

# Bayesian Analysis of Population Mixture and Admixture

Eric C. Anderson

*Interdisciplinary Program in Quantitative  
Ecology and Resource Management*

University of Washington, Seattle, WA, USA

Jonathan K. Pritchard

*Department of Statistics*

University of Oxford, UK

This Research Supported by:  
NSF BIR – 9807747

# Overview

§1 A Motivating Problem—*Felis sylvestris* in Scotland

§2 A model for population mixture

§3 A model for population admixture

- Block updating Gibbs sampler
  - A Baum *et al.* type of computation

§4 Simultaneous mixture/admixture analysis

§4 Results

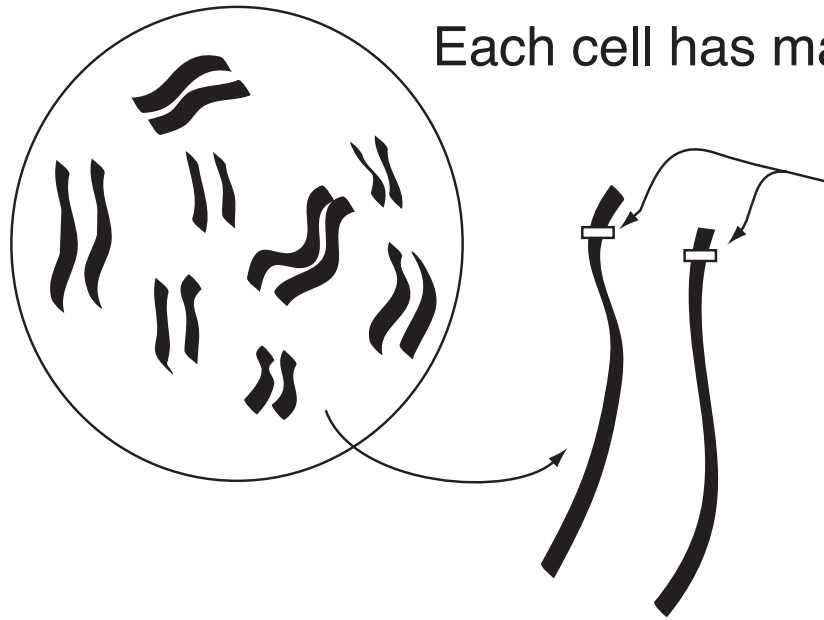
## *Felis sylvestris*: History and Genetic Data



Data Provided by Mark A. Beaumont (University of Reading, UK):

- 230 Wild-Living Cats Genotyped at 8 Microsatellite Loci

# Genetics Background



Each cell has many pairs of chromosomes

Very precise locations in the genome may be reliably found and analyzed.

Such a location is called a LOCUS (plural = LOCI).

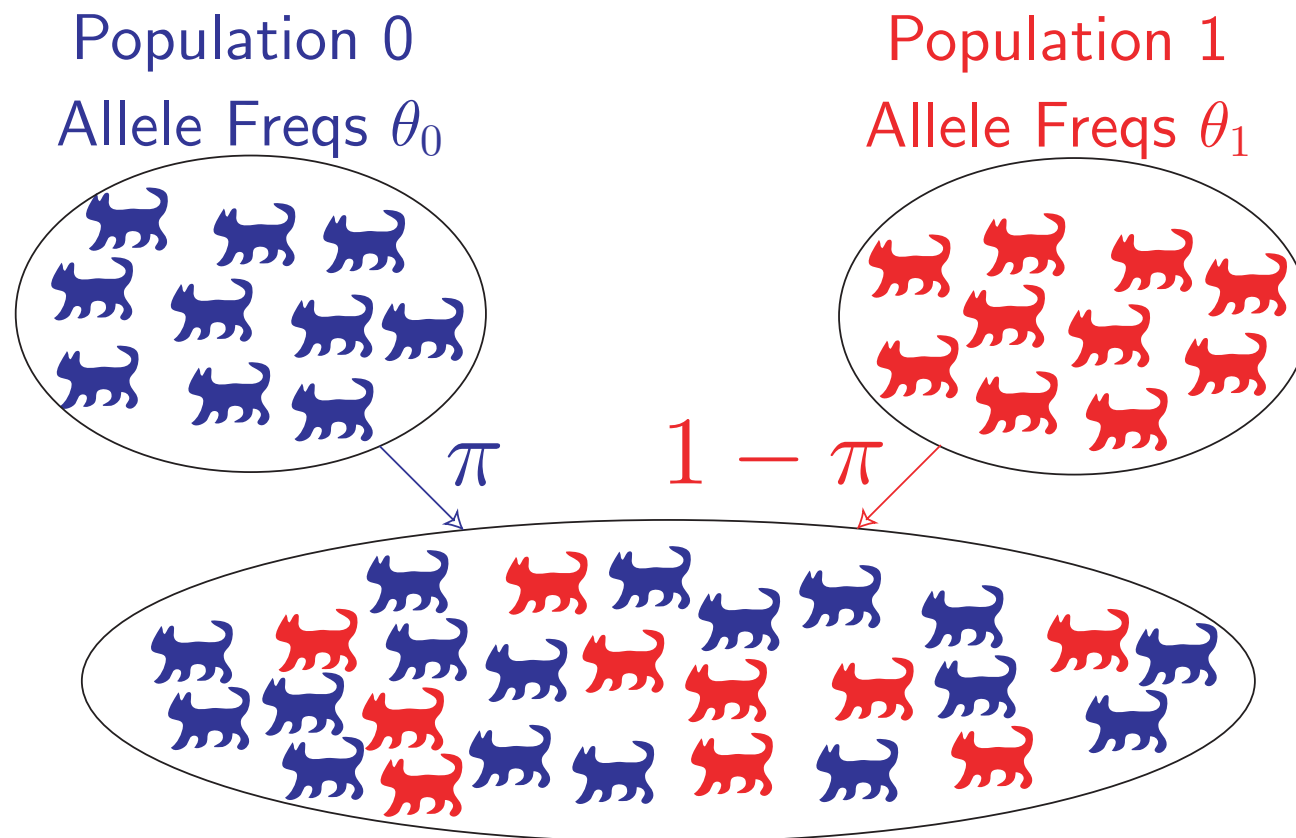
Genetic variants at a locus are known as alleles.

