Reaching haplotopia: the promise and practice of short-read microhaplotypes in molecular ecology

Eric C. Anderson and the NMFS-SWFSC-MEGA Team

Durham University Ecology Seminar
14 NOV 2016

Collaborators and Acknowledgments

NOAA/SWFSC/UCSC

- Carlos Garza
- Anthony Clemento
- Thomas Ng (UCSC grad student)
- Diana Baetscher (UCSC grad student)

UCSC Rockfish Project:

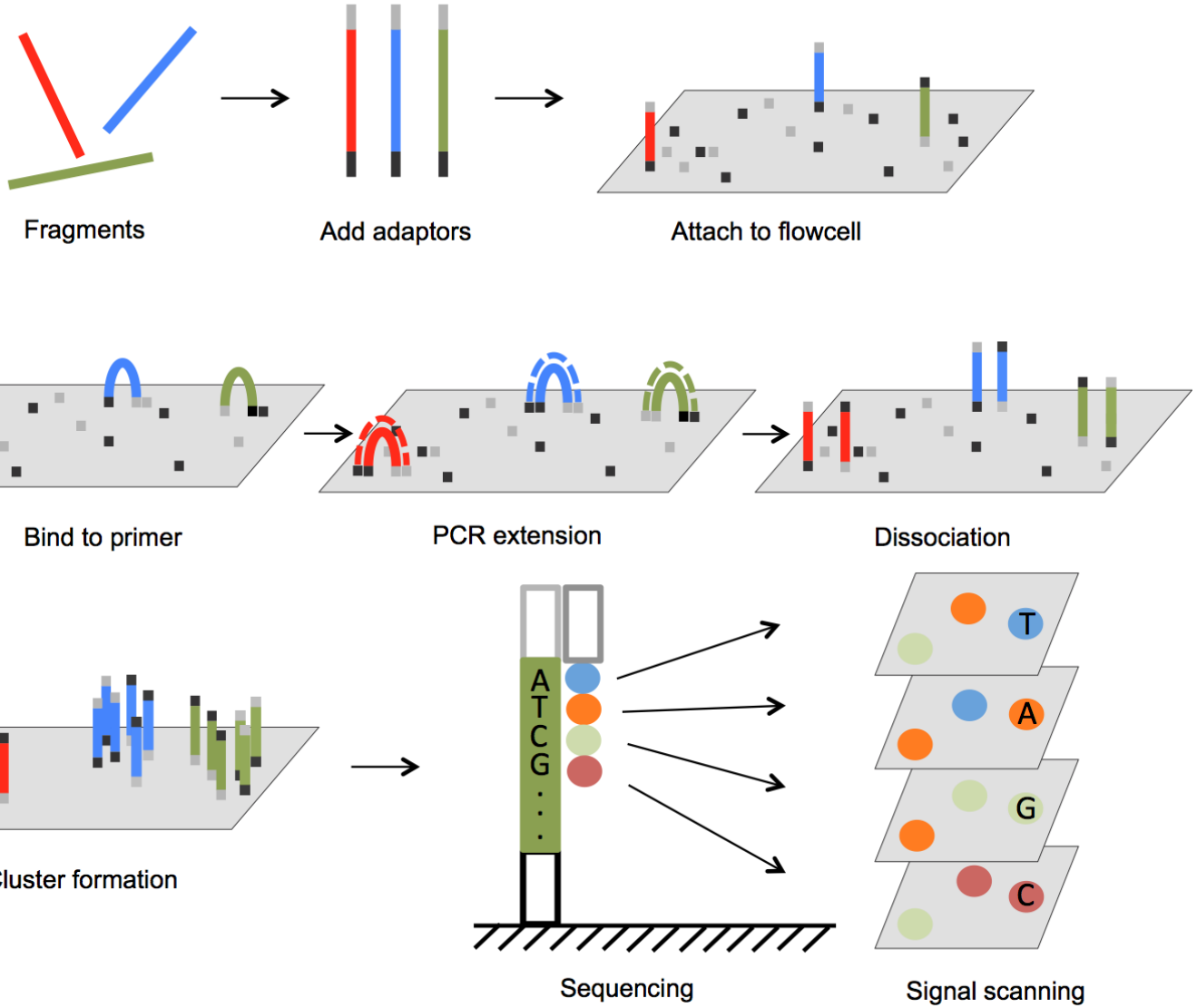
- Mark Carr
- Chris Edwards
- Dan Malone
- Emily Saarman
- Patrick Drake
- Anna Lowe

Overview

- Next Generation Sequencing Methods
  - sequencing, assembly, alignment
  - genome representation reduction
  - amplicon sequencing
- Microhaplotypes
  - available by reanalysis of most data sets
  - advantages
- Example problem: Dispersal of larval rockfish
  - parentage and sibling inference in a multispecies context
- Tools from our lab:
  - CKMRsim
  - haPLOType

