BAPS: Bayesian Analysis of Population Structure

Manual v. 6.0

NOTE: ANY INQUIRIES CONCERNING THE PROGRAM SHOULD BE SENT TO JUKKA CORANDER (first.last'at'helsinki.fi).
http://www.helsinki.fi/bsg/software/BAPS/

Latest update of this manual on 14.2.2013

Jukka Corander, Lu Cheng, Pekka Marttinen, Jukka Sirén and Jing Tang

Department of Mathematics and statistics
University of Helsinki 00014
Finland

Table of contents:

New features in a nutshell

1. Spatial clustering of DNA sequences, output can be directly integrated with Google Maps using http://www.spatialepidemiology.net/.
2. Trained clustering (i.e. semi-supervised classification) of DNA sequence data.
3. Tandem command line program hierBAPS for clustering DNA sequence data in a hierarchical manner and for visualization of the results up to whole genome scale.

Introduction

BAPS 6 (Bayesian Analysis of Population Structure) is a program for Bayesian inference of the genetic structure in a population. BAPS treats both the allele frequencies of the molecular markers (or nucleotide frequencies for DNA sequence data) and the number of genetically diverged groups in population as random variables. However, analyses and model comparisons can also be performed using a fixed number of genetically diverged groups or pre-specified population structures.

With BAPS 6 you can cluster molecular data and perform admixture analyses. Genetic mixture analysis can be done either at:

(1) group level (typically corresponds to clustering of sample populations), or at
(2) individual level.

In fact, in many applications with relatively sparse molecular data it is wise to do analyses at both these levels, when biologically relevant auxiliary information is available to define the groups before the mixture analysis (see, e.g. Corander and Marttinen 2006, for further information). Both types of genetic mixture analyses can be done either by using a:

(1) non-spatial, or
(2) spatial model,

for genetic discontinuities in populations. The spatial model requires that coordinate data is available for the clustered units (groups or individuals), however, these may contain also missing values. In '**Trained clustering**' you can make use of individuals whose origin is known, in order to find the best clustering of individuals with unknown origins. In '**Clustering with linked loci**' a genetic mixture analysis is done either for haploid sequence data, phased diploid/tetraploid sequence data, or for linked marker data for which a *single allele* is recorded per locus. The latter can thus be *e.g.* dominant markers such as AFLPs for a diploid organism, or SNPs for a haploid organism.

Given the results from any of the models for genetic mixture analysis, or any given division of a set of individuals to a number of 'Populations', you can also do inference about admixture events.

The different modules incorporated in BAPS 6 have been introduced in a number of scientific papers, and therefore, the following table provides guidance about which should be cited when publishing results produced with the software. When new methods/publications become available, the BAPS webpage will be updated with the detailed scientific citation information.

Software module:	Scientific citation:
Admixture analysis	Corander J, Marttinen P. Bayesian identification of admixture events using multi-locus molecular markers. Molecular Ecology, 2006, 15, 2833-2843. Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics, 2008, 9:539.
Non-spatial genetic mixture analysis, including 'Trained clustering'	Corander J, Marttinen P, Mäntyniemi S. Bayesian identification of stock mixtures from molecular marker data. Fishery Bulletin, 2006, 104, 550-558. Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics, 2008, 9:539. Cheng L, Connor TR, Aanensen DM, Spratt BG, Corander J (2011) Bayesian semi-supervised classification of bacterial samples using MLST databases. BMC Bioinformatics, 12:302.
Spatial genetic mixture analysis	Corander J, Sirén J, Arjas E. Bayesian Spatial Modelling of Genetic Population Structure. 2008. Computational Statistics 23, 111-129. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Molecular Biology and Evolution, 2013, doi: 10.1093/molbev/mst028
Hierarchical clustering	Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with

	BAPS software. Molecular Biology and Evolution, 2013, doi: 10.1093/molbev/mst028
Genetic mixture analysis with sequences or linked loci	Corander J, Tang J. (2007). Bayesian analysis of population structure based on linked molecular information. Mathematical Biosciences, 205, 19-31. Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics, 2008, 9:539.
Estimates and graphics for gene flow among inferred populations	Tang J, Hanage WP, Fraser C, Corander J. (2009). Identifying currents in the gene pool for bacterial populations using an integrative approach. PLoS Computational Biology, 5(8): e1000455.
Other BAPS papers (versions 1.0& 2.x and the mathematical foundations):	Corander, J., Waldmann, P. and M.J. Sillanpää. 2003. Bayesian analysis of genetic differentiation between populations. Genetics 163: 367-374. Corander J, Waldmann P, Marttinen P, Sillanpää MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics 20: 2363-2369. Corander, J., Gyllenberg, M. and Koski, T. Random Partition models and Exchangeability for Bayesian Identification of Population Structure. Bulletin of Mathematical Biology, 2007, 69, 797-815.

The idea in BAPS GUI is simple, any analysis is performed by clicking the corresponding button, and providing the necessary input to the algorithms by choosing certain files and by feeding values to any fields opened by the GUI. After the analysis, program writes the numerical results to the log-window and to a result file, if such has been specified by the user. If a result file name is not specified by the user, BAPS will automatically write the results in a txt-file carrying a name similar to the name of the used data file. Depending on situation, the program may also provide a visual representation of genetic mixture or admixture results (this is usually automatically produced). The graphics can be saved and opened in the program using the internal format, but they can also be exported to a variety of different formats.

An internal format (binary) result file is produced each time you run a clustering or an admixture analysis and choose to save the results when the program asks you to (this file contains the numerical values needed e.g. for the graphics and subsequent

analyses). Notice that this does not refer to the output file specified in the File menu, where results are stored in text format.

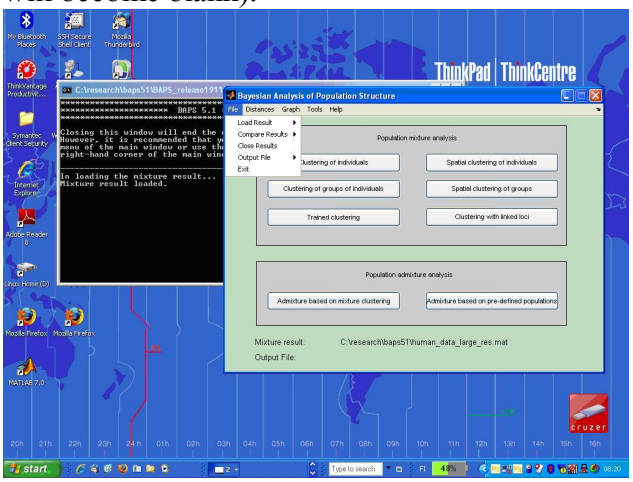
Hint: We recommend that you check the "pre-processed data" option described in the data format section. This is a particularly useful time-saving option for analyses of large data sets. The pre-processed files are compatible between spatial and non-spatial mixture clustering modules, however, if you load a pre-processed data set without coordinates to the spatial clustering module, the program will ask for a coordinate file.

Basic software GUI features

The are five menus in the GUI: File, Distances, Graph, Tools and Help. The documentation part in the Help menu is currently inactive, please refer to this manual instead.

In the File menu (see the image below) you can:

1. Set or remove the text output file where the program writes the numerical results. Notice that when several analyses are performed without changing the output file, the new results are always appended after the old ones. If you don't specify an output file, the text formatted output will be written to file using a default name based on the name of the data file.
2. Load results from earlier analyses to reproduce and/or modify graphics. Notice that these files are in a binary format and that certain functions are not available for files saved with earlier versions of BAPS.
3. Summarize results from parallel analyses of the same data set spread over multiple computers.
4. Close a result file, which means that a previously loaded result information is removed from the program memory (the field under the buttons will become blank).

