Spatial Mapping of Malaria Parasite Genetics

Challenges and Opportunities of High Diversity Genetic Loci

Maxwell Murphy – UC Berkeley/UCSF

4/17/2019

Outline

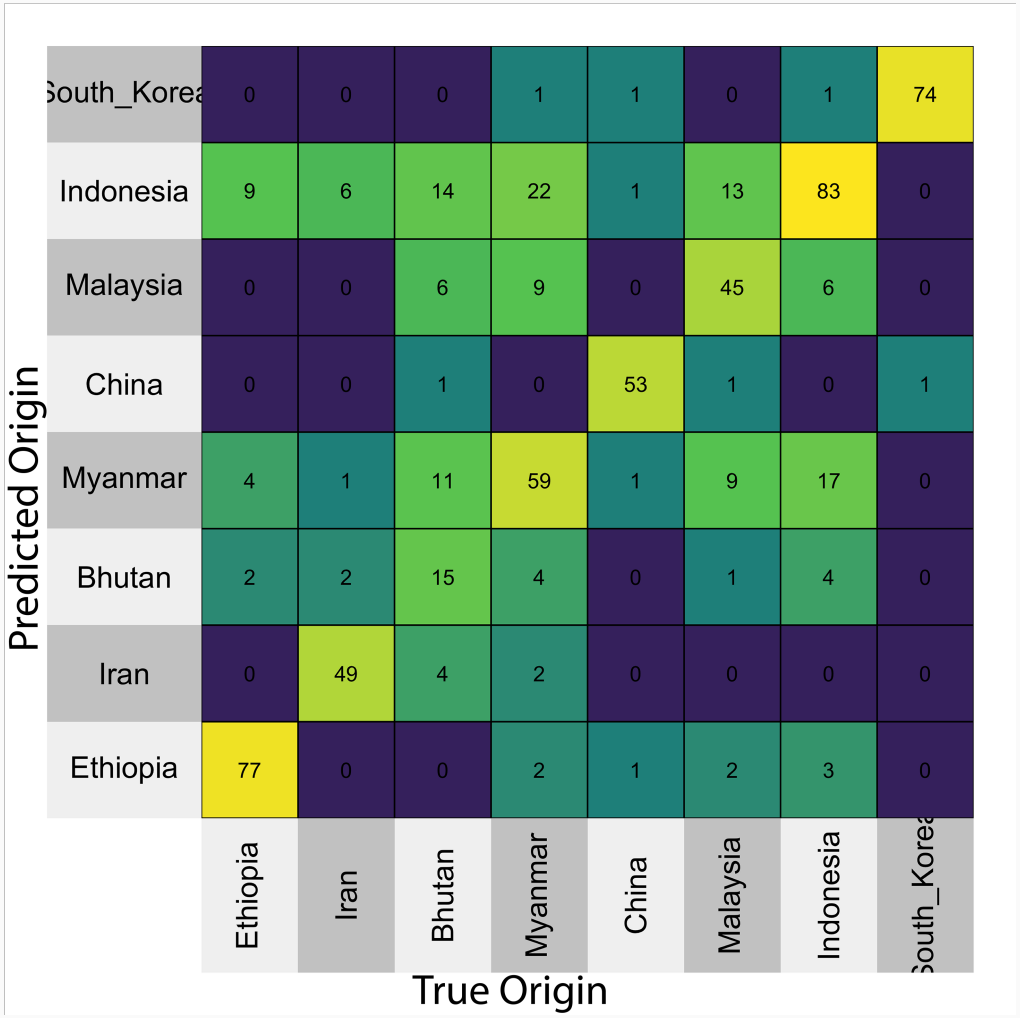
- Applying gaussian process regression to estimate allele frequency surfaces and classify origin of clinical infections
- Challenges of utilizing polygenomic data
- Approaches to utilize polygenomic data with high diversity genetic loci

