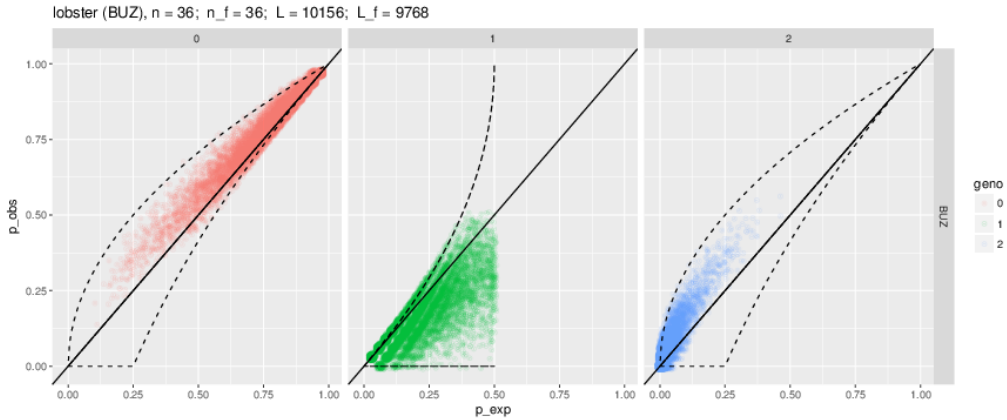
Estimates of genotyping error from published reduced representation genotype-by-sequence data

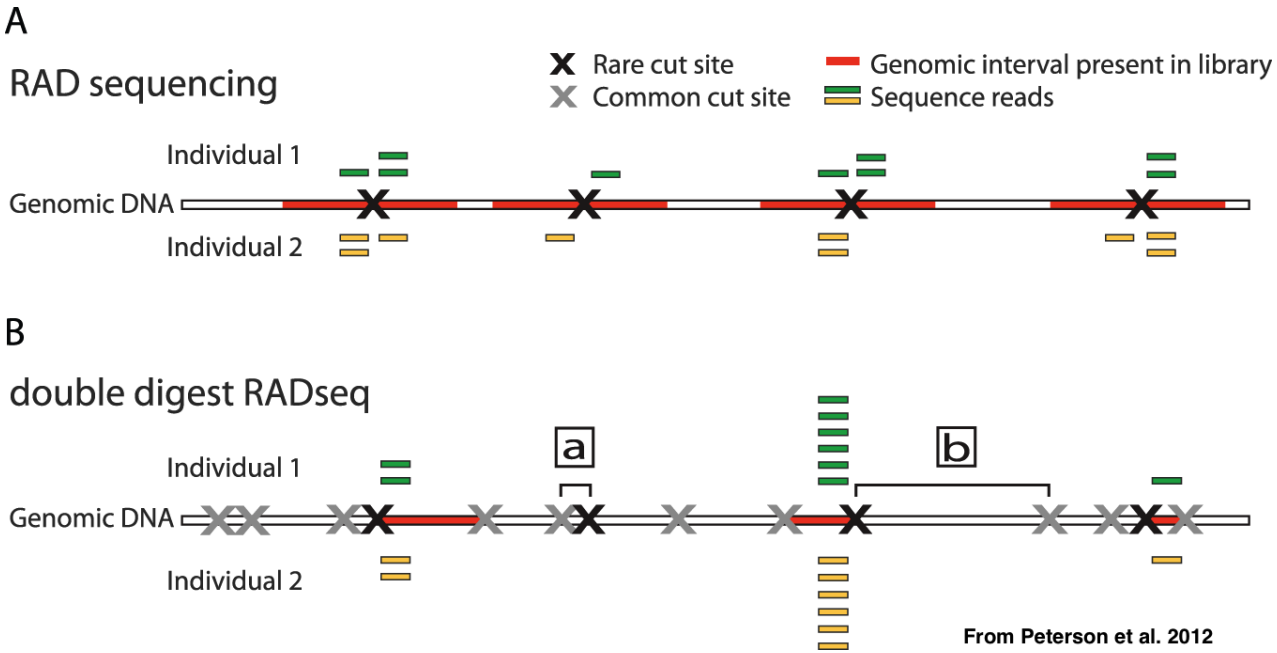
Eric C. Anderson

Fisheries Ecology Division, Southwest Fisheries Science Center, Santa Cruz



Bay Area Population Genomics Meeting
21 APR 2018

A Variety of RAD-seq Methods



- Can genotype many individuals
- No genome needed(*) for non-model organisms
- High read depths should provide accuracy of genotypes. (Julian Catchen says, "$> 10X$ for heterozygotes").

RAD-bashing

There have been some criticisms. . .

Allelic dropout / null alleles and other biases:

RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling

B. ARNOLD,[1] R. B. CORBETT-DETIG,[1] D. HARTL and K. BOMBLIES
Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

The effect of RAD allele dropout on the estimation of genetic variation within and between populations

MATHIEU GAUTIER,* KARIM GHARBI,† TIMOTHEE CEZARD,† JULIEN FOUCAUD,* CAROLE KERDELHUÉ,* PIERRE PUDLO,*‡ JEAN-MARIE CORNUET* and ARNAUD ESTOUP*

Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences

HUATENG HUANG* AND L. LACEY KNOWLES

Insufficient genome extent:

OPINION

Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation

DAVID B. LOWRY,*† SEAN HOBAN,‡§ JOANNA L. KELLEY,¶ KATIE E. LOTTERHOS,** LAURA K. REED,†† MICHAEL F. ANTOLIN‡‡ and ANDREW STORFER¶

RAD "Genotyping Accuracy" Studies

Typically explorations of different bioinformatic settings and filters

Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference

A. MASTRETTA-YANES,* N. ARRIGO,† N. ALVAREZ,† T. H. JORGENSEN,‡ D. PIÑERO§ and B. C. EMERSON*¶

Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8**, 907–917 doi: 10.1111/2041-210X.12700

Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference

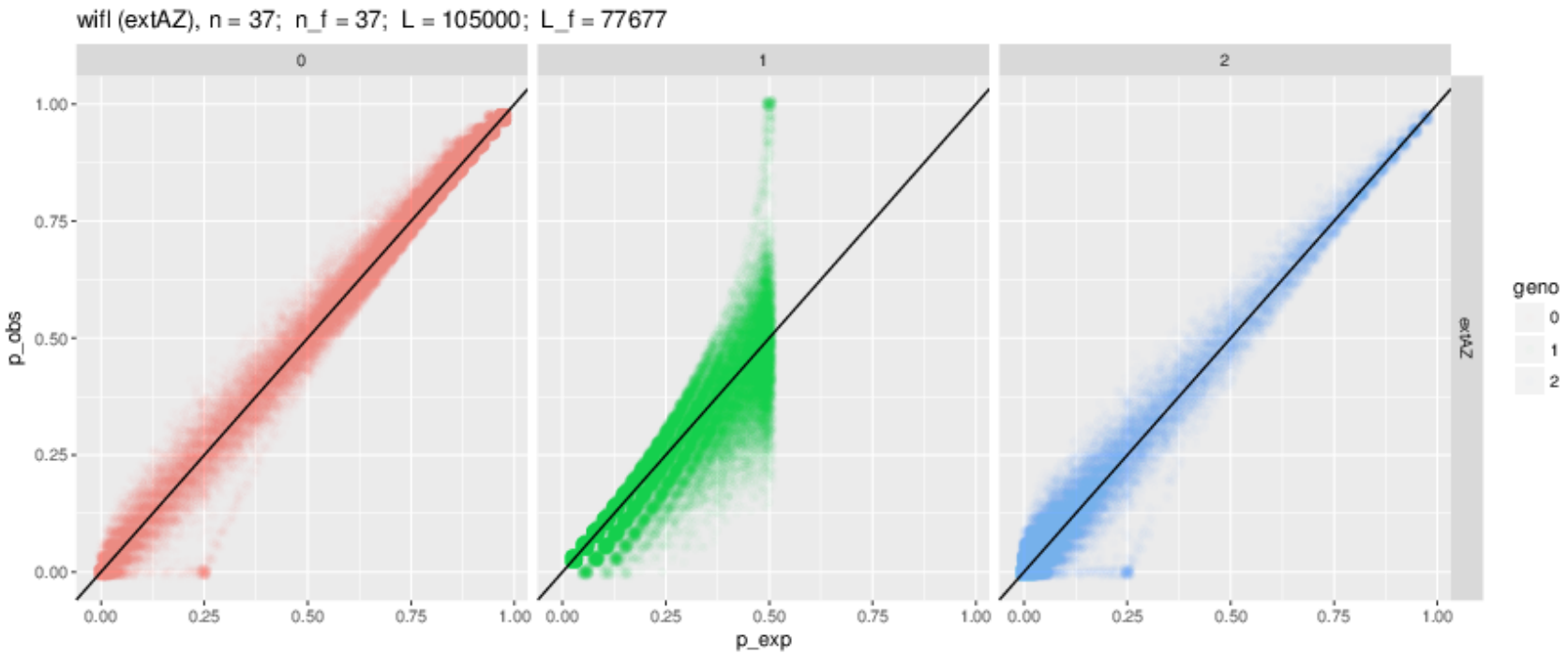
Aaron B. A. Shafer†,[1,2], Claire R. Peart†,[1], Sergio Tusso[1], Inbar Maayan[1], Alan Brelsford[3], Christopher W. Wheat[4] and Jochen B. W. Wolf*,[1,5]