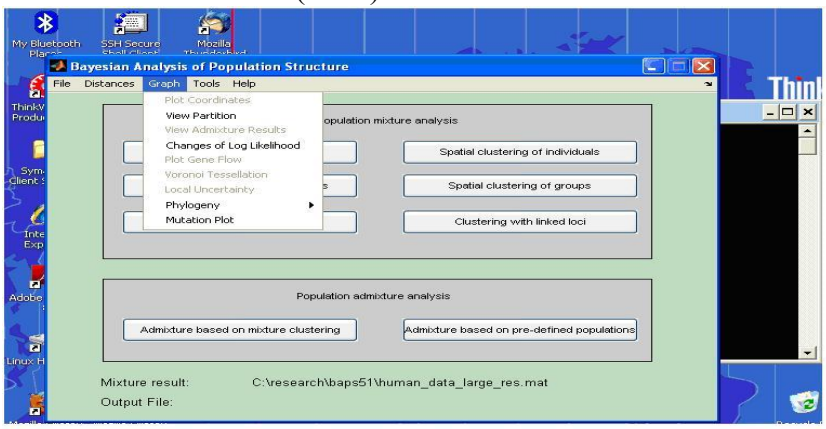


In the Distances menu you can produce a variety of genetic distance matrices between the clusters obtained in a BAPS analysis.

In the Graph menu (see the image below) you can produce a wide variety of graphics, some of which are available only for certain types of analyses. Notice that all graphics

can be saved either in the internal format (.fig) or exported to a variety of file formats using the menu in the graphics window. The graphics options are explained more in detail in the Results section. The following functions are available after loading a result file (binary) to BAPS:

1. Reproduce the partition image showing a clustering solution where the clustered sampling units are shown as colored vertical bars with the color determining the cluster membership.
2. Reproduce the admixture image showing for each individual the proportion of genome estimated to have ancestry in a particular cluster. The proportions are shown as colored segments of a vertical bar where the color determines the origin of a segment.
3. Produce graphics showing the 'genetic shapes' of the clusters with respect to each other, see Tang et al. (2009).
4. Produce a gene flow network for the clusters estimated from admixture results, see Tang et al. (2009).
5. Display a spatial partition from the results of spatial cluster analysis using a Voronoi tessellation.
6. Display 3D-graphics showing how strongly the posterior is peaked locally for the Voronoi tessellation cells (high peaks mean considerable uncertainty about the origin of the particular cell).
7. Draw a phylogenetic tree for the clusters using any of the three available distance measures.
8. Produce a 'Mutation plot' where alleles having support for a different ancestry given a clustering solution can be explored using a threshold for log Bayes factor set by the user. If statistical support exceeding the threshold is found, the corresponding alleles are written to the log-window and the result file, see Corander et al. (2008).



In the Tools menu you can:

1. Specify the clustering to be done using the number of clusters you wish. Notice that the program goes to a 'Fixed K' mode by choosing this alternative, and then any cluster analyses will ask for the fixed number clusters to be used. If

one wishes to do clustering using a range of values of K (#clusters), one simply runs the cluster analysis multiple times each time providing the appropriate input. Notice also, that the program asks for the number of replicate runs to be used in the estimation with any fixed K value. This increases the probability of finding the posterior optimal clustering with that value. The output from the clustering with a fixed number of clusters contains the log(ml) value which can be used for comparison of the clustering solutions. For details on log(ml), see the BAPS papers. NB 'Fixed K' option is not supported for spatial analysis of DNA sequence data. In the 'Not Fixed K' mode, the program will treat the number of clusters unknown and will ask the user to specify an upper limit (or a range of upper limits) for the number of clusters (see the Inputs section below for the details).

2. Compare the posterior probabilities of any number of pre-specified clusterings of the data, e.g. suggested by some biological theory. The clusterings should be provided to the program using an external text file with the following format. Assume that K clusterings are to be compared. Then, the first row of the file should contain K non-zero prior probabilities which sum up to one (a uniform prior would be a typical option). Thereafter, each column of the file will define a partition of the sampling units (individuals or sample groups) you have in the data file. An example of the file for a data set consisting of three sample populations is given below. Here one wishes to cluster the sample populations:

0.333	0.333	0.334
1	1	1
2	1	1
3	1	2

In the first clustering all sample populations are claimed to be genetically distinct, in the second clustering they are claimed all to represent a single population, and in the third clustering the last sample population is claimed to be distinct from the first two, which are homogeneous. Similarly, if the clustering is to be done on the level of individuals for the same data set, and there are three individuals from each sample population, an example of a partition file is the following:

0.333	0.333	0.334
1	1	1
2	1	1
3	1	2
4	1	2
5	1	1
6	1	1
7	2	3
8	2	3
9	2	3

3. Load figures produced and saved earlier using the .fig file format. Notice again that all graphics can also be exported to a range of formats using the menu in the graphics window.

Inputs for BAPS

Here you find information about what the program expects as input when you click any of the analysis buttons, when the program is not on the 'Fixed K' mode. For that mode, see also the information provided above for the Tools menu. For each analysis module there are examples files available on the BAPS website for all possible supported data formats. Examples include haploid, diploid and tetraploid cases.

Inputting the maximum number of populations, K

In all clustering modules of the 'Population mixture analysis' you need to tell the program your opinion about the maximum number of genetically diverged groups, say K. In 'Trained clustering' K must naturally be greater or equal to the number of reference clusters. The prior distribution for clusterings is then either: (1) uniform in the space of clusterings having at most K clusters for the non-spatial model, or (2) non-uniform in the space of clusterings having at most K clusters for the spatial model, depending on the spatial pattern of the observed data. Generally, the larger the K, the longer the execution of the estimation algorithm will take. We recommend that you experiment with different values of K and run the analysis several times (it typically goes fairly fast), to see whether the results change noticeably (which you can always see by comparing their corresponding "logmls"). As an example, you might have a sample of 100 individuals genotyped at 10 microsatellite loci, and on the basis of your biological reasoning, you expect to find 3 divergent clusters. However, you might also have outlier individuals in the data, not representing any of these 3 groups, so a careful strategy would e.g. be to run the estimation with K = 5, K = 10, K =15. If K is set to be extremely large, e.g. close to the number of observed individuals, then the algorithm may easier get stuck to a local mode. Clearly, if K is too small, the "true" structure cannot be detected. If you get the estimated number of clusters equal to K, then you should try with a higher K, because otherwise the results can be misleading.

In practice the easiest way to run the program several times with different K is to provide a vector of values when the program askes you for the maximum number of clusters. For example if you want to run the estimation with K=5, K=10, and K=15, the correct input would be: '5 10 15' (without ':s).

Hint! You can fill as many K values into the input window as you wish, e.g. like two hundred. The input field simply shifts to the right when it becomes filled. So the input could look like:
10 10 10 10 10 10 11 11 12 12 12 12 12 12 12 13 13 13 13 13 13 13 13 13 13 13 etc.
What the program does for each K value (even the replicates of the same value) is to find the optimal partitions with $k \leq K$, it stores these internally, and the after all K values have been processed, it merges the stored results according to the logml values. Even if small k values are considered a priori possible, they can have extremely bad fit compared to the larger values, in which case they are ignored in the results.

Starting from different values K is important also in the sense that K affects the initial assignment of the simulation and thus the possibility of finding only a local mode is reduced when simulation is started many times from different initial assignments. It makes sense even to start the simulation many times with the same K, since the optimization algorithm is stochastic and can therefore end up in different solutions in separate runs. If a vector of K values is provided, BAPS will write its output on the basis of the best solution that was found. Some caution with K is necessary, e.g. use of K=200 for a data set of 250 individuals would not in general be a wise strategy.

See also 'pre-processed data' option, which is very useful if you wish to run the program many times and you have a large data set whose pre-processing takes a long time.