Next Generation Sequencing – I

Illumina Sequencing By Synthesis



http://www.intechopen.com

Next Generation Sequencing – II

de novo "genome" assembly

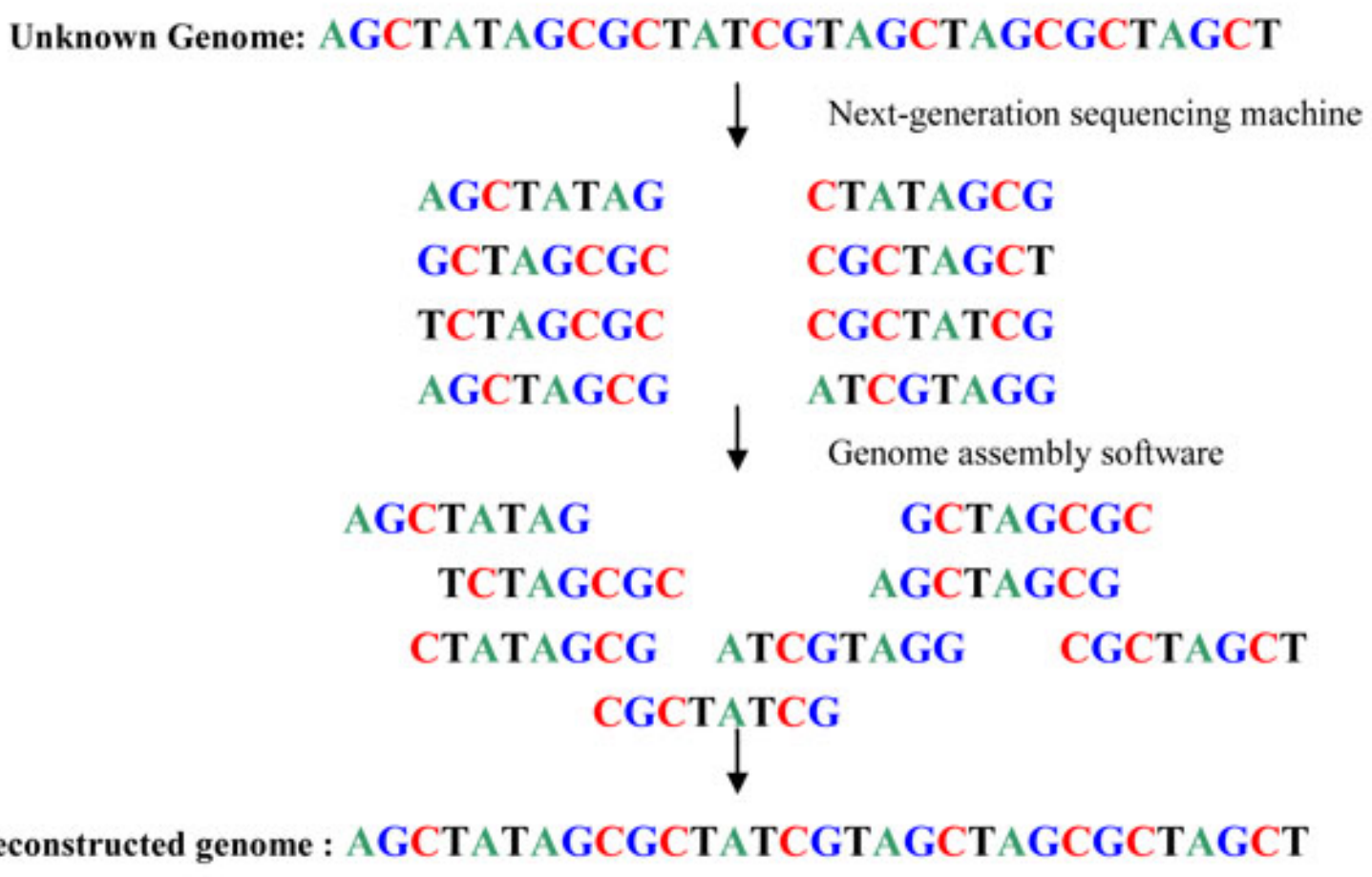
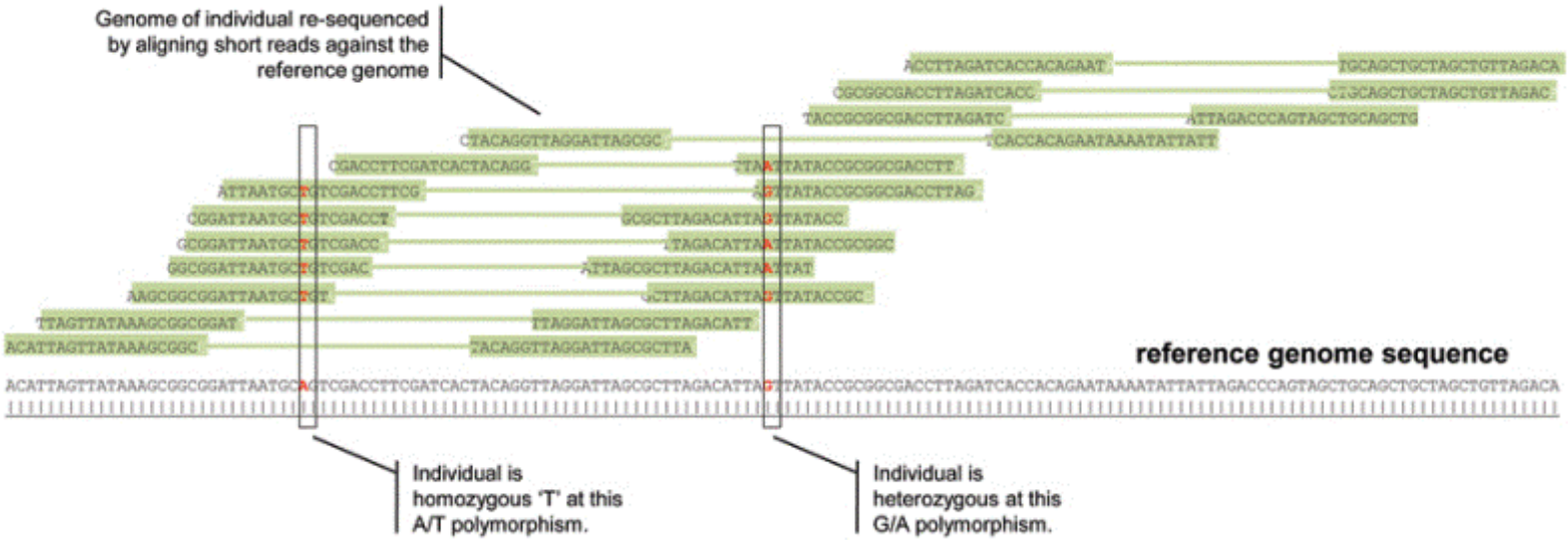


Figure 1. Workflow of discovering the genome of a species

http://www.cs.hku.hk

Next Generation Sequencing – III

Alignment of sequencing reads to “genome” allows identification of polymorphisms and genotyping



http://www.historyofnimr.org.uk

Sequencing Capacities

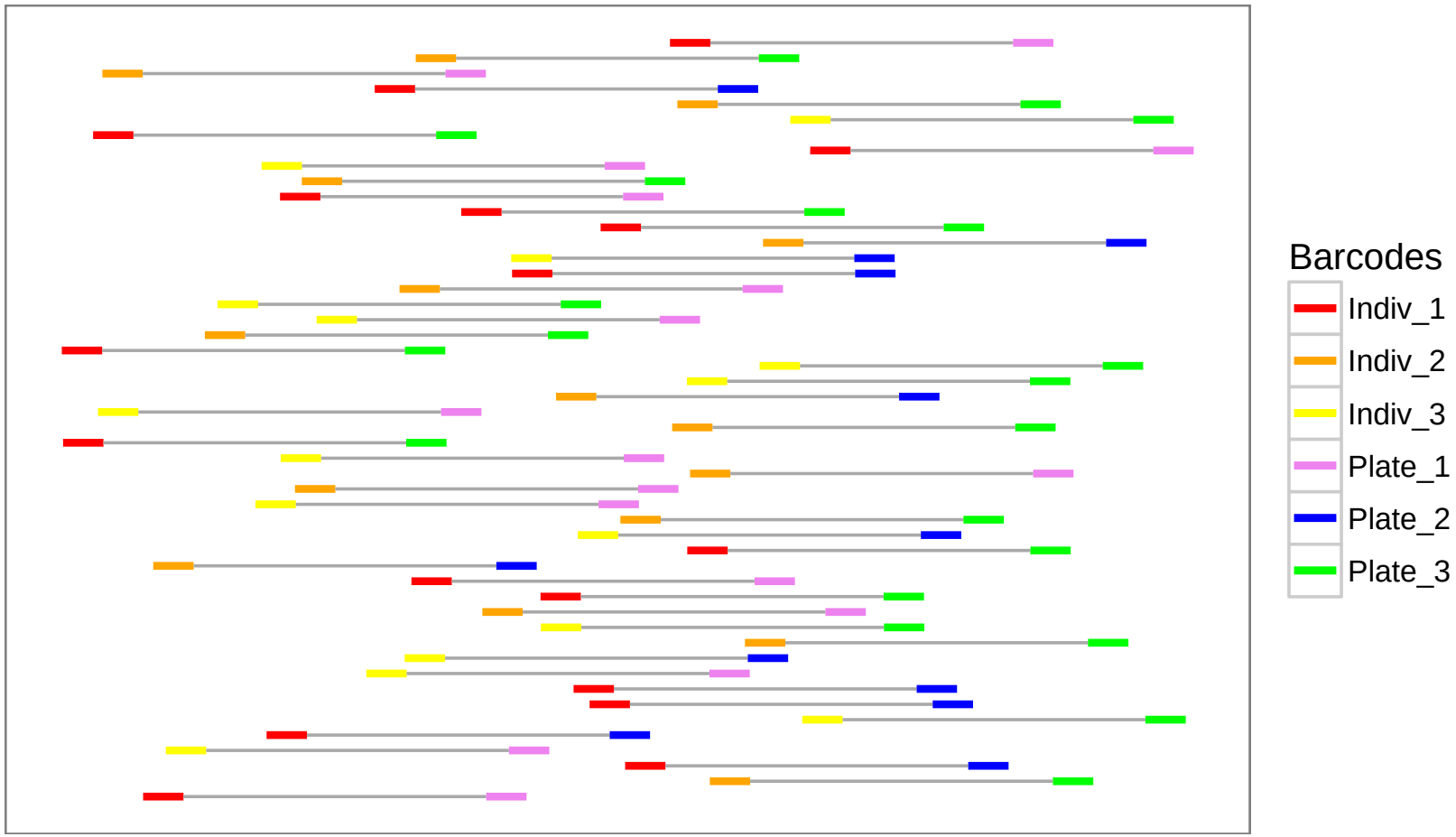
Continually increasing. Examples from Illumina:

- Small Benchtop MiSeq sequencer:
  - 25 million reads/run ($\approx$ \$1,000 reagent cost)
  - up to 2 x 300 bp
- Large Core facility HiSeq sequencer:
  - 500 million reads per lane ($\approx$ \$3,400)
  - up to 2 x 150 bp

- *Many* individuals can be sequenced together via combinatorial barcoding.

