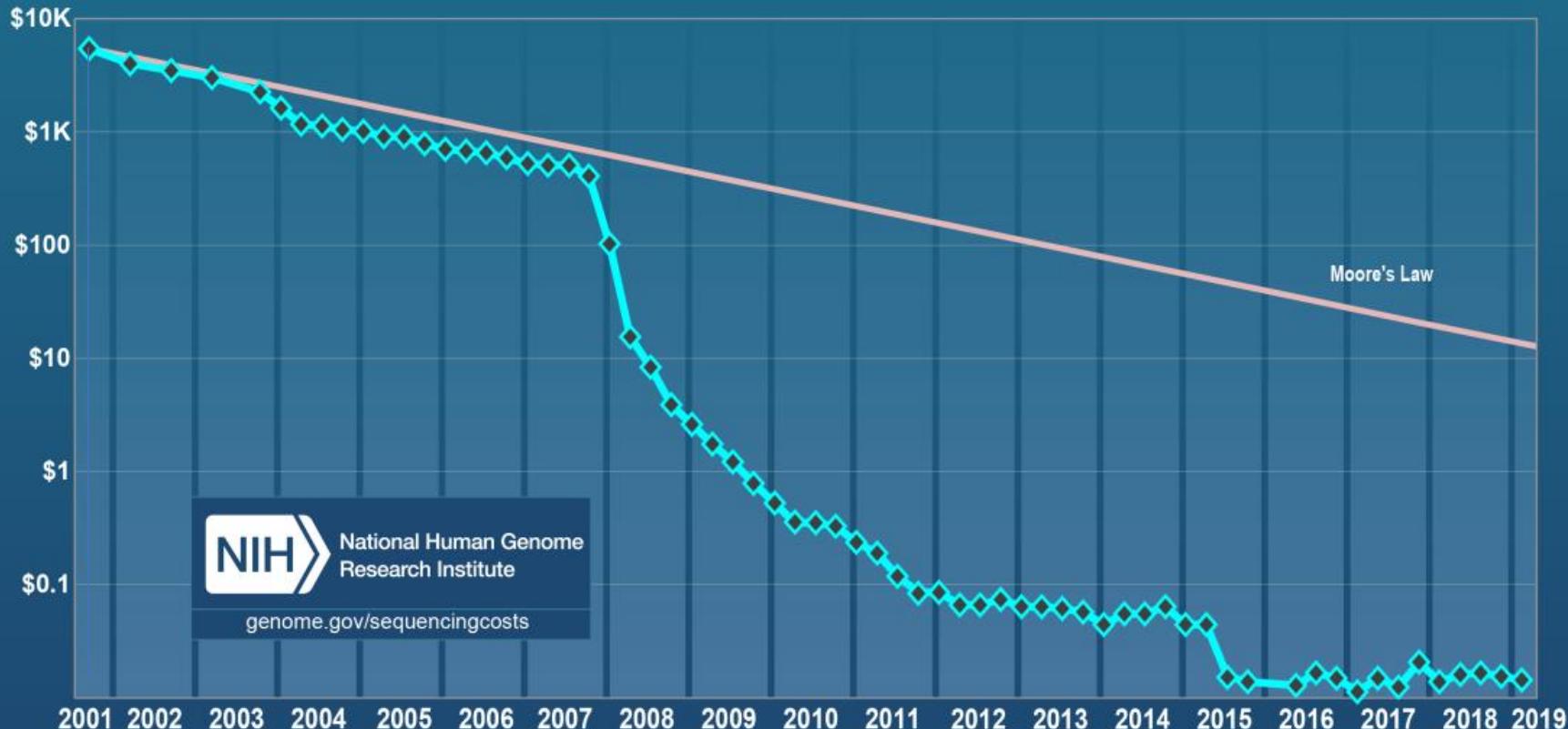


Introduction to low-coverage whole genome sequencing (lcWGS)

Plan for today

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

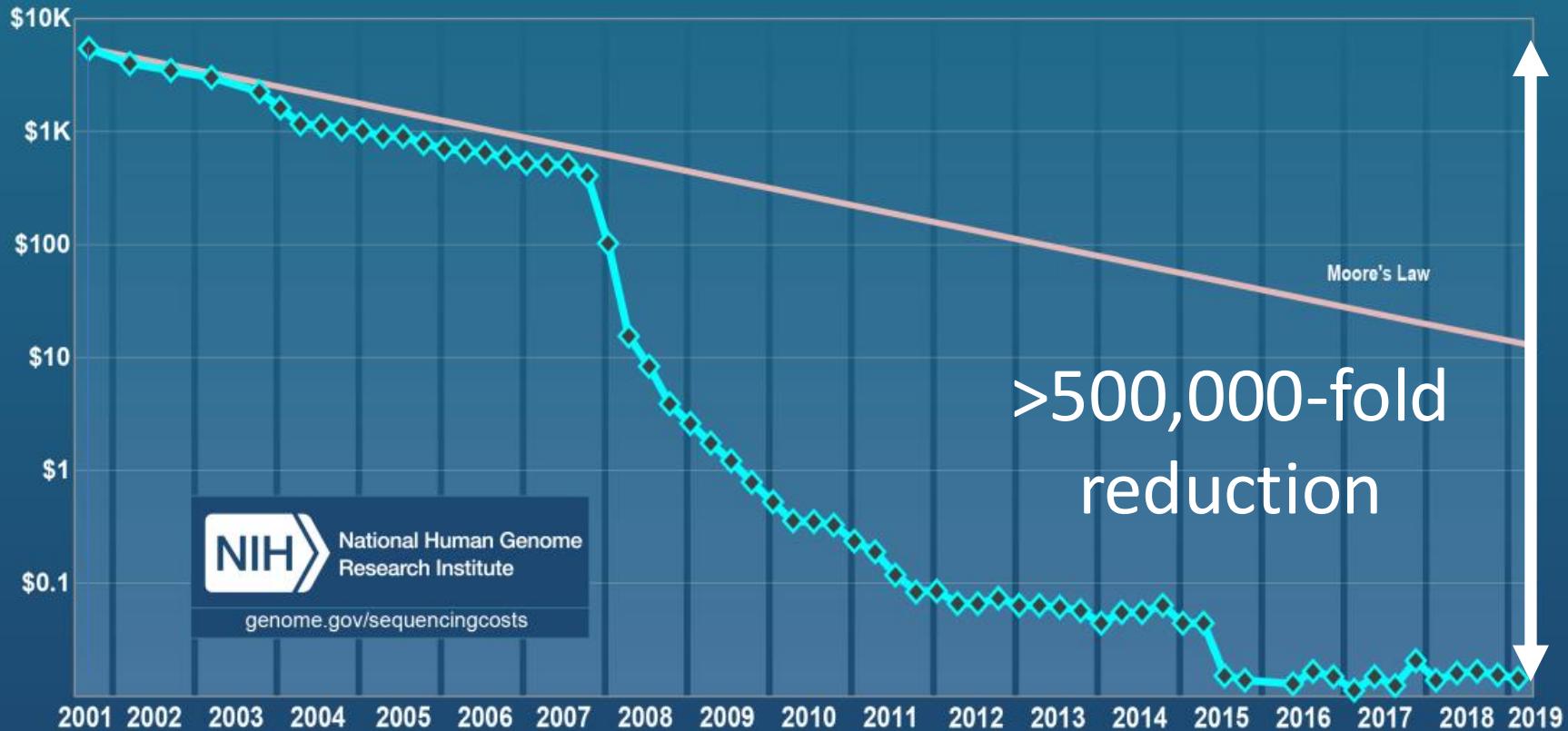
Cost per Raw Megabase of DNA Sequence



National Human Genome
Research Institute

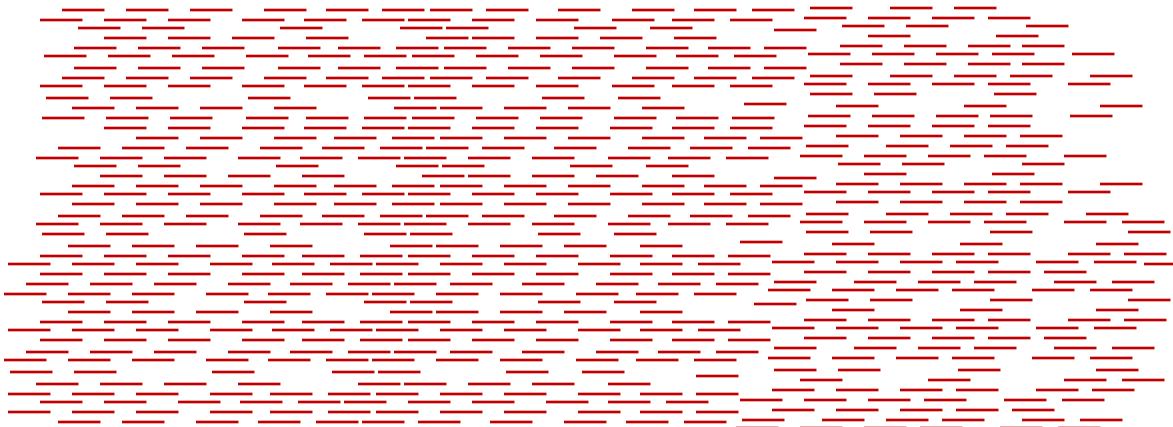
genome.gov/sequencingcosts

Cost per Raw Megabase of DNA Sequence

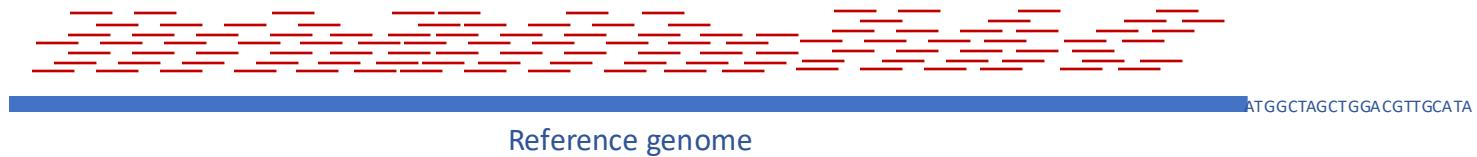


How to distribute sequencing effort?

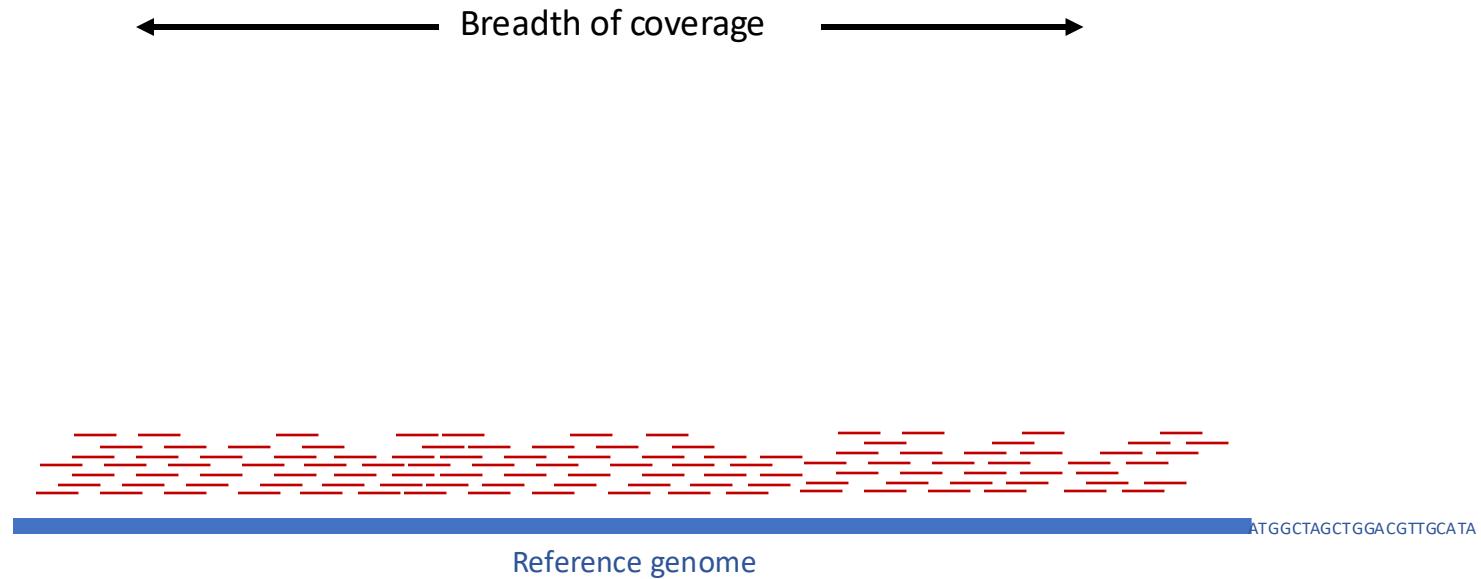
Raw sequence reads



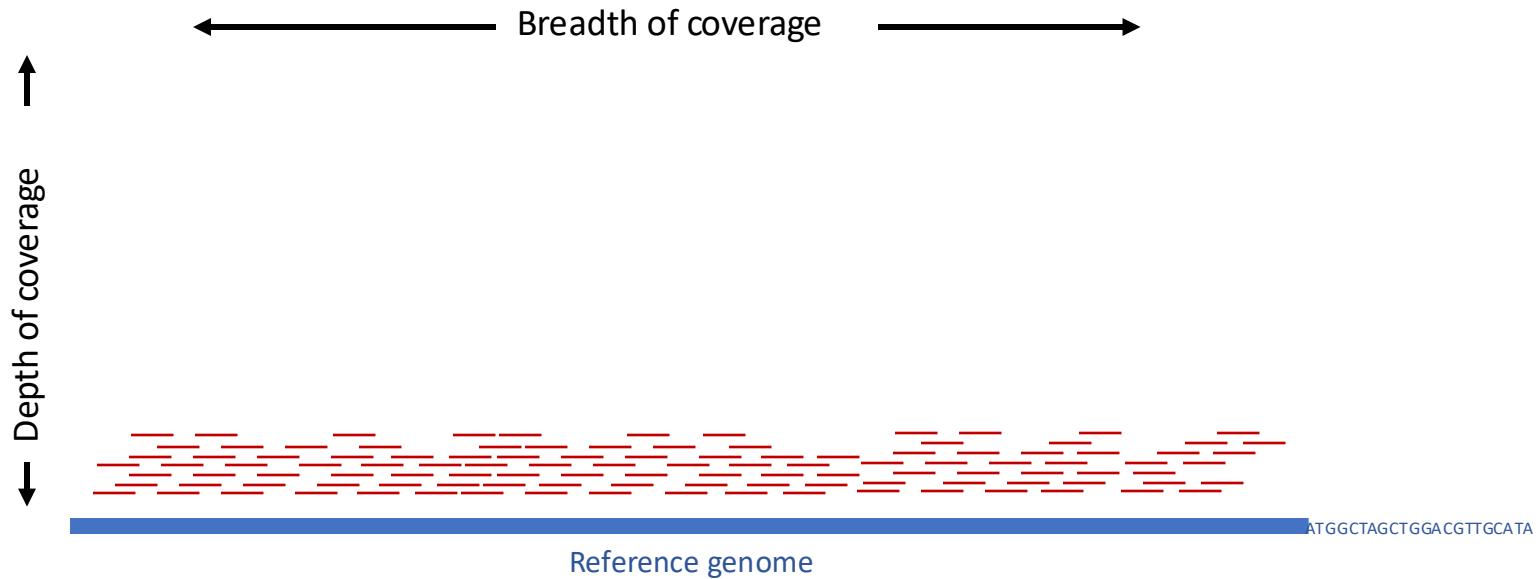
How to distribute sequencing effort?



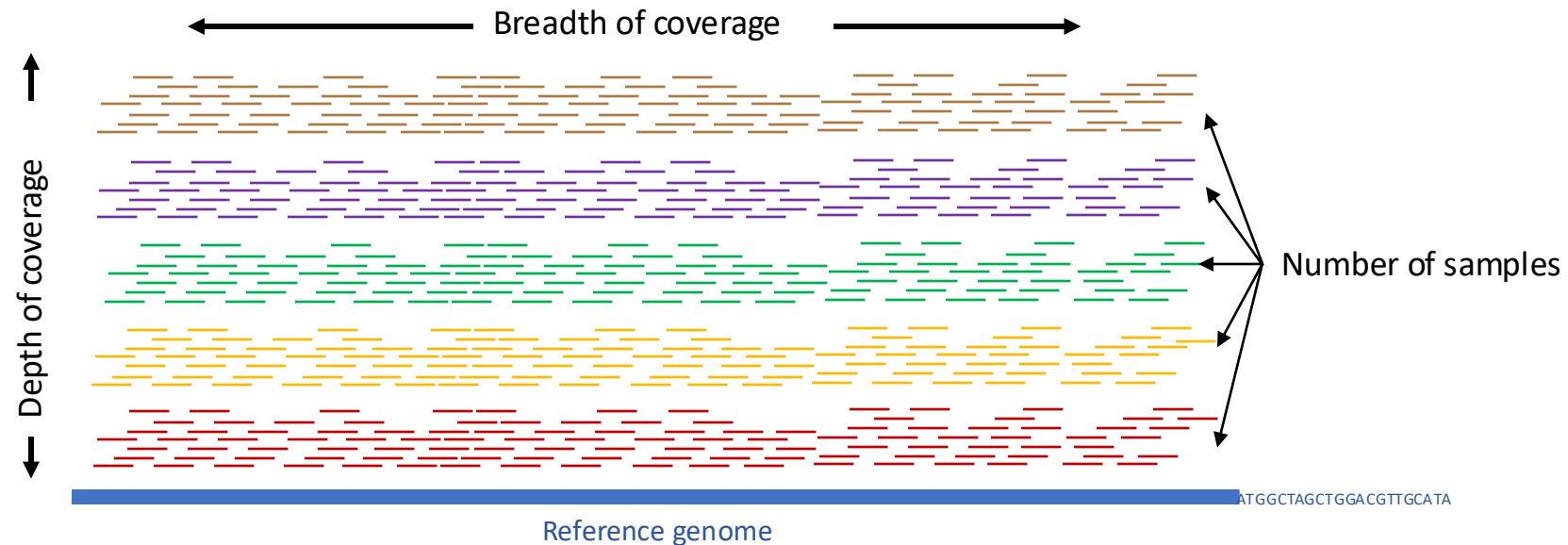
How to distribute sequencing effort?



