# A Bayesian Framework for Analyses of Demographic DNA Sequence Data

Ioanna Manolopoulou[1]

I.Manolopoulou@statslab.cam.ac.uk

Simon Tavaré[1], Steve Brooks[1], Lorenza Legarreta[2]

[1]University of Cambridge

[2]University of East Anglia

UNIVERSITY OF CAMBRIDGE

## The Problem

We have aligned sections of the DNA sequences of $n$ individuals along with their geographical locations. Example:

| Individual | DNA seq | location |
|---|---|---|
| A | ATCGA | (1.3, 2.5) |
| B | ATCGA | (1.7, 3.9) |
| C | ATCCA | (2.9, 0.1) |
| D | CTTGA | (3.1, 6.1) |
| E | CTGAG | (1.3, 2.5) |

One of the questions asked by biologists is how to split our data into significant clusters in terms of their geographical location, so that the results are consistent with the genetic history, and what demographic events occurred in history (e.g. colonisation), including the Most Recent Common Ancestor.

The main method used at the moment is based upon deciding on a unique mutation tree, and identifying significantly different clusters by using an ANOVA-type analysis.

## The Mutation Process

Firstly we model the mutation process. Each sequence of $l$ nucleotides A, G, C, T can be represented by $l$ parallel independent Markov Processes where each position $j$ =1.. $l$ ) has transition matrix:

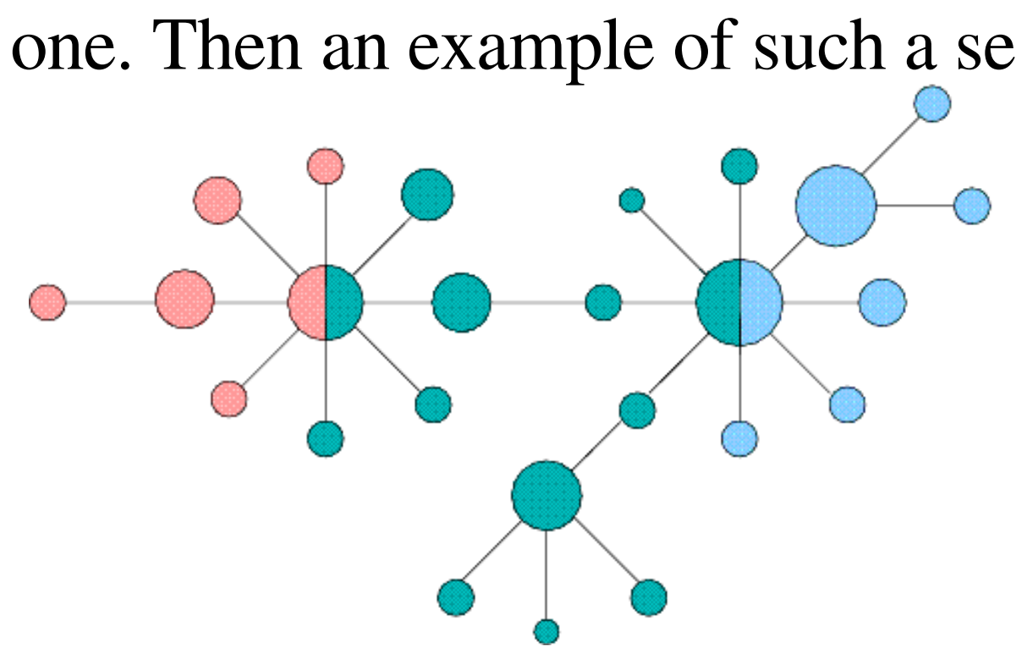
$$Q = \begin{pmatrix} \cdot & \phi_j\pi_G\alpha & \phi_j\pi_C\beta & \phi_j\pi_T\gamma \\ \phi_j\pi_A\alpha & \cdot & \phi_j\pi_C\delta & \phi_j\pi_T\epsilon \\ \phi_j\pi_A\beta & \phi_j\pi_G\delta & \cdot & \phi_j\pi_T\zeta \\ \phi_j\pi_A\gamma & \phi_j\pi_G\epsilon & \phi_j\pi_C\zeta & \cdot \end{pmatrix}$$