Combinatorial Barcodes

- Imagine that you prepared individuals in batches (plates) of three at time.
- Individual barcode on one end; plate barcode on the other.



Fundamental Tradeoff

- Accurate determination of the sequence *at a genomic position* and *in an individual* requires a sufficient number of reads sequencing that position in that individual.
- So, either:
  - One (or a few) individuals across large swaths of genome
  - Many individuals across a reduced portion of the genome
- Methods for *reduced representation* sequencing:
  - RNAseq (sequence just the transcriptome)
  - RADseq (Restriction Associated DNA sequencing)
  - GBS (Genotyping by Sequencing)
  - Capture/Bait methods (MyBaits, RAPTURE, etc.)
  - Amplicon Sequencing (GTseq)

GTseq amplicon sequencing

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) doi: 10.1111/1755-0998.12357

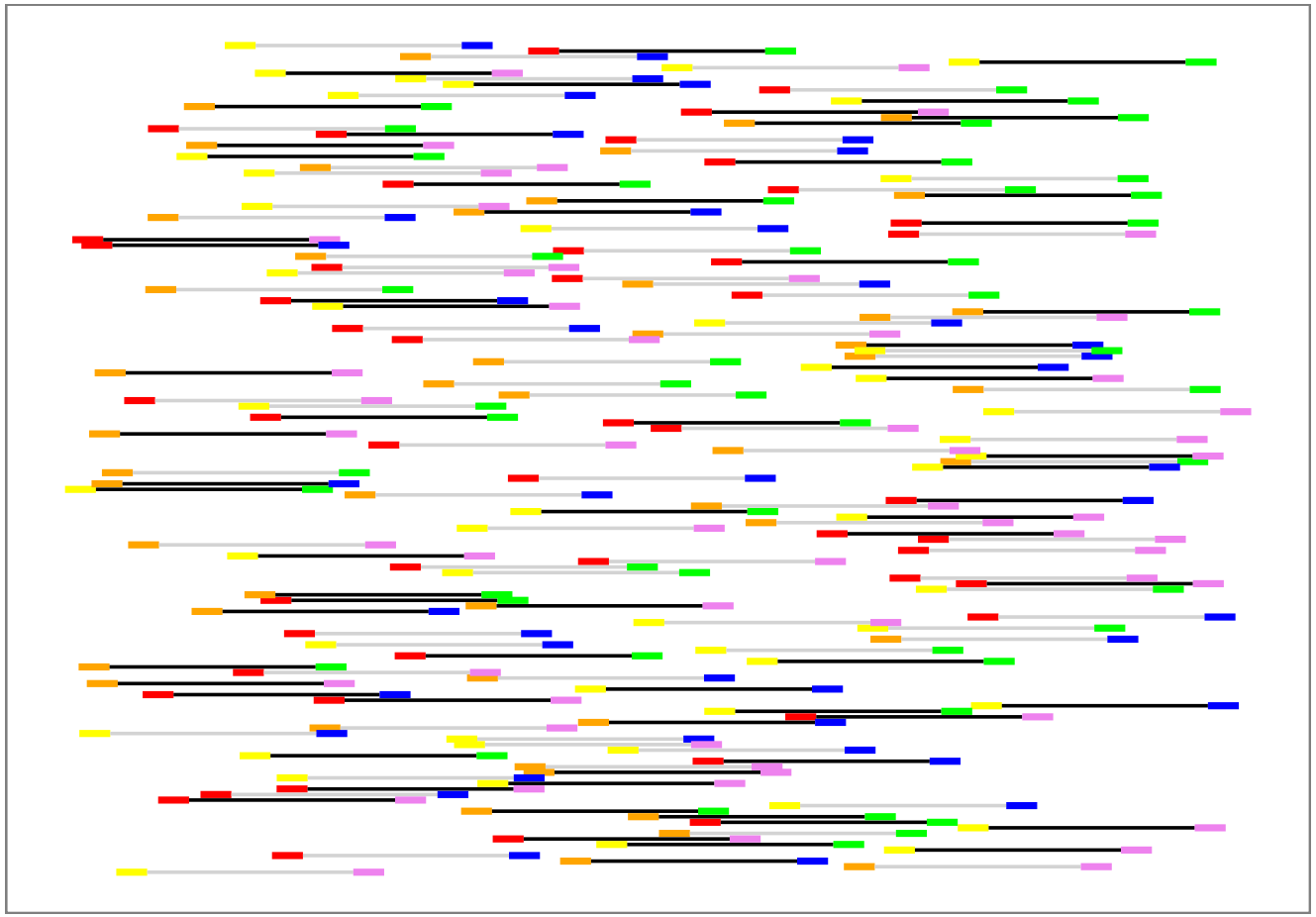
Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing

NATHAN R. CAMPBELL STEPHANIE A. HARMON and SHAWN R. NARUM
Columbia River Inter-Tribal Fish Commission, 3059F National Fish Hatchery Road, Hagerman, ID 83332,USA

- Multiplexed PCR primers amplify regions of interest.
- Amplicon sequencing of 200–500 regions in 500–2,000 individuals
- $\approx$ \$7/individual
- Converted Fluidigm SNP assays. Tens of 1,000s of salmon each year.

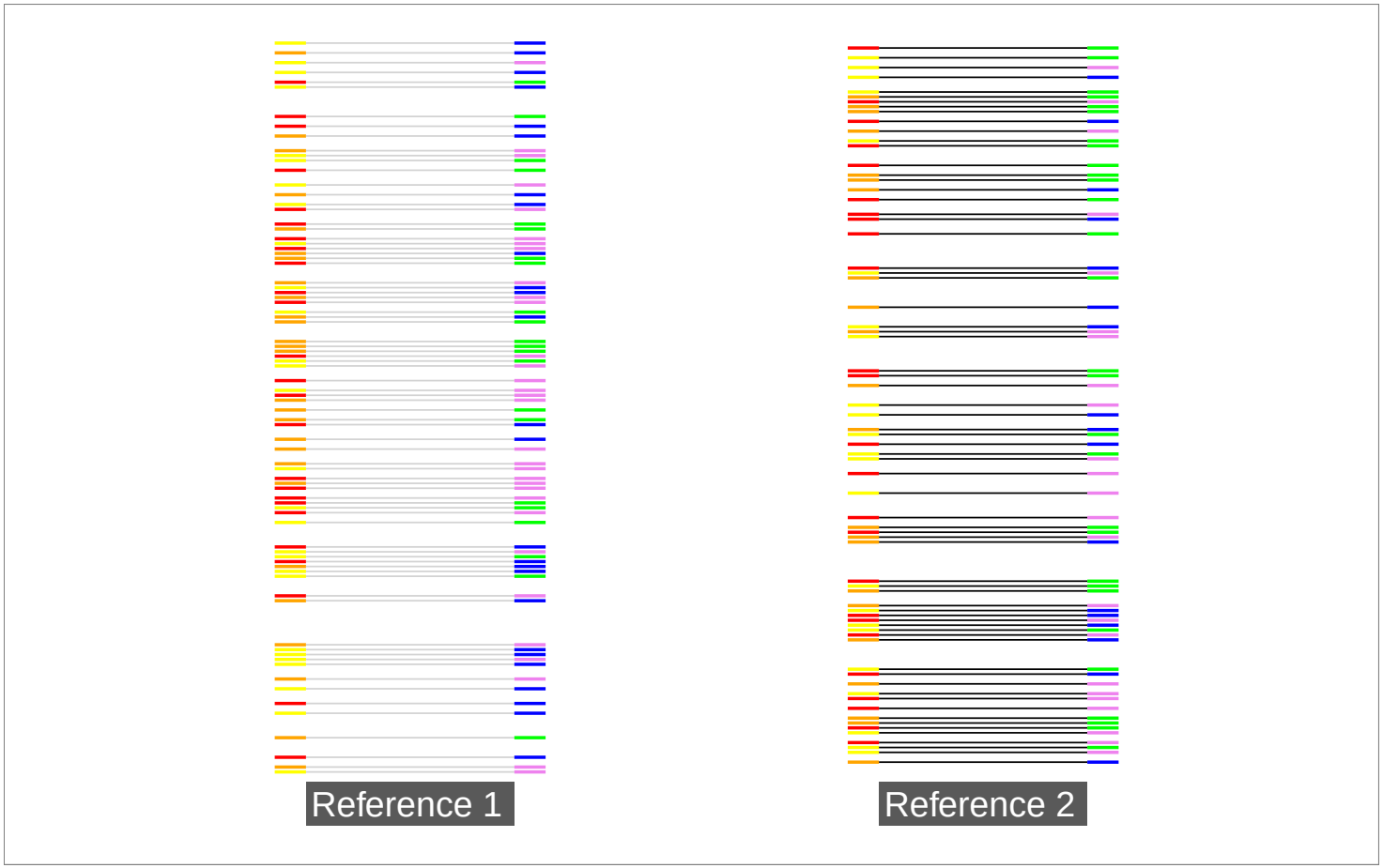
GTseq – I (coming off the sequencer. . .)