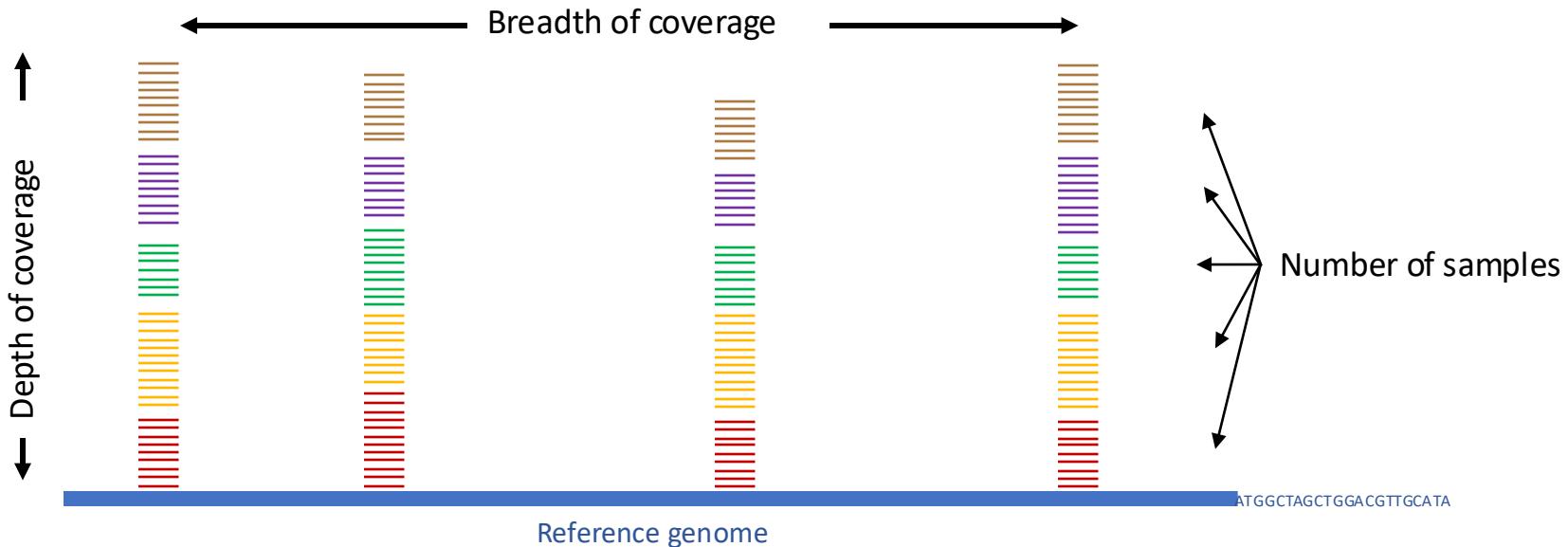
How to distribute sequencing effort?



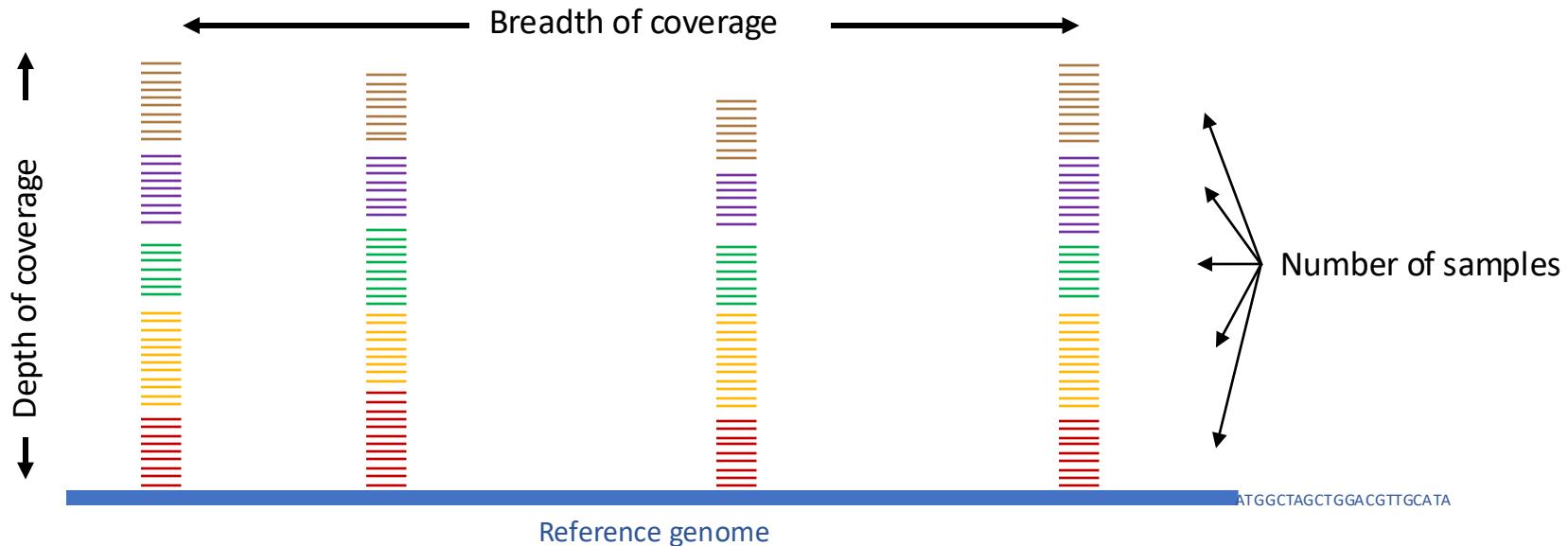
How to distribute sequencing effort?



RAD-seq approach

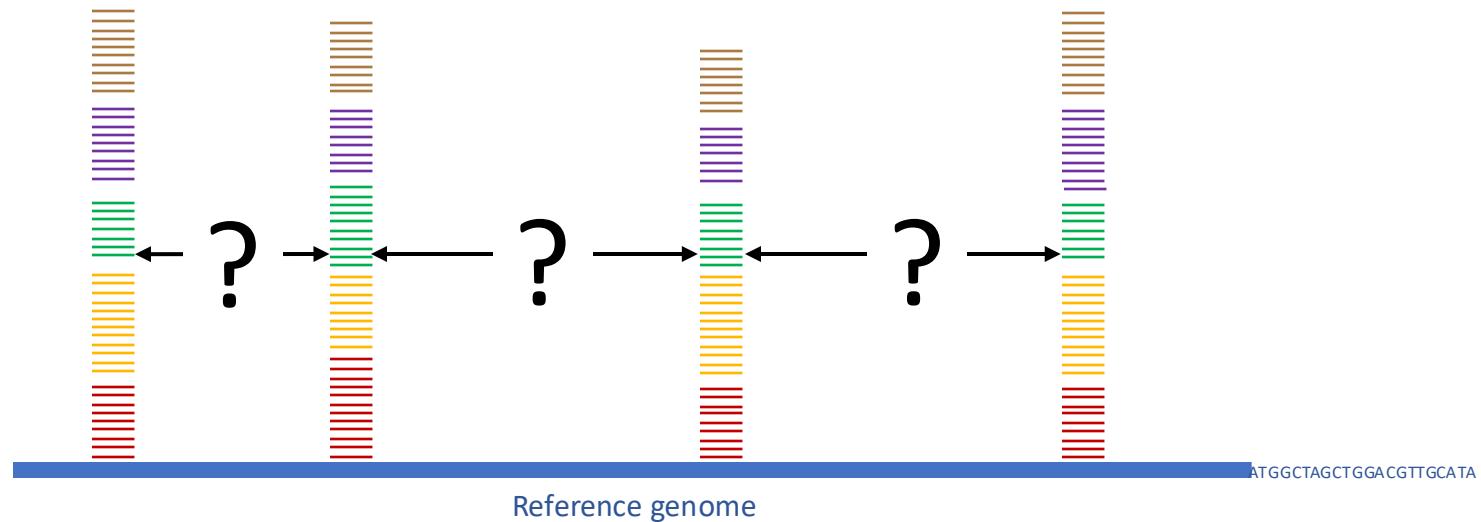


RAD-seq approach or SNP chip



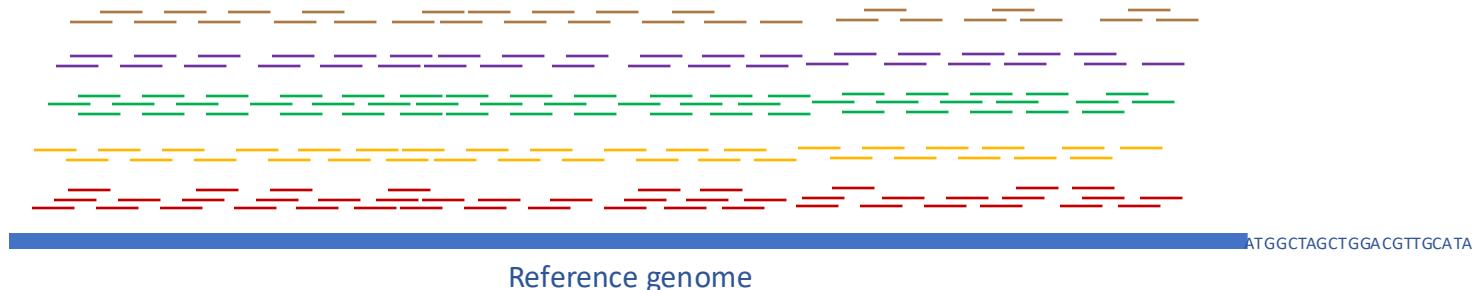
RAD-seq approach or SNP chip

What are we missing?



The value of trading depth for breadth

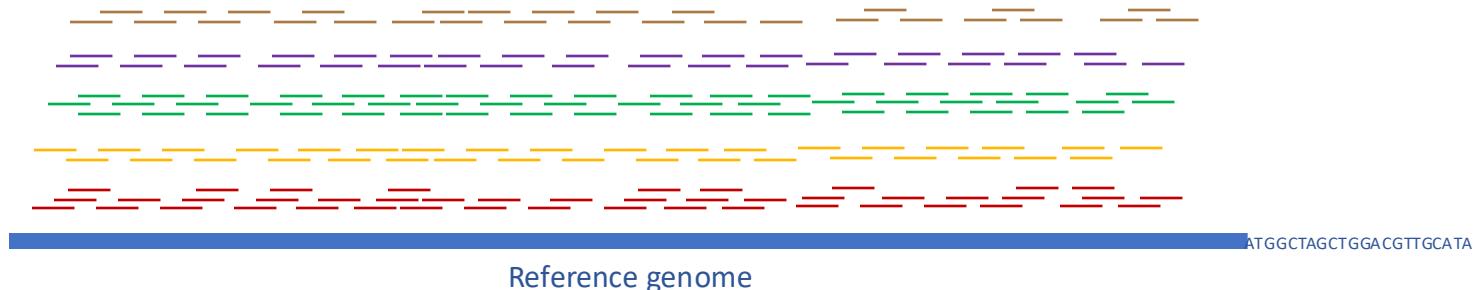
Maximizing the information content in our sequence data



The value of trading depth for breadth

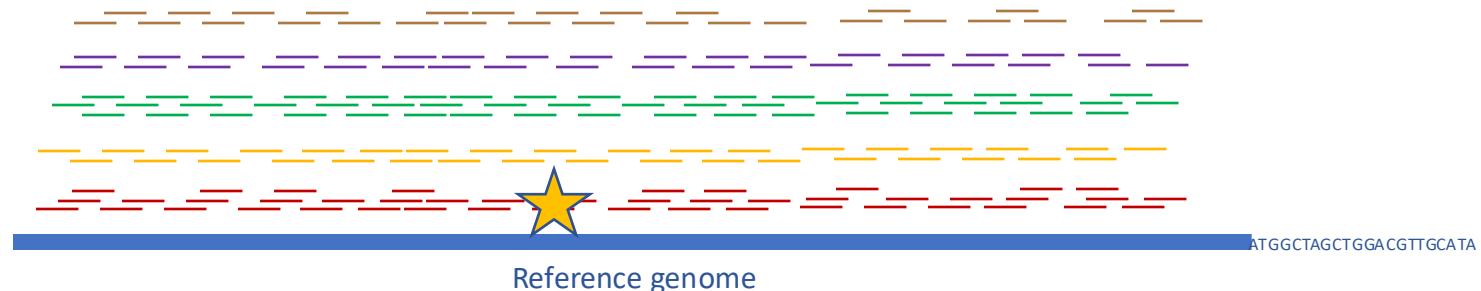
Maximizing the information content in our sequence data

But low-coverage precludes robust genotype calls



The value of trading depth for breadth

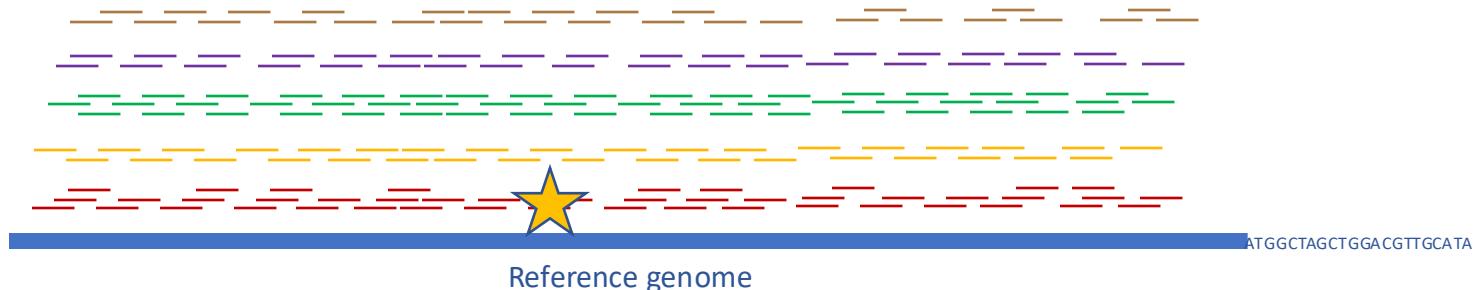
Do we care about the **genotype** at a **particular SNP** in a **particular individual**?



The value of trading depth for breadth

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

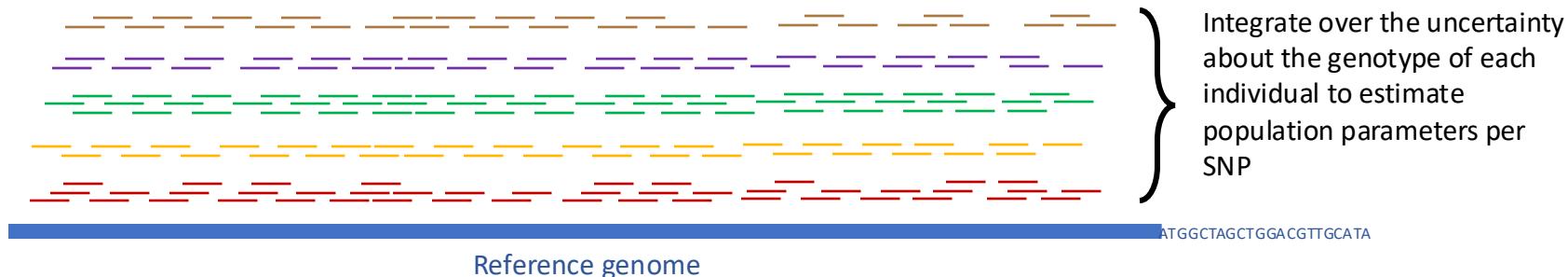
Often not



The value of trading depth for breadth

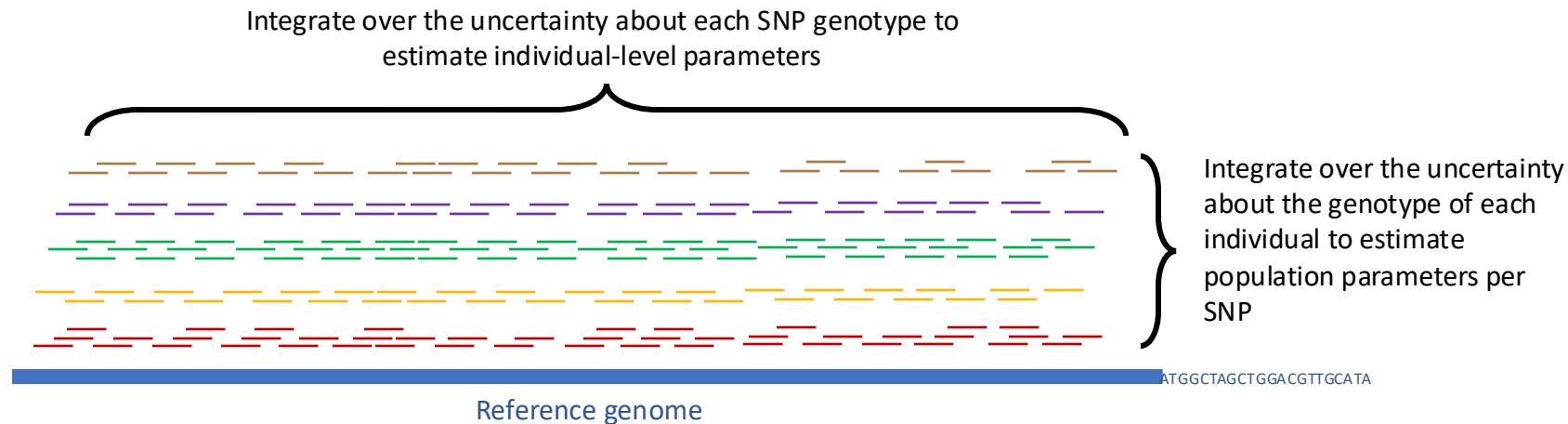
Do we care about the **genotype** at a **particular SNP** in a **particular individual**?

