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

Eric C. Anderson

Conservation Genomics Workshop

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

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

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

Try to study larval dispersal by linking larvae to parents

"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)

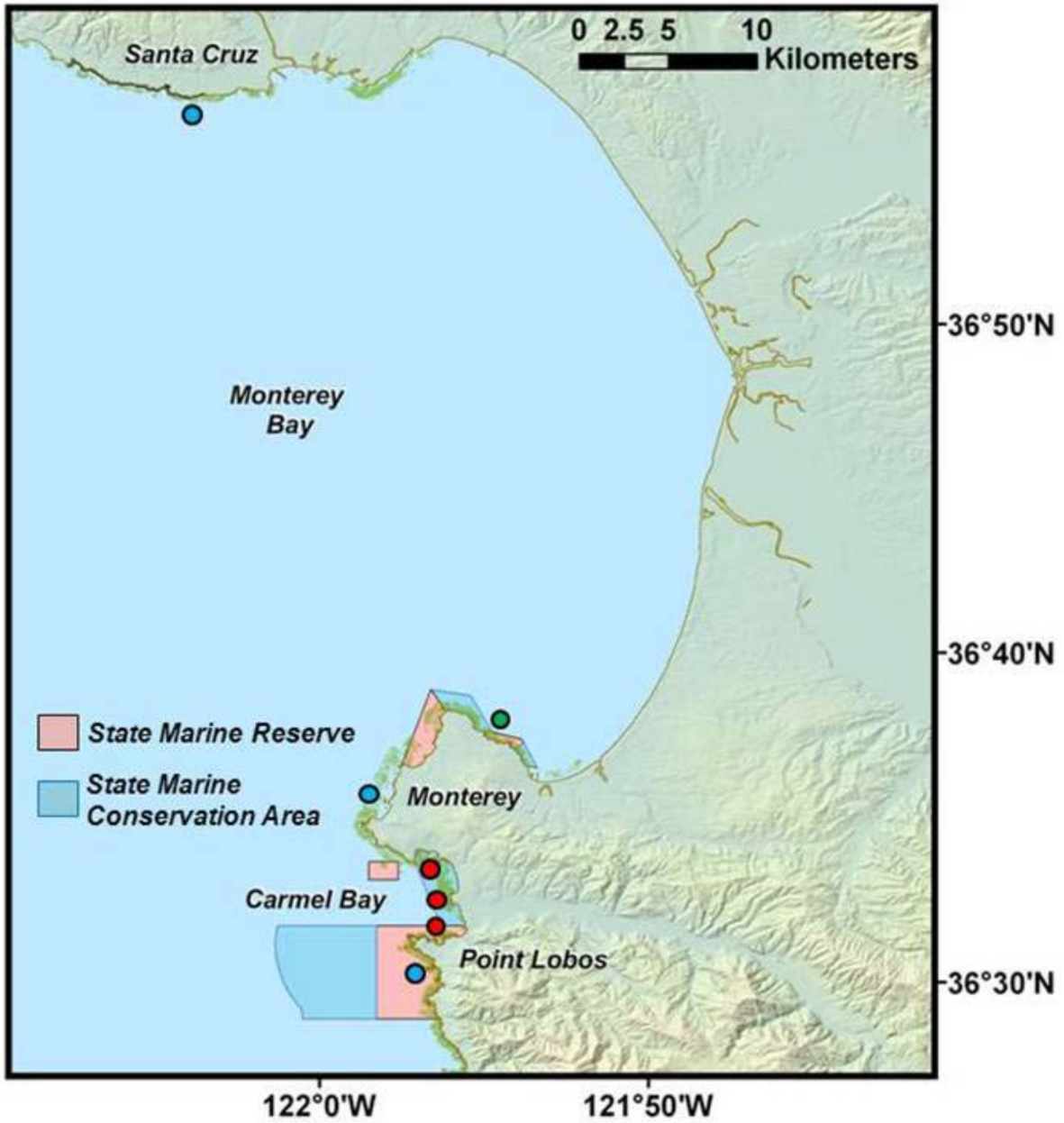


- **Goal:** use parentage inference to accurately estimate the degree of self-recruitment of kelp rockfish in Carmel Bay, and integrate this with fine-scale current modeling.
- I will be talking about the genetic tools we have developed to do this.

Kelp rockfish (*Sebastes atrovirens*)

Strongly associated with *Macrocystis*; ≈ Santa Cruz to Baja; Life-span 20-25 years; Mature ≈ 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



