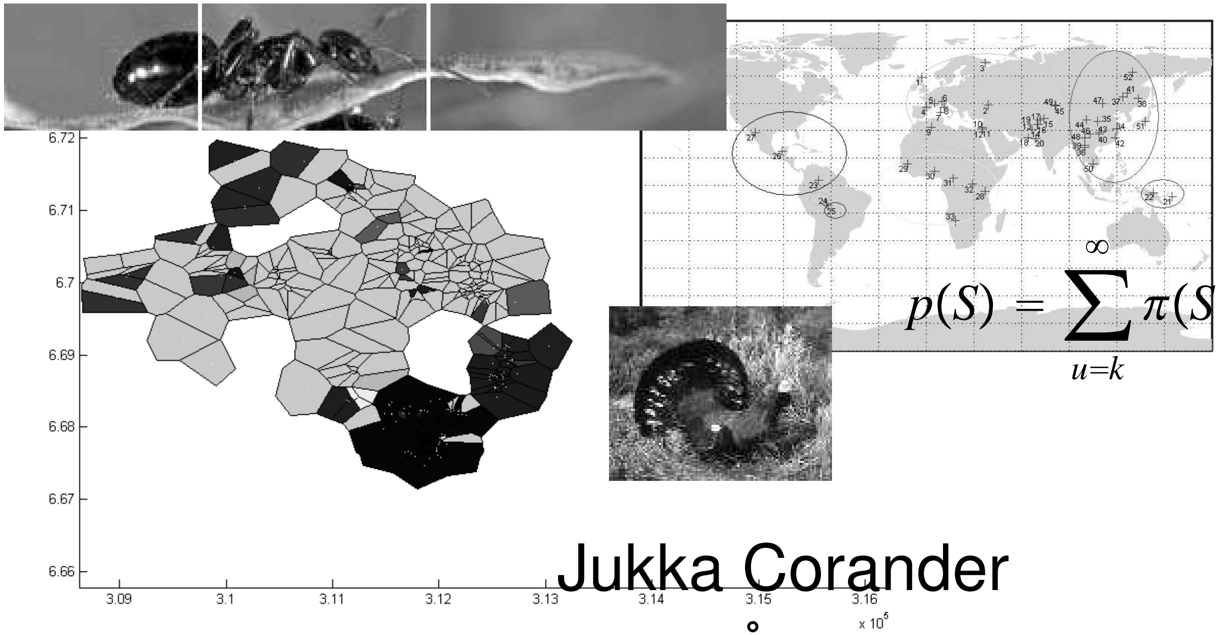
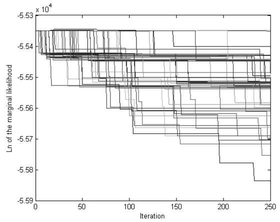
Bayesian analysis of genetic population structure using BAPS





$$p(S) = \sum_{u=k}^{\infty} \pi(S|u)p(u), S \in \mathcal{S},$$

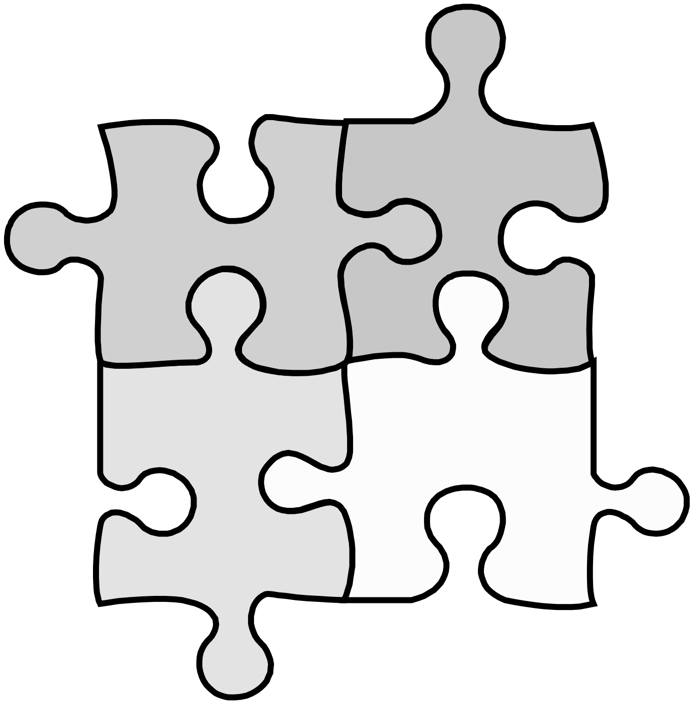
Jukka Corander

Department of Mathematics, Åbo Akademi University, Finland

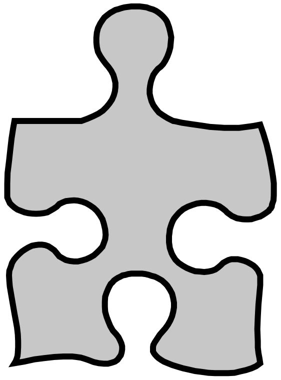
- The Bayesian approach to inferring genetic population structure using DNA sequence or molecular marker data has attained a considerable interest among biologists for almost a decade.
- Numerous models and software exist to date, such as: BAPS, BAYES, BayesAss+, GENECLUST, GENELAND, InStruct, NEWHYBRIDS, PARTITION, STRUCTURAMA, STRUCTURE, TESS...
- The likelihood core of these methods is perhaps surprisingly similar, however, the explicit model assumptions, fine details and adopted computational strategies to performing inference vary to a large extent.

Basis of Bayesian learning of genetic population structure

- Assume that the target population is potentially genetically structured, such that boundaries limiting gene flow exist (or have existed). The extent and shape of such substructuring is typically unknown for natural populations.



- Bayesian models capture genetic population structure by describing the molecular variation in each subpopulation using a separate joint probability distribution over the observed sequence sites or loci.



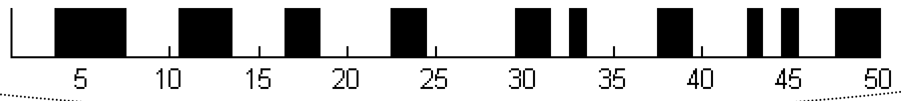
A Bayesian model describing a genetic mixture in data

- Let x_{ij} be an allele (e.g. AFLP, SNP, microsatellite) observed for individual *i* at locus *j*, $j = 1,\ldots,N_L$ (N_L is the total #loci).
- Assume that the investigated data represents a mixture of *k* subpopulations $c = 1,\ldots,k$ (*k* is typically unknown).
- A mixture model specifies the probability $p_c(x_{ij})$ that x_{ij} is observed if the individual *i* comes from the subpopulation *c*.
- Such probabilities are defined for all possible alleles at all loci, and they are assumed as distinct model parameters for all *k* subpopulations.
- These parameters represent the *a priori* unknown allele frequencies of the subpopulations.

- The basic mixture models operate under the assumption that the subpopulations are in HWE.
- The model gets statistical power from the joint consideration of multiple loci, i.e. it can compare the probabilities of occurrence for a particular multi-locus allelic (genotype) profile between the putative ancestral sources.
- The unknown mixture model parameters also include the assignment of individuals to the subpopulations (the explicit assumptions concerning the probabilistic assignment vary over the suggested models in the literature).
- The model attempts to create *k* groups of individuals, such that those allocated in the same group resemble each other genetically as much as possible.
- Notice that mixture models based on a similar reasoning have been in use in other fields (e.g. robotics, machine learning) long before they appeared in population genetics.

