MONASH University

Science

# Using STRUCTURE to estimate Population Genetic Structure

Anders Gonçalves da Silva

Population and landscape genomics workshop – Day 2

25 March 2014

CBA – ANU

# Outline

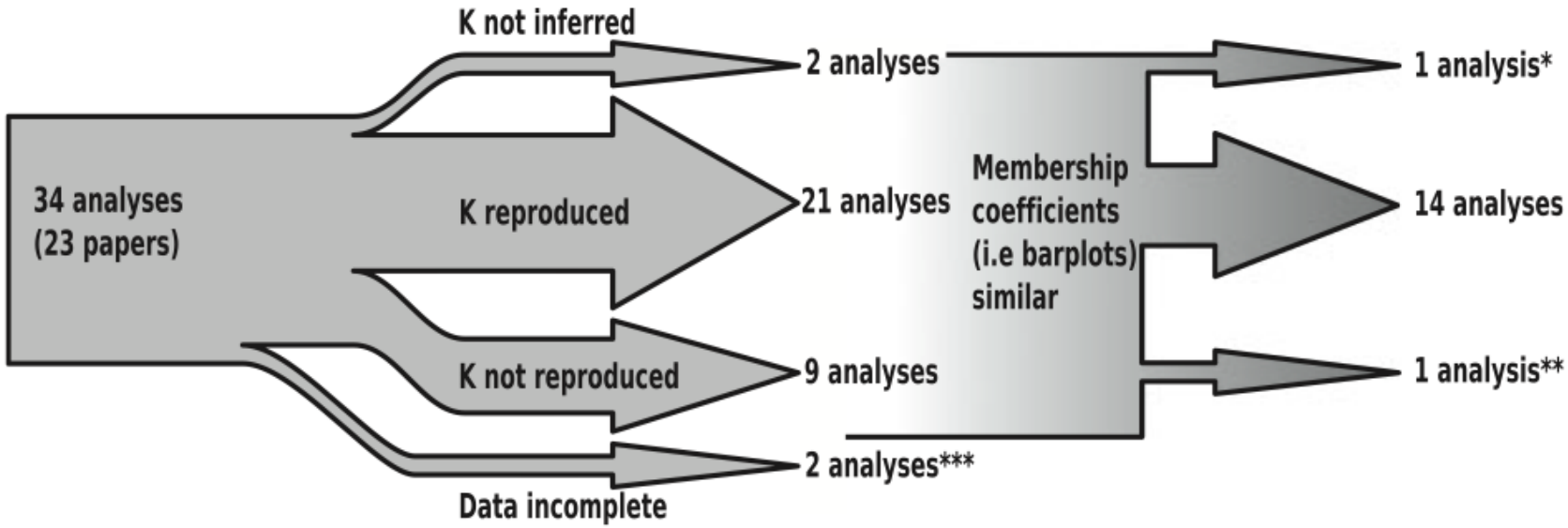
- The problems of reproducibility
- STRUCTURE concept
- Bayesian inference
- The inner workings of STRUCTURE
- The Dirichlet distribution
- Preparing an input file
- Choosing parameter values – going beyond default values
- Collating data, interpreting results

# Question?

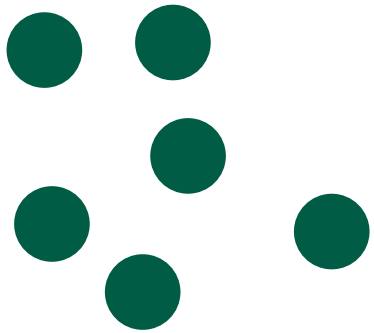
- What is a prior?

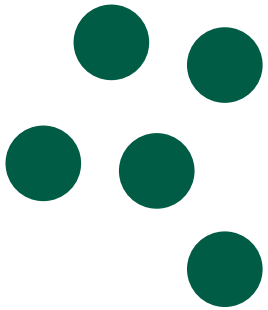
- What is your prior?

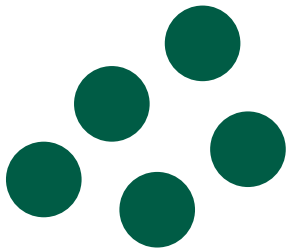
# Gilbert *et al*. (2012) *Mol. Ecol.* 21: 4925



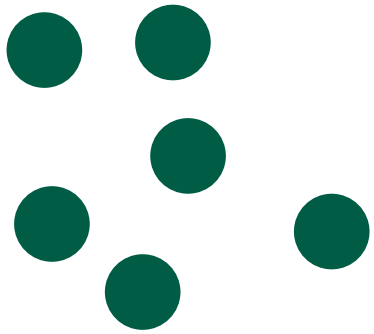
# What does STRUCTURE do?



