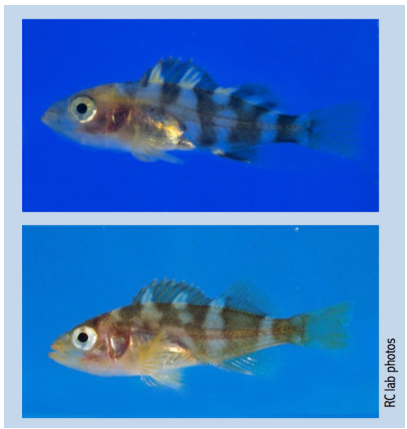
Next-generation genetic technologies for studying marine larval dispersal: Microhaplotypes and kelp rockfish

Eric C. Anderson, NMFS-SWFSC-MEGA Team



USC Marine and Environmental Biology Seminar
21 February 2017

Overview

1. Larval dispersal and genetic patterns
2. Kelp-rockfish recruitment project
3. Next-generation sequencing and microhaplotypes
4. Confirmation of our data quality
5. Preliminary relationship inference work

Collaborators and Acknowledgments

NOAA/SWFSC/UCSC

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

UCSC Rockfish Project:

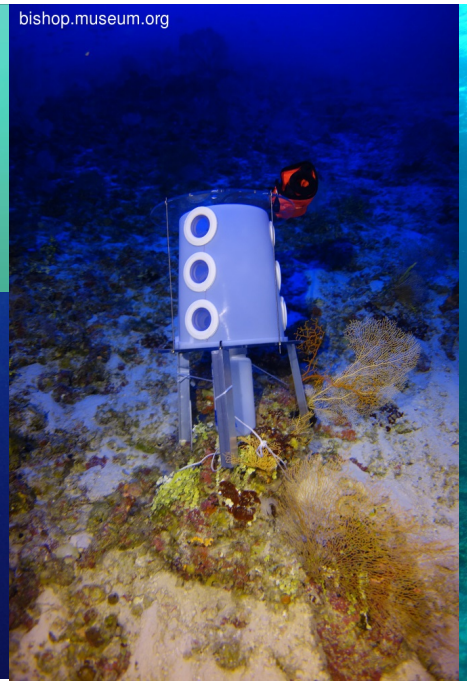
- Mark Carr
- Chris Edwards
- Dan Malone
- Emily Saarman
- Patrick Drake
- Anna Lowe

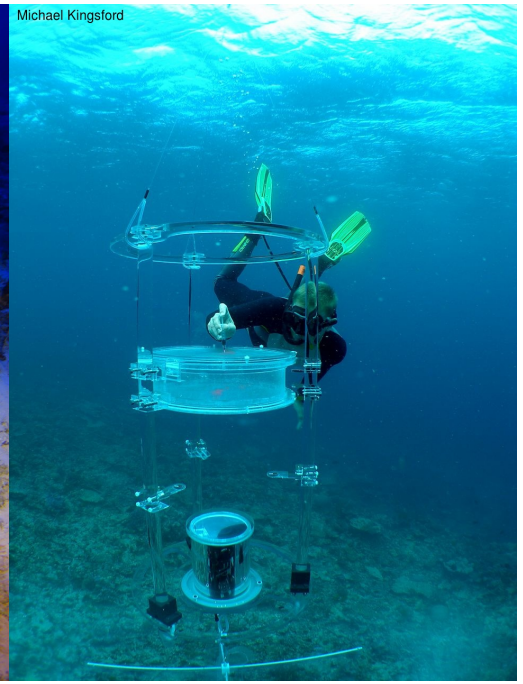
Larval dispersal patterns influence many things

- Dynamics of recruitment and populations
- The effects of fishing
- The utility of different MPA designs
- . . . and it is just fascinating from a behavioral ecology perspective

Studying larval settlement

Lots of interesting apparatus for collecting settling larvae.





But none answer the question “where did these larvae come from?”
For that, people have tried:

1. current modeling / drifters
2. isotope methods
3. genetic methods

Genetic approaches to studying recruitment

- Early approaches compared allele frequencies from collections made between:
  - recruits and adults
  - recruits in different areas
  - recruits at different times
- These are somewhat ‘Indirect” approaches because they are inferring historical patterns of connectivity based on currently observed patterns.
- Persistent, surprising findings of more genetic structure than expected: *chaotic genetic patchiness* (CGP).
- Some hypotheses for this: Sweepstakes hypothesis (Hedgecock & Pudovkin 2011); Unexpectedly high larval self-recruitment; kinship aggregation.

Kinship-based genetic recruitment studies

- Method 1: Parentage “Mark-recapture” by way of identifying parent-offspring connections between individuals.
- Method 2: less-direct method: identifying closely related kin within juvenile recruit samples.

Example of a Method-2-based conclusion

Selwyn et al. (2016) *PLOS One*. marine goby

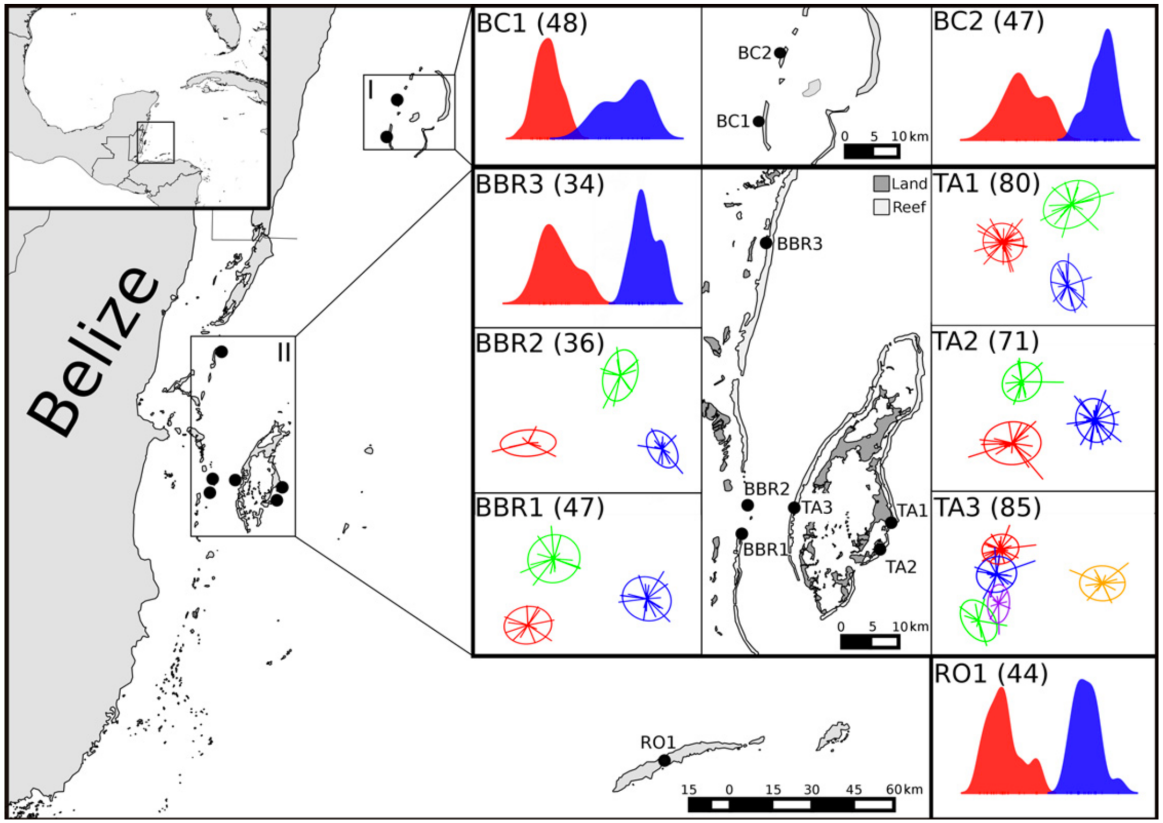


Fig 1. The Mesoamerican Barrier Reef study sites: BC–Banco Chinchorro, Mexico; BBR–Belizean barrier reef, Belize; TA–Turneffe Atoll, Belize; RO–Roatán, Honduras. Insets show scatter plots (and density plots in the case of two clusters) of clusters from DAPC analysis within sampling locations. The axes of the plots are the first two discriminant functions used to delineate clusters with inertia ellipses representing 67% of the variance. The end of the lines connected to the centre of each inertia ellipse represent individuals plotted on each discriminant function and denotes cluster membership. In locations with only two clusters present there is only one discriminant function, as such density plots of proportion of individuals present at each value of the discriminant function were included to show cluster separation. Numbers in parentheses indicate how many individuals were collected from each location.