Often not



The value of trading depth for breadth

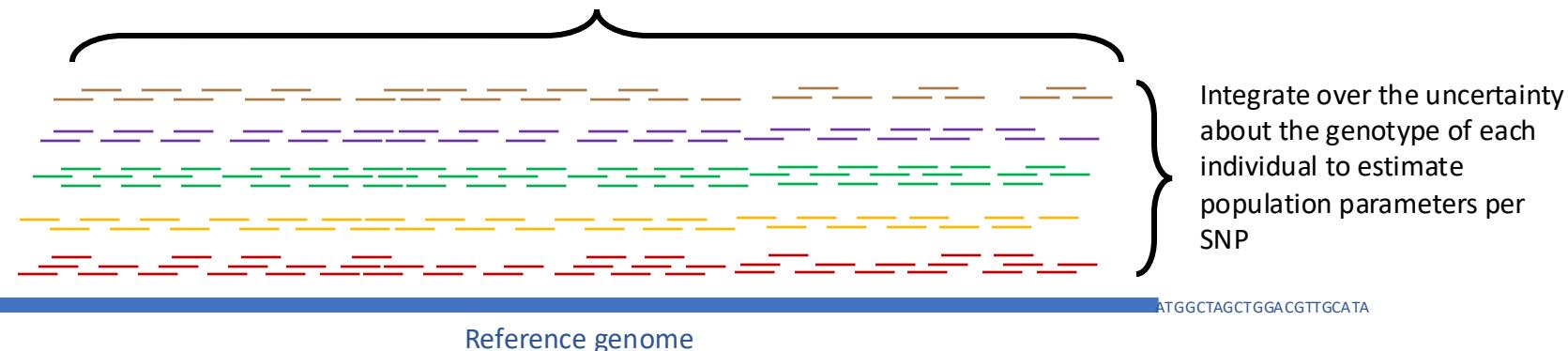
Do we care about the **genotype** at a **particular SNP** in a **particular individual**?



The value of trading depth for breadth

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

Integrate over the uncertainty about each SNP genotype to estimate individual-level parameters