After mixture clustering analysis has finished you should save the result file in order to use it later for reproducing graphics and for performing admixture analysis. If you used group level mixture clustering, BAPS will need to know how many rows from one individual are presented in the data. This will be asked from you before the saving is done. It is on user's responsibility to make sure that the original data really contained the given number of rows per individual.

Clustering of individuals

There are three alternative ways of providing input for BAPS when you wish to do the clustering of individuals with unlinked marker data.

BAPS format:

The BAPS formatted data files should be plain ASCII text files. The input file contains a data matrix where columns are either separated by blanks or tabs. The columns of the matrix correspond to loci at which the individuals were observed. The rows of the matrix correspond to the individuals. There is an additional column in the right end of the matrix that contains on each row the index of the individual whose alleles are presented on the row. There can be more than one row per individual. For example if individuals are diploid, there should be two rows per each individual corresponding to two alleles that can be observed at each locus.

Alleles can be indexed with any non-negative integer value, thus, for example with microsatellites you can use either directly the repeat lengths or some alternative coding of the alleles. The indices of individuals, however, should start with 1 for the first individual and end with the value that corresponds to the total number of individuals. Missing allele at some locus is denoted by any negative integer, e.g. –999 or -9 (these are widely used).

If the sampling populations of the individuals are known, you can input them by giving two additional files: one containing the names of the populations, the other containing the indices of the first individuals of each sampling population. This information will make the numerical and graphics output easier to investigate.

The following example files with 10 microsatellite loci and 5 individuals are included in the package ExamplesDataFormatting.zip available on the BAPS website (http://www.helsinki.fi/bsg/software/BAPS/). The last two files provide examples about the sampling population information, which can be used with the BAPS data format. In the example there are 3 sample populations, such that individuals 1-2 belong to Example population 1, individuals 3-4 to Example population 2, and individual 5 to Example population 3.

Example data in BAPS format for clustering of haploid individuals.txt
Example data in BAPS format for clustering of diploid individuals.txt
Example data in BAPS format for clustering of tetraploid individuals.txt
Example of sample population names file to be used with BAPS formatted data.txt
Example of sample population index file to be used with BAPS formatted data.txt

GENEPOP format:
See http://wbiomed.curtin.edu.au/genepop/help_input.html for general information about GENEPOP format. The data must obey strictly the rules mentioned in the referred page concerning punctuation and empty spaces. Otherwise BAPS may be behave unexpectedly. You can use both 2- and 3-digit allele codes, BAPS will investigate the format of the data and act accordingly. However, all allele codes in one data file must have the same number of digits. Note that when you use data that is in GENEPOP format you always have to provide two alleles for each individual at each locus. If you wish to cluster haploid individuals you must mark the other allele of each individual at each locus as missing (00 or 000). BAPS uses the labels of the first individuals of the populations as names for the populations.
NB! Do not begin the sample population labels with the word 'pop', because it will then be erroneously interpreted as the sample population separator.

Pre-processed data:
Before the model fitting can start, all data must be pre-processed by BAPS. For large data sets the pre-processing may take quite a long time, e.g. more than half an hour. If you wish to analyze such data many times starting from different initial assignments, it saves some time to save the data after pre-processing it once. Next time, instead of starting from the original data file, start with the pre-processed file that you previously saved.
NB! Data pre-processed within some mixture analysis module (any of the 6 buttons) must be used ONLY within the same module! For example, if a data file pre-processed within 'Clustering of individuals' is used for 'Clustering of groups of individuals', BAPS may produce an error message, or the analysis may produce erroneous results, even if no errors are displayed in the log-window.

Clustering of groups of individuals

There are three alternative ways of providing input for BAPS when you wish to do the clustering of groups of individuals.

BAPS format:
The data file is very similar to the data file used in clustering of individuals, the only difference being that instead of specifying the individual, the last column contains the index of the group that is the origin of the alleles on the particular row.

The following example files with 10 microsatellite loci and 5 individuals are included in the package ExamplesDataFormatting.zip available on the BAPS website. The contents of these files are otherwise equal to those provided for 'Clustering of individuals', except that the last column now indicates from which sample population a particular row of data are taken. The last file contains the sample group names. Thus, in the example there are again 3 sample populations, such that individuals 1-2 belong to Example population 1, individuals 3-4 to Example population 2, and individual 5 to Example population 3. Notice that no index file is needed for this type of analysis.

Example data in BAPS format for group-wise clustering of haploid individuals.txt
Example data in BAPS format for group-wise clustering of diploid individuals.txt
Example data in BAPS format for group-wise clustering of tetraploid individuals.txt
Example of sample population names file to be used with BAPS formatted data.txt

GENEPOP format:

See GENEPOP format above in the clustering of individuals. The populations in the data define the groups to be clustered.

Pre-processed data:

Instead of starting from the original group level clustering data file, you can start with the pre-processed file that you have saved after once pre-processing the original data. NB! You should not use here files pre-processed under 'Clustering of individuals'.

Trained clustering

In order to do the trained clustering of individuals you must provide two data files: one containing the reference individuals or samples whose origins are known/labeled, the other containing the sampling units (individuals or groups of individuals) that you wish to cluster. The trained clustering option is available for both marker and DNA sequence data types.

For DNA sequence data both files must be in the concatenated Excel (xls) format used also in the module 'Clustering with linked loci'. The following example files (included in ExamplesDataFormatting.zip available on the BAPS website) contain training sequence data for 7 genes from 11 baseline 'populations', and test data with 11 samples of unknown origin (the latter file):

Example training data in XLS format for Trained clustering.xls
Example test data in XLS format for Trained clustering.xls

In this example every baseline population is represented by a single sample in the test data file. The column with the header 'ST' in the two files contains the ID of each sample and must be a positive integer such that no two rows have the same value. The column with the header 'Cluster' in the training data file indicates the baseline population for each sample and the values must be integers running from 1 to the number of populations. For further information about trained/supervised clustering of DNA sequence data, see Cheng et al. (2011).

For marker data both files must be in GENEPOP format (see GENEPOP format in Clustering of Individuals above). Individuals in one population (separated by a word 'pop') in reference data file correspond to individuals from a single origin. In the other file the word pop separates the sampling units. Thus, if you wish to cluster unknown individuals one by one, you should write the word 'pop' above every line that specifies an individual in the sampling unit data file.

In both data files all individuals should be given names. These names will be needed by the program when the output is written.

The following example files (included in ExamplesDataFormatting.zip available on the BAPS website) contain baseline microsatellite data for 10 loci from 5 baseline populations, and sample data with 10 individuals of unknown origin (the latter file):

Example baseline data in GENEPOP format for Trained clustering.txt
Example sample data in GENEPOP format for Trained clustering.txt

If there is some auxiliary information available, which allows a pre-grouping of the sample data in trained clustering to occur before the mixture analysis, this can be used in BAPS by formatting the GENEPOP sample data file such that the pre-groups are separated by the word 'pop'. This means that BAPS forces always all individuals within a single pre-group to be assigned to the same population (either a baseline or a novel population, depending on the marker data and the values of K used as an input to the analysis). Use of this strategy is discussed in Corander et al. (2006). An example of a sample data pre-grouped into four groups is contained in the following file, where the molecular information is otherwise the same as in the above example sample data file:

Example pre-grouped sample data in GENEPOP format for Trained clustering.txt

The rationale behind the use of pre-grouping (or 'sampling units') is that, if there is some biologically relevant information available that tells us that some individuals must have the same, yet unknown origin, then by clustering them together we are able to increase the statistical power to detect the correct origin. The availability of such information is very species dependent, and its reliability must be determined by the user from case to case.