(Disclaimer: I'm something of a skeptic)

A steady drumbeat of kin-based larval aggregation

Ecology, 87(12), 2006, pp. 3082–3094

CURRENT SHIFTS AND KIN AGGREGATION EXPLAIN GENETIC PATCHINESS IN FISH RECRUITS

MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 3476–3494 doi: 10.1111/mec.12341

Combined analyses of kinship and F_{ST} suggest potential drivers of chaotic genetic patchiness in high gene-flow populations

PLOS | ONE

RESEARCH ARTICLE

Kin-Aggregations Explain Chaotic Genetic Patchiness, a Commonly Observed Genetic Pattern, in a Marine Fish

Long-term aggregation of larval fish siblings during dispersal along an open coast

Let's do this, but go BIG

"Integrative evaluation of larval dispersal and delivery in kelp rockfish using *inter-generational genetic tagging*, demography and oceanography"

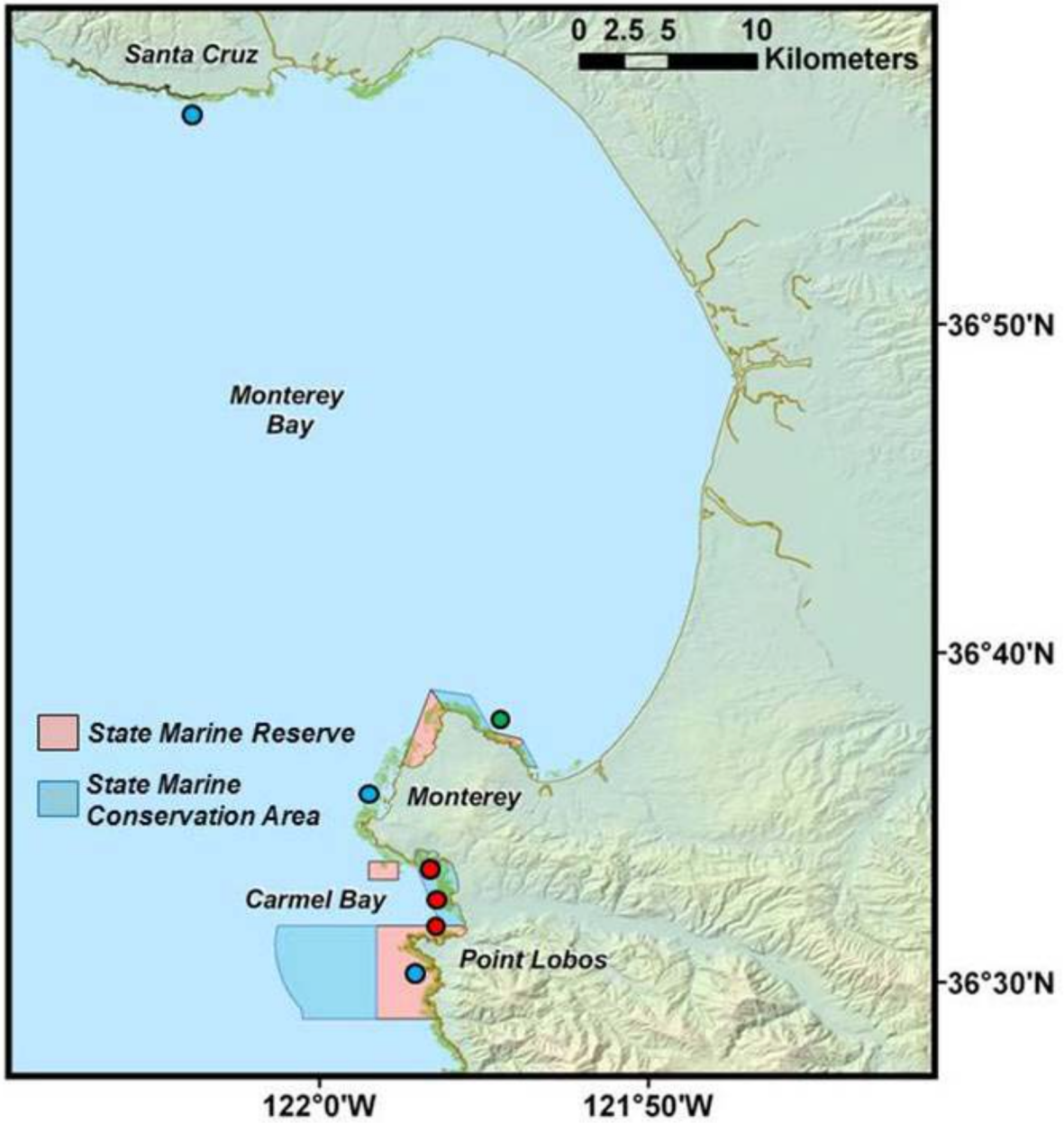
Mark Carr (PI), Chris Edwards, Carlos Garza, Eric Anderson (Co-PIs)

- **Goal:** use parentage inference to accurately estimate the degree of self-recruitment of kelp rockfish in Carmel Bay, and integrate this with fine-scale current modeling.
- I will be talking about the genetic tools we have developed to do this.

Kelp rockfish (*Sebastes atrovirens*)

Strongly associated with *Macrocystis*; $\approx$ Santa Cruz to Baja; Life-span 20-25 years; Mature $\approx$ 5 years; Live-bearing, internal fertlization with multiple paternity; Fecundity in the 100,000's; 2-3 months pelagic larval duration. Juveniles recruit to kelp beds in summer; Adults highly sedentary; no detectable population structure along coast (Gilbert-Horvath et al 2006).

Study Area: Carmel Bay



Carmel Bay in the context of coastal circulation

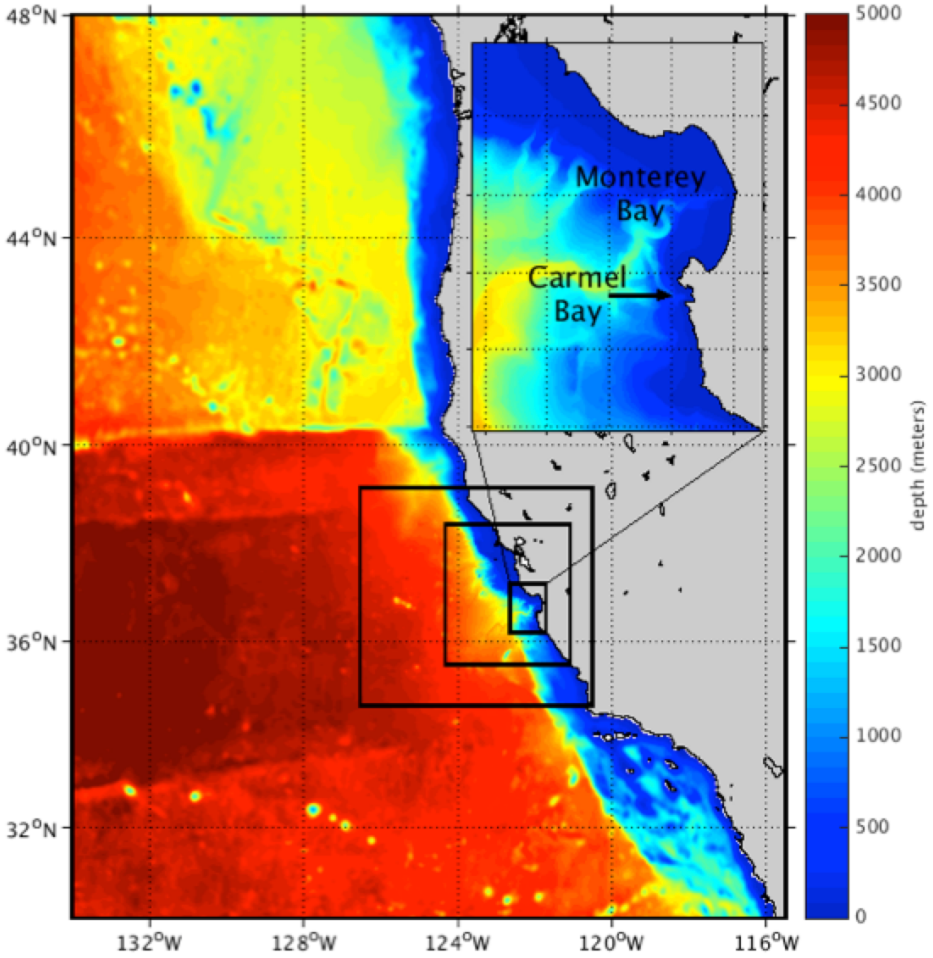
Anna Lowe (UCSC grad student) current modeling approach

ROMS (Regional Ocean Modeling System), realistically configured for the California Current System

Multiple, one-way nesting of model grid

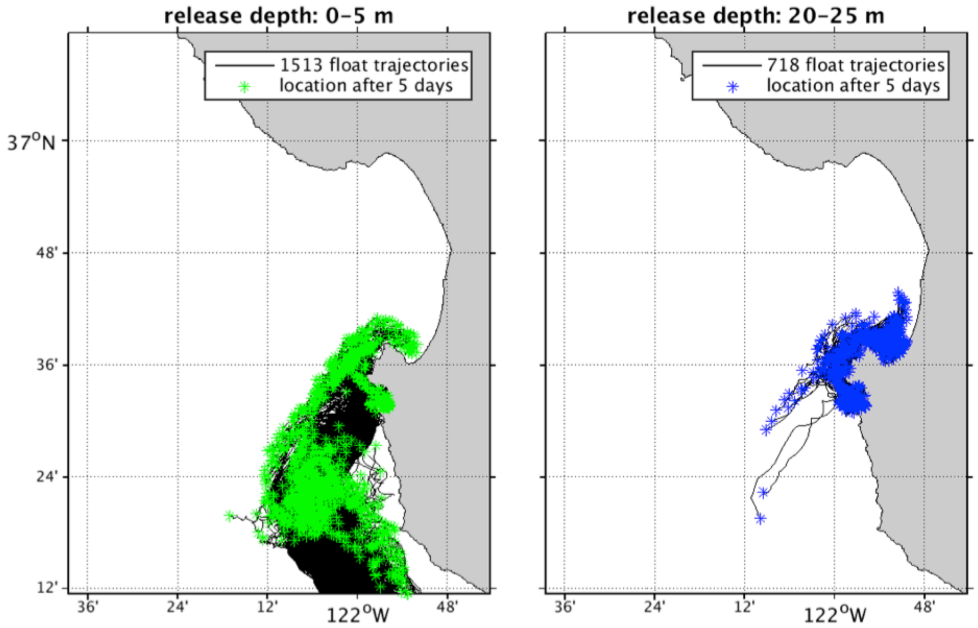
Factor of 3 increase with each nest

① 1/30 (~3.5 km)
② 1/90 (~1.2 km)
③ 1/270 (~400 m)
④ 1/810 (~130 m)