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



Gemma Clucas



Nicolas Lou



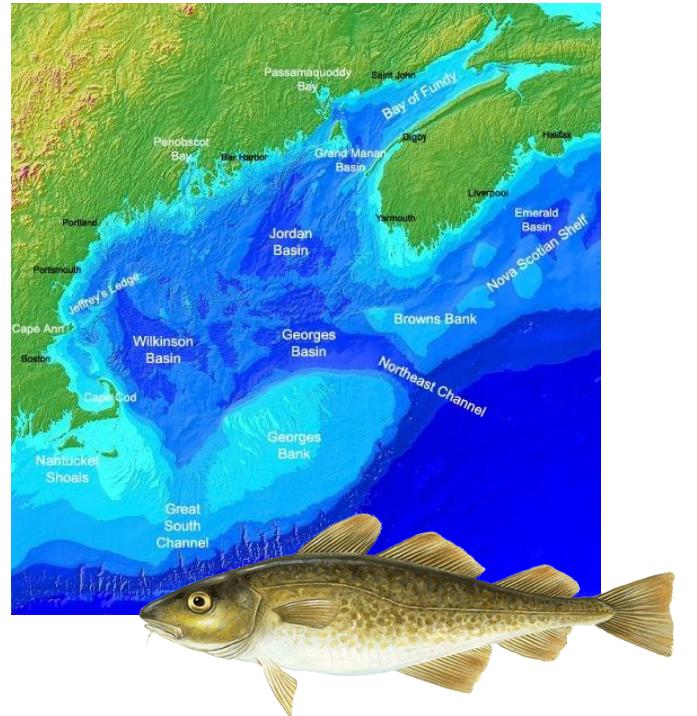
University of
New Hampshire



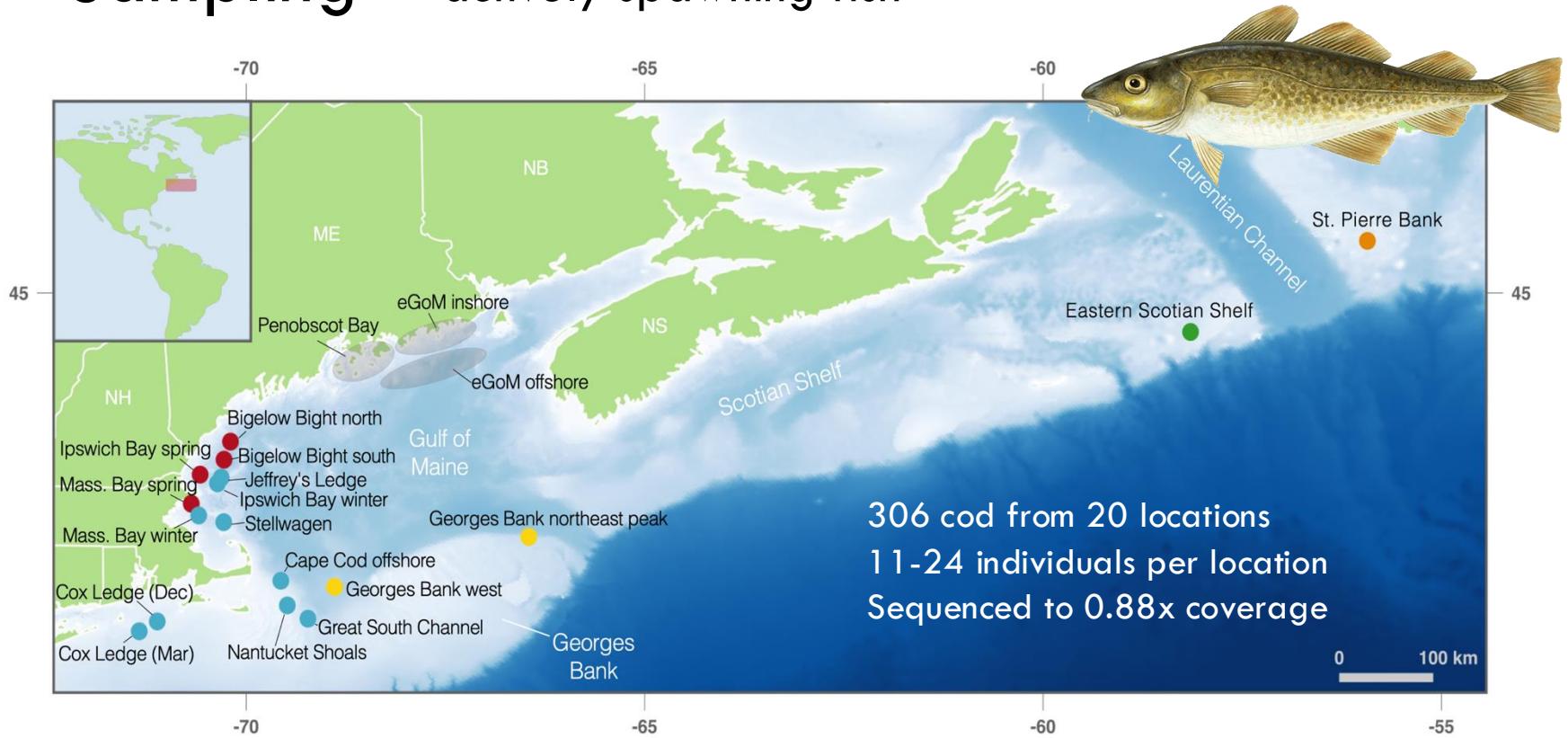
University of
New Hampshire



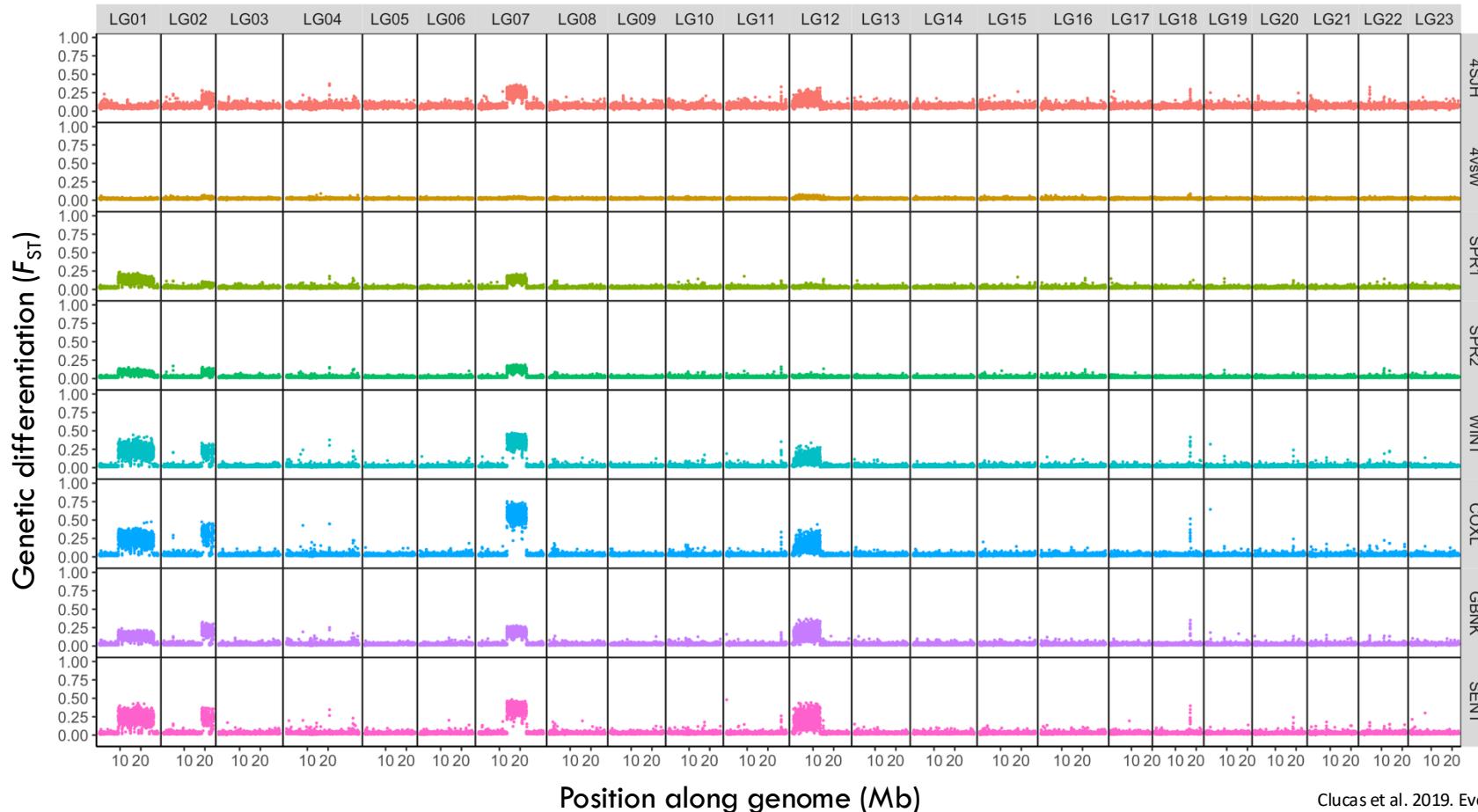
Cornell University



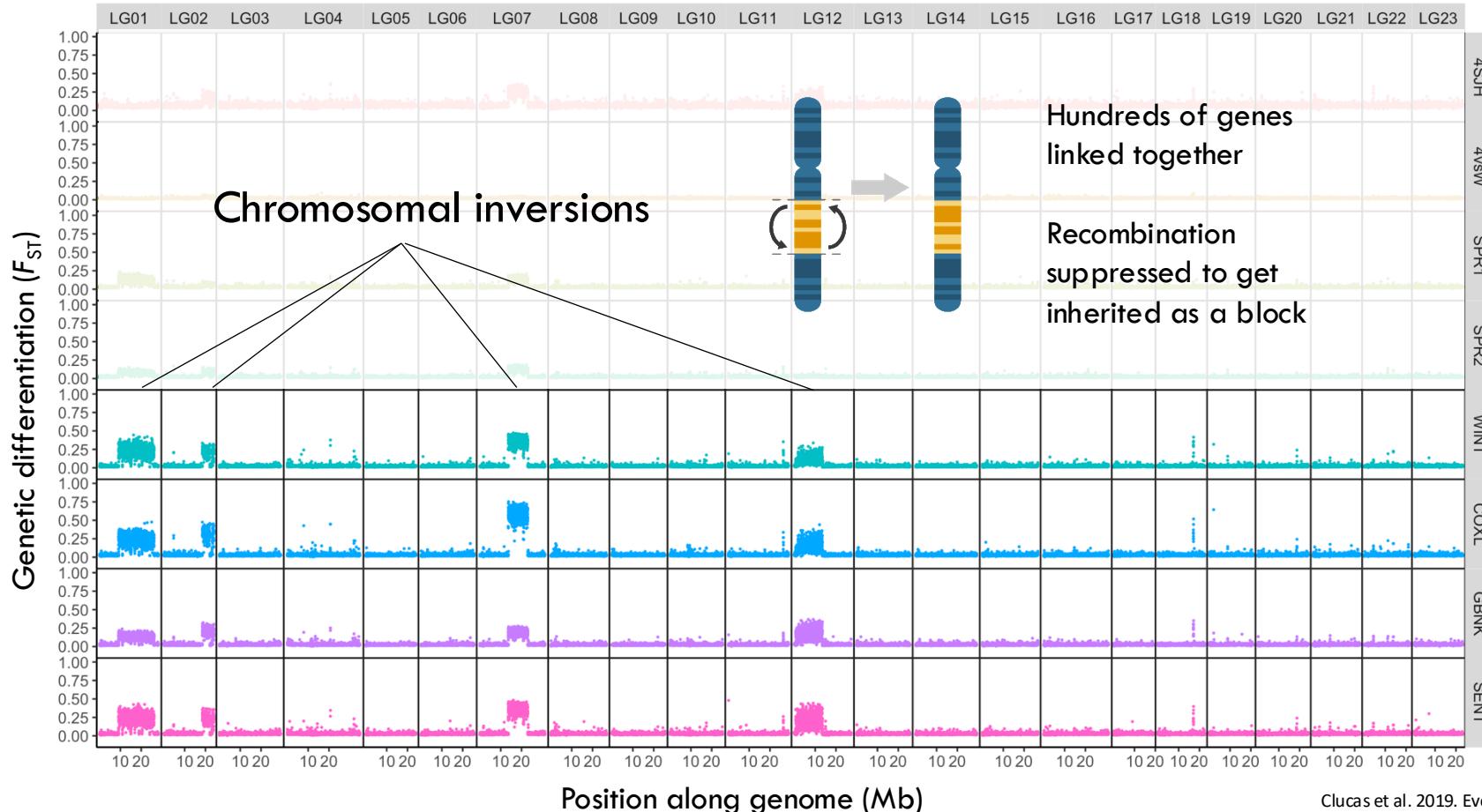
Sampling — actively spawning fish



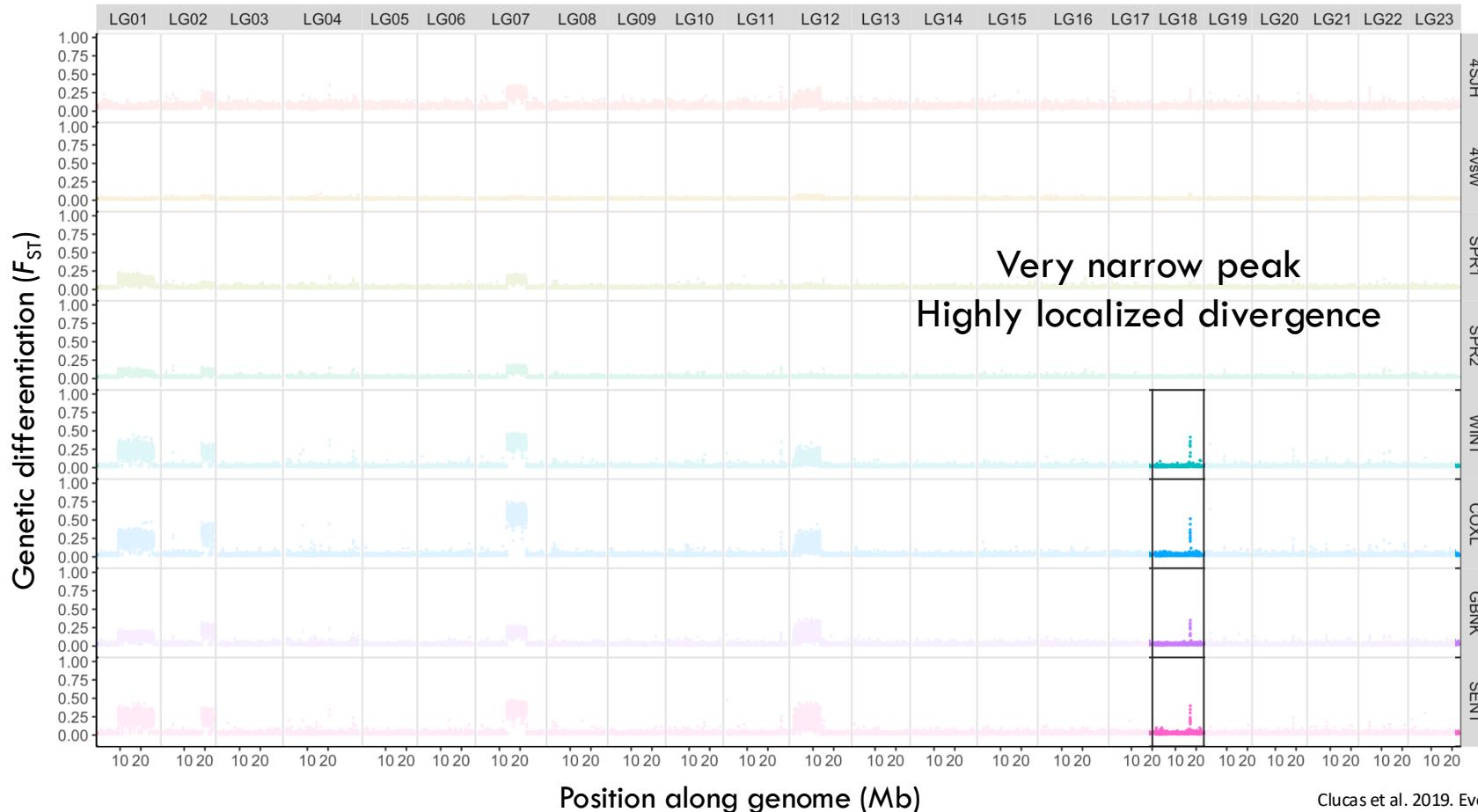
Pairwise comparisons to St. Pierre Bank, Canada (15 kb windows)



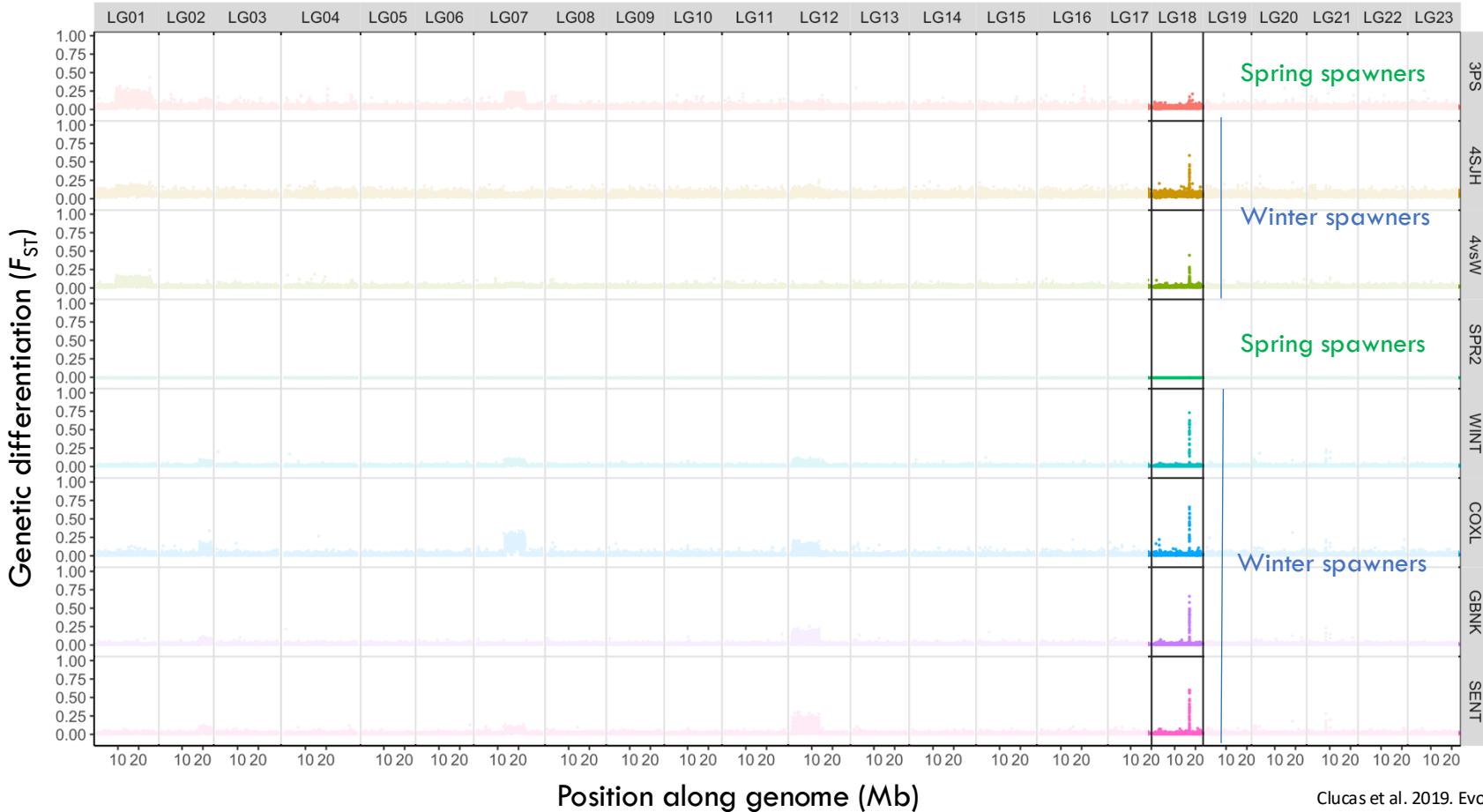
Pairwise comparisons to St. Pierre Bank, Canada (15 kb windows)



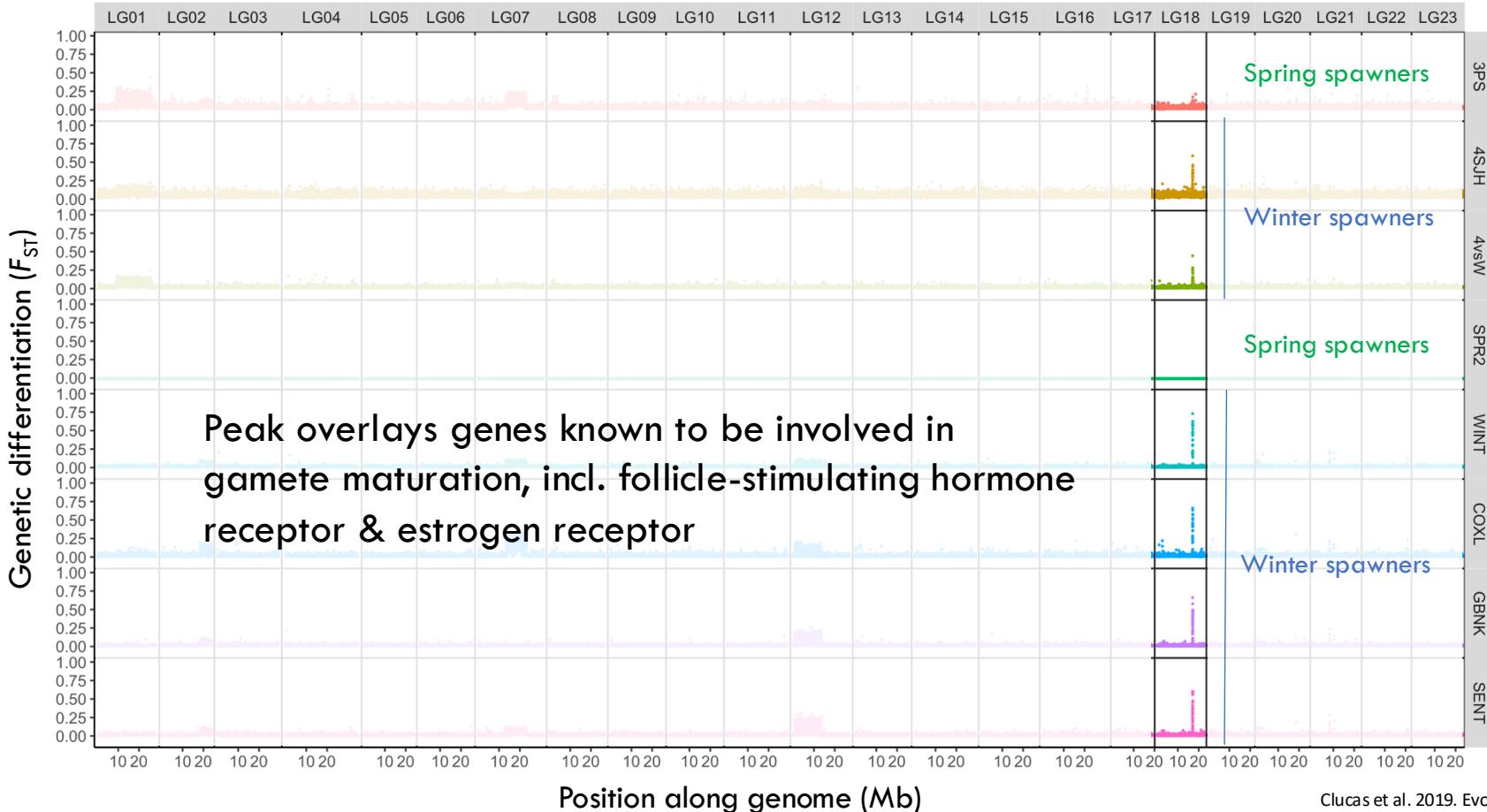
Pairwise comparisons to St. Pierre Bank, Canada (15 kb windows)



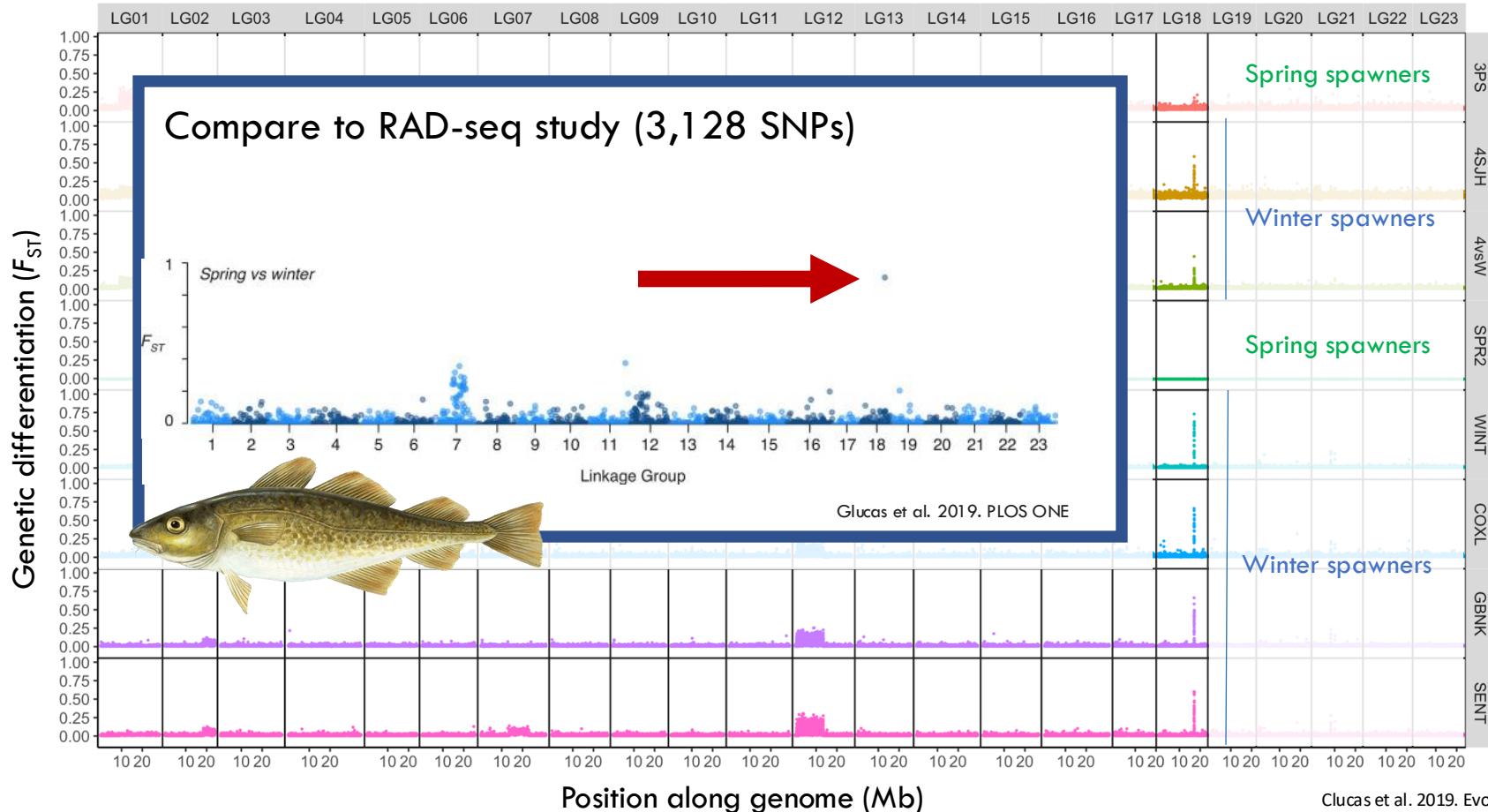
Pairwise comparisons to the wGoM spring spawners (5 kb windows)



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

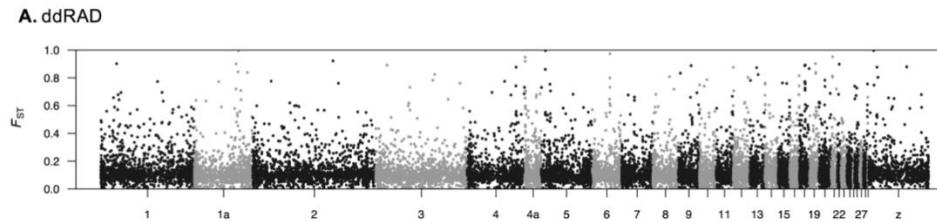
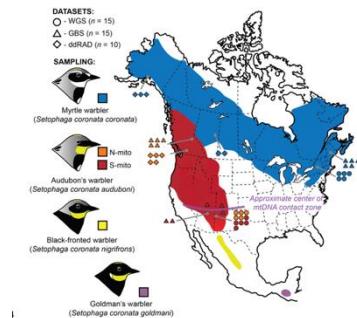


Pairwise comparisons to the wGoM spring spawners (5 kb windows)

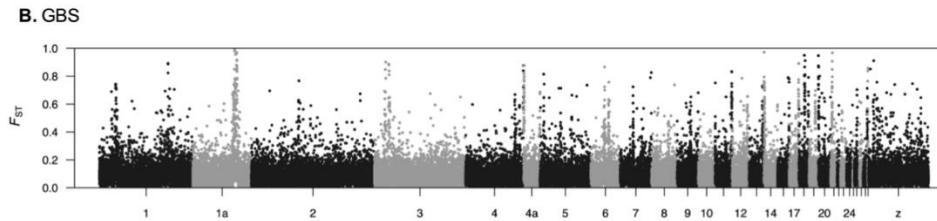


Direct comparison of ddRAD, GBS, and WGS data

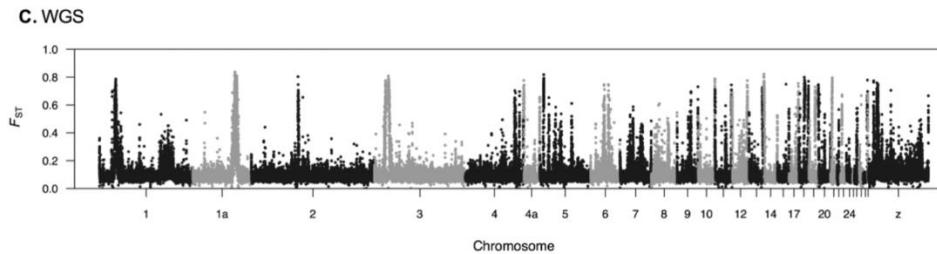
Comparing subspecies of yellow-rumped warbler