Each individual has two copies of genetic material at a locus which determine its single-locus genotype.

The probability that an individual carries a particular allele at a locus depends on how frequent that allele is in the population.

For an individual from a population in *equilibrium*, the alleles carried are independent of one another within and between loci.

# Model For Genetic Mixture



Using a sample from the mixture the goals are to:

1. Estimate the allele frequencies in Populations 0 and 1
2. Estimate the mixing proportion  $\pi$
3. For each individual in the sample, compute the posterior probability that it is from Population 0 or 1

- Goals 1 and 2 would be made very easy if we could observe for each cat a variable  $z_i$ :

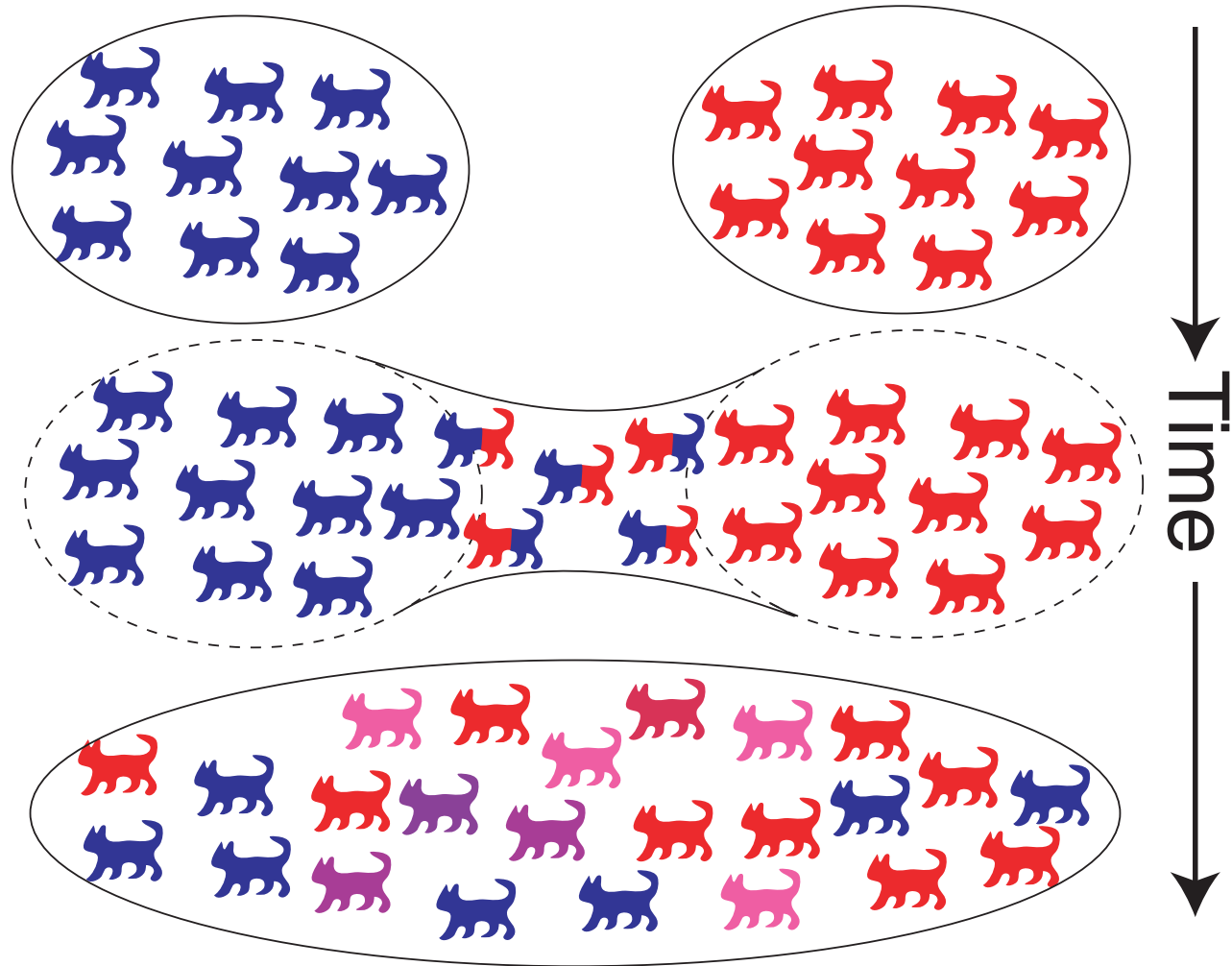
$$z_i = \begin{cases} 0 & \text{if } i^{\text{th}} \text{ cat is from Pop. 0} \\ 1 & \text{if } i^{\text{th}} \text{ cat is from Pop. 1} \end{cases}$$