- Some nice studies
- A whole lot of computation
- Not much immediate feedback for individual RAD users or the question of, "How we doin' here?"

A Simple Visualization from Called RAD Genotypes

Some willow flycatcher data I was working with

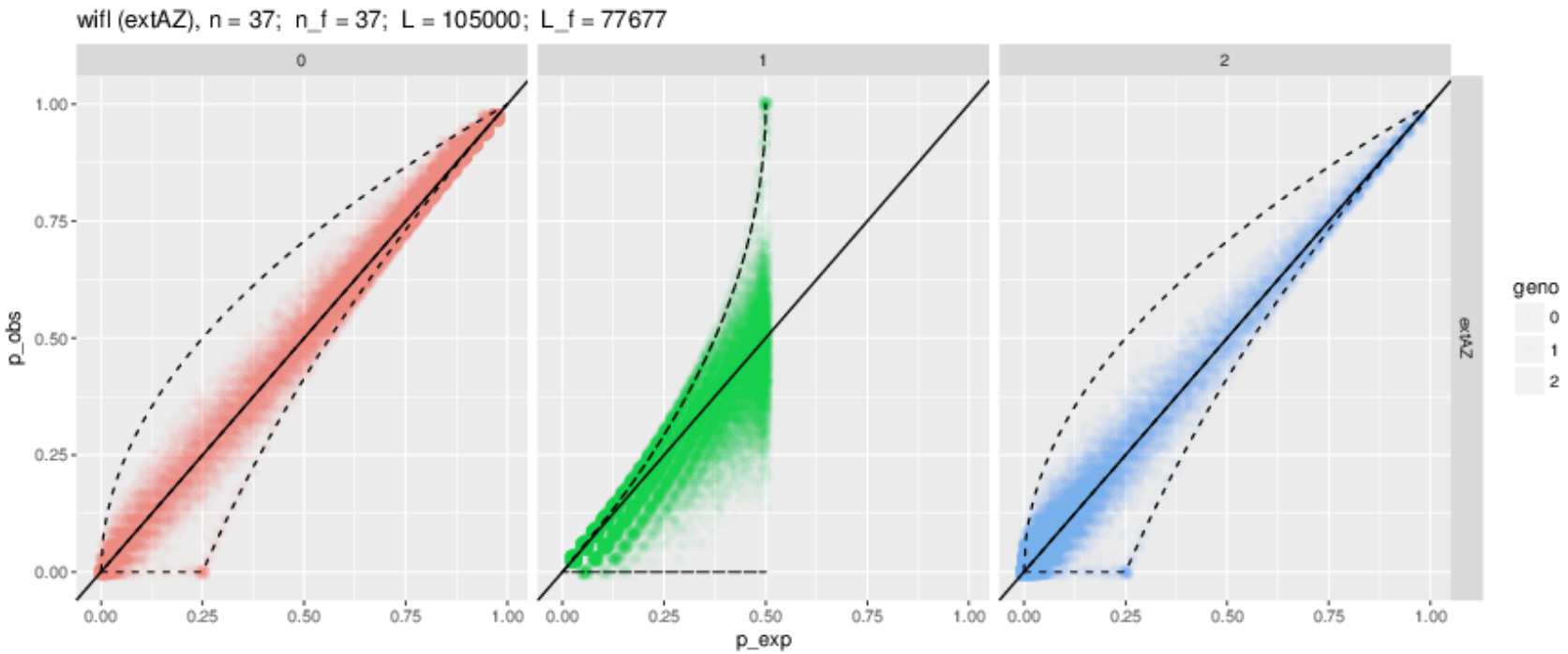
Just plot the observed frequency of genotypes against their expected frequency given the allele frequencies and Hardy Weinberg Equilibrium



A Simple Visualization from Called RAD Genotypes

Some willow flycatcher data I was working with

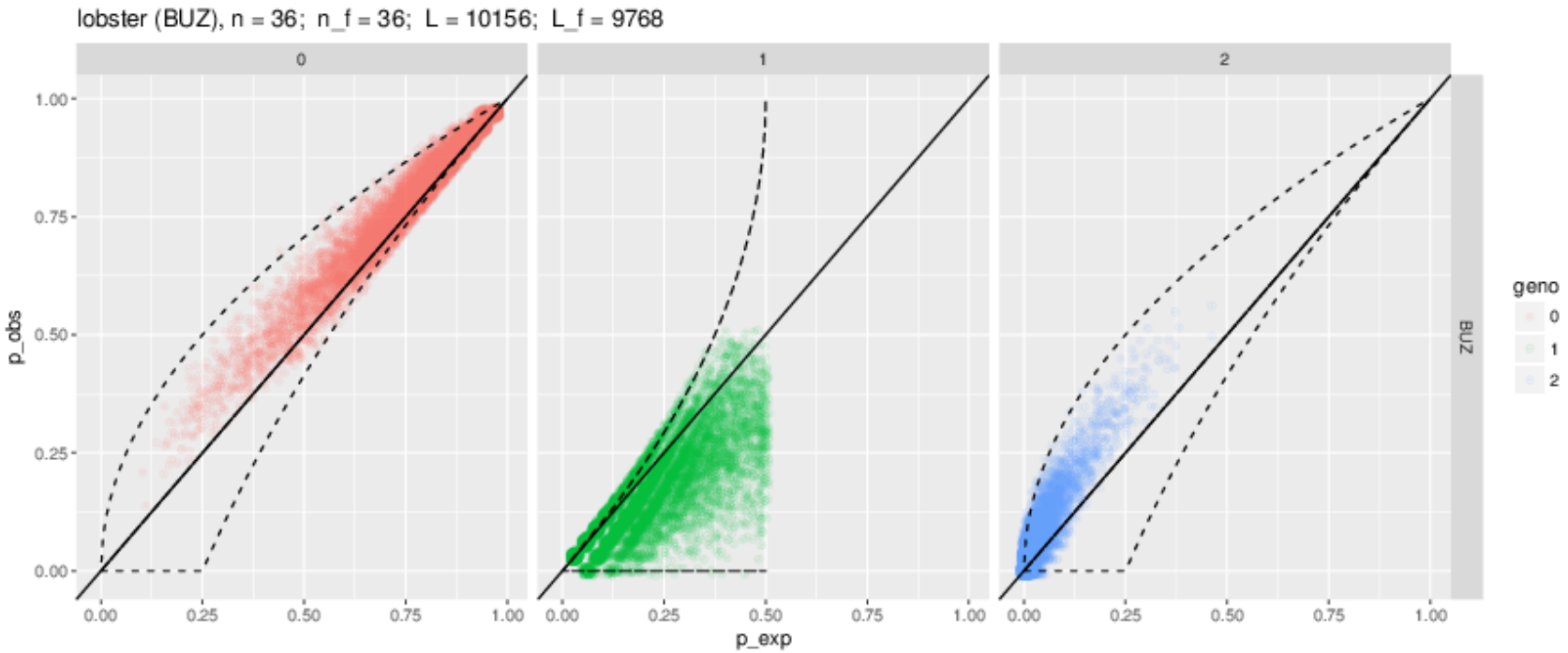
Just plot the observed frequency of genotypes against their expected frequency given the allele frequencies and Hardy Weinberg Equilibrium