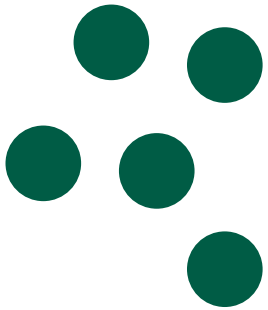


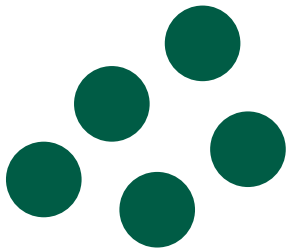


# What does STRUCTURE do?

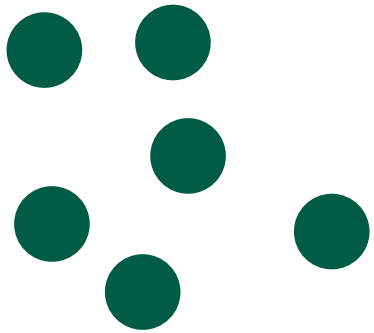


P(K=1|Data)

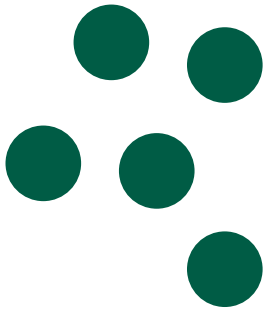


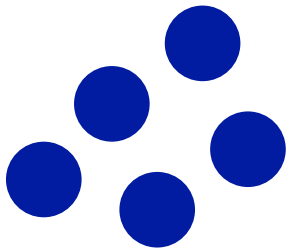


# What does STRUCTURE do?

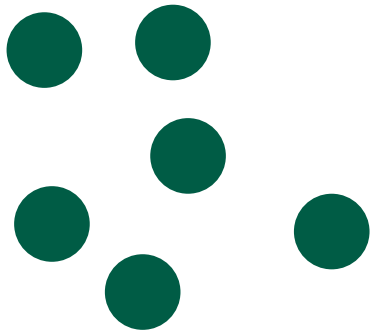


P(K=2|Data)

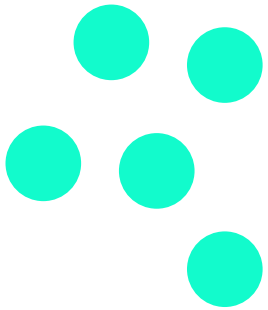


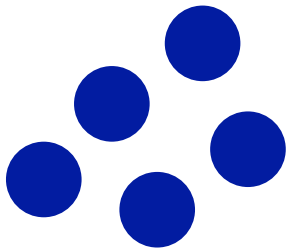


# What does STRUCTURE do?

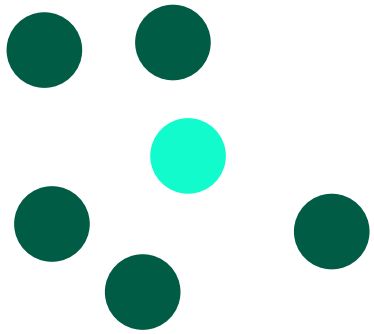


P(K=3|Data)

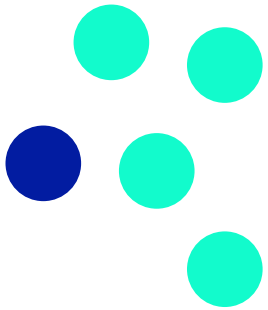


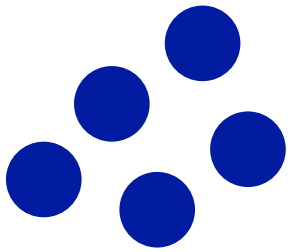


# What does STRUCTURE do?

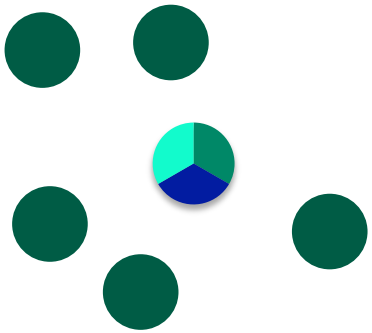


P(K=3|Data)