- Of course, we do not know  $z_i$ , it is a *latent variable*.
- However, if we knew the allele frequencies and the mixing proportions, we could compute the probability distribution for  $z_i$  given the  $i^{\text{th}}$  cat's multilocus genotype:

$$P(z_i = 0 | \theta_0, \theta_1, \pi, \text{gtyp}_i) = \frac{\pi P(\text{gtyp}_i | \theta_0, z_i = 0)}{\pi P(\text{gtyp}_i | \theta_0, z_i = 0) + (1 - \pi) P(\text{gtyp}_i | \theta_1, z_i = 1)}$$

- Taking Dirichlet priors for  $\theta_0$ ,  $\theta_1$  and  $\pi$ , the inclusion of the variables  $z_i$  makes Gibbs sampling straightforward in this model.
- Bayesian inference following DIEBOLT & ROBERT (1994)

## A Schematic of Genetic Admixture



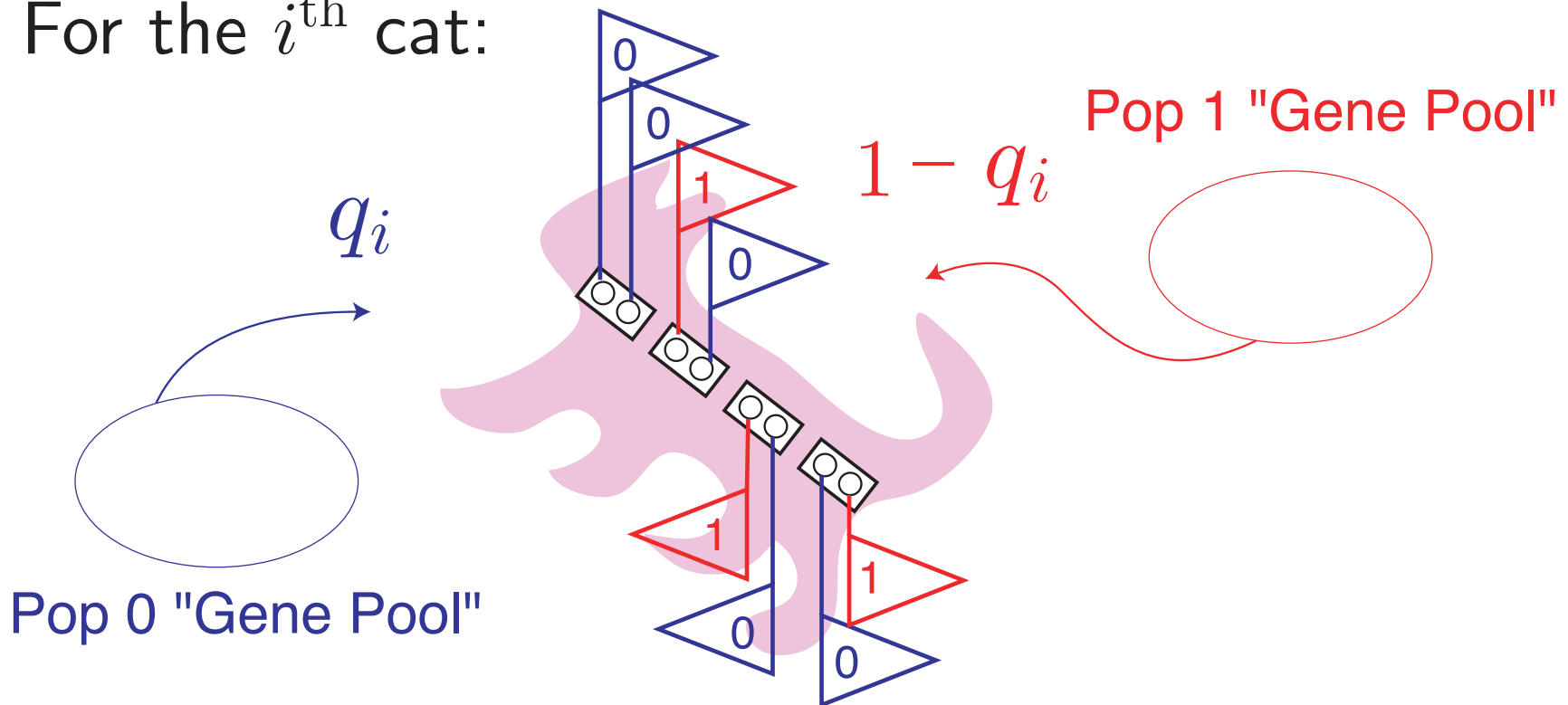
- This requires a different probability model with different latent variables