Things don't always look so peachy

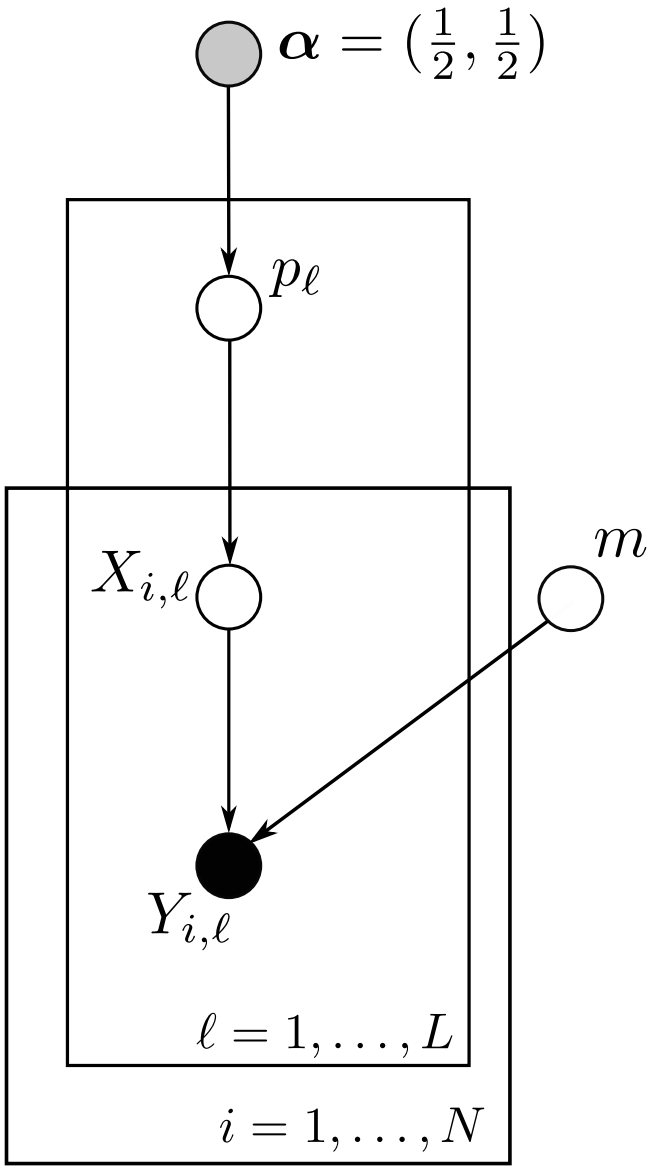
Published data from lobsters (Benestan et al. 2015)

This is a sample of individuals that were all collected from the same place that should have been in HWE. . .



A Genotyping Error Model

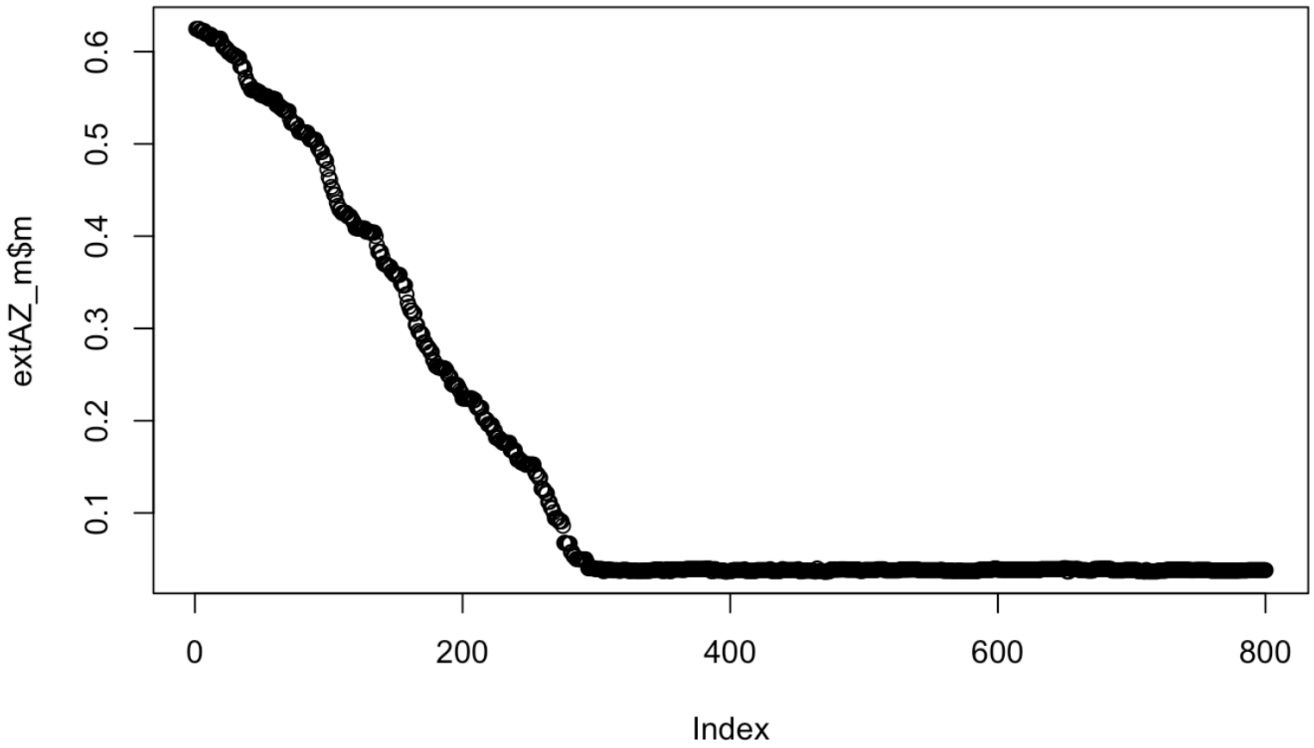
How much error must there be to look that bad?



- α beta prior parameters for allele frequencies
- p_ℓ unknown frequency of alternate allele at SNP ℓ
- $X_{i,\ell}$ true, underlying, but unobserved, genotype of individual *i* at SNP ℓ
- $Y_{i,\ell}$ the observed (called/scored) but possibly incorrect genotype of individual *i* at SNP ℓ
- m the rate at which true heterozygotes are incorrectly called as homozygotes

