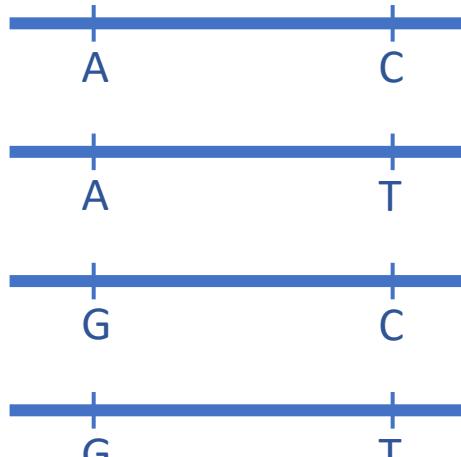


LD and LD pruning

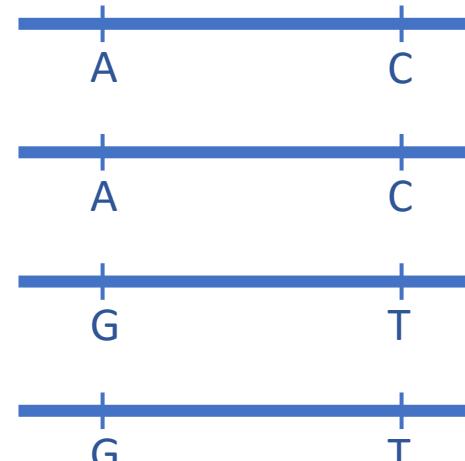
Physalia course
Nina Overgaard Therkildsen

Linkage disequilibrium (LD)

Nonrandom association of alleles at different loci



Equilibrium



Disequilibrium

What affects LD?

- Genome-wide
 - Population history
 - Breeding system
 - Pattern of geographic subdivision
- Within each genomic region
 - The history of natural selection, gene conversion, mutation and other forces that cause gene-frequency evolution
- Effect depends on local recombination rates

Why estimate LD?

- Discuss in breakout rooms

Why estimate LD?

- Identify genomic regions with elevated LD
- Examine the degree of LD between particular variants/regions of interest
- Understand LD decay patterns (and whether they are different between populations)
- Identify genomic blocks that may bias inference of population structure

Good resource for learning more

Nature Reviews Genetics, 9(6), 477–485 (2008)

REVIEWS



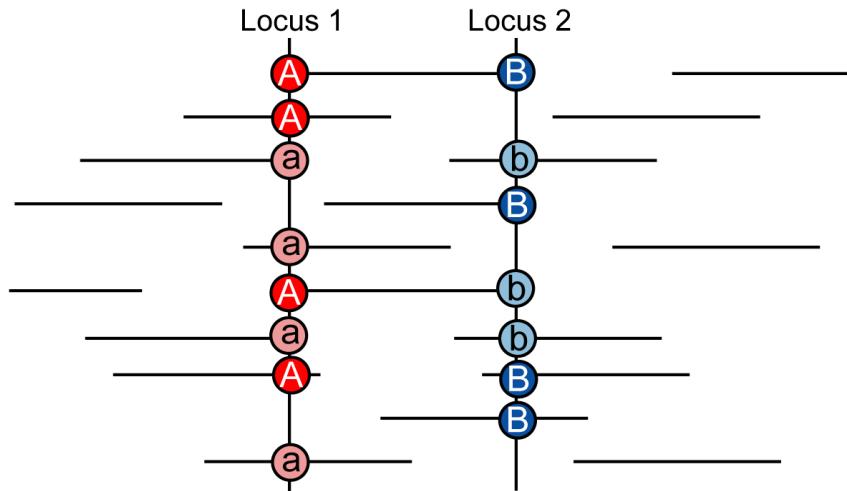
FUNDAMENTAL CONCEPTS IN GENETICS

Linkage disequilibrium — understanding the evolutionary past and mapping the medical future

Montgomery Slatkin

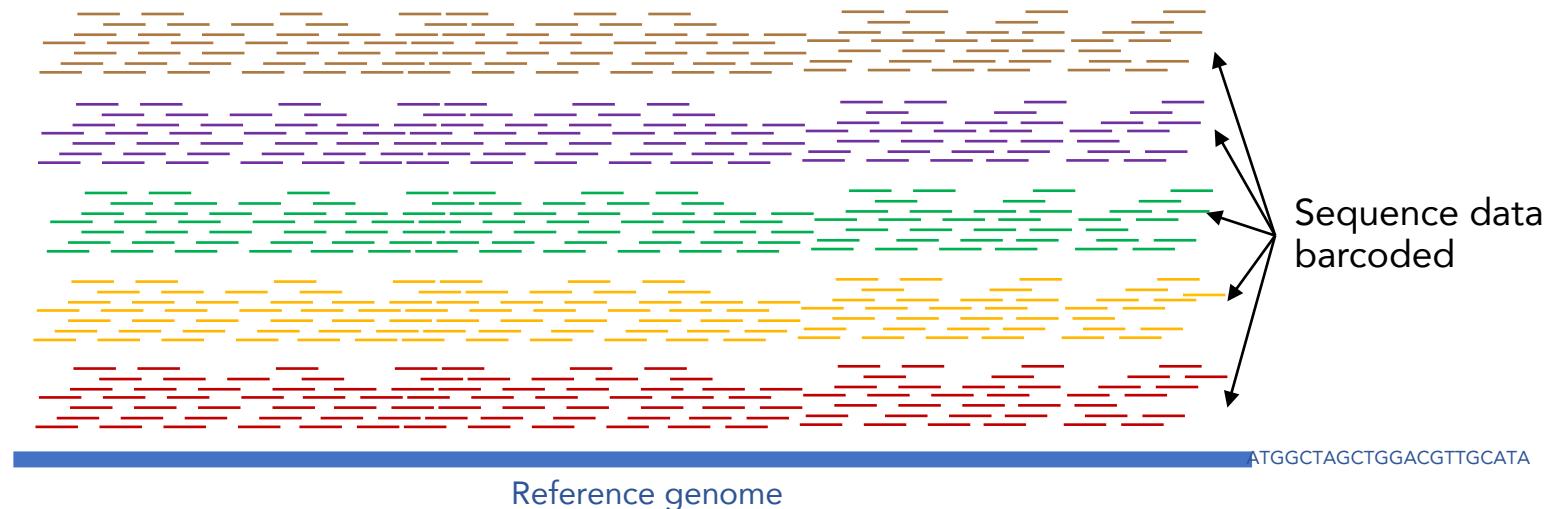
Abstract | Linkage disequilibrium — the nonrandom association of alleles at different loci — is a sensitive indicator of the population genetic forces that structure a genome. Because of the explosive growth of methods for assessing genetic variation at a fine scale, evolutionary biologists and human geneticists are increasingly exploiting linkage disequilibrium in order to understand past evolutionary and demographic events, to map genes that are associated with quantitative characters and inherited diseases, and to understand the joint evolution of linked sets of genes. This article introduces linkage disequilibrium, reviews the population genetic processes that affect it and describes some of its uses. At present, linkage disequilibrium is used much more extensively in the

With Poolseq data, we can only estimate LD across the length of reads



Can only estimate LD between SNPs spanned by single read pairs (i.e. <1000bp)

With IcWGS data, we preserve information about which individual carried each observed allele across the entire genome