# Latent Data, $q$ and $w$ for the Admixture Model

$$q_i \sim \text{Beta}(\alpha, \alpha)$$

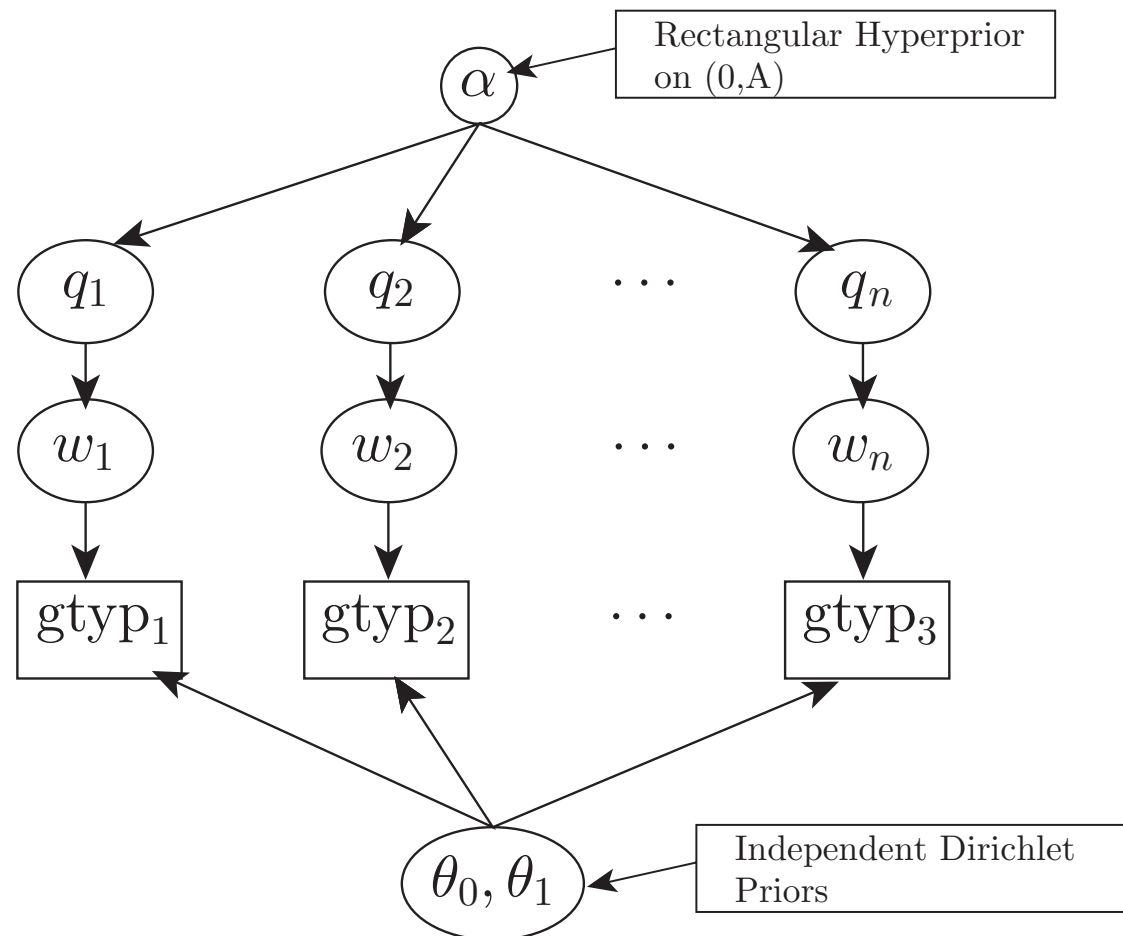
Pritchard et al. (2000)