41 SNPs/10 kb



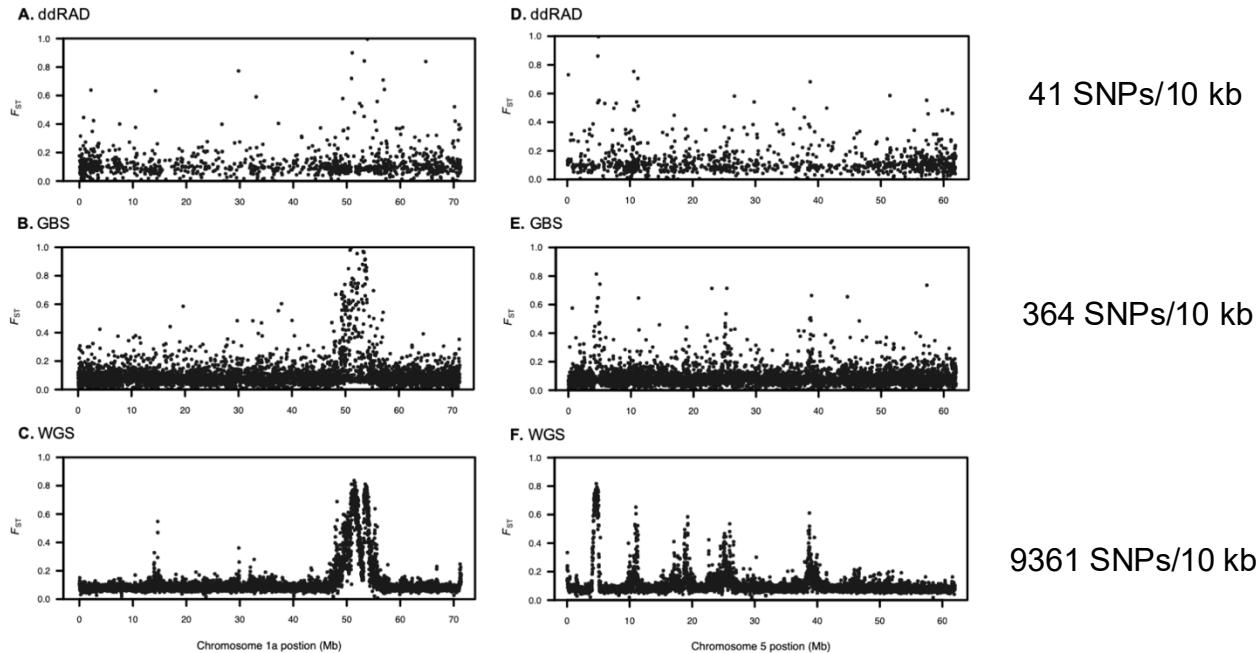
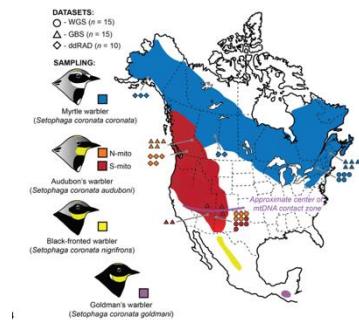
364 SNPs/10 kb



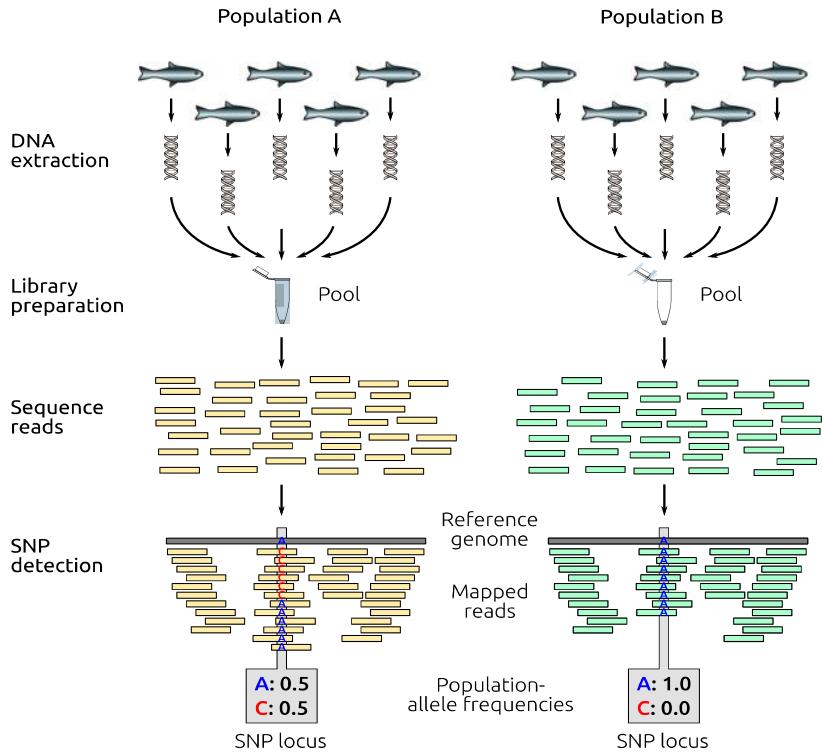
9361 SNPs/10 kb

Direct comparison of ddRAD, GBS, and WGS data

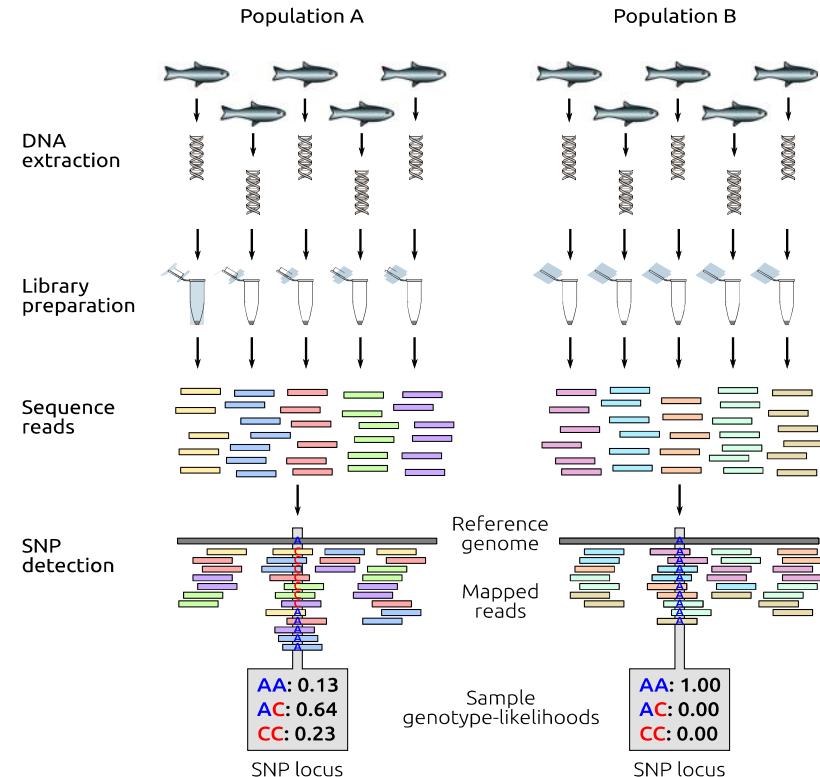
Comparing subspecies of yellow-rumped warbler



Pool-seq



Low-coverage WGS



Advantages of IcWGS 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 disequilibrium

Atlantic silverside *Menidia menidia*



Photo: Jacob Snyder

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

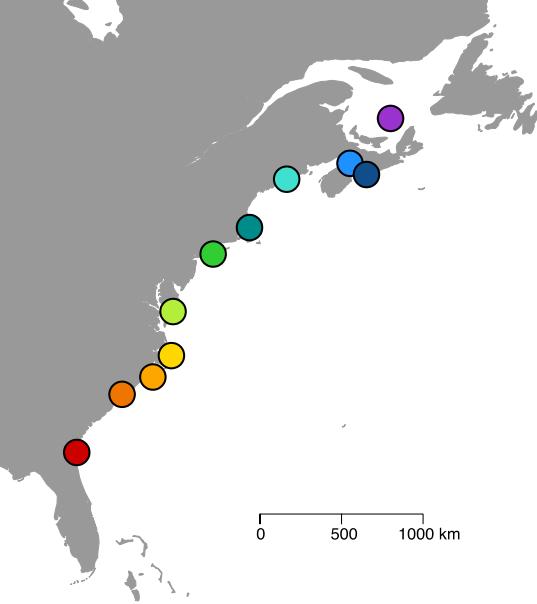


David Conover
U. of Oregon

Funding:



BIO-OCE 1434325
BIO-OCE 1756316

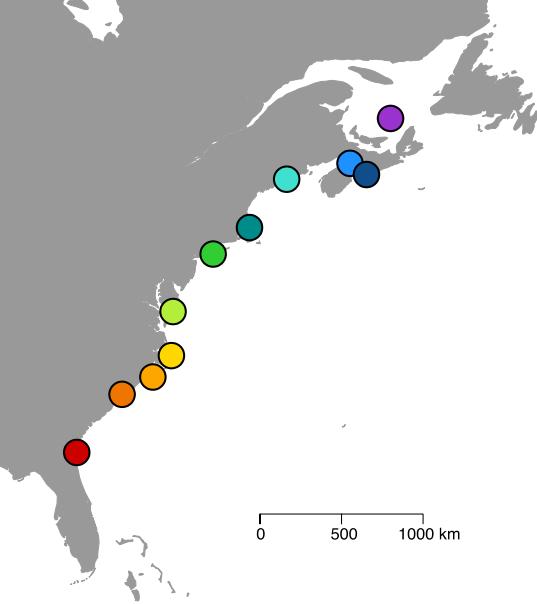


Individual-level analysis

Enables analysis of the genetic distance among individuals

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

Low-coverage ‘in silico’
exome capture

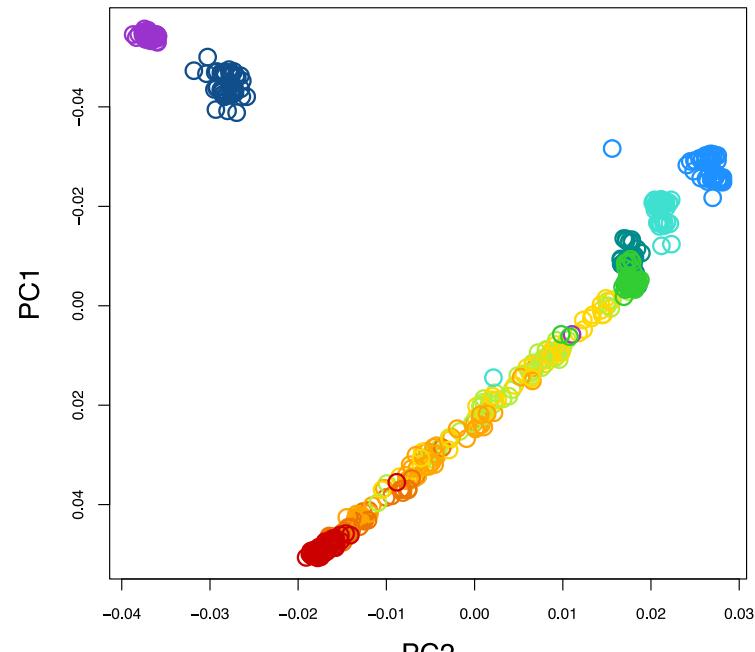


n=50 silversides per site
(sampled during spring
spawning)

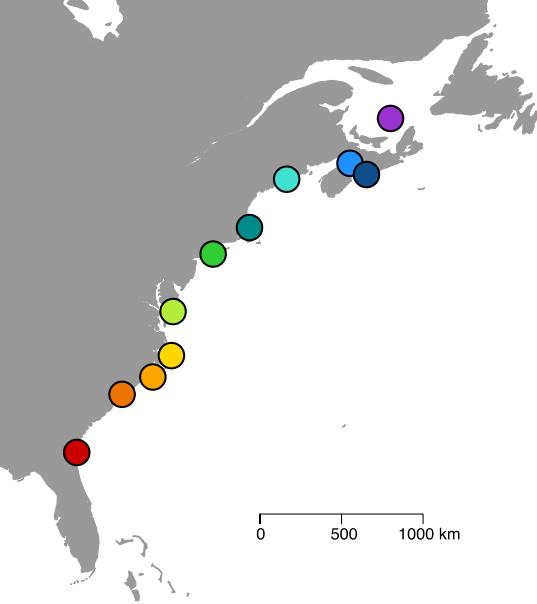
Low-coverage 'in silico'
exome capture (~1.3x)

Individual-level analysis

Enables analysis of the genetic distance among individuals



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

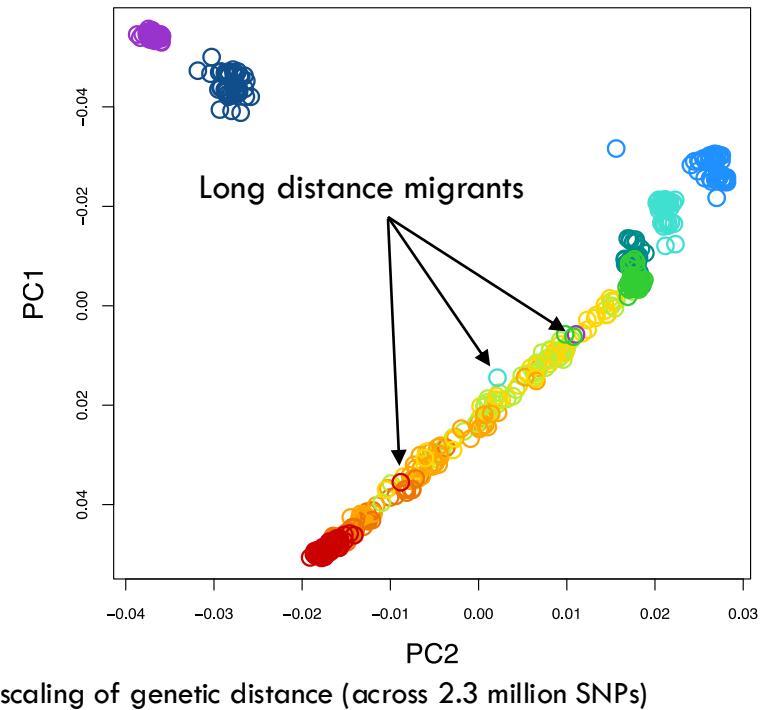


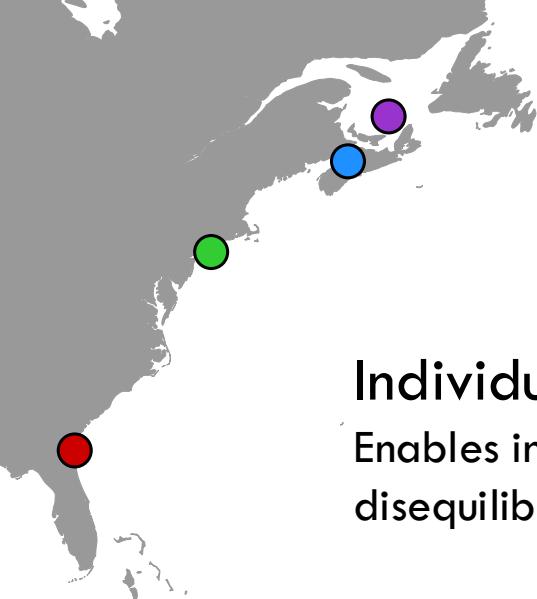
n=50 silversides per site
(sampled during spring
spawning)

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

Individual-level analysis

Enables identification of divergent individuals and mixed samples



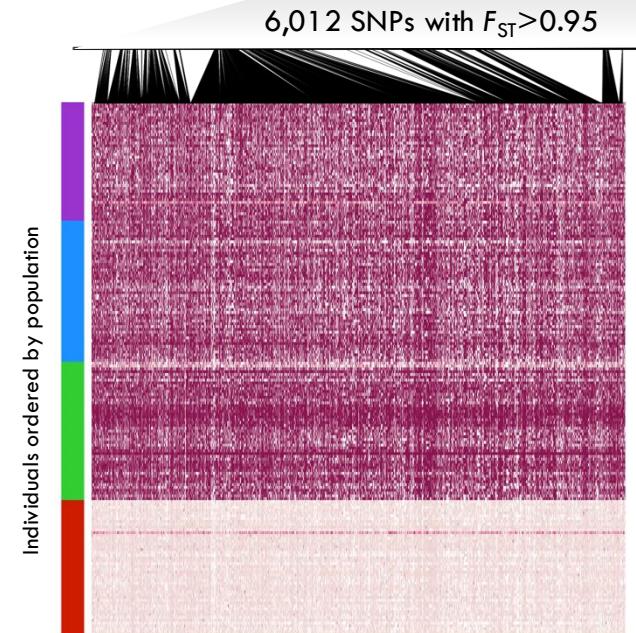
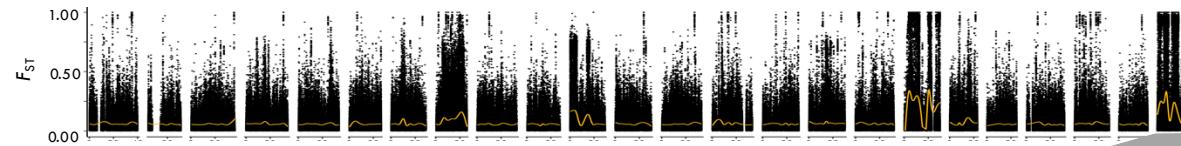