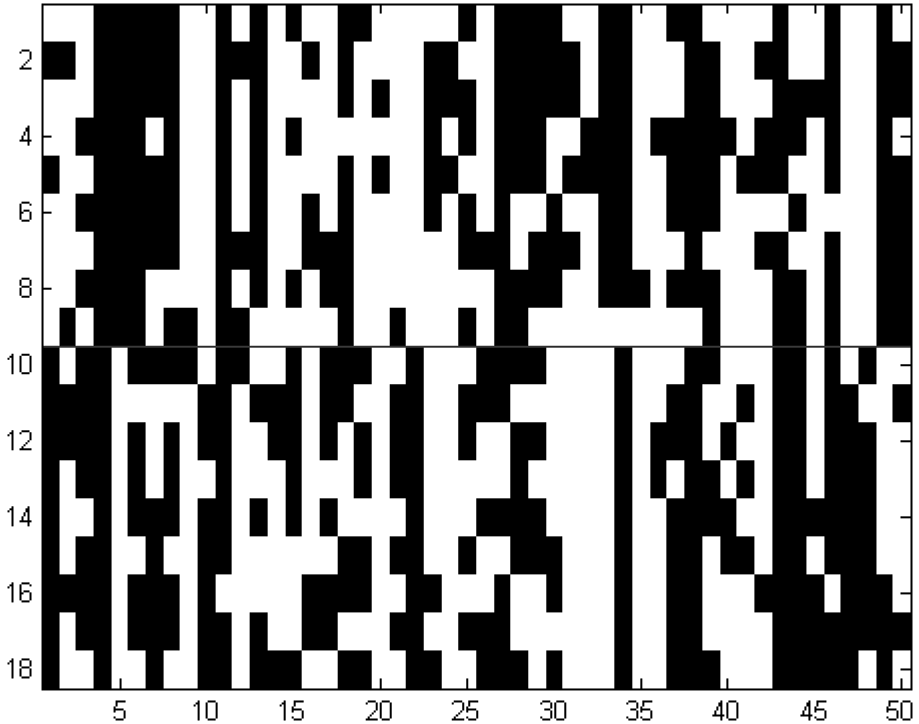
An example of how a mixture model reasons:

Allelic profile of a haploid individual *i* for 50 biallelic loci.



Black/white correspond to the two allelic forms.

Allelic profiles of 9 individuals currently assigned to subpopulation 1.



Allelic profiles of 9 individuals currently assigned to subpopulation 2.

The model calculates the probability of occurrence of an allelic profile conditional on the observed profiles of those individuals already allocated to a particular subpopulation (here 1 or 2).

This way, inference can be made about where allocate individual *i*.

How to do inference with these mixture models?

- Most of the methods in the literature rely on standard Bayesian computation, i.e. MCMC using the Gibbs sampler algorithm.
- This algorithm simulates draws from the posterior distribution of the model parameters conditional on the observed data.
- It cycles between the allele frequency and allocation parameters, and it can also handle missing alleles by data augmentation.
- In the MCMC literature Gibbs sampler is known to have convergence problems, in particular when the model complexity increases (#individuals, #subpopulations, #loci).
- Also, Gibbs sampler is computationally slow and as it assumes a fixed value of *k*, inference about the number subpopulations supported by a data set becomes more difficult.

What about BAPS?

- The genetic mixture modelling options in the current BAPS software are built on a quite different approach compared to the ordinary latent class model.
- The BAPS mixture model is derived using novel Bayesian predictive classification theory, applied to the population genetics context.
- Also, the computational approach is different and it utilizes the results on nonreversible Metropolis-Hastings algorithm introduced by Corander et al. (Statistics & Computing 2006).
- A variety of different prior assumptions about the molecular data can be utilized in BAPS to make inferences.

The partition-based mixture model in BAPS

- Let $S = (s_1,\ldots, s_k)$ represent a partition of the n observed individuals into k non-empty classes (clusters).
- Let the complete set of observed alleles over all individuals be denoted by $\mathbf{x}^{(N)}$, i.e. this set reflects the presence of any missing alleles among the individuals.
- In the Bayesian framework it is acknowledged that all uncertainty faced in any particular situation is to be quantified probabilistically.
- Such a quantification may be specified as the following probability measure for the observed molecular data, where the summation is over the space of partitions:

$$p(\mathbf{x}^{(N)}) = \sum_{S\in\mathbf{S}} p(\mathbf{x}^{(N)} \mid S)\, p(S)$$

- Here $P(S)$ describes the *a priori* uncertainty about *S, i.e.* the genetic structure parameter
- Further $p(\mathbf{x}^{(N)}|S)$ is the (prior) predictive distribution of the marker data given the genetic structure
- This means that the statistical uncertainty about the allele frequency parameters for each subpopulation in S (the $p_c(x_{ij})$'s) has been acknowledged in $p(\mathbf{x}^{(N)}|S)$ by integration w.r.t a prior probability distribution for them (product Dirichlet distribution).
- From the perspective of statistical learning concerning the genetic structure, we are primarily interested in the conditional distribution of S given the marker data (*i.e.* the posterior), which is determined by the Bayes' rule:

$$p(S|\mathbf{x}^{(N)}) = \frac{p(\mathbf{x}^{(N)}|S)p(S)}{\sum_{S\in\mathcal{S}} p(\mathbf{x}^{(N)}|S)p(S)}.$$

- The prior predictive distribution $p(\mathbf{x}^{(N)}|S)$ for the observed allelic data given a genetic structure S can be derived exactly (Corander et al. 2007, BMB) under the following assumptions:

(1) Assume that each class (cluster) of S represents a "random mating unit".

(2) Assume loci to be unlinked and to be representable by a fixed number of alleles.

- Then, $p(\mathbf{x}^{(N)}|S)$ has the form:

$$p(\mathbf{x}^{(N)} \mid S) = \prod_{c=1}^{k} \prod_{j=1}^{N_L} \left[\frac{\Gamma(\sum_l \alpha_{jl})}{\Gamma(\sum_l (\alpha_{jl} + n_{ijl}))} \prod_{l=1}^{N_{A(j)}} \frac{\Gamma(\alpha_{jl} + n_{ijl})}{\Gamma(\alpha_{jl})} \right]$$

Explaining the model further…

- The model investigates whether there is evidence for subgroups that have genetically drifted apart.
- This reflects the assumption of neutrality of the considered molecular information.
- The Bayesian predictive nature of the model means that both the level of model complexity and predictive power are always taken into account, when two putative genetic structures are compared to each other.
- The presence of missing alleles in the allelic profile of an individual *i* is reflected by a flatter marginal posterior distribution over the possible subpopulation assignments for *i*.
- Handling the posterior is computationally an extreme problem in general.

The number of partitions as a function of n.

n		#S	n		#S
1		1	11		678570
2		2	12		4213597
3		5	13		27644437
4		15	14		190899322
5		52	15		1382958545
6		203	16		10480142147
7		877	17		82864869804
8		4140	18		682076806159
9		21147	19		5832742205057
10		11597	20		51724158235372

Some advantages of the stochastic partition -based approach:

- Avoids Monte Carlo errors related to the allele frequency parameters, which is particularly important for small subpopulations.
- Enables directly inference about the #subpopulations, as *k* is not assumed fixed in the model.
- Allows for fast adaptive computation in learning of the genetic population structures supported by the data.
- Enables a flexible use of different types of priors for the structure and allele frequency parameters.

Currently available mixture learning options in BAPS (v5.1):

1. User specifies an upper bound *K* (or a set of upper bound values) for #subpopulations *k*, whereafter the algorithm attempts to identify the posterior mode partition in the range 1≤*k*≤*K* (or 1≤*k*≤max(*K*) if multiple values are used).
2. User specifies a fixed *k* and the algorithm attempts to identify the *a posteriori* most probable partition having exactly *k* subpopulations.
3. User specifies a finite set of any *a priori* hypotheses about S (i.e. distinct configurations of the genetic structure) and the program calculates the posterior probabilities for each of them.
4. User provides an additional auxiliary data from a set of known baseline populations and specifies an upper bound *K* for the #subpopulations *k*, whereafter the algorithm attempts to identify for the current data the posterior mode partition conditionally on the baseline data (in the same range of values as in the option 1.)

Genetic mixture estimation in BAPS

- Given either the upper bound K or the ’Fixed k’ specification, BAPS uses repeatedly the following stochastic search operators to identify posterior mode partition:
- Given the current partition, attempt to re-allocate every individual in a stochastic order to improve $p(\mathbf{x}^{(N)}|S)$.
- Given the current partition, attempt to merge clusters to improve $p(\mathbf{x}^{(N)}|S)$.
- Given the current partition, attempt to split clusters in an intelligent manner to improve $p(\mathbf{x}^{(N)}|S)$ (random splits extremely inefficient!)