SMURF traps

Standard Monitoring Unit for the Recruitment of Reef Fishes

Carr-Raimondi lab photo

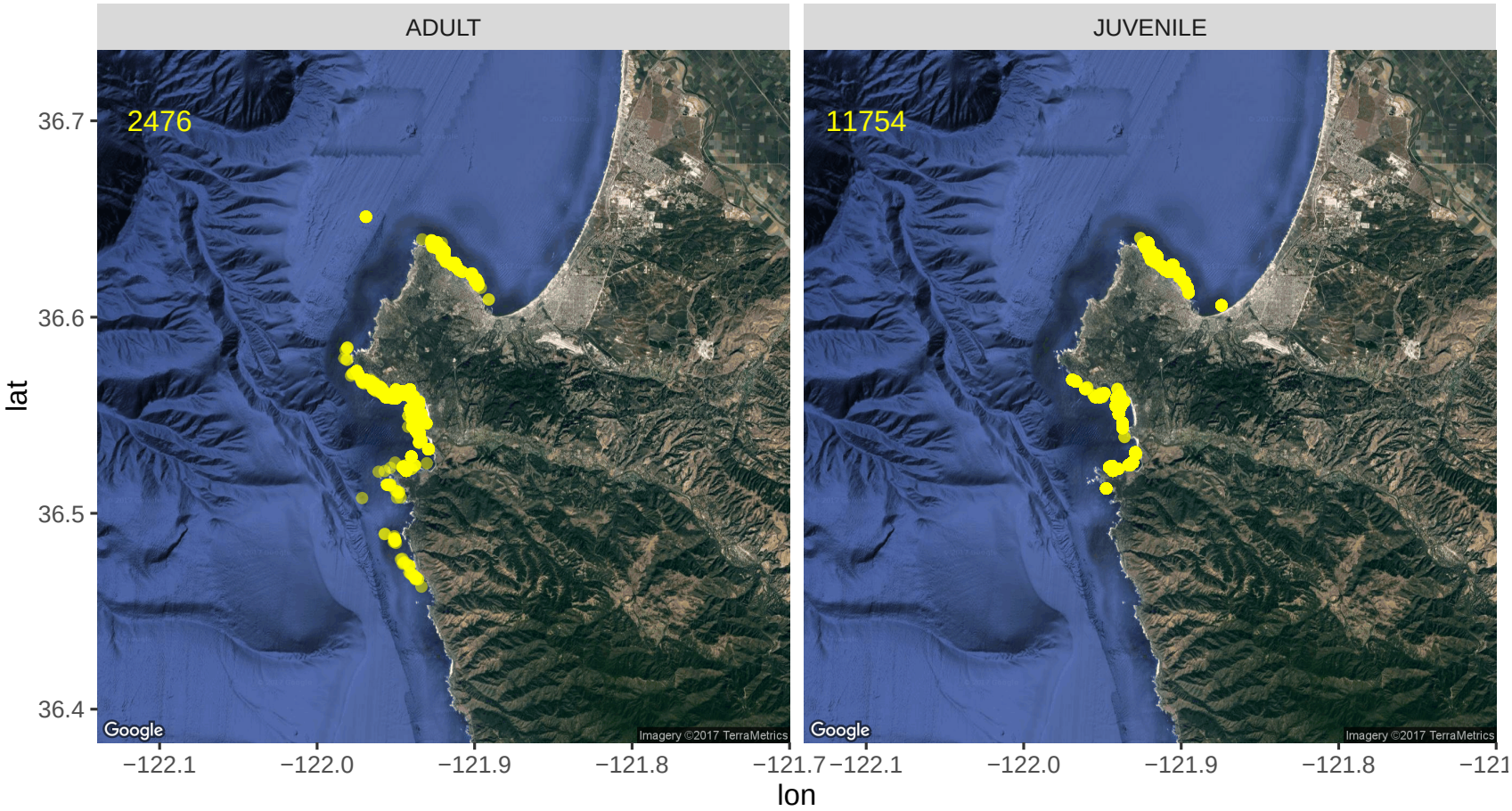
SMURF traps

Standardized protocol for juvenile collection

Carr-Raimondi lab photo

Sampling of adult and juvenile rockfish

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



Emily Saarman 

Dan Malone 

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?

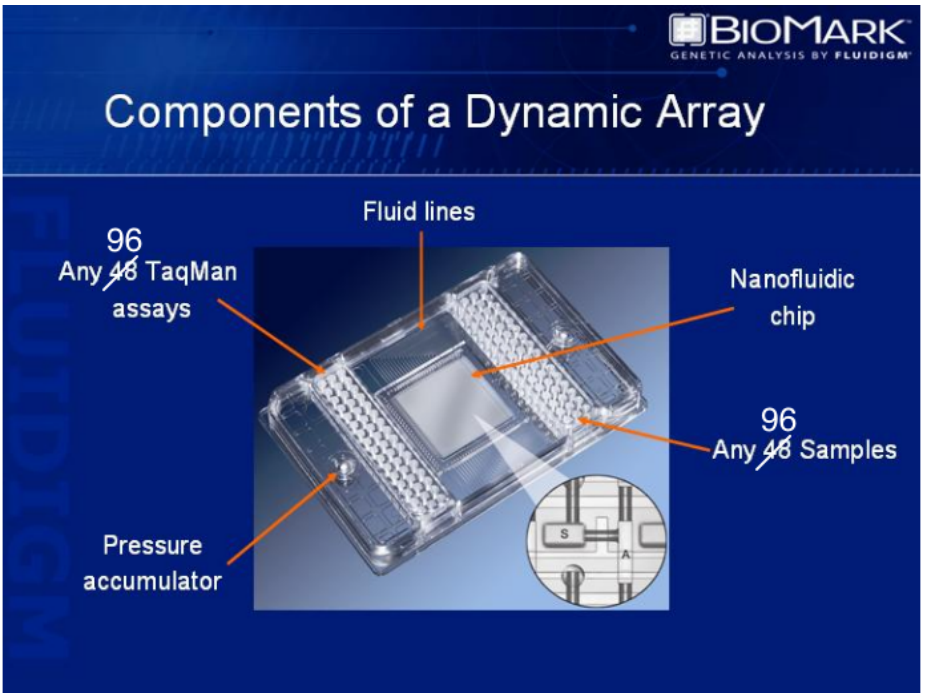
Digression: Why did we think we could do this?

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

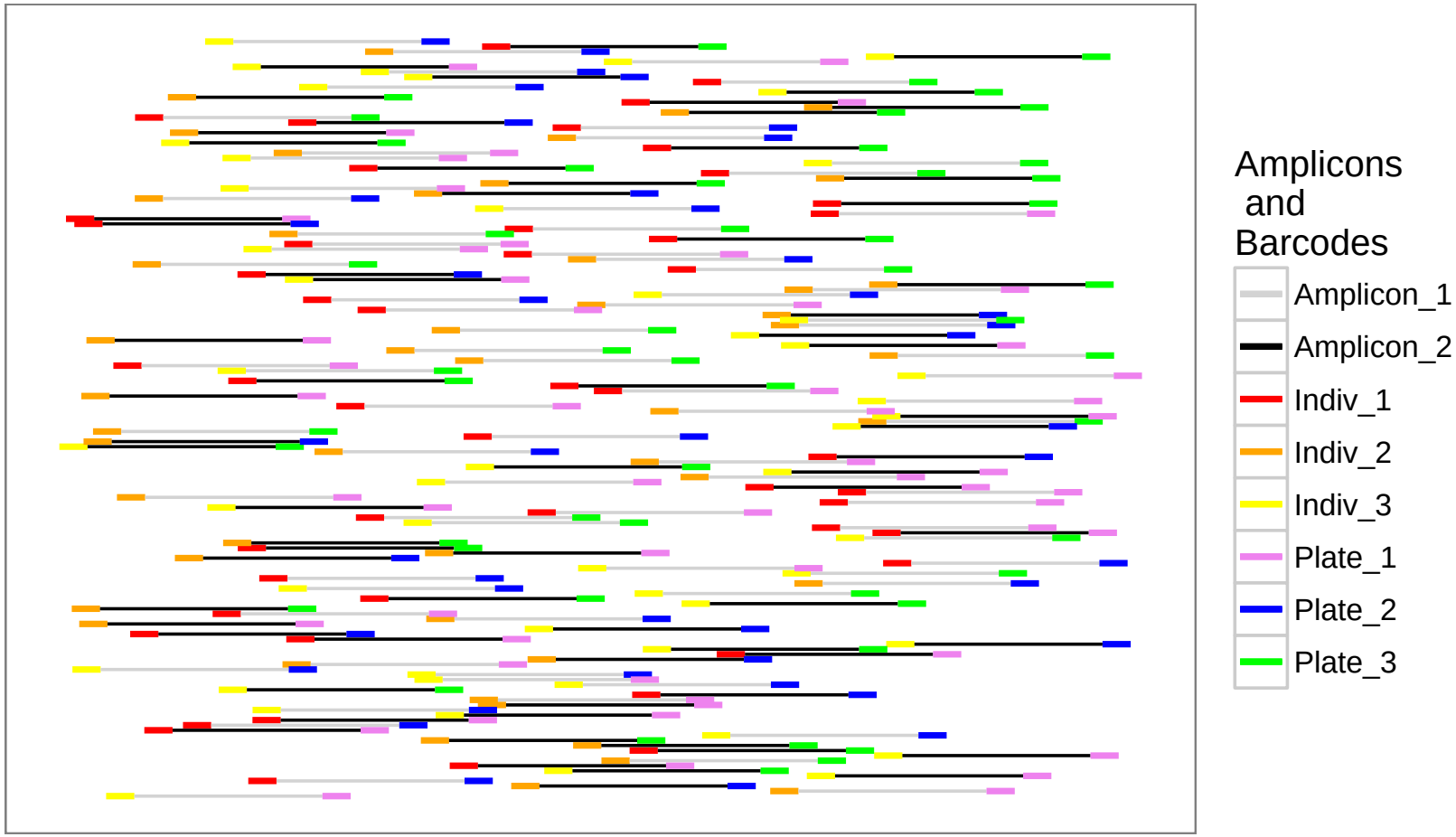
Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing

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.