Spatial clustering

The input to the spatial clustering modules is otherwise exactly the same as in the above cases of 'Clustering of individuals' and 'Clustering of groups of individuals', except for the coordinate values that need to be given in a separate file. The coordinate file should be plain ASCII with as many rows as there are individuals ('Spatial clustering of individuals') or groups ('Spatial clustering of groups') in the molecular data set. If the coordinates are missing for an individual or a group, this should be indicated by a corresponding line in the coordinate file containing two consecutive *zeros*. Columns in the coordinate file should always be tab separated. Notice that negative coordinate values are also acceptable, but zeros are reserved for the cases with missing coordinates. The following example file (included in ExamplesDataFormatting.zip available on the BAPS website) contains coordinates for 10 cases, such that the coordinate values for the last case are unknown:

Example coordinates for the spatial clustering.txt

Note that the numbers in the above example file are in the scientific e-notation, but BAPS accepts ordinary decimal numbers as well. As explained in Corander et al. (2008a), the rationale of using spatial information is to assign a biologically relevant non-uniform prior distribution over the space of clustering solutions, which expects that underlying clusters are spatially smooth at least to some extent. This increases the power to correctly detect the underlying population structure and can be used to investigate the population structure also visually. When the molecular data are very extensive, the spatial and non-spatial clustering models are expected to yield highly similar results.

In BAPS 6 the spatial clustering model has been implemented also for DNA sequence data, see Cheng et al. (2012). The input sequences can be in FASTA format, see example below.

Example data in FASTA format for clustering of haploid individuals.fasta
Example coordinates for the spatial clustering of DNA sequences.txt

BAPS outputs a txt file which can used to map the clusters on Google Maps, using http://www.spatialepidemiology.net/, choose Create User Maps and insert the spatial information.

Clustering of linked molecular data

Here the rationale of the analysis is the same as for the genetic mixture analysis using unlinked markers, except that the used Bayesian model accounts for dependences present between the "loci" (either marker loci or sites within aligned sequences). The outputs from this software module can be further used in the admixture module. BAPS automatically recognizes from the result file whether the linkage model or the independent loci –model was used in the genetic mixture analysis, and chooses the appropriate admixture model when the result file is loaded in an admixture analysis.

All example files mentioned in this section are included in the package ExamplesDataFormatting.zip available on the BAPS website. Four distinct options are available for linked data input: MLST-format (either as separate fasta files or as a single Excel file), BAPS-numeric format, BAPS-sequence format, and pre-processed data. The MLST option is intended for sequence data for haploid organisms only, such that the number of genes used for clustering can be conveniently specified by the user. The BAPS format can be used for both analyzing sequence and linked molecular marker data. All used sequence data must be multiple aligned and have equal length for all individuals. When data represent multiple genes, the length of any individual gene sequence can be arbitrary.

The simplest option for using haploid DNA sequence data is to store the sequences in an Excel file. Sequence gaps and missing nucleotides should be denoted by a dash (-). The first column should have the header ST on the first row, whereafter the individuals are labeled by linearly increasing integers on the consecutive rows (from 1 to n with n individuals in the data set). Each gene will be represented by one column in the Excel sheet, such that the first row contains the gene labels. Thus, the cell on row i at column j+1 should contain the sequence data for individual i for the jth gene.

The following example file contains data for 6 individuals over 3 genes.

Example MLST DNA sequence data in concatenated Excel format.xls

The 2nd option for reading in MLST type sequence data, is to use a similar formatting as applied in the MLST databases, together with separate fasta-formatted files for each gene. With this format it is convenient to run several different analyses with different subsets of genes included. Two types of files are needed: 1) profile file similar to those obtained by MLST database queries, 2) fasta-formatted sequence files for each gene.

The following example file contains the profile for 4 samples, which are sequenced for a total of 6 genes:

Example of an MLST profile file for 6 genes.txt

NB! The profile file must be tab delimited, with an equal number of tabs between columns.

An example file for one of the genes in the above profile file (recA) is the following:

Example of a fasta formatted sequence file for gene recA.txt

NB! The sequence identifier, e.g. >RecA-1 must match with the corresponding gene name in the profile and with the label of the individual (these two are separated by a dash).

After loading the profile file to BAPS, program asks which species are to be included in the analysis. By clicking on the Select all –option, all rows of the data set are included. After this, a user has the possibility to choose a range of isolates (all or a subset of them). When the selection of isolates is completed (all isolates to be included are in the right hand side window), click OK. Then, a window for selecting the genes for the analysis appears. For each chosen gene, BAPS requires the user to input a corresponding Fasta file containing the aligned sequences for all included isolates. Missing nucleotides in the sequence are by default denoted by the question mark ("?") and sequence gaps denoted by the dash symbol ("-"). However, both incomplete information types will be decoded as the same after the data is loaded and are thus not distinguishable. For convenience it is possible to use either one of the two symbols to denote any unknown bases.

After all the data have been provided for BAPS, it starts preprocessing the files. We recommend that you save the pre-processed data by answering Yes to the question, since this saves a lot of time in repeated analyses of the same data set. BAPS will ask the user to specify the linkage model, and for sequence type data it will be mostly relevant to use the Codon linkage model. After the linkage model is specified, BAPS asks whether to save the fully pre-processed data (again we recommend the user to do this). This question is raised because the chosen linkage model will affect the internal data formatting depending on the eventual presence of missing values. When running repeated analyses with the same data, the user can load the fully pre-processed data set by using the 'Pre-processed' option when clicking on the button 'Clustering with linked loci'. The final option is to choose the prior upper bound K by typically inserting a range of values into the window. For details on this see the 'Inputting the maximum number of populations, K' section in the beginning of this manual.

As the alternative option to MLST formatted data, it is possible to read in linked data in BAPS format. When the BAPS data format is used, the sequence data should be formatted either: (1) as haploid marker data for the other clustering modules (see the previous sections of this manual), which corresponds to a single data row per individual, or (2) as diploid marker data which is *phased within each considered gene*, which corresponds to two rows of data per individual, or (3) as tetraploid marker data which is *phased within each considered gene*, which corresponds to four rows of data per individual.

NB! Notice that in contrast to the MLST format you need under the BAPS format to concatenate the sequences from all considered genes into a single one, and tell the program about the gene boundaries in a separate file (see below).

You can either use a direct sequence (character) based format, or a numeric data input format. The numeric format is obtained by replacing each of A,C,G,T with a unique integer, and by replacing the eventual dashes with a negative integer such as '-9'. Thus, this formatting will be exactly the same as the one used for BAPS formatted

unlinked marker data (see the previous sections). The following example files show how the BAPS sequence (character) formatted data files should look like:

Example of a BAPS formatted haploid sequence file for clustering with linked loci.txt
Example of a BAPS formatted diploid sequence file for clustering with linked loci.txt
Example of a BAPS formatted tetraploid sequence file for clustering with linked loci.txt

NB! Notice that there must be a space between the last element of the concatenated sequence and the individual index.

In the BAPS format it is also necessary to tell the program about the gene boundaries. This is done by providing a separate file where the number of rows equals the number of genes. At each row, the integers refer to those columns of the data matrix that correspond to the sites of the gene in the concatenated sequence. Additional zeros are used to fill the rows to have an equal number of columns. An example of a file specifying the gene boundaries for concatenated sequence of a total length 750 bases, from three genes (200 sites in the 1st gene, 250 sites in the 2nd gene, and 300 sites in the 3rd gene) is the following:

Example of a file specifying gene boundaries for a concatenated sequence in BAPS format.txt

Finally, linked (and phased) marker data should be formatted as the sequence data in the numeric BAPS format, i.e. analogously to the unlinked markers. The loci representing the same linkage group should be ordered linearly in accordance with the previous example concerning representation of sequences over a number of genes. Thus, for marker data each "gene" in the previous example should be replaced by a linkage group, and the other aspects of the formatting are kept equal. The "linkage map" file should be formatted exactly as the above example file for gene boundaries, such that a sequence site is replaced by a locus. The appropriate modeling option for linked marker data is typically the 'linear' linkage model (for sequence data 'codon' option is recommended).