Current modeling around Carmel Bay

The potential for self-recruitment exists.



courtesy: A.B. Lowe

Genetic Sampling/Genotyping Goals

- Across 3 to 4 years (and perhaps more if we can continue this) get non-lethal samples from $\approx$ 5,000 adults and $\approx$ 5,000 recruiting kelp rockfish.
- Accurately identify parent offspring pairs from amongst the 25 million possible pairs.
- That is a lot of chances to make a mistake! Why did we think we could do this?

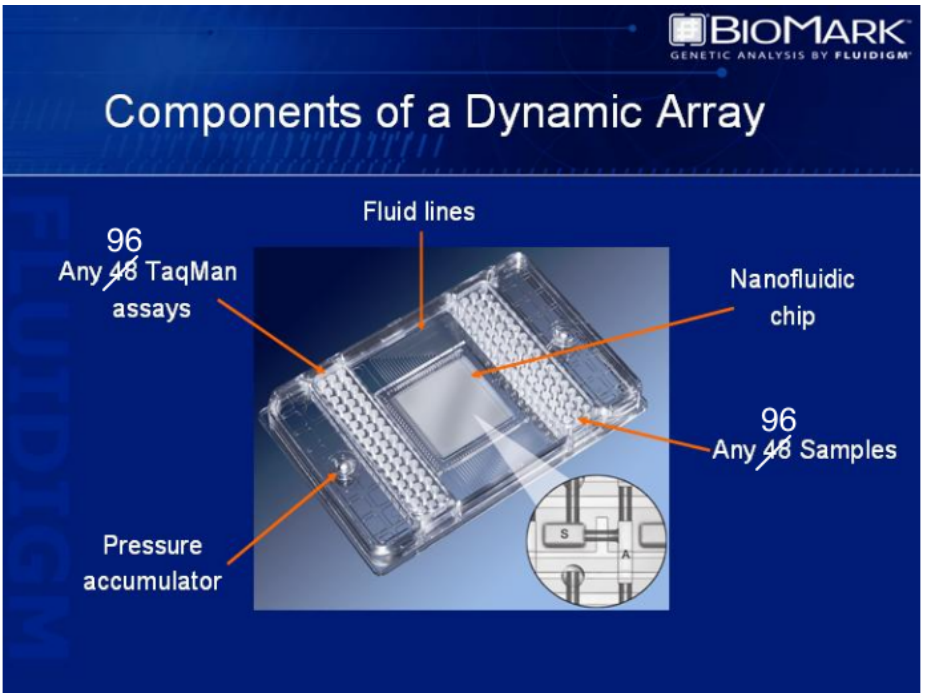
Digression: Why did we think we could do this?

We had pioneered similar techniques for replacing Coded Wire Tags for the monitoring of hatchery salmon populations

Our lab routinely genotypes tens of thousands of fish each year.

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: $<$\$10/fish

Microfluidic chips, a known quantity

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

The juvenile ID bombshell

- Below a certain size, kelp, gopher, and black-and-yellow rockfish are not visually dinstinguishable.
- Upwards of 50% of juveniles could be non-kelp!
- Non-kelp rockfish could be identified with our SNP assay
- But, the SNP data on those other species would be essentially worthless
- Solution: hijack NGS tech to create markers useful for all three species

GTseq amplicon sequencing

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

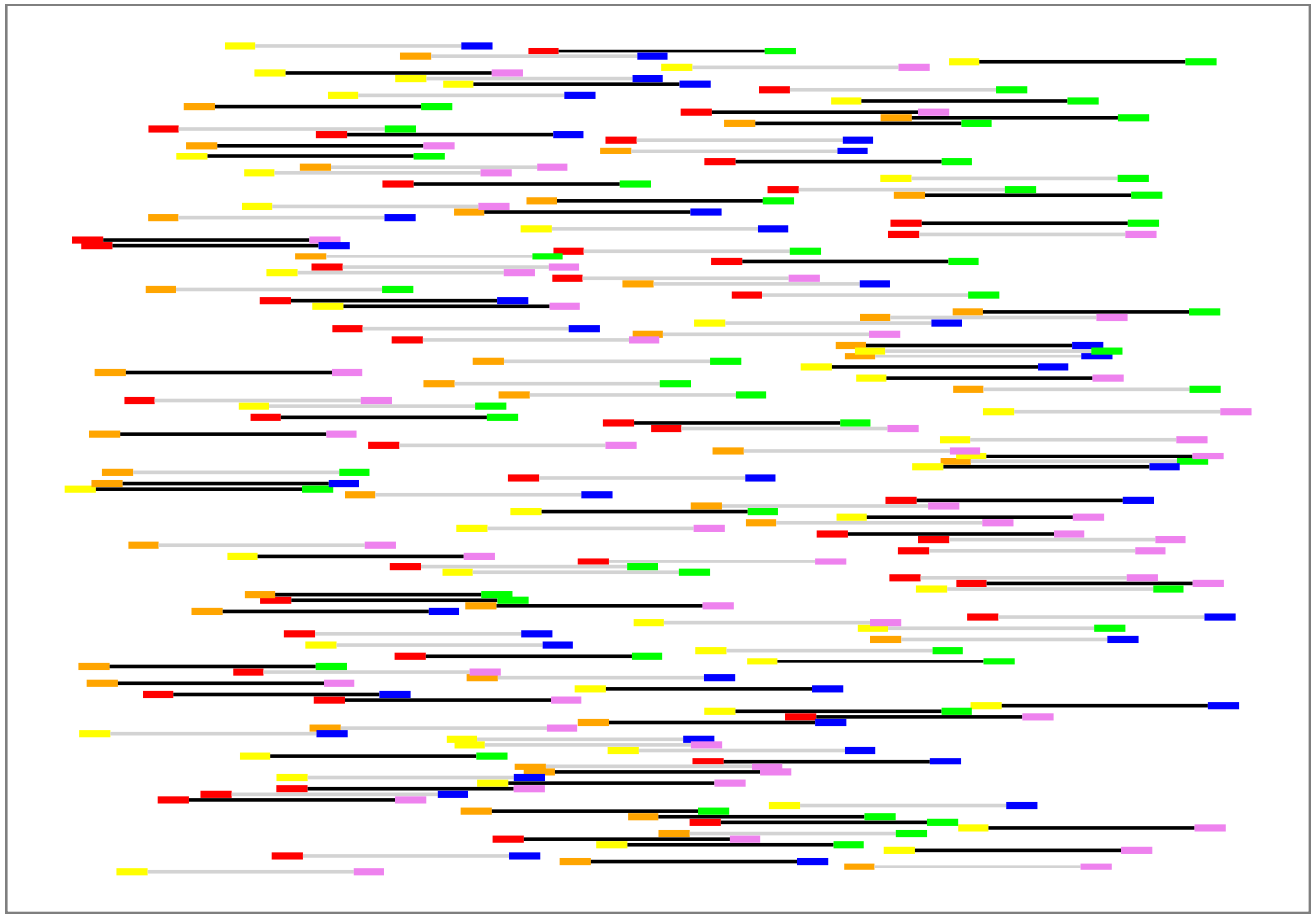
Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing

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 100–500 regions in 300–2,000 individuals
- $\approx$ \$7/individual
- Converted Fluidigm SNP assays. Tens of 1,000s of salmon each year.