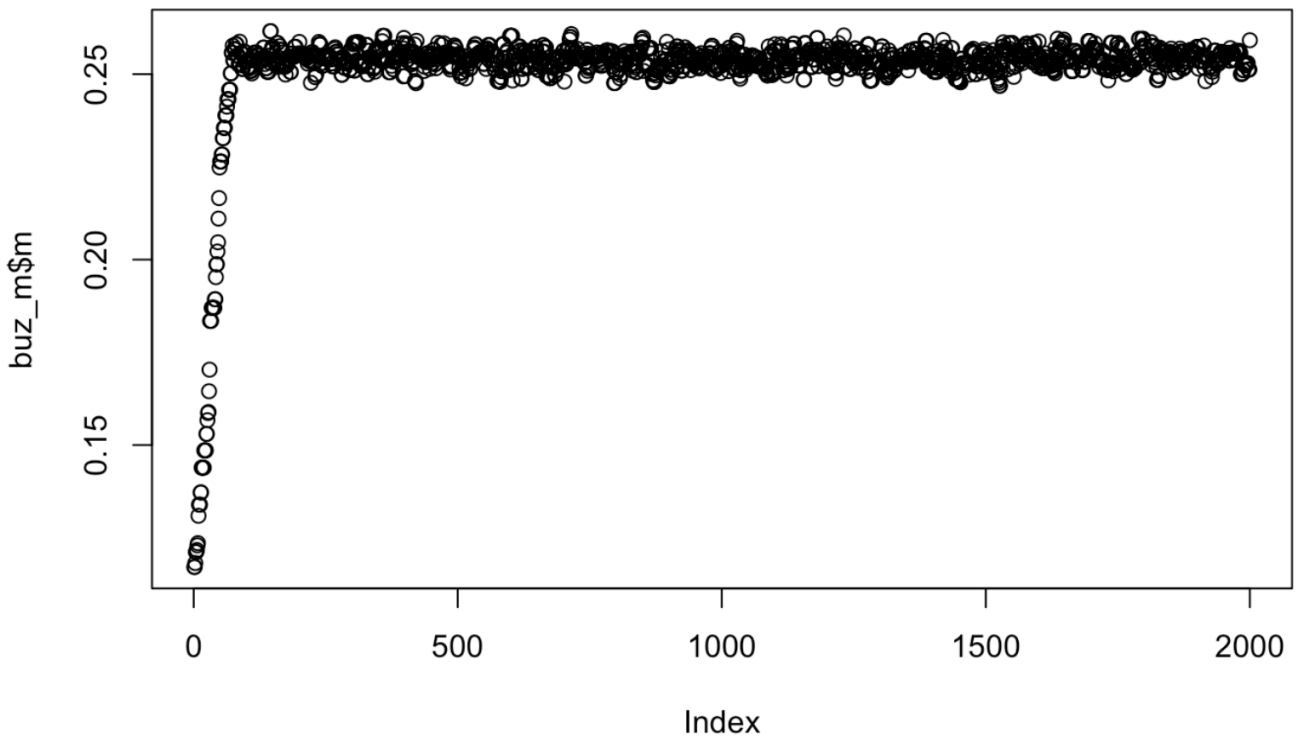
Estimating *m* via MCMC

Willow Flycatchers



Estimating *m* via MCMC

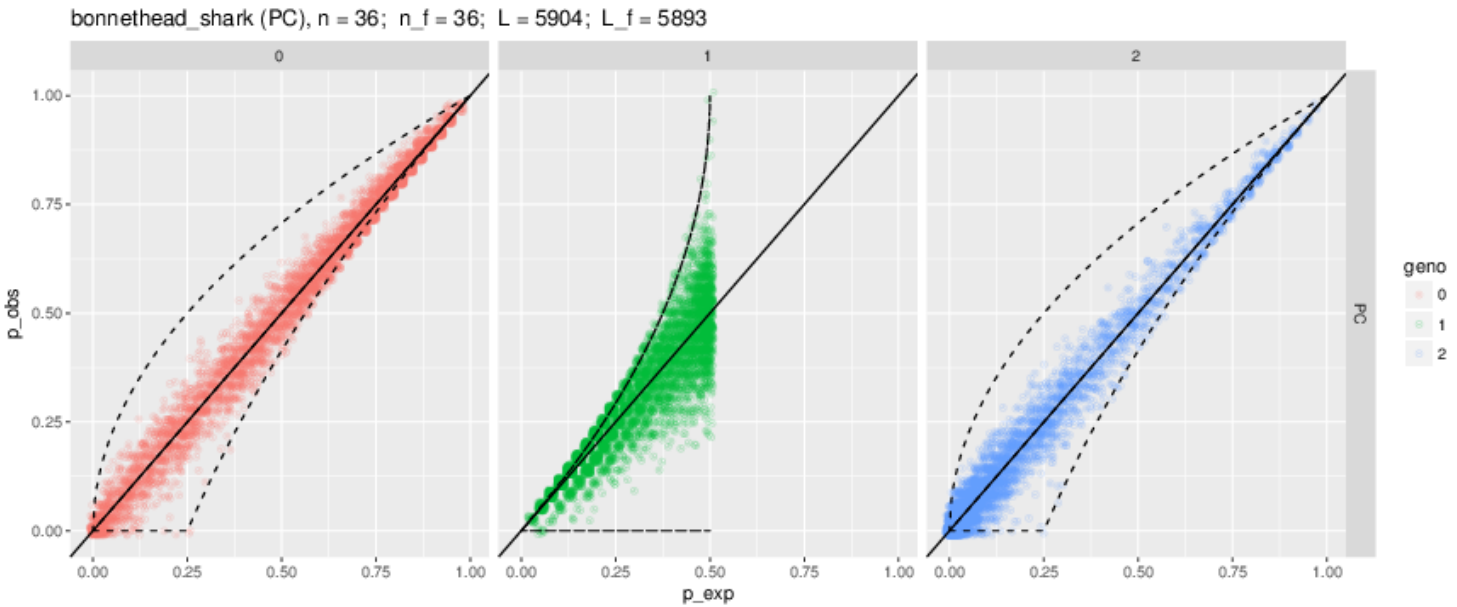
Lobsters



Holy Moly!