NB! The BAPS data format offers also the option to do clustering of groups of individuals in the module 'Clustering with linked loci'. This can be done by replacing the individual integer labels in the last column of the data matrix by integer labels of the corresponding populations.

Admixture of individuals based on mixture clustering

The input file for admixture analysis under this option is the binary result file of mixture clustering (saved by the user at the end of any mixture analysis module).
NB! Notice that this is not the same thing as the ASCII output file, where the result summaries are written in a format readable by text editors (Word, WordPad etc)!

All mixture clustering modules (non-spatial, trained, spatial, linkage) produce files that are compatible with the admixture module. If you have done multiple cluster analyses with the same data set, e.g. using different numbers of clusters in the 'Fixed K' mode, you can run the admixture analysis separately for each of them.

Before the admixture analysis can start you will be requested to input the minimum size of a population that will be taken into account when admixture is estimated.

BAPS will then remove individuals who belong to a cluster whose size is less than the given number. The removed outlier individuals are displayed on the screen.

You will also be asked the following things: 1) the number of iterations that are used to estimate the admixture coefficients for the individuals, 2) the number of reference individuals from each population, 3) the number of iterations that are used to estimate the admixture coefficients for the reference individuals. These three things affect the accuracy of the estimation. The first input determines the number of times the individuals in the data are analyzed using different simulated allele frequencies. The higher this input is, the better the uncertainty in the allele frequencies is taken into account. A good value would be for instance 100. For really extensive data sets lower values can be used according to time available for the analysis. The second and the third input are needed in simulation and estimation of reference individuals. These individuals are used to estimate the level of spurious admixture that can be assigned to the molecular variation in the population estimated in genetic mixture analysis (see Corander and Marttinen 2006). Good value for the number of reference individuals from one population would be for instance 200. Because the accuracy of admixture coefficients for the reference individuals is not of great importance, the number of iterations used to analyze these individuals can be much lower than what was used in the analysis of the individuals in the data. Reasonable values would be for instance somewhere between 5 and 20. It is usually a good idea to first test with small inputs to see how long the analysis takes and then do a re-analysis using higher values.

Admixture based on pre-defined clustering

If the user has a good idea of how the individuals should be clustered, e.g. with data from pure and hybrid species, the admixture analysis can be done on the basis of a partition of individuals given by the user. In this module there are again two alternatives for providing the input for the program: BAPS and GENEPOP formats. If you use BAPS format, the input files are exactly similar to those used in the clustering of individuals. However, you also have to provide an additional file that contains the partition of the individuals. The partition file contains as many rows as there are individuals in the data. On each row there is an index identifying the cluster to which the individual belongs. The indices of the clusters must range from 1 to the total number of clusters. The file below contains an example partition for a data set with 10 individuals (they can be of any ploidy level) which are *a priori* assigned into 3 populations.

Example of a partition file for admixture analysis based on user-specified populations.txt

If the GENEPOP format is used, the input file is again similar to that used in the clustering of individuals. However, here the populations of individuals in the data (as separated by the word ‘pop’) are used to define the partition of individuals on which the admixture analysis will be based, instead of defining the sampling populations, as in ‘clustering of individuals’.

NB! It is also possible to estimate the admixture of individuals with respect to different origins even if the placing of those particular individuals in different clusters is not known, and thus they do not contribute to allele frequencies of any particular population. As an example, suppose that you have two groups of individuals from two different origins and a third group of individuals who are suspected to be admixed between known two origins (e.g. hybrids). Now you wish to know the admixture proportions of the individuals in the third group. To do this kind of analysis in BAPS

format you only have to mark the cluster (in partition file) of those individuals who are not pre-assigned to any cluster as -1. In GENEPOP format the analysis is as easy. You just add an extra population in the end of the data file that specifies the individuals that have not been pre-assigned to clusters. Before the analysis starts the program will ask you if the last population in the data file will be used to define one more cluster with respect to which admixture proportions will be estimated, or if the last population consists of individuals who do not contribute to allele frequencies of any cluster.

NB! In order to this kind of admixture analysis being reasonable you should make sure that populations specified by the user really are genetically distinct, at least to some extent.

As in the admixture analysis based on mixture clustering, you will also now be asked for the minimum size of a population that will be taken into account when admixture is estimated. In BAPS you will also have to input the values that determine the numbers of iterations and the number of reference individuals. (see: Admixture based on mixture clustering.)

Hierarchical clustering and visualization of DNA sequence data using the command line tandem program hierBAPS

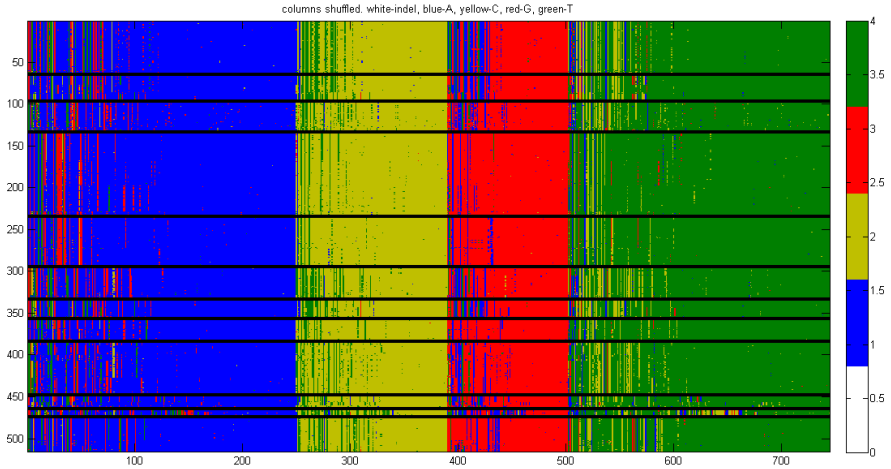
As illustrated in Willems et al. (2012), BAPS clustering of DNA sequences in a hierarchical manner can provide an increased resolution in the estimate of genetic population structure. A substructure may remain undetected when all sequences are simultaneously analyzed e.g. because of the presence of a strongly divergent lineage introducing many SNPs that are monomorphic outside the lineage. To detect substructure within BAPS clusters, one can use each cluster data separately as an input to a subsequent analysis, which results in a hierarchical or nested set of clusters. The larger and more heterogeneous the original input data are, the more the subsequent stages of clustering after the first 'global' clustering are expected to improve the resolution on detecting the underlying lineages. To automate such an analysis, Cheng et al. (2013) introduced a command line program hierBAPS which can be used in tandem with BAPS. Links to hierBAPS and its bundled tools are found on the BAPS website.

hierBAPS can be used to analyze multiple sequence alignments using the BAPS clustering model in an iterative manner, such that the user specifies the number of layers in the hierarchy and the program estimates the maximum a posteriori partition of the sequences at each layer. hierBAPS is bundled with a simple tool (exData) for converting a fasta-formatted alignment file or a concatenated Excel-file into a data matrix in an internal format which is used in the hierBAPS analysis. Additionally, another tool (drawSnpMat) is introduced to visualize the sequence variation within and between the inferred clusters in an efficient manner.

The two example files below represent input to hierBAPS in the two different formats.

Example of hierBAPS input in FASTA format.fa
Example of hierBAPS input in XLS format.xls

The figure below produced by drawSnpMat shows the hierarchical clustering result for 515 sequences with two successive layers of clusters. drawSnpMat permutes by default the original SNP site order to maximize visual separation of the clusters, but an unpermuted version of the figure can also be produced. Horizontal black lines are drawn to separate the clusters obtained in the first level of clustering. Cluster substructure is well visible e.g. in the second top most cluster of the first level, where the sequence variation is reduced within blocks of alignment rows.