2 Amplicons; 3 individuals on each of 3 plates (9 individuals total)



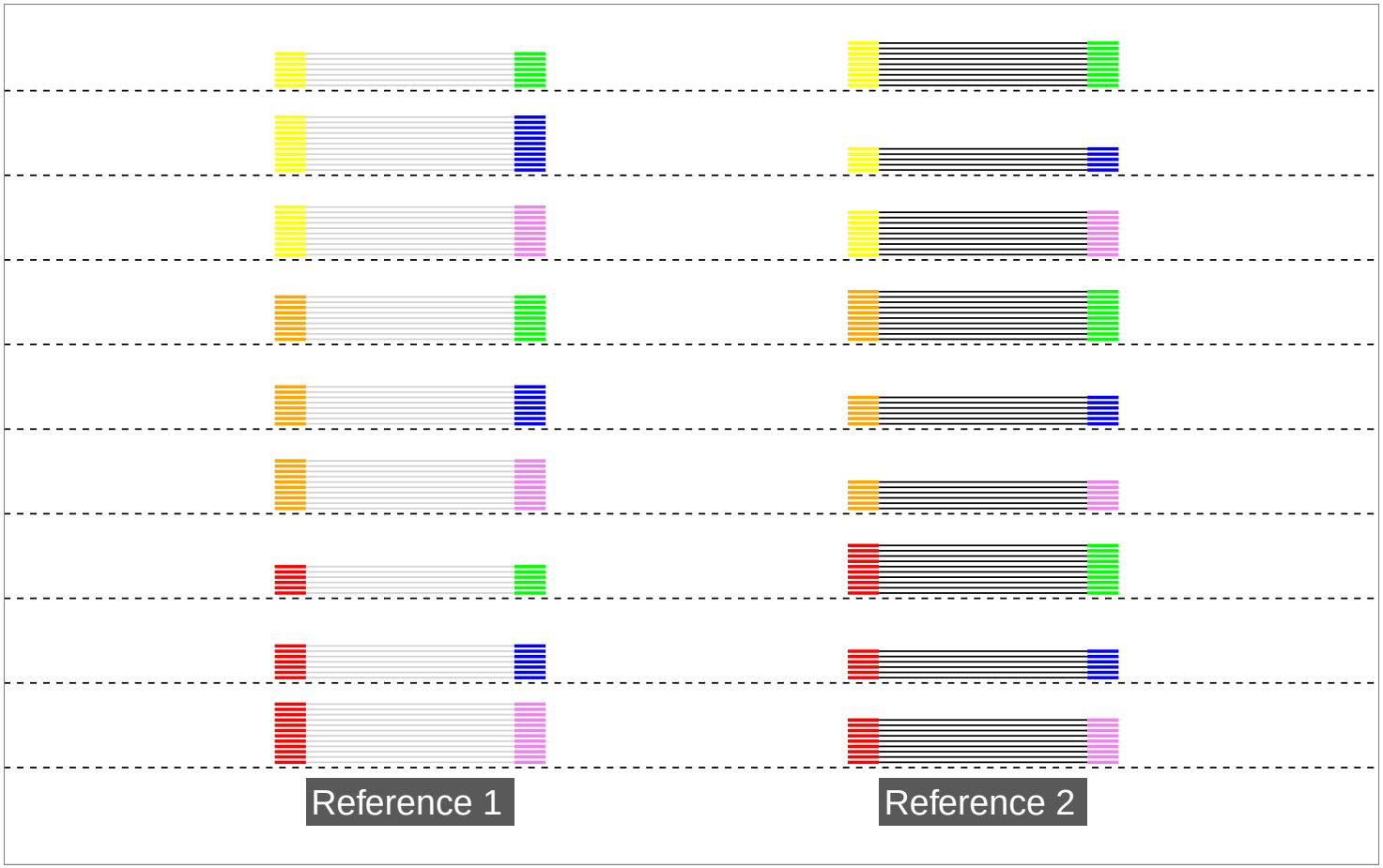
GTseq – II (aligned amplicons)

Amplicon sequences are known. Alignment *in silico* is straightforward.



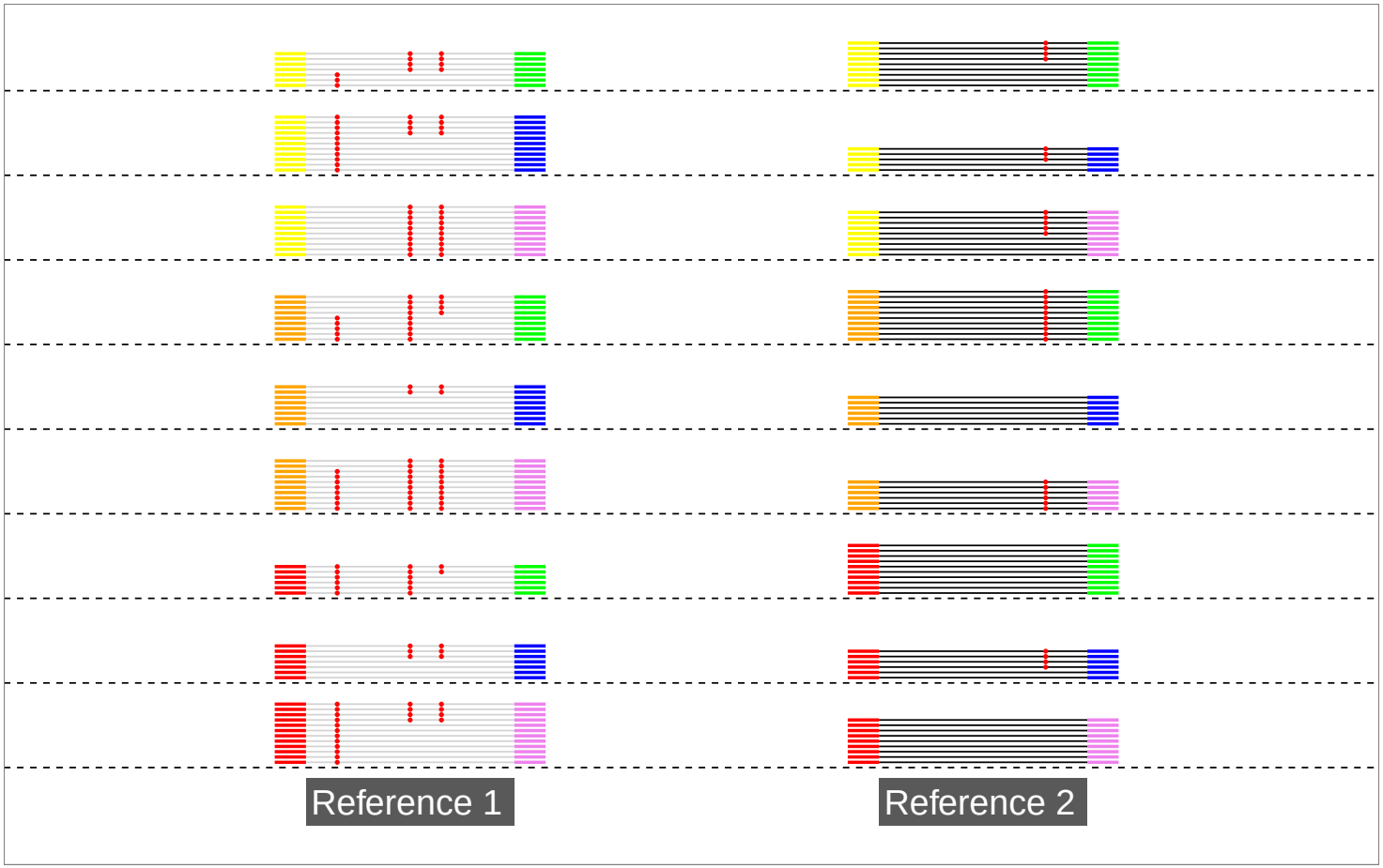
GTseq – III ("demultiplexing")

Identify individual origin of reads via combinatorial barcodes



GTseq – IV (identify SNPs in sequence)

Variants are easy to identify



Multiple SNPs in amplicons/sequences. . .