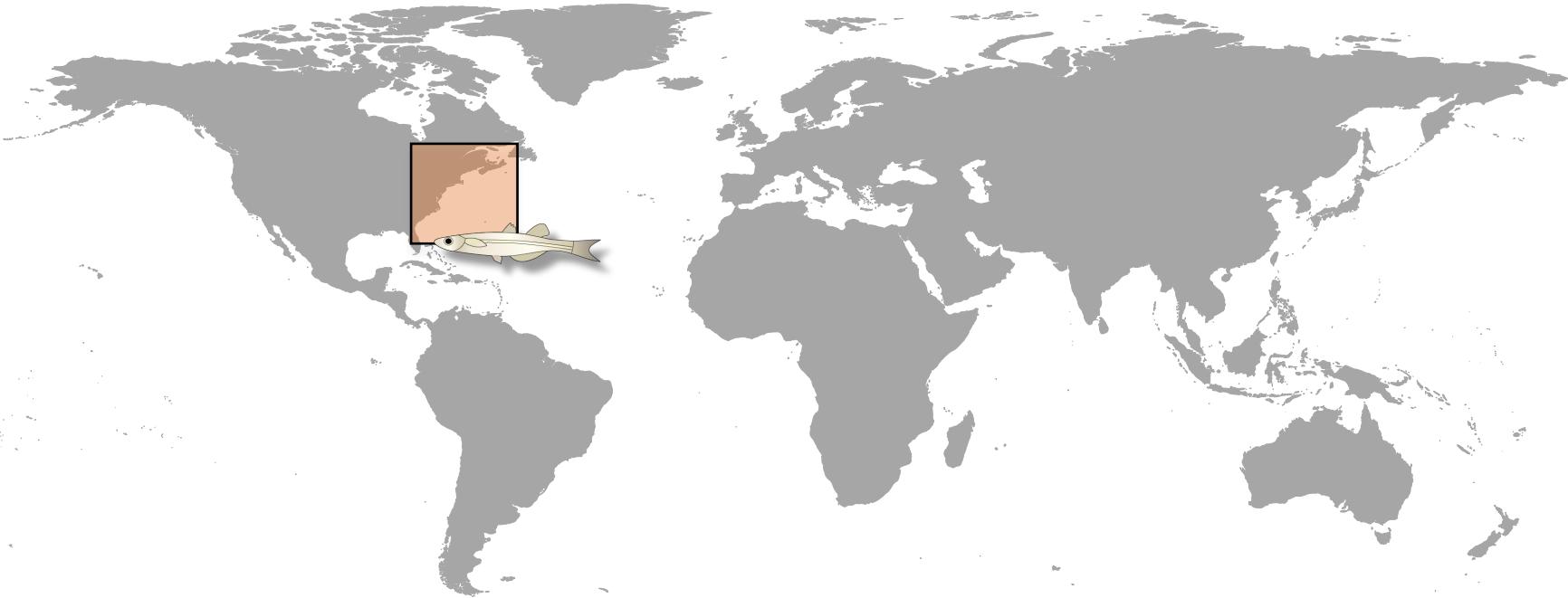


Some examples

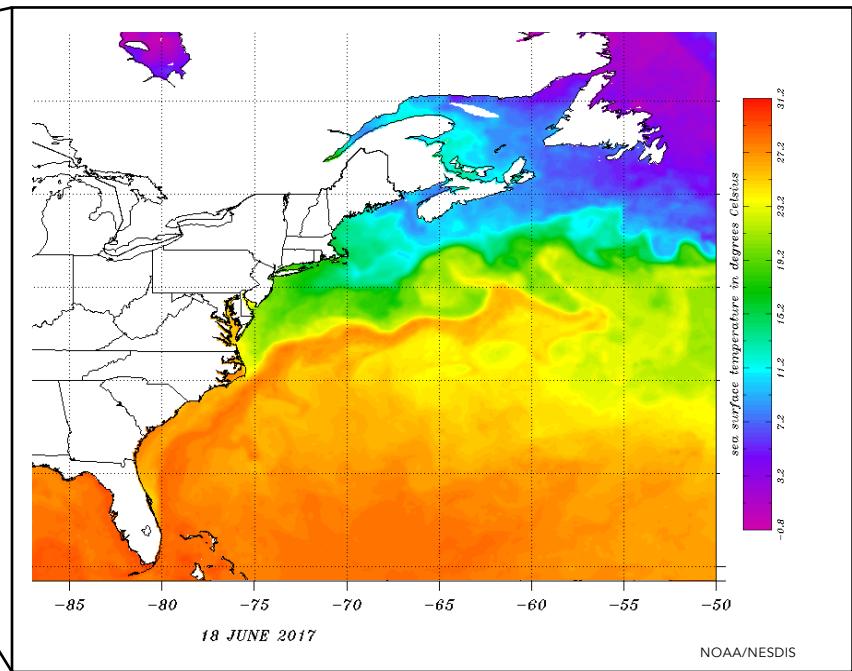
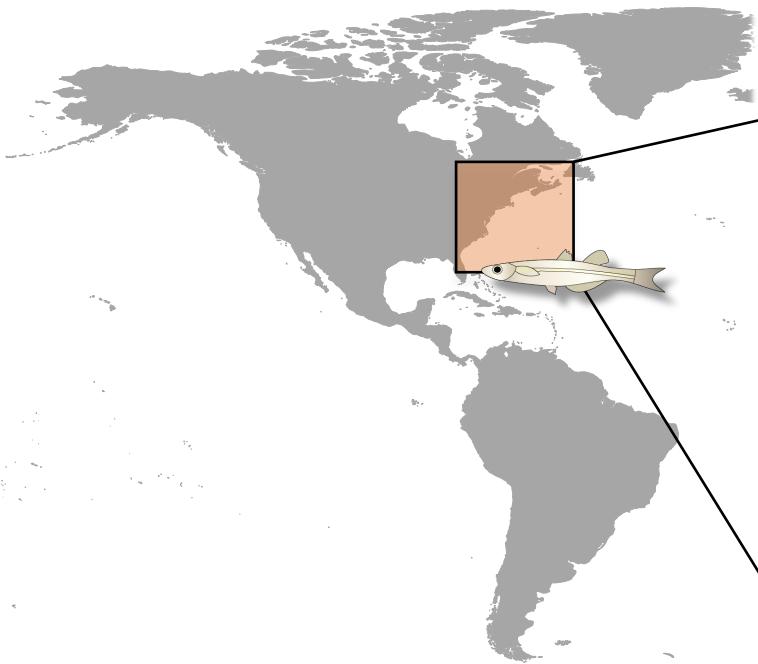
Atlantic silverside *Menidia menidia*



Photo: Jacob Snyder

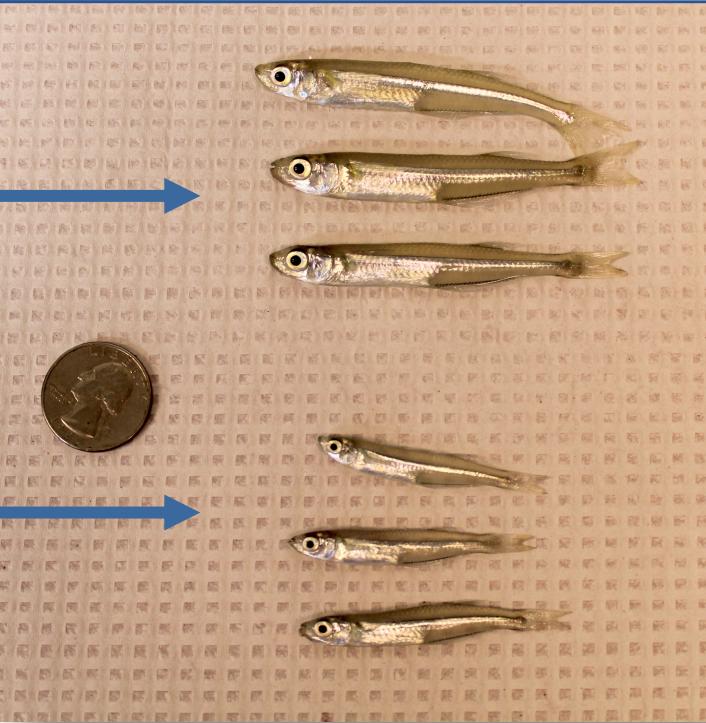


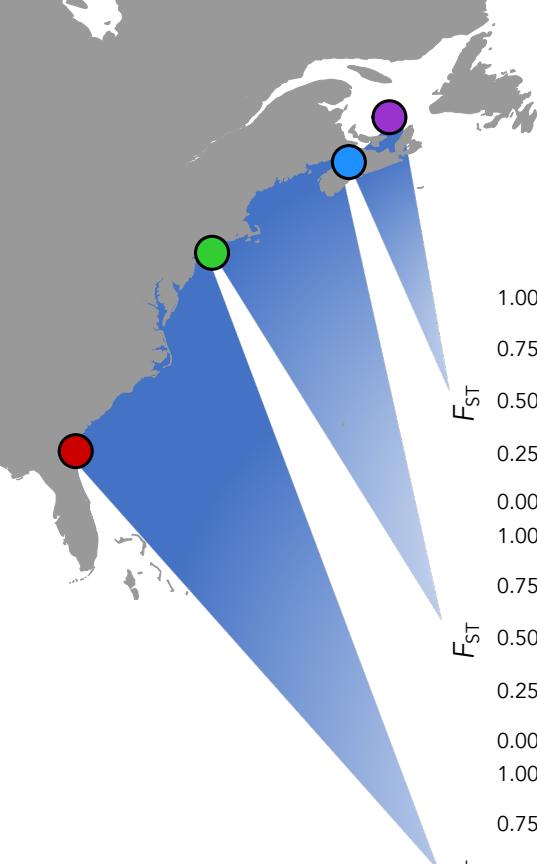
One of the world's steepest
thermal gradients