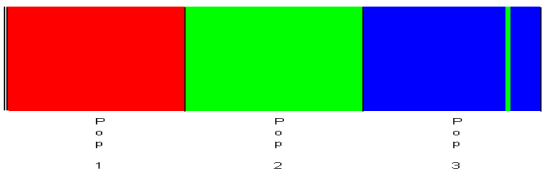


About results

Some graphics are automatically produced by BAPS after an analysis is finished. However, to access all graphics options, the user can load a result file in the menu File-Load result, after which various options will become available in the menu Graph, depending on the type of the loaded file.

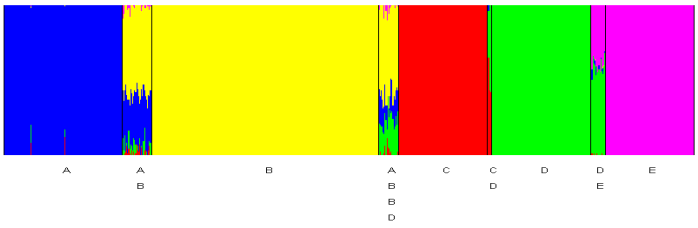
Mixture partition

The mixture clustering graphical output is a colored partition of the clustered units (see the example image below), which is automatically produced when there are at most 30 clusters. This output is also available from Figures menu as the View Partition –option. Each cluster is assigned a unique color in the graphic, but the color ordering is arbitrary, so you cannot compare colors between analyses. Each “sampling unit” (an individual or a group) that was clustered, is represented by a vertical bar having the color corresponding to the cluster where it was placed. Width of the bars depends on how many of them need to be drawn. If the names of the sampled populations have been provided to the program (see the input format section), these are printed below the colored bars to denote the sample origins. The names appear in the same order as in the data, and are printed in the middle of the set of bars representing the particular sample population. In 'Trained clustering' result picture the individuals are in such order that first (from the left) come the reference individuals with known origins and after them come all other individuals.



Admixture partition

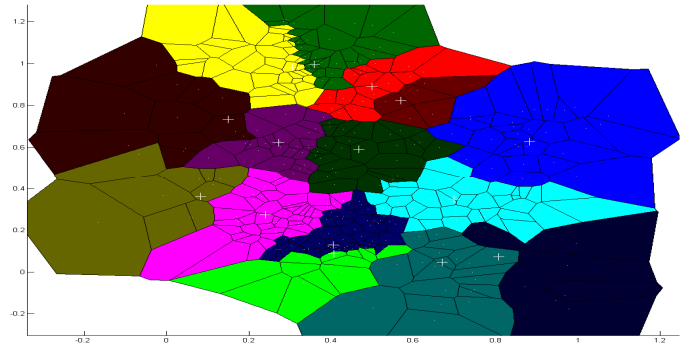
The admixture clustering graphical output is in most analyses also automatically produced when there are at most 30 clusters (admixture using linked data is an exception). Each cluster is assigned a unique color in the graphic, but the color ordering is arbitrary between analyses. Notice, that here every vertical colored bar necessarily corresponds to an individual, in the same order as in the original data provided by the user. The vertical bars are split into several colors when there is evidence for the admixture, such that each color corresponds to an ancestral source (a cluster), and the proportion of a particular color in the vertical bar corresponds to the proportion of the genome estimated to be represented by that source. The image below provides an example admixture partition. A graphical tool for optimizing the admixture graphs for visual clarity is included in the software.

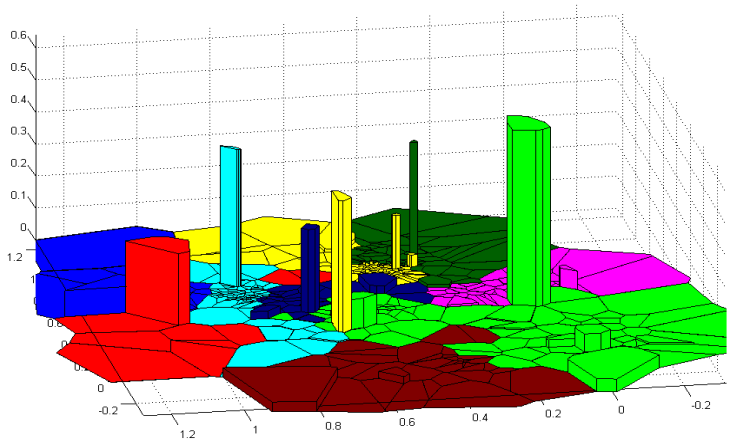


NB! The admixture graphics displayed by default after the estimation show the posterior estimates for all cases, irrespective of the simulated p-values. Admixture graphics contain also the possibility of showing only significant admixture estimates using a user-specified threshold for the *p*-values (Use first File-Load result-Admixture result to load the result file to BAPS and then Graph – View admixture result and set the p-value threshold according to your preferences).

Voronoi tessellation and Local uncertainty

Voronoi tessellations are produced by the spatial clustering module when there are at most 30 clusters in the data (also accessible through the Graph menu with or without data labels). A cell of the tessellation corresponds to the physical neighborhood of an observed data point, and is colored according to the cluster membership. A graphical 3D-representation of the local posterior uncertainty in the estimated tessellation is available from the Graph menu as 'Local uncertainty'. These graphics are considered in detail in Corander et al. (2008a). Examples are provided by the images below.



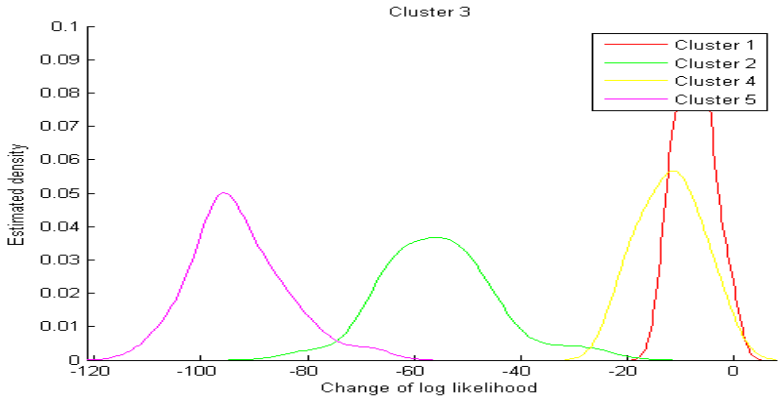


‘Genetic shapes’ of populations

The function ‘Changes of log likehood’ in the Graph menu can be used to do a model-based investigation of the ‘genetic shapes’ of the estimation populations (clusters). In this tool it is possible to choose a ‘source’ cluster (a single cluster or several ones) and then specify a set of ‘target’ clusters, such that BAPS calculates the changes in the log marginal likelihood of the used mixture clustering model occurring when an individual is re-allocated from the source cluster to a target cluster. These values are calculated for all individuals of the chosen clusters, and they are related to the genetic affinities among the clusters. Also, the genetic composition of a cluster (i.e. the estimated underlying population) will affect the shape of the distribution of the values. To provide visually interpretable clues for investigating how the populations relate to each other according to the mixture model, and how heterogeneous a population is in this respect, BAPS estimates non-parametric density curves from the obtained set of log marginal likelihood changes.

An example image is shown below, where there are 5 estimated populations (clusters) and cluster 3 is chosen as the ‘source’. Negative changes in the log marginal likelihood close to zero indicate that the mixture model judges both assignments (source cluster and target cluster) to be statistically reasonable for an individual. In contrast, values further away from zero indicate that the relative genetic affinity of the clusters has decreased. In the example, it is seen, for instance, that the cluster 3 is genetically much closer to clusters 1 & 4, than to 2 & 5. Also, the higher level of peakedness in the red curve shows that the genetic affinities of the individuals in cluster 3 towards those in cluster 1 are highly similar, because the values of the log marginal likelihood changes are concentrated over a rather short interval. In contrast, affinities towards cluster 2 (green curve) are more widely distributed. The curves can

thus be used to investigate the 'genetic shapes' of the estimated populations with respect to each other. For instance, a bi-modal curve reveals that the estimated population consists of two parts with distinct genetic affinity towards another population. For more examples, see Tang et al. (2009).



