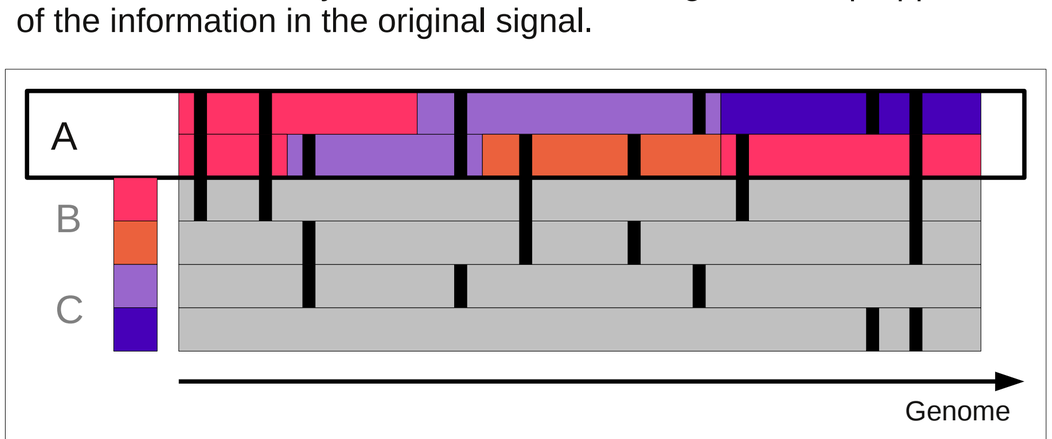
Inference of population structure using dense genotype data

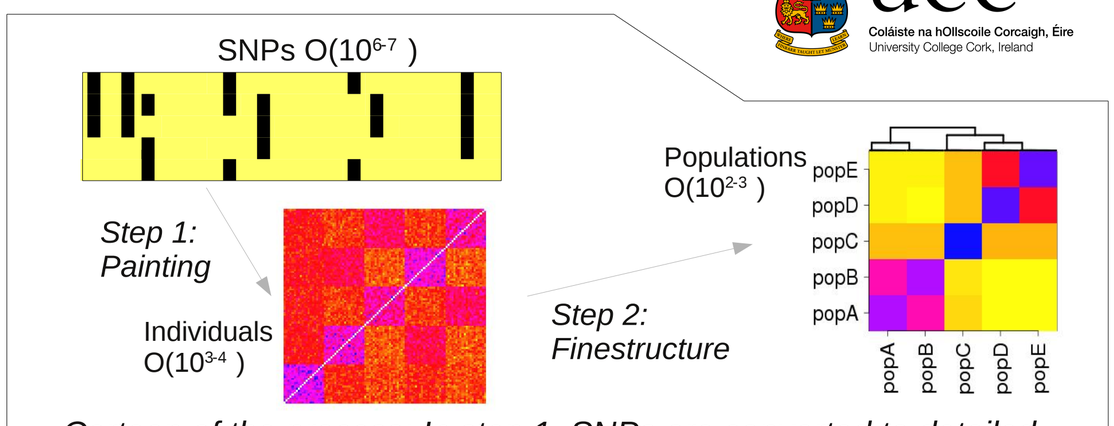
Daniel Lawson (Bristol), Garrett Hellenthal (Oxford), Simon Myers (Oxford), Daniel Falush (Cork)

The Problem:

With millions of Single Nucleotide Polymorphisms (SNPs) and hundreds of individuals, many population samples cannot be dealt with using standard techniques (e.g. STRUCTURE, Pritchard Et al. 2000) which treat each SNP separately. Dense SNPs will be linked by common descent and so provide correlated information about ancestry. What is needed is a data reduction strategy capturing the essential signal, together with a model for ancestry suitable for that data. Population level information can then be used for further analysis. We solve this using a two step approach that keeps almost all of the information in the original signal.



One possible painting for A over a short genome segment with 2 other individuals treated as fixed. A has 'copied' 4 chunks from B and 3 from C. We take the expectation over all paintings and recombination events.