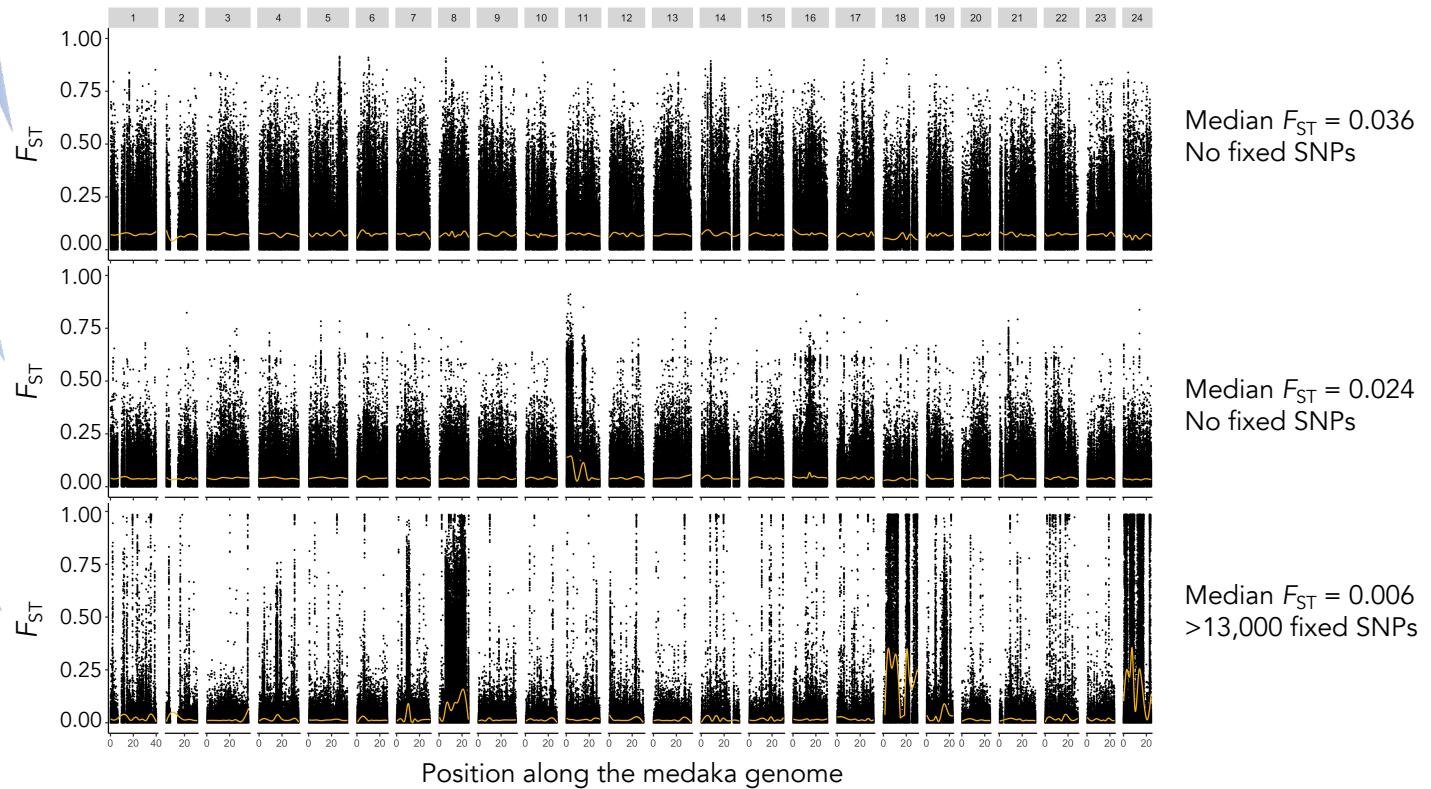


Same age,
Common lab environment

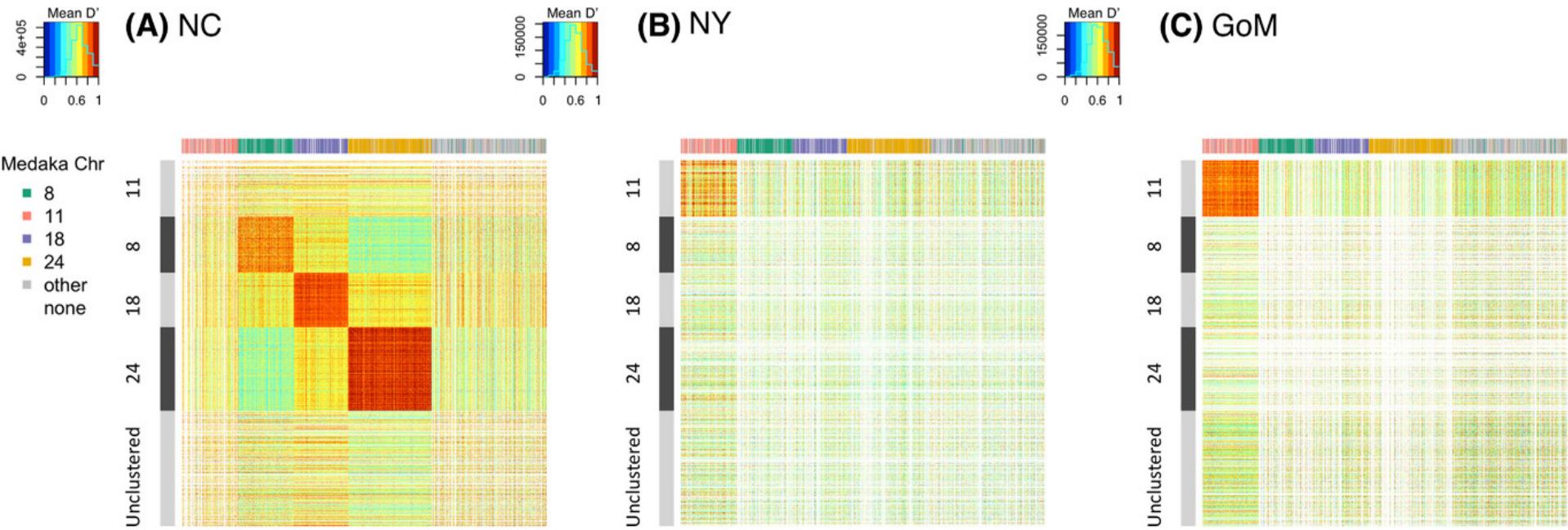


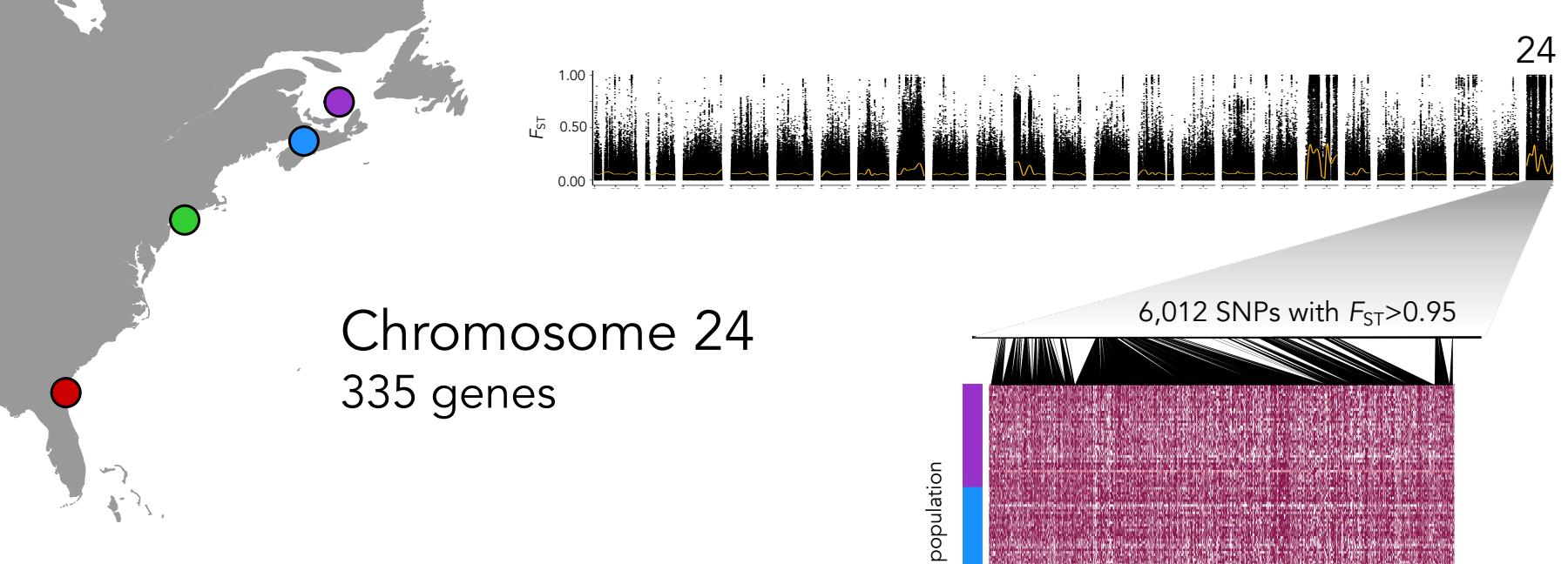


Heterogeneous gene flow and divergence patterns across latitudes



LD patterns among outliers



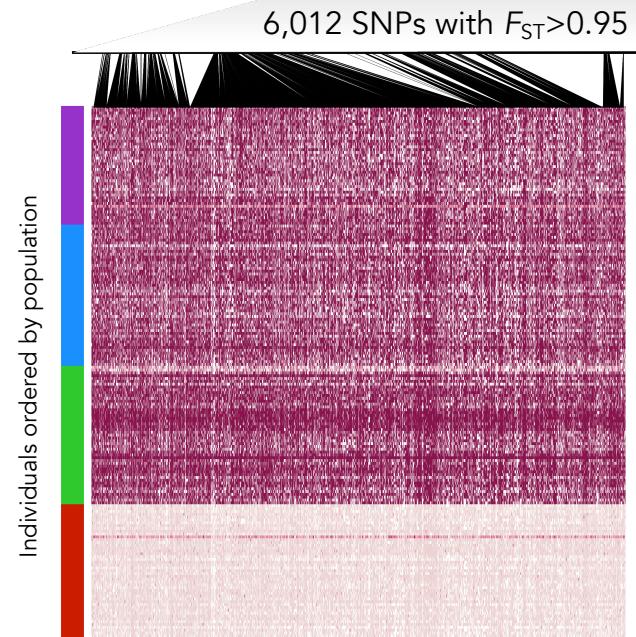


Chromosome 24

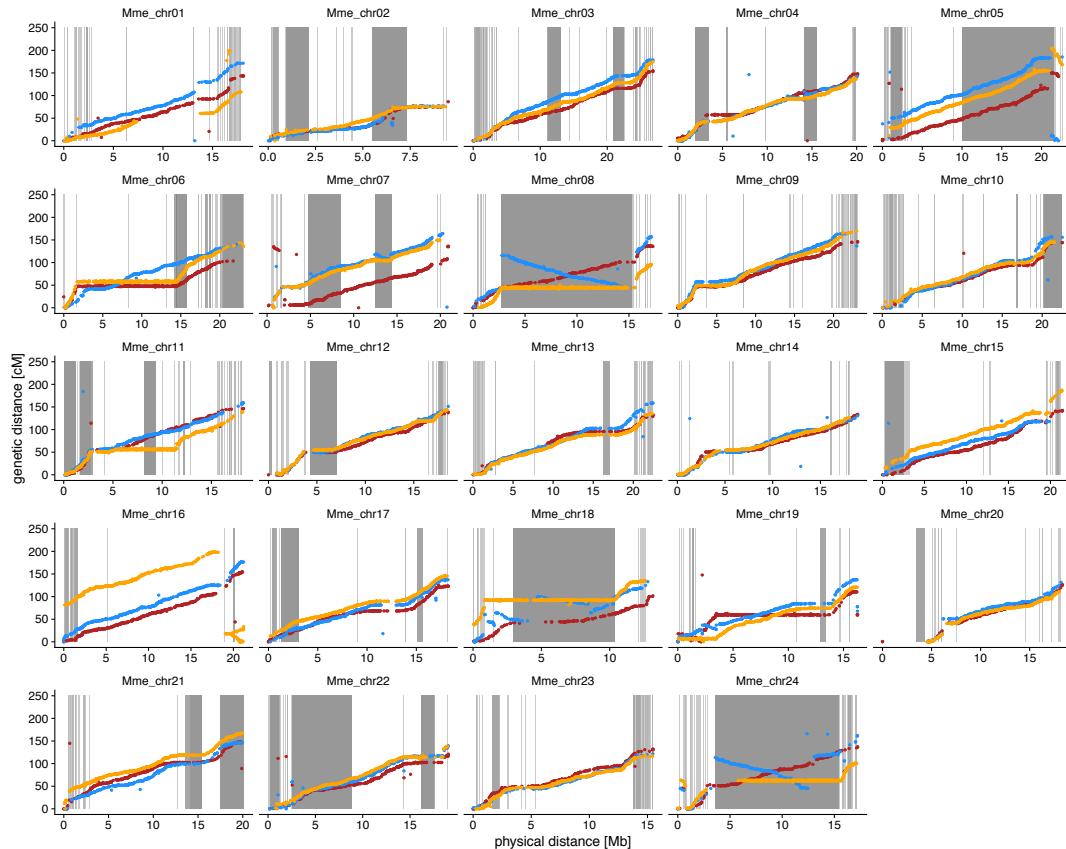
335 genes

Most likely genotype for each individual

- Northern homozygote
- Heterozygote