Where the $\pi$ represent the stationary probabilities of the chain and the $\phi$ represent the relative mutation rates for each nucleotide position. It can easily be checked that this chain is time-reversible and that $\pi$ is indeed the stationary distribution. So, for each sequence present (with repeats if a sequence is present more than once) we have $l$ simultaneous chains. On top of the mutation process comes the splitting process, representing the reproduction of sequences within the population. Typically the splitting process is much faster than the mutation process, so that the number of times a sequence is found is larger than the number of times is mutates.

By considering a finite set of parsimonious and nearly parsimonious mutational histories, we draw inferences about the mutational tree by introducing it as one of the parameters.

## The Colonisation Model

Now, for a colonisation event, we assume that one or more sequences left a geographical population to start a new one. Then an example of such a set of sequences would be:



Each node here represents a sequence and lines represent mutations. The colour represents the geographical cluster. So, in the figure above, the node which is green/red is a sequence which initially was present in one population (the green), but then colonised to a new one (the red), so that the individuals with that sequence have distinct descendant sequences in the 2 geographical clusters.

The geographical clusters are assumed to have a 2-dimensional Normal Distribution, and hence colonised sequences will have a mixture of Normals distribution

$$\mathrm{MVN}_2(\boldsymbol{\mu}_i, \Sigma_i)$$

Clearly, the parameter space of possible allocations of sequences to clusters increases rapidly with the amount of sampled individuals.

## Problems and Methods

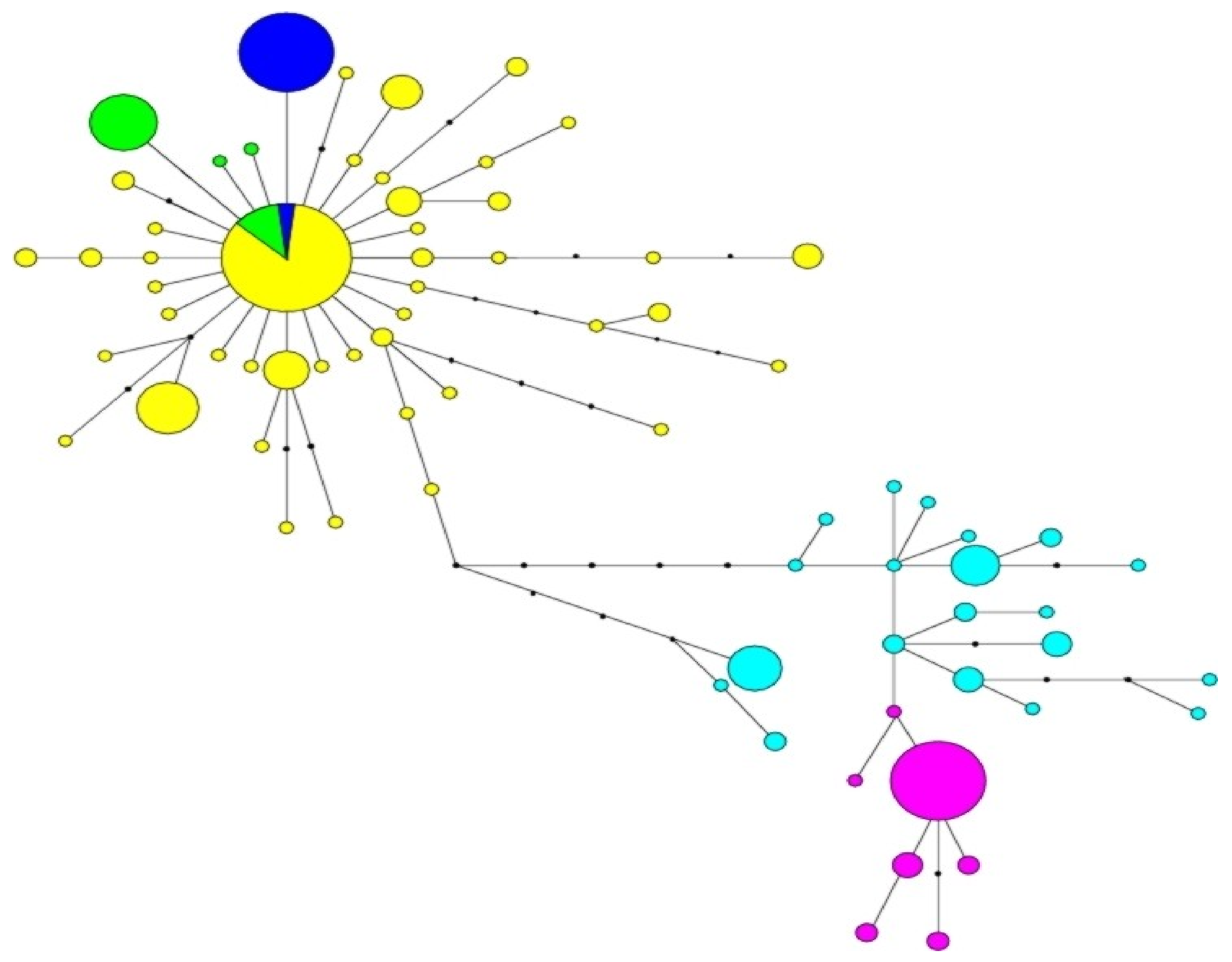
Clearly, the exact mutational history of the sequences is not known. Our probability model for the coalescence process allows us to express the likelihood of a specific scenario, based on the precise order of all splitting and mutation events. A vast number of different such scenarios may result in the same network we are interested in, hence the inference of the precise history is only necessary in order to obtain a measure of the likelihood.