Estimating Allele Frequencies Using Gaussian Processes

Motivation

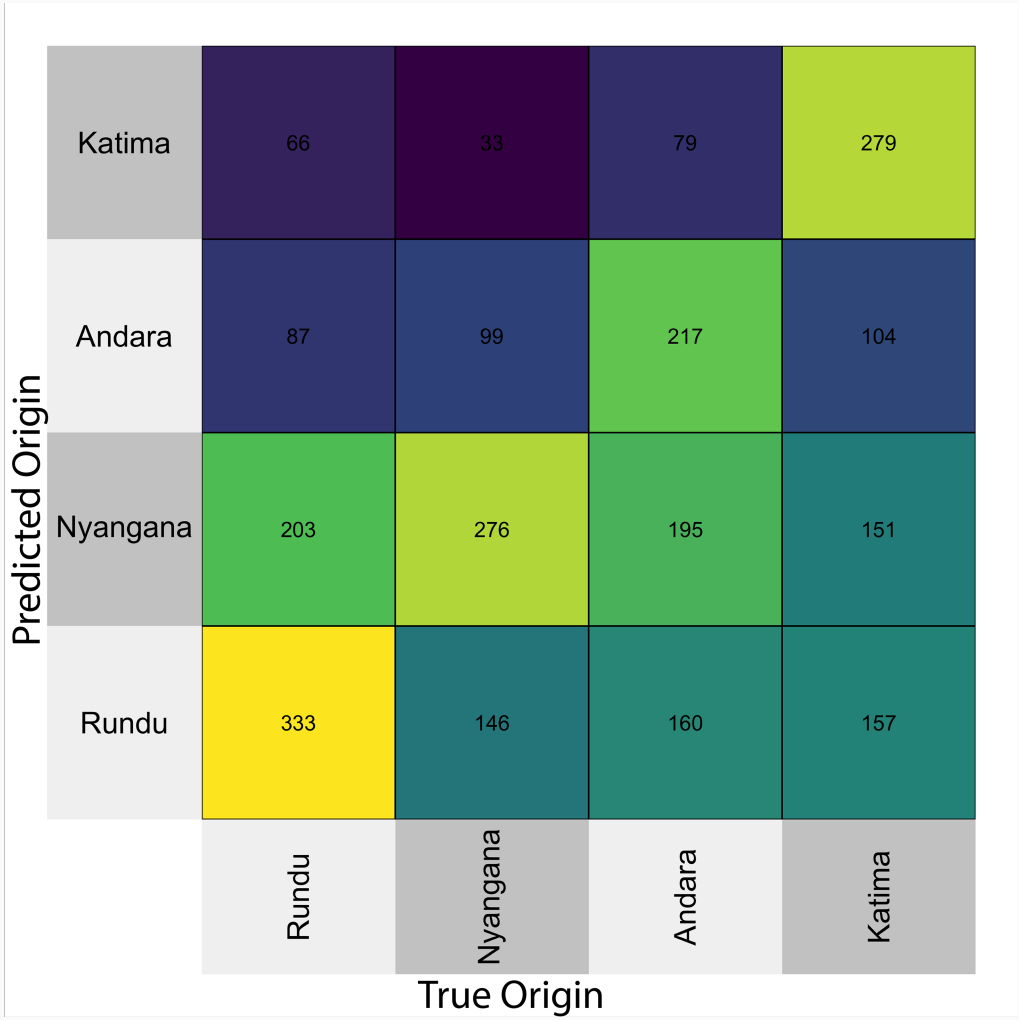
- Classifying malaria cases as local vs imported and determining origin of infection is of great interest
- Spatial models of parasite diversity and distributions would make for interesting inputs into other spatial models
- Integrating spatial information with genetic data remains challenging

Application to VivaxGEN Data



- Publicly available microsatellite data from VivaxGEN
- 617 Samples genotyped at 9 microsatellites
- Polyclonal samples were restricted to dominant alleles

Application to Regional Clinical Data



- 2585 samples collected from the northern region of Namibia (Tessema et al., eLife 2019)
- Genotyped at 26 microsatellite markers
- Polyclonal samples were restricted to dominant alleles