- Southern homozygote



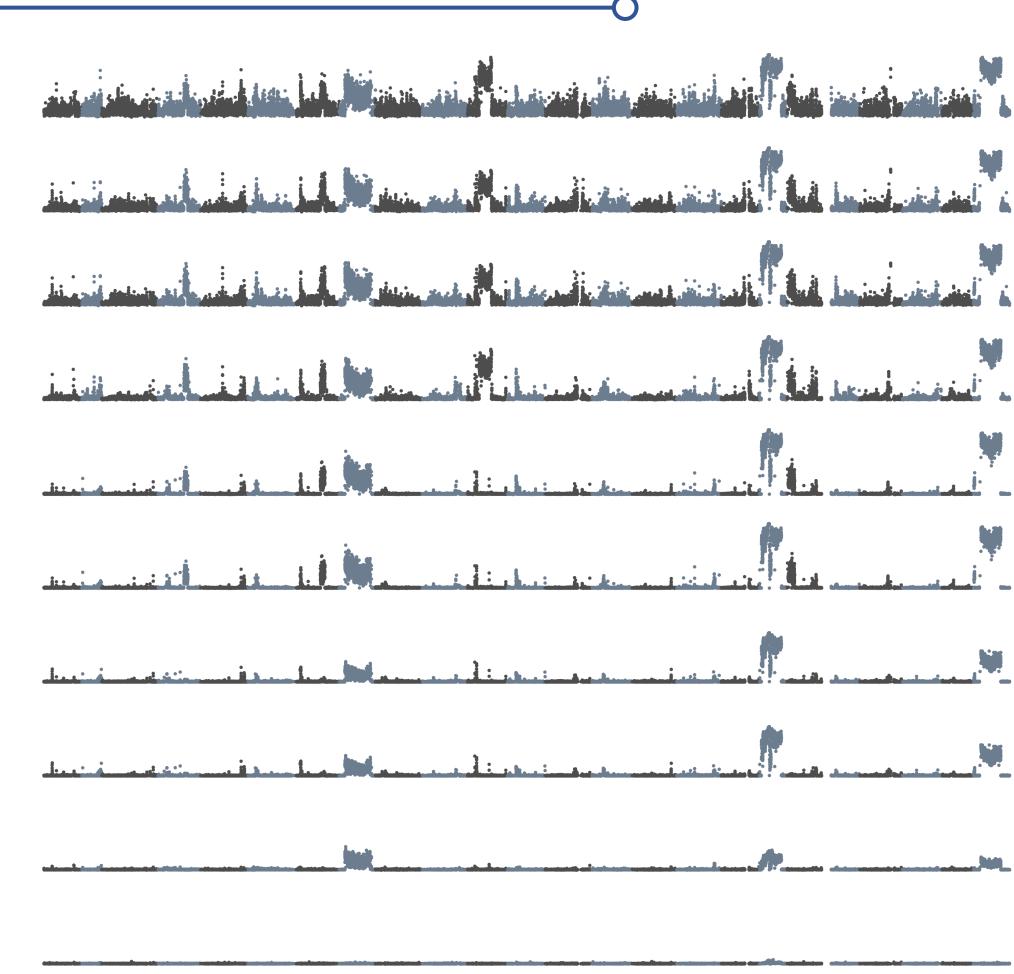
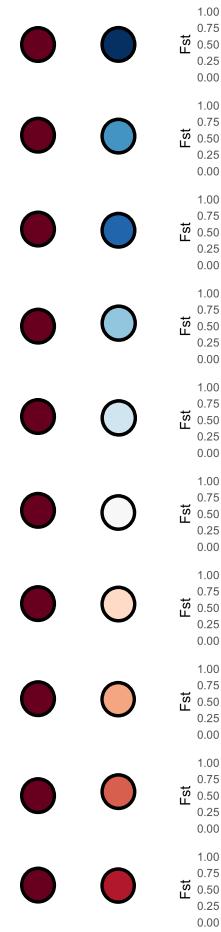
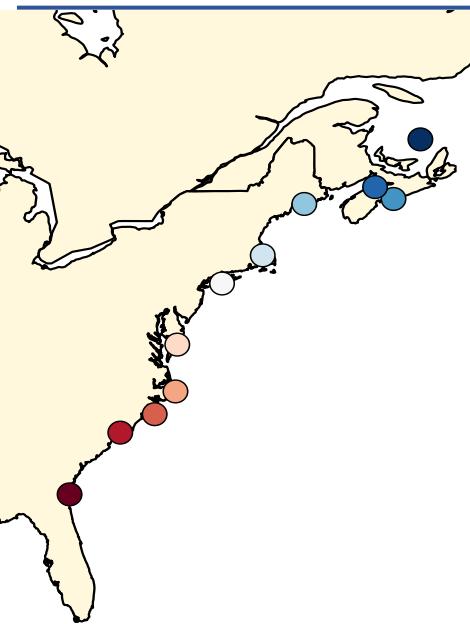
Linkage mapping confirms presence of massive inversions



Maria Akopyan and Arne Jacobs
(unpublished results)

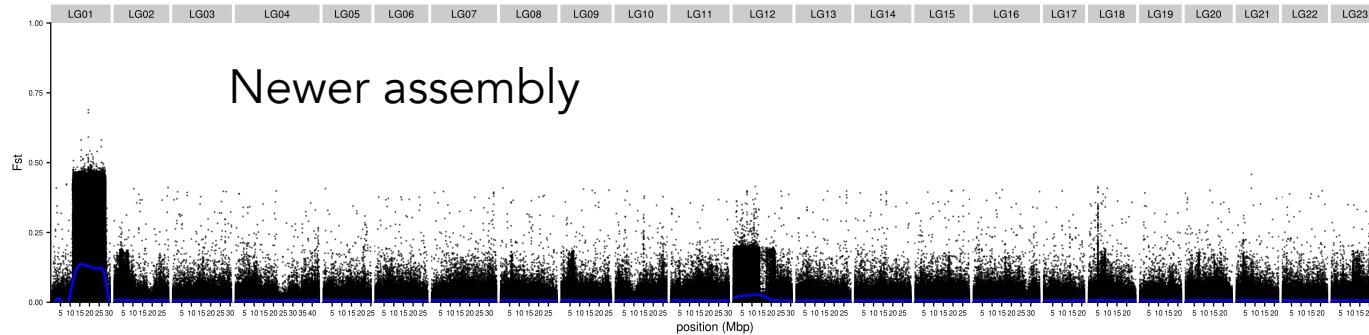
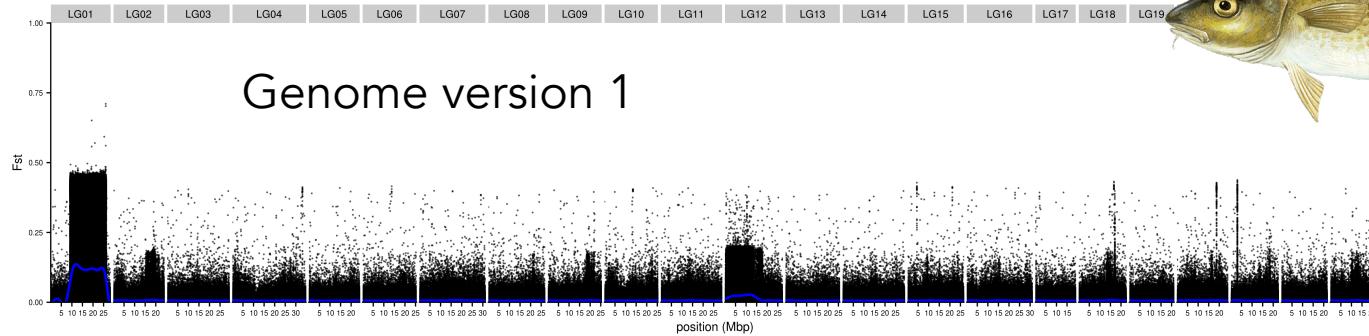
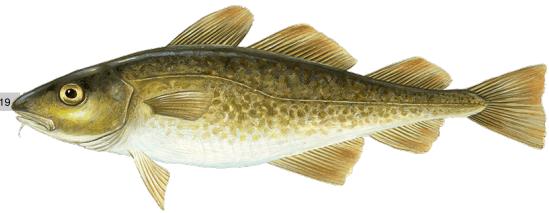
Genetic differentiation increases with latitude

Arne Jacobs (unpublished results)

