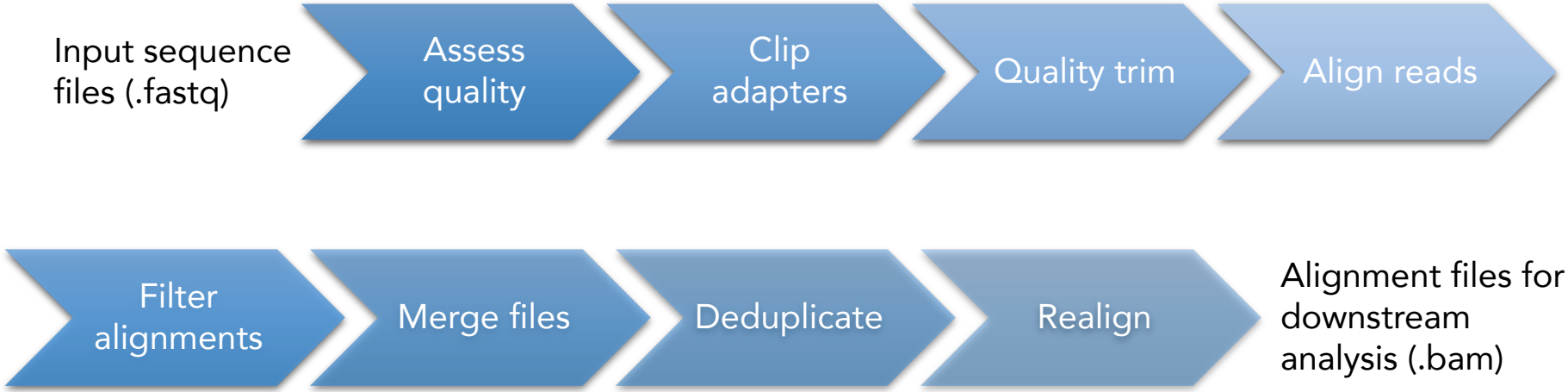
From fastq to bam

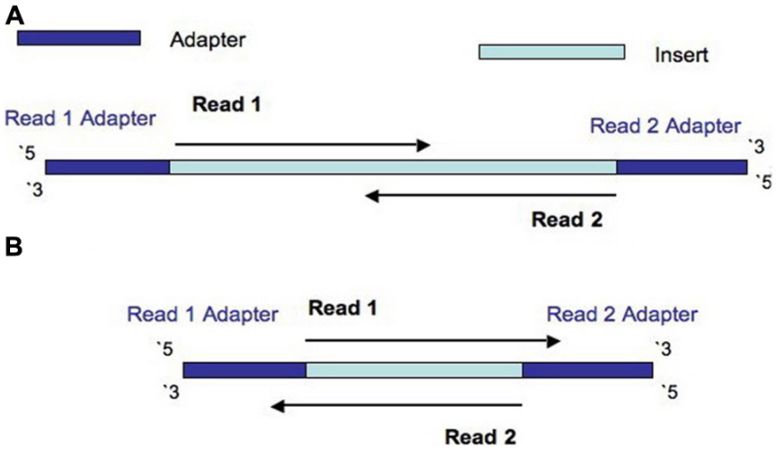
Bioinformatic pipeline



FastQC

Nice tool for diagnosing problems with your data

Read adapter read-through sequence

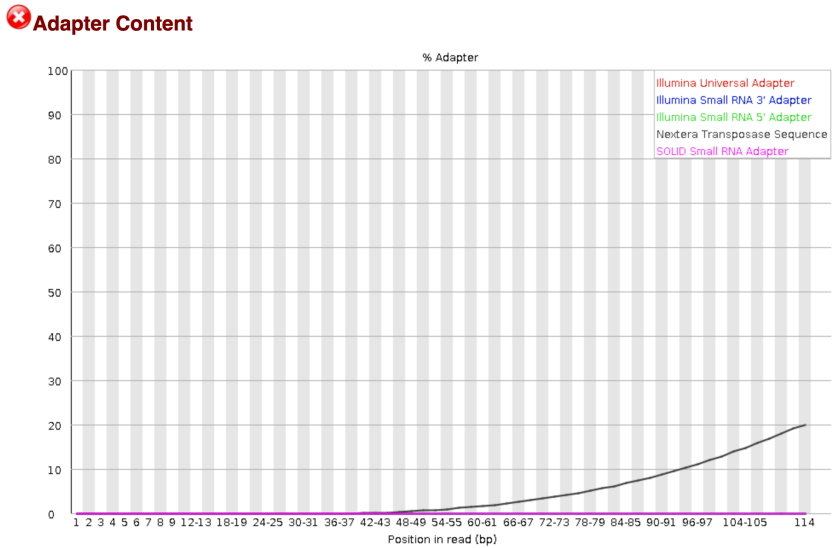


https://www.frontiersin.org/articles/10.3389/fgene.2014.00005/full

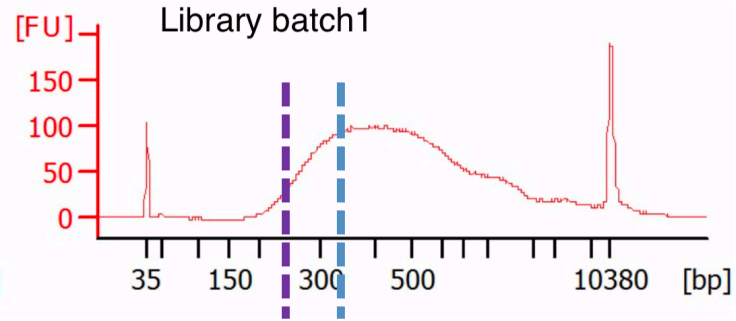
- If the library insert is shorter than the read length, the end of the read will be adapter sequence
- Adapter sequence can interfere with mapping and variant calling