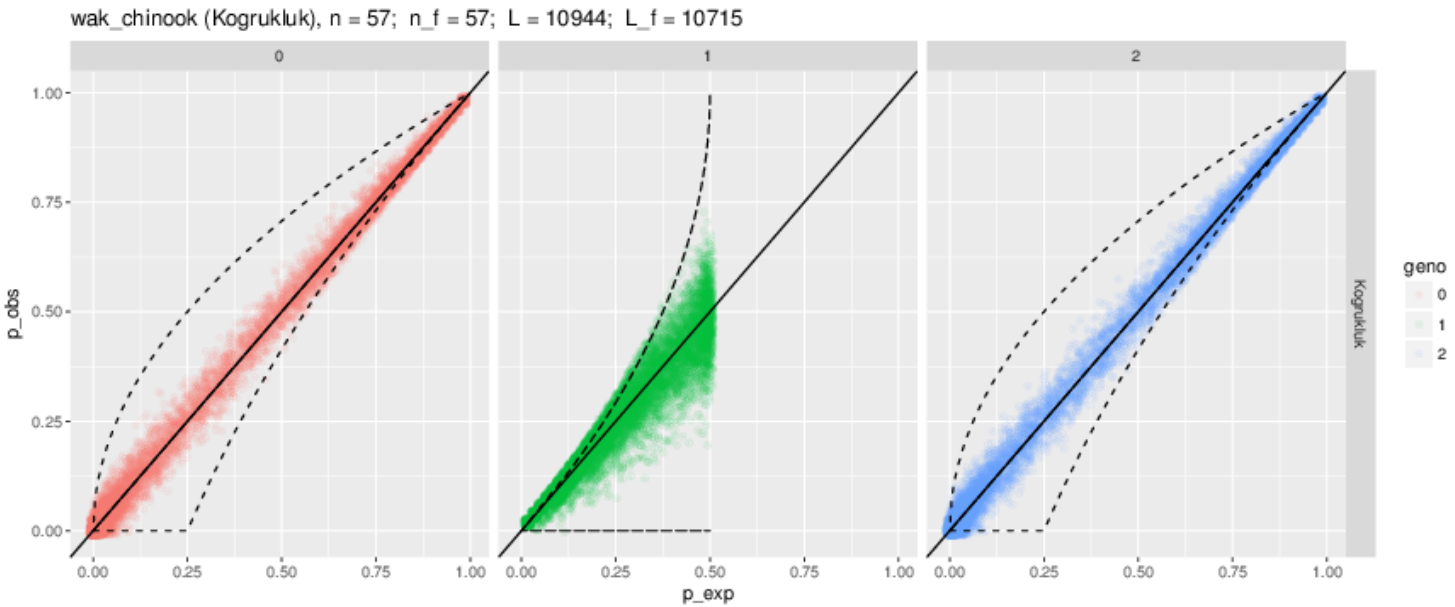
A Survey of Some Published Data Sets

Bonnethead Shark $\hat{m} = 0.01$



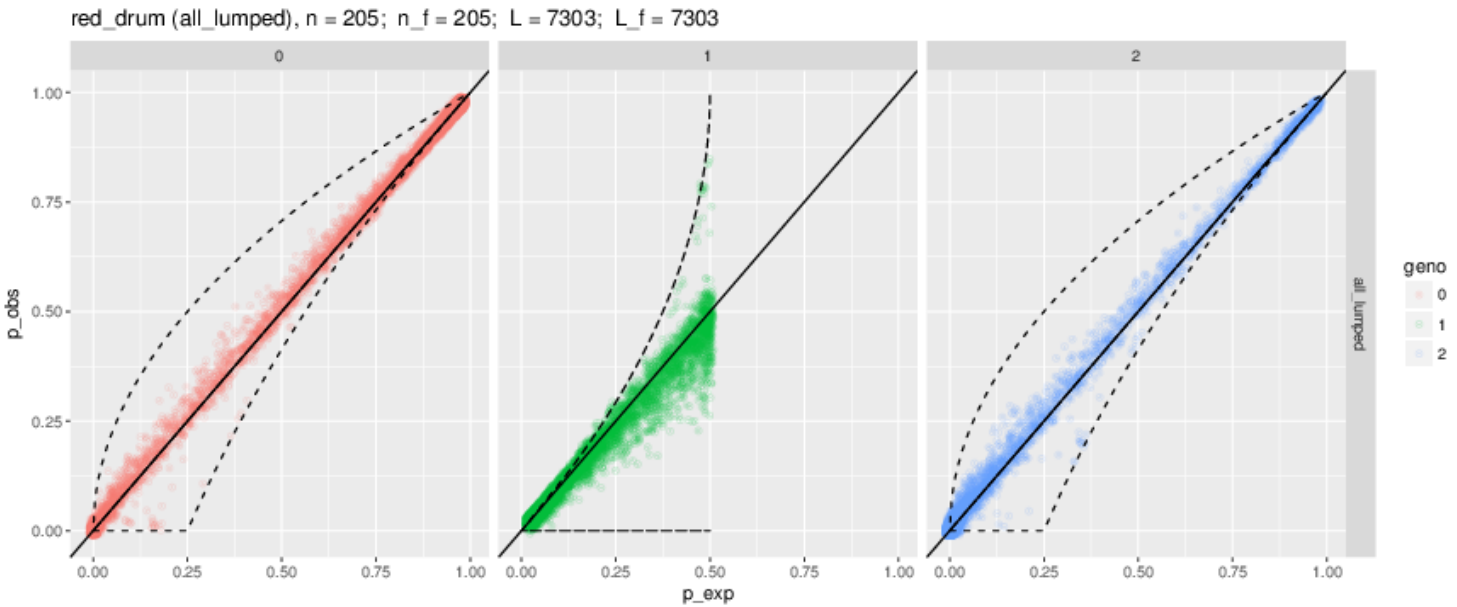
A Survey of Some Published Data Sets

Western Alaska Chinook $\hat{m} = 0.02$



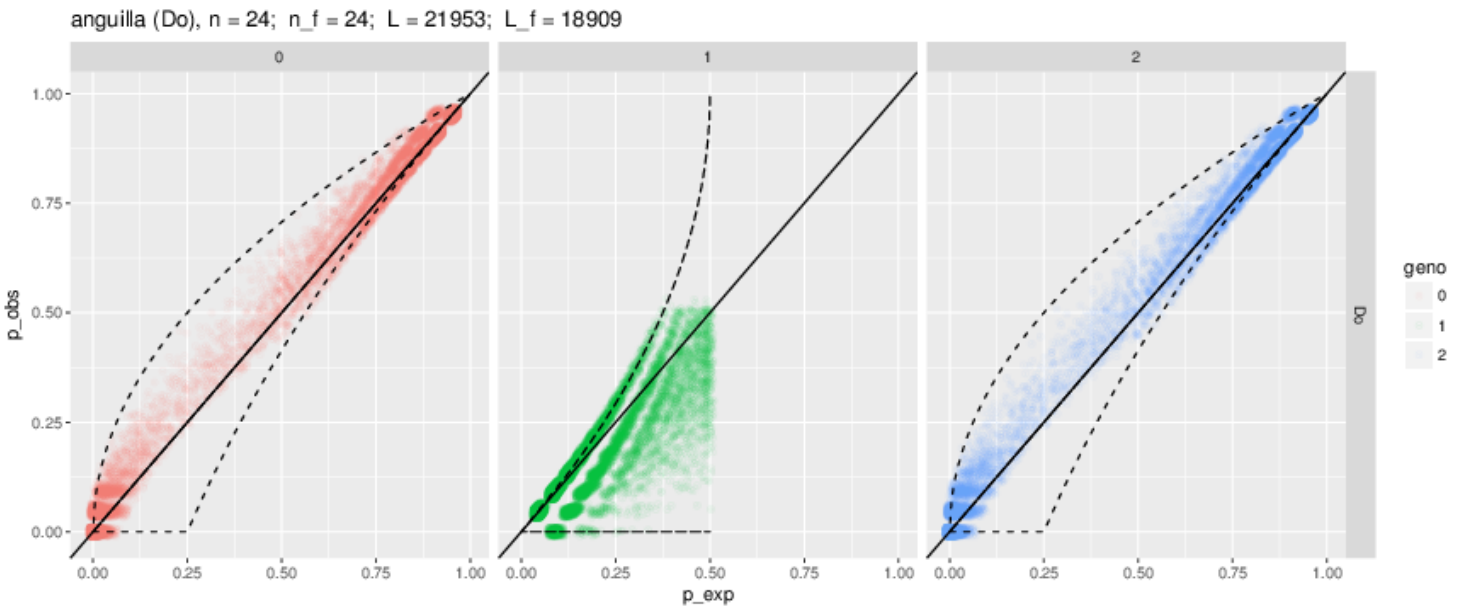
A Survey of Some Published Data Sets

Red Drum $\hat{m} = 0.05$



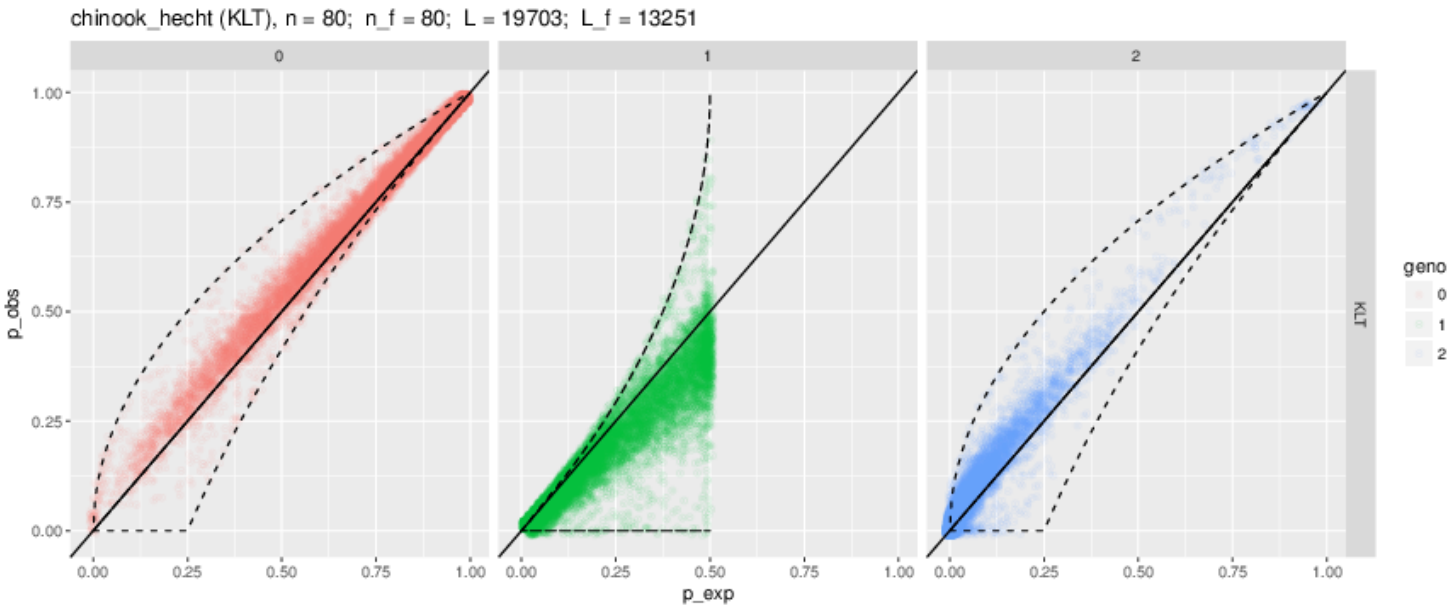
A Survey of Some Published Data Sets

