

How different sources of error affect the accuracy of genetic stock identification

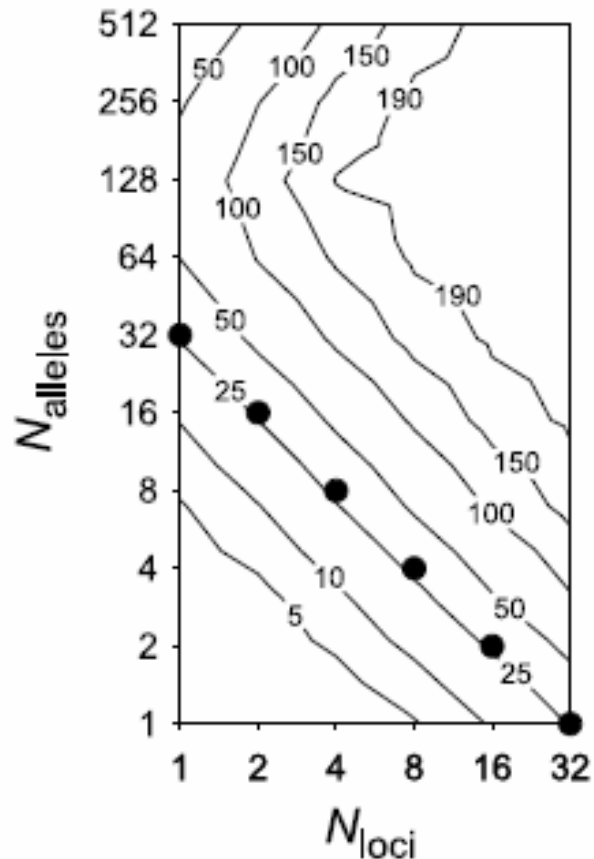
Steven Kalinowski, Kezia Manlove, Mark Taper

Department of Ecology

Montana State University



How do we determine the most effective methods for accurate GSI?



Kalinowski (2004) used completely simulated data to predict best ways to perform GSI

Now that more empirical data is available, it should be used for GSI study design

Consider some hypothetical GSI mixture estimates:

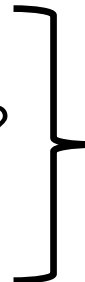
Stock	ACTUAL proportion in fishery	ESTIMATED proportion
West Vancouver	40%	33%
Fraser River	10%	8%
WA Coast	10%	13%
Columbia River	15%	12%
Etc.		

What causes this error?

Too few fish sampled from fishery?

Not enough loci genotyped

Small baseline sample sizes?

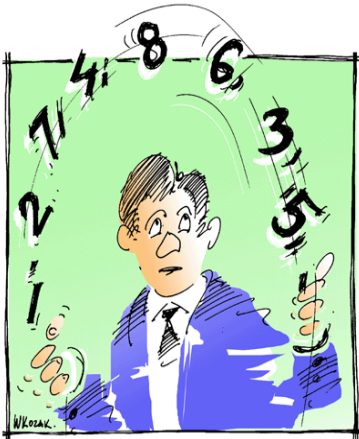


Answering these questions would help us reduce error

Sources of GSI Estimation Error

Source of Error	Currently in model?
Non random sampling of fishery	No
Random sampling of fishery	Yes
Genotyping a finite number of loci	Yes
Genotyping error	No
Sampling individuals from baseline populations while estimating allele frequencies	Yes
Contributing populations not in baseline	No

Talk Outline



1. Sources of GSI Error
2. Error decomposition
 - Statistics
 - More statistics
3. Results
4. Discussion: What next?

Expected Squared Error:

A convenient measure of how much estimates are expected to be wrong

$$ESE(\hat{\theta}_i) = E \left[(\theta_i - \hat{\theta}_i)^2 \right]$$

Like a variance, but includes effect of bias.