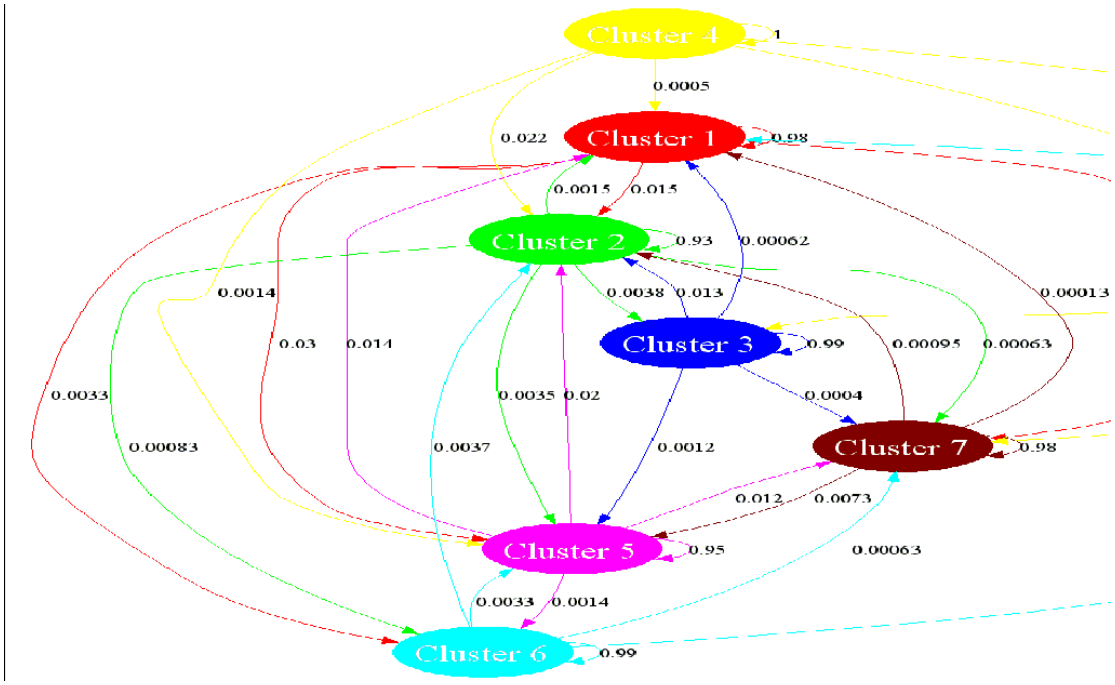
Plot gene flow

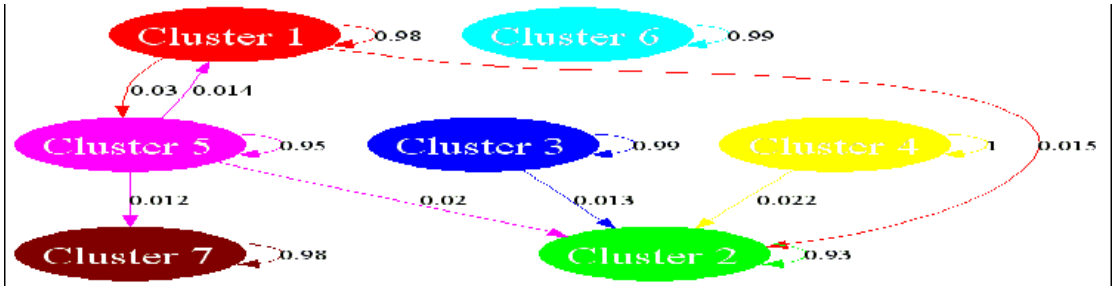
The function Plot Gene Flow in the Graph menu estimates and draws a network of clusters where gene flow is indicated by weighted arrows, such that the weights equal relative average amounts of ancestry in the source cluster among the individuals assigned to the target cluster. This function becomes accessible, when a result file from an admixture analysis is loaded through the File menu.

To use this function, it is necessary to install the GraphViz package which is called by BAPS to draw the graphics. GraphViz can be downloaded from this site: www.graphviz.org. When choosing the Plot Gene Flow –function, BAPS will first ask the user to specify a threshold for the significance of p-values of the admixture estimates. The default option is 0.05, which is likely to be satisfactory for a majority of analyses. After this BAPS opens a dialog where the folder containing the GraphViz program (dot.exe) is given. On a typical Windows installation this path is \Program Files\ATT\GraphViz\bin. BAPS produces now a graphics window, where the estimated gene flow network is drawn.

NB! Due to inappropriate image scaling, the window may be only partially visible for certain result files. To see the entire network, drag the bottom of the window and resize it until the upper blue pane is visible. The window can be moved on the screen by pulling it with the mouse on the blue pane. If the network is still only partially visible, or looks too messy, it can be cleaned up by pruning edges (arrows). This is done via the Attributes – Prune edges menu of the graphics window. By choosing the menu, a dialog box opens where the minimal edge weight to be shown can be set. The default value is the smallest weight estimated from the data larger than zero. For instance, a value 0.01 is likely to work well for many cases, and it states that the average relative gene flow between two populations (clusters) must be at least 0.01 before an edge between them is included in the network. After inserting the threshold value, the network will be redrawn. In some cases it will be necessary to resize and re-position the graphics window again, before the whole image is clearly visible. If the level of pruning appears insufficient, the procedure can be repeated with a different threshold value. The image can be exported to a range of different graphics formats by clicking on the disk icon.

Two example networks are shown below, one with the default minimum threshold on the edge inclusion, and the other shows the same network with pruning based on the threshold value 0.01. For more examples, see Tang et al. (2009).





Mutation plots:

The function 'Mutation plot' in the Graph menu offers the possibility to do for any individual a screening for alleles showing conclusive evidence for a deviating ancestry. This function becomes accessible when a Mixture clustering result file is loaded via File – Load result. The function opens first a dialog box, where the index of the individual and a 'significance' threshold should be given. The default threshold is 2.3 (on a log scale), which correspond to a Bayes factor of at least 10 in favor of an ancestry deviating from the population (i.e. cluster) to which the individual was assigned in the genetic mixture analysis. These graphics are considered in detail in Corander et al. (2008b).

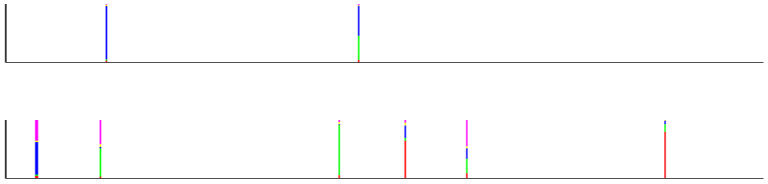
An example image is shown below, where two panes are shown if the data are diploid, such that the lower pane corresponds to the first allele, and the upper pane to the second allele at each locus. The loci are shown in the order they are in the original data file. A colored bar is shown for the loci with evidence larger than the threshold for the hypothesis that the allele in question has its ancestry elsewhere than in the population (i.e. cluster) to which the individual was assigned in the genetic mixture

analysis. The colored segments in the bar correspond to the posterior probabilities of each possible ancestral origin of the allele under a uniform prior distribution. For example, if an individual was assigned to a blue cluster in the genetic mixture analysis, and the colored bar is entirely green, it corresponds to the case where the Bayes factor shows for this particular allele an overwhelming evidence for the ancestry in the green population.

