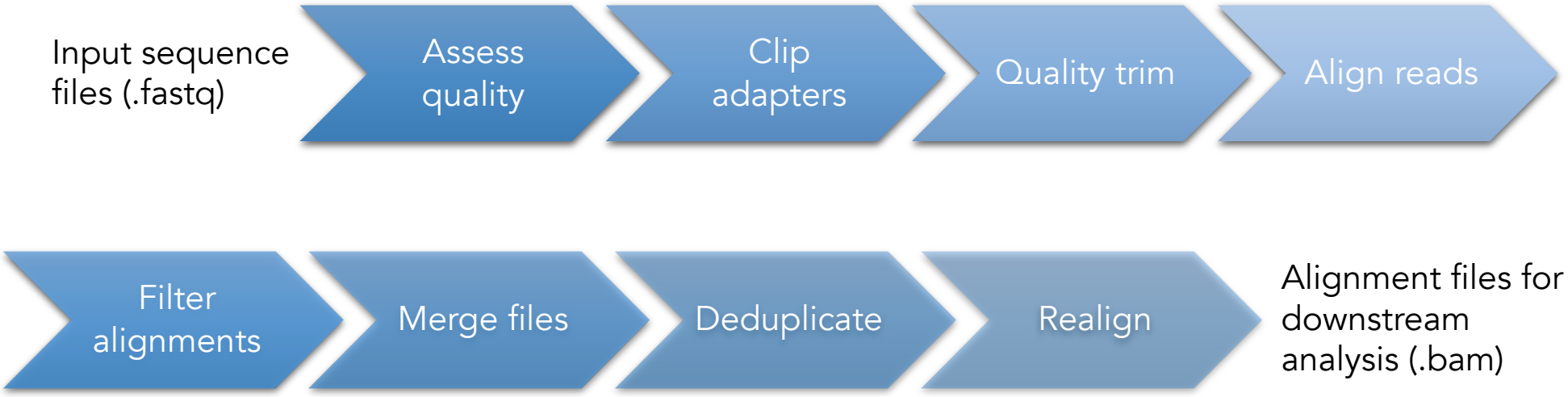
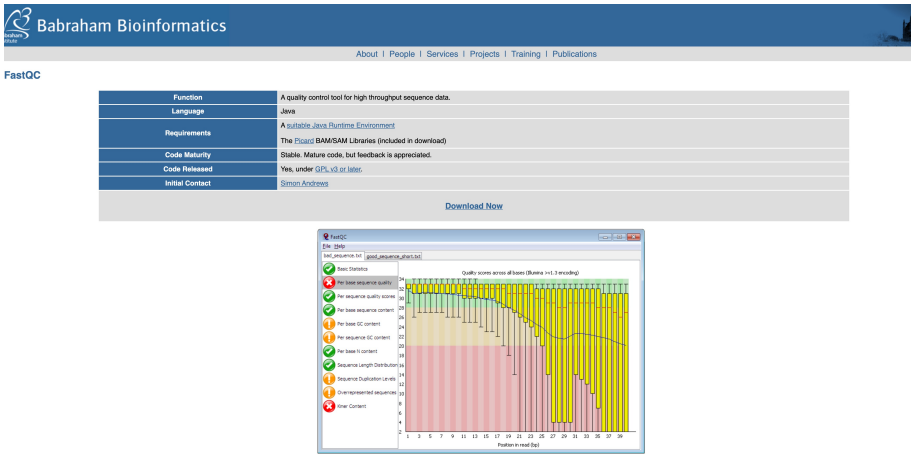
From fastq to bam

