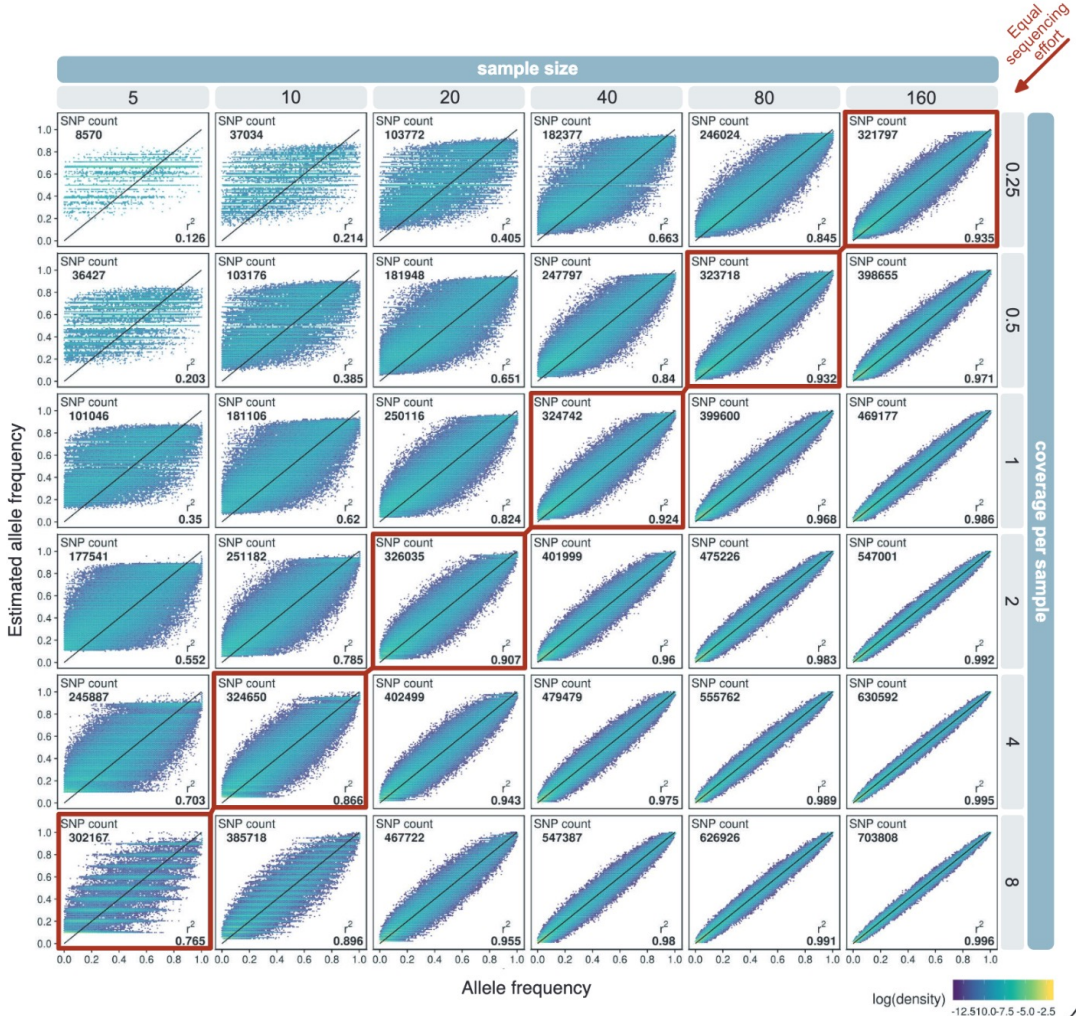
Correspondence
Runyang Nicolas Lou and Nina Overgaard Therkildsen, Department of Natural

Abstract

Low-coverage whole genome sequencing (lcWGS) has emerged as a powerful and cost-effective approach for population genomic studies in both model and nonmodel species. However, with read depths too low to confidently call individual genotypes, lcWGS requires specialized analysis tools that explicitly account for genotype uncer-