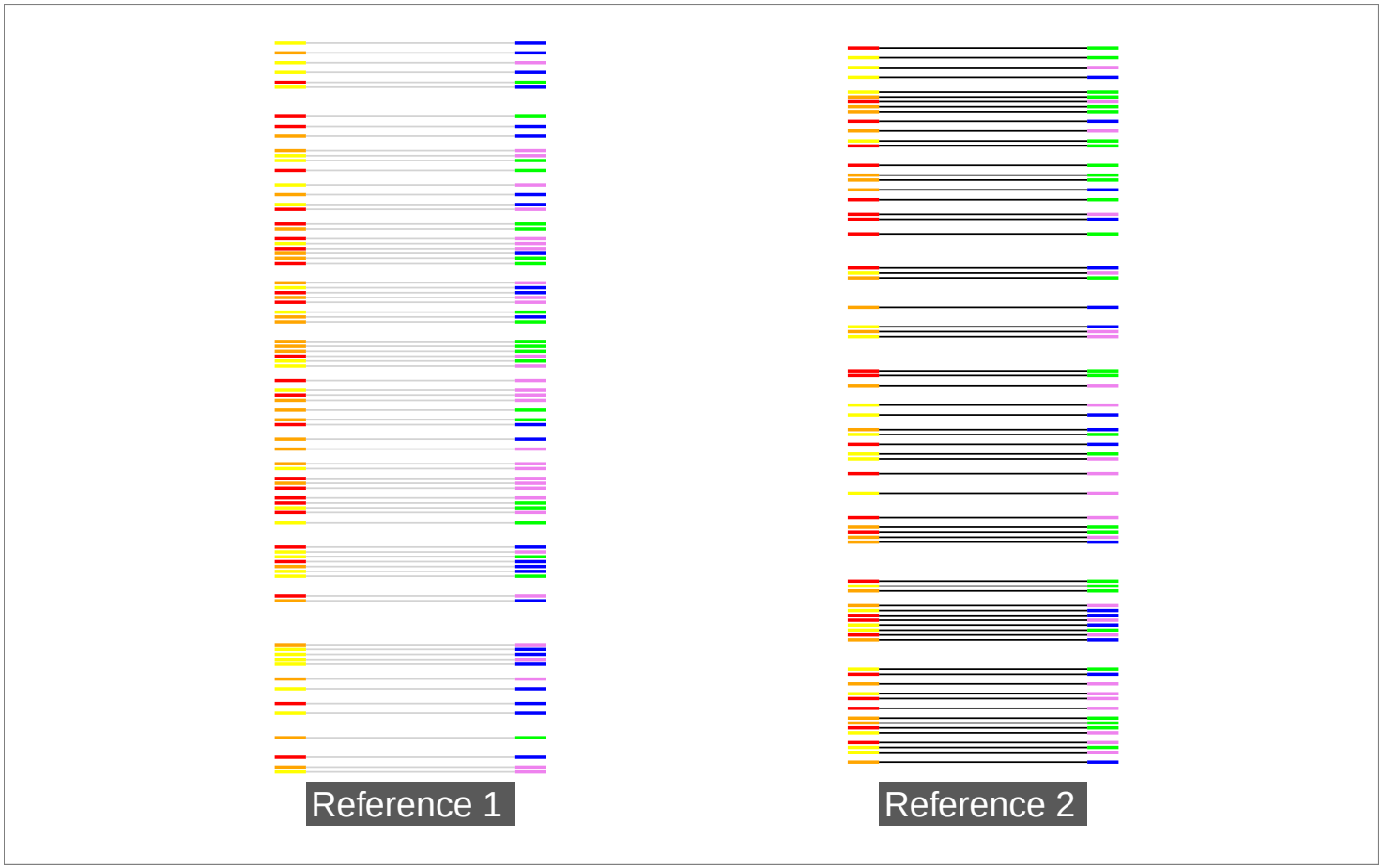
GTseq – I (coming off the sequencer. . .)

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



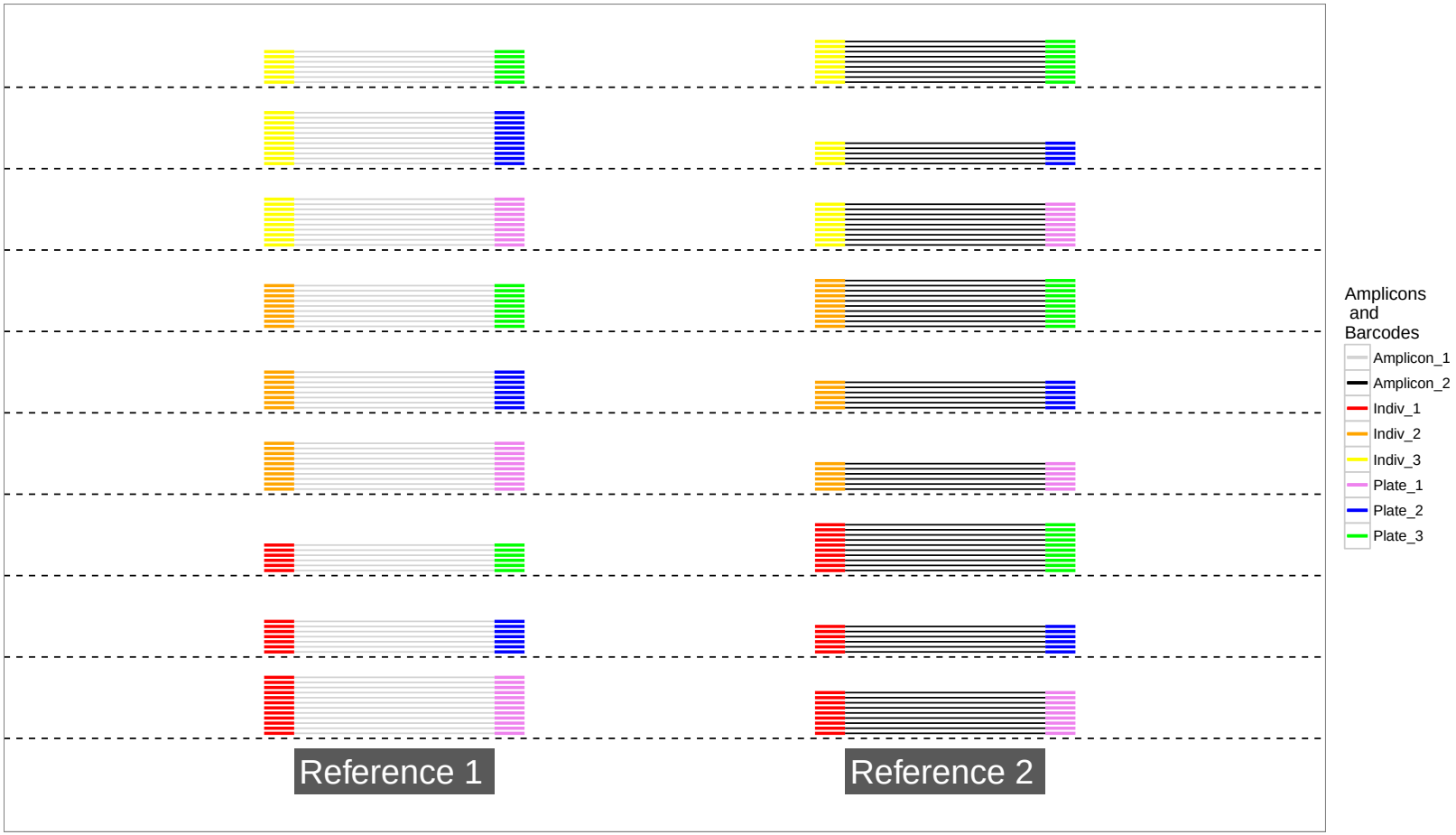
GTseq – II (aligned amplicons)