Genetic mixture estimation in BAPS

- The stochastic search for the posterior mode should be repeated to decrease the probability of identifying a local mode.
- No search method exists (apart from complete enumeration) that can be guaranteed to find posterior optimum in a FINITE number of steps (limit results).
- Given our accumulated experience, BAPS is highly efficient even for very large and complex data sets (say ~5500 individuals and ~130 clusters).
- In the output BAPS provides a measure of local uncertainty around the optimal partition (log changes of $p(\mathbf{x}^{(N)}|S)$ when moving data into other clusters).
- The ‘Fixed *k*’ option may also be utilized to find posterior optima for alternative values of *k* around the estimate of the global optimum.

Choosing an appropriate prior for the partition parameter (= choosing clustering model in BAPS)

- BAPS software contains five variations of the genetic mixture model, which are based on different biological sampling scenarios:

1. Individuals sampled dispersely from the population without any relevant geographical information. Choose ‘Clustering of individuals’, (or ‘Clustering with linked loci’ depending on data).
2. Individuals sampled from a number of chosen geographically limited areas (commonly used in population genetics to enable use of F-statistics). Choose either ‘Clustering of groups of individuals’ or ‘Clustering of individuals’ (or even both), depending on the properties of the molecular data. If very small #loci is available, then ‘Clustering of individuals’ is not a statistically sound option. The most extreme case is ‘#loci = 1’, where a mixture model clustering just individuals is not even identifiable. The option ‘Clustering with linked loci’ can also handle pre-grouping of the data.

Choosing an appropriate prior for the partition parameter (= choosing clustering model in BAPS)

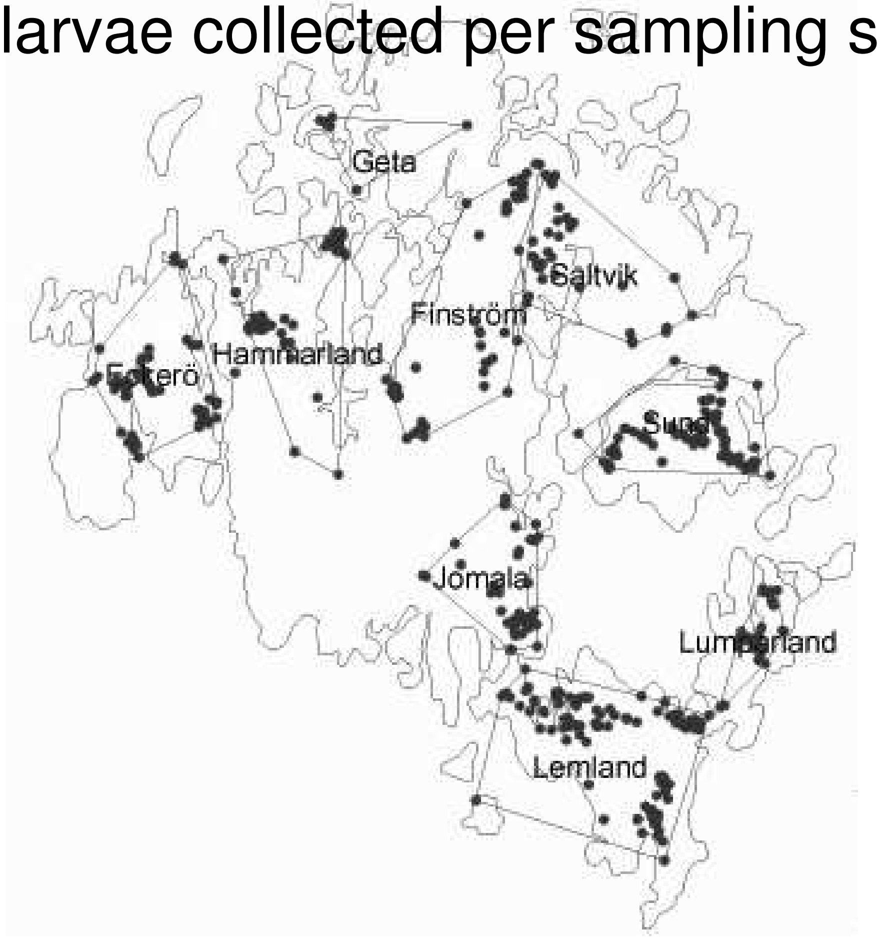
3. Individuals sampled fairly continuously from the population with relevant known geographical coordinates. Choose ’Spatial clustering of individuals’.
4. Groups of individuals known to belong to the same deme are sampled with relevant known geographical coordinates for the group. Choose ’Spatial clustering of groups’. As in the 2nd option, this choice enables a statistically better handling of small #loci.
5. Two types of samples available, one with known genetic origins (baseline sample) and one without (current sample). Choose ’Trained clustering’ to let the baseline data to be used for updating knowledge about allele frequencies in the baseline populations. Notice that individuals in the current sample may either be forced to be assigned to some of the baseline populations or allowed to become members of new clusters outside the baseline.

Why bother with the choice of the prior?

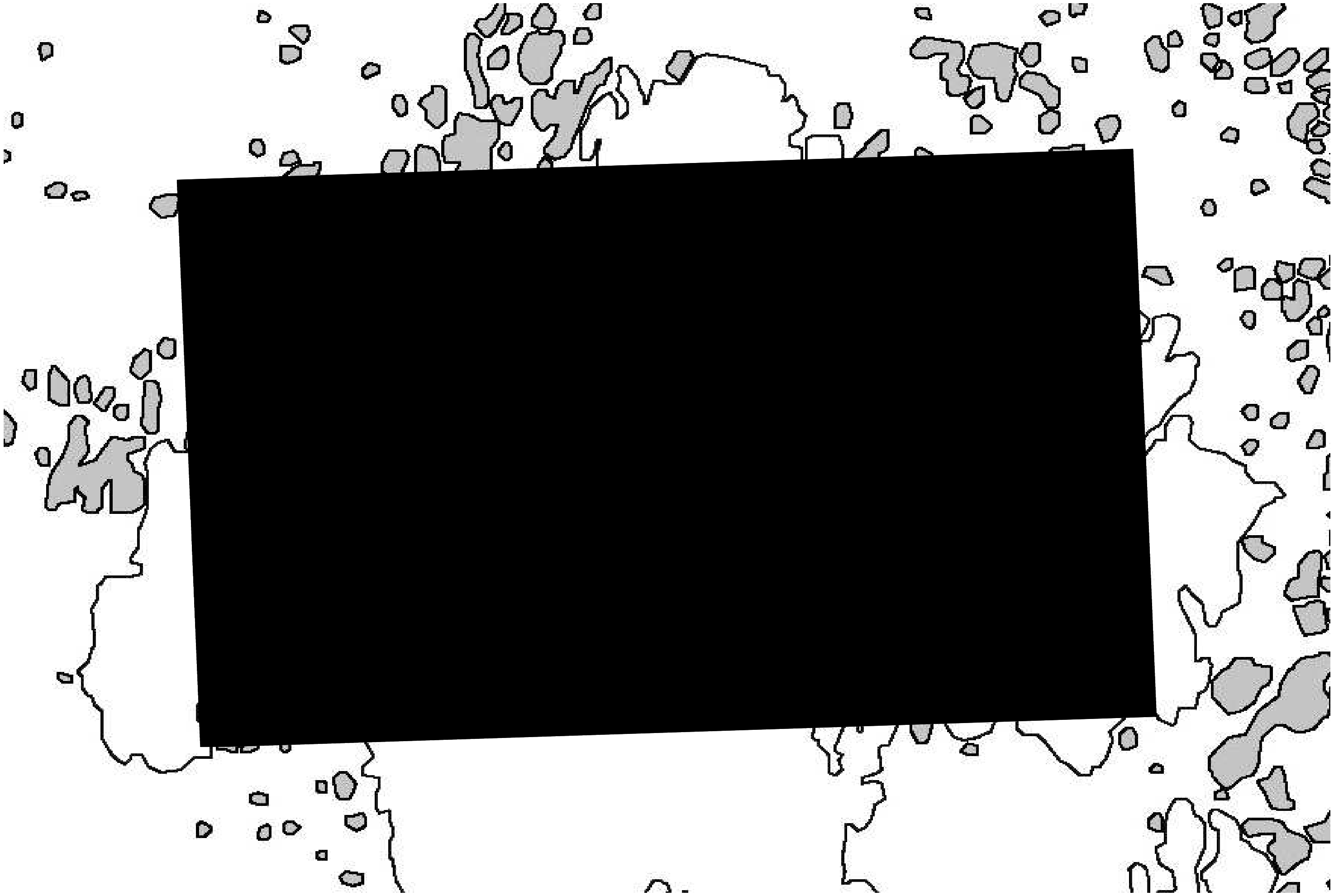
- Use of an appropriate prior may strengthen the inferences considerably.
- For sparsely informative genetic markers, the geographical information is highly useful.
- This applies both to the ’Spatial clustering’ as well to ’Clustering of groups of individuals’.
- The rationale for the latter is that the prior can bind individuals together in the mixture model when the markers are too weak to do that appropriately.
- In the ’Clustering of groups’ options the mixture model investigates the observed allele frequencies (counts) of the pre-grouped data and targets to identify those groups where there is enough statistical evidence to claim that the underlying allele frequencies differ.