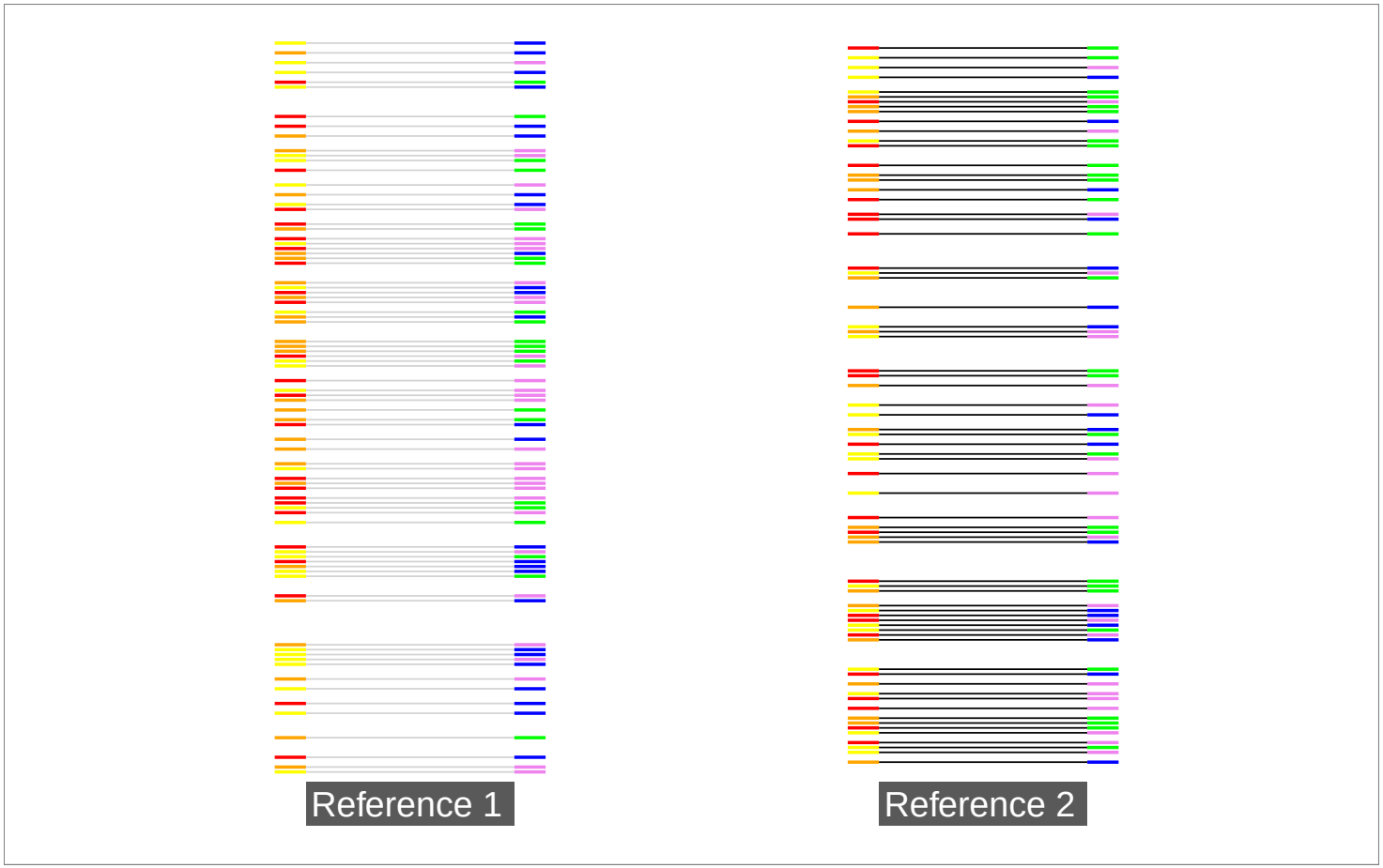
GTseq – I (coming off the sequencer. . .)

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



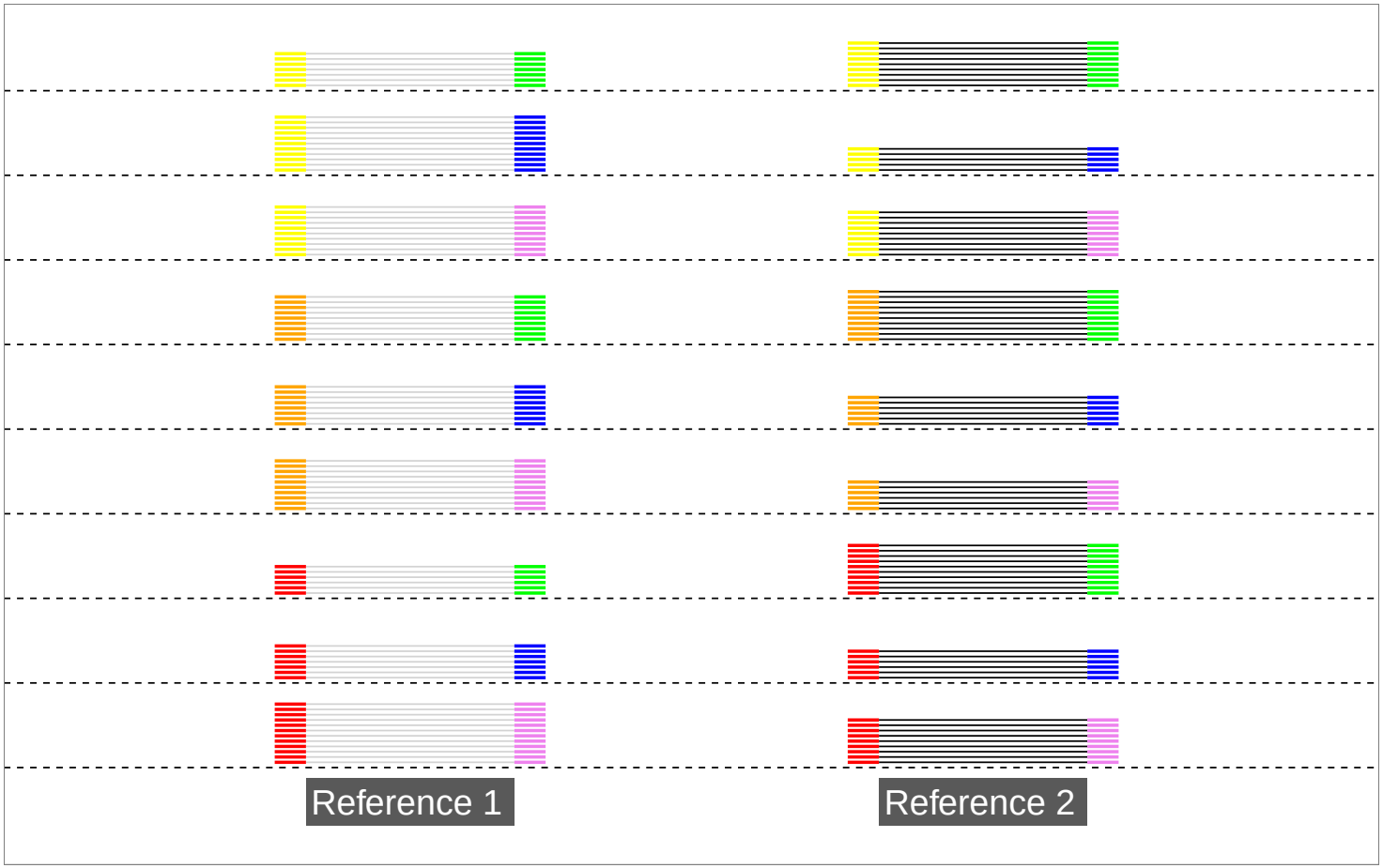
GTseq – II (aligned amplicons)

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



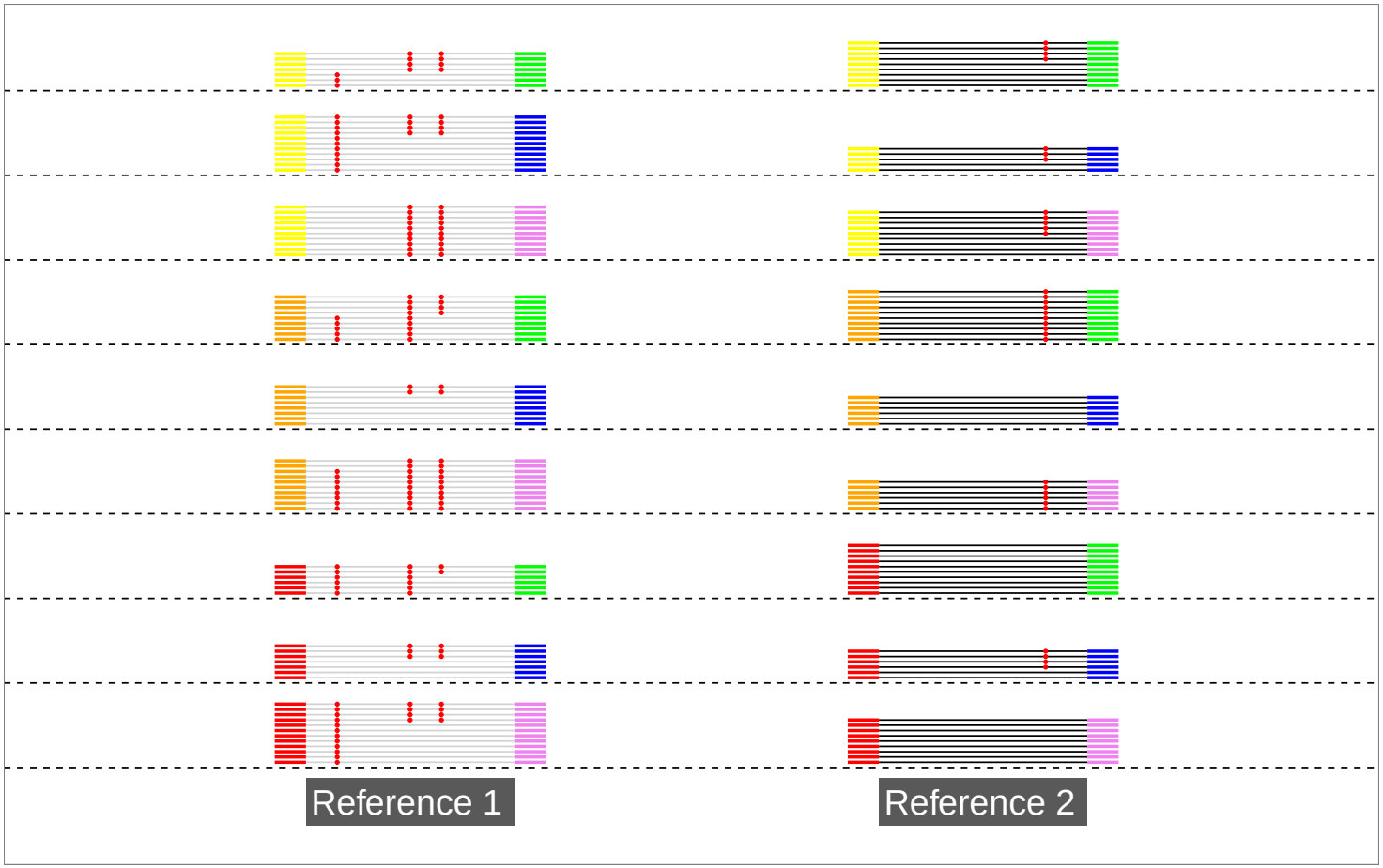
GTseq – III ("demultiplexing")