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. . .

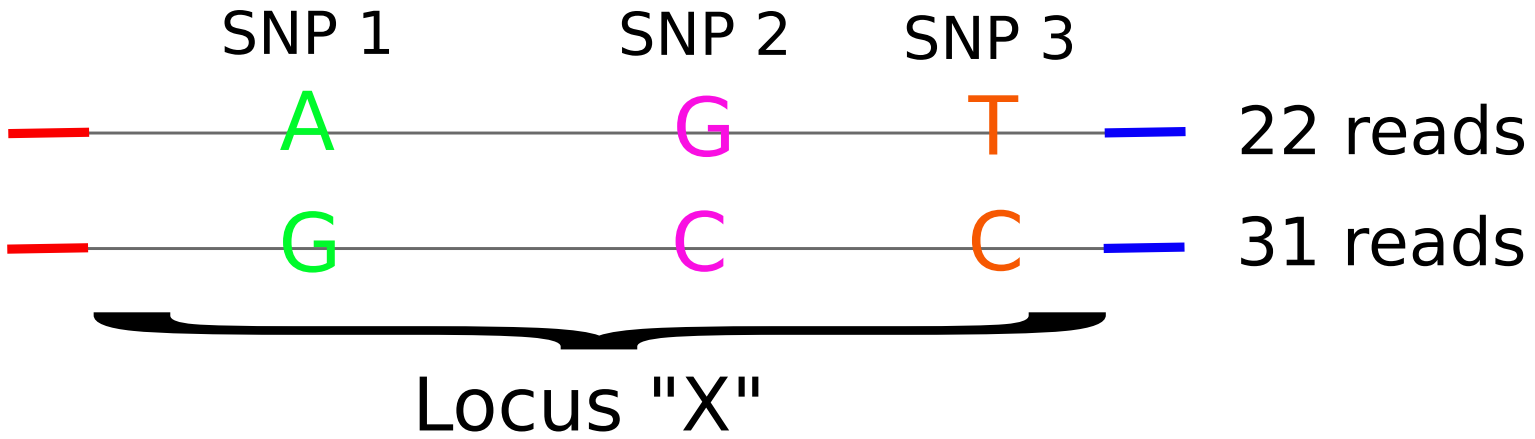
. . . are almost universally ignored

- 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

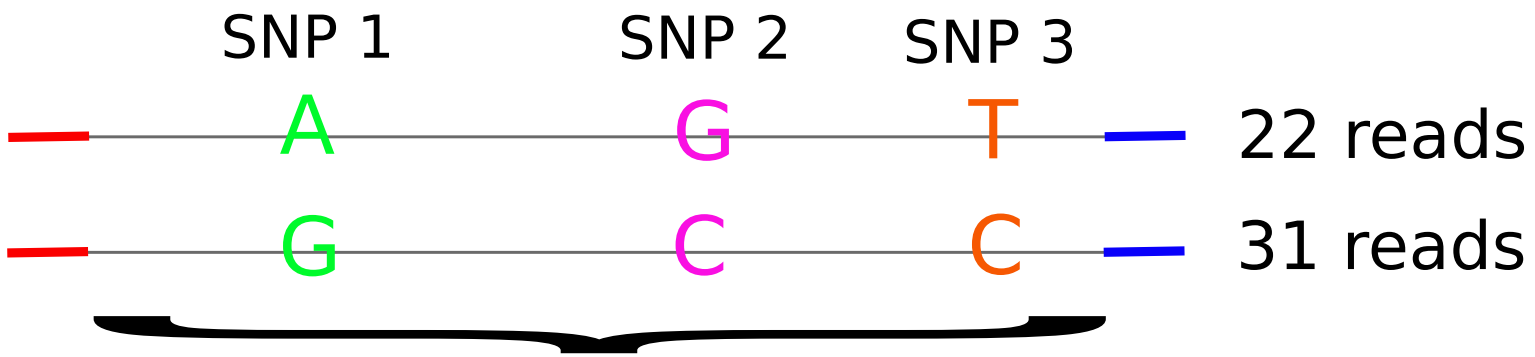
SNP 2: CG

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.

Microhaplotype Genotyping

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

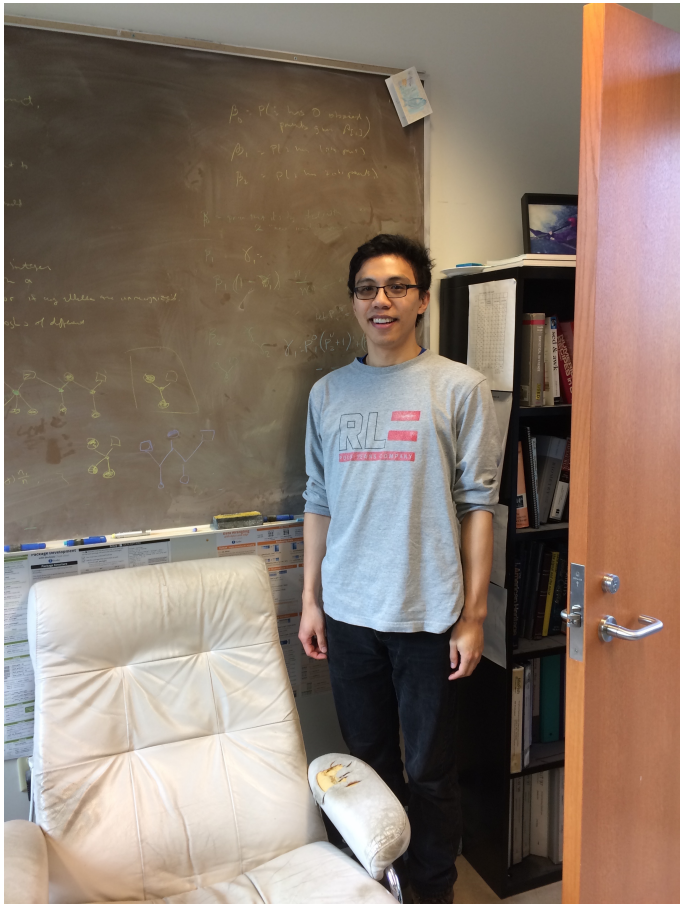
Diana Baetscher (UCSC Grad Student)