Bioinformatic pipeline



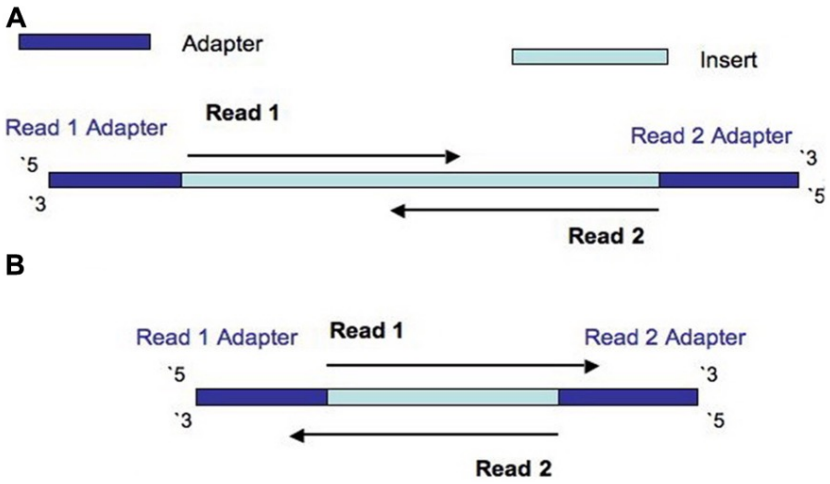
FastQC

Nice tool for diagnosing problems with your data



Assess quality | Clip adapters | Quality trim | Align reads

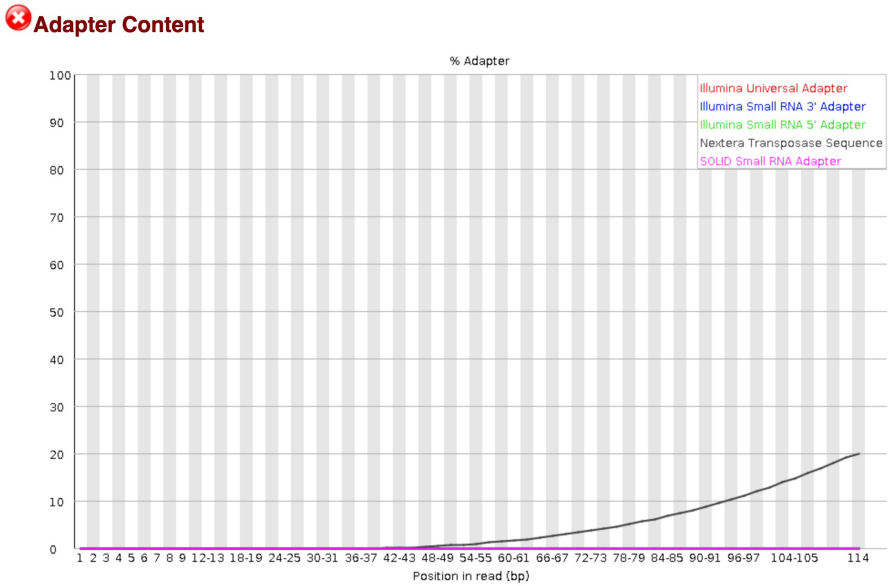
Read adapter read-through sequence



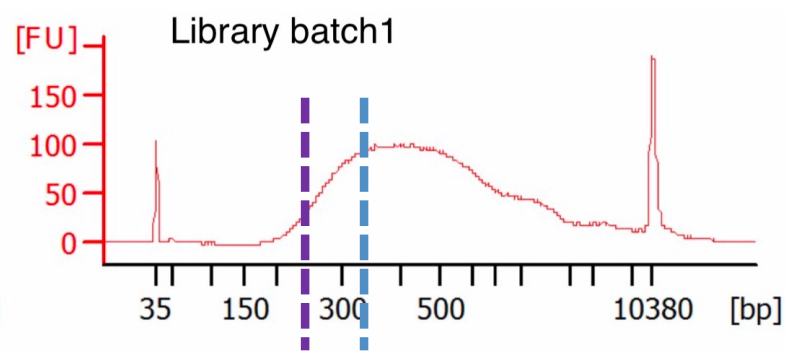
https://www.frontiersin.org/articles/10.3389/fgene.2014.00005/full

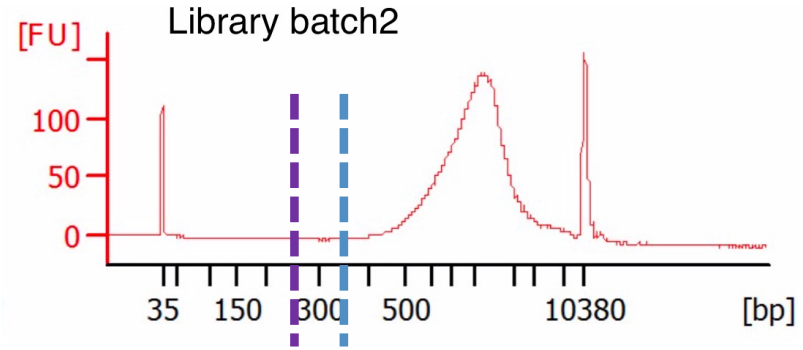
- If the library insert is shorter than the read length, the end of the read will be adapter sequence
- Adapter sequence can interfere with mapping and variant calling

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

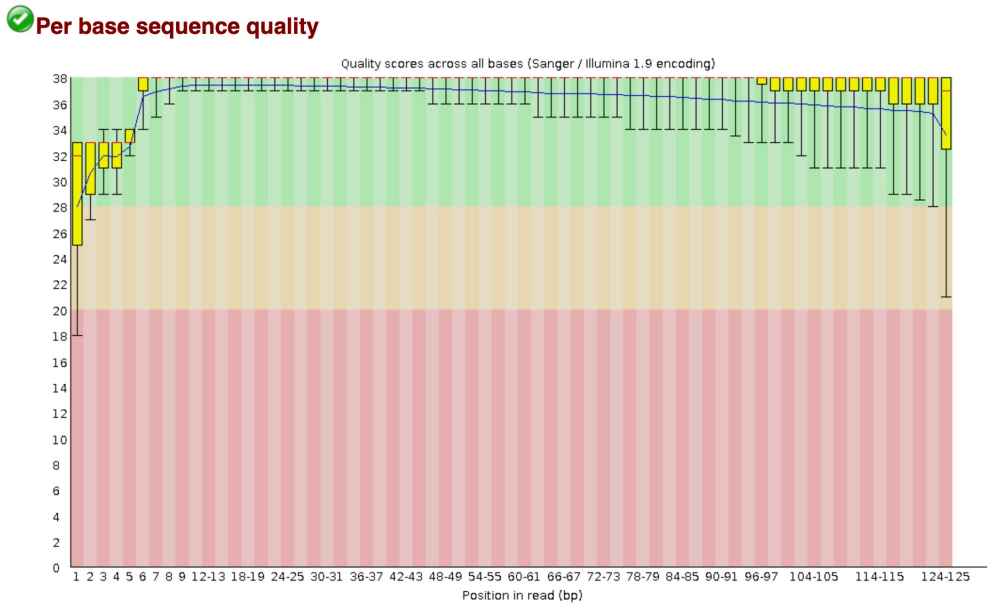
$$Q = -10\log_{10}(e)$$

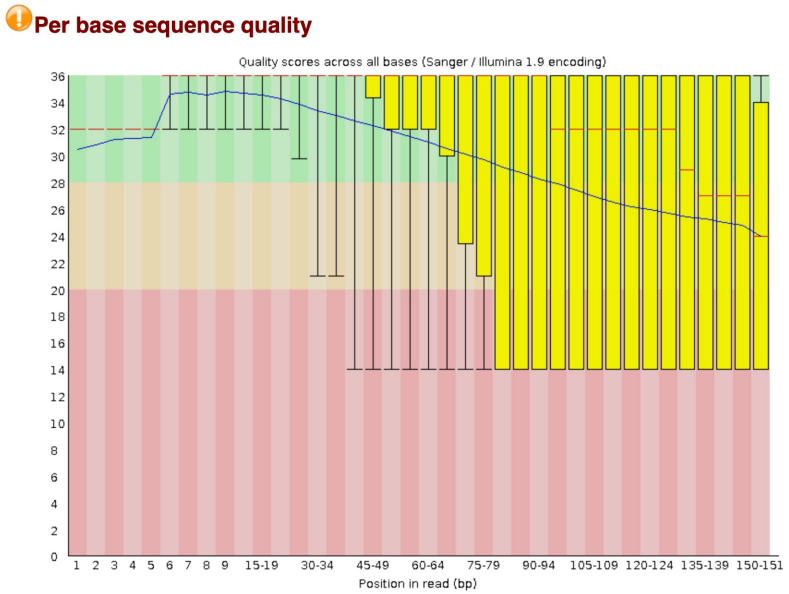
where e is the estimated probability of the base call being wrong