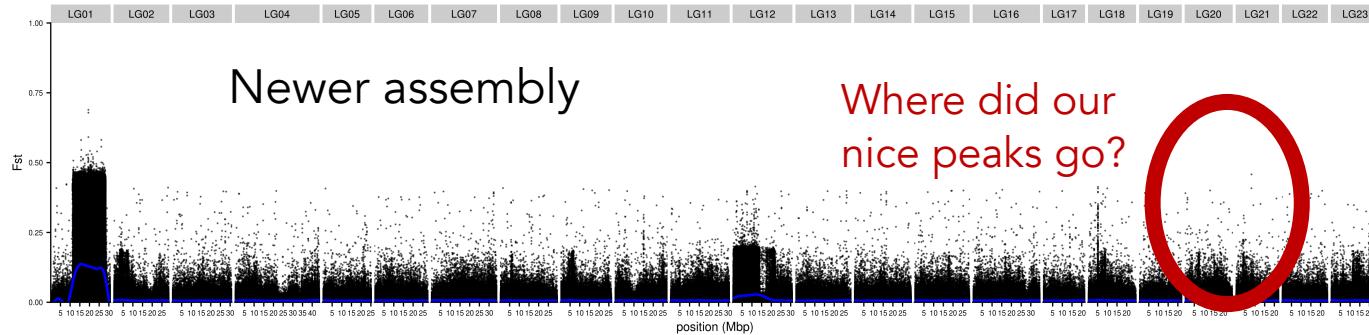
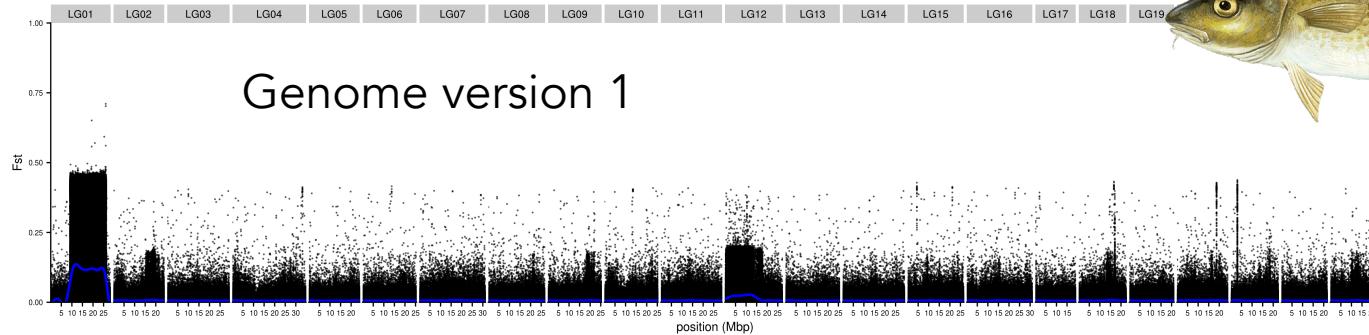
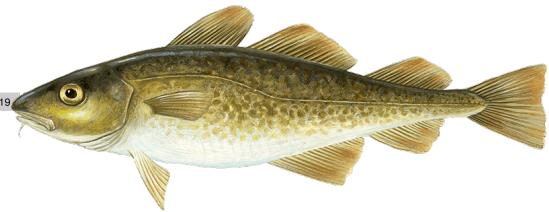


Fst in 10kb windows

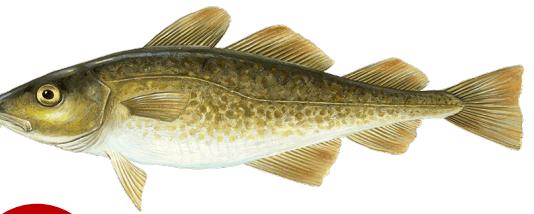
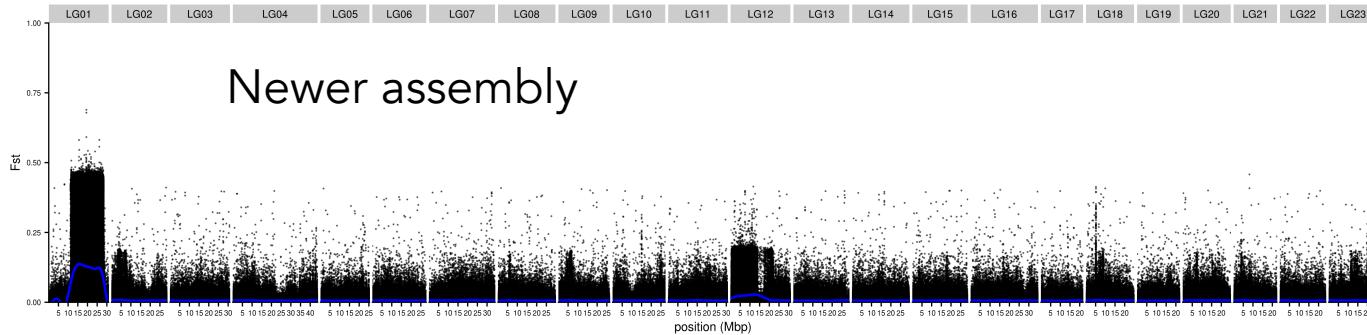
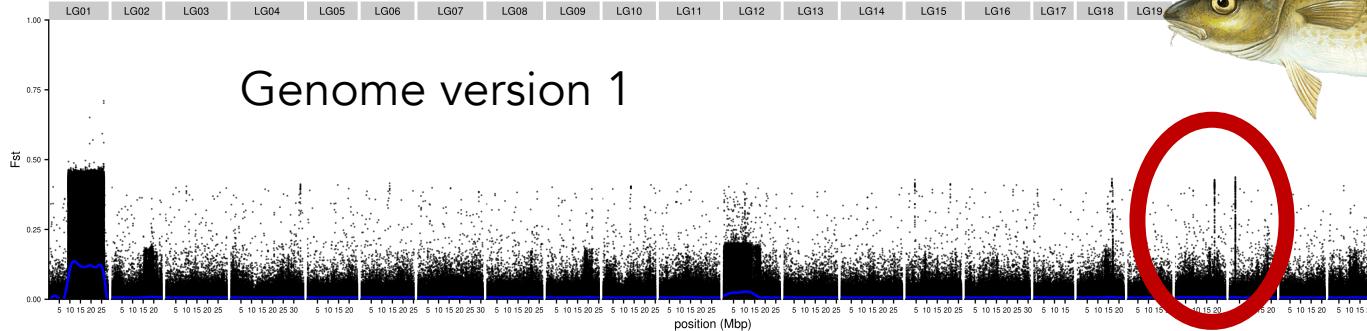
Examples from Atlantic cod



Examples from Atlantic cod



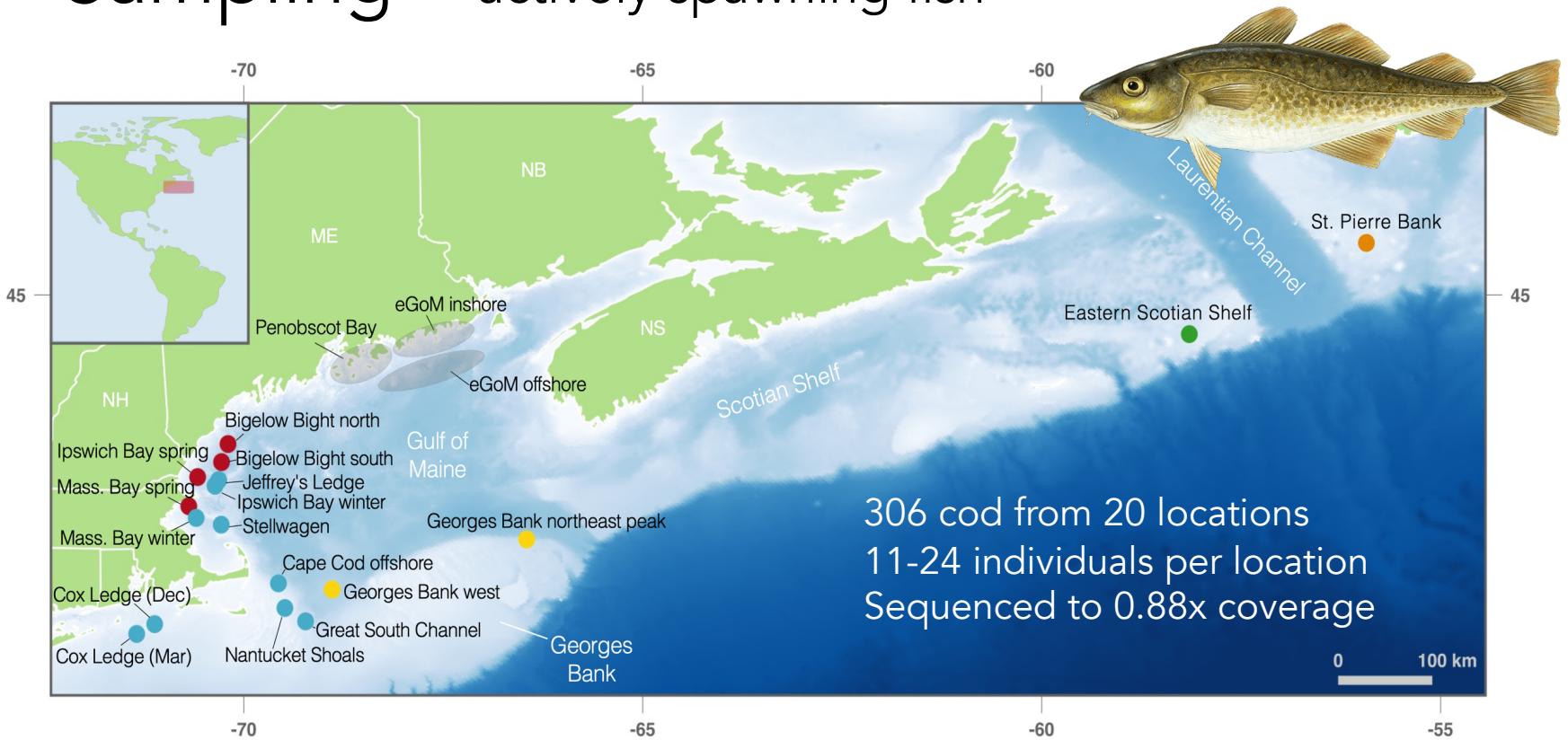
Examples from Atlantic cod



Are the isolated peaks in LD with SNPs in the inversions?