Some examples of genetic mixture analyses

Spatial clustering of groups of individuals, where each group corresponds to a small breeding region (meadow) for the Glanville fritillary metapopulation at Åland Islands (Ilkka Hanski’s group). DNA extracted from a small number of larvae collected per sampling site.



Genetic structure estimated using 4 microsatellites and 10 SNPs (Orsini et al. Mol Ecol 2008).

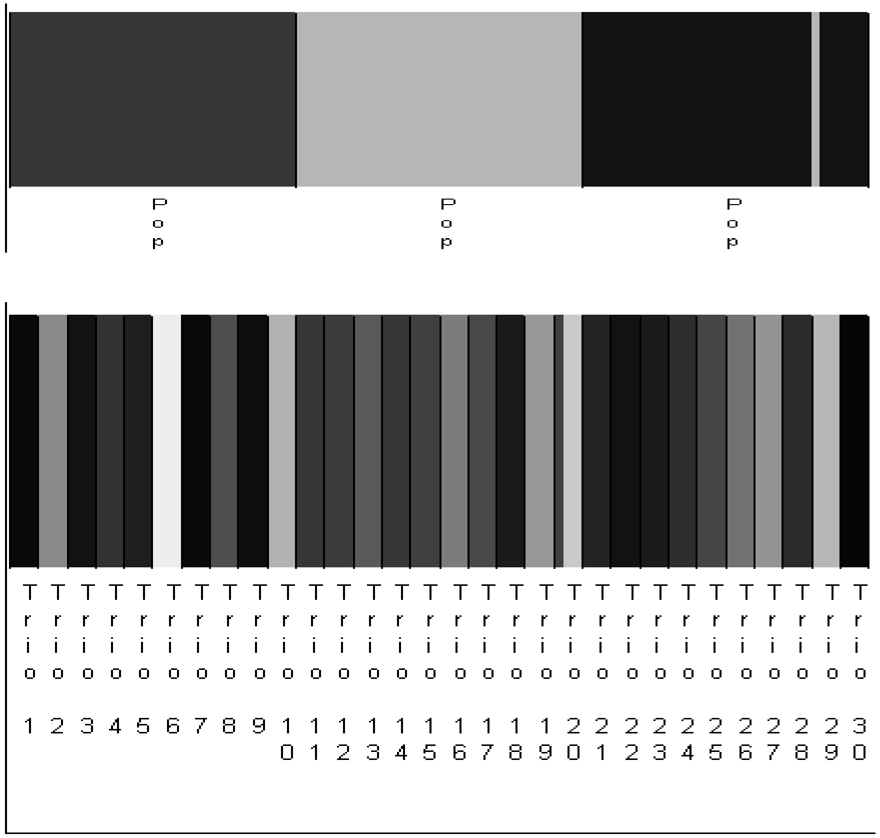


[redacted]

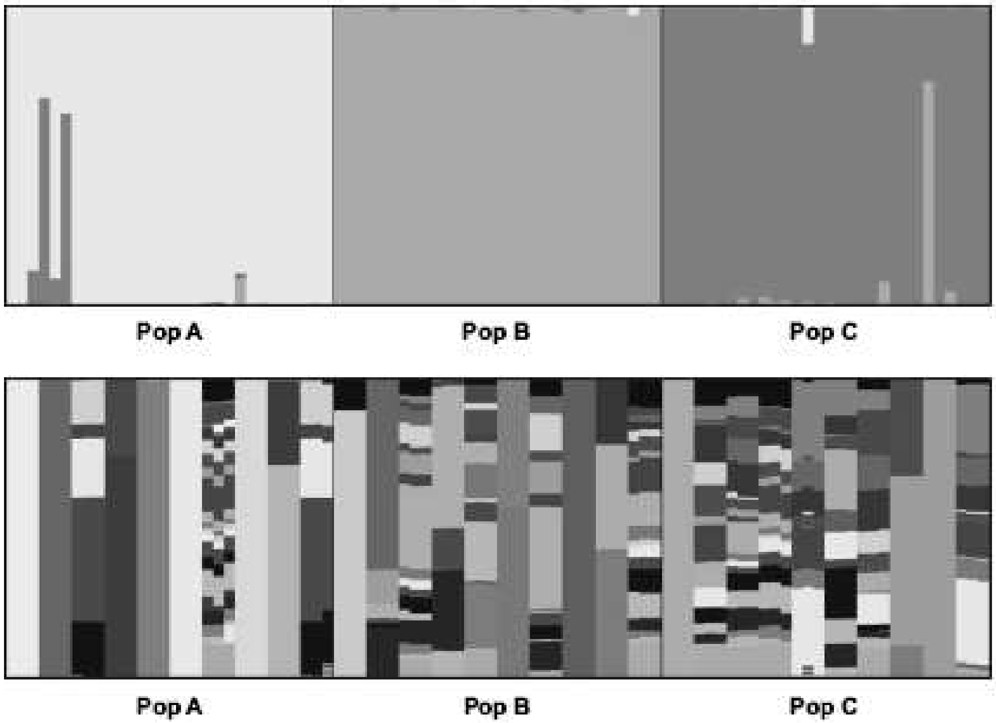
Results for simulated human data using the ‘Fixed k’ option. Data taken from Gasbarra et al. Theor Pop Biol 2007; consists of 15 unlinked microsatellites, 3 populations, each consisting of 10 trios of siblings.