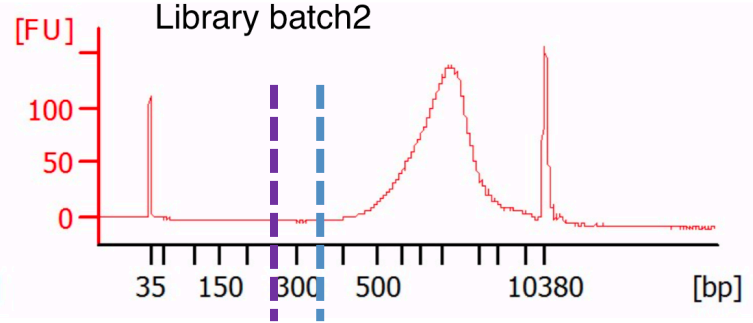
Assess quality | **Clip adapters** | Quality trim | Align reads

Read adapter read-through sequence



Two examples of our library pools





The length of Nextera adapters is 138 bp and libraries were sequenced with 2*125bp reads

→ Minimum fragment length to avoid overlap 383bp
→ Minimum fragment length to avoid adapter read-through 250bp

Base call quality scores

Measure the probability that a base is called incorrectly

$$Q = -10\log_{10}(e)$$

where e is the estimated probability of the base call being wrong

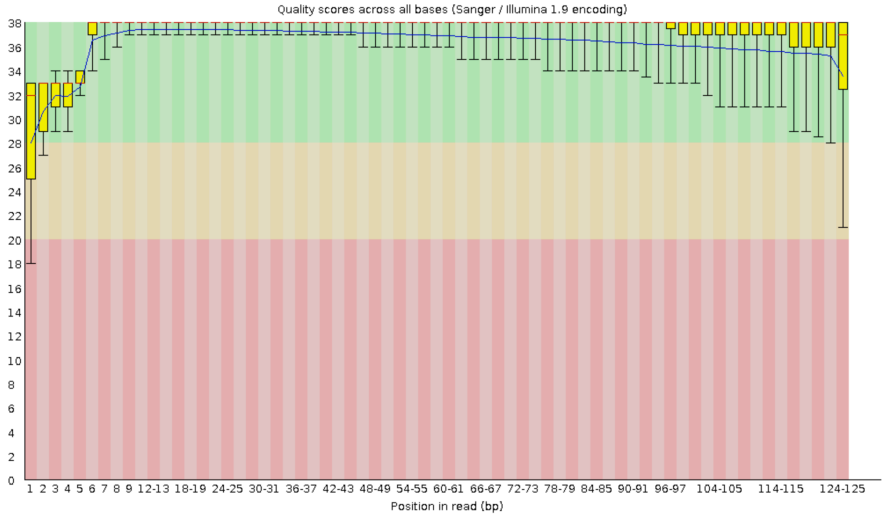
- **Higher Q scores** indicate a smaller probability of error
- **Lower Q scores** may lead to false variant calls

Quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

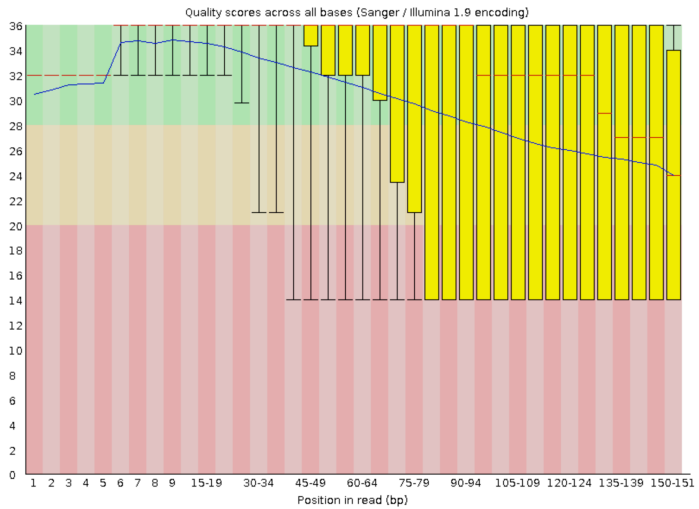
Assess quality | Clip adapters | **Quality trim** | Align reads