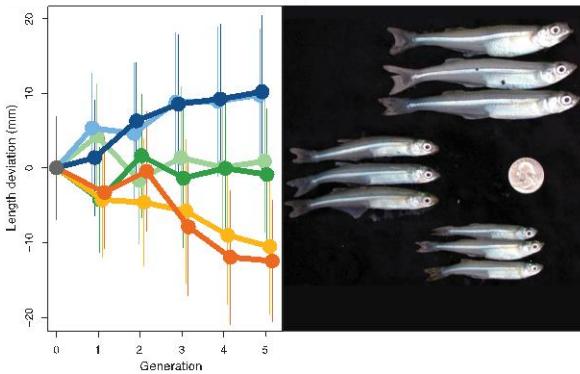
Individual analysis
Enables inference of linkage
disequilibrium

Most likely genotype for each

individual

- Northern homozygote
- Heterozygote
- Southern homozygote





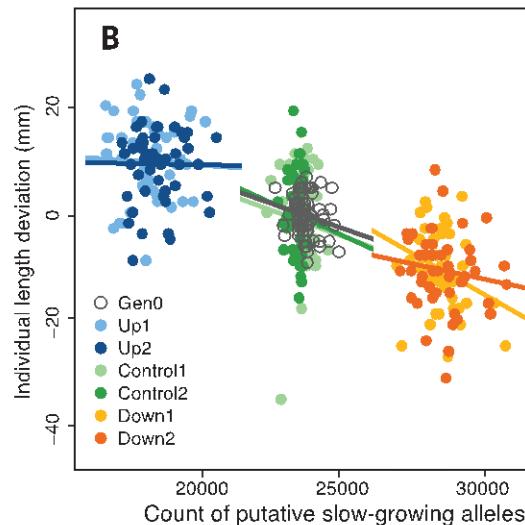
~50 silversides per population in generations 0 and 5

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

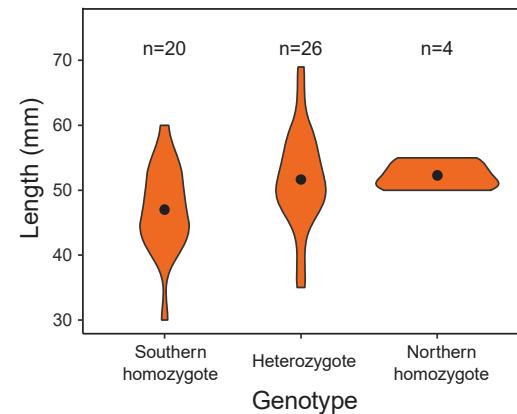
Individual-level analysis

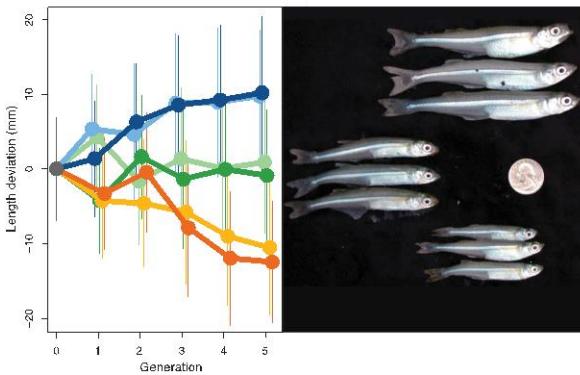
Enables individual genotype-phenotype association analysis

Polygenic score



Haplotyping inversions





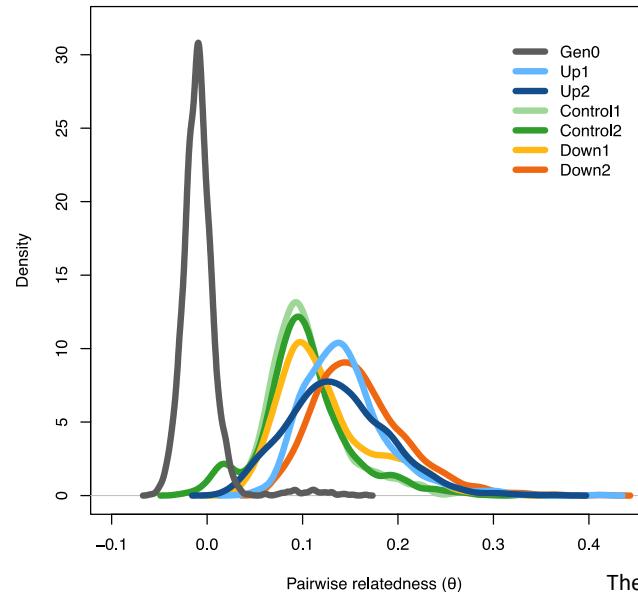
~50 silversides per population in generations 0 and 5

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

Individual-level analysis

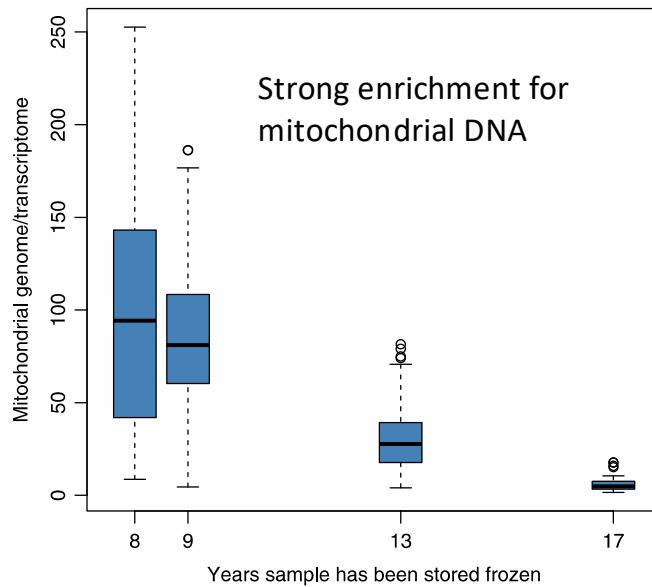
Enables computation of relatedness and individual diversity statistics

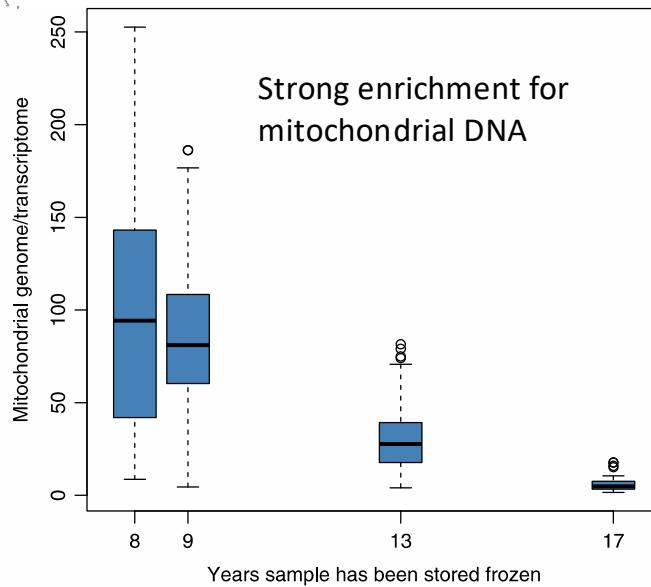
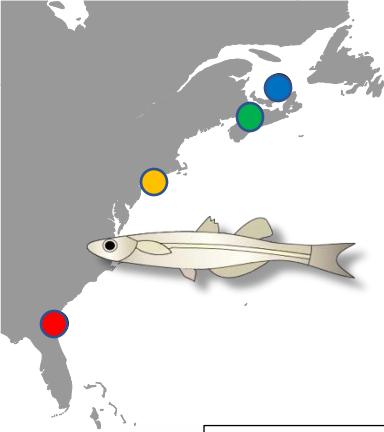
Pairwise relatedness



Individual-level analysis

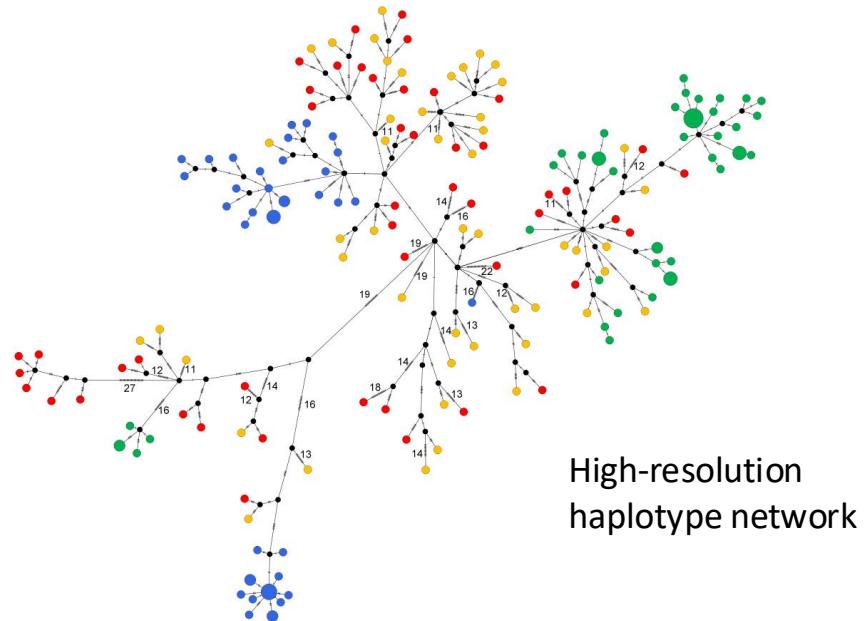
Enables recovery of full mitochondrial genome sequences



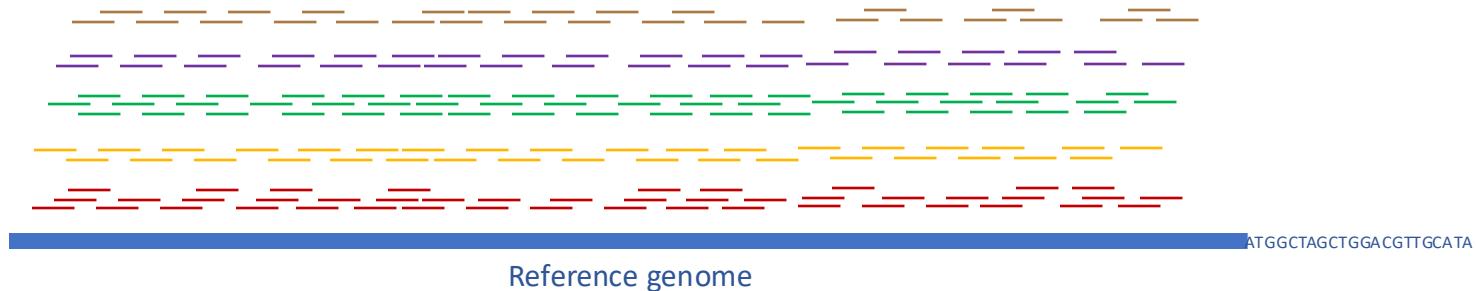


Individual-level analysis

Enables recovery of full mitochondrial genome sequences



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?

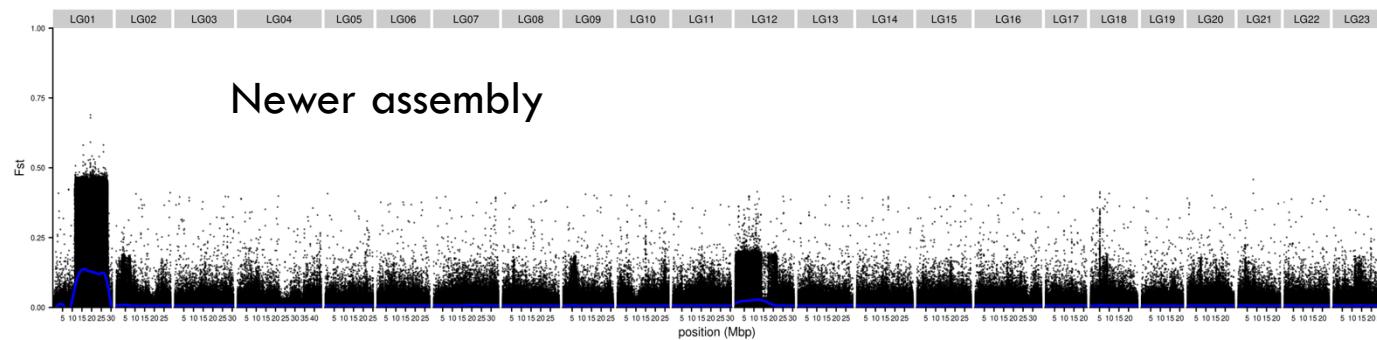
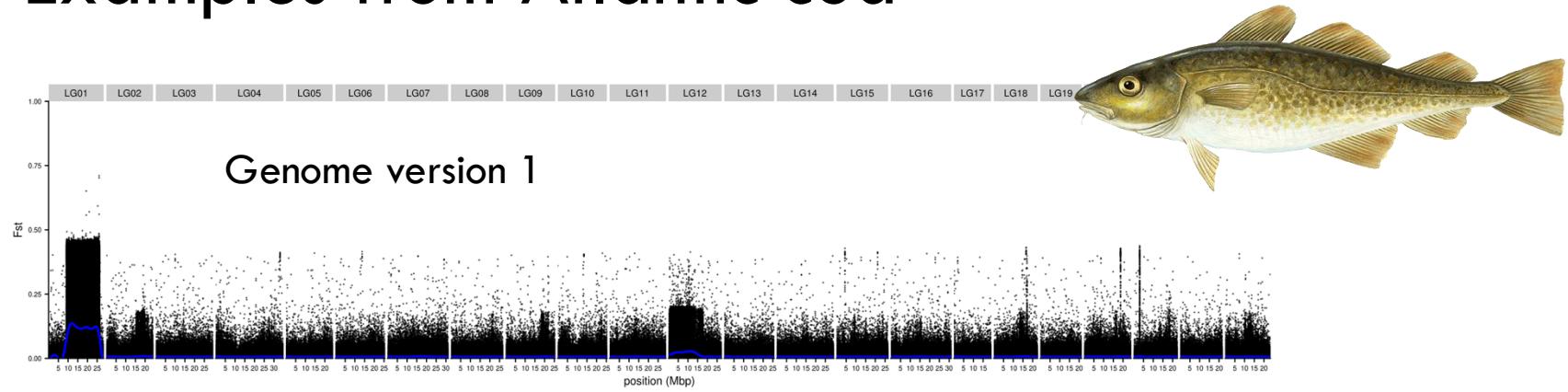
19 responses



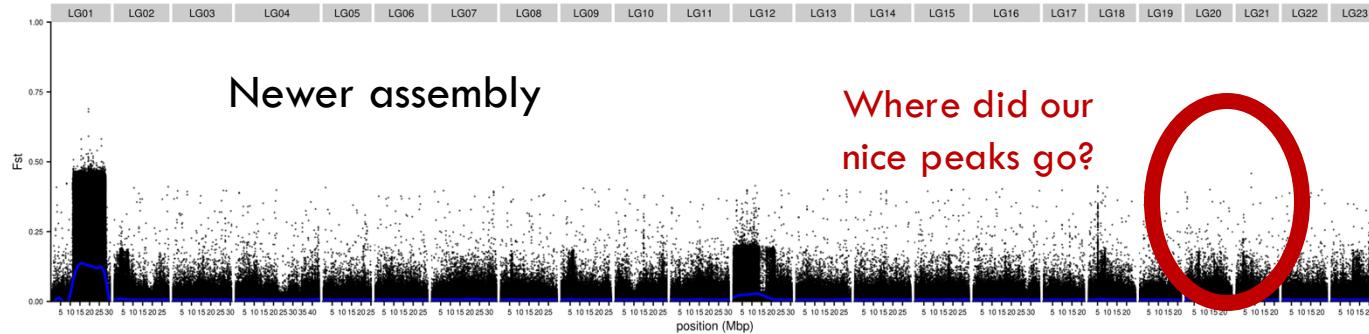
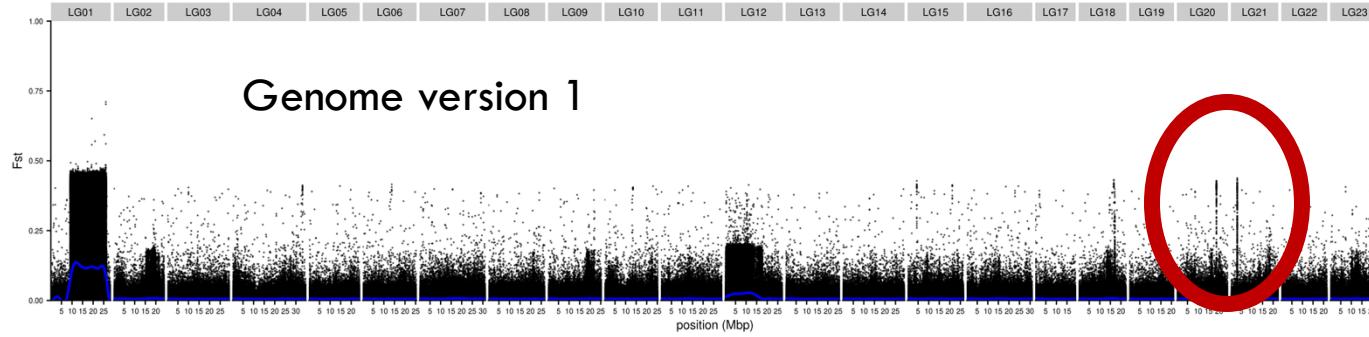
Discussion in breakout rooms

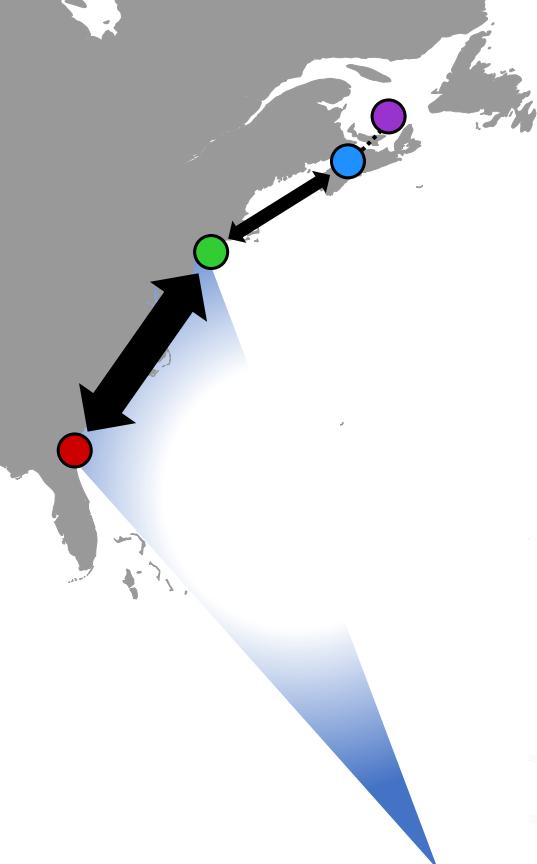
- How important is it to have a high-quality genome assembly for successful analysis of low-coverage whole genome re-sequencing data?