k = 3

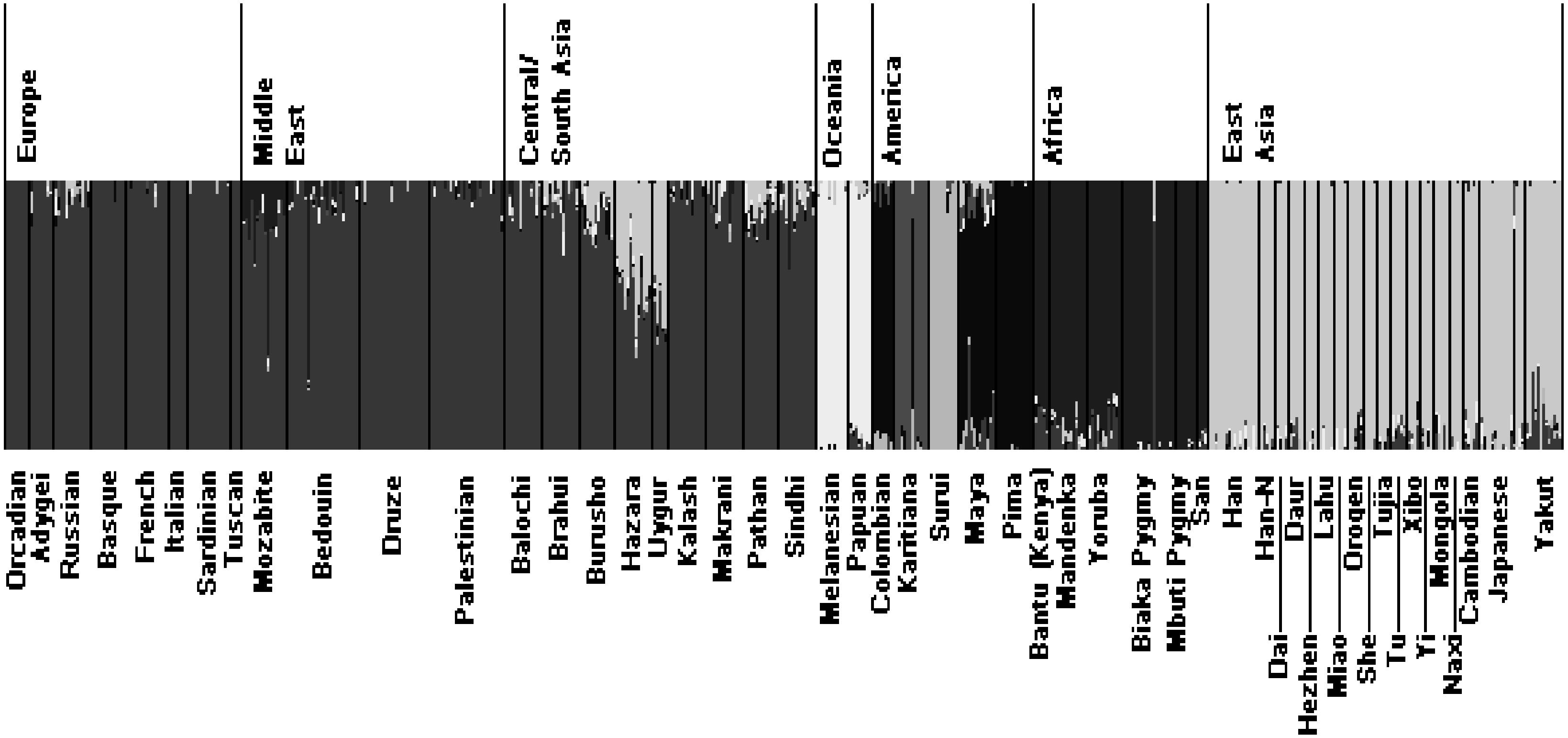
k = 30



For comparison purposes the STRUCTURE results obtained by Gasbarra et al. for the same data (no admixture model):



Results of ‘Clustering of individuals’ for large human data from Corander and Marttinen 2006, Mol Ecol. The data contains 1056 individuals and 377 microsatellite loci. Admixture results are shown below.

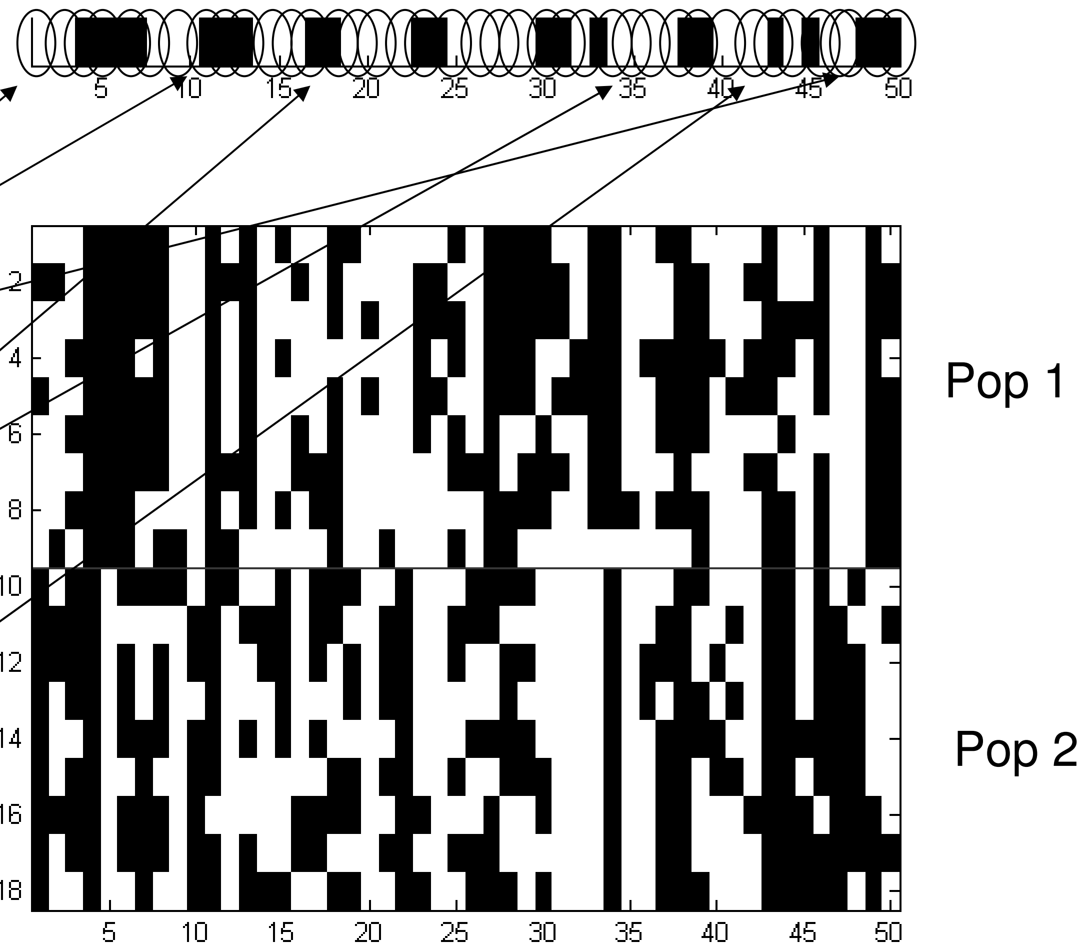


Going over to admixture models…

- If the amount of available molecular information is sufficiently large, it is biologically relevant to consider also questions of admixed ancestry, in addition to genetic mixture modelling.
- Some examples in the literature show that admixture inference may behave very spuriously when pushing the limits too far (say only 6 loci available).
- Let us try to understand the statistical rationale behind genetic admixture modelling...

An example of how an admixture model reasons:

The 50 observed alleles of a haploid individual *i* for 50 biallelic loci.



Examples of alleles having likely ancestry in subpopulation 1.

Examples of alleles having likely ancestry in subpopulation 2.

Examples of alleles having equally likely ancestry in subpopulations 1 and 2.

Each allele is assigned to an ancestral subpopulation according to the conditional probability of observing it there. The conditioning is based on the alleles already allocated to that particular subpopulation (here 1 or 2).

This way, inference can be made about where to allocate all the alleles of individual *i*.

More about admixture models

A (latent class) admixture model considers in a sense the *k* putative ancestral origins as baskets, where the alleles of an individual may be placed.

The proportion of alleles in each such basket may be represented by an admixture coefficient, such that the sum of them equals unity.

$$q_i = (q_{i1}, \ldots, q_{ik}), \sum_{c=1}^{k} q_{ic} = 1$$

If we consider a diploid individual and N_L loci, each allele represents 100/(2N_L)% of the observed genomic characters. Thus, the amount of loci determines the resolution at which an admixed ancestry can be considered.

For instance, with 10 microsatellite loci for a diploid species, each allele represents 5% of the genome in the admixture inference. Thus, with a small number of loci, it is not feasible to reliably make detailed statements at very fine scale

How to do inference with such admixture models?

- MCMC-based approach using the Gibbs sampler algorithm cycles between the allele frequency and allocation parameters for each allele (earlier allocation was done at the level of individuals).
- The posterior estimates of the admixture coefficients are then typically based on the relative number of times an individual's alleles visit a particular basket (=ancestral subpopulation) during the simulation.
- This enables a numerical approximation of the posterior probability distribution for the admixture coefficients.