Anguilla $\hat{m} = 0.14$



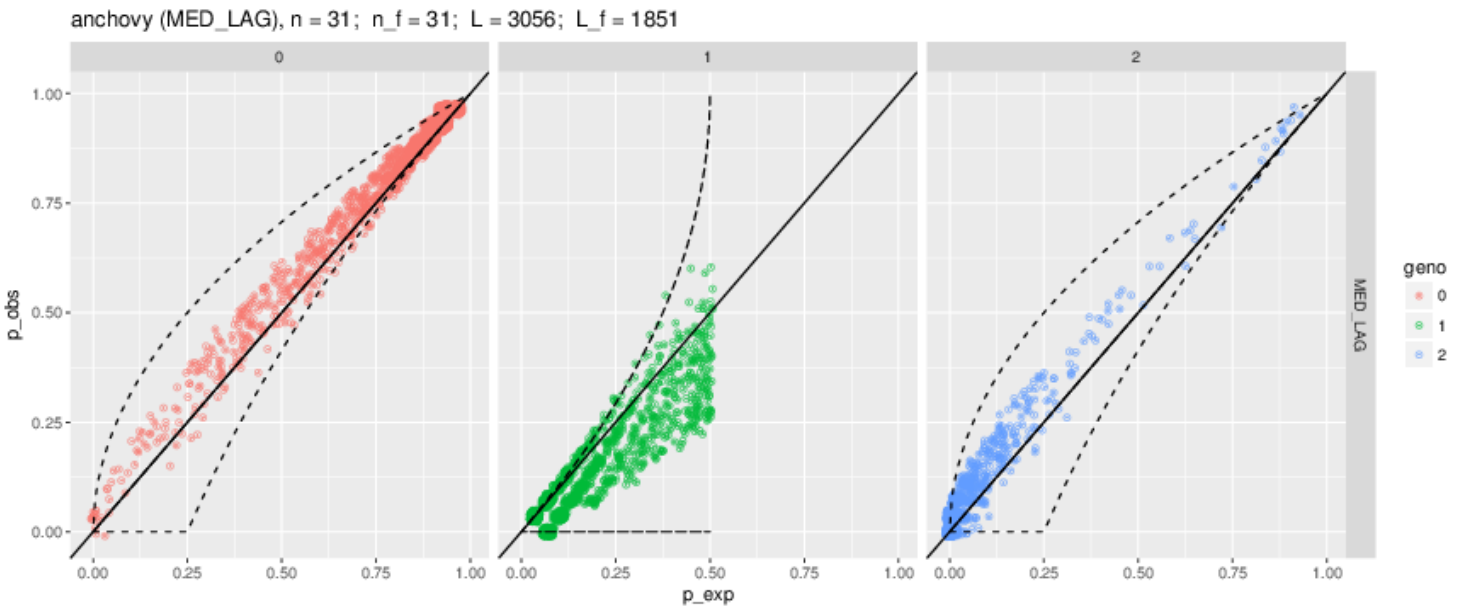
A Survey of Some Published Data Sets

Columbia River Chinook $\hat{m} = 0.17$



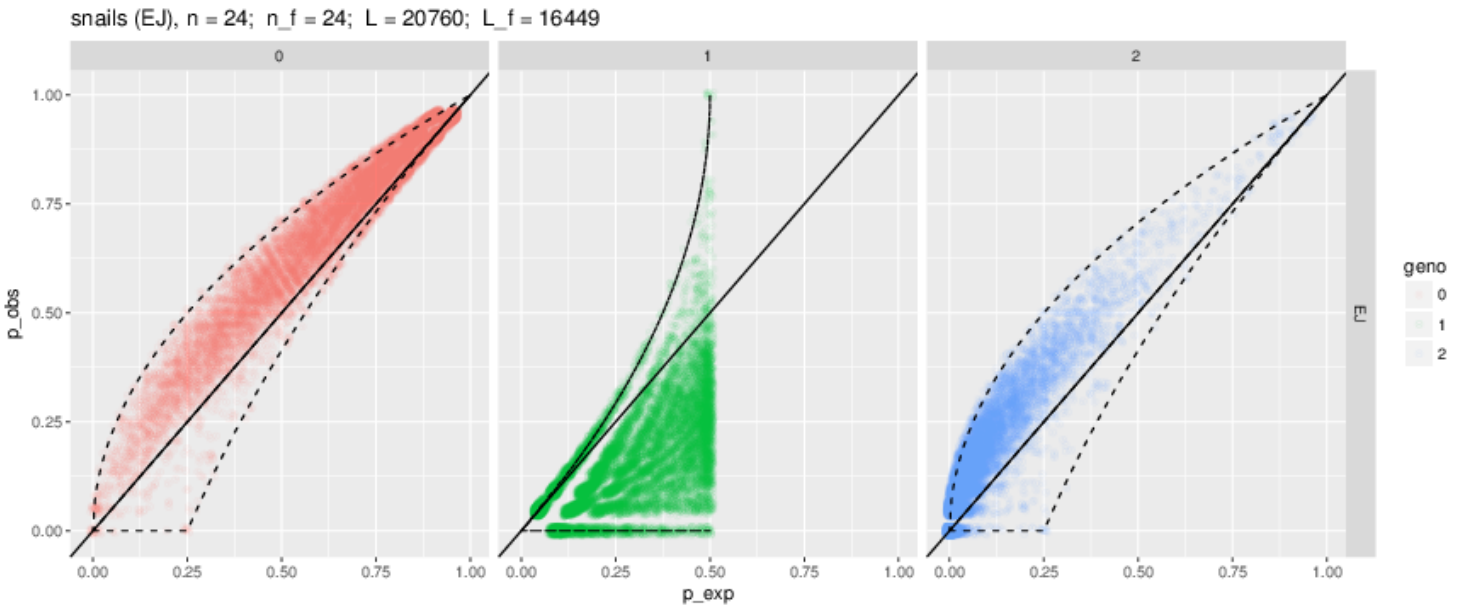
A Survey of Some Published Data Sets

Anchovy $\hat{m} = 0.28$



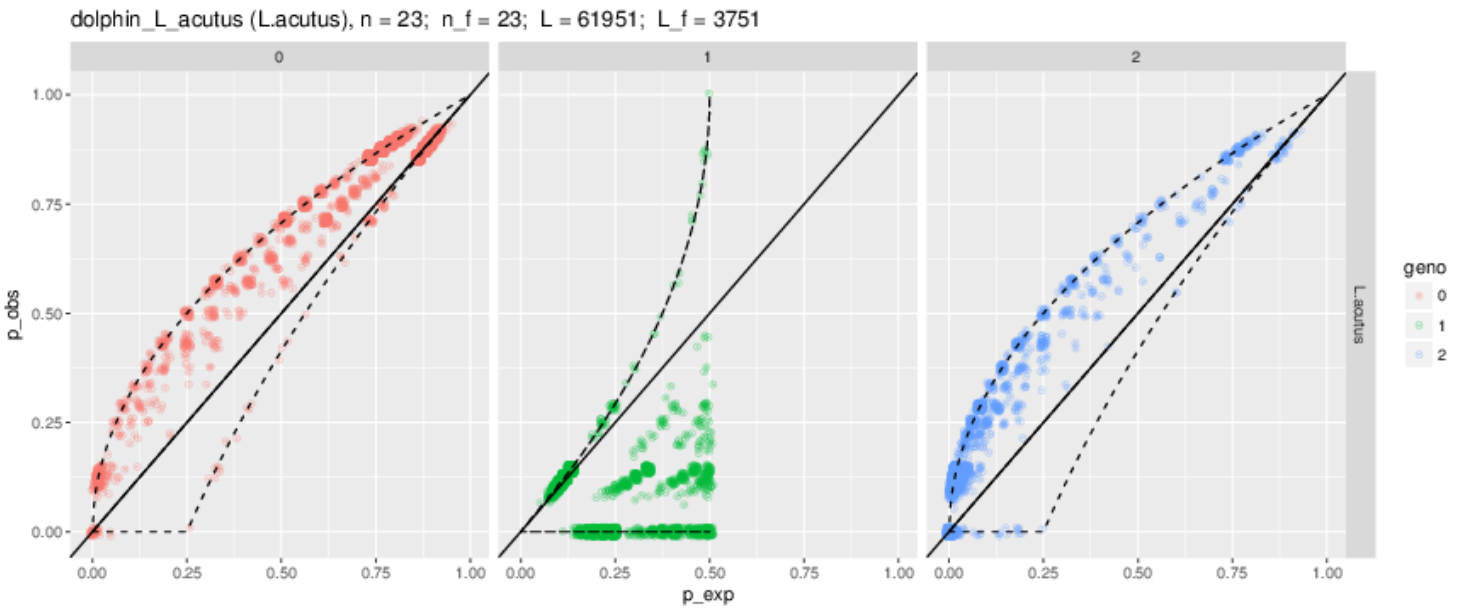
A Survey of Some Published Data Sets

Snails $\hat{m} = 0.45$



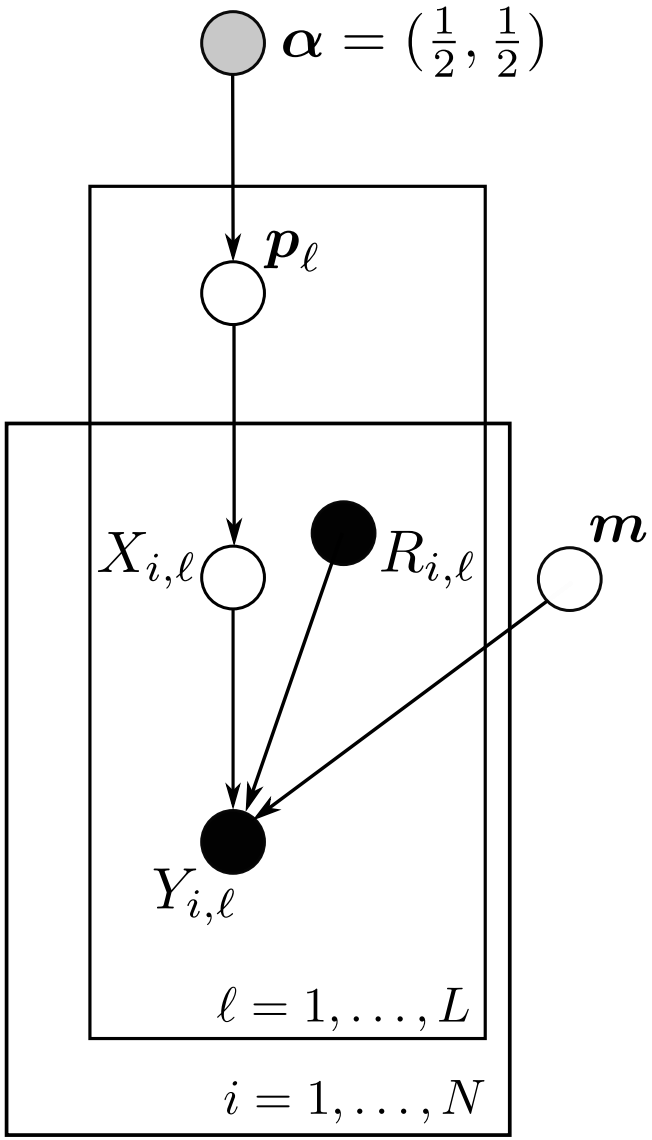