Sequence quality can vary dramatically between sequencing runs

Per base sequence quality



Per base sequence quality

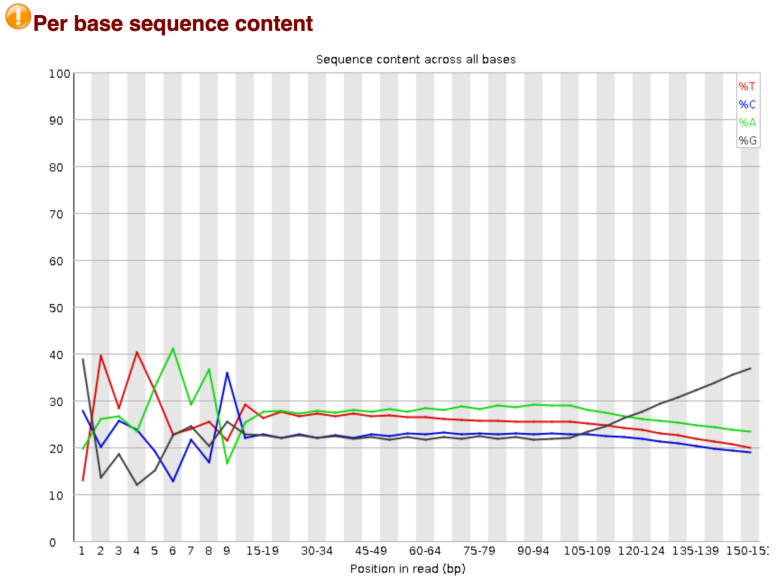


Quality trimming is optional

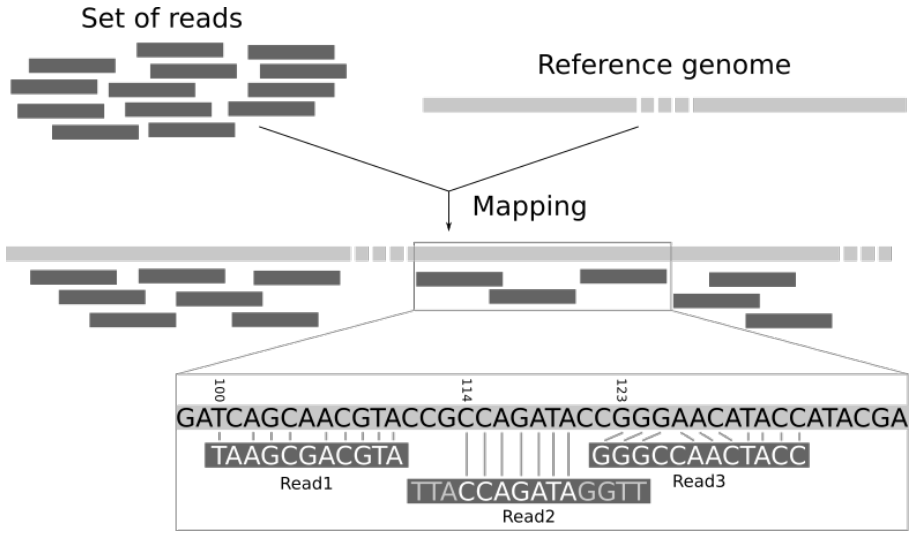
- Probabilistic downstream analysis frameworks take base quality scores into account
  - Retaining base calls with a 99% probably of being correct may be more valuable than discarding it (zero information)
- However, base call quality scores may be mis-calibrated (i.e. not reflecting the true probability of being correct)
- Low quality data can add noise to analysis

Assess quality → Clip adapters → **Quality trim** → Align reads

Other potential quality issue: poly-G tails in two-channel sequencers



Mapping to a reference sequence



Filter alignments

- Not always possible to map reads to its point of origin in the genome
  - E.g. duplicated and repetitive sequence
- Each alignment given a mapping quality (MapQ)

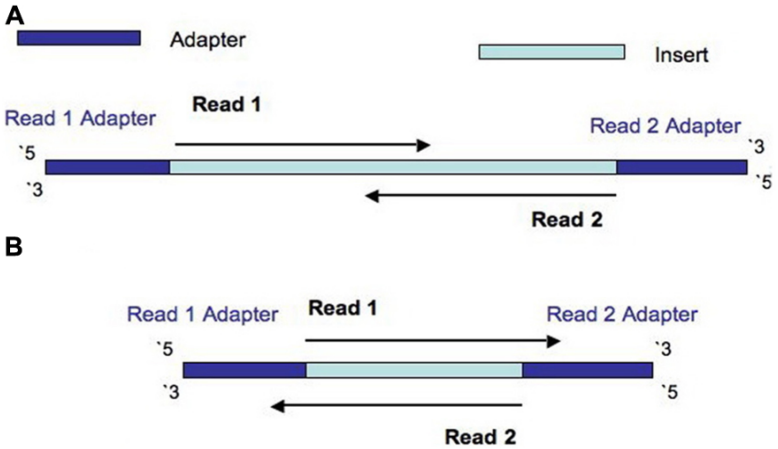
  $Q = -10\log_{10}(p)$,

  where p is an estimate of the probability that the alignment does not correspond to the read's true point of origin

- Mapping quality related to uniqueness
- Reads that map in multiple places can bias analysis

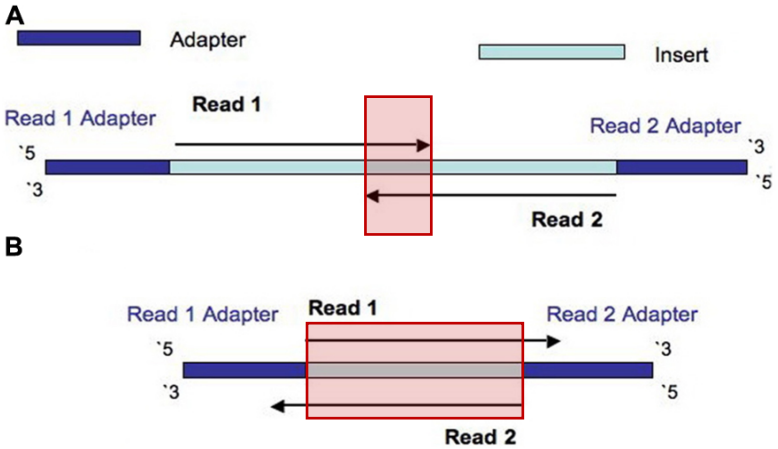
MapQ	Probability that read truly originated from different place
10	1 in 10
20	1 in 100
30	1 in 1000

Clip read overlap



https://www.frontiersin.org/articles/10.3389/fgene.2014.00005/full

Clip read overlap



https://www.frontiersin.org/articles/10.3389/fgene.2014.00005/full

Overlapping reads of the same DNA fragment

Avoid double-counting by soft-clipping the overlap

Merge all alignment files with sequence from the same library

- Reads from separate sequencing runs should be mapped separately to keep read group identifier

- But we need
  - A single bam file per library for deduplication
  - A single bam per individual for downstream analysis

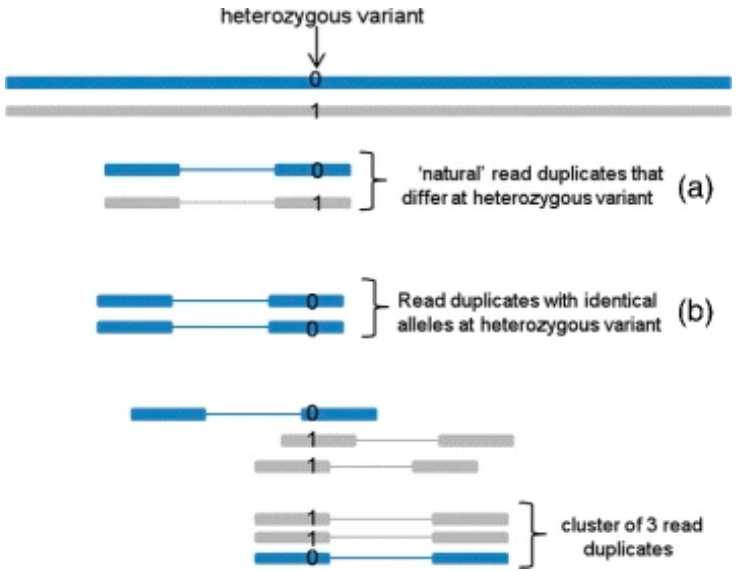
Remove duplicate sequence

- Duplicate reads are defined as originating from a single fragment of DNA
  - PCR duplicates (arise during library preparation)
  - Optical duplicates (arise during sequencing)