Identify individual origin of reads via combinatorial barcodes



GTseq – IV (identify SNPs in sequence)

Variants are easy to identify



Multiple SNPs in amplicons/sequences. . .

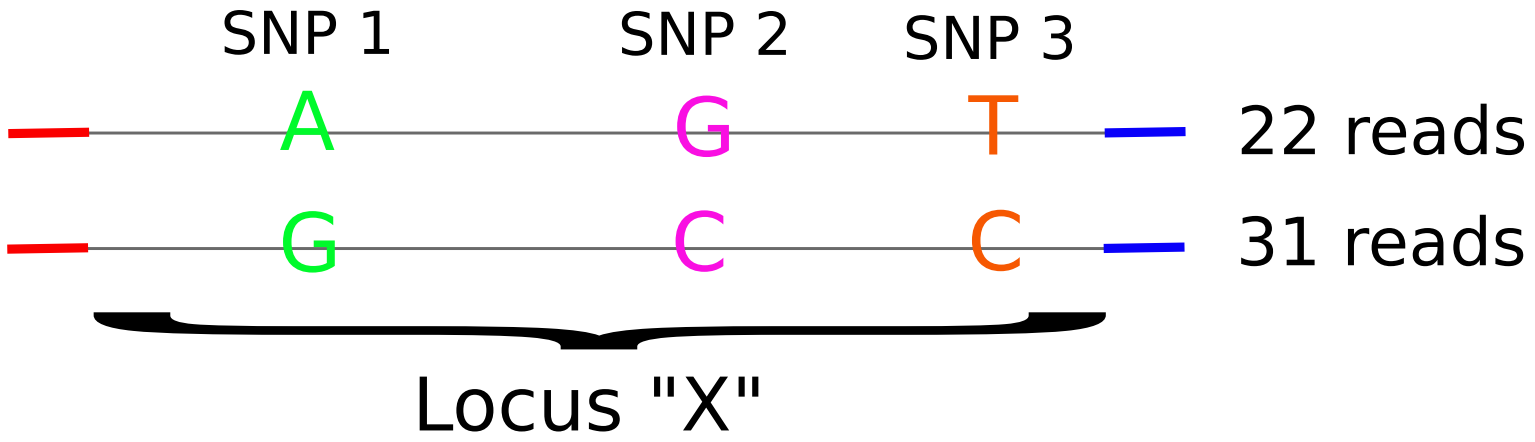
. . . are almost universally ignored

- 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

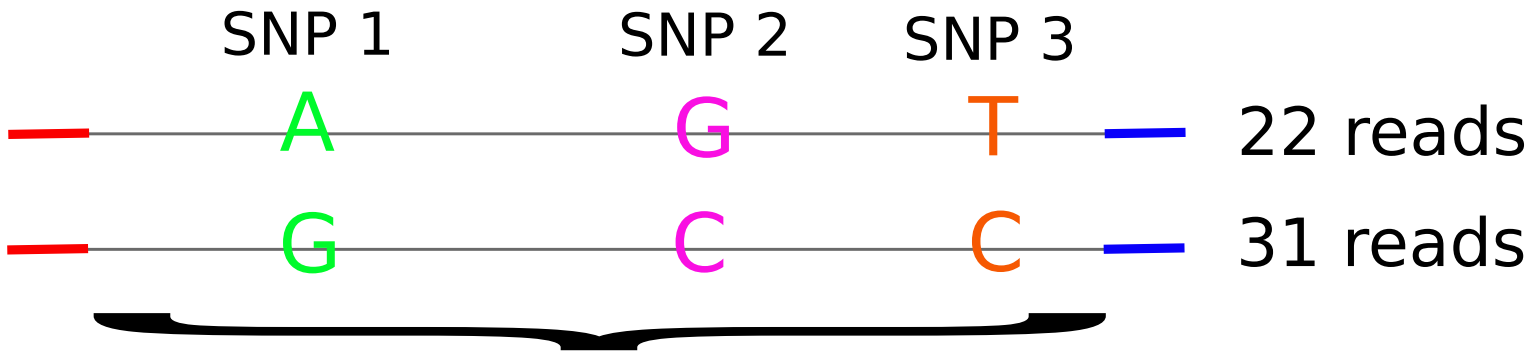
SNP 2: CG

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.

Microhaplotypes
Potential advantages

- 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.
  - So, we opened up sampling to more species
- We designed 96 microhaplotype amplicons for genotyping on an Illumina Mi-Seq in batches of 384 individuals.