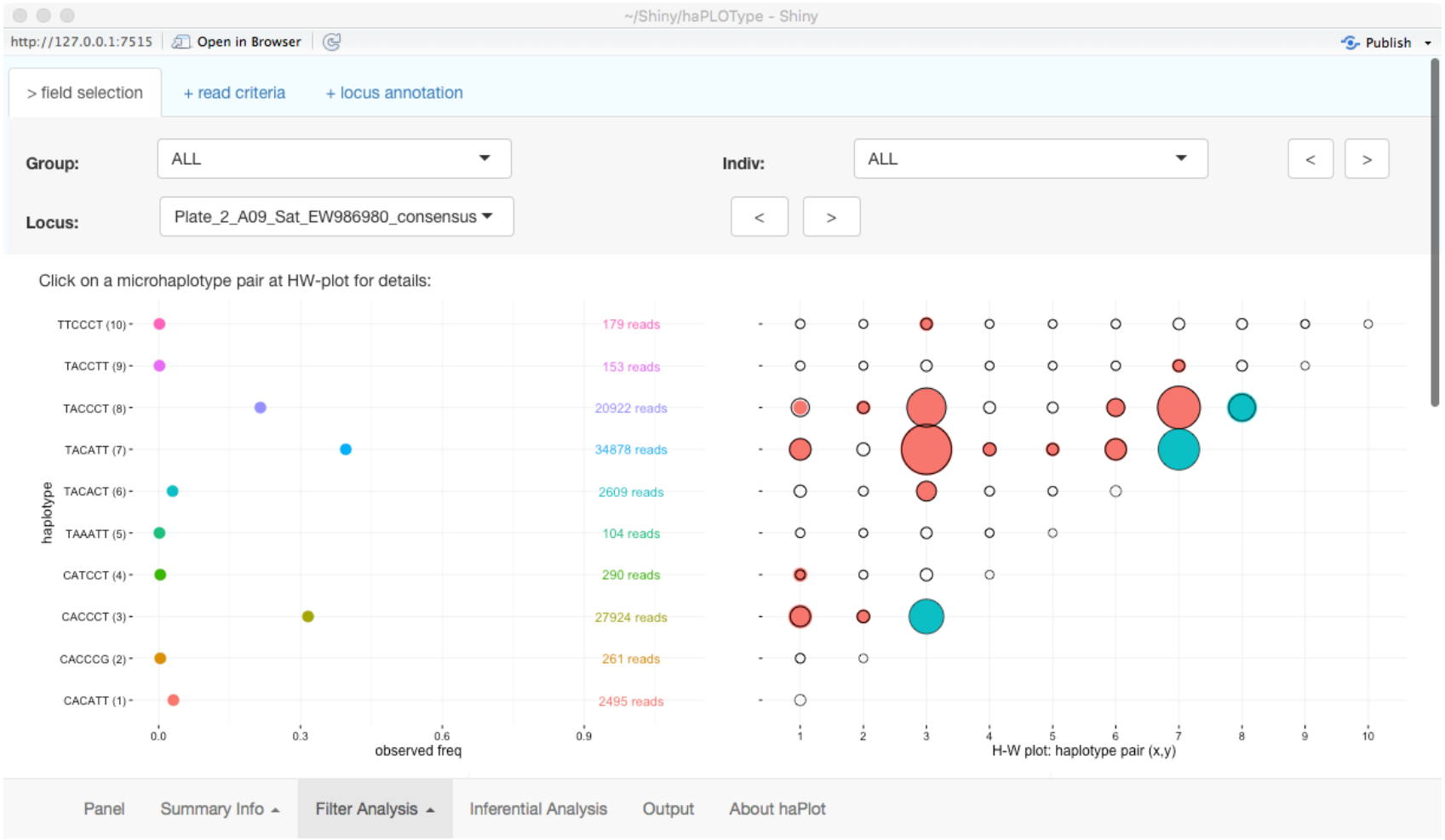
Extracting microhaplotypes from alignments