- Duplicates can bias our downstream analysis

Filter alignments → Merge files → **Deduplicate** → Realign

Remove duplicate sequence



https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1471-9

Realign around indels

- Genome aligners can only consider each read independently, and the scoring strategies they use to align reads relative to the reference limit their ability to align reads well in the presence of indels

example: 1000G PhaseI low coverage
chr15:81551110, ref:CTCTC alt:ATATA

Interpreted as 3 SNPs

Interpreted as microsatellite expansion/contraction

Slide from Erik Garrison

example: 1000G PhaseI low coverage
chr20:708257, ref:AGC alt:CGA

```
ref: TATAGAGAGAGAGAGAGAGCGAGAGAGAGAGAGAGAGGGAGAGACGGAGTT
alt: TATAGAGAGAGAGAGAGCGAGAGAGAGAGAGAGAGAGGGAGAGACGGAGTT
```

```
ref: TATAGAGAGAGAGAGAGAGC--GAGAGAGAGAGAGAGAGGGAGAGACGGAGTT
alt: TATAGAGAGAGAGAGAG--CGAGAGAGAGAGAGAGAGGGAGAGACGGAGTT
```

Slide from Erik Garrison

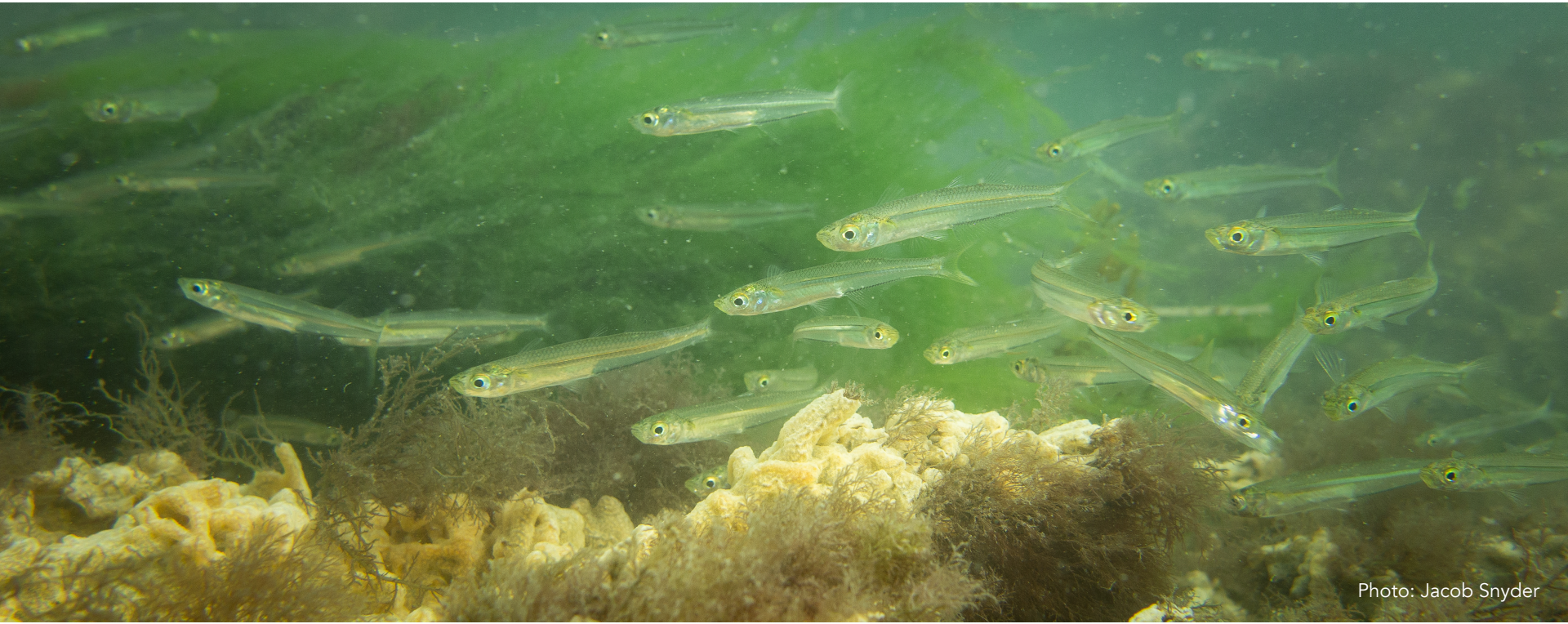
Realign around indels

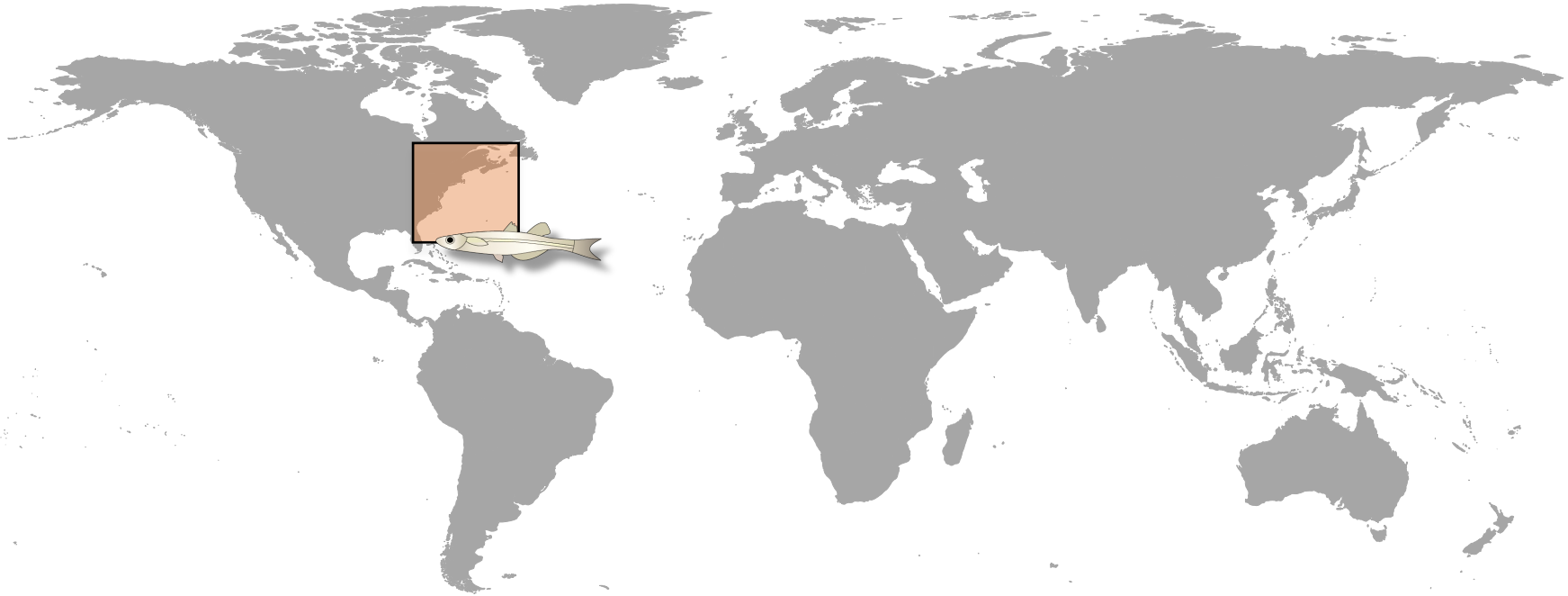
- Genome aligners can only consider each read independently, and the scoring strategies they use to align reads relative to the reference limit their ability to align reads well in the presence of indels