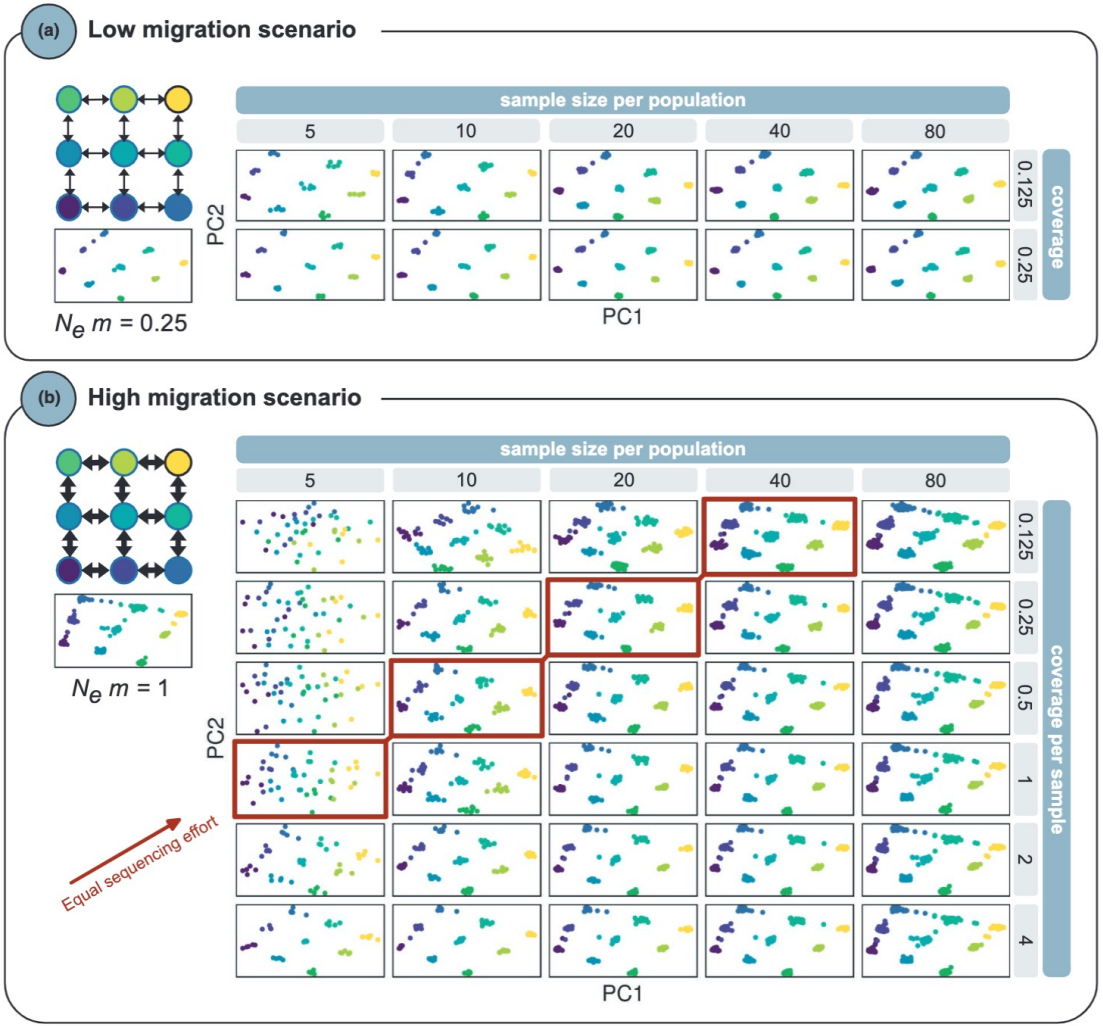
Allele frequency

Lower coverage of more individuals provides more accurate estimates