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/microhaplot



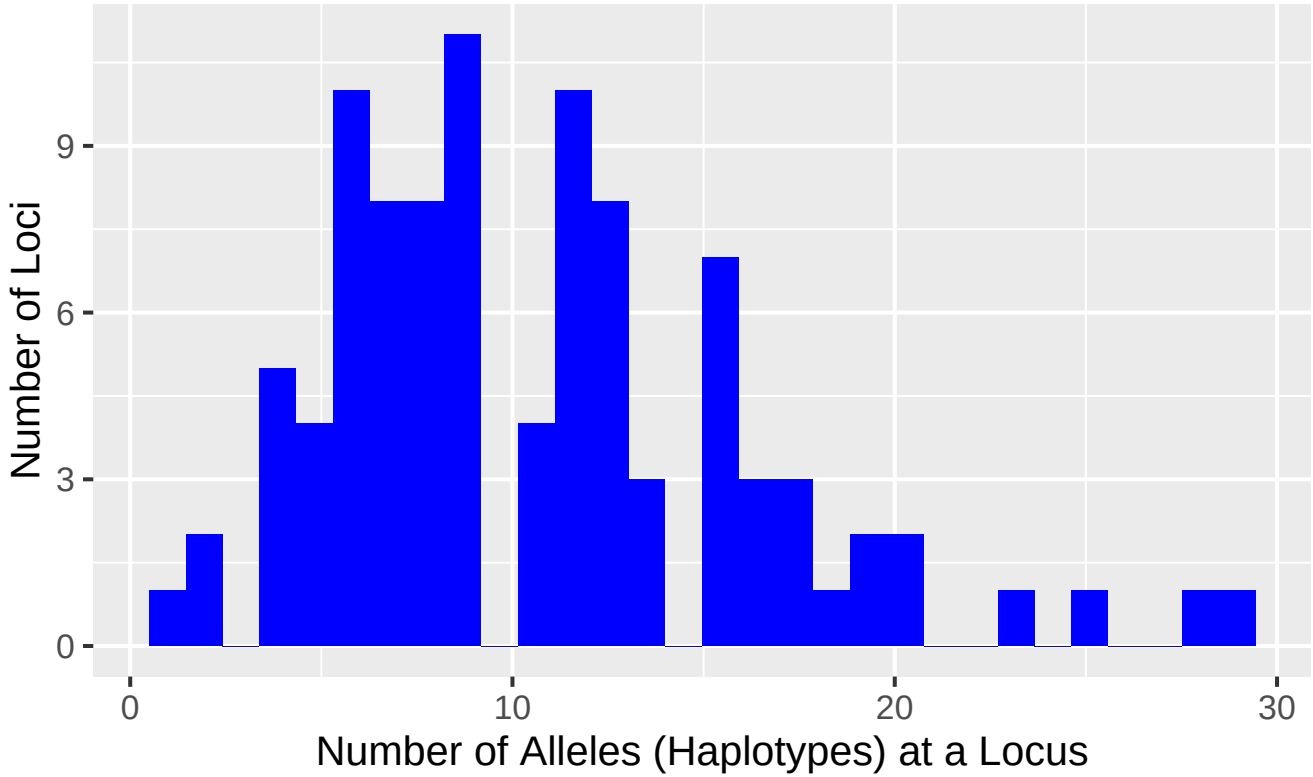

~/Shiny/haPLOType - Shiny
http://127.0.0.1:7515
Open in Browser
Publish
> field selection
+ read criteria
+ locus annotation
Group:
ALL
Indiv:
ALL
Locus:
Plate_2_A09_Sat_EW986980_consensus
Click on a microhaplotype pair at HW-plot for details:
haplotype
TTCCCT (10)
TACCTT (9)
TACCCT (8)
TACATT (7)
TACACT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

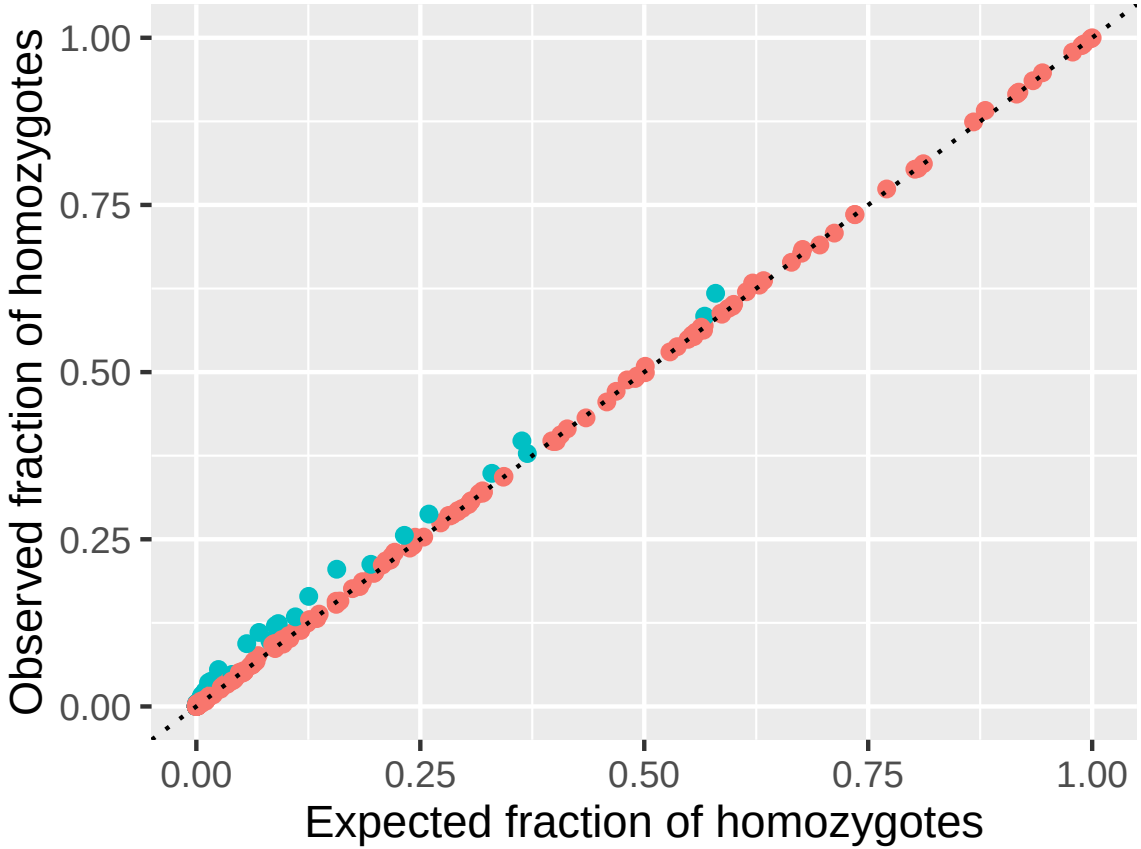
Across ≈15,000 individuals and at all 96 loci, called-allele read depths were as follows:

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

Expected-vs-observed rate of homozygotes...

Allele-specific Homozygosities from ≈ 6,000 kelp rockfish

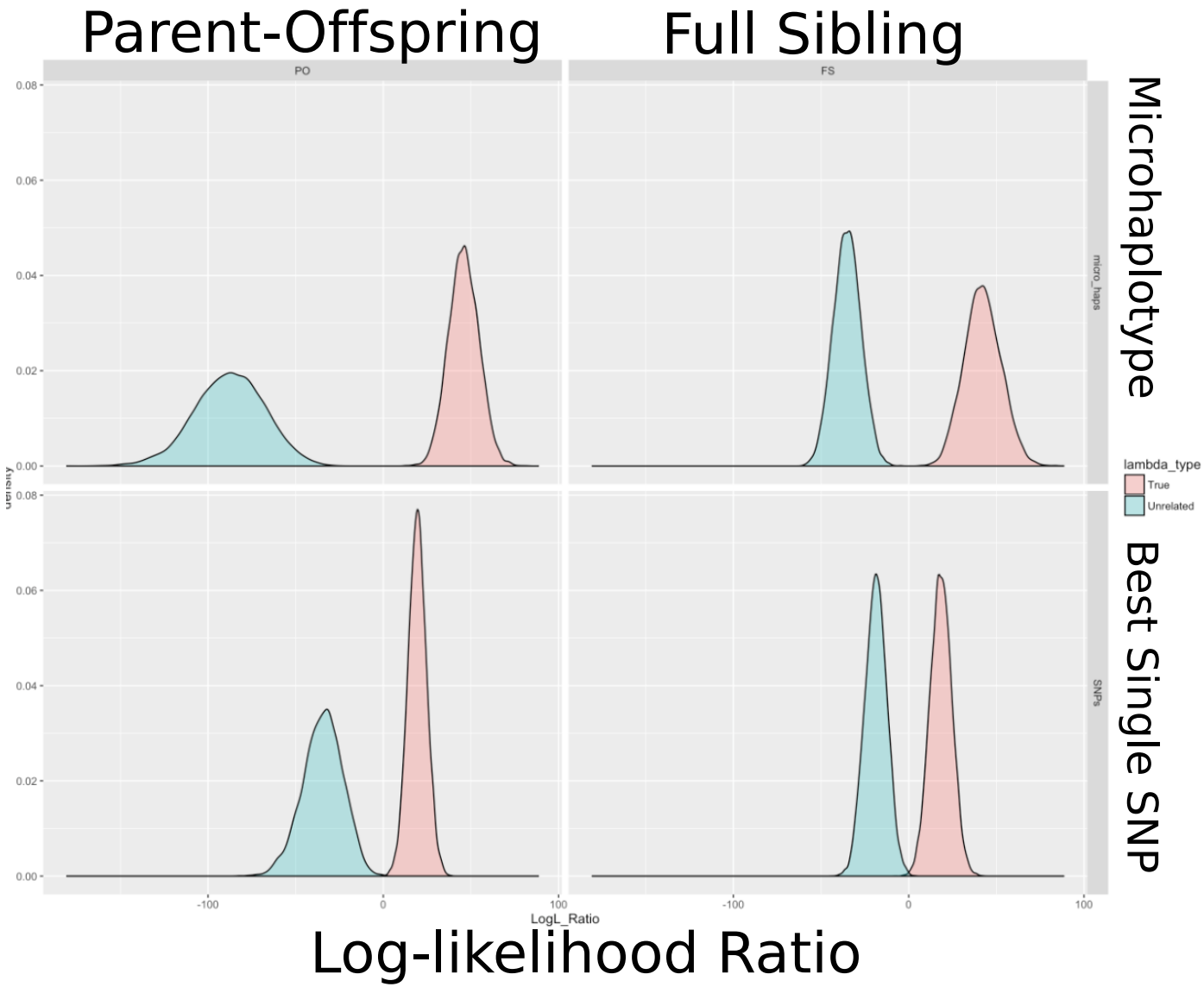
Green = 10 loci with null alleles. Orange = 86 remaining loci