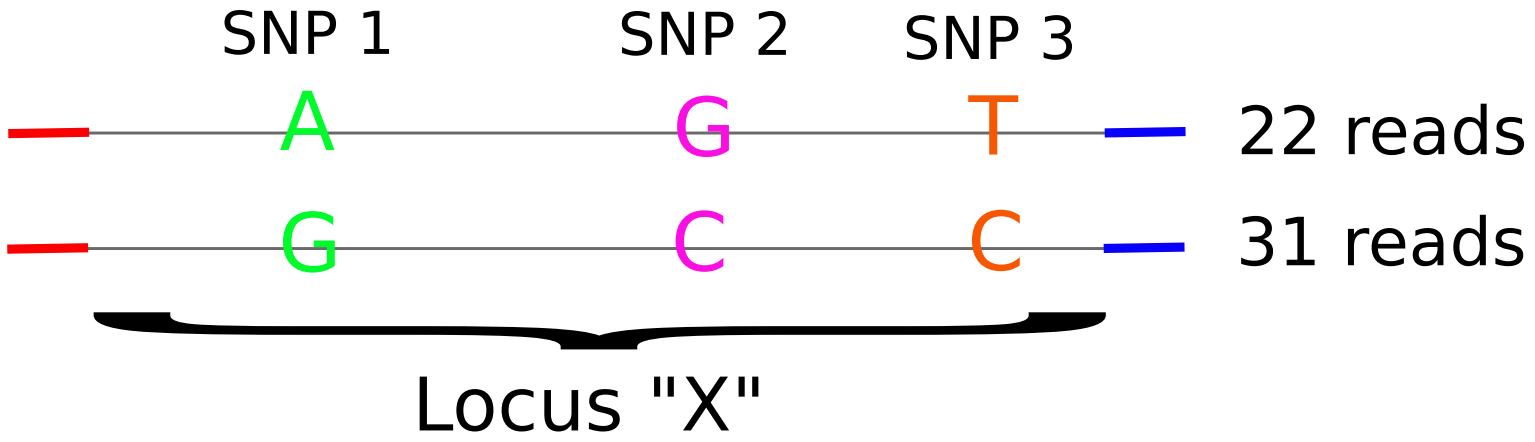
. . . are almost universally ignored

- Multiple SNPs in each amplicon/RAD-locus are typically scored
- But then either:
  1. Only a single SNP from each amplicon/locus is used
  2. OR, all SNPs are treated as unlinked

  Depending on the analyses, the result is either a lack of power or (potentially) incorrect inference.

Phase of SNPs on reads is almost universally ignored

If this is observed:



It will often be treated as, either:

SNP 1: AG

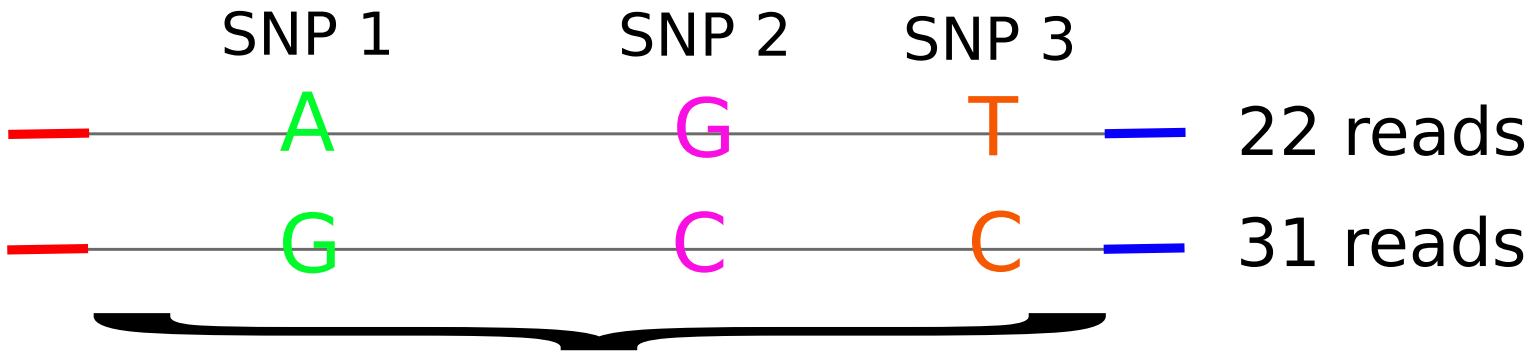
SNP 2: CG

SNP 3: CT

OR

Locus "X": CG

If this is observed:



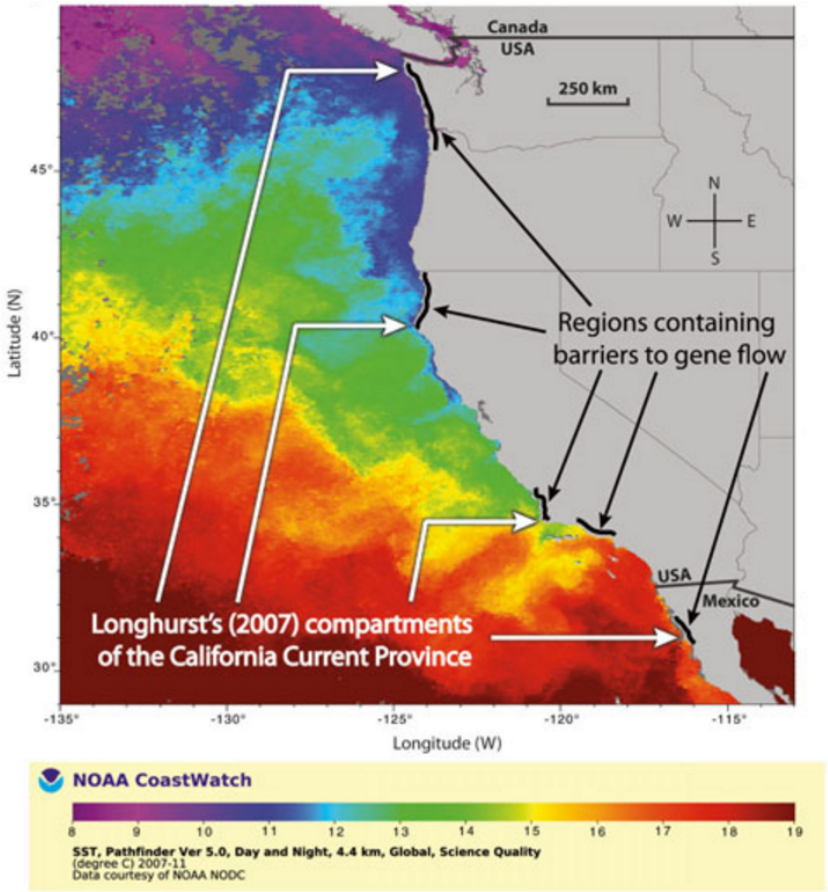
It might be beneficial to treat it as:

Locus "X": $\frac{\text{AGT}}{\text{GCC}}$

With each haplotypic combination being recorded as an allele.