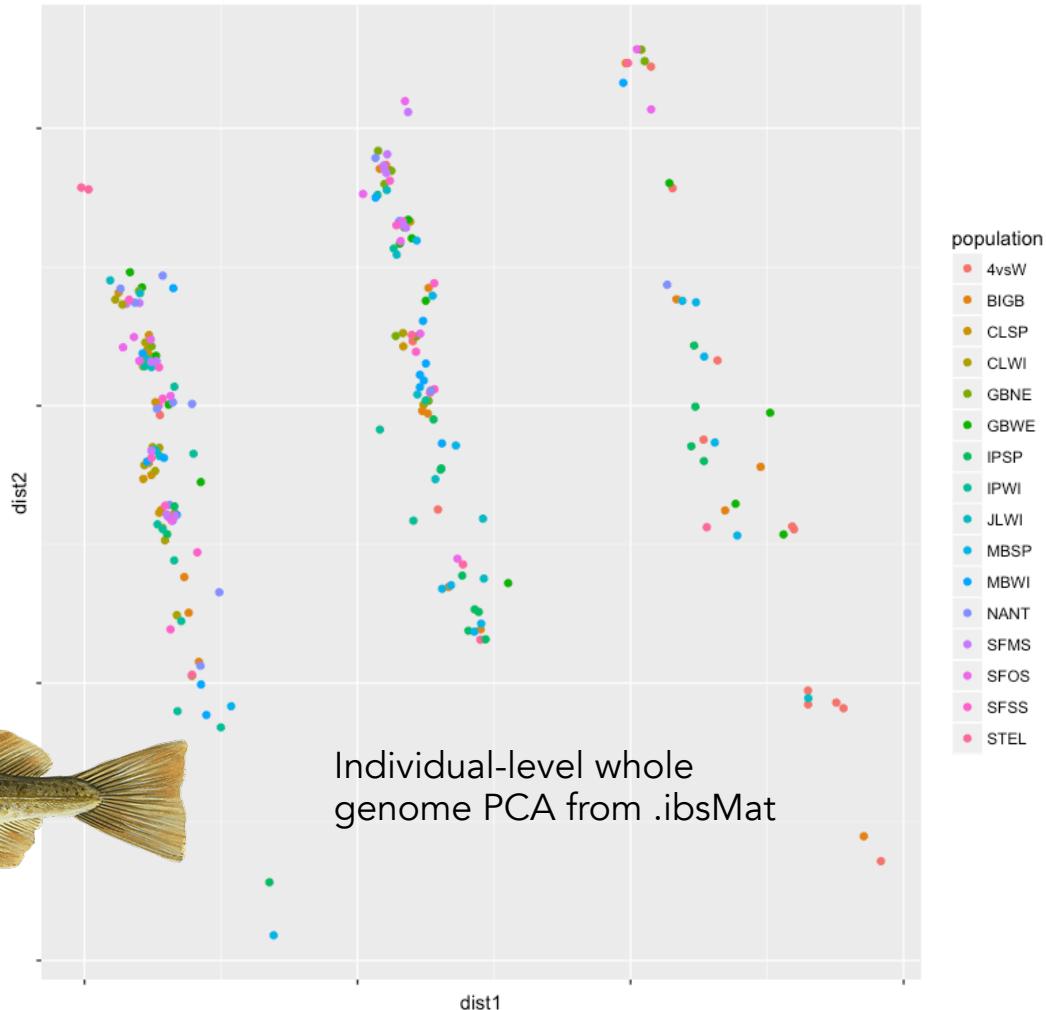
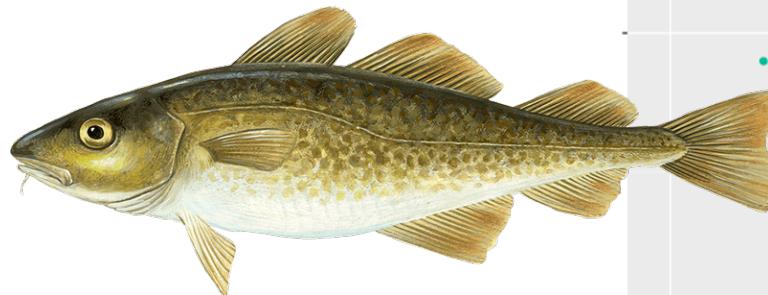
LD can also affect inference of population structure