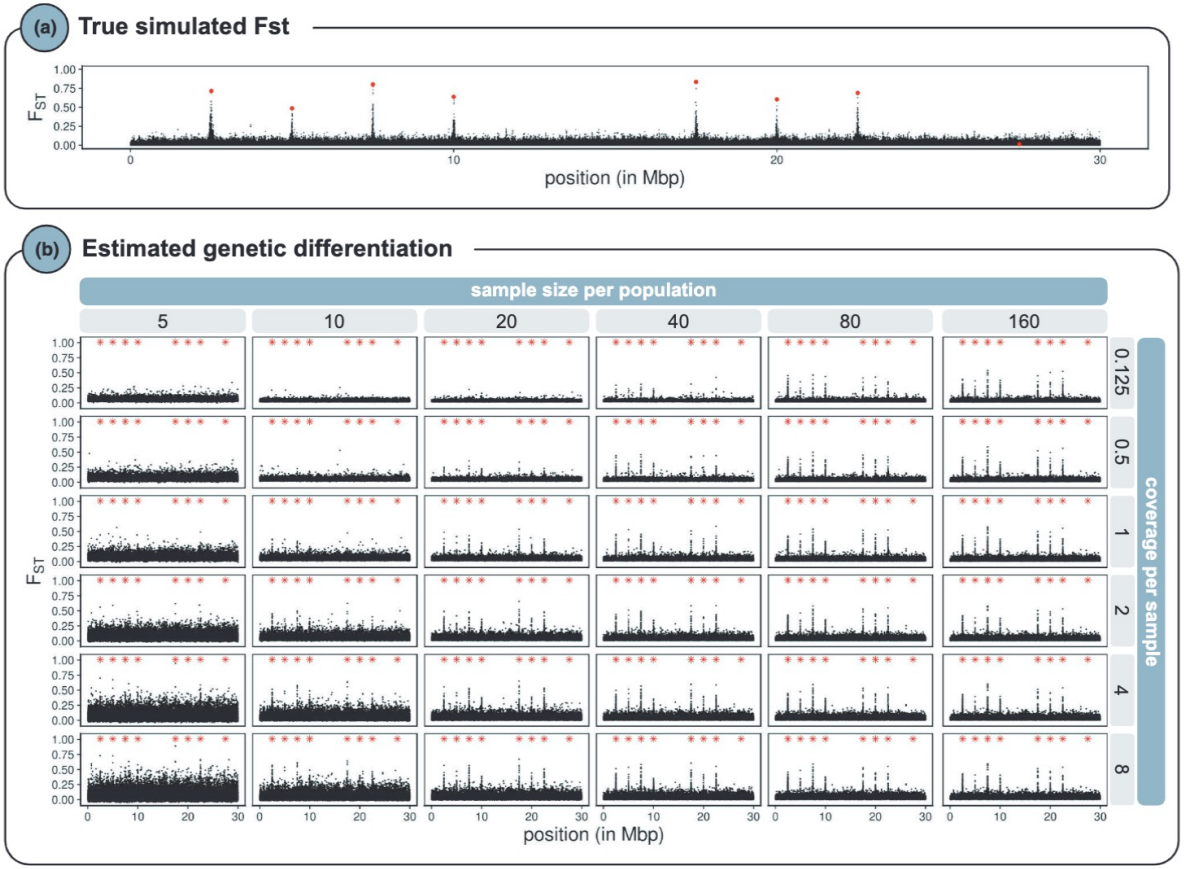
PCA

As long as the sample size is sufficient, population structure is detected even at 0.125x coverage