- Multiallelic loci
  - More power for relationship inference / pedigree reconstruction
- Need not discard SNPs from certain loci
  - Retain low-frequency variants. Useful for population structure in recently diverged populations.
- Amplicons typically cross-amplify between closely-related species
  - Unlike single SNP assays, the microhaplotype data collection method, unmodified, can yield useful data for non-target species.

Marine Population Structure

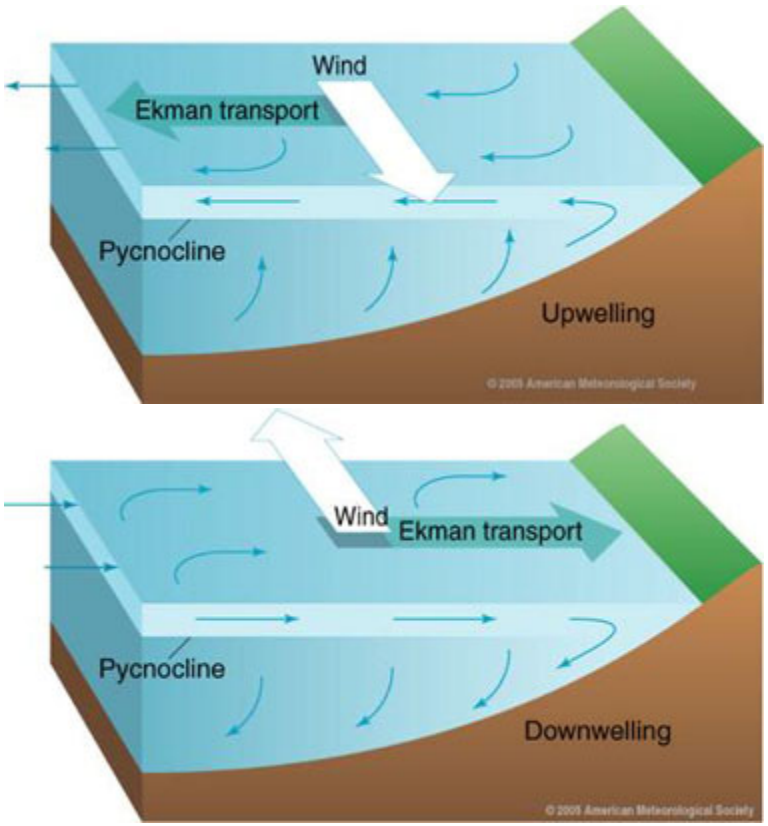


- California Current
  - Upwelling; nutrients; long larval duration
- Barriers to gene flow
  - Suspected from patterns of genetic variation
  - Difficult to estimate actual migration/dispersal rates
- Considerable interest in larval dispersal/retention
  - Design of MPAs, etc.

Hyde and Vetter (2009) CJFAS

Larval Dispersal off California

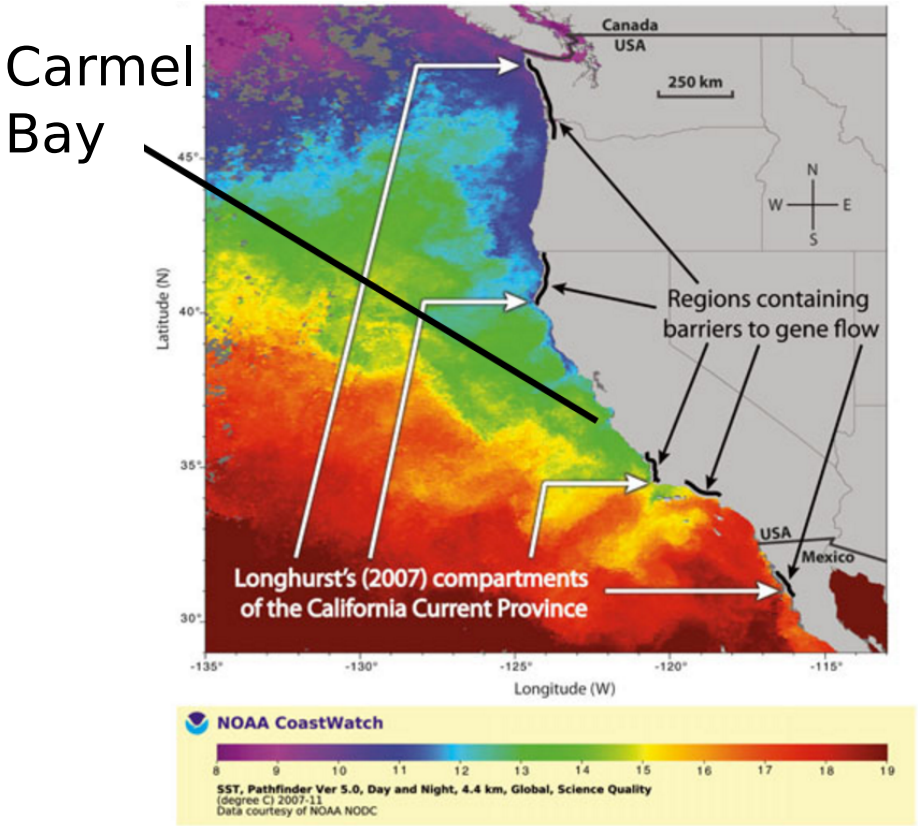
Heavily influenced by Ekman transport



Amer. Meterol. Soc.

- Typical Pattern:
  - Winds blowing south down coast
  - Water transport offshore
- Periods of relaxation:
  - Winds blowing to the north
  - Water transport toward shore
- Larval recruit origins
  - Coarse current models suggest most come from far away
  - BUT, Fine-scale current irregularities might allow local larval retention

Question:
What is the degree of larval self-recruitment around Carmel Bay?



- Experimental Approach with Kelp Rockfish
  1. Non-lethally sample a boatload ($\approx 5,000$) of adult fish
  2. Genotype them
  3. Collect tissues from $\approx 5,000$ larval recruits
  4. Genotype them
  5. Find parent-offspring pairs

Hyde and Vetter (2009) CJFAS

Kelp Rockfish (*Sebastes atrovirens*)

photo: Wikimedia

Adults stationary. Sampled by biopsy spear or hook-and-line.
Very active volunteer “recreational-sampling” component.

SMURF traps

Standard Monitoring Unit for the Recruitment of Reef Fishes

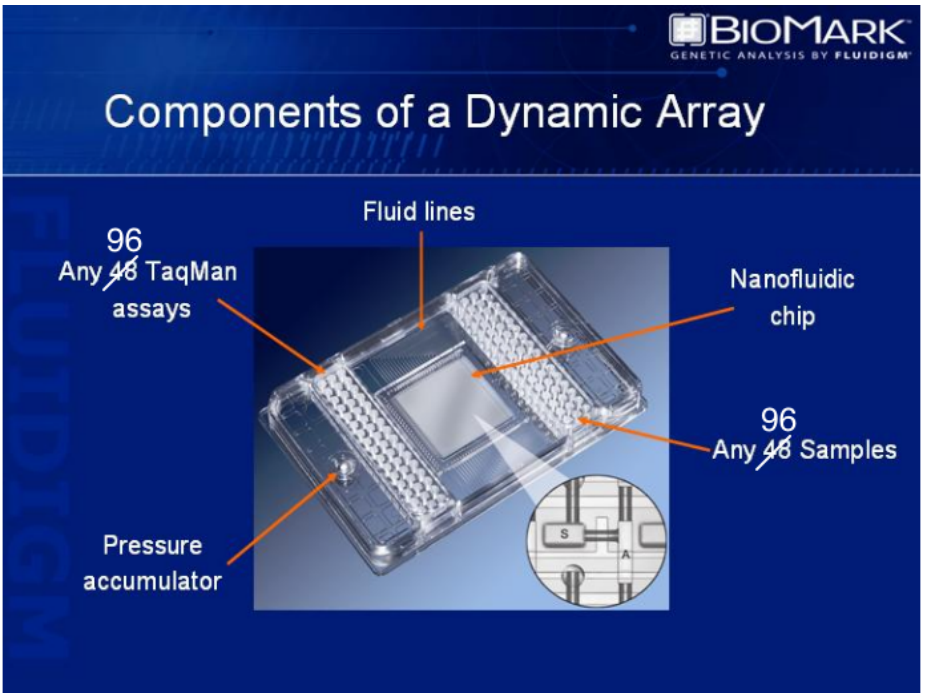
Carr-Raimondi lab photo

SMURF traps
Standardized protocol for juvenile collection

Carr-Raimondi lab photo

Originally-proposed approach

Large-scale parentage inference with microfluidic SNP assays



With 96.96 Arrays and only one controller and thermal cycler, can genotype almost 300 fish per day w/96 SNPs.