Takeaways

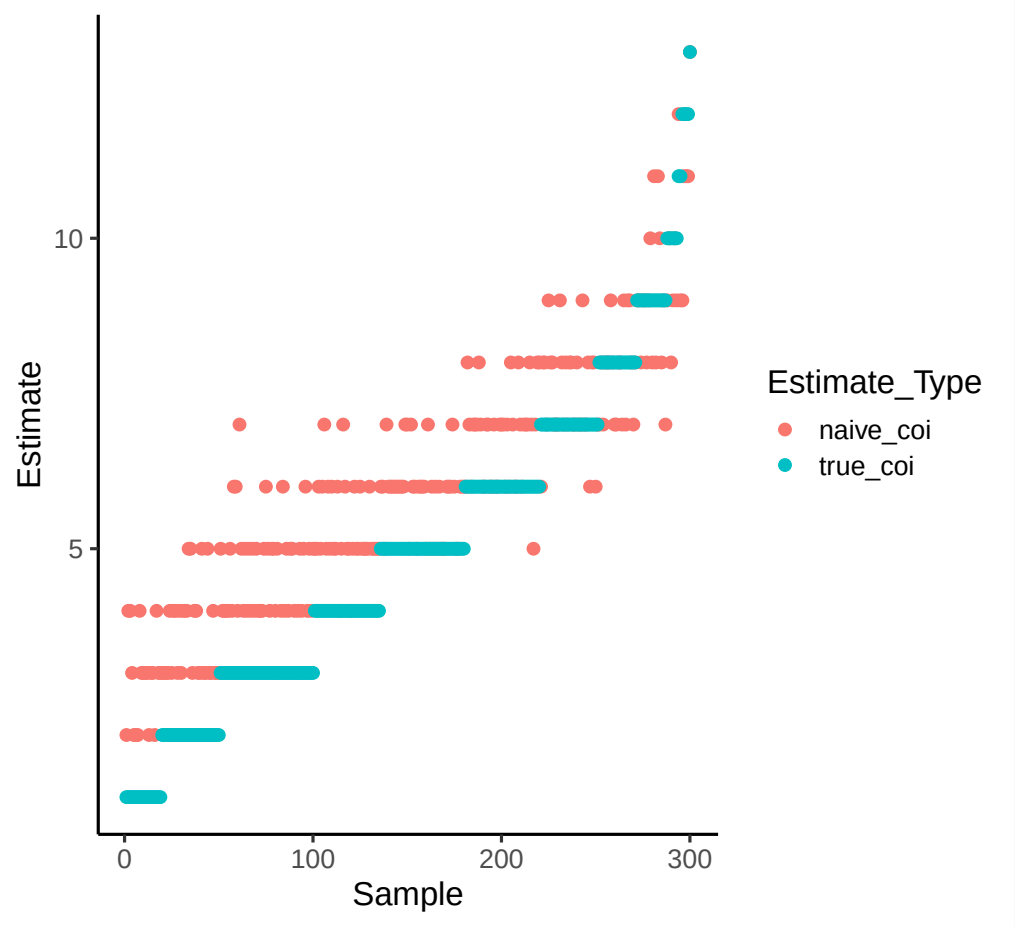
- Using vanilla GP regression with an exponential kernel and deep enough sampling, spatial signal can be extracted that is useful for origin classification
- More nuanced approaches to modeling spatial covariance would likely be very fruitful in exposing spatial connectivity of parasite populations
- Publicly available regional databases of genetic data will be critical in developing allele frequency maps

The Dirty Little Secret of Malaria Genomics

Challenges of Complex Infections and High Diversity Genetic Loci

- "Polyclonal samples were restricted to dominant alleles"
- "Analysis was restricted to monoclonal infections"
- Primarily motivated by using statistics that were not designed to be used in the context of mixed DNA populations
- Also a consequence of noisiness of genotyping method
  - e.g. microsatellite data
- Statistical convenience is being prioritized at the consequence of bias, either due to sampling or because of properties of estimator