Outlier detection

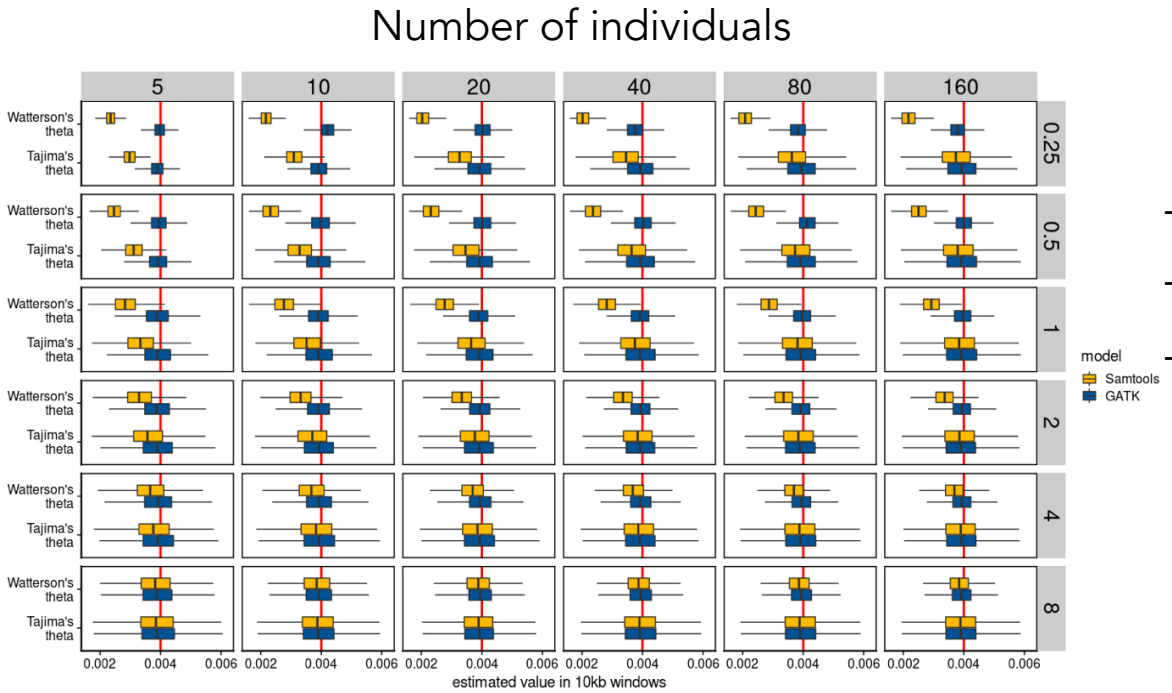
Peaks can be detected even at low coverage when the number of individuals is sufficient



Diversity statistics

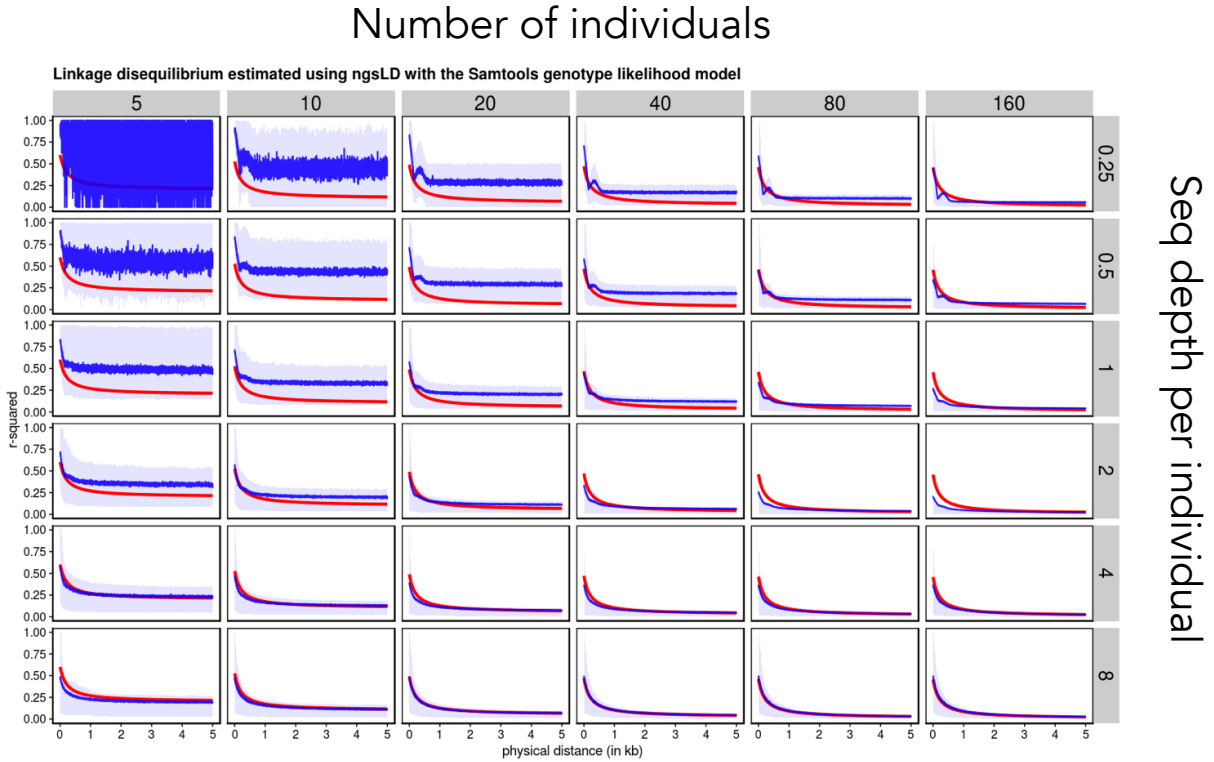
Estimates can be sensitive to the genotype likelihood (GL) model when depth is < ~2-4x