A Survey of Some Published Data Sets

Dolphin $\hat{m} = 0.72$



Include Read Depth in the Genotyping Error Model

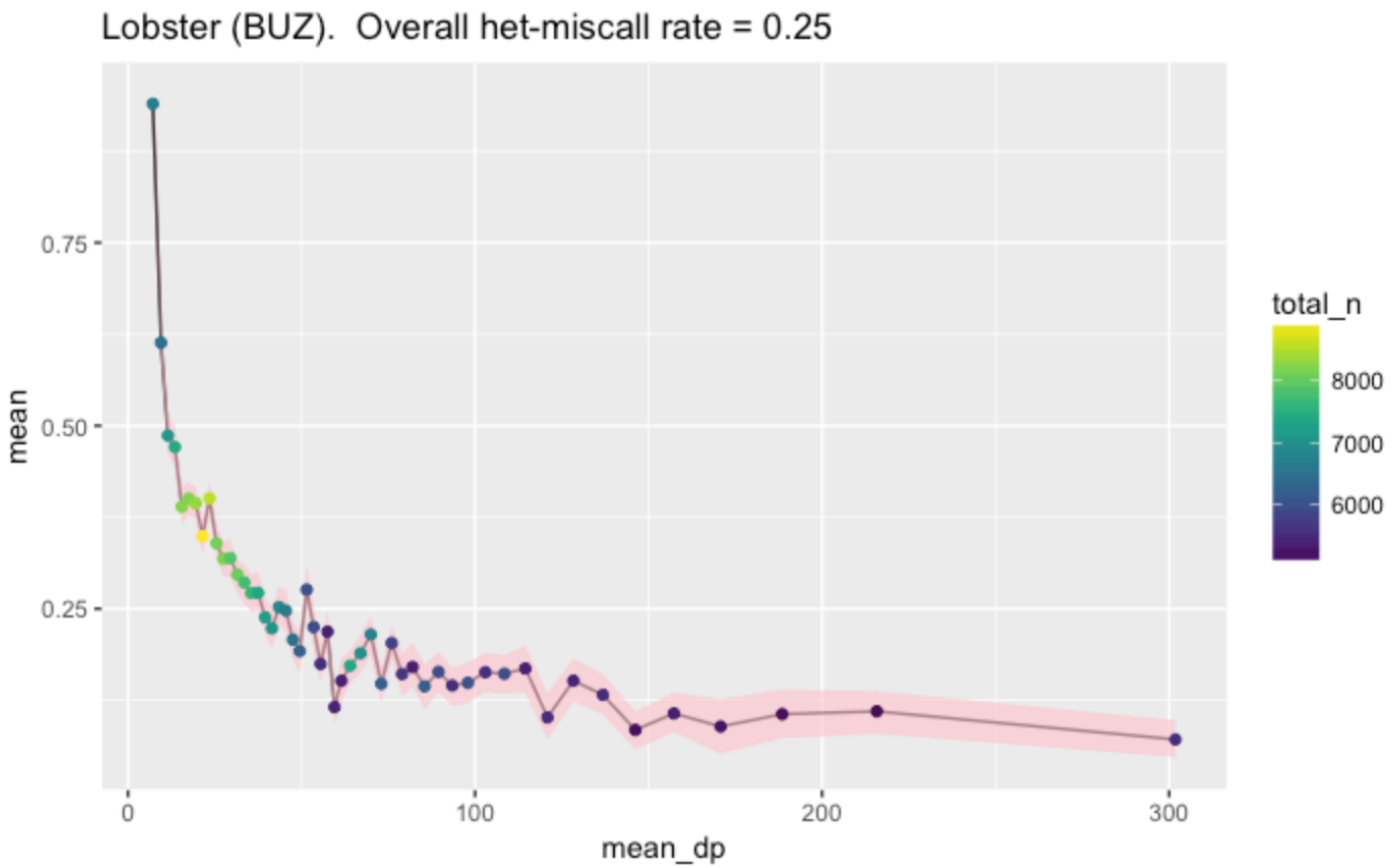
Is this error mostly a consequence of inaccuracy at low read depths?



- $R_{i,\ell}$ the read-depth category or bin of the ℓ^{th} SNP in the i^{th} individual.
- $\boldsymbol{m}$ This is now a vector—a separate heterozygote miscall rate for each read depth category.