Cost: $<$\$15/fish

Microfluidic chips, a known quantity

- However, 96 SNPs are not enough for single parent assignments,
- and species ID in juveniles is unreliable

Visual Species ID of juveniles is difficult

Kelp, Gopher, Black-and-yellow indistinguishable at small size

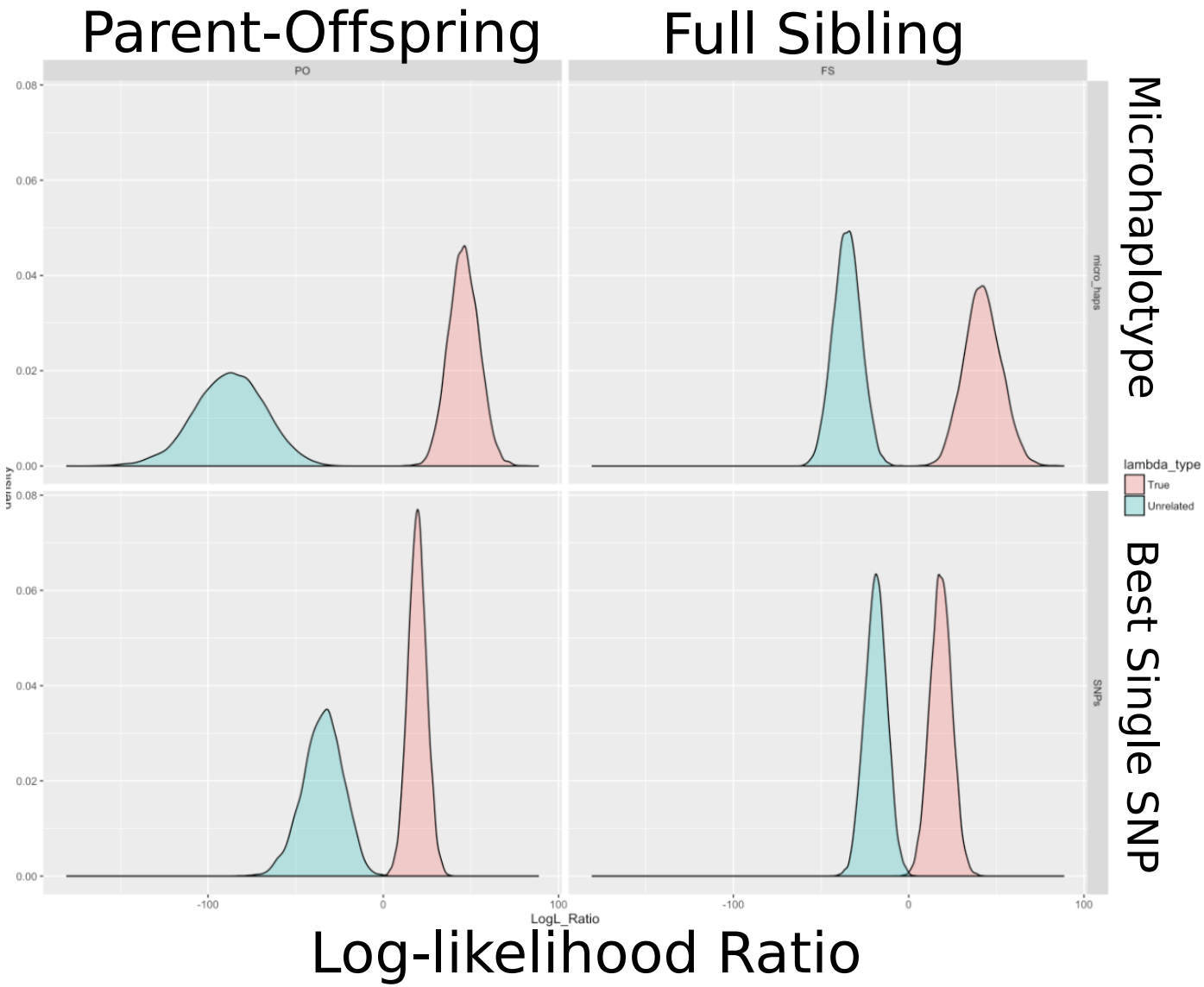
- Up to 40% of juvenile samples might not be kelp rockfish
- SNP assays in non-target species tend to be monomorphic, and hence useless for any inference.
- It hurts to contemplate throwing away that much genotyping effort.

Developing GTseq Amplicons for Rockfish

- ddRAD seq on 16 individuals
- $\approx 4,000$ genome fragments sequenced
- 200 developed into amplicons, and tested
- 165 amplicons retained
- Amplified and sequenced from 240 individuals.
- Allele/Haplotype frequencies estimated
  - 825 alleles (average of 5 haplotypes per locus)
- Power for Parentage and Full sibling inference computed