SMURF traps

Standard Monitoring Unit for the Recruitment of Reef Fishes

Carr-Raimondi lab photo

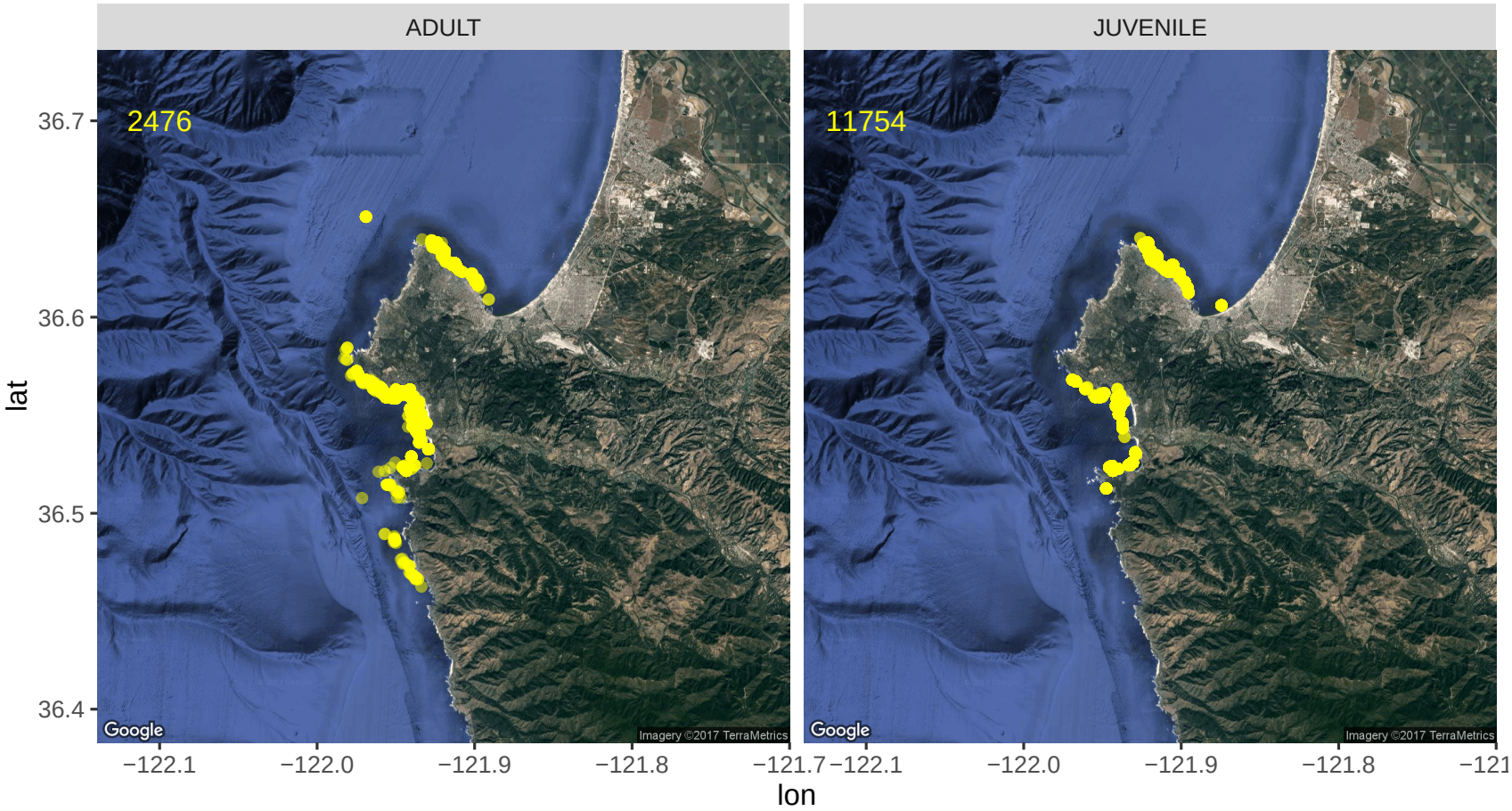
BINCKE nets

Benthic Ichthyofauna Net for Coral/Kelp Environments

Carr-Raimondi lab photo

Sampling of adult and juvenile rockfish

Legions of undergraduates fishing and on SCUBA using biopsy-dart pole spears



Emily Saarman  Dan Malone 

Microhaplotype Genotyping

Whoa! 15,000 fish genotyped in <4 months. Total materials cost ≈$6/fish

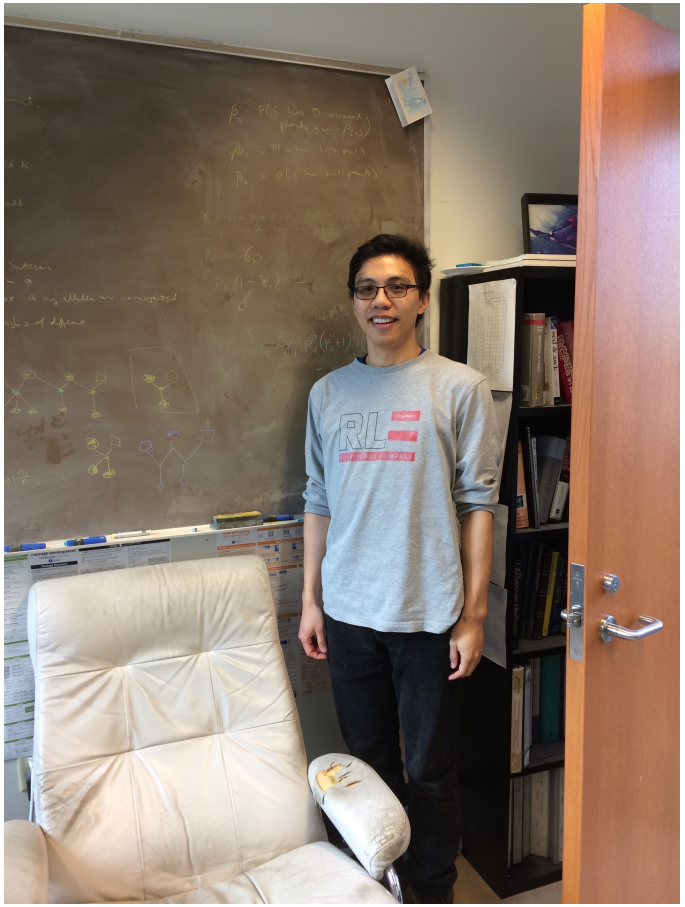
Diana Baetscher (UCSC Grad Student)

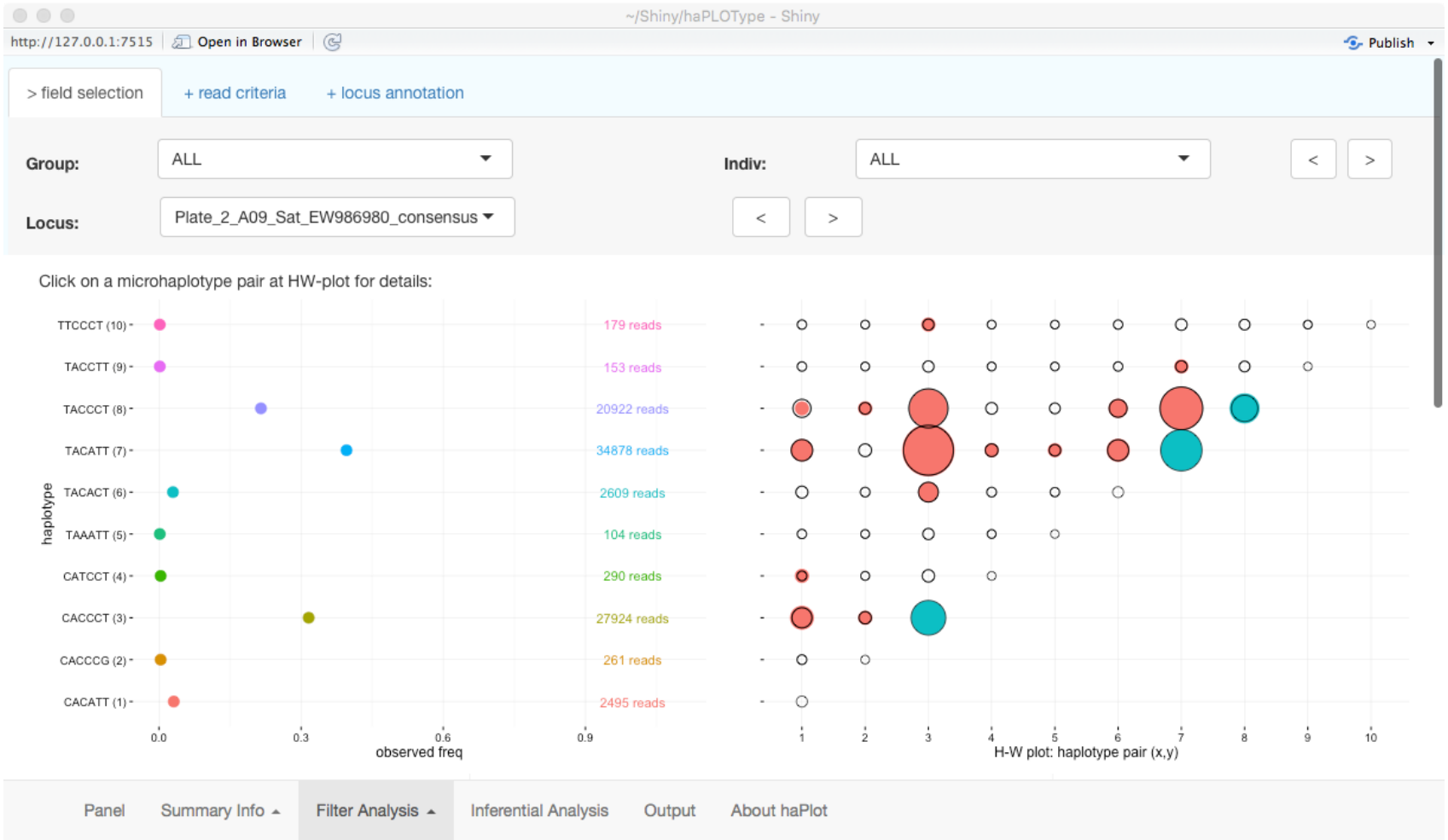
Extracting microhaplotypes from alignments

Bioinformatics software from our lab — MICROHAPLOT

- R package by Thomas Ng (UCSC grad student)
- 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