Examples from Atlantic cod



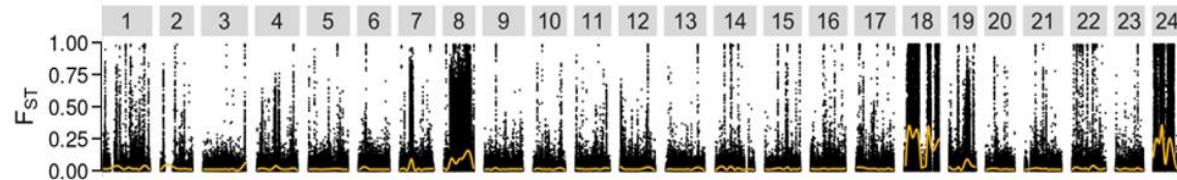
Examples from Atlantic cod



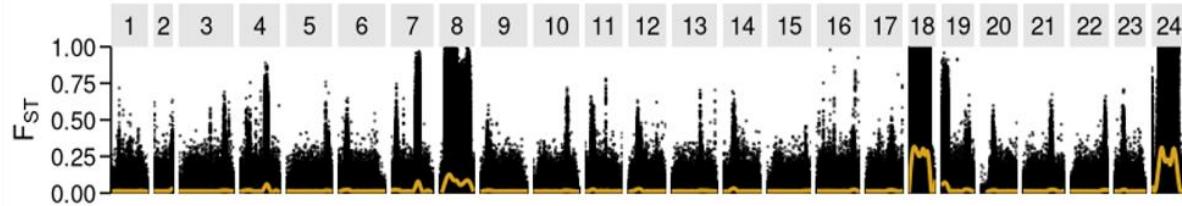


Effect of using species-specific reference genome

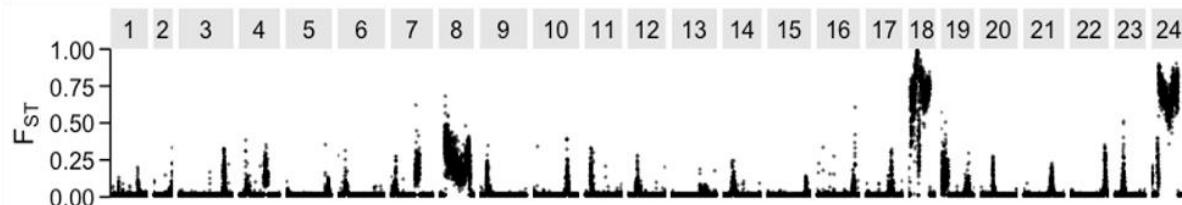
transcriptome SNPs mapped to medaka genome



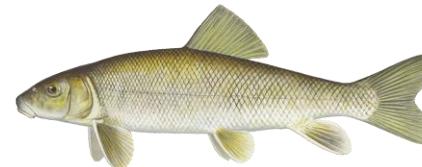
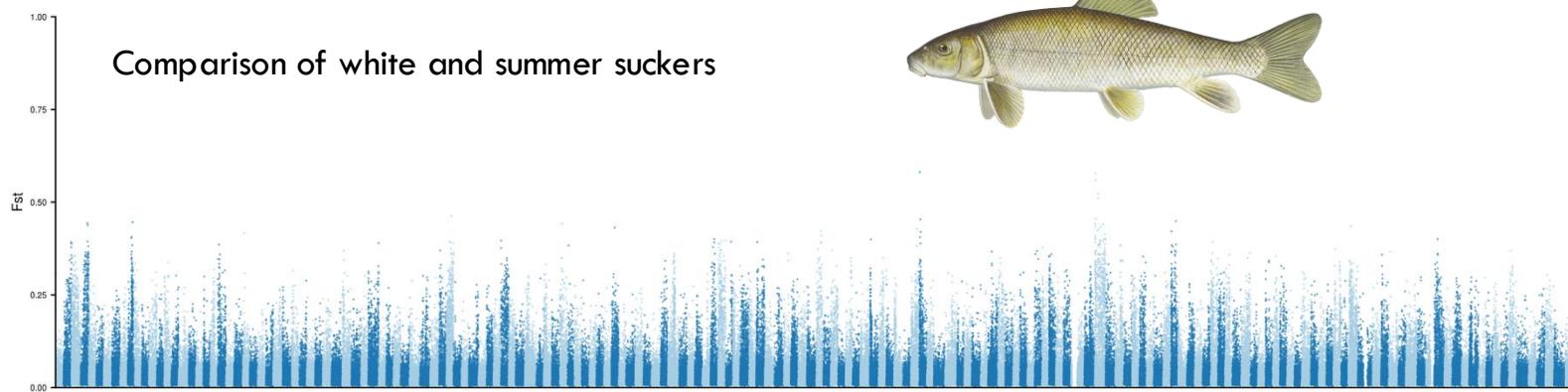
lcwgs SNPs mapped to silverside genome



lcwgs SNPs 10kb windows mapped to silverside genome

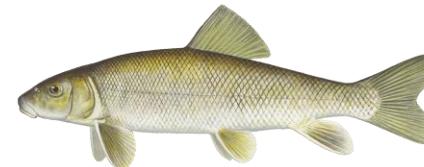
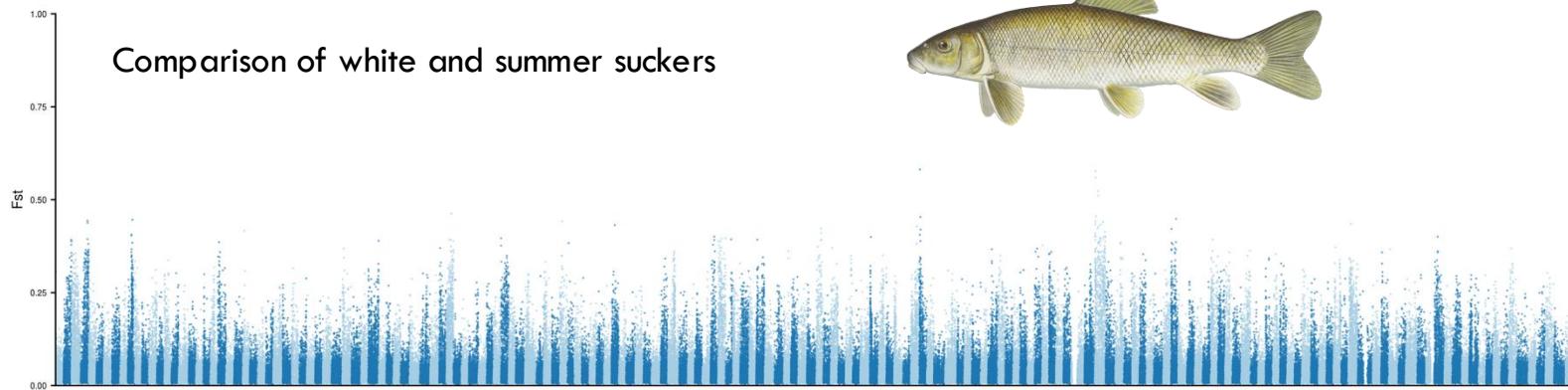


Highly fragmented genome assembly



Highly fragmented genome assembly

How do we even begin to make sense of this?



New technologies facilitate highly contiguous genome assemblies

- Long-read (PacBio and Nanopore) sequencing
- Hi-C
- Optical mapping

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

VGP ENSEMBL

Annotations

Set of primary assemblies

UCSC VGP GB

Annotations

VGP GENBANK



ABOUT EBP REPORTS COMMITTEES MEDIA CONTACT

CREATING A NEW FOUNDATION FOR BIOLOGY

Sequencing Life for the Future of Life

[VIEW EBP SEQUENCING PROGRESS](#)

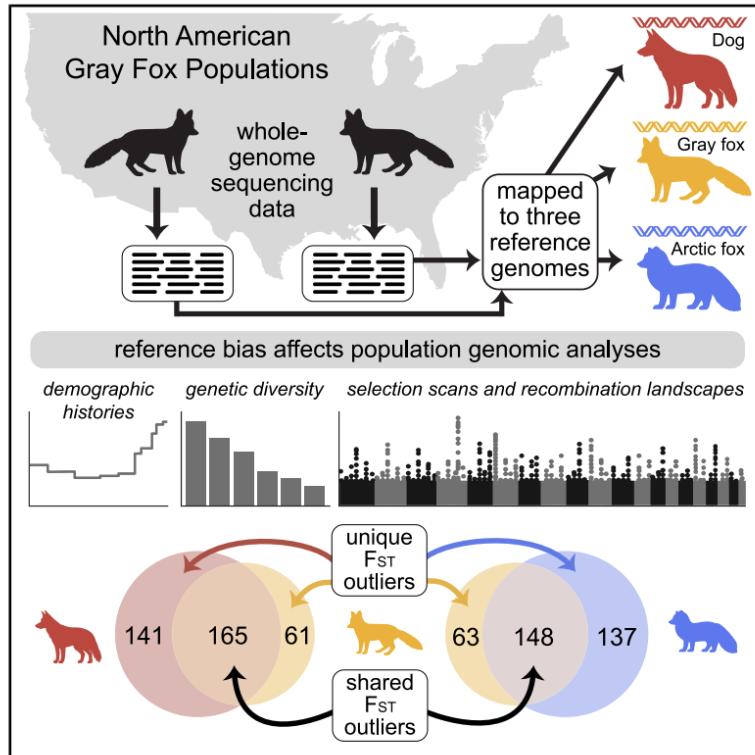
[HOW TO JOIN THE EBP](#)

Using reference genomes from other species

- Mapping bias can affect results

Reference genome choice compromises population genetic analyses

Graphical abstract



Authors

Maria Akopyan, Matthew Genchev,
Ellie E. Armstrong, Jazlyn A. Mooney

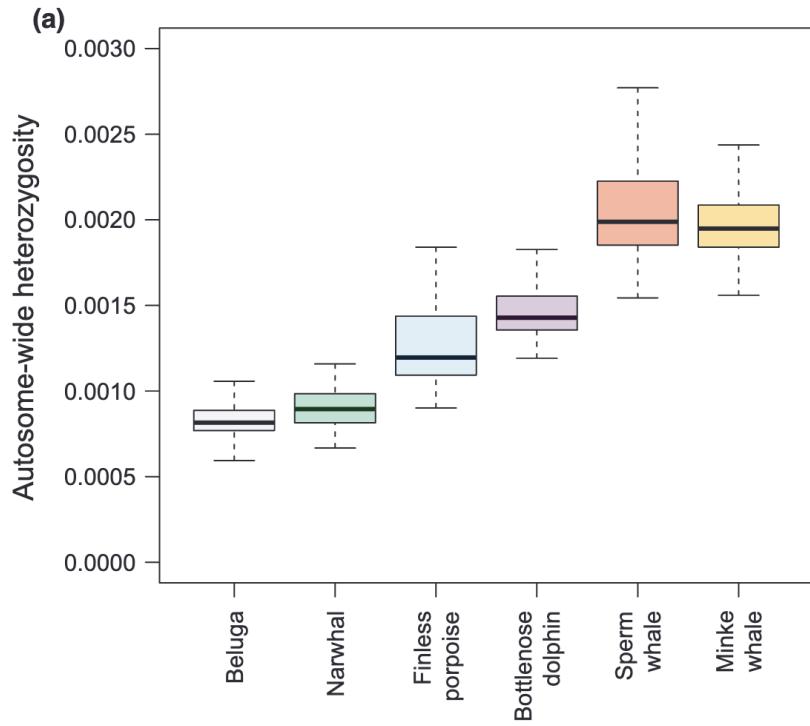
Correspondence

jazlynmo@usc.edu

In brief

Using a species-specific reference genome significantly improves the accuracy of population genetic analyses, as exemplified by mapping gray fox sequence data to genomes of several canid species. Mapping to a conspecific genome provides more reliable estimates of genetic variation, diversity, and evolutionary history.

Akopyan, M., Genchev, M., Armstrong, E. E., & Mooney, J. A. (2025). Reference genome choice compromises population genetic analyses. *Cell*. <https://doi.org/10.1016/j.cell.2025.08.034>



Prasad, A., Lorenzen, E. D., & Westbury, M. V. (2022). Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *Molecular Ecology Resources*, 22(1), 45–55.

Mapping bias

- Heterospecific references typically
 - Reduce mapped read rates
 - Under-call non-reference alleles
 - Result in lower π /heterozygosity estimates
 - Can shift FST and outlier scans—sometimes a lot

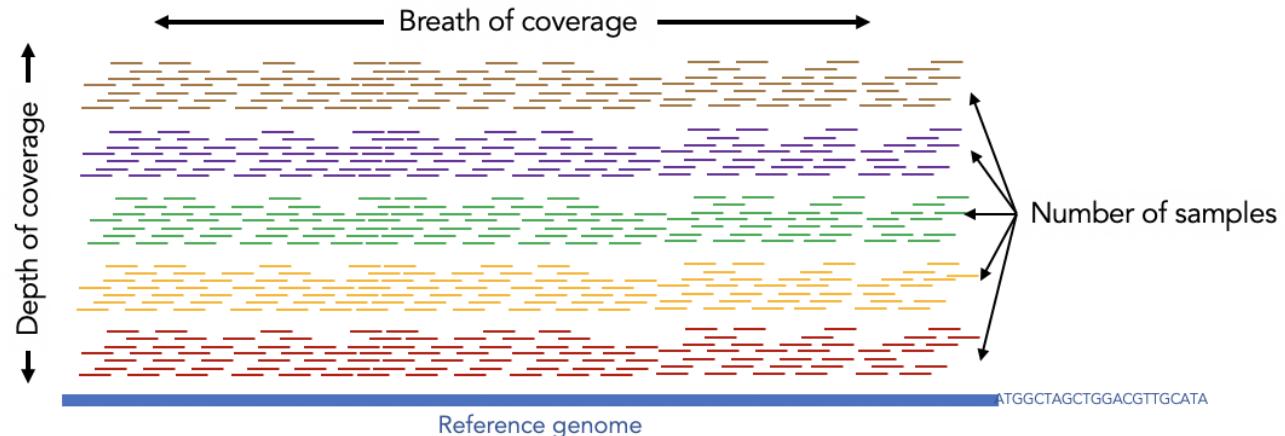
Validate key findings against a conspecific or “local” reference if one becomes available

Diagnostic and mitigation steps

- Run bias checks (REF/ALT balance at heterozygotes)
- Use major-allele or local pseudo-references to reduce REF-allele favoritism.
- Mask/omit low-complexity and poorly mappable regions before downstream stats
- Cross-map sanity checks
 - Map a subset to both the related-species reference and a draft/pseudo-reference built from your data
 - Compare π , FST, SFS, and outlier windows—large discrepancies are a red flag

Experimental design for IcWGS

- 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?