Our goal:

Partition ESE into 3 components

$$ESE(\hat{\theta}_i)_{total} = ESE(\hat{\theta}_i)_{fishery} + ESE(\hat{\theta}_i)_{genotypic} + ESE(\hat{\theta}_i)_{baseline}$$

Knowing the relative magnitude of each error would be valuable

$$ESE(\hat{\theta}_i)_{total} = ESE(\hat{\theta}_i)_{fishery} + ESE(\hat{\theta}_i)_{genotypic} + ESE(\hat{\theta}_i)_{baseline}$$

E.g. If $ESE_{baseline}$ is small relative to other errors, increasing baseline sample sizes will not be useful

ESE_{Total} Estimated via simulation
using “conventional” method
that assumes allele frequencies in baseline
populations are known

$$ESE(\hat{\theta}_i)_{total} \hat{=} \frac{1}{R} \sum \left[(\theta_i - \hat{\theta}_i)^2 \right]$$

ESE_{Fishery} can be calculated from
binomial variance

$$ESE(\hat{\theta}_i)_{\text{fishery}} = \frac{\theta_i(1-\theta_i)}{N}$$

ESE_{Baseline} requires knowing
baseline allele frequencies

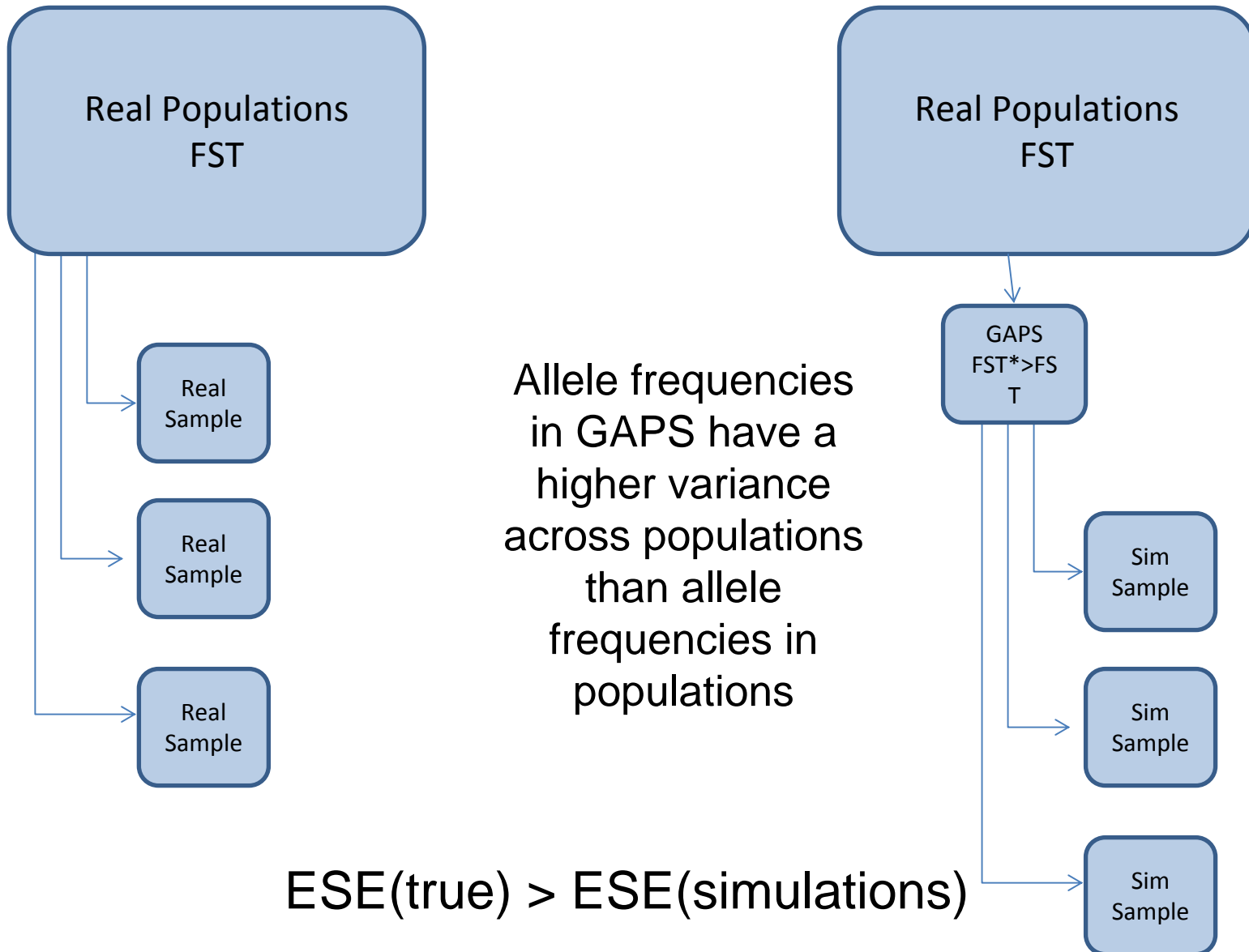
$\hat{\theta}_i^*$ An estimate obtained using
parametric baseline allele
frequencies (possible only in
simulations)