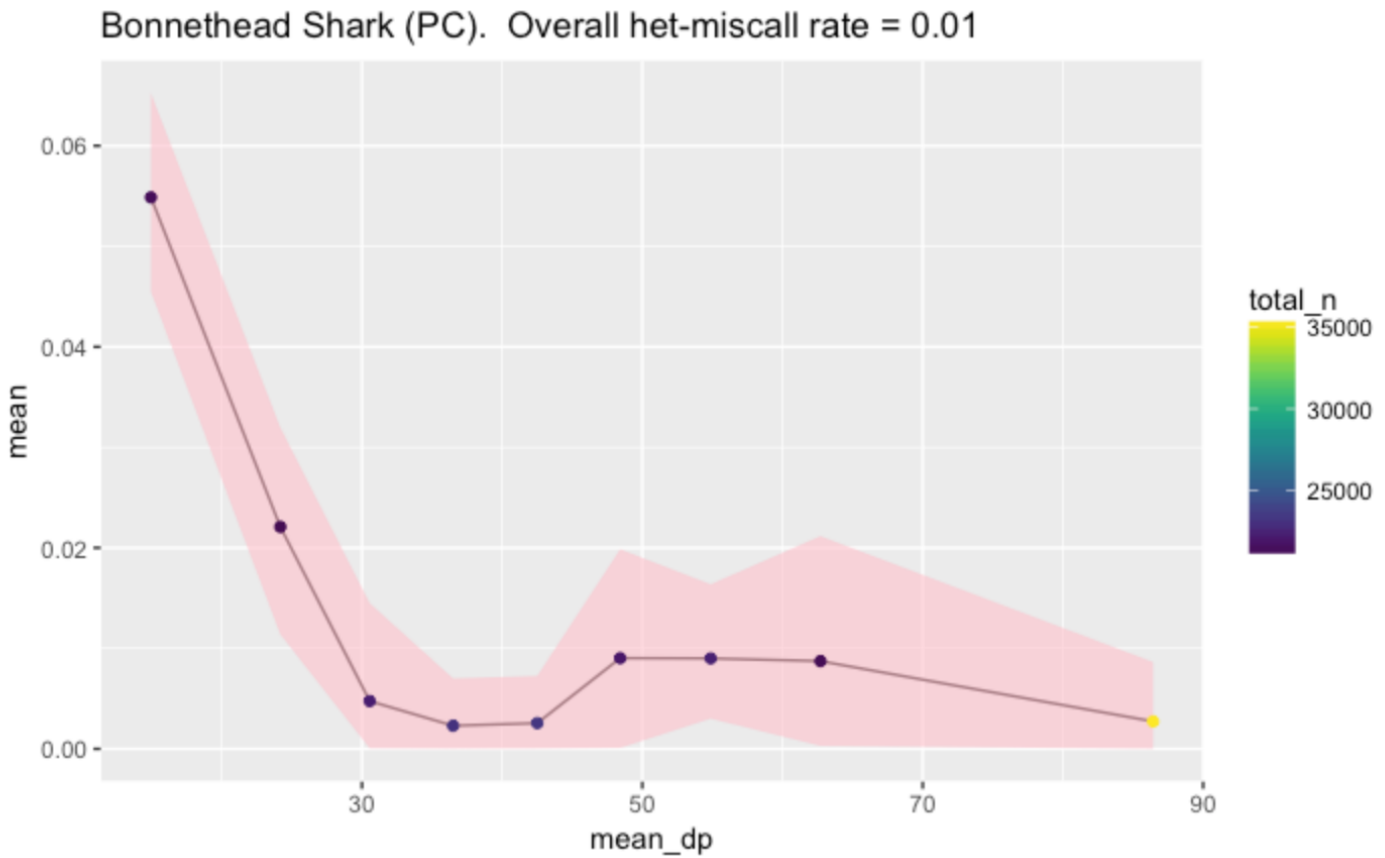
Het Miscall Rate Higher at Low Read Depth

Lobster overall $\hat{m} = 0.25$



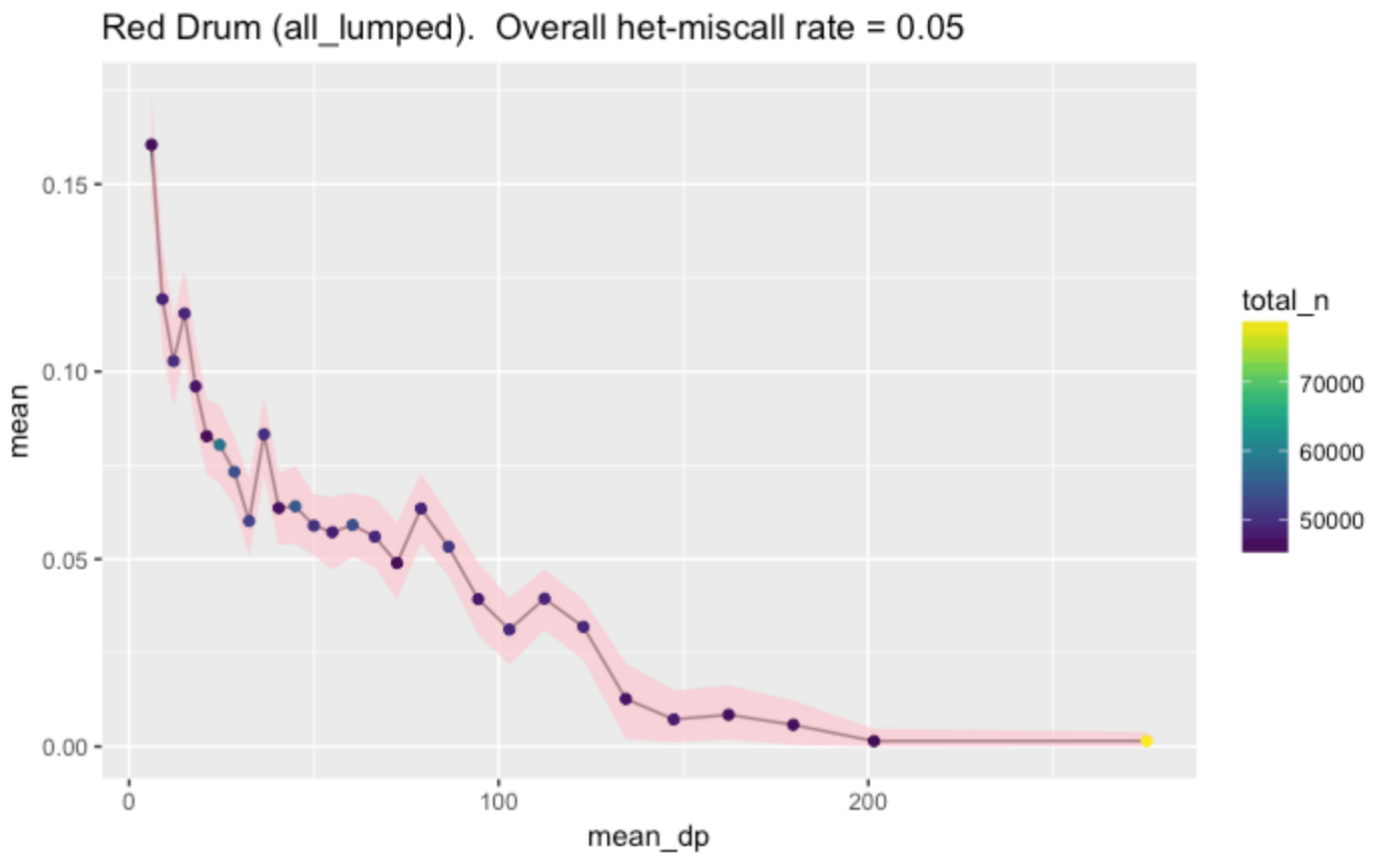
This Trend Seen Even in "High Accuracy" Data Sets

Bonnethead Shark overall $\hat{m} = 0.01$



This Trend Seen Even in "High Accuracy" Data Sets

Red Drum overall $\hat{m} = 0.05$



Wrap Up

- Working on a paper with Gordon Luikart and Thierry Gosselin doing a more complete survey.
- While RAD suffers some known biases, there are also problems with straight-up genotyping error.
- A primary driver seems to be insufficient read depth to call heterozygotes
- Effects on downstream analysis depend on what you are doing:
  - Allele frequency estimation (not too bad)
  - Relationship inference (disastrous)
  - Identification of F_{ST} outliers (potentially problematic)
  - etc.
- Seems to be begging for a probabilistic genotype calling approach, incorporating a prior on genotype frequencies