Challenges with the admixture inference

- Convergence problems for the Gibbs sampler applied to admixture models are even more serious than for genetic mixture models.
- The admixture model is burdened by a serious unidentifiability problem, when both the admixture coefficients AND the #ancestral sources (=*k*) are inferred simultaneously.
- This causes a strong dependence on the prior and can lead to overfitting (too large a *k* is inferred).

Challenges with the admixture inference

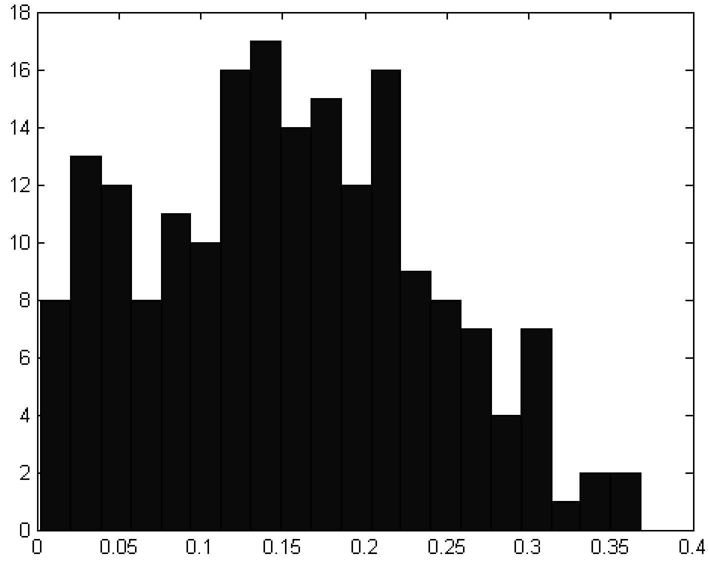
- Posterior distribution of the admixture coefficients does not necessarily reflect directly their statistical significance, i.e. the posterior may be centered away from zero for a particular ancestral subpopulation, although the individual is not admixed.
- The reason for this behavior is actually quite simple.
- Consider two ancestral populations that have diverged only moderately in genetic terms (e.g. Fst ~.05 -.10).

Challenges with the admixture inference

- Assume that 20 microsatellites are used for inferring the parameters of the admixture model for a set of diploid individuals.
- Consider any particular individual *i* with non-admixed ancestry in subpopulation 1, whose alleles are to be assigned to ancestral origins in an iteration of the Gibbs sampler algorithm.
- Now, given the moderate difference between the ancestral origins, it is quite unlikely that for EVERY allele x_{ij} of *i,* the probability $p_c(x_{ij})$ is considerably higher for $c = 1$ than for $c = 2$.
- In fact, it is not unreasonable to expect that $p_2(x_{ij})$ is higher than $p_1(x_{ij})$ for, say 4, alleles out of the 40 observed.

Challenges with the admixture inference

- Then, it is likely that these alleles are assigned to subpopulation 2, which would correspond to $q_2 = .1$.
- Unfortunately, even higher spurious values of the admixture coefficients can be expected by chance.
- Observe that the same phenomenon would persist even if the #loci is very large, e.g. we have investigated this in detail by a simulation scenario comparable to the 377 microsatellite human data set shown earlier.
- Below is an example of the posterior distribution (histogram approximation) for a spurious admixture coefficient.

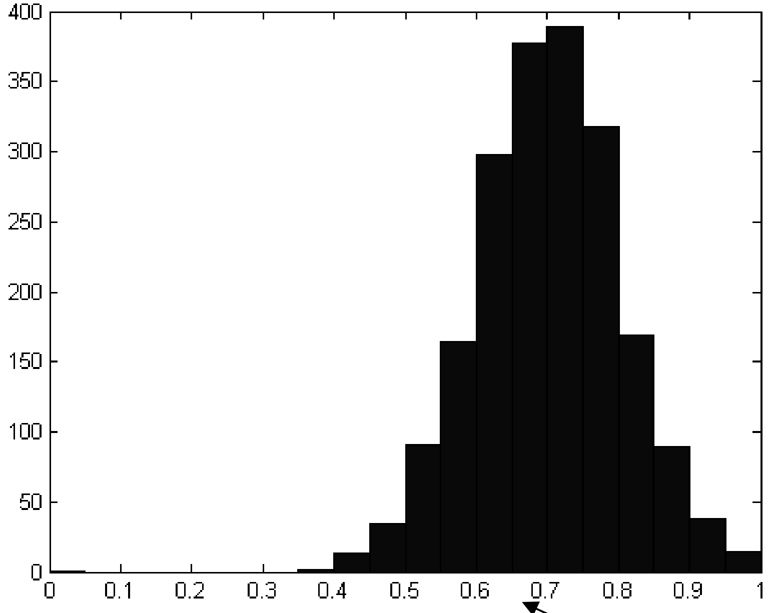


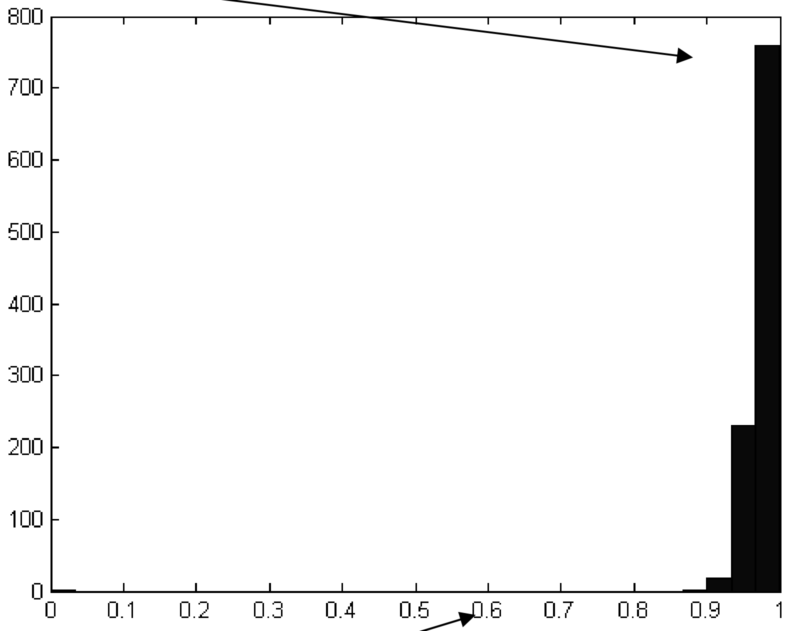
How do we deal with these problems in BAPS?

- Genetic mixture, i.e. *k* and the corresponding clustering are estimated first. Alternatively, user can specify the underlying ancestral populations (BUT THEY SHOULD THEN REALLY BE GENETICALLY DISTINCT!).
- This solves the problem with the weak admixture model identifiability.
- Conditional on the inferred or given ancestral populations, admixture coefficients are inferred for each individual using the posterior mode estimate calculated with a Monte Carlo simulation combined with standard numerical optimization.
- Significance level for the admixture is obtained using a new Monte Carlo simulation, where non-admixed reference individuals are simulated from the populations and admixture coefficients are estimated for them.
- Eventual missing data is dealt with by randomly deleting approapariate amount of alleles from the multilocus genotypes of the reference individuals.

How do we deal with these problems in BAPS?

This way a null distribution is obtained for the amount of admixture expected by chance and the resulting p-value is simple to calculate.