$$ESE(\hat{\theta}_i)_{\text{baseline}} = ESE(\hat{\theta}_i)_{\text{total}} - ESE(\hat{\theta}_i^*)$$

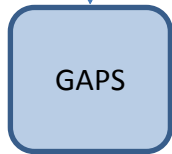
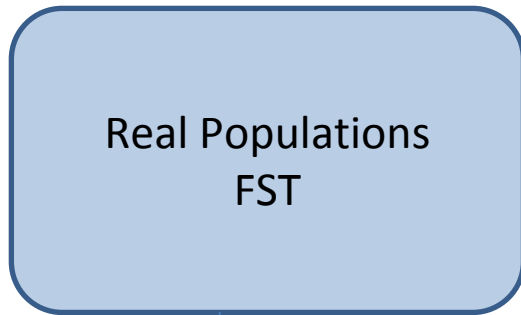
$ESE_{\text{Genotypic}}$ estimated by subtraction

$$ESE(\hat{\theta}_i)_{total} = ESE(\hat{\theta}_i)_{fishery} + ESE(\hat{\theta}_i)_{genotypic} + ESE(\hat{\theta}_i)_{baseline}$$

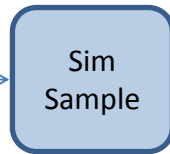
A Problem: We don't know allele frequencies in baseline populations




A potential solution



Adjust allele frequencies in baseline (to compensate for sampling) before performing simulations

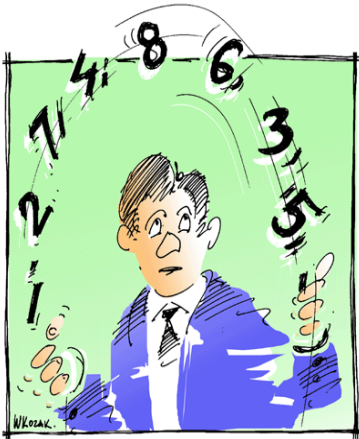


Allele frequencies can be adjusted towards the mean to reduce variance


$$\left\{ \begin{array}{l} \text{If } \hat{p}_i > \bar{p}, \quad \tilde{p}_i = \bar{p} + \sqrt{(\hat{p}_i - \bar{p})^2 - \left(1 - \frac{1}{k}\right) \frac{\hat{p}_i(1-\hat{p}_i)}{n_i}} \\ \text{If } \hat{p}_i < \bar{p}, \quad \tilde{p}_i = \bar{p} - \sqrt{(\hat{p}_i - \bar{p})^2 - \left(1 - \frac{1}{k}\right) \frac{\hat{p}_i(1-\hat{p}_i)}{n_i}} \end{array} \right.$$

Simulations suggest that this works well.

Talk Outline



1. Sources of GSI Error
2. Error decomposition
 - Statistics
 - More statistics
3. Results
4. Discussion: What next?

Results:

Error decomposition for a real fishery

Fishery: SE Alaska, winter*
Baseline: GAPS 2.1 microsatellites
Estimation: Population level

Source of Error	Proportion
Fishery	9.5%
Genotypic	2.7%
Baseline	87.5%

*Thanks to ADFG for providing realistic mixture proportions for this analysis

Discussion

What does this

Source of Error	Proportion
Fishery	9.5%
Genotypic	2.7%
Baseline	87.5%

mean?