- Local realignment considers all reads spanning a given position.
  - Can more confidently infer indel polymorphisms

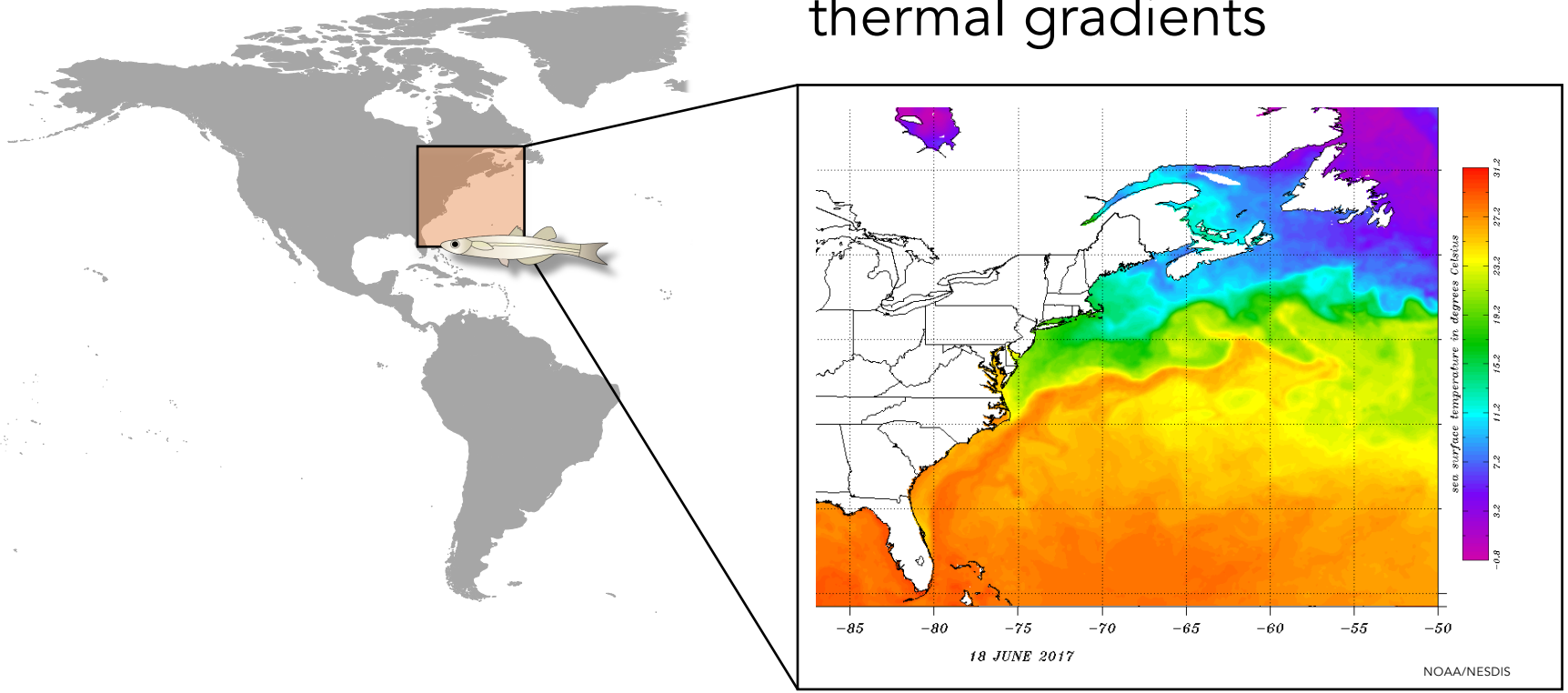
- Then all reads can be realigned around known indels