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 Lou¹  | Arne Jacobs¹  | Aryn P. Wilder²  |
Nina Overgaard Therkildsen¹ 

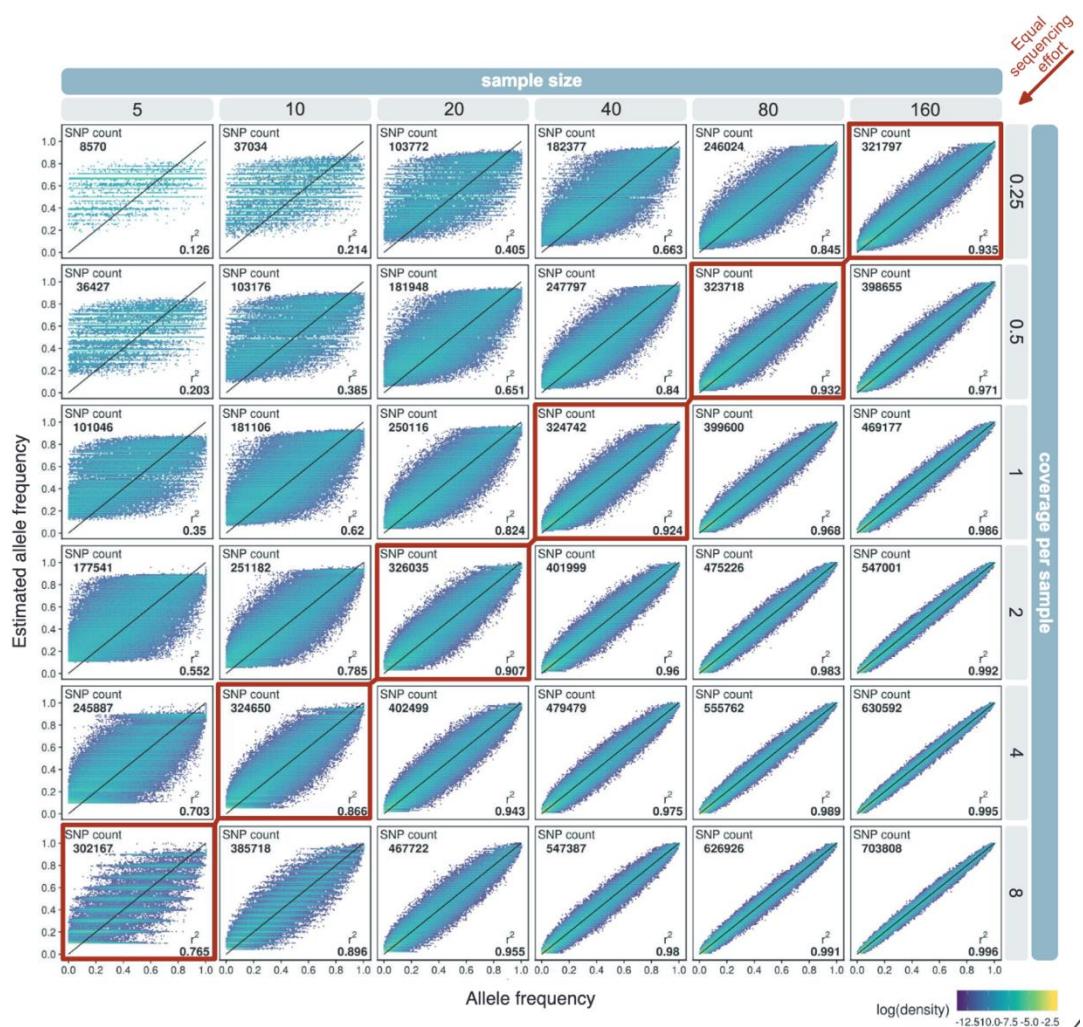
¹Department of Natural Resources and the Environment, Cornell University, Ithaca, New York, USA
²San Diego Zoo Wildlife Alliance, Escondido, California, USA

Correspondence
Runyang Nicolas Lou and Nina Overgaard Therkildsen, Department of Natural Resources and the Environment, Cornell University, Ithaca, New York, USA

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 uncertainty. A growing number of such tools have become available, but it can be difficult

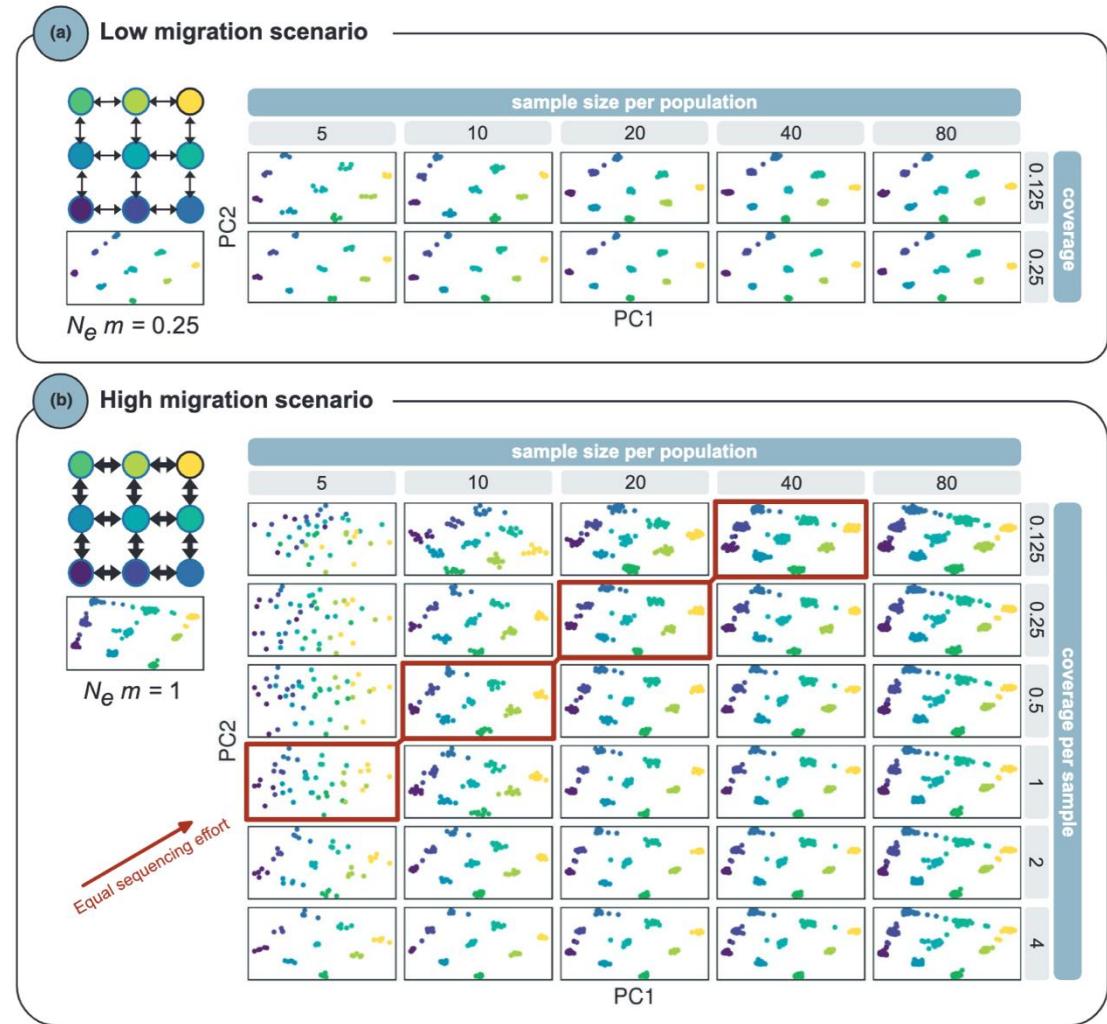
Allele frequency

Lower coverage of more individuals provides more accurate estimates



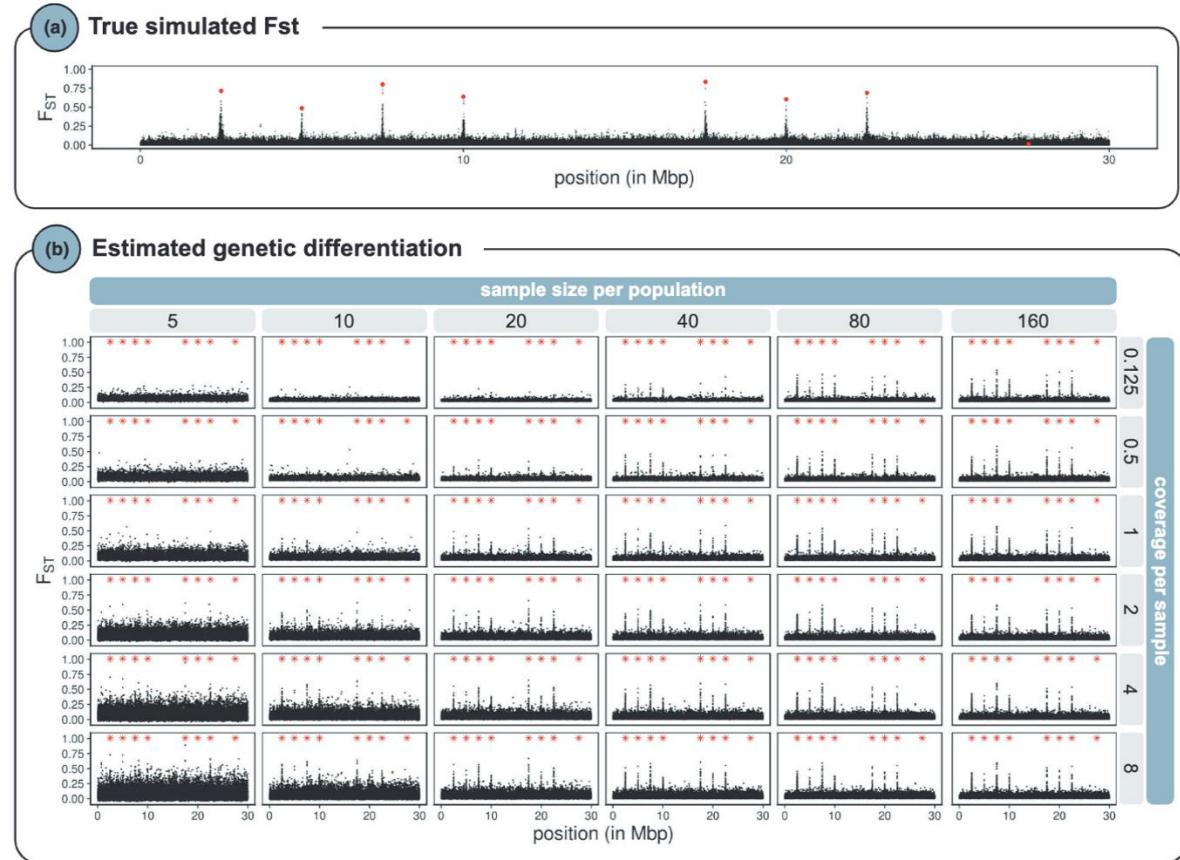
PCA

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



Outlier detection

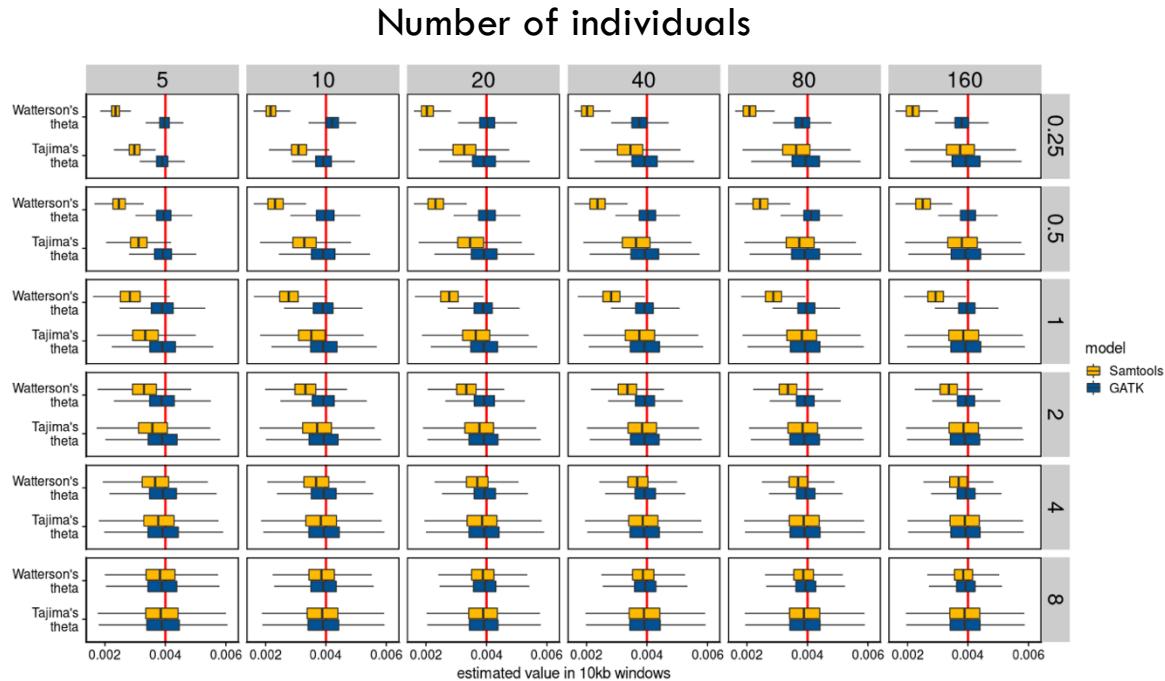
Peaks can be detected even at low coverage when the number of individuals is sufficient



Diversity statistics

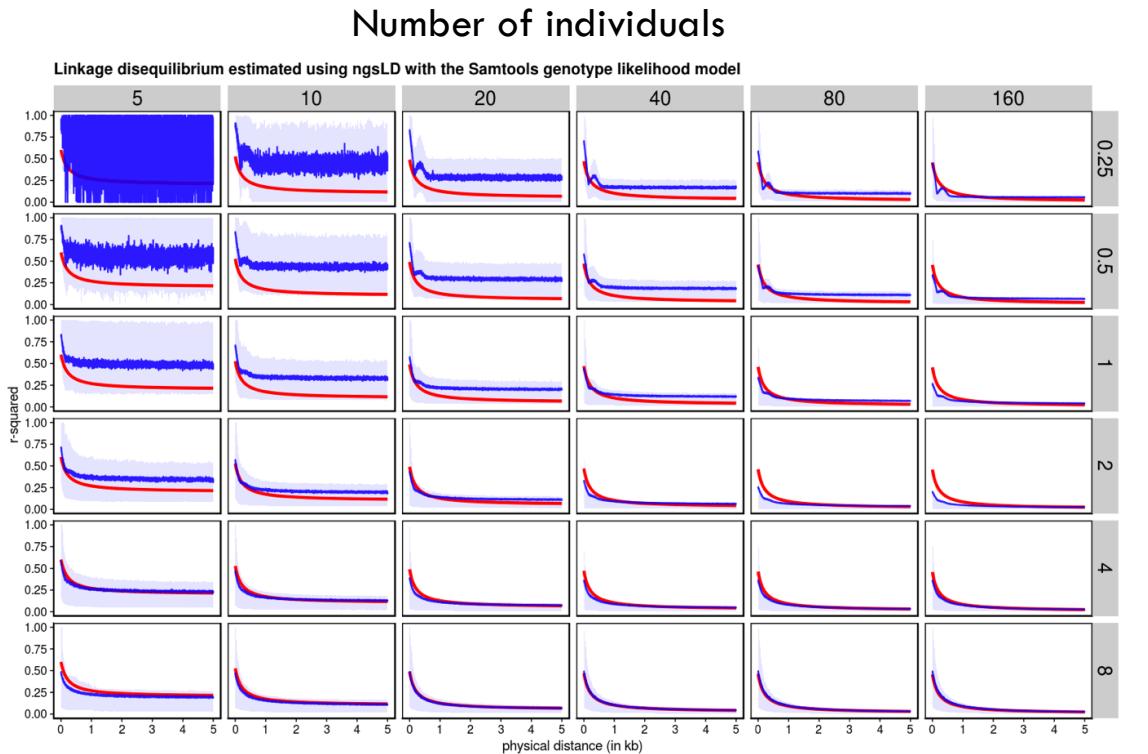
Estimates can be sensitive to the genotype likelihood (GL) model when depth is $< \sim 2\text{-}4x$

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 $\sim 2\text{-}4\times$, 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



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 Lou¹ | Arne Jacobs¹ | Aryn P. Wilder² |
Nina Overgaard Therkildsen¹

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

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

Correspondence

Runyang Nicolas Lou and Nina Overgaard Therkildsen, Department of Natural Resources and the Environment, Cornell University, Ithaca, NY, USA.
Emails: rl683@cornell.edu (R.N.L.); nt246@cornell.edu (N.O.T.)

Present address

Arne Jacobs, Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, UK

Funding information

National Science Foundation, Grant/Award Number: OCE-1756316

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 uncertainty. A growing number of such tools have become available, but it can be difficult to get an overview of what types of analyses can be performed reliably with lcWGS data, and how the distribution of sequencing effort between the number of samples analysed and per-sample sequencing depths affects inference accuracy. In this introductory guide to lcWGS, we first illustrate how the per-sample cost for lcWGS is now comparable to RAD-seq and Pool-seq in many systems. We then provide an overview of software packages that explicitly account for genotype uncertainty in different types of population genomic inference. Next, we use both simulated and empirical data to assess the accuracy of allele frequency, genetic diversity, and linkage disequilibrium estimation, detection of population structure, and selection scans under different sequencing strategies. Our results show that spreading a given amount of

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, $\geq 10x$ 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, $>4x$ 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, $\geq 10x$ 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., $0.125x$, Figure 4; Figures S6 and S11) or highly uneven per-sample coverage (e.g., ranging from $0.5x$ to $6x$, Skotte et al., 2013) can be viable.
Absolute estimation of linkage disequilibrium	LD decay rate, demographic inference	Prioritize higher coverage per sample, $\geq 4x$ 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 $1x$ could be viable, $\geq 20x$ coverage per population (Figures S4 and S5).
Genotype imputation without reference panels	STITCH, BEAGLE	STITCH: prioritize larger sample size (≥ 500) over per-sample coverage ($1x$ could be sufficient) BEAGLE: prioritize higher per-sample coverage ($\geq 2x$) over sample size (≤ 250 could be sufficient) (Figure B4).