Example: Estimating Complexity of Infection

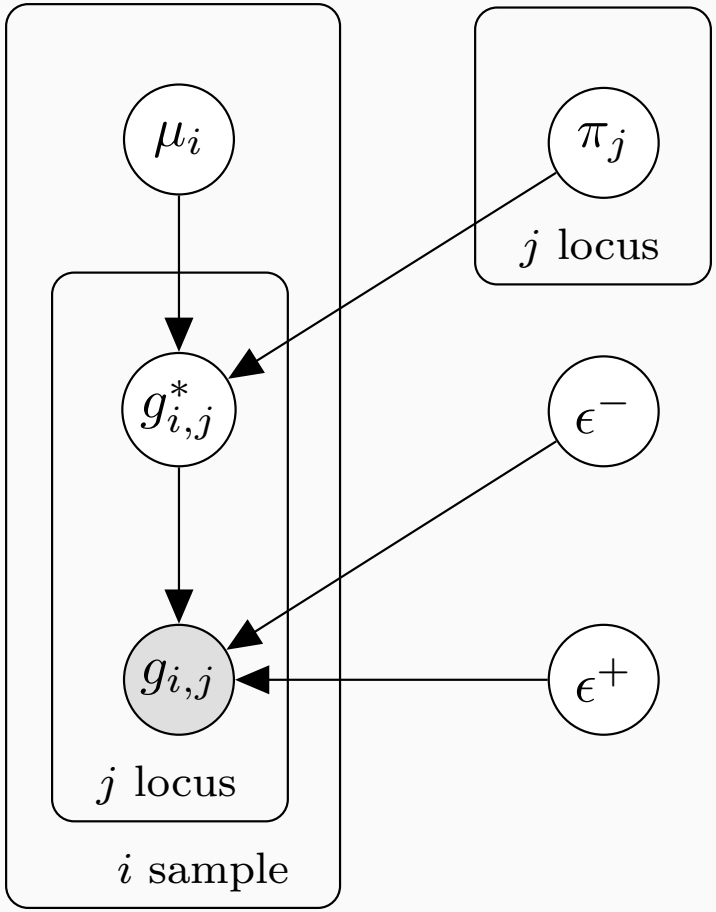


- Simulated data using 12 Loci ranging from 5 to 25 alleles
- FP Rate: .03
- FN Rate: .1
- Complexity of infection estimated by taking maximum observed number of alleles

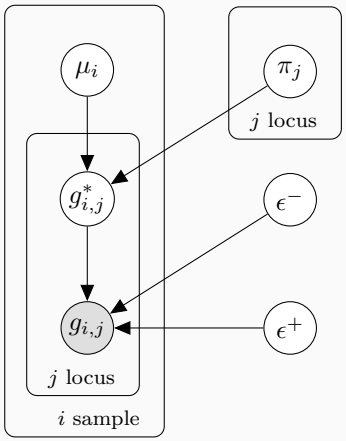
Challenges of Complex Infections and High Diversity Genetic Loci

- Some tools exist to estimate parameters such as complexity of infection (COI), allele frequencies, and population structure
- COIL, THE REAL McCOIL
  - Restricted to SNP data, incorporates genotyping error model
- MALECOT
  - Supports multi-allelic data, does not incorporate genotyping error model

Constructing Models to Account for Complex Infections Using Multi-allelic Data



Constructing Models to Account for Complex Infections Using Multi-allelic Data



$$\mathcal{L}(\pi, \epsilon^-, \epsilon^+, \mu | G) = \prod_{i=1}^{n} \prod_{j=1}^{k} \sum_{g^* \in G^*} P(g_{i,j} | g^*_{i,j}, \epsilon^+, \epsilon^-) P(g^*_{i,j} | \mu_i, \pi_j)$$

$$P(g_{i,j} | g^*_{i,j}, \epsilon^+, \epsilon^-) = \prod_{k=1}^{a} \begin{cases} (1-\epsilon^-)^{g^*_{i,j,k}} & \text{if } g_{i,j,k} = 1 \text{ and } g^*_{i,j,k} > 0 \\ (\epsilon^-)^{g^*_{i,j,k}} & \text{if } g_{i,j,k} = 0 \text{ and } g^*_{i,j,k} > 0 \\ (\epsilon^+) & \text{if } g_{i,j,k} = 1 \text{ and } g^*_{i,j,k} = 0 \\ (1-\epsilon^+) & \text{if } g_{i,j,k} = 0 \text{ and } g^*_{i,j,k} = 0 \end{cases}$$

$$P(g^*_{i,j} | \mu_i, \pi_j) = \frac{\mu_i!}{g^*_{i,j,1}! \cdots g^*_{i,j,k}!} \pi_{j,1}^{g^*_{i,j,1}} \cdots \pi_{j,k}^{g^*_{i,j,k}}$$

Computational Complexity

$$\mathcal{L}(\pi, \epsilon^-, \epsilon^+, \mu | G) = \prod_{i=1}^{n} \prod_{j=1}^{k} \sum_{g^* \in G^*} P(g_{i,j} | g^*_{i,j}, \epsilon^+, \epsilon^-) P(g^*_{i,j} | \mu_i, \pi_j)$$

Computational Complexity

$$\mathcal{L}(\pi, \epsilon^-, \epsilon^+, \mu | g) = \prod_{i=1}^{n} \prod_{j=1}^{k} \sum_{g^* \in G^*} P(g_{i,j} | g^*_{i,j}, \epsilon^+, \epsilon^-) P(g^*_{i,j} | \mu_i, \pi_j)$$

	1	2	3	4	5	6
2	2	3	4	5	6	7
4	4	10	20	35	56	84
8	8	36	120	330	792	1716
16	16	136	816	3876	15504	54264
32	32	528	5984	52360	376992	2324784
64	64	2080	45760	766480	10424128	119877472
128	128	8256	357760	11716640	309319296	6856577728