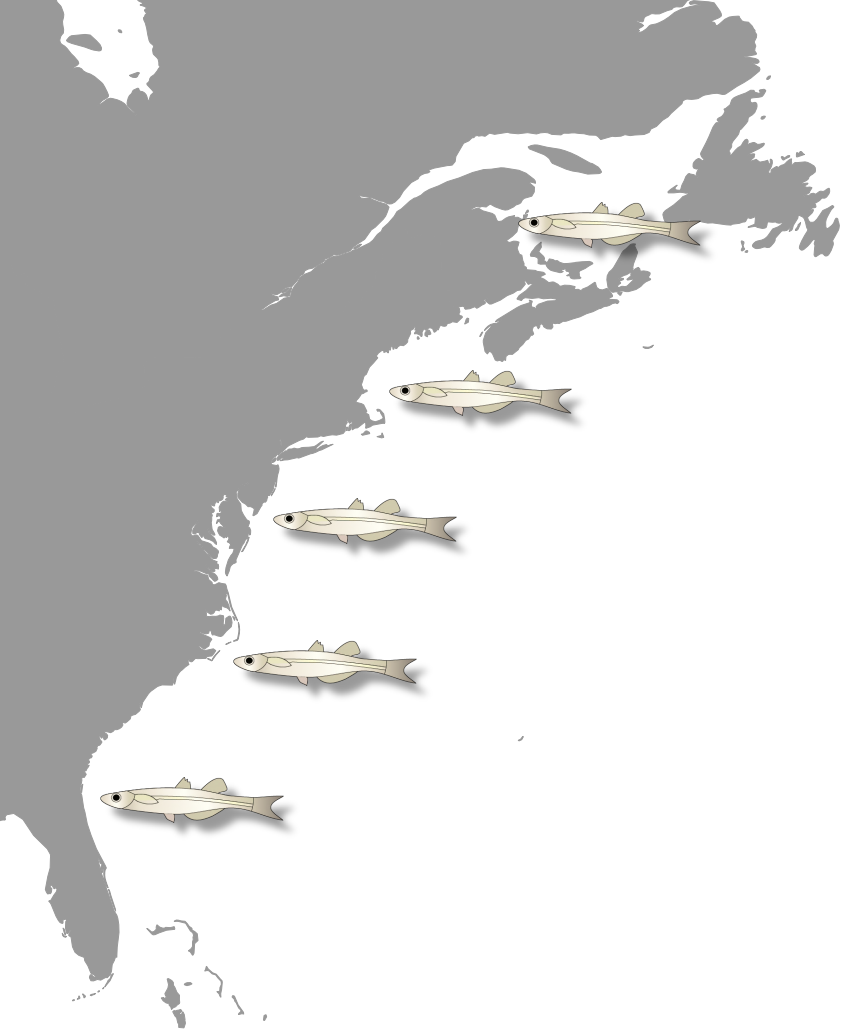
Atlantic silverside *Menidia menidia*



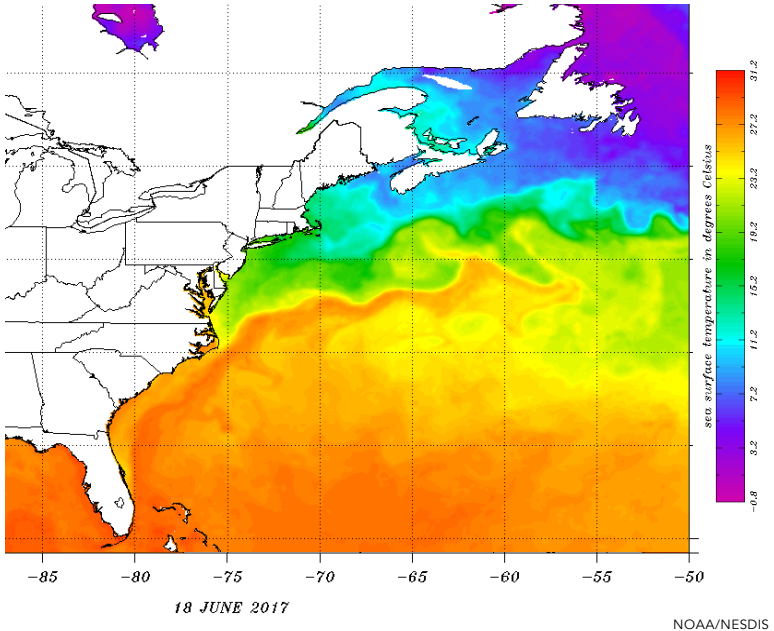


One of the world's steepest thermal gradients

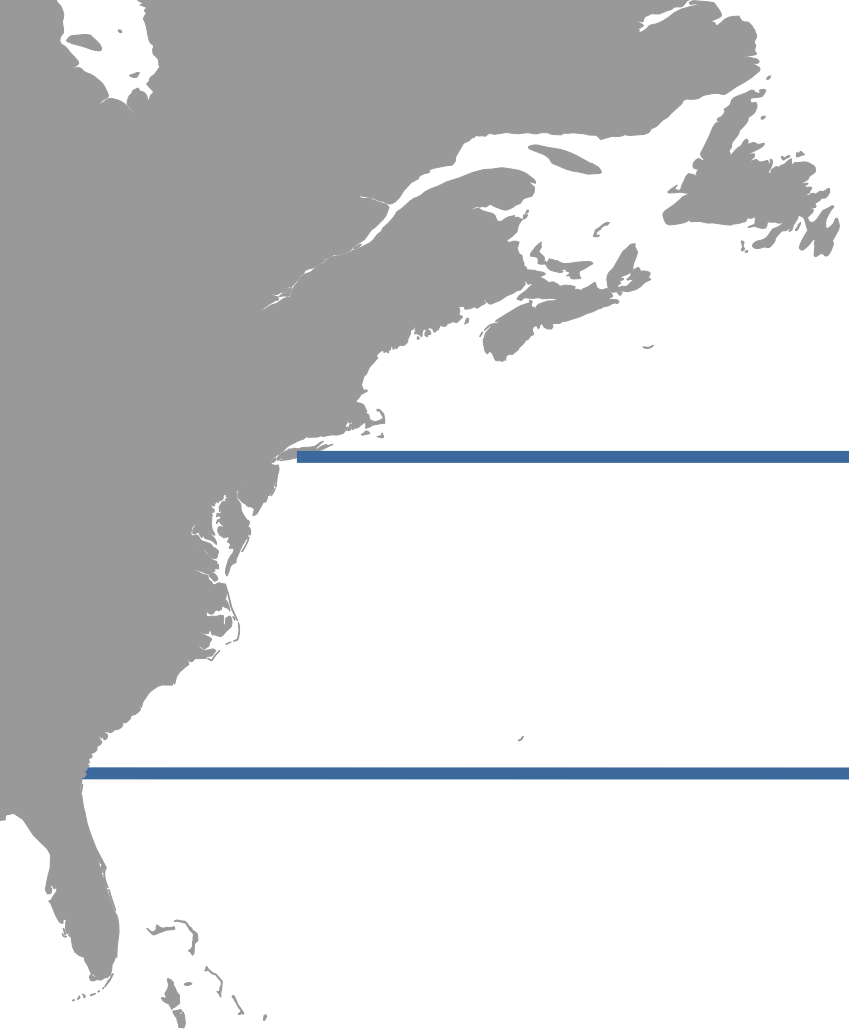




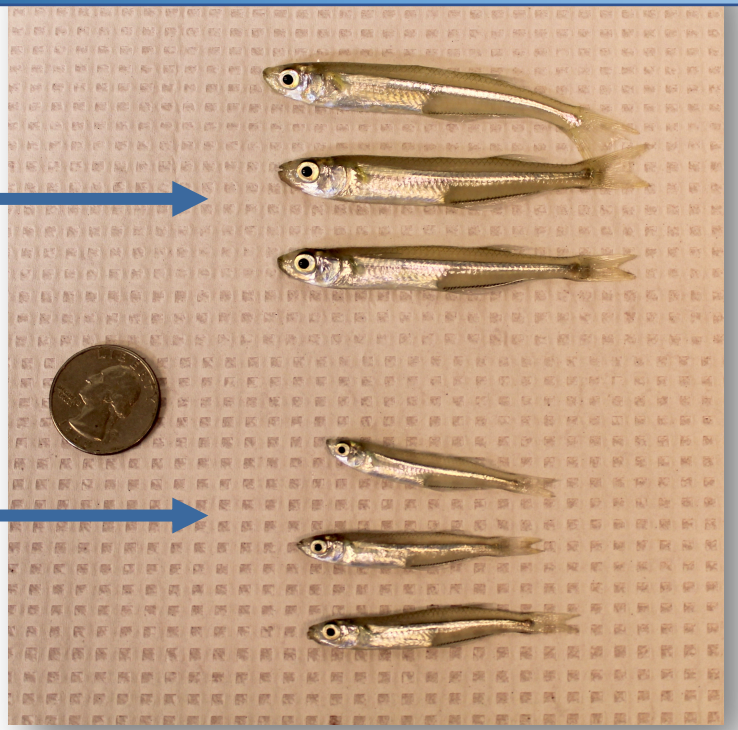