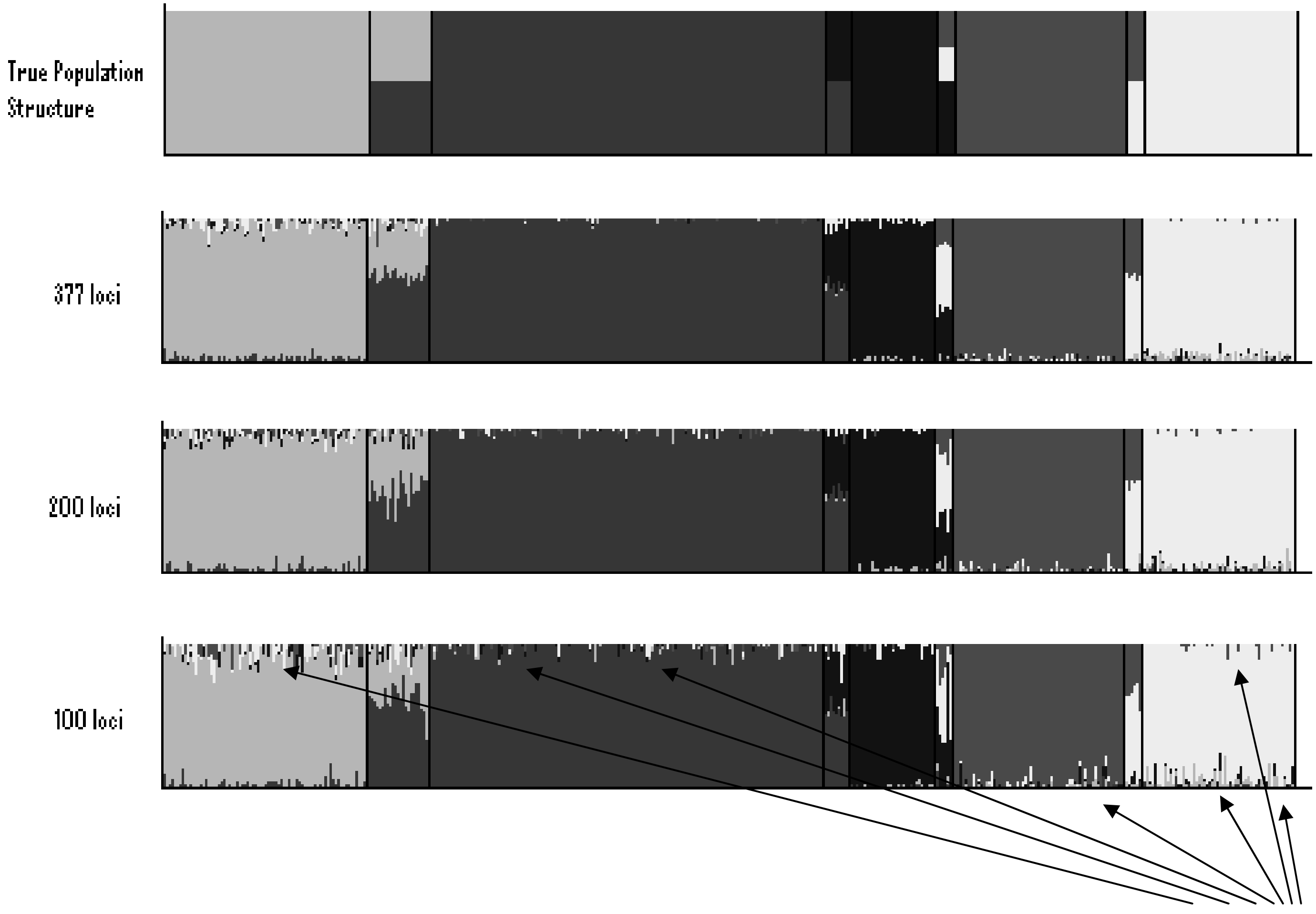


Estimated value of q_i for the ”home” population, i.e. the population where individual *i* was assigned in the inferred genetic mixture.

Non-significant admixture case.

Significant admixture case.

An example of admixture estimation results from Corander and Marttinen (2006).



We manage to maintain a low level of false positives, as most of THESE cases are inferred to be non-significant at 5% level.

How to account for linkage?

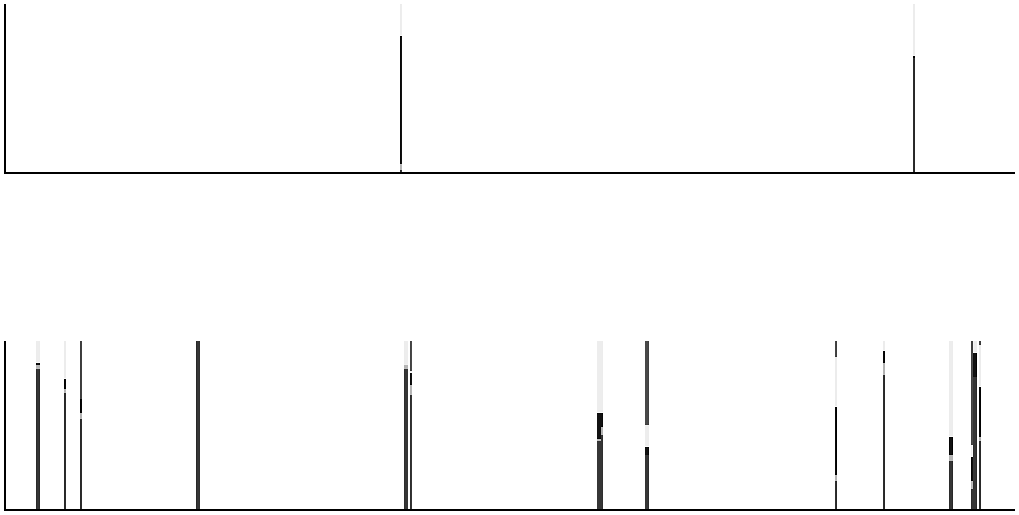
- In BAPS one can utilize two distinct dependence models introduced by Corander and Tang (2007).
- Assume data are available from linked loci residing in *m* chromosomes (marker data) or *m* narrow genomic areas (sequences)
- First and second order Markov dependence structures can be incorporated to the predictive likelihood $p(\mathbf{x}^{(n)}|S)$ conditional on the partition *S* (‘Clustering with linked loci option’).

New possibilities in BAPS 5.1

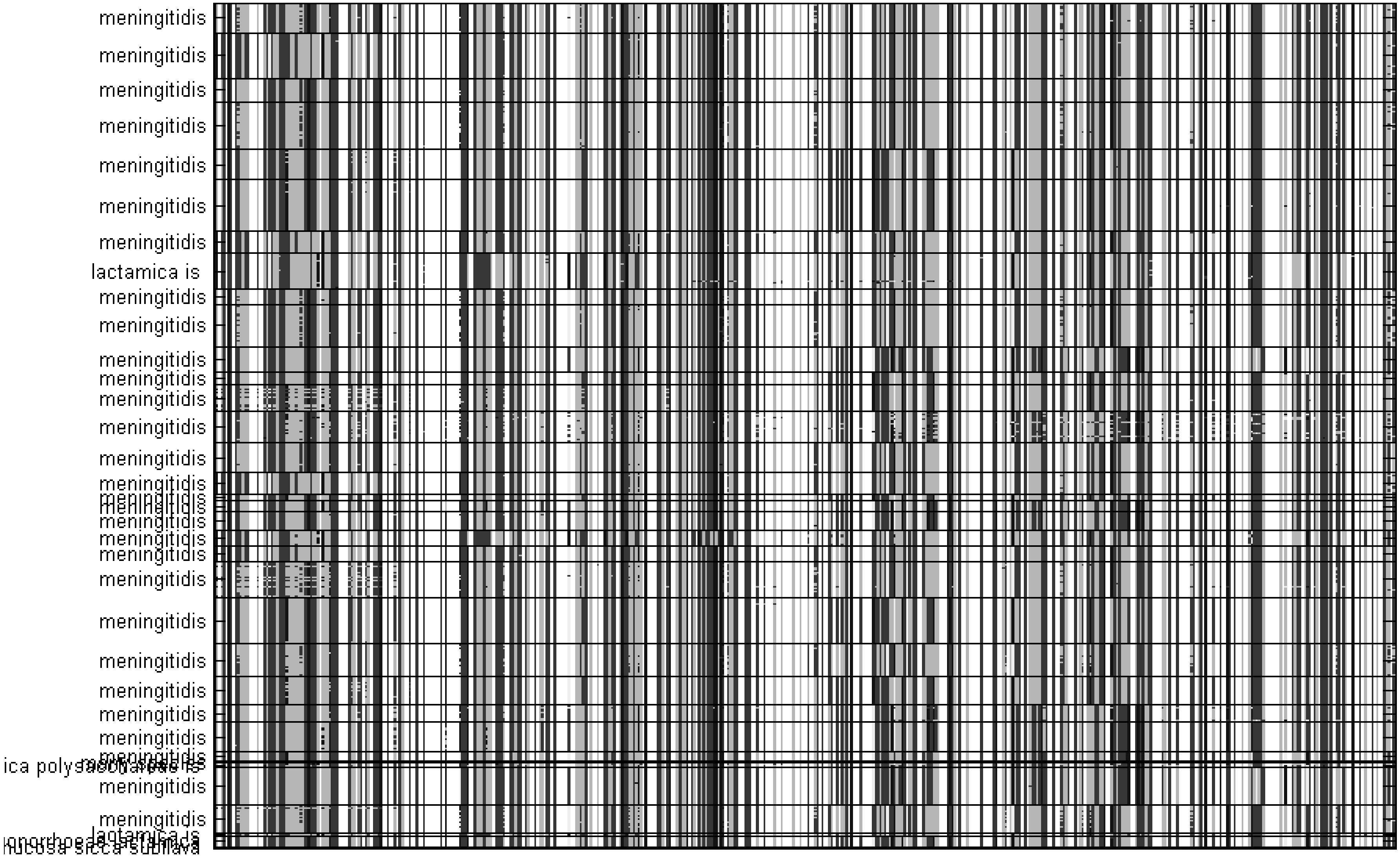
- Analyses can be spread over multiple computers.
- The user has the possibility to use scripts to automate the analyses by calling BAPS 5 from a command line.
- After analyses have been in run separate computers, they can be summarized using a function in the GUI menu system.
- This enables the investigation of much larger data sets.