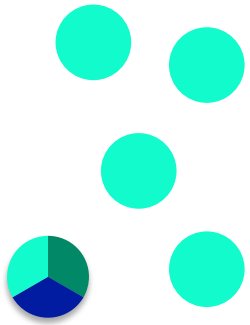


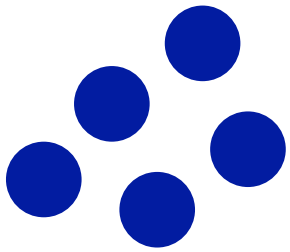


# What does STRUCTURE do?



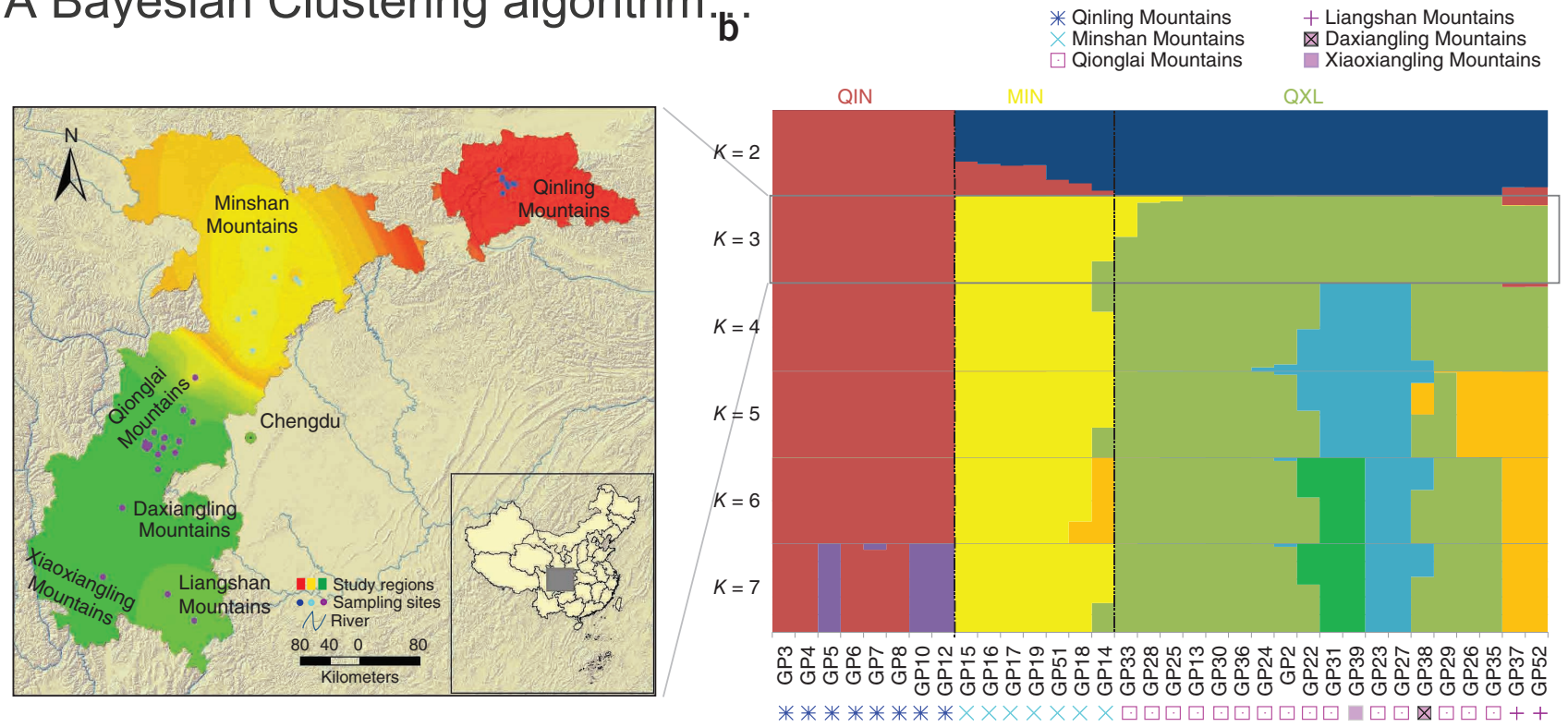
P(K=3|Data)