One of the world's steepest thermal gradients



NOAA/NESDIS

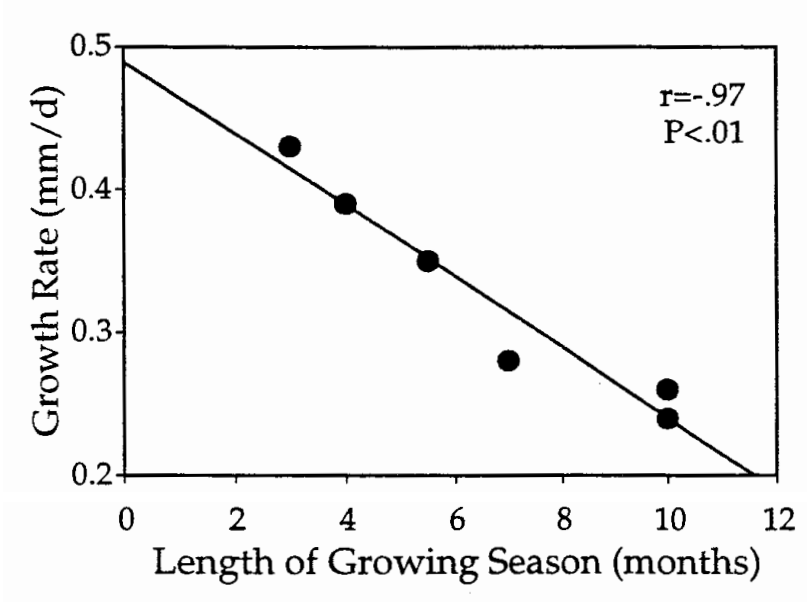


Same age,
Common lab environment



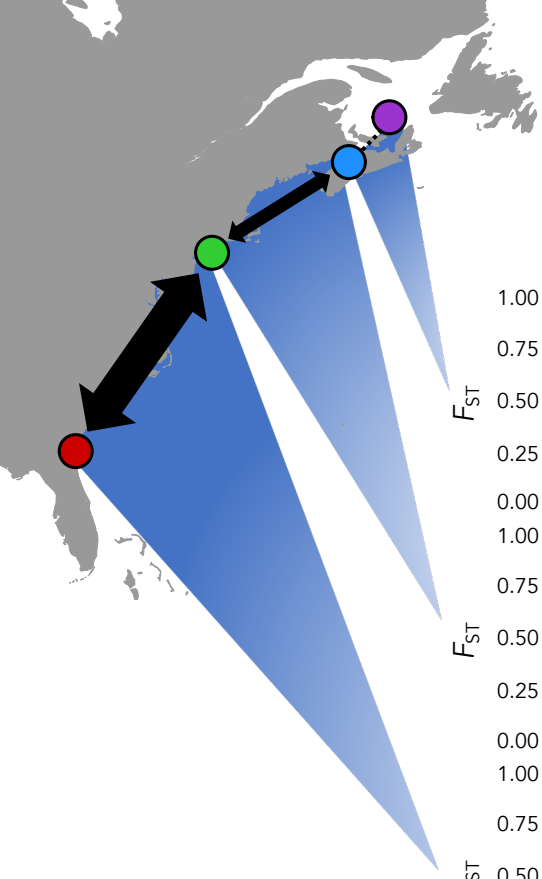
Growth capacity is tightly correlated with latitude