For the  $i^{\text{th}}$  cat:



- Each gene copy comes from Pop 0, independently, with probability  $q_i$
- The  $t^{\text{th}}$  gene copy in the  $i^{\text{th}}$  cat gets  $w_{it} = 0$  or  $1$  (Flags in Diagram)

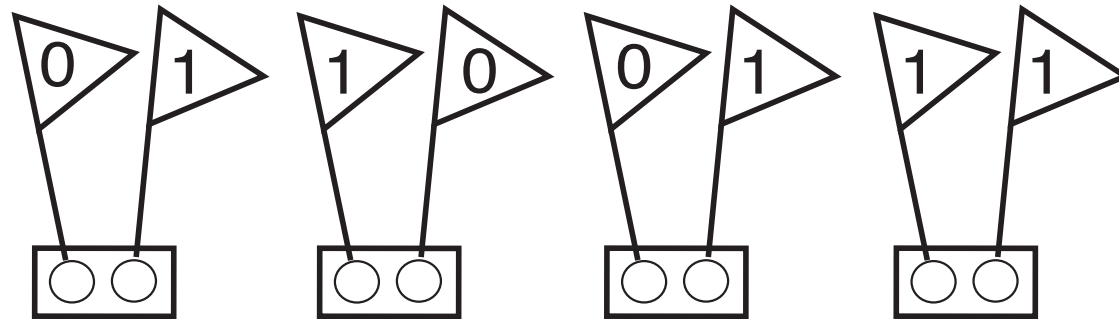
# Hierarchical Structure of the Admixture Model



- Allows straightforward Gibbs sampling for  $\theta$ ,  $w$ , and  $q$
- Metropolis-Hastings update for  $\alpha$  (slow mixing)

## Eliminating the $q_i$ 's

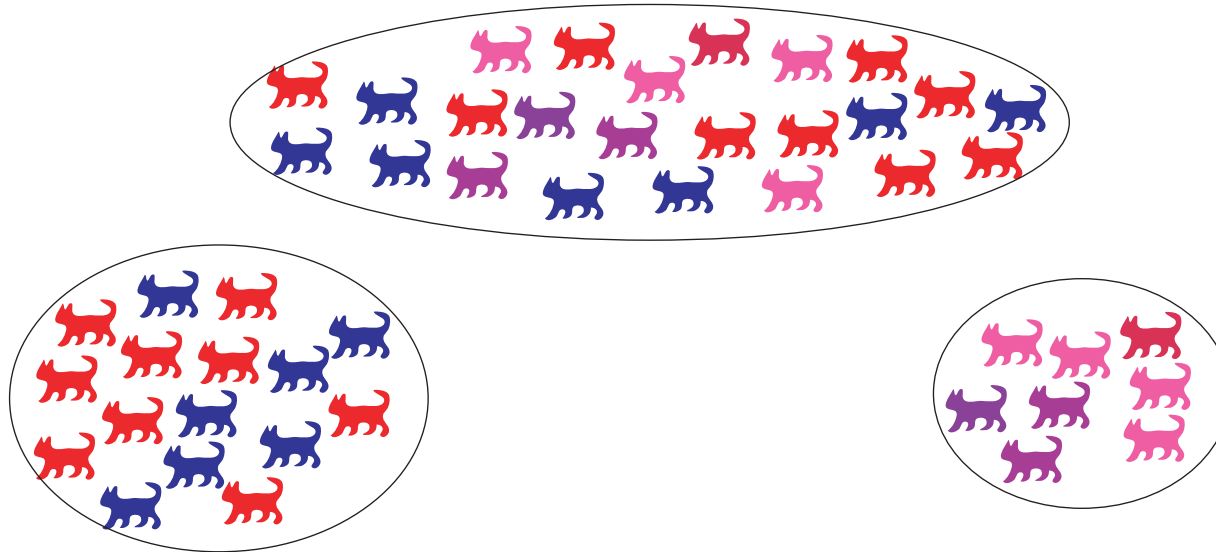
- After integrating out  $q_i$ , the  $w_{it}$  within the  $i^{\text{th}}$  cat have a *labelled* beta-binomial distribution with parameters  $(\alpha, \alpha)$
- This has an interpretation as a Pólya-Eggenberger urn scheme
  - This, in turn, has a Markov chain interpretation