But note that Tajima's theta, aka nucleotide diversity (π) is only sensitive to the GL model at very low depths. Other diversity statistics more strongly dependent on rare alleles (including the site frequency spectrum) are more sensitive



Linkage disequilibrium

The absolute value can be overestimated when depth is < ~2-4x, but relative values can be informative at lower depths



There is no design that is ideal for all purposes

There is no design that is ideal for all purposes

- Tradeoffs – for example

  - Accuracy in allele frequency estimation
    (e.g. for analysis of population structure or outlier detection)
    [spreading sequencing over as many individuals as possible is better]

    and

  - Accuracy in estimation of diversity statistics dependent on rare alleles
    [ensuring ~4x depth per sample is better]

There is no design that is ideal for all purposes

- Tradeoffs – for example
  - Accuracy in allele frequency estimation
    (e.g. for analysis of population structure or outlier detection)
    [spreading sequencing over as many individuals as possible is better]

    and

  - Accuracy in estimation of diversity statistics dependent on rare alleles
    [ensuring ~4x depth per sample is better]

Think about what the primary goal of your study is
and whether absolute vs. relative estimates are needed

Guide to experimental design Lou et al. 2021. Mol Ecol

Type of analyses	Examples	Recommendations on experimental design
Allele frequency and differentiation	Population allele frequencies, most genotype–environment association analysis (GEA) methods, F_{ST} (as implemented, e.g., in VCFLIB), pF_{ST}	Prioritize larger sample sizes, ≥10 samples per population, ≥10× coverage per population, (Figures 3 and B2). Avoid uneven sample size for estimation of F_{ST} (Berner, 2019)
SFS-based analyses (absolute estimation of rare-allele-dependent metrics)	Absolute estimation of Watterson's θ, Tajima's D, individual heterozygosity. Reconstruction of demographic history (e.g., DADI)	Prioritize higher coverage per sample, >4× coverage per sample, ≥5 samples per population (Figures S2 and S3).
SFS-based analyses (relative estimation of rare-allele-dependent metrics, or nonrare-allele-dependent metrics)	Relative estimation of Watterson's θ and Tajima's D (e.g., for outlier scans). π, d_{xy}, F_{ST} (as implemented in ANGSD)	Prioritize larger sample sizes, ≥10 samples per population, ≥10× coverage per population, (Figure 5; Figures S2, S3, S7, S14–S17). Avoid uneven sample size for estimation of F_{ST} (Figure S7, see also Berner, 2019).
Population structure	PCA, admixture analysis	Prioritize larger sample sizes, ≥10 samples per population, extremely low per-sample coverage (e.g., 0.125×, Figure 4; Figures S6 and S11) or highly uneven per-sample coverage (e.g., ranging from 0.5× to 6×, Skotte et al., 2013) can be viable.
Absolute estimation of linkage disequilibrium	LD decay rate, demographic inference	Prioritize higher coverage per sample, ≥4× coverage per sample, ≥20 samples per population, (Figures S4 and S5; Bilton et al., 2018; Fox et al., 2019; Maruki & Lynch, 2014).
Relative estimation of linkage disequilibrium	LD pruning, LD block identification	Per-sample coverage as low as 1× could be viable, ≥20× coverage per population (Figures S4 and S5).
Genotype imputation without reference panels	STITCH, BEAGLE	STITCH: prioritize larger sample size (≥500) over per-sample coverage (1× could be sufficient) BEAGLE: prioritize higher per-sample coverage (≥2×) over sample size (≤250 could be sufficient) (Figure B4).