- **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%

Sequence quality can vary dramatically between sequencing runs

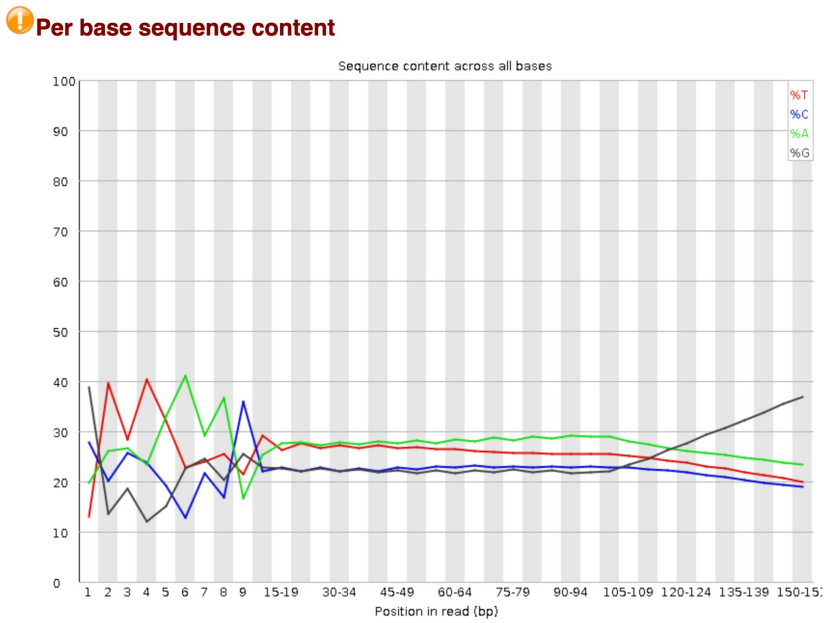




Quality trimming is optional

- 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

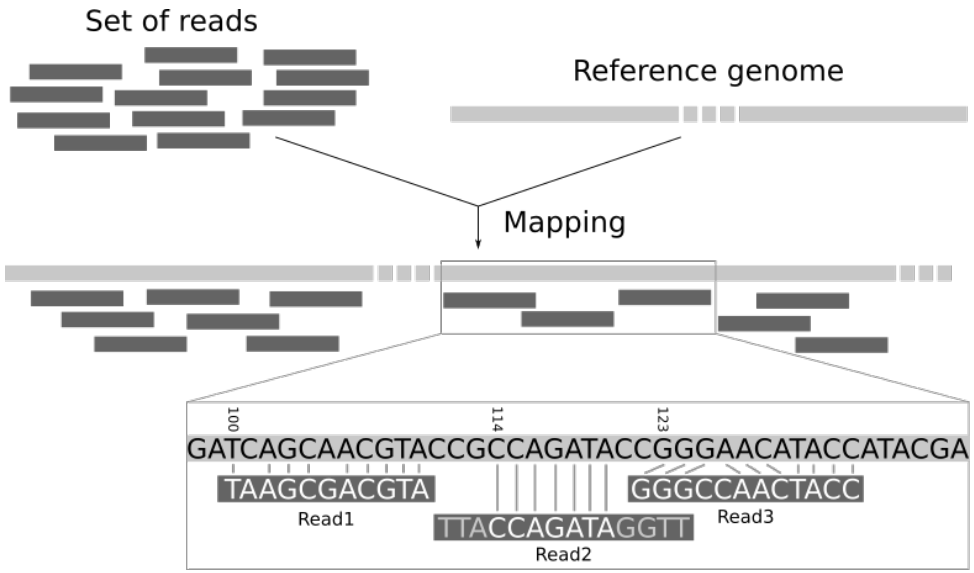
Other potential quality issue: poly-G tails in two-channel sequencers



Can remove with dedicated poly-G trimming software, but we have seen better results from simple sliding window quality score trimming

Lou*, R. N. & Therkildsen, N. O. 2021. Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, identification, and mitigation. **Authorea** doi: 10.22541/au.162791857.78788821/v2.

Mapping to a reference sequence



Filter alignments

- Not always possible to map reads to their 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

Filter alignments

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

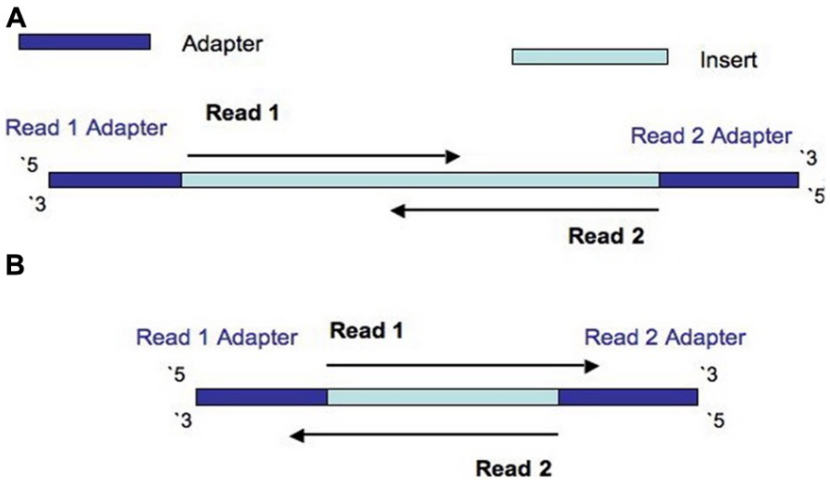
  $Q = -10\log_{10}(p)$,

  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

Mapping quality is not considered in estimation of genotype likelihoods in current software
(so you may want to be conservative in filtering)