Sampling – actively spawning fish

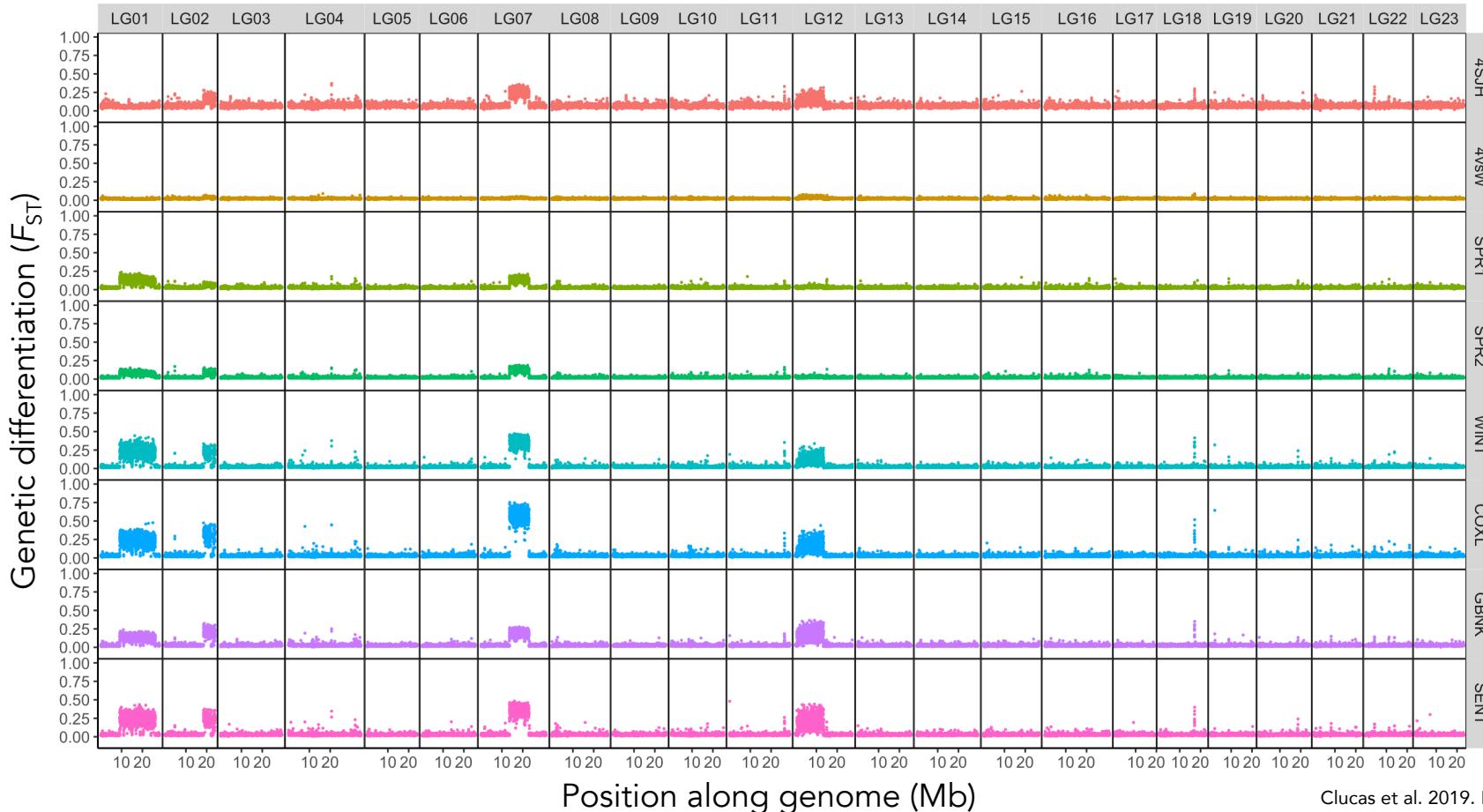


PCA

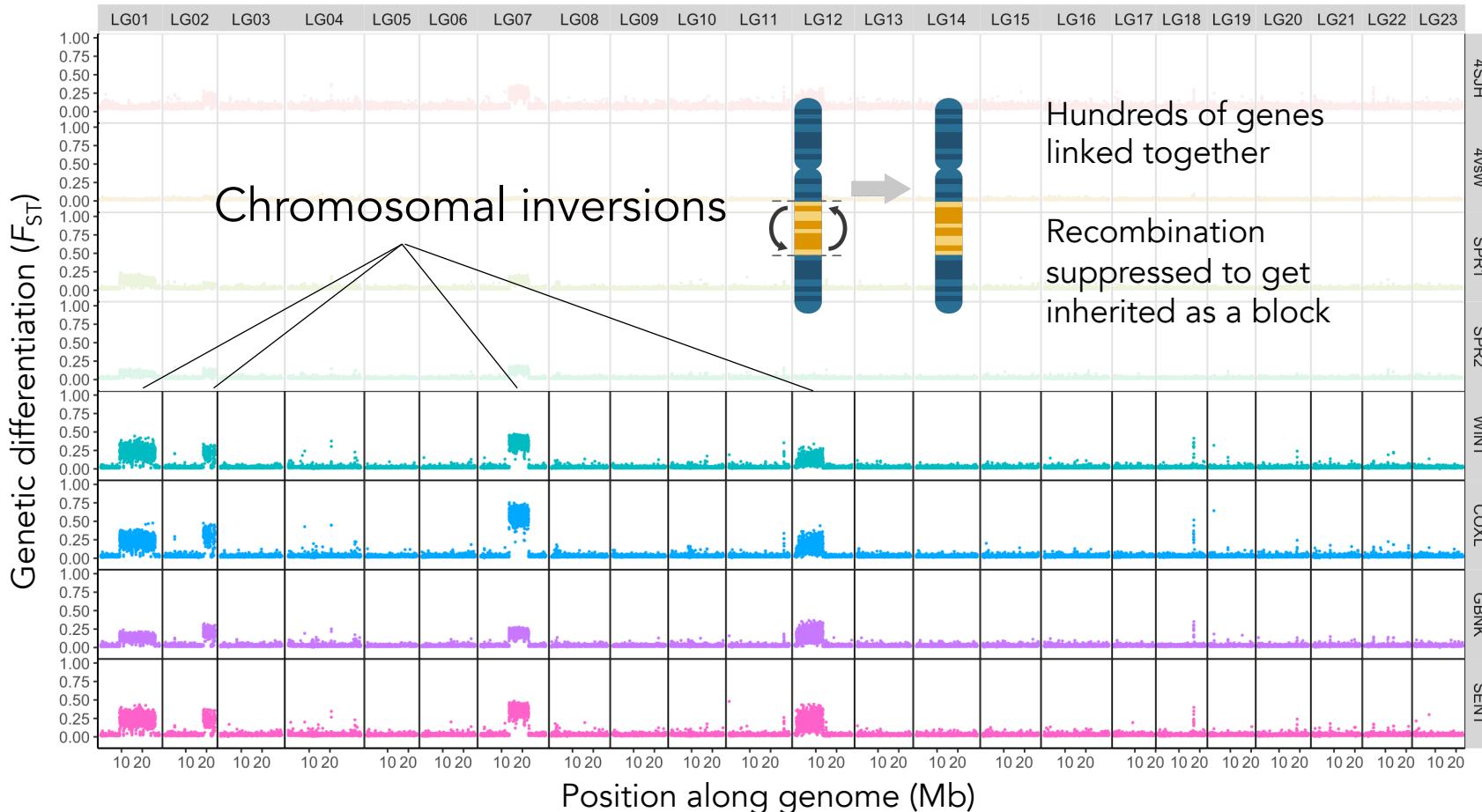
- What could explain this pattern?



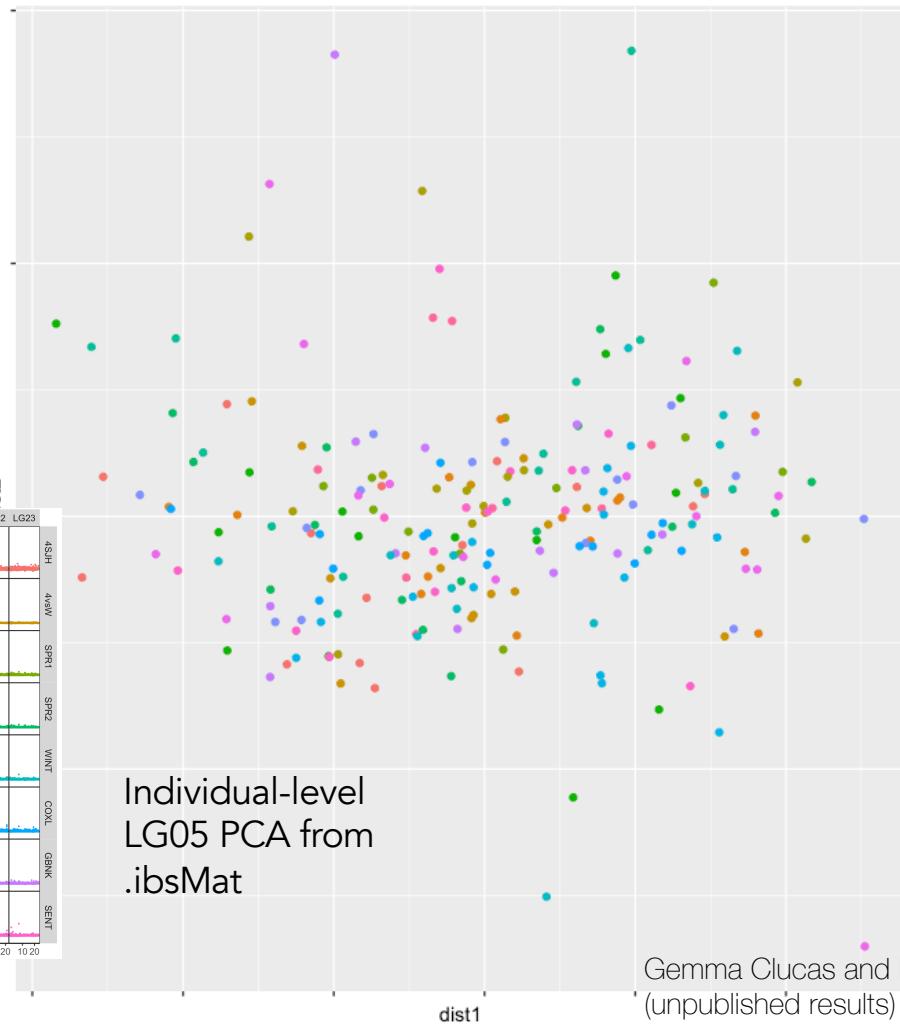
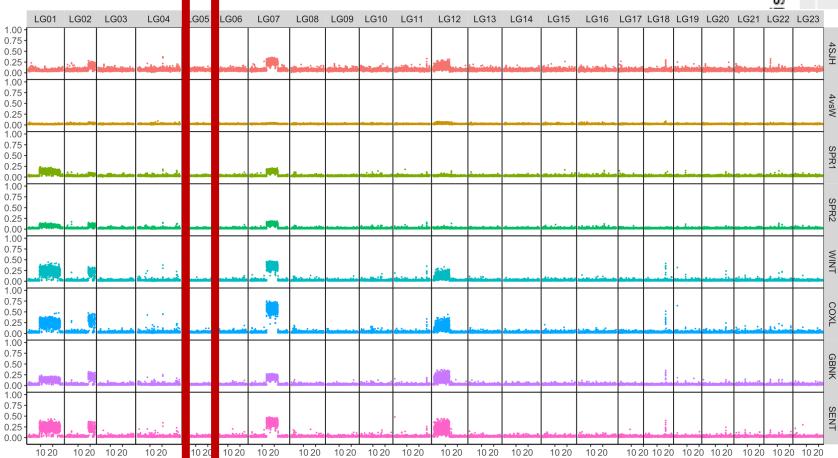
Pairwise comparisons to St. Pierre Bank, Canada (15 kb windows)



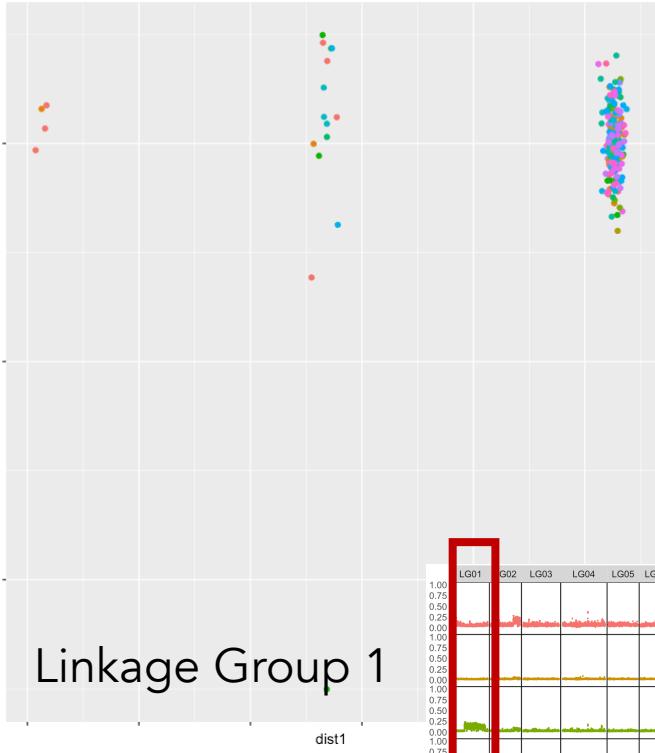
Pairwise comparisons to St. Pierre Bank, Canada (15 kb windows)



PCA with SNPs only on Linkage Group 5



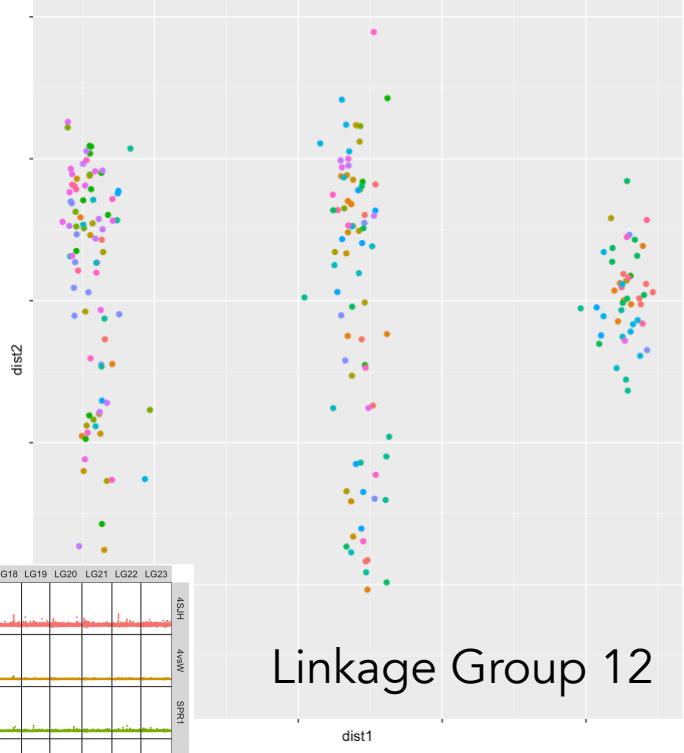
dist2



Plots of .ibsMat

Gemma Clucas and Nicolas Lou
(unpublished results)