Tackling Complex Infections with High Diversity Genetic Loci

Estimating Marginal Likelihoods

- We do not need to calculate the exact marginal likelihood for each sample
- An unbiased estimate of the probability density results in a Markov Chain with the exact target as its stationary distribution (Andrieu and Roberts, 2009)

Computationally Intractable

$$r := \frac{P(x')g(x|x')}{P(x)g(x'|x)}$$

Tractable

$$\hat{r} := \frac{\hat{P}(x')g(x|x')}{\hat{P}(x)g(x'|x)}$$

Estimating Marginal Likelihoods

- For each sample at each locus, we can generate an unbiased estimate of the likelihood instead of calculating an exact likelihood

$$P(g|\pi, \epsilon^-, \epsilon^+, \mu) = \sum_{g^* \in G^*} P(g|g^*, \epsilon^+, \epsilon^-) P(g^*|\mu_i, \pi_j)$$

Importance Sampling

- Importance Sampling provides a computationally tractable method to generate unbiased estimates of the marginal likelihood of a given data point
- Intuitively makes sense, as the vast majority of potential true genotypes contribute very little to the likelihood

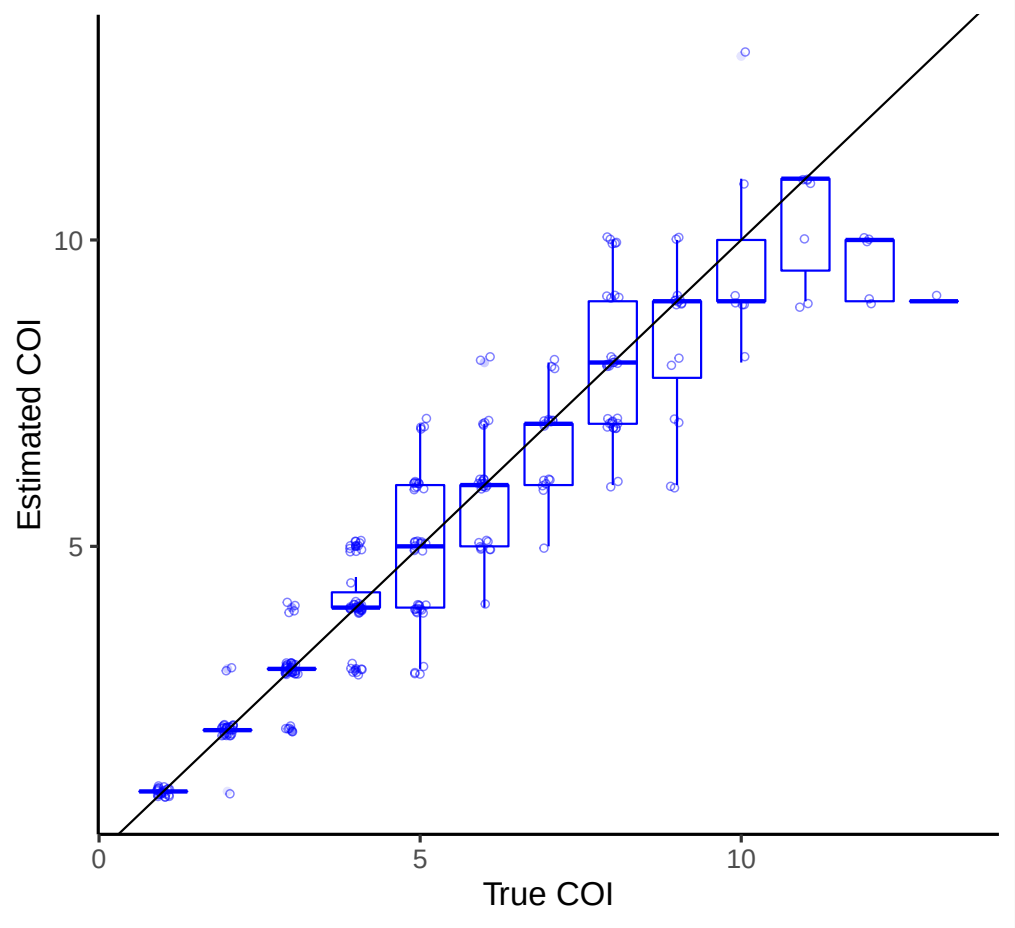
$$\hat{P}(g|\pi, \epsilon^-, \epsilon^+, \mu) = \frac{1}{K} \sum_{k=1}^{K} \frac{P(g|g_k^*, \epsilon^+, \epsilon^-) P(g_k^*|\mu_i, \pi_j)}{q(g^*)}$$