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

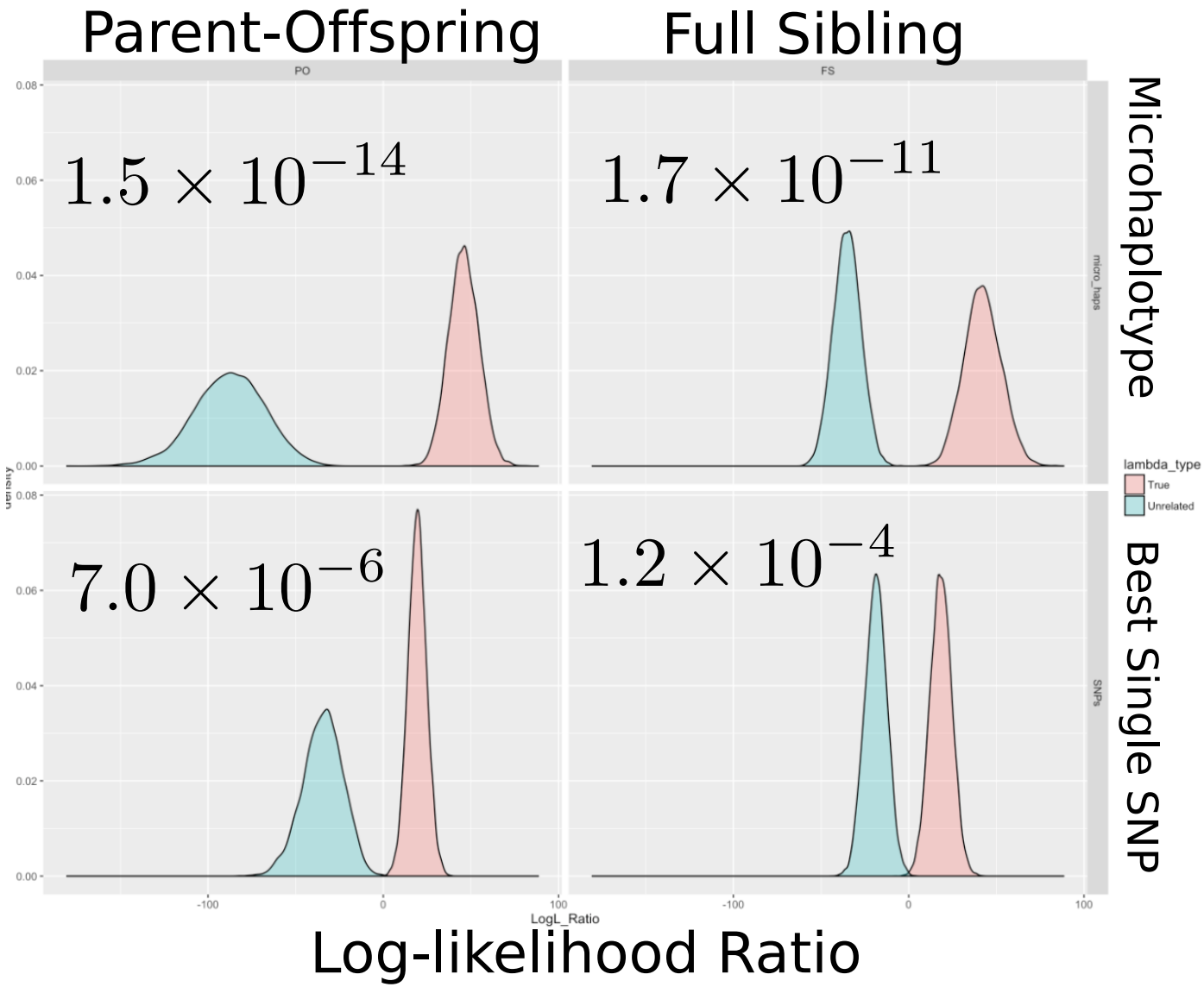
Log-likelihood ratio distribution

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)

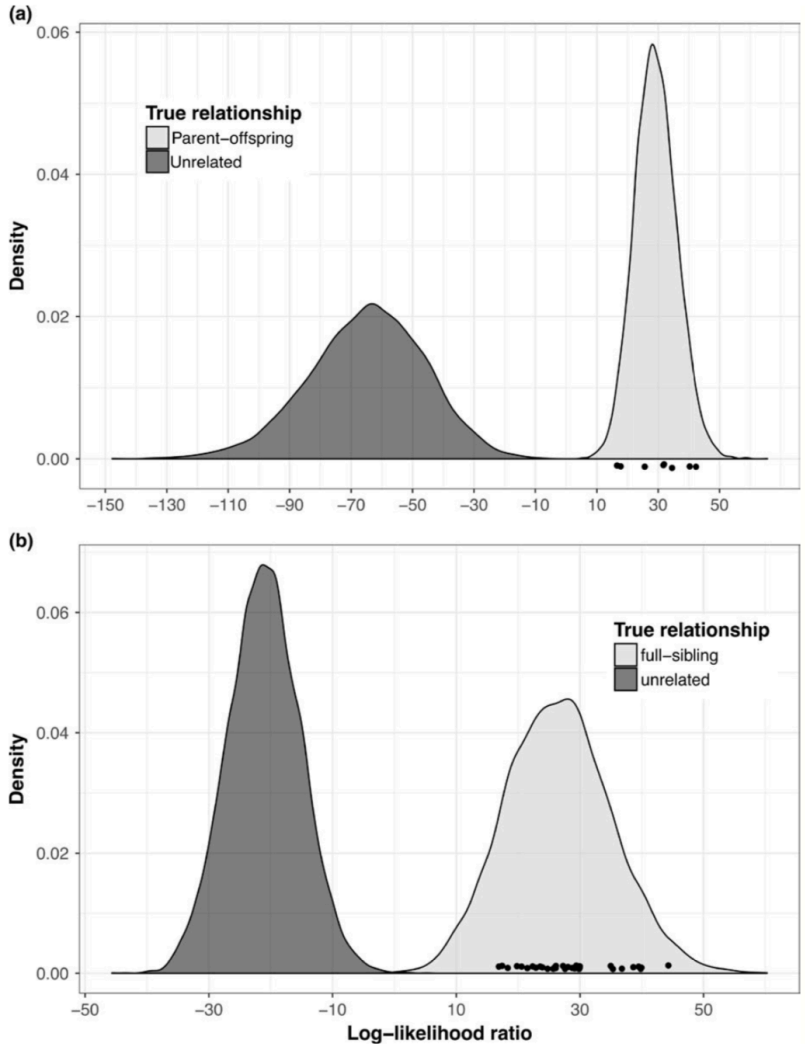
Statistical Science
2016, Vol. 31, No. 2, 259–274
DOI: 10.1214/16-STS552