- Forward-Backward algorithms for Hidden Markov Chains allow:
  - Joint updating of the  $w_{it}$ 's from their full conditional dsns within the  $i^{\text{th}}$  cat
  - Better-mixing Metropolis updates for  $\alpha$
  - Efficient calculation of  $P(\text{gtyp}_i | \alpha, \theta)$

# Simultaneous Mixture/Admixture Analysis

- If possible we would like to separate our sample into two groups:

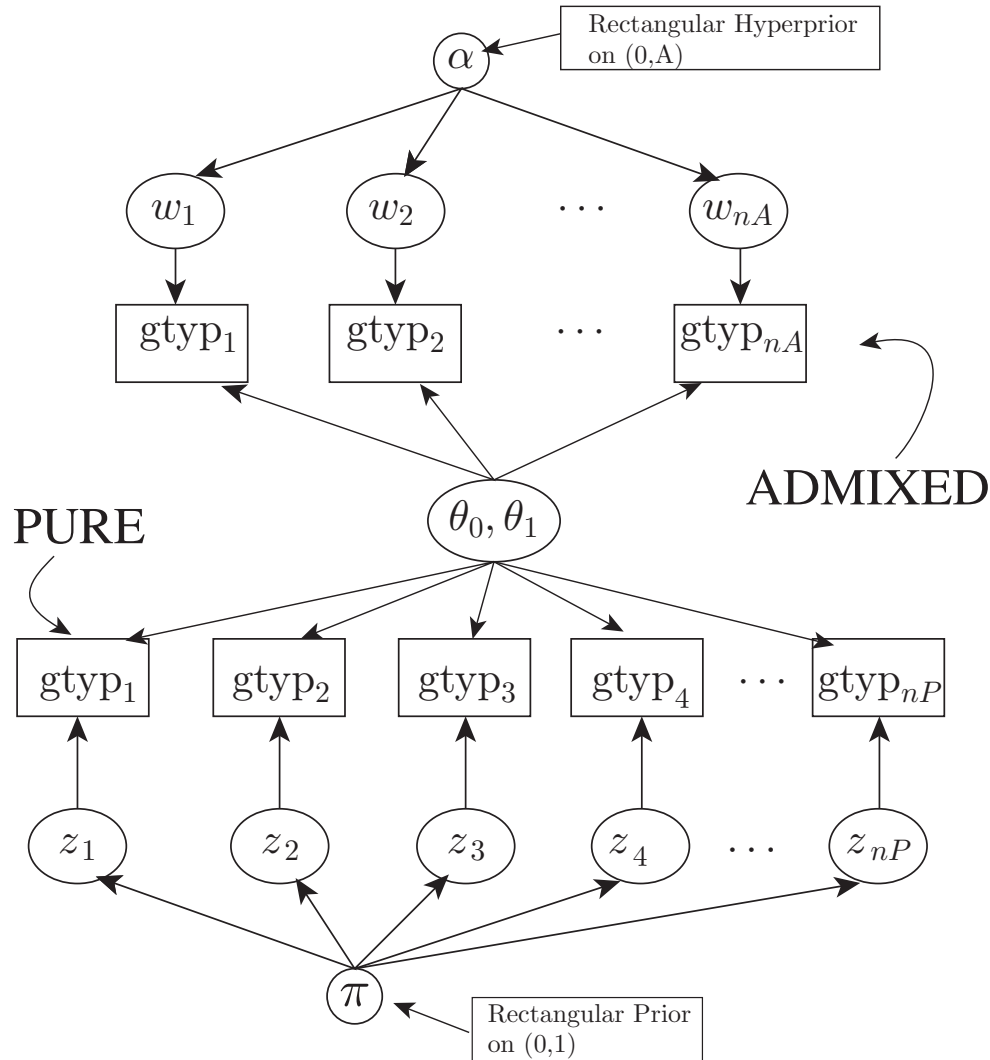


Pure individuals in a mixture  
governed by  $\pi$

Admixed individuals with admixture  
proportions governed by  $\alpha$ .

- But we don't know for certain which individuals are "Pure" and which are "Admixed."
- Different partitions of the sample into the Pure and the Admixed groups correspond to different models that we must average over.

# ADG for Simultaneous Mixture/Admixture Analysis



- Model at left corresponds to one partition of the cats in the sample into Pure and Admixed groups.

- Green (1995) describes reversible-jump methodology for general sampling over such partitions of data.

- However, since we are able to integrate out the  $q_i$ 's, we may employ Gibbs sampling over the partitions.