$$g^* \sim q$$

Choosing an Importance Sampling Distribution

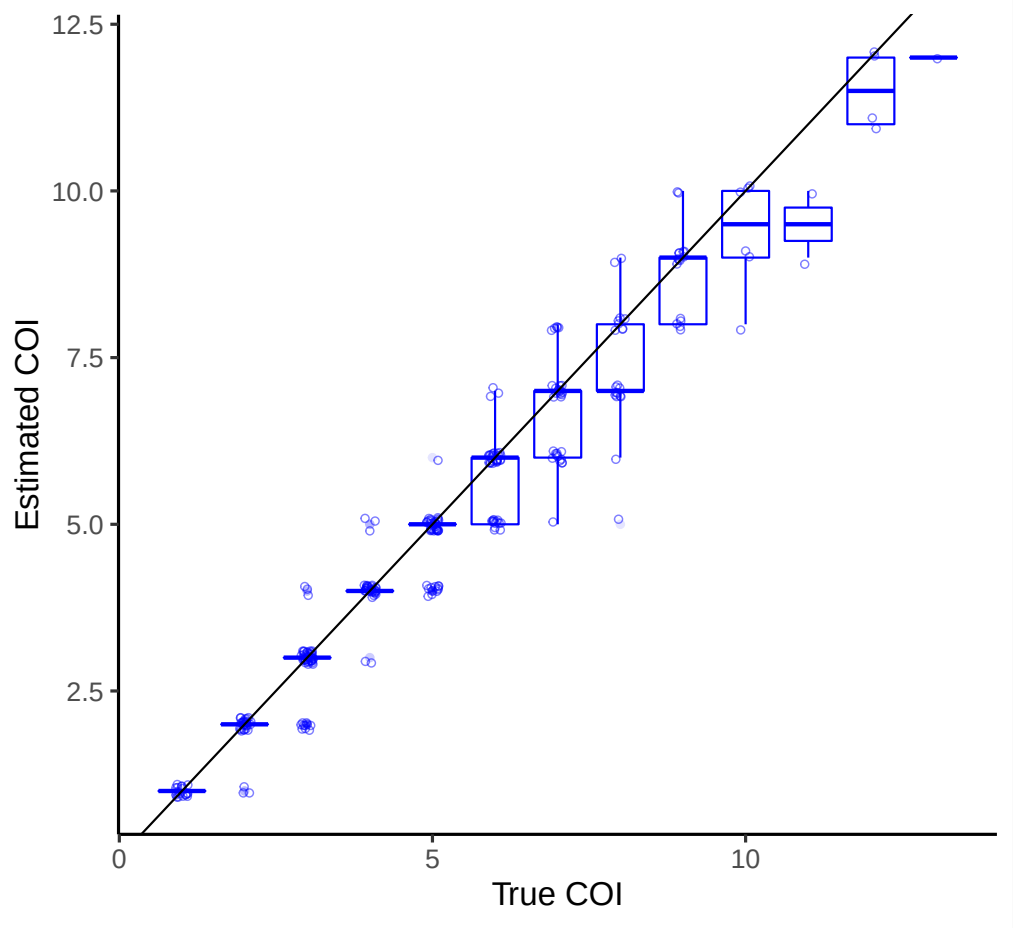
- Choice of Importance Sampling Distribution has significant impact on variance of estimate, and subsequently the efficiency of sampling
- A good heuristic for choosing a sampling distribution is to reweight the estimated allele frequency distribution based on the observed genotype and false negative rate
- Ex:
  - $\pi = [.1, .2, .3, .4]$
  - $\epsilon^{-} = .1$
  - $g = [1, 0, 0, 0]$
  - $q = [.9, .02, .03, .04]$

Demonstration: Low Diversity



- Simulated data using 12 Loci with 5 alleles
- FP Rate: .03
- FN Rate: .1

Demonstration: High Diversity



- Simulated data using 12 Loci with between 5 and 25 alleles
- FP Rate: .03
- FN Rate: .1

Future Directions

Future Directions

- With this computational framework, we will extend our spatial modeling of allele frequencies to incorporate all observed genetic data
- Further develop this framework for estimating other parameters of interest

Thank you

maxwell.murphy@ucsf.edu

github.com/m-murphy