# STRUCTURE is…

- A Bayesian Clustering algorithm…




Zhao et al. (2013) Nature Genetics 45: 67

# ANATOMY OF BAYESIAN INFERENCE

# Anatomy of Bayesian inference

$$P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{P(Data)}$$

# Anatomy of Bayesian inference

$$P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{P(Data)}$$

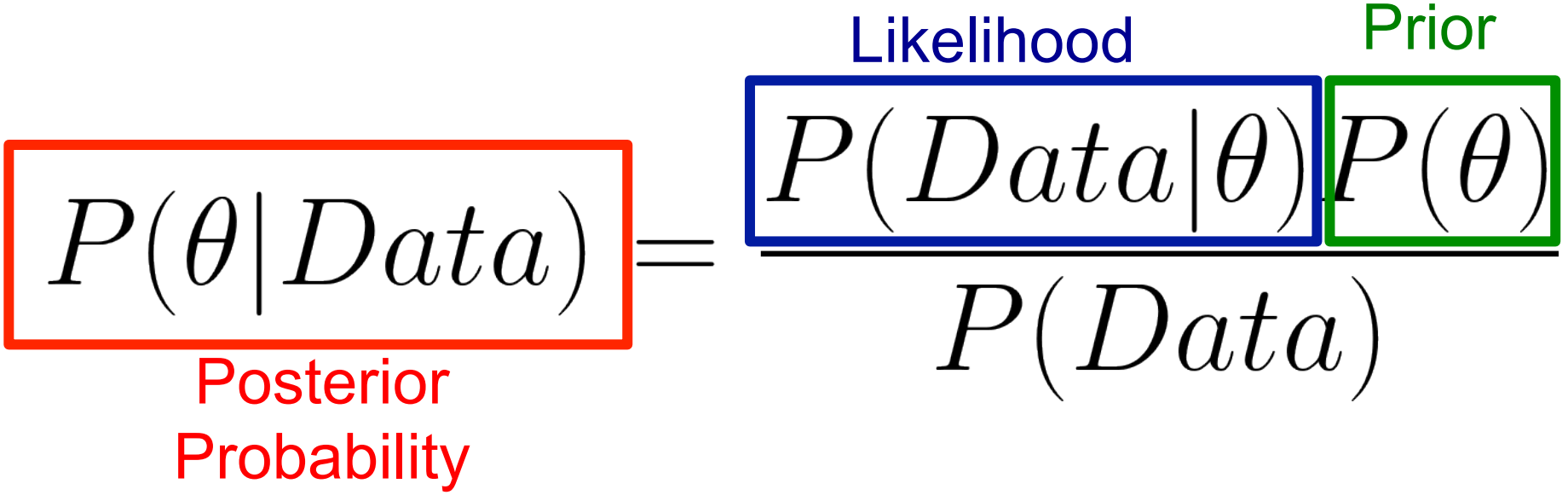
Posterior Probability

# Anatomy of Bayesian inference

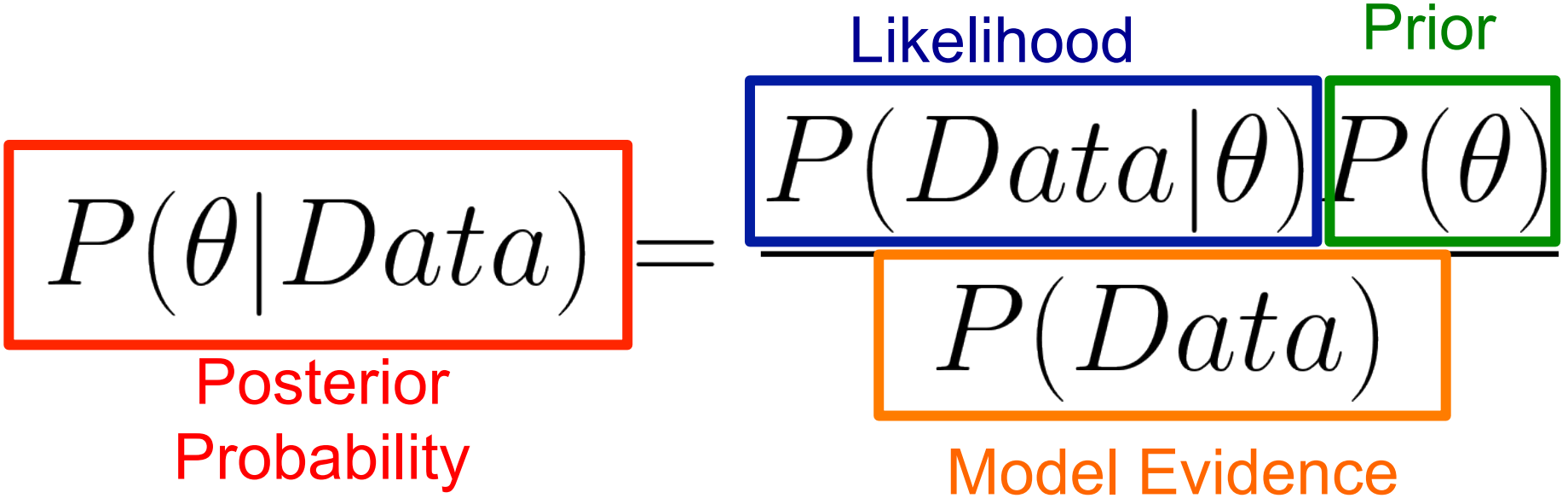
Likelihood

$$\underbrace{P(\theta|Data)}_{\text{Posterior Probability}} = \frac{P(Data|\theta)P(\theta)}{P(Data)}$$