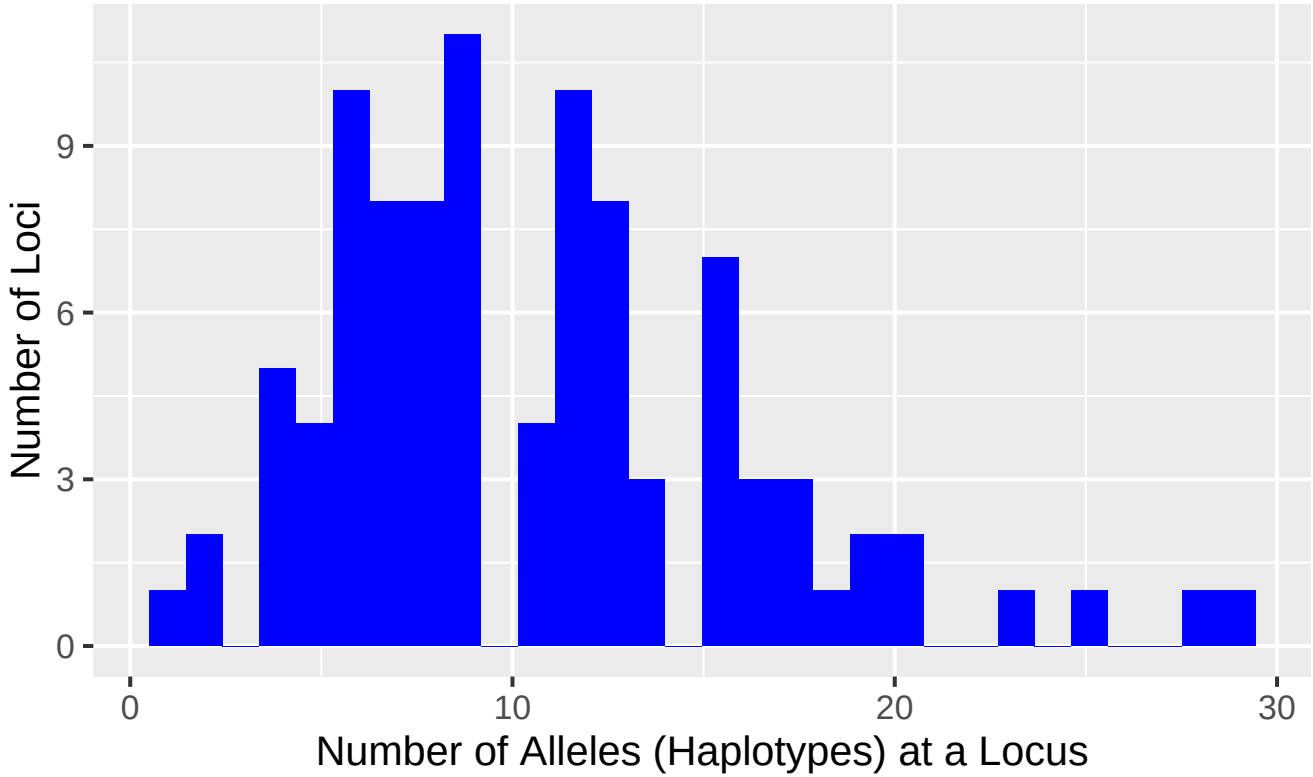

~/Shiny/haPLOType - Shiny
http://127.0.0.1:7515
Open in Browser
Publish
> field selection
+ read criteria
+ locus annotation
Group:
ALL
Indiv:
ALL
Locus:
Plate_2_A09_Sat_EW986980_consensus
Click on a microhaplotype pair at HW-plot for details:
haplotype
TTCCCT (10)
TACCTT (9)
TACCCT (8)
TACATT (7)
TACACT (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

How Well Does It Work? — Roadmap

- Works on multiple species? Yes! grad student Hayley Nuetzel just completed species ID work with 24 species. With very little missing data, and highly accurate species assignments.
- Haplotype Diversity in Kelp Rockfish
- Genotype Quality / Accuracy:
  - Sequence Read Depths
  - Distortions from Hardy-Weinberg Proportions
  - Regenotype Discordance Rate
- Preliminary Results from Relationship Inference

Allele/Haplotype Diversity

In 96 Loci in Kelp Rockfish: 1039 Unique Alleles in total!



Read Depths For Calling Alleles

Much higher than low-coverage genome work

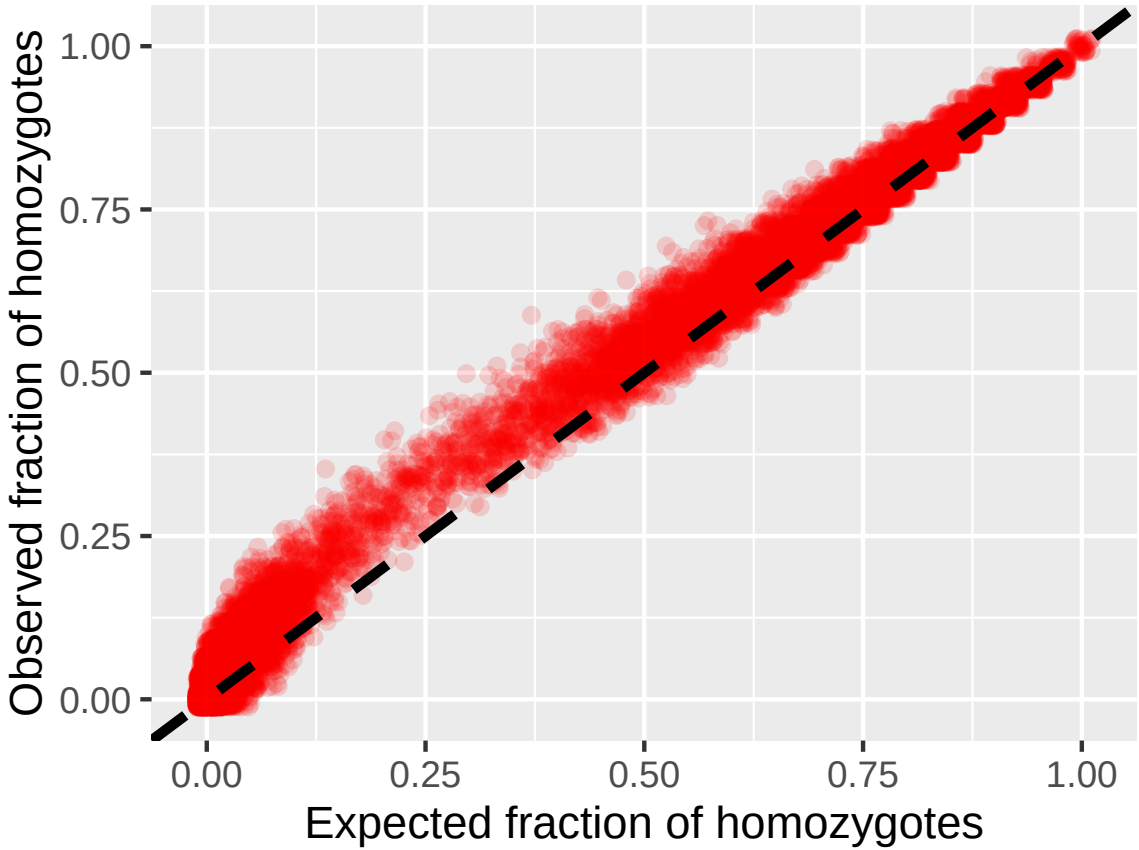
Across ≈15,000 individuals and at all 96 loci, called-allele read depths were as follows:

Read Depth Bin	n	Percentage	Cumulative
(0,10]	3859	0.2	0.2
(10,20]	20509	1.1	1.3
(20,30]	34903	1.9	3.2
(30,40]	44896	2.4	5.6
(40,50]	50964	2.8	8.4
(50,75]	138593	7.5	15.9
(75,100]	136191	7.4	23.3
(100,250]	637913	34.5	57.8
(250,500]	484226	26.2	83.9
(500,1e+03]	234141	12.7	96.6
(1e+03,1e+06]	62604	3.4	100.0

Some types of next-gen data *can* be a little dodgy...

Published lobster data from Benestan et al. 2016

Homozygosities at 10,156 SNPs from RAD-Seq

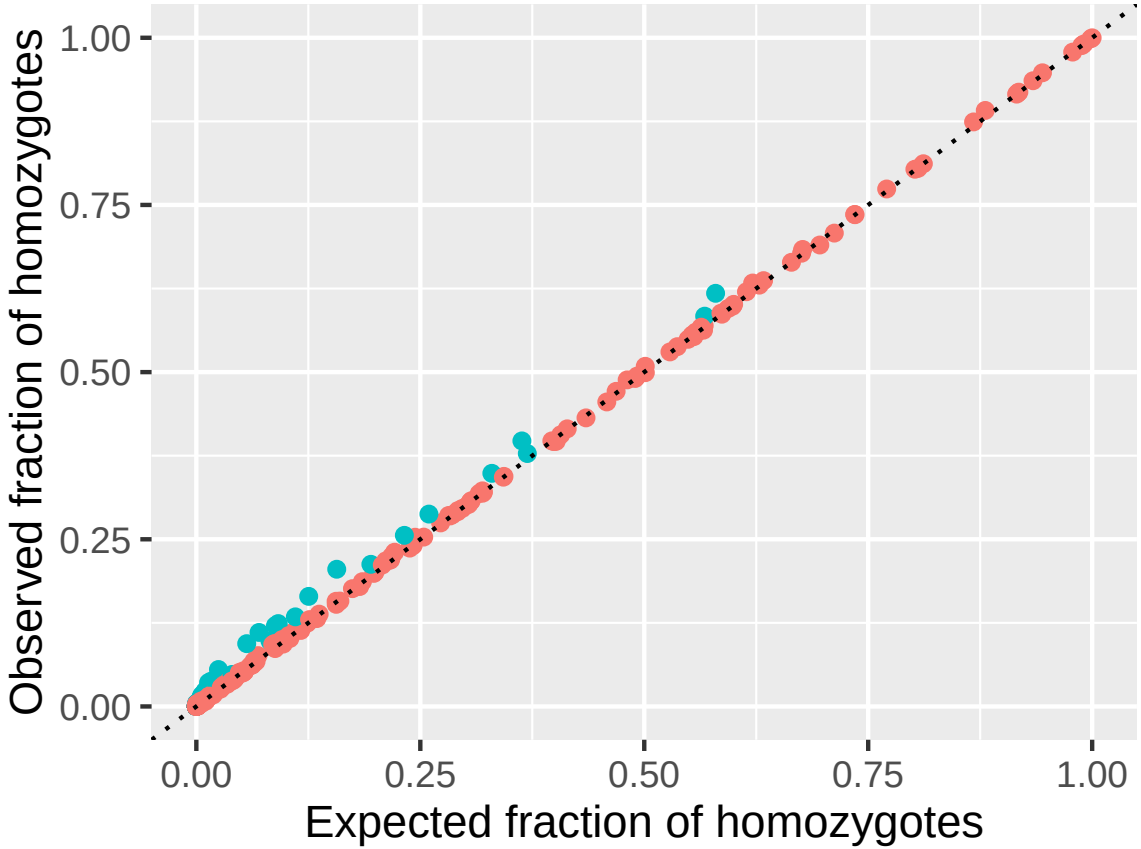


Rad-Seq and Whole-genome Seq: Low read depth and allelic dropout can cause heterozygotes to be erroneously called as homozygotes.