Close-Kin Mark-Recapture

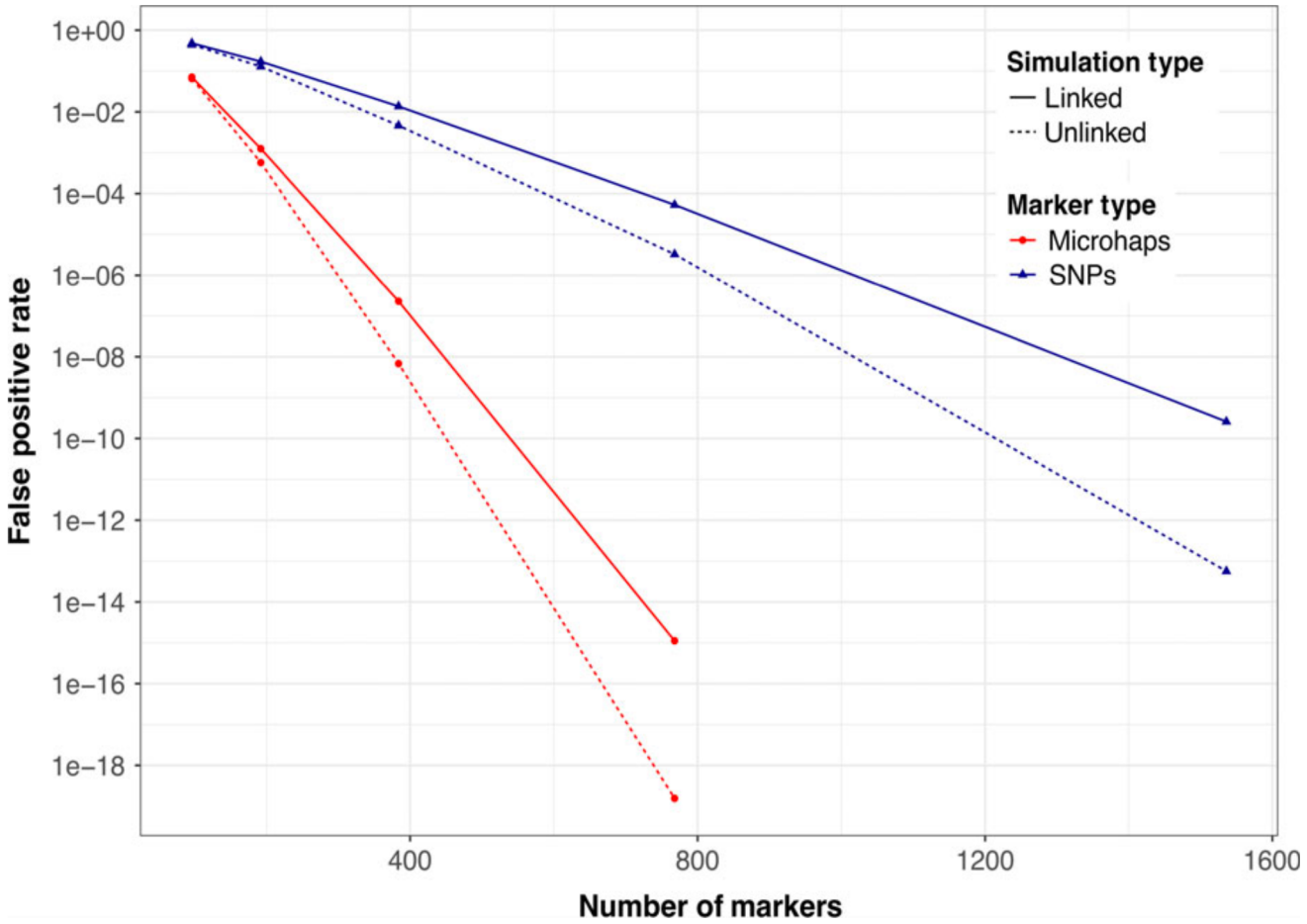
Mark V. Bravington, Hans J. Skaug and Eric C. Anderson

Relationship Inference in Baetscher et al. (2019)

- Kelp Rockfish: 1,912 Adults and 4,250 juveniles ($\approx$ 8.1 million pairs)
- 8 parent offspring pairs, showing evidence of dispersal between MPAs and from MPAs to unregulated zones.
- 25 pairs of full-siblings identified. Including 2 from different years.



The Prospects for Half-Sibling Identification



`microhaplot` overview of workflow

R package by Thomas Ng

- PCR amplify amplicons, or otherwise select a small portion of the genome (100 to 5000 chunks, each about 150 bp long) and sequence those regions in many barcoded individuals.
- Align those reads to a "genome," (which might even just be the sequences that you expect from the regions that you have amplified.)
- Call SNPs in those regions. Put those variant calls in a VCF file.
- Filter those variants into ones that you are confident about and create what we can call our "input VCF" file. This is the file that tells `microhaplot` at which positions it should look to find and record variants.

- Feed the input VCF and the SAM files (alignment files) into `microhaplot`. It then parses the SAMs and records the bases at each position that is named in the input VCF. Nucleotides that occur together on the same read are the "microhaplotypes."
  - Note: there is no statistical phasing going on here.
  - The tricky part in this is parsing the CIGAR string to figure out which positions in the reads correspond to the desired positions in the reference.
- Visualize, filter, interpret, and output data while using `microhaplot`
  - This is what we will do together today (we won't work so much with the previous bioinformatics steps, but, rather, we just want to familiarize people with the `microhaplot` interface.)
- This will not run on the ConGen RStudio Server, but it is all configured to run on the Remote Desktop. Full directions back at `https://eriqande.github.io/con-gen-2020/`