Cartoon of the process: In step 1, SNPs are converted to detailed co-inheritance information. In step 2 we analyse the population structure.

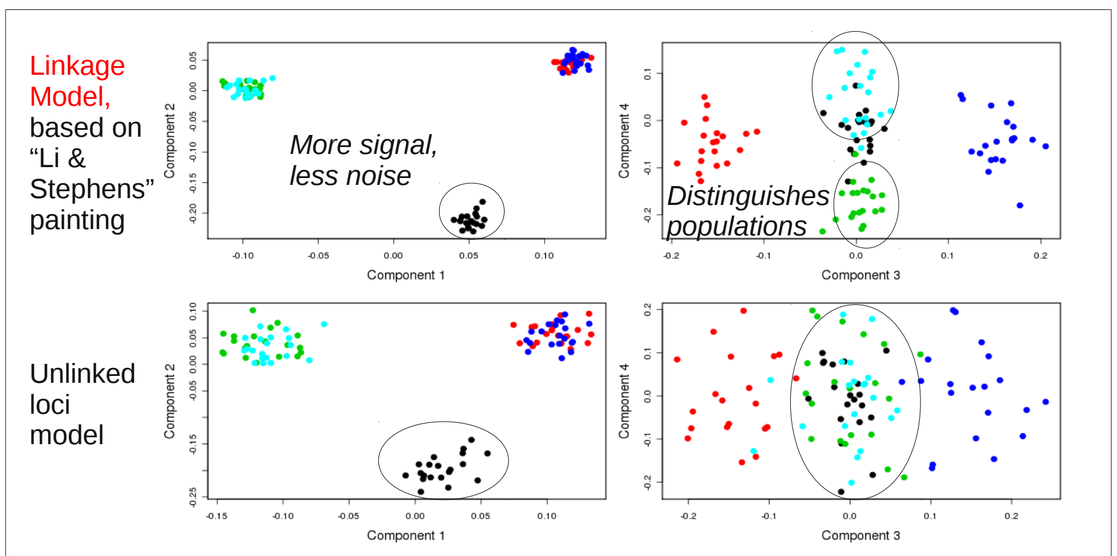
The Solution: step 1, painting

Ancestry information for sexually reproducing organisms is primarily stored as the frequency that particular chunks of DNA are inherited, because recombination occurs more frequently than mutations. We therefore describe an individual by the frequency with which it co-inherits ("copies") a segment of DNA with another individual in the sample. We use a "painting" algorithm following Li & Stephens (2003) to find the recombination points and the individual for which each chunk has the most recent common ancestor. Unlike their algorithm we don't impose an ordering, and we take the expectation over all possible pairings, weighting according to the mutation process.