Microhaplotype data are *not* dodgy. . .

Allele-specific Homozygosities from ≈ 6,000 kelp rockfish

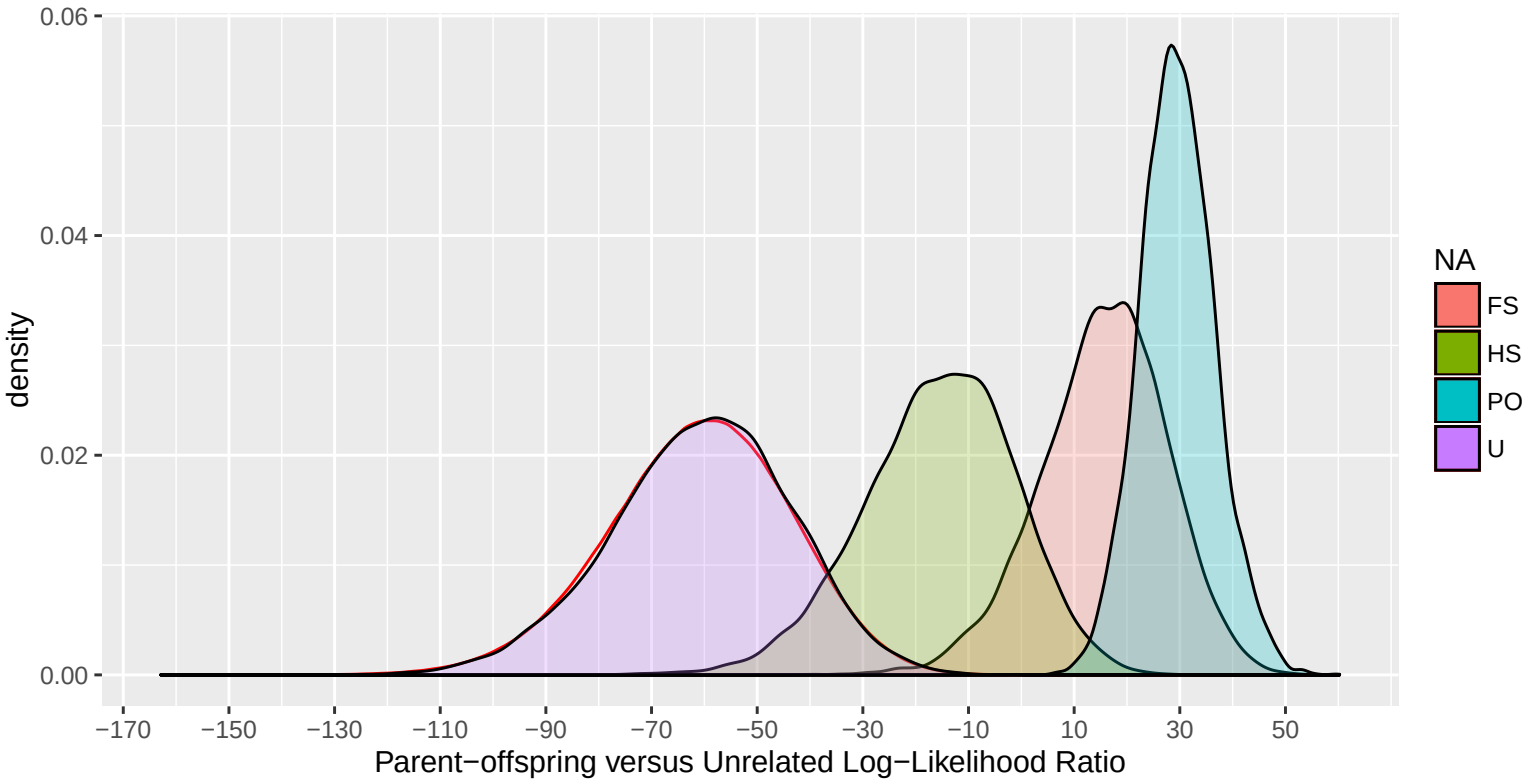
Green = 10 loci with null alleles. Orange = 86 remaining loci



- Null alleles can be treated systematically.
- From 75 kelp rockfish genotyped twice, the per-locus discordance rate was 3/1000.

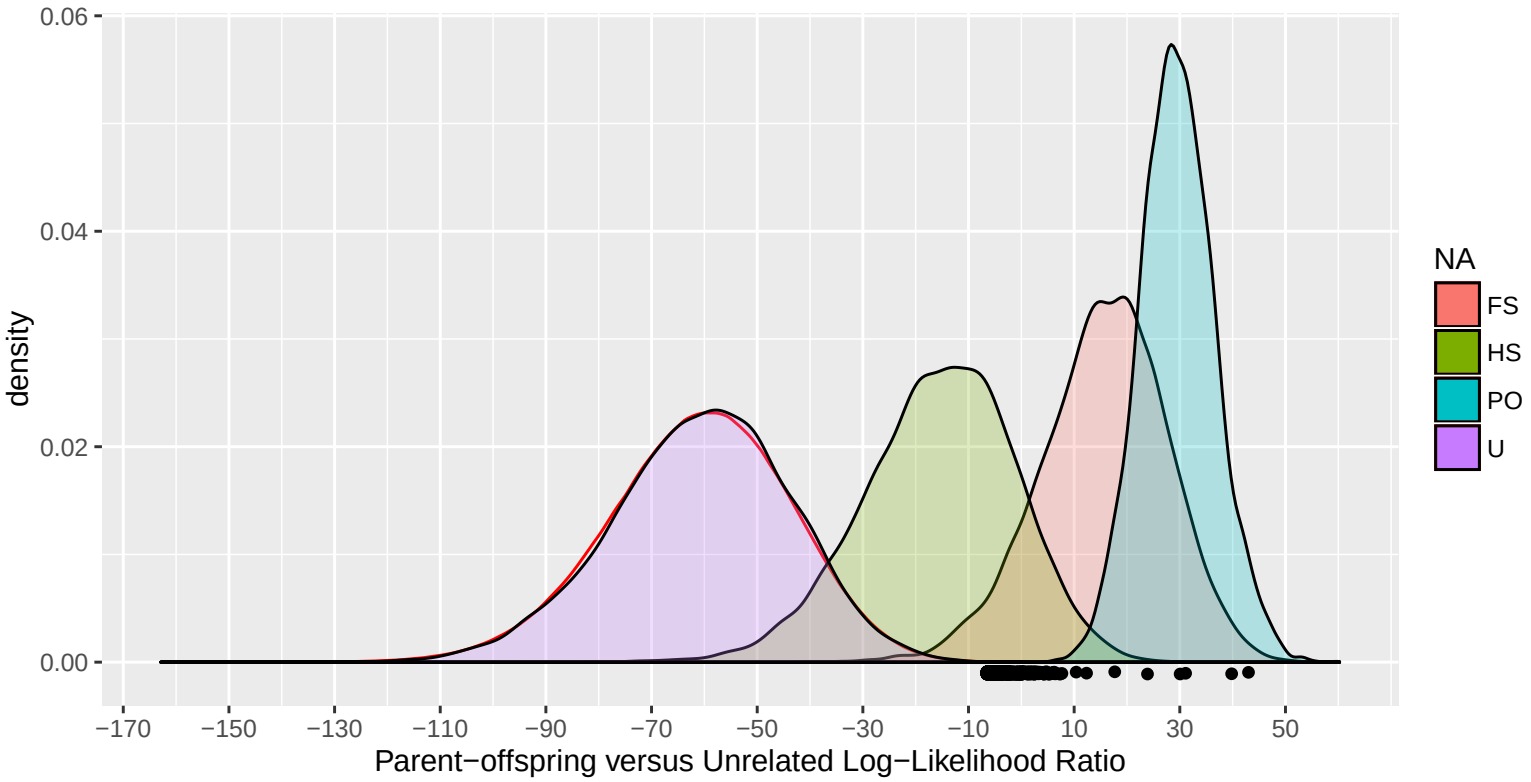
Microhaplotypes for relationship inference

Lots of power for parentage



And, preliminarily, we find 6 high-confidence parent-offspring pairs

Out of 4177 juvenile and 1889 adult kelp rockfish.



Calculating False Positive Rates And Efficiently Doing Pairwise Relationship Inference

Software from our lab – CKMRsim

`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`

Identifying Full Siblings
Again, preliminarily...

- We found 38 high-confidence full-sibling pairs
- None of these were in large full-sibships as reported by some previous studies
- They were all in simple full-sib pairs
- A funny story about species misidentification and apparent large full sibling groups...

Conclusions

- Next-generation-sequenced microhaplotypes—a powerful and economical system for high-throughput genotyping.
- The first case, to our knowledge, of identifying the source location of pelagically dispersed larvae of a marine animal on the open West Coast of North America
- We demonstrate the feasibility of this approach to identify patterns of population connectivity for marine fishes along the open coast
- Will demonstrate and inform how young produced within MPAs contribute to local self replenishment, to fished populations outside an MPA, and to other MPAs across a network
- By using microhaplotypes we are able to extend our project to three species with almost no increase in cost.