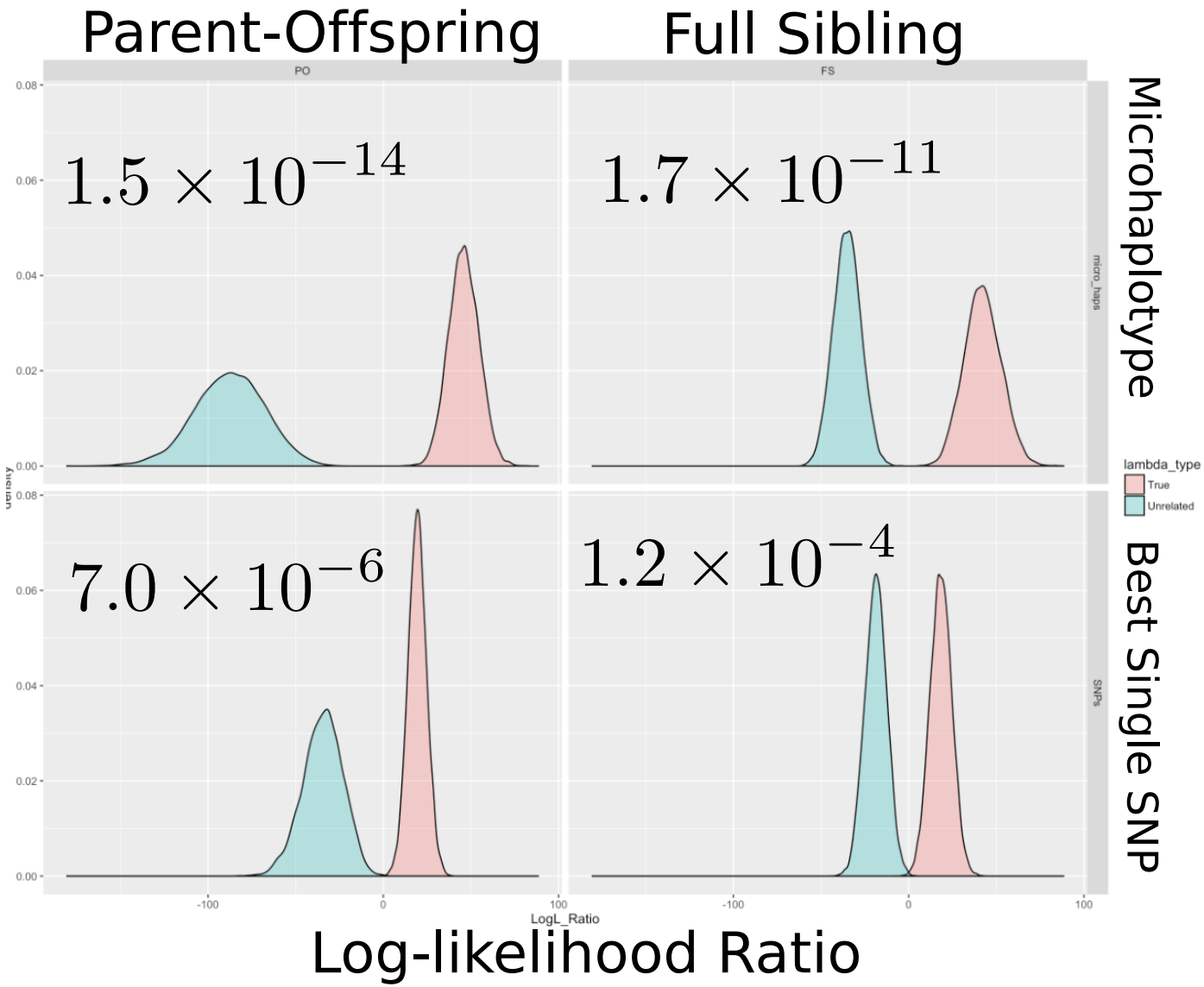
Log-likelihood ratio distribution

For Parent-offspring and full-sib pairs vs Unrelated



Log-likelihood ratio distribution

False Positive Rates for Unrelated individuals at False Negative Rate = 1%



Outstanding for Relationship Inference

- A false positive rate of 1.5×10^{-14} means you could compare 1 million candidate parents to 1 million unrelated candidate offspring and not expect any errors.
- Using just a single SNP from each locus you would expect 7 errors when comparing 1,000 candidate parents to 1,000 unrelated candidate offspring.
- Genotyping costs dropping toward $<$£5 per sample.
- Very exciting for "close-kin mark-recapture" *Statistical Science* 31:259–274. (2016)

Statistical Science
2016, Vol. 31, No. 2, 259–274
DOI: 10.1214/16-STS552
In the Public Domain

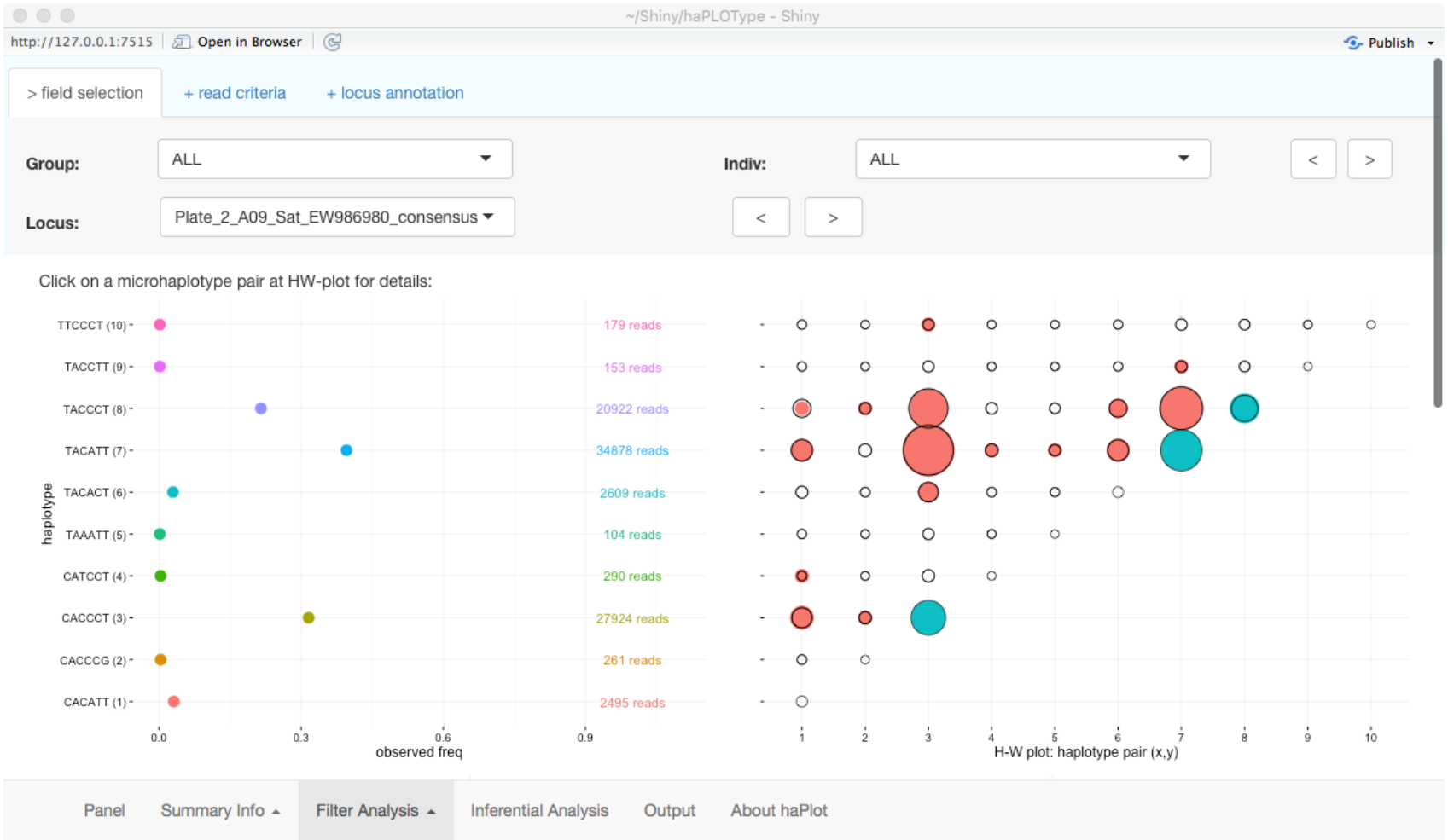
Close-Kin Mark-Recapture

Mark V. Bravington, Hans J. Skaug and Eric C. Anderson

`haplot` `(microhaplot, haplotr, haPLOType?)`

- an R package by Thomas Ng
- Appoint SNPs in haplotypes with a VCF file
- Extract haplotypes from SAM files
- Filter on Read Depth / Allelic Balance ratio
- Partially completed Bayesian inference of haplotypes
- Full Shiny App for Visualization.

`https://github.com/ngthomas/haplot`


~/Shiny/haPLOType - Shiny
http://127.0.0.1:7515
Open in Browser
Publish
> field selection
+ read criteria
+ locus annotation
Group:
ALL
Indiv:
ALL
Locus:
Plate_2_A09_Sat_EW986980_consensus
Click on a microhaplotype pair at HW-plot for details:
haplotype
TTCCCT (10)
TACCTT (9)
TACCCT (8)
TACATT (7)
TACACT (6)
TAAATT (5)
CATCCT (4)
CACCCT (3)
CACCCG (2)
CACATT (1)
179 reads
153 reads
20922 reads
34878 reads
2609 reads
104 reads
290 reads
27924 reads
261 reads
2495 reads
observed freq
H-W plot: haplotype pair (x,y)
Panel
Summary Info
Filter Analysis
Inferential Analysis
Output
About haPlot

`CKMRsim`

- an R package by Eric C. Anderson
- Estimate false positive rates for pairwise relationship inference
- Built to accommodate general genotyping error models
- Importance sampling algorithm to estimate very small probabilities
- Integration with Mendel to simulate linked markers
- Key parts written in C++ for speed.

`https://github.com/eriqande/CKMRsim`

Conclusions

- NGS technologies provide opportunities for low-cost high-throughput genotyping of many individuals
- Typical pipelines for GTseq and amplicon sequencing don't extract as much information as possible.
- Preserving the haplotypic phase of SNPs on reads yields multiallelic markers...
- ...that provide substantially more power for relationship inference (at least in high-diversity species)
- Tools for handling these microhaplotype data are available from our lab

This talk available at:
`https://github.com/eriqande/TALKS-microhap`