Filter alignments → Merge files → Deduplicate → Realign

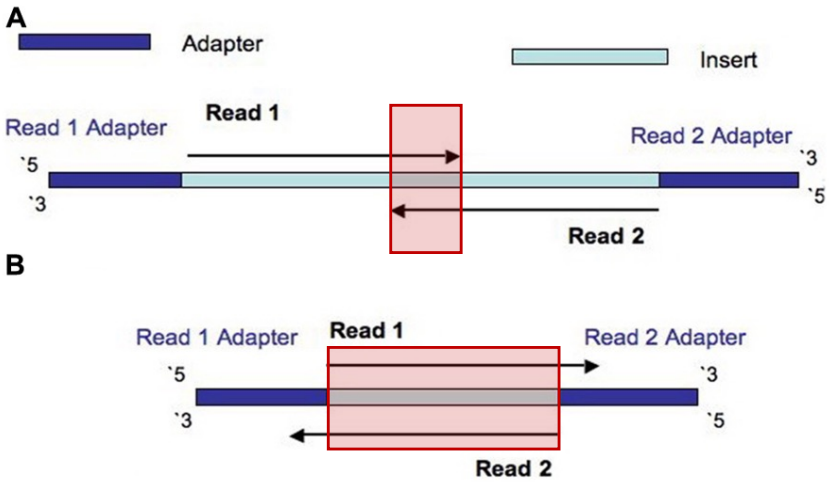
Clip read overlap



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

Filter alignments | Merge files | Deduplicate | Realign

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

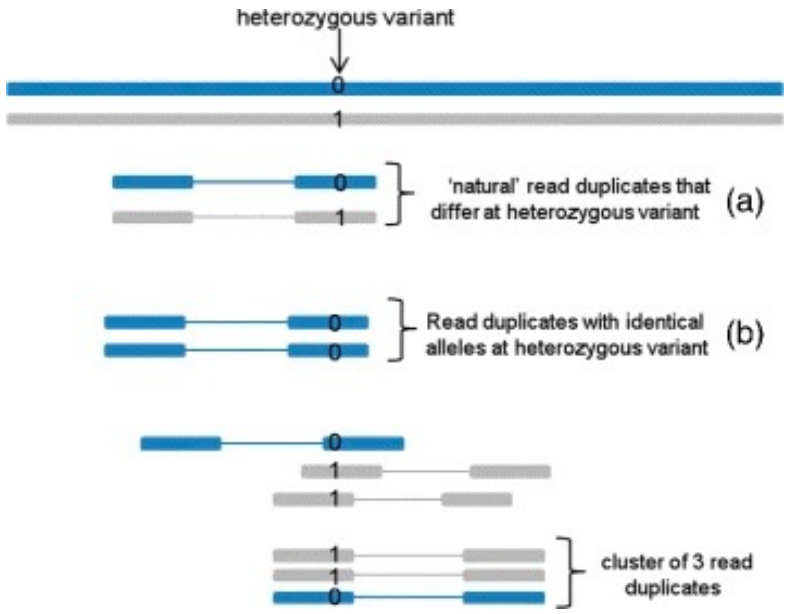
Remove duplicate sequence

- 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

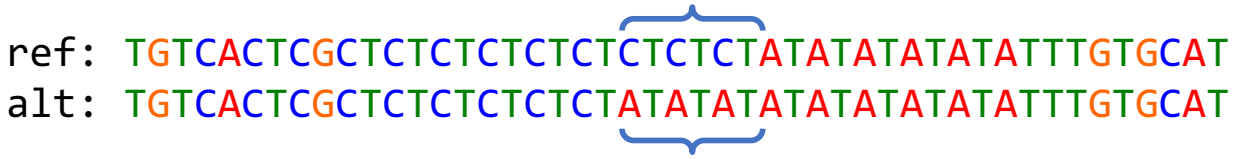


https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1471-9

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 Phase1 low coverage
chr15:81551110, ref:CTCTC alt:ATATA



Interpreted as 3 SNPs

Interpreted as microsatellite expansion/contraction

Slide from Erik Garrison

example: 1000G Phase1 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

General recommendations on filtering

But appropriate filters depends on the dataset and project – no one size fits all!

TABLE 3 Key data filters to consider in the analysis of lcWGS data

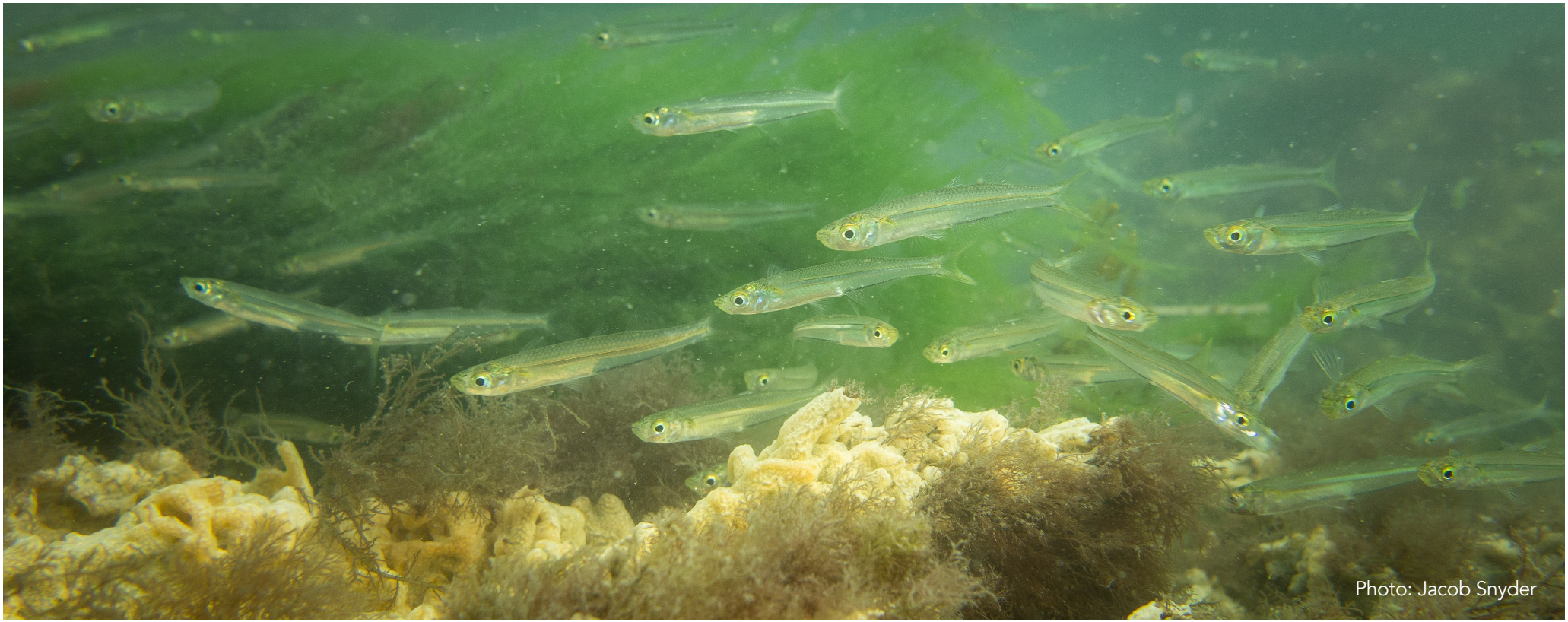
Category	Filter	Recommendation
General filters	Base quality	Base quality scores are factored into the calculation of genotype likelihoods, so if they accurately reflect the probability of sequencing error, bases with low scores also carry useful information. However, base quality scores are sometimes miscalibrated, so noise may be reduced if bases with scores below a threshold (e.g., 20) are either trimmed off prior to analysis or ignored. Alternatively, all base quality scores can be recalibrated based on estimated error profiles in the data (see Section 3.1).
	Mapping quality	Mapping quality is *not* considered in genotype likelihood estimation in currently available tools, so it is often advisable to remove low-confidence and/or nonuniquely mapped reads prior to analysis (e.g., reads with mapping quality <20). Filtering out reads that do not map in proper pairs should also further increase confidence in reads being mapped to the correct location, but could cause biases in regions with structural variation.
	Minimum depth and/or number of individuals	To avoid sites with low or confounding data support in downstream analysis, minimum depth and/or minimum number of individual filters can be used to exclude sites with much reduced sequencing coverage compared to the rest of the genome (e.g., regions with low unique mapping rates, such as repetitive sequences). Appropriate thresholds will vary between data sets, but could, for example, exclude sites with read data for <50% of individuals (globally or within each population), or with <0.8× average depth across individuals (after filtering on mapping quality)
	Maximum depth	Maximum depth filters are used to exclude sites with exceptionally high coverage (e.g., regions that are susceptible to dubious mapping, such as copy number variants). Common maximum depth thresholds could be one or two standard deviations above the median genome-wide depth.
	Duplicate reads	PCR and optical duplicates can give inflated impressions of how many unique molecules have been