A number of problems arise and are resolved in the following ways:

- The parameter space of precise histories is vast, hence we narrow it down to "feasible" scenarios.
- The exact likelihood of each network is computationally impossible to calculate, and we use a valid approximation which preserves the reversibility of our MCMC. We approximate $\frac{f(D|\theta)}{f(D|\theta')}$ by $\frac{\hat{f}^2(D|\theta)}{\hat{f}^2(D|\theta')} \times \frac{1}{\mathbb{E}\left(\frac{\hat{f}(D|\theta)}{\hat{f}(D|\theta')}\right)}$.
- The parameter space of possible clusterings is vast, so we need efficient proposals: instead of proposing to add individual datapoints to a cluster as in standard clustering problems, we propose to add whole branches.
- There is no obvious way to label and store each possible history, hence we use hashing by chaining based on a one-to-one representation of histories with low-dimensional vectors.
- The number of clusters is unknown, so we use RJMCMC.
- Label-switching affects out analysis, so we use an appropriate labelling criterion.

## Results

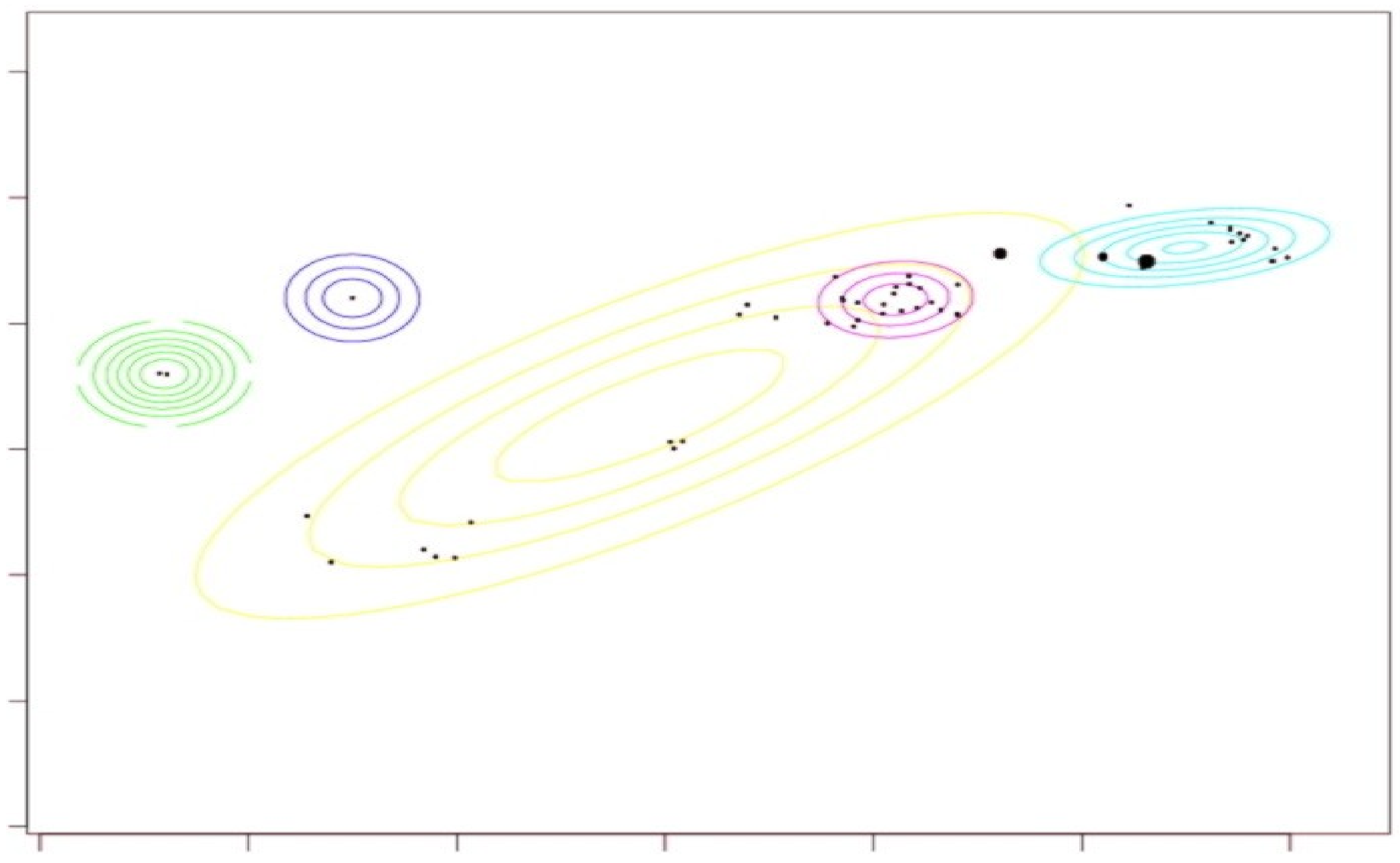
We use a dataset of DNA sequences from weevils in South-West Europe and obtain the following Maximum A Posteriori estimate for the mutational tree:



We see that 2 clusters are distinctly separated, whereas the blue, yellow and green share a node, representing the sequence which colonised the populations.

## Conclusions

Looking at the contour plot of the posterior estimates for the 5 clusters below, and taking into account the ancestral locations (represented by larger dots) we see that the area was colonised from the Rhône area easterly into the alpine region and westerly into the iberian region. The North-West locations are clearly totally isolated from all others.



Our approach gives results accurate with biologists' prior beliefs and provides a quantitative comparison to possible scenarios.

## References & Acknowledgements

[1] Brooks, S.P., Manolopoulou, I. and Emerson, B.C. (2007) "Assessing the Effect of Genetic Mutation – A Bayesian Framework for Determining Population History from DNA Sequence Data". *Bayesian Statistics 8.*

[2] I. Manolopoulou, S. P. Brooks and L. Legarreta. (2007) "A Bayesian Framework for Analyses of Demographic DNA Sequence Data". Proceedings of the 20th Panhellenic Statistics Conference 2007 "Statistics and Society", to appear.

Many thanks to Trinity College for funding this research and B. C. Emerson for his invaluable contribution.