- Genetics is limiting accuracy of GSI, not sampling from fishery
- Adding more loci is not necessarily required to improve accuracy
- Accuracy could be improved by
 - adding more loci
 - increasing baseline sample sizes
 - improving estimates of allele frequencies with more sophisticated data analysis



How we might improve estimates of allele frequencies in baseline populations without sampling more fish

- Generalized Expectation Maximization Algorithm
 - Use mixture samples to improve estimates of allele frequencies in baseline population
- Spatial models of allele frequencies
 - Populations near each other tend to be similar
 - Mathematical models in epidemiology can be adapted

Software for error decomposition will soon be available at

Website: www.montana.edu/kalinowski

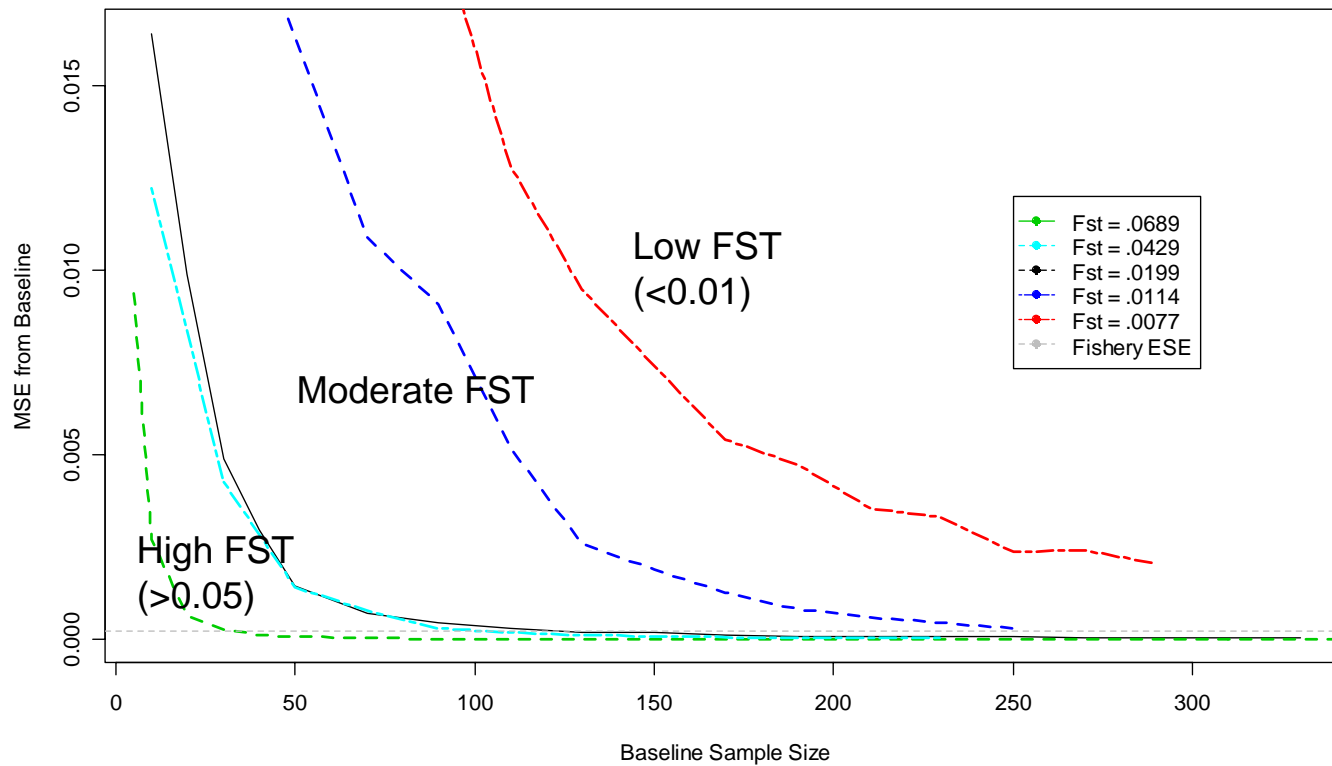
Program name: ONCOR

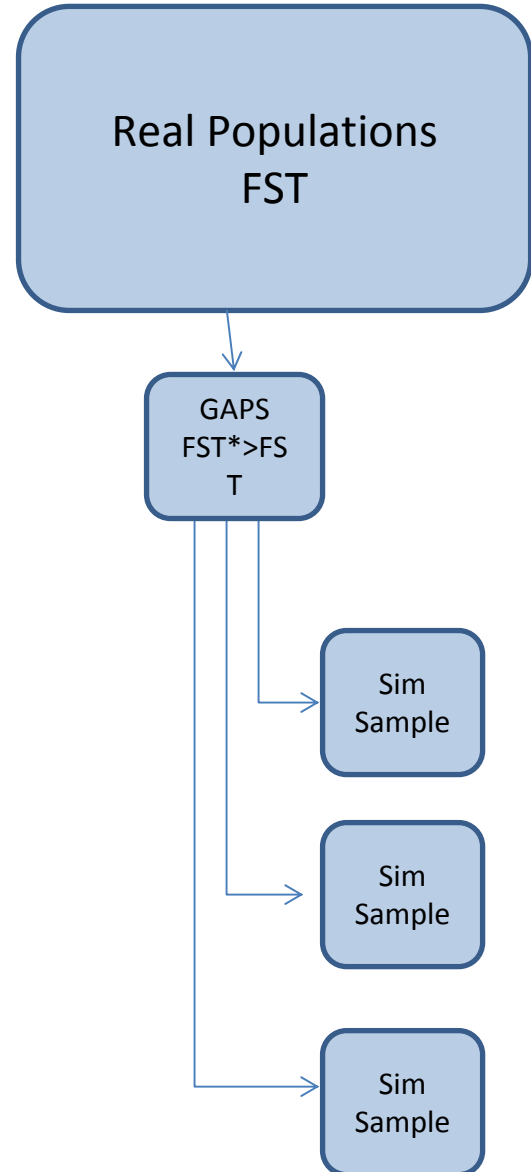
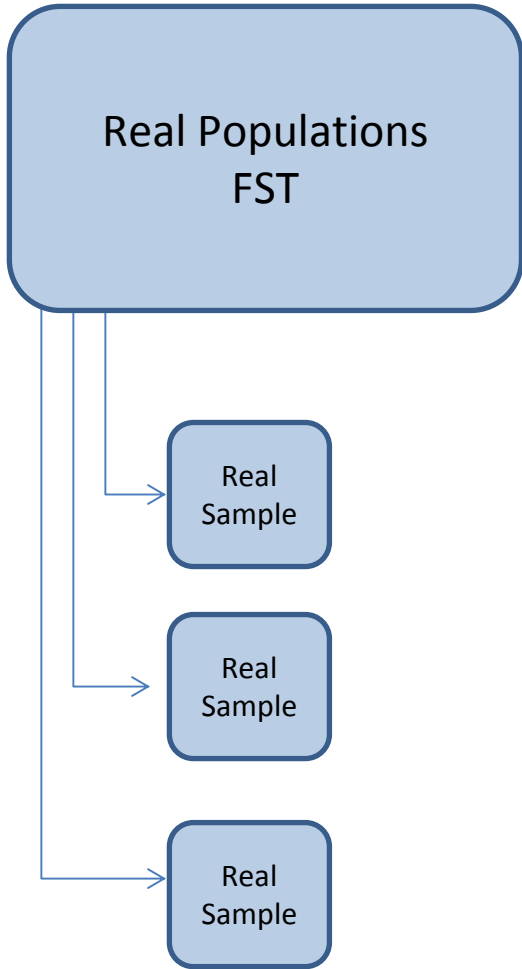
Thanks to



Preliminary Result:

Large baselines justified when FST small





Estimates from 100% Simulations

	AndrewCr	Kowatua	LTahltanR	NakinaR	Tatsatua	UNahlinR	All pops
Expected average	0.9088	0.7287	0.7874	0.6618	0.7983	0.8284	0.786
Anderson	0.8875	0.7321	0.7767	0.6735	0.7813	0.7994	0.775
Reduced variance	0.9021	0.7031	0.7673	0.6288	0.7799	0.8138	0.766
NonZero	0.9499	0.8535	0.8870	0.8192	0.8868	0.8969	0.882
SampleFreqs	0.9560	0.8674	0.8994	0.8371	0.8987	0.9071	0.894

Preliminary conclusion: Method seems to work