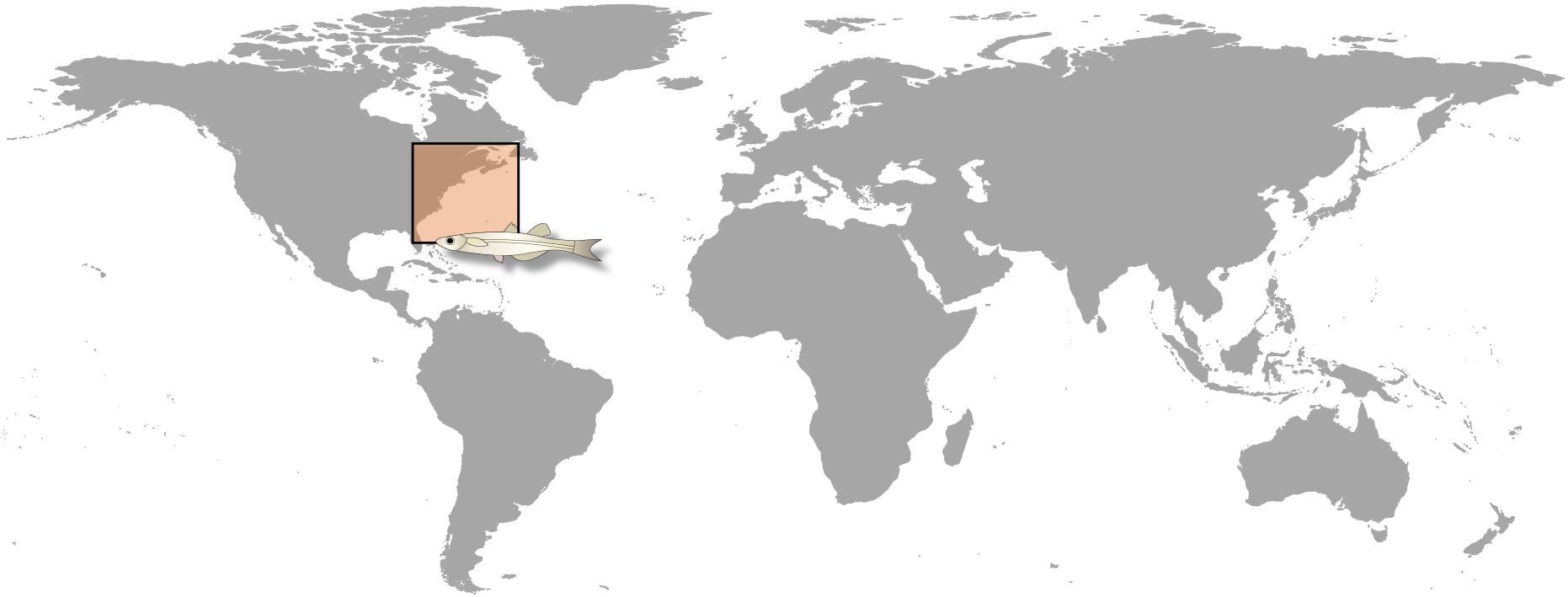
Lou et al. 2021. Mol. Ecol.

Now, your turn!