# Anatomy of Bayesian inference



# Anatomy of Bayesian inference

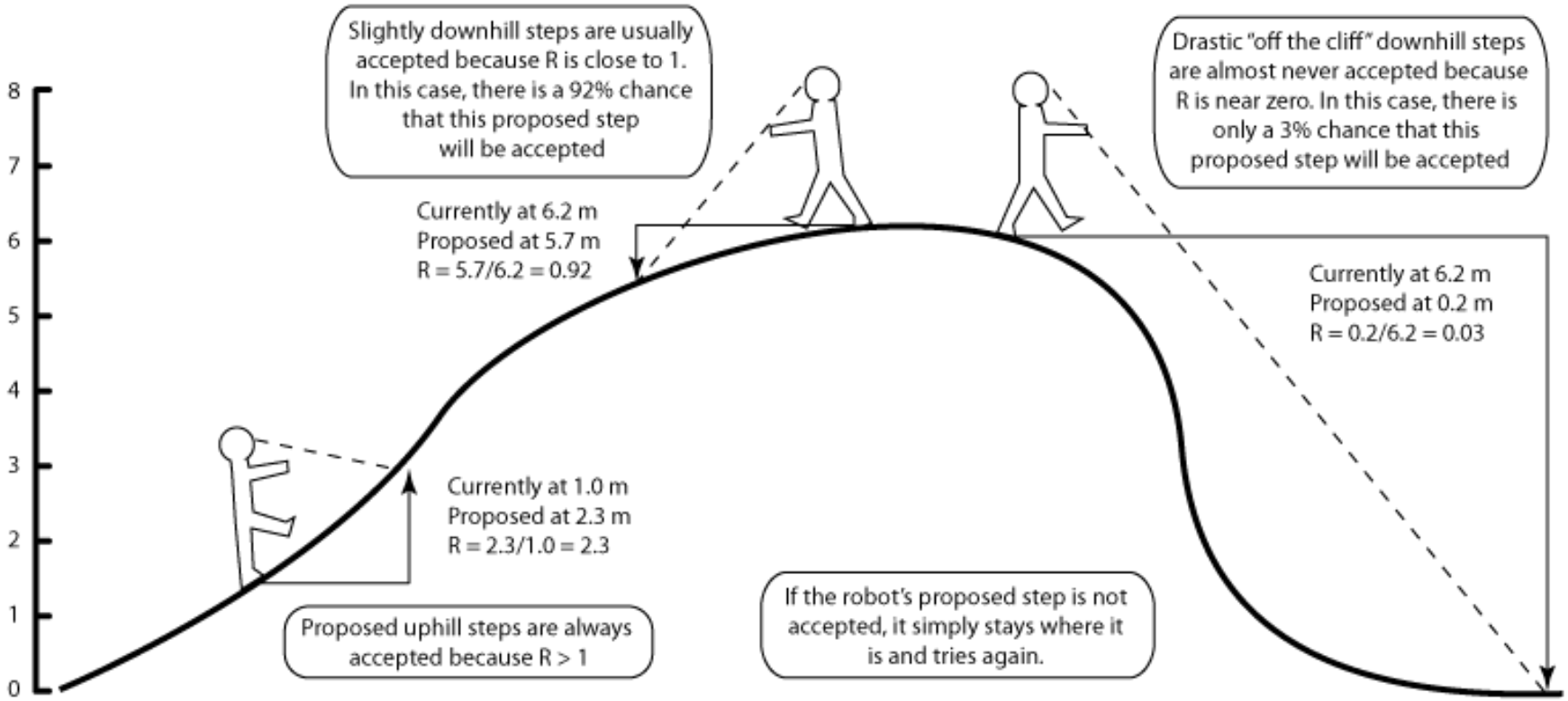


# Trouble getting to the posterior…

$$P(\theta|Data) \propto P(Data|\theta)P(\theta)$$

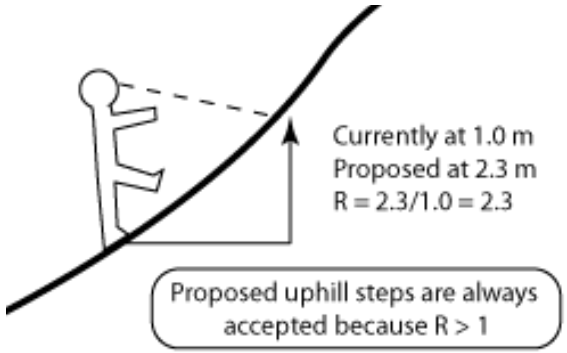
Proportional to

# Inner workings of MCMC



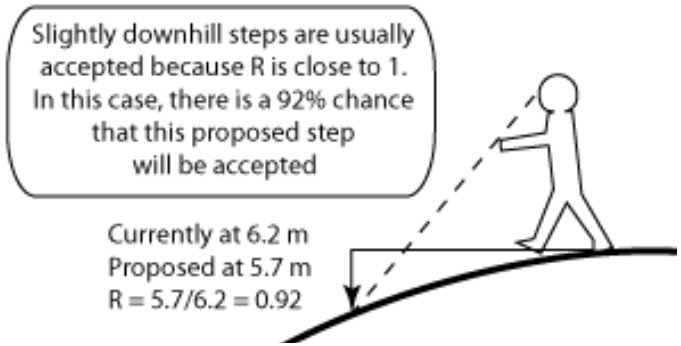

Created by Paul Lewis (U Conn) – http://marple.eeb.uconn.edu/mcmcrobot/?page_id=24