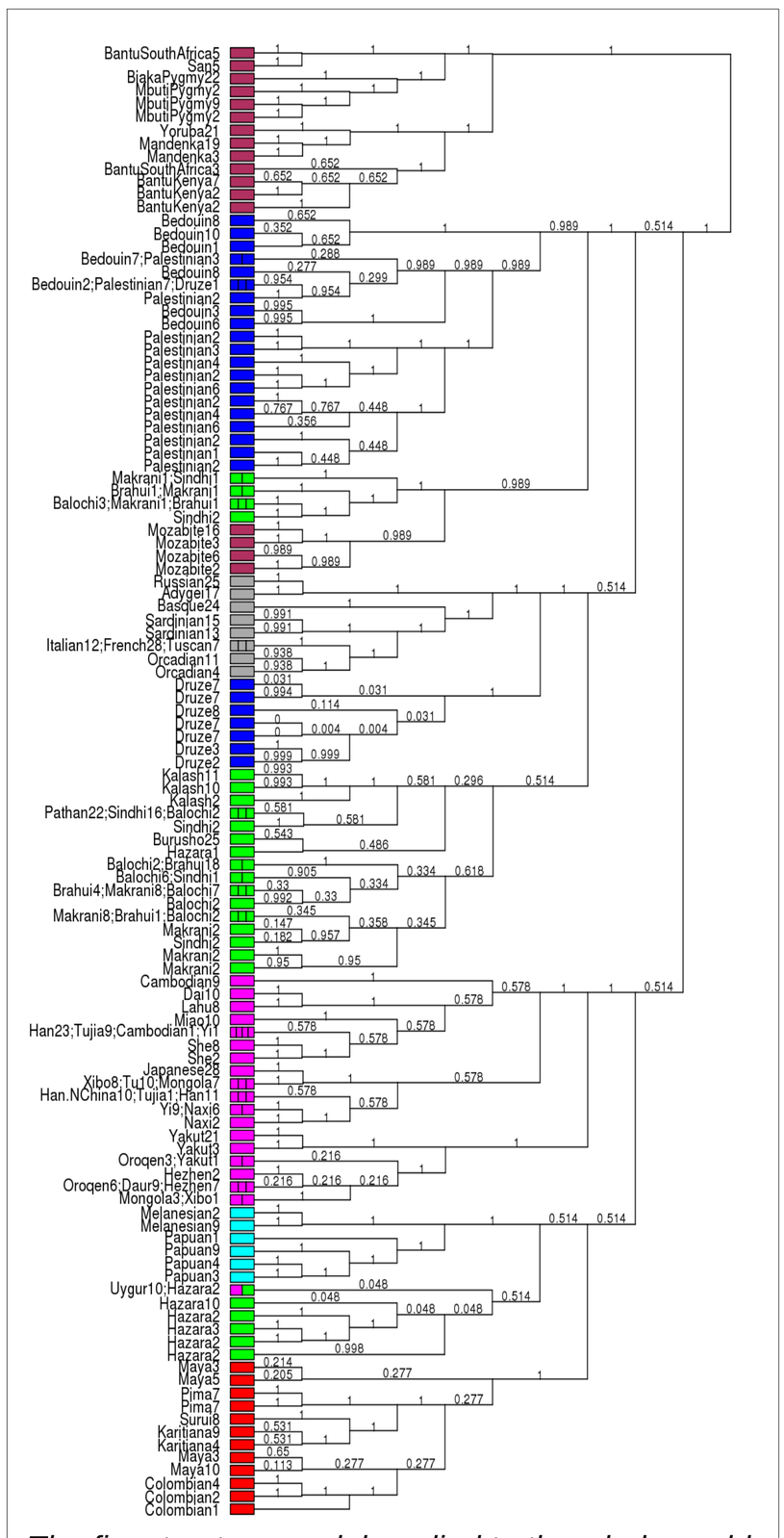
Data representation:

The painting process gives three square matrices of size *NxN* (*N*=number of individuals): the frequency individuals copy chunks from each other, the average length of those chunks, and the mutation rate. These contain different aspects of the ancestry history, and almost completely summarize population ancestry.

We focus on the chunk count matrix. When the loci are unlinked, this matrix is a linear transform of the standard Principal Components Analysis (PCA) matrix. When the loci are linked, our painting correctly accounts for this and produces a "linkage corrected" matrix with greater contrast between populations.



PCA on dense simulated data with population structure. Our "painting" linkage model allows for better resolution of populations.



The finestructure model applied to the whole world. Colours are continents. Labelled populations are shown, as is bipartition certainty.

Step 2, Population structure model

A big advantage of our representation over PCA is that the matrix elements are interpretable. We are looking for a set of populations, each of which has a distinctive chunk copying frequency vector $\boldsymbol{P}_a$, and chunks are copied independently. Therefore the rows of the copy count matrix $\boldsymbol{X}$ are multinomial:

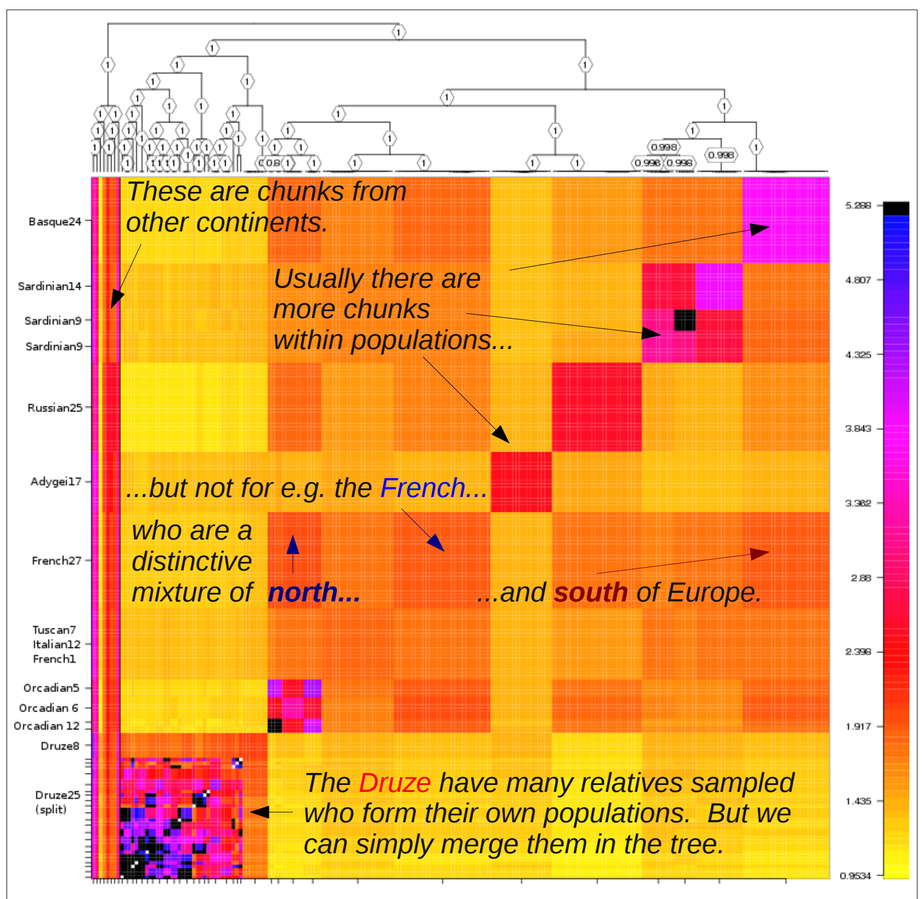
$$p(X|P) = \sum_{i,j=1}^{N} \left(\frac{P_{q_i q_j}}{\hat{n}_{q_j}} \right)^{x_{ij}}$$

q_i =Population assignment if i

Number of individuals in population q_j

and we can apply a (conjugate) Dirichlet Prior for P_a in order to integrate them out (Pella & Masuda 2003) and obtain the probability of a given population split. Markov-Chain Monte-Carlo allows us to integrate over possible population assignments, and is efficient because few parameters must be inferred. The model has the same* power and accuracy as STRUCTURE in large datasets.

* The models are equivalent to first order in N when loci are unlinked, drift is weak and genotyped SNPs are not very rare.



Results for Europe. On the continental scale, our model finds more populations and allows for readable summaries of the signal in the chunk count matrix.

Results: Finestructure

We can apply our method to huge datasets. The HGDP dataset has 938 individuals and ~650,000 SNPs. The painting is relatively fast and is parallelisable over individuals. The population structure model takes $O(N^2)$ iterations to fully converge, hours for a single continent on a desktop PC and a few days for the whole dataset. We find over 180 populations in the world HGDP data, confirmed as real by splitting the genomic data into two halves. We can group them using 'similarity under our model' by forcing merges to obtain a tree; this finds populations that are similar but still distinct. The data can be represented as a 'population level' matrix to see signals of ancestry, and the rich information contained in the 3 data matrices describe may hide historical population structure such as bottlenecks, inbreeding, both ancient and recent admixture, and more.

Important future work: modelling admixture between populations, and accounting for correlations in the drift rate between populations (to find all populations in one go).

Acknowledgements: DL and DF are funded by the Wellcome Trust, DF also by the Science Foundation of Ireland, grant number 05/FE1/B882. This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol - http://www.bris.ac.uk/acrc/. The project also benefited greatly from computing resources funded by a grant from the Wolfson Foundation to Peter Donnelly.