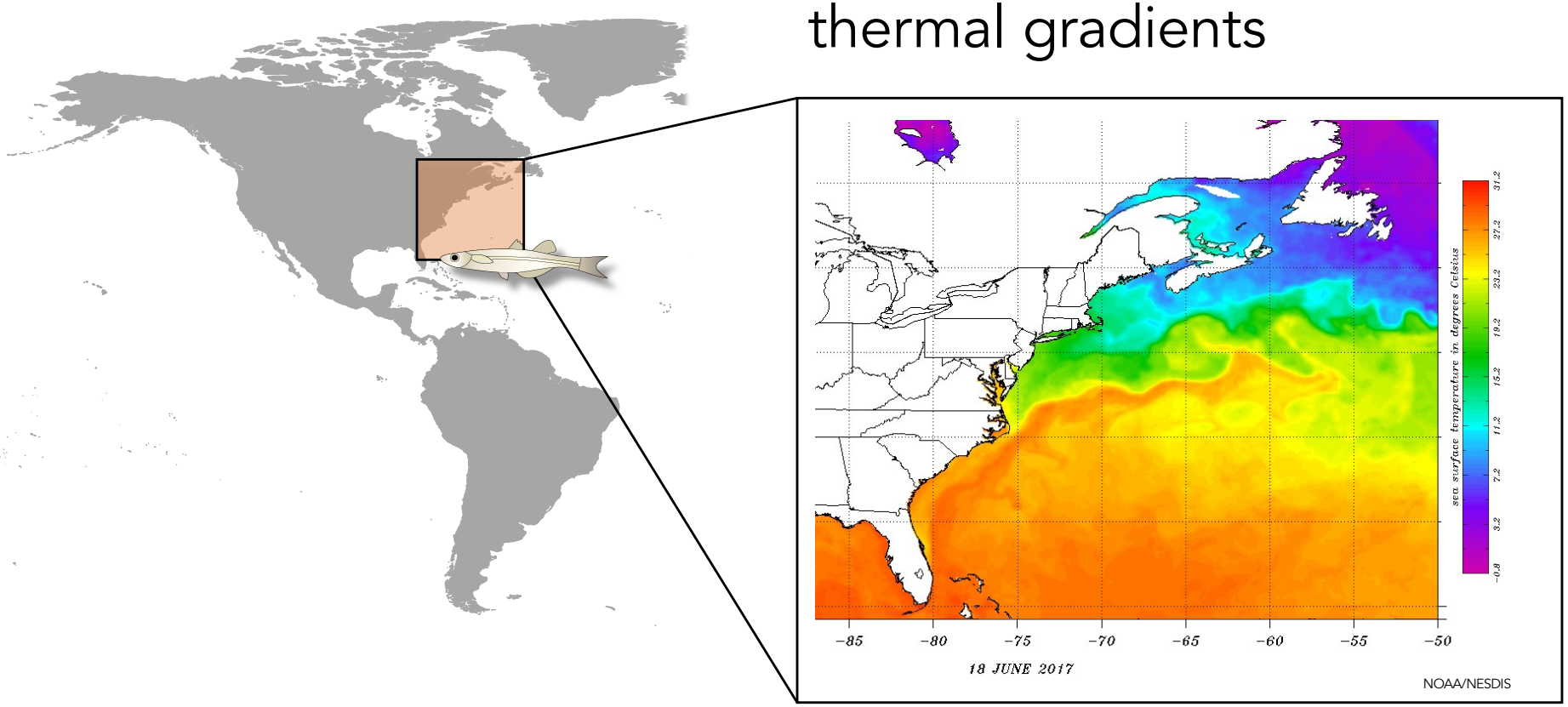
Example data for practicals

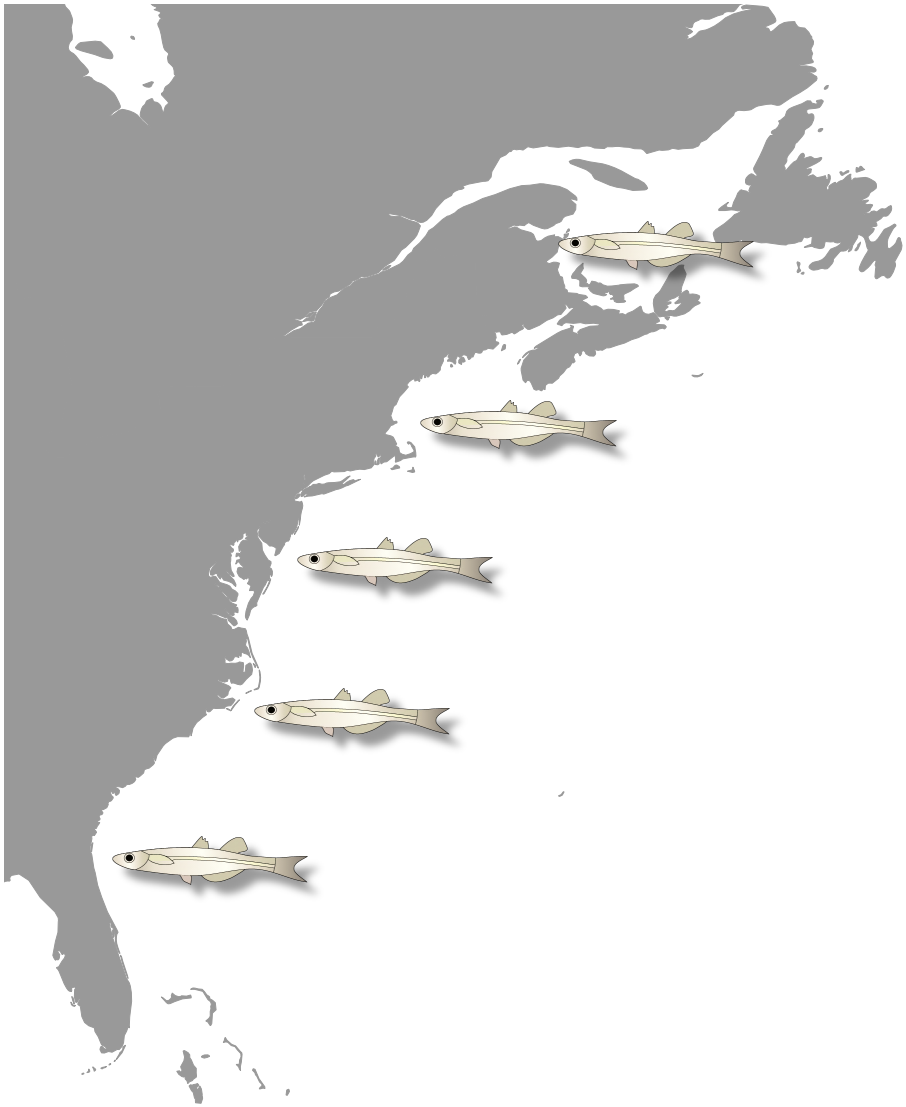
Atlantic silverside *Menidia menidia*



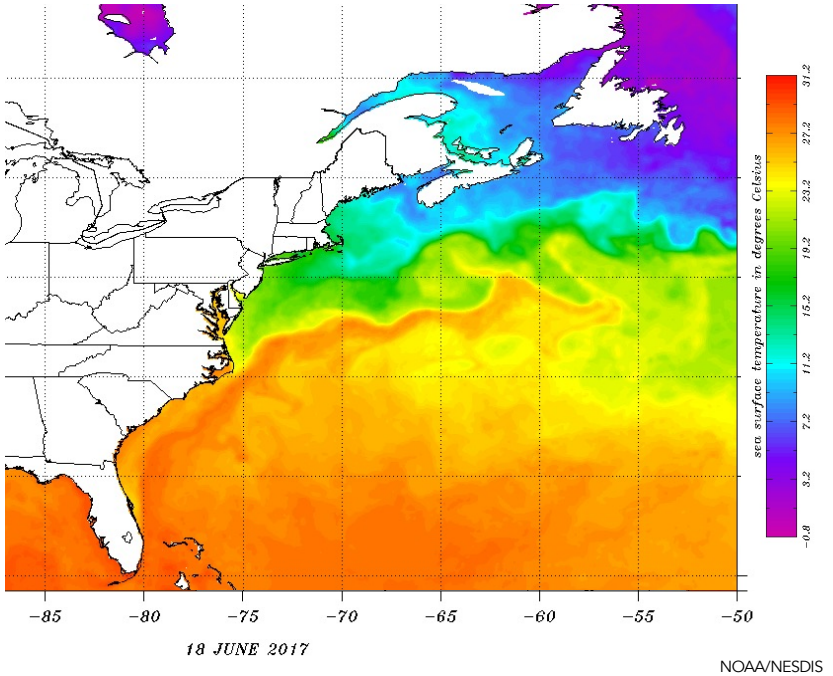


One of the world's steepest thermal gradients