# Inner workings of MCMC



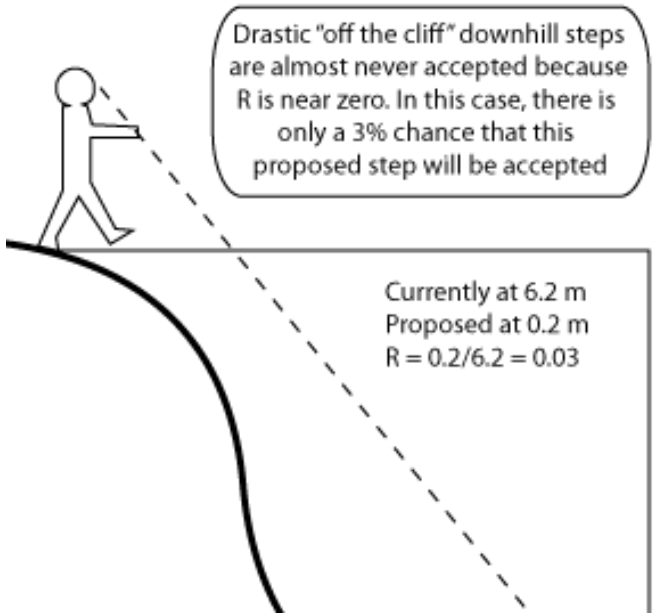

Created by Paul Lewis (U Conn) – http://marple.eeb.uconn.edu/mcmcrobot/?page_id=24

# Inner workings of MCMC



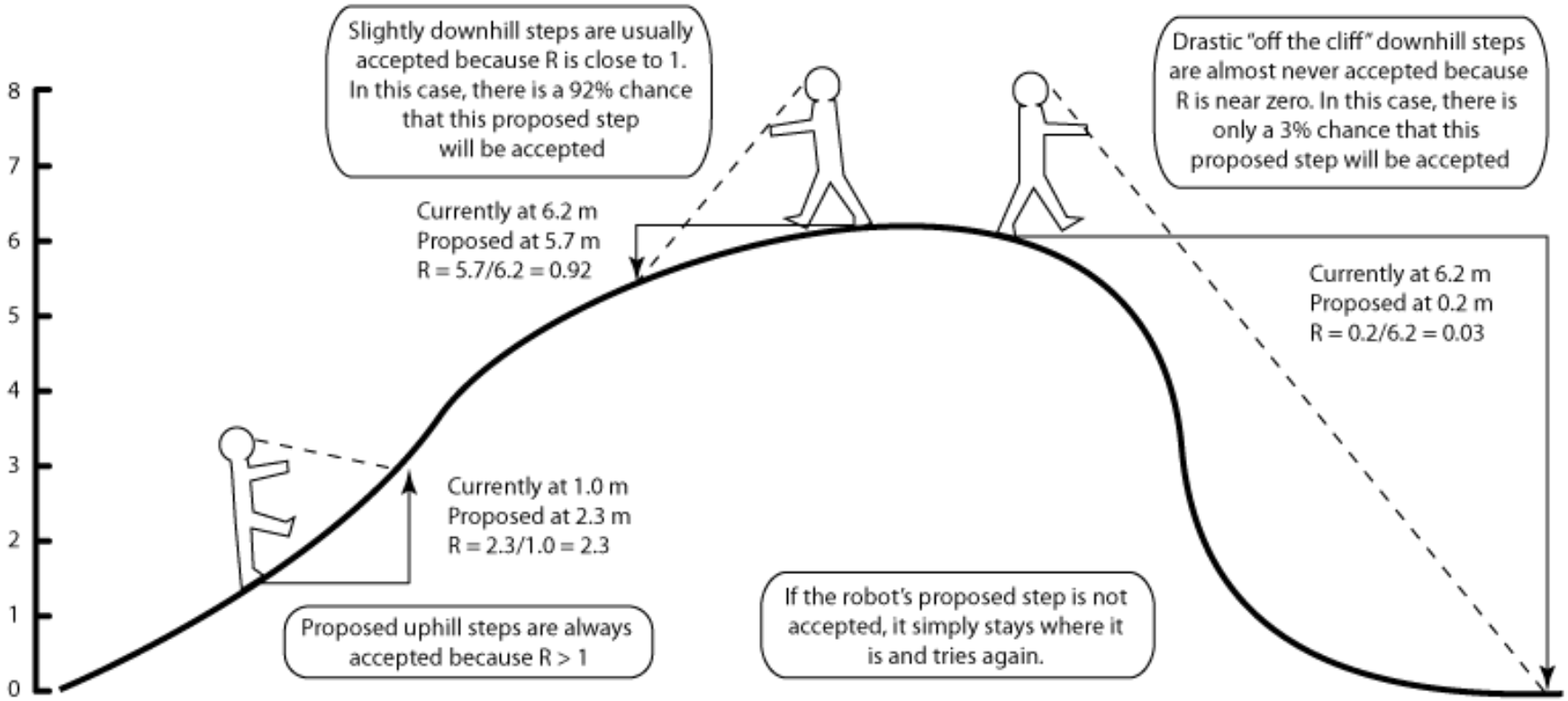

Created by Paul Lewis (U Conn) – http://marple.eeb.uconn.edu/mcmcrobot/?page_id=24