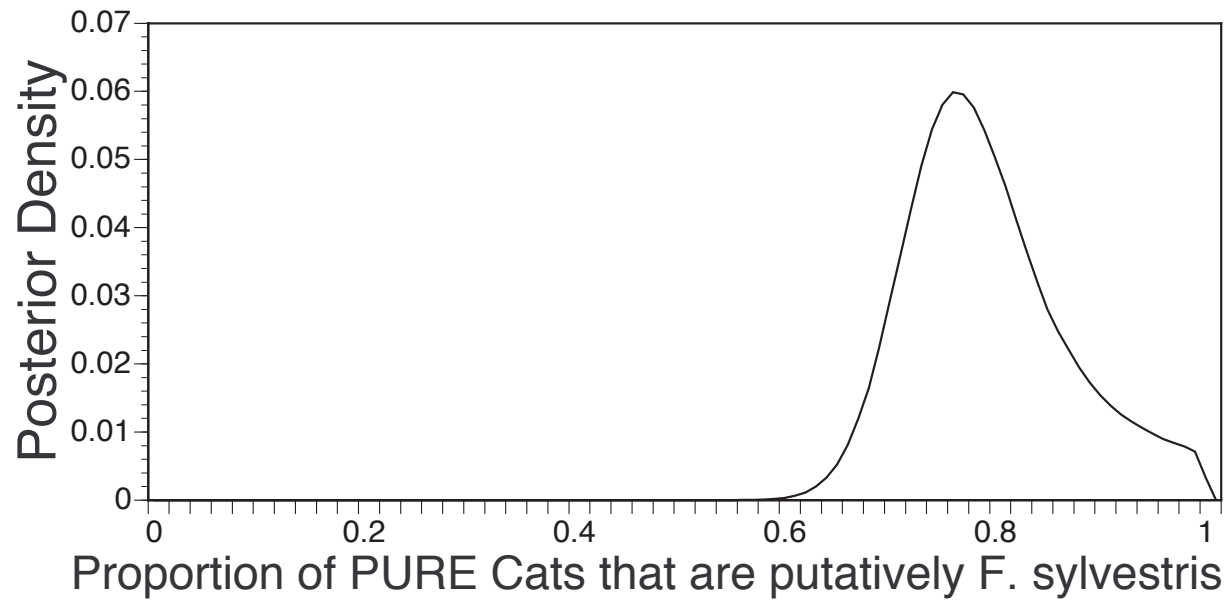
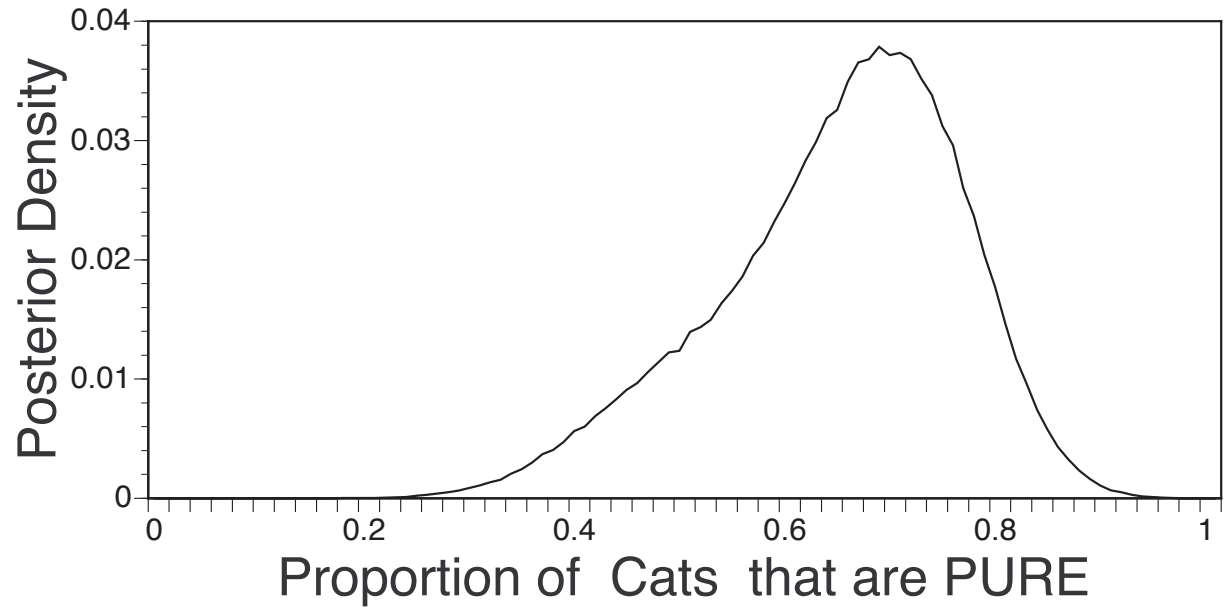
This gives us:

- Very fast mixing between partitions of the data (cats mix appropriately quickly between Pure and Admixed groups)
- Rao-Blackwellized Monte Carlo estimates of the posterior probability that a sampled cat is Pure or Admixed.