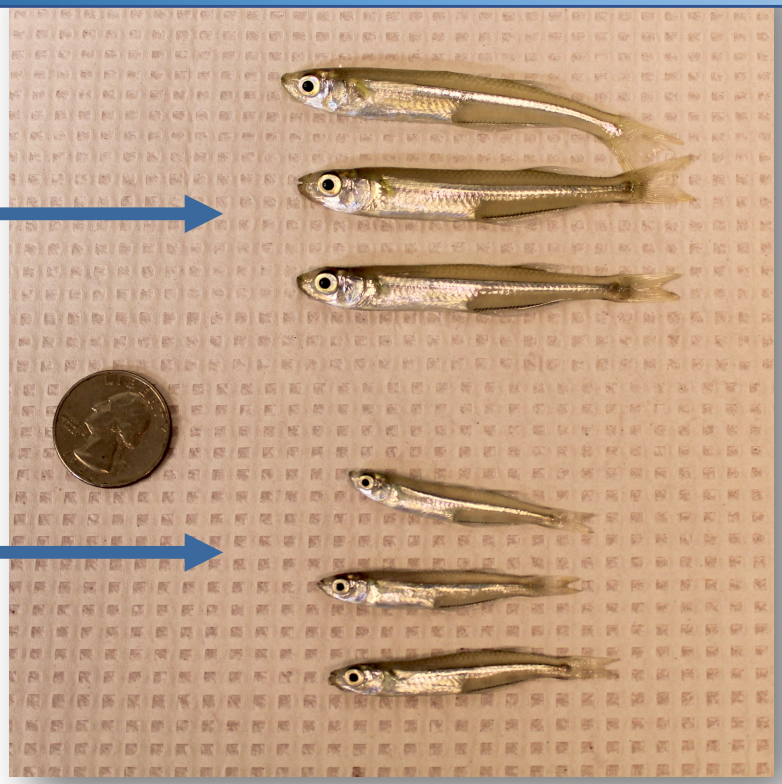


One of the world's steepest thermal gradients



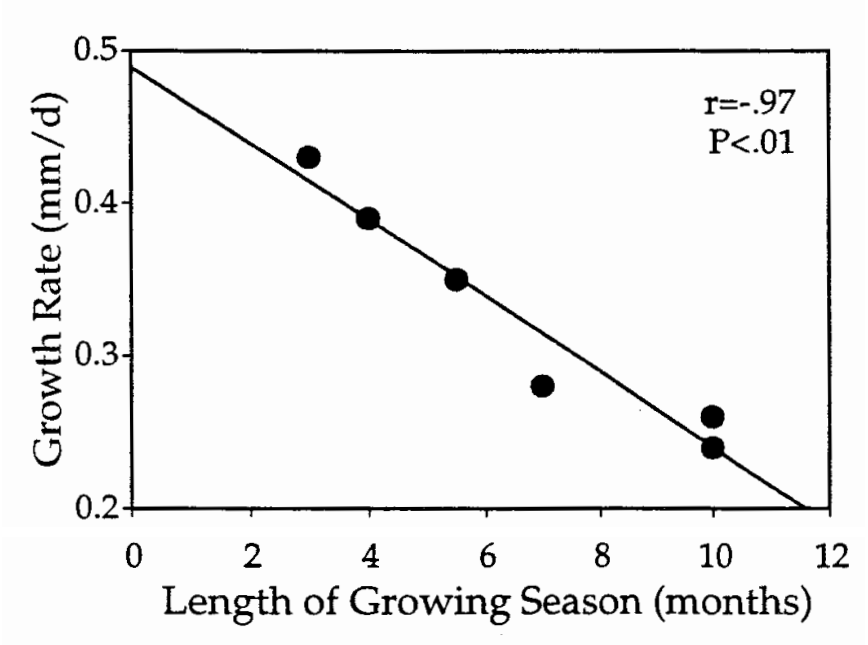
NOAA/NESDIS

Same age,
Common lab environment



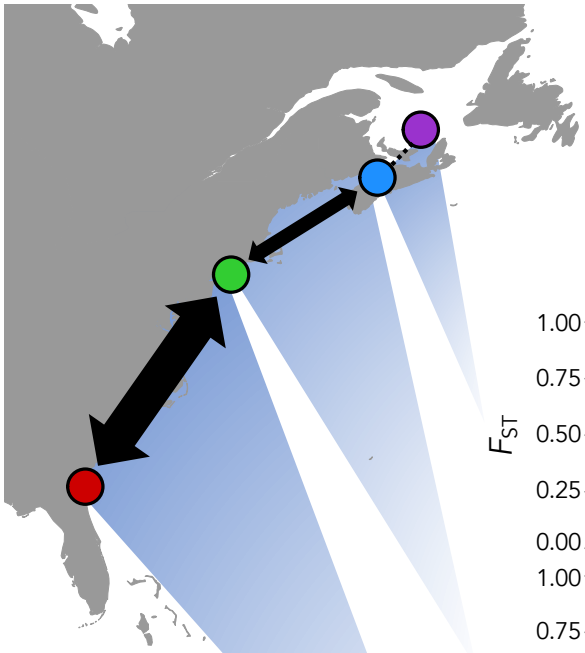
Growth capacity is tightly correlated with latitude