# Inner workings of MCMC




Created by Paul Lewis (U Conn) – http://marple.eeb.uconn.edu/mcmcrobot/?page_id=24

# Inner workings of MCMC




Created by Paul Lewis (U Conn) – http://marple.eeb.uconn.edu/mcmcrobot/?page_id=24

# Trouble getting to the posterior

- http://www.r-bloggers.com/animating-the-metropolis-algorithm/

# GIVING BAYESIAN INFERENCE A GO!

# Exercise

- Is a coin a fair coin?

- DATA:
  - Counts of ‘heads’ and ‘tails’ from *n* coin flips
- MODEL:
  - $C_{HEADS}$ ~ Binomial(size=*n*, probability=θ)
  - θ ~ Beta(α,β)
- OUTPUT:
  - P( θ | $C_{HEADS}$ )

# STRUCTURE is a Bayesian model

- No admixture model:
  - Probability of **Z** and **P** given the data is proportional to the Probability of the Data given **Z** and **P** times the Probability of **Z** times the Probability of **P**

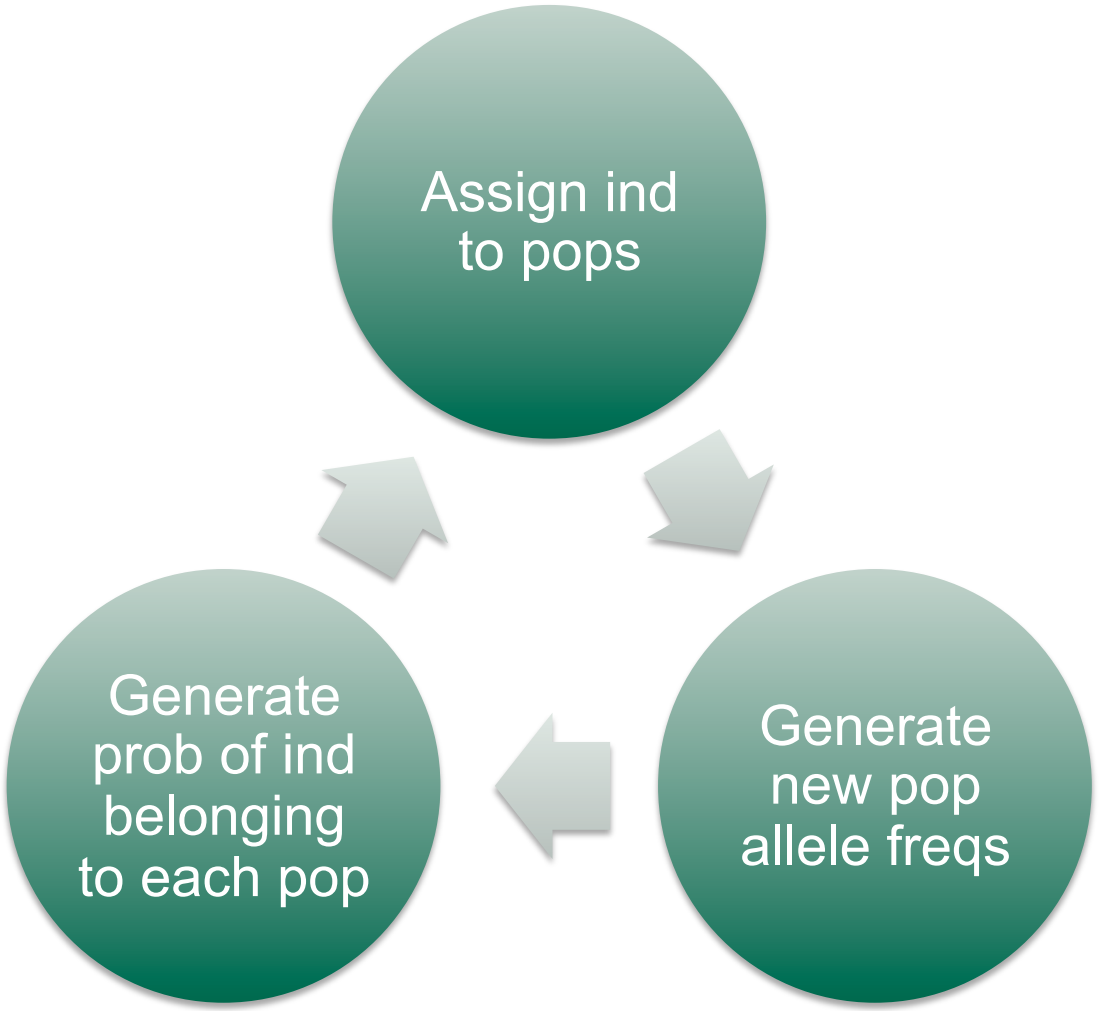
- With admixture model:
  - Probability of **Q**, **Z**, and **P** given the data is proportional to the probability of the Data given **Q**, **Z**, and **P** times the Probability of **Q** times the Probability of **Z** and the Probability of **P**