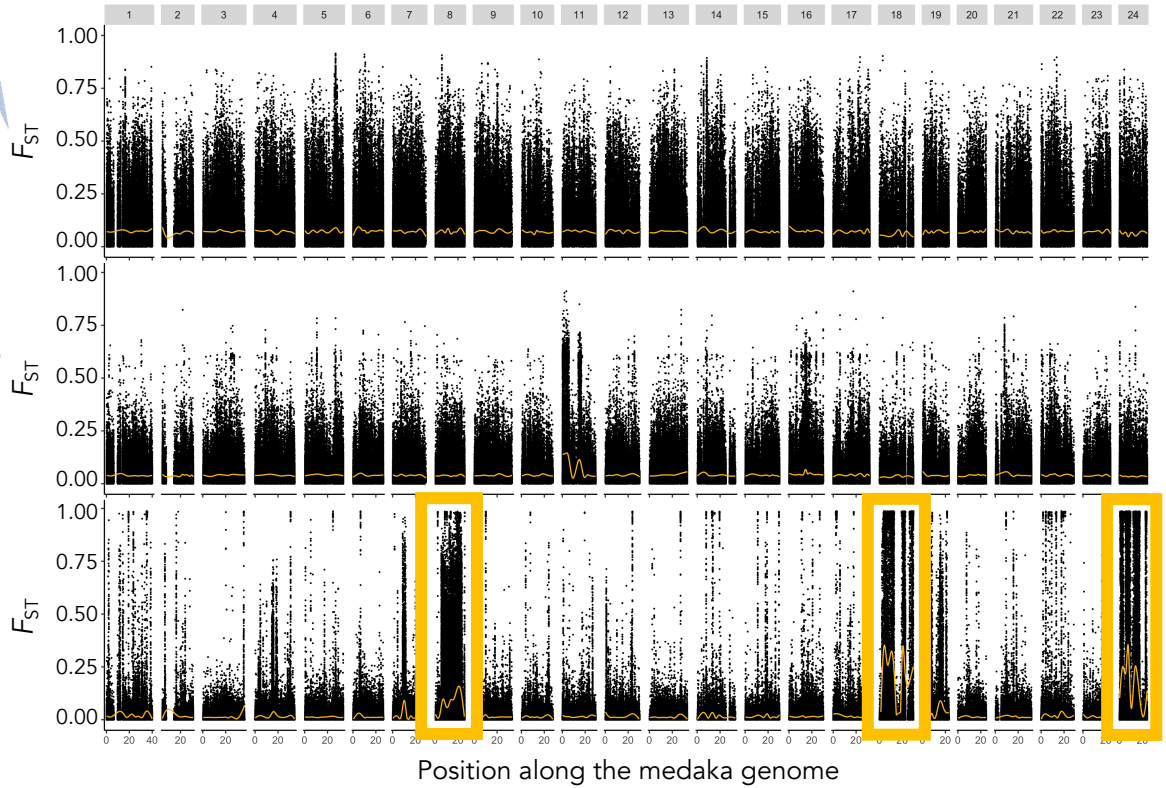
David Conover et al.



Reproduced from Conover. 1998. Bull. Mar. Sci.

Heterogeneous gene flow and divergence patterns across latitudes

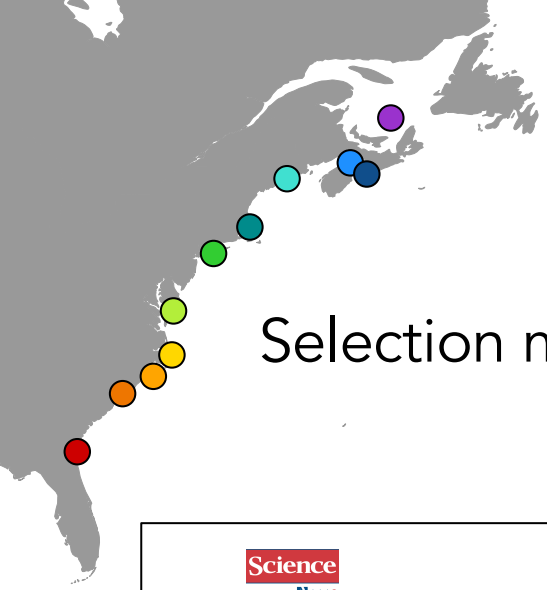


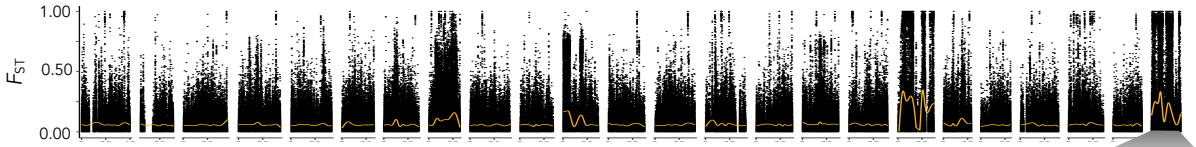


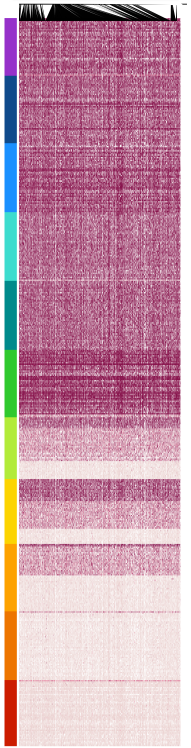
Median F_{ST} = 0.036
No fixed SNPs

Median F_{ST} = 0.024
No fixed SNPs

Median F_{ST} = 0.006
>13,000 fixed SNPs







Selection mapping through a fisheries experiment

Science
AAAS

REPORTS

Sustaining Fisheries Yields Over Evolutionary Time Scales

David O. Conover* and Stephan B. Munch

Fishery management plans ignore the potential for evolutionary change in harvestable biomass. We subjected populations of an exploited fish (*Menidia menidia*) to large, small, or random size-selective harvest of adults over four generations. Harvested biomass evolved rapidly in directions counter to the size-dependent force of fishing mortality. Large-harvested populations initially produced the highest catch but quickly evolved a lower yield than controls. Small-harvested populations did the reverse. These shifts were caused by selection of genotypes with slower or faster rates of growth. Management tools that preserve natural genetic variation are necessary for long-term sustainable yield.

It is well established that wild pest and pathogen populations may evolve in response to anthropogenic forces of mortality (*1*), but is the same true of fisheries? Fishing mortality is highly selective. Exploited stocks typically display greatly truncated size and age distributions that lack larger and/or older individuals (*2*–*4*). This occurs not only because fishers may seek to exploit large individuals but also because regulatory measures often impose minimum size or gear regulations that ensure selective harvest of

ing behavior], with one major exception. The short generation time of *M. menidia* (1 year) coupled with the ease with which large populations can be maintained in captivity enable experimental designs that would otherwise be impossible. Second, *M. menidia* from different latitudes display clinal adaptive genetic variation in somatic growth rate (*12*), a geographical pattern common to other harvested species (*13*–*16*). Hence, a key production trait (somatic growth rate) appears capable of evolving in the