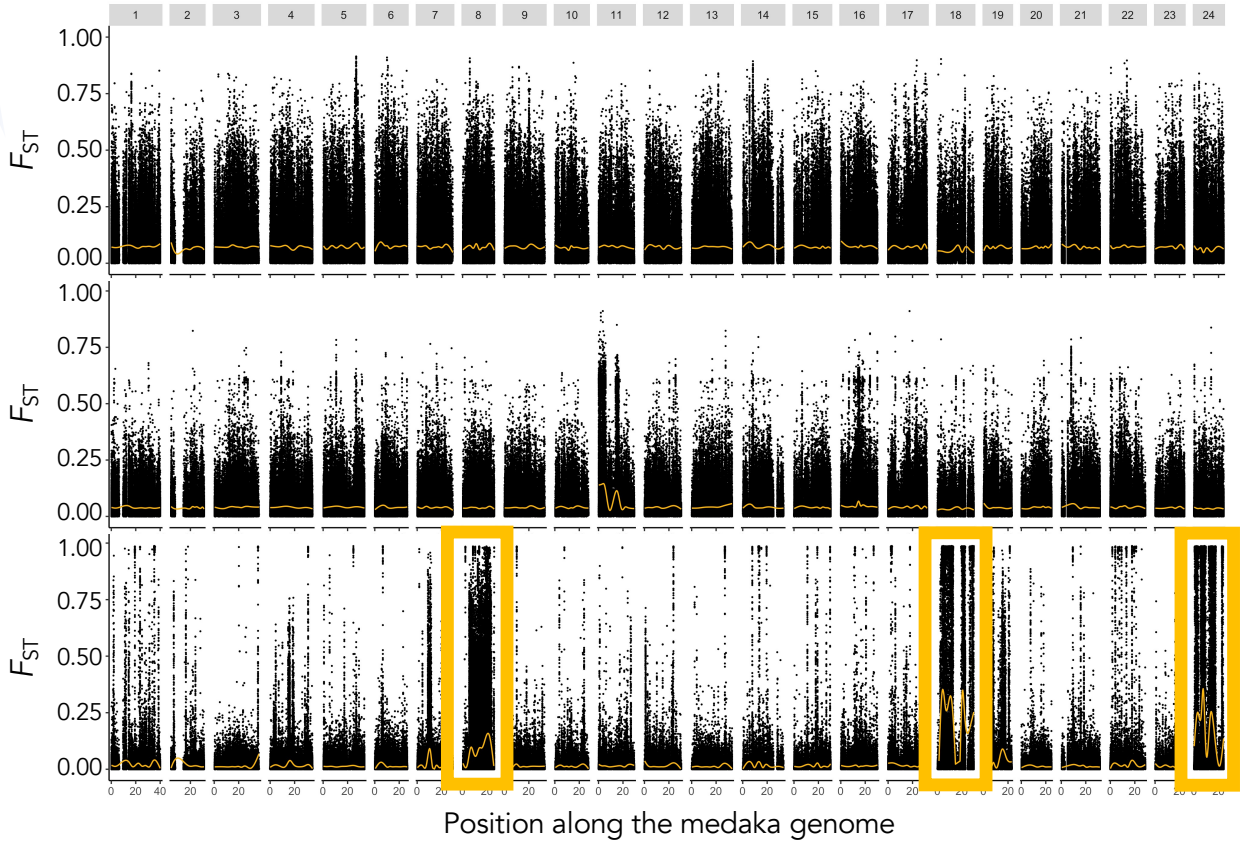
David Conover et al.



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

Heterogeneous gene flow and divergence patterns across latitudes

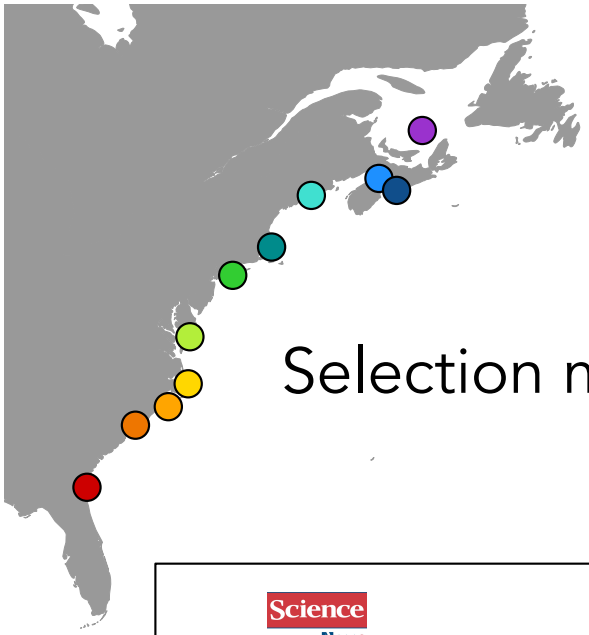


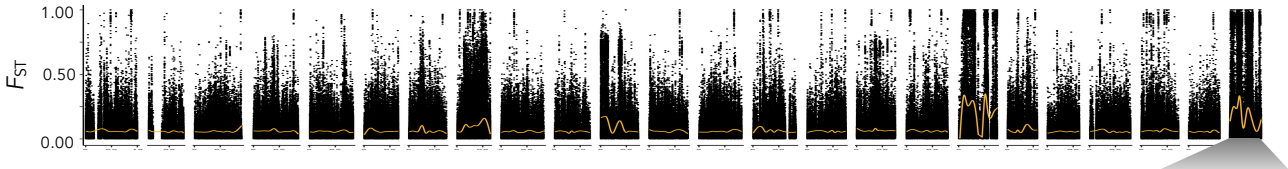


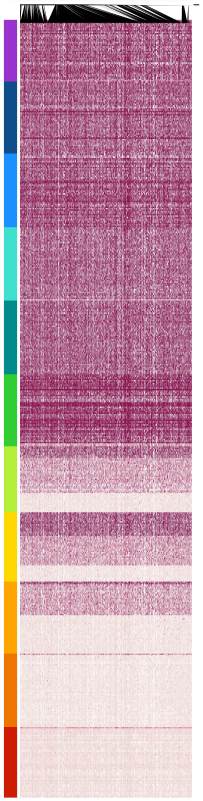
Median F_{ST} = 0.036
No fixed SNPs

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

on the basis of one of
cific rules: (i) in two po
than the 10th percentile
90%) were harvested (
two other populations,
90th percentile (the sma
ed (small-harvested); a
were controls in which 9
with respect to size (ra
vors ($n \approx 100$) were in
riod manipulations to sp
were collected and rea
ditions over multiple ge
our methods in the supp

Cross-generation tre
vested populations stro
pothesis (Fig. 1). Large
initially produced the
mean weight of fish bu
harvested populations
and then increased. By
selection, the biomass
weight of harvested in
harvested lines was n
large-harvested lines. Moreover, the spawning stock biomass differed even more. The mean weight of individual spawners (i.e., the survivors) in generation 4 was 1.05, 3.17, and 6.47 g in the large-, random-, and small-harvested