# RUNNING STRUCTURE…

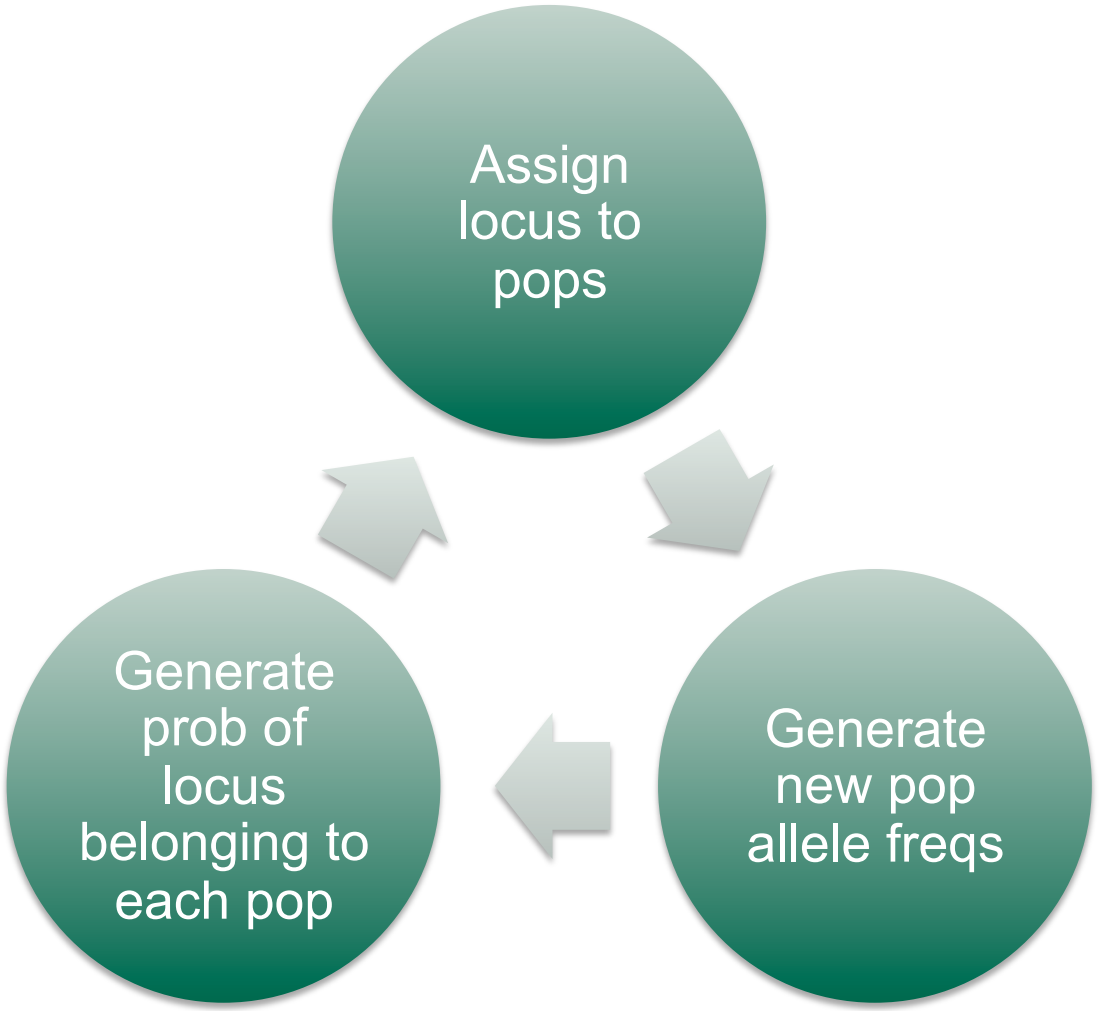
# Further reading


- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.
- Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567–1587.
- Hubisz, M. J., Falush, D., Stephens, M., & Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, *9*(5), 1322–1332.
- Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., & Lareu, M. (n.d.). An overview of STRUCTURE: applications, parameter settings and supporting software. *Frontiers in Genetics*, *4*. doi:10.3389/fgene.2013.00098

# No-admixture model



# Admixture model



- http://www.r-bloggers.com/animating-the-metropolis-algorithm/