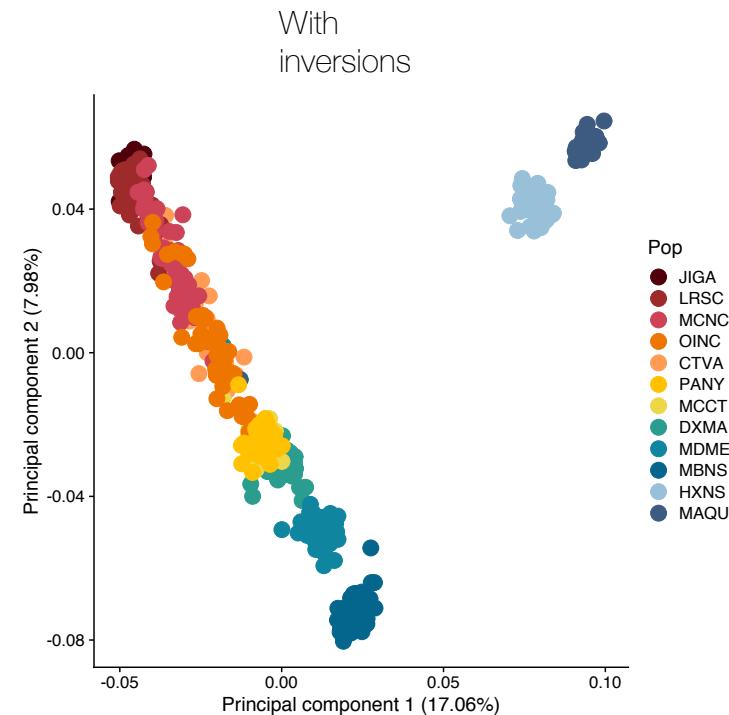
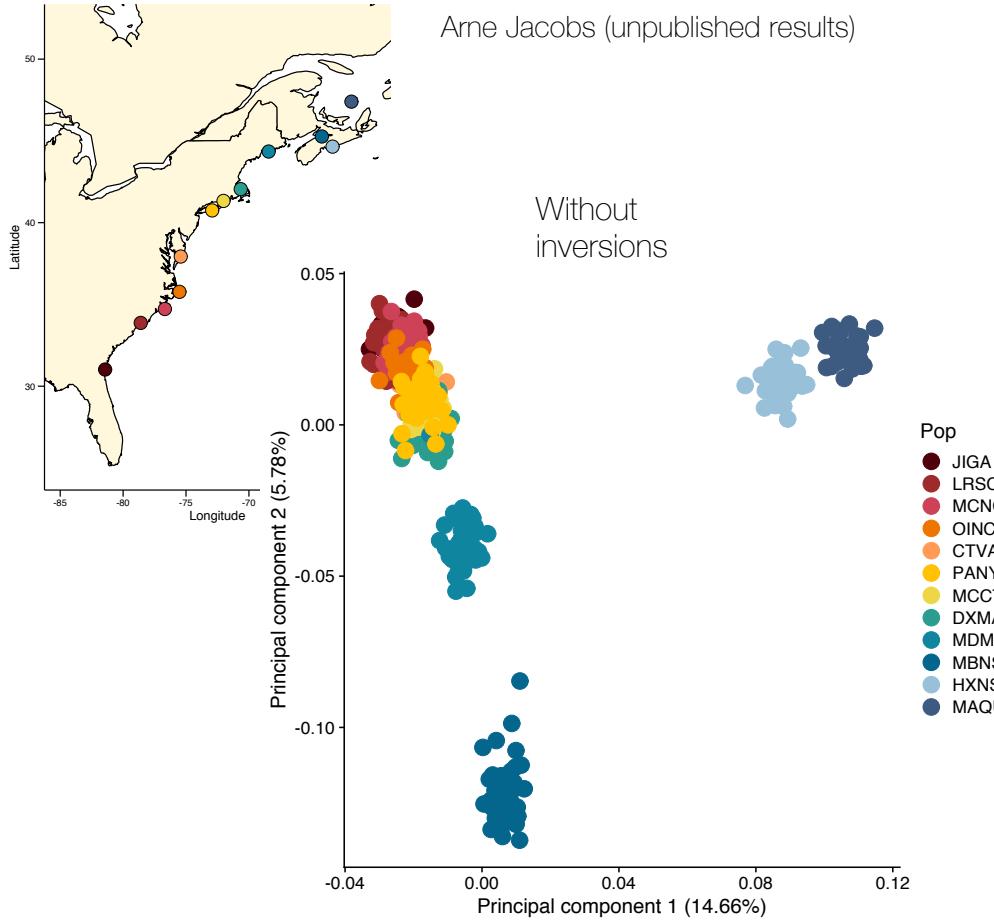
dist2



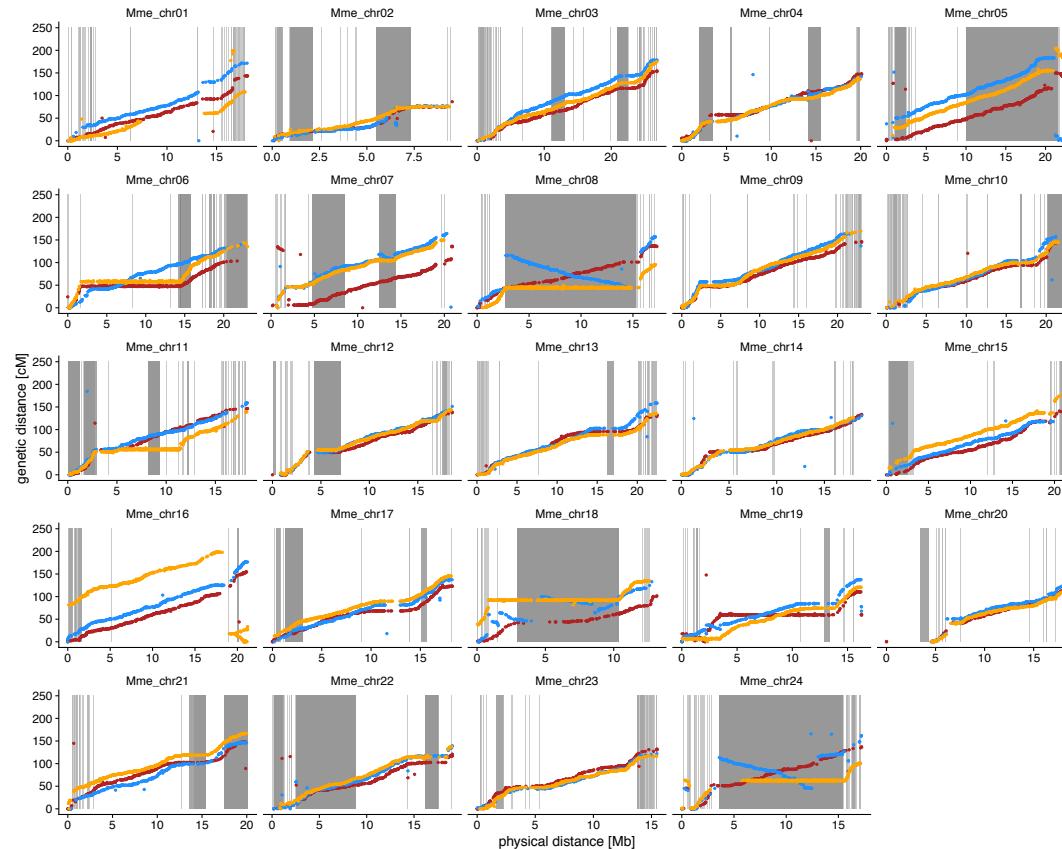
The need for LD pruning data sets

- Strongly linked regions are inherited as a single block, i.e. are effectively a single genetic marker
- If it contains many SNPs, this “single marker” will be way over-represented in analyses of geographic variation
- Many downstream analysis approaches assumes independence among markers

Another example – inversions don't always result in clustering into three genotype classes



Lots of inversions in silversides!





Genetics and population analysis

nsgLD: evaluating linkage disequilibrium using genotype likelihoods

Emma A. Fox¹, Alison E. Wright², Matteo Fumagalli¹ and Filipe G. Vieira^{3,*}

¹Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot, UK, ²Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK and ³Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on May 9, 2018; revised on December 27, 2018; editorial decision on March 17, 2019; accepted on March 20, 2019

Abstract

Motivation: Linkage disequilibrium (LD) measures the correlation between genetic loci and is highly informative for association mapping and population genetics. As many studies rely on called genotypes for estimating LD, their results can be affected by data uncertainty, especially when employing a low read depth sequencing strategy. Furthermore, there is a manifest lack of tools for the analysis of large-scale, low-depth and short-read sequencing data from non-model organisms with limited sample sizes.

Results: *nsgLD* addresses these issues by estimating LD directly from genotype likelihoods in a fast, reliable and user-friendly implementation. This method makes use of the full information available from sequencing data and provides accurate estimates of linkage disequilibrium patterns in both simulated and real data sets. We demonstrate its performance in a variety of applications, including the analysis of sequencing data from non-model organisms with limited sample sizes.