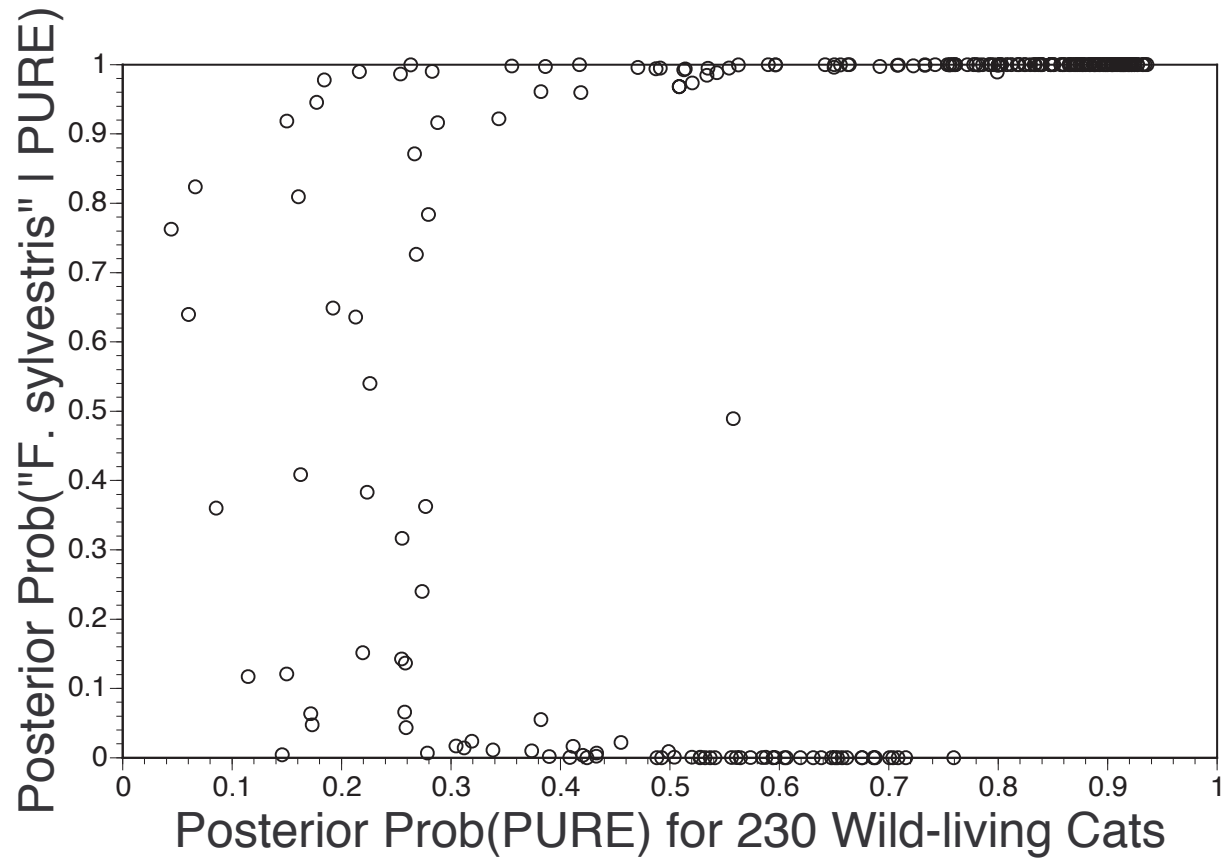
And allows inference of other interesting quantities:

- The proportion of Pure/Admixed cats in the population from which the sample was drawn
- The proportion of Pure *Sylvestris* cats in the population
- The proportion of Pure housecats
- The allele frequencies in the two putative gene pools

# Results for Scottish Cats I



## Results for Scottish Cats II



- \* **Note:** Known house cats, if included, cluster with the others on the bottom right half.
- \* **Also:** Estimated allele frequencies for the “Non-*Sylvestris*” gene pool are very close to those of English housecats.

# Summary

- Genetic mixture model
- Pritchard *et al.*'s genetic admixture model
- Novel computations that improve MCMC in the admixture model
- Simultaneous consideration of mixture and admixture models
- Example Dataset: *Felis sylvestris* in Scotland
  - 43% to 81% of the cats may be of “pure” origin
  - Between 6% and 31% of those may be feral housecats
  - Individuals may be classified on the basis of their posterior probability of being “pure” or “admixed.”
- We anticipate that these methods will be widely useful in studying natural populations.