New possibilities in BAPS 5.1 (ctd)

Discovering alleles with a deviating ancestry using Bayes factors (‘Mutation plot’ function). An example profile for a single admixed (‘green’) individual in the above simulated data.

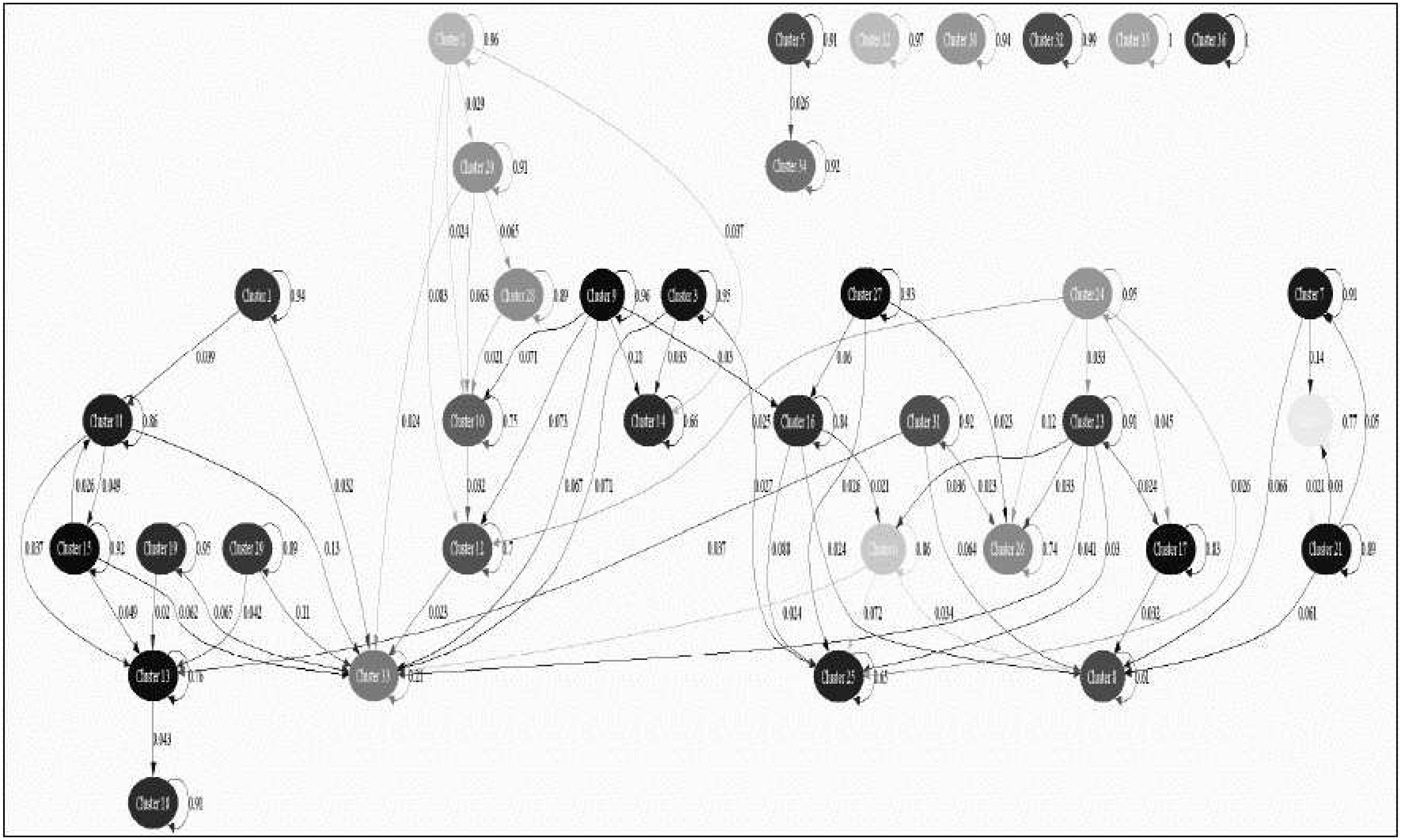


Example of bringing order into a large-scale chaos (5175 bacterial multi-locus DNA sequences). An inferred genetic mixture with 35 clusters. Cluster limits shown as horizontal lines, columns are sites of aligned DNA sequences, colors represent different bases, rows of the image are the individuals.



- However, given that 35 clusters were discovered, it’s not easy to grasp the results in a single screenshot.
- To assist in the exploration of the results, BAPS5 contains new visualization tools describing cluster homogeneity and the genetic relationships between them.
- These tools attempt to describe ’currents in the gene pool’.
- It took appr. 2 days to run this analysis in BAPS 5.1.
- As a comparison, our experiments tell us that it would be extremely time-consuming to run a comparable analysis using ordinary Gibbs sampler or Reversible-Jump MCMC methods.

Example of a gene pool map



Methodological ”BAPS” papers:

- Corander, J. Waldmann, P. and Sillanpää, M.J. (2003). Bayesian analysis of genetic differentiation between populations. Genetics, 163, 367-374.
- Corander, J., Waldmann, P., Marttinen, P. and Sillanpää, M.J. (2004). BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics, 20, 2363-2369.
- Corander, J., Marttinen, P. and Mäntyniemi, S. (2006). Bayesian identification of stock mixtures from molecular marker data. Fishery Bulletin, 104, 550-558.
- Corander J, Marttinen, P. Bayesian identification of admixture events using multi-locus molecular markers. Molecular Ecology, 2006, 15, 2833-2843.
- Corander, J., Gyllenberg, M. and Koski, T. (2006). Random Partition models and Exchangeability for Bayesian Identification of Population Structure. Bulletin of Mathematical Biology, 69, 797-815.
- Corander, J. and Tang, J. (2007). Bayesian analysis of population structure based on linked molecular information. Mathematical Biosciences, 205, 19-31.
- Corander, J., Sirén, J. and Arjas, E. (2006). Bayesian Spatial Modelling of Genetic Population Structure. Computational Statistics, 23, 111-129.
- + some submitted ones...

Thanks to:

- My PhD students Pekka Marttinen, Jukka Sirén and Jing Tang, who have done a great job in the BAPS development process!
- All other collaborators (far too many to be listed precisely).

BAPS software can be found at:

http://www.abo.fi/fak/mnf/mate/jc/software/baps.html