By specifying the BAPS Ouput file via File – Output File, the results of the allele screening are written to the file (they are also shown in the log window). The results for the example image are as follows:

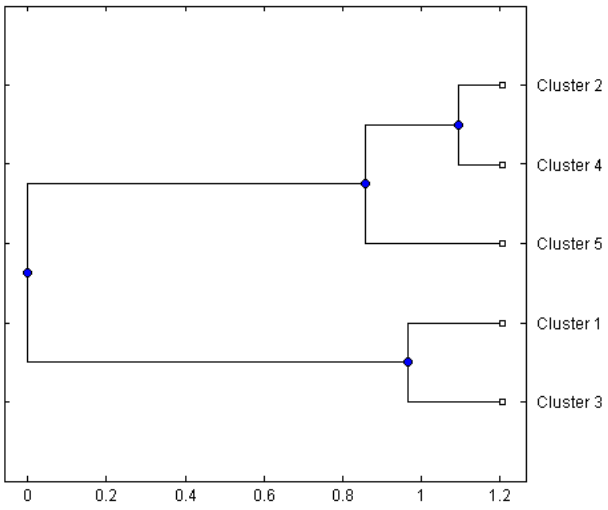
```
Putative locations for mutations:

locus,  haplotype, possible origins

16      1          3(3.1763) 5(2.767)

48      1          2(2.4618) 5(2.3355)

167     1          2(3.3364)

200     1          1(2.669)

231     1          5(2.3993)

330     1          1(4.8763) 2(3.1397)

51      2          3(4.0413)

177     2          2(3.0487) 3(3.2618)
```

The alternative origins of an allele are shown as cluster indices, with the strength of evidence (log Bayes factor) in parentheses. For instance, in the example case the log Bayes factor is ~4.04 in favor of ancestral origin in population 3 for the second allele at locus 51, which corresponds to the nearly entirely blue bar in the upper pane of the image below.



Displaying trees of clusters

The function 'Phylogeny' is accessible in the Graph menu when a mixture clustering result file has been loaded into BAPS. Two types of trees are available: Neighbor-Joining and UPGMA. These can be used to quickly explore the relationships among the identified clusters. An example UPGMA tree based on Nei's distances (averaged over loci) is shown below. The visual appearance of the tree can be changed through the Attributes menu in the graphics window.



Numerical results in the output file for mixture clustering:

An example of the result file for individual level clustering is given below. The reported changes in "logml" if individual i is moved to group j, refer to how much worse the solution would become with the indicated change. The number is the logarithm of the Bayes factor, so that exp("absolute value of the change") tells how many times better the optimal partition (i.e. clustering) is. The value is zero for the cluster where the individual is in the optimal solution. Very small absolute values of the change (<2.3, see Kass and Raftery, 1995), indicate that the individual could possibly be allocated to the alternative cluster as well. For detailed explanation of Bayes factors, see Kass and Raftery (1995). The K-L divergence matrix refers to the estimated Kullback-Leibler divergence between the clusters.

NB! It is possible to easily convert the the table of log marginal likelihood changes into conditional posterior probabilities for each individual. This can be done by transforming each element, say x_{ic}, on a particular row i according to the formula:
$y_{ic} = \exp(x_{ic})/[\Sigma_{c=1,\ldots,k} \exp(x_{ic})]$, where $\Sigma_{c=1,\ldots,k}$ refers to the sum over the k columns in the table (the k estimated clusters). The value y_{ic} refers to the conditional posterior probability of assigning the individual i into cluster c according to the data. Some modules of BAPS do directly output these assignment probabilities.

A list of sizes of ten best visited partitions with their log(ml) values is displayed. These values can be used to estimate the "correct" number of clusters. Also a posterior probability for the number of clusters is provided. This probability is based on the log(ml):s of the partitions that were visited during the current run. The probability is meant to be only a rough estimate. To get a better picture of the probabilities of different numbers of clusters you should run the program many times giving a vector of values for K in the beginning (see: Inputting the maximum number of populations, K). Then the probability would be computed on the basis of best partitions that were visited during all the runs. Also, the option of using the 'Fixed K' mode is useful when there is much uncertainty concerning the number of clusters, as the you can easily explore a range of different fixed K-values and the associated log(ml) values.

RESULTS OF INDIVIDUAL LEVEL MIXTURE ANALYSIS:
Data file: my_funny_data_set.mat

Number of clustered individuals: 42
Number of groups in optimal partition: 6
Log(marginal likelihood) of optimal partition: -1706.9897

Best Partition:
Cluster 1: {7, 8, 31, 32, 33}
Cluster 2: {1, 2, 3, 6, 9, 10, 11, 12, 13, 14, 15, 16, 17,
18, 19, 20, 24, 25, 26, 27, 28, 29, 30, 34, 35,
36}
Cluster 3: {4, 5}
Cluster 4: {37, 38, 39}
Cluster 5: {21, 22, 23}
Cluster 6: {40, 41, 42}

Changes in log(marginal likelihood) if individual i is moved to group j:

ind	1	2	3	4	5	6
1:	-36.9	.0	-50.3	-68.4	-67.9	-106.4
2:	-37.8	.0	-54.6	-72.8	-72.3	-107.9
.						
.						
.						
41:	-110.9	-167.0	-101.4	-103.9	-120.0	.0
42:	-103.5	-164.5	-100.4	-102.3	-120.8	.0

KL-divergence matrix:

	1	2	3	4	5	6
1						
2	0.311					
3	0.419	0.516				
4	0.543	0.632	0.355			
5	0.598	0.683	0.394	0.505		
6	0.710	1.021	0.667	0.677	0.807	

Numerical results in the output file for admixture analysis:

An example of the result file for admixture analysis performed by BAPS is given below. There were 5 clusters found by the clustering algorithm. The number in column i after the individual ID label, is the Bayesian posterior mean estimate of the proportion of the genome represented by the cluster i. Here, the first individual has estimated admixture coefficients .05, .87, and .08 for the 1st, 3rd and 4th clusters. The final column gives the *p*-value for the individual. This value tells the proportion of reference individuals simulated from the population in which the individual was originally clustered having the admixture coefficient to the cluster smaller than or equal to the individual. For instance, the p-value for the first individual in the example is 0.43 meaning that 43 per cent of the reference individuals simulated from population 3 (the population in which individual 1 was first clustered) had admixture coefficient to population 3 less than or equal to .87. Individuals having *p*-value larger than 0.05 are by default considered as having "non-significant" evidence for the admixture. Here, individual 1 has not "significant" admixture, whereas individuals 31, 561, and 562 have "significant" admixture. If a user wishes to apply a more stringent

"significance" limit, one can simply use a lower threshold for the values in the final column.

NB! Remember that when the admixture graphics are drawn by default, they display all posterior estimates of the admixture coefficients, irrespectively of the p-value. As explained in the Graphics section of this manual, the graphics can be re-drawn via Graph menu using a threshold (default 0.05), such that individuals with a p-value larger than the threshold are shown with a single-colored bar in the image (i.e. no evidence for the admixture).

```
RESULTS OF ADMIXTURE ANALYSIS BASED
ON MIXTURE CLUSTERING OF INDIVIDUALS
Data file: data_example_five_populations.mat
Number of individuals: 600
Results based on 50 simulations from posterior allele frequencies.

1:    0.05  0.00  0.87  0.08  0.00:  0.43
2:    0.00  0.02  0.84  0.05  0.09:  0.315
3:    0.05  0.04  0.91  0.00  0.00:  0.615
4:    0.00  0.00  0.99  0.01  0.00:  0.95
.
.
.
27:   0.01  0.00  0.92  0.02  0.05:  0.65
28:   0.09  0.00  0.88  0.03  0.00:  0.465
29:   0.07  0.00  0.77  0.16  0.00:  0.11
30:   0.00  0.00  0.98  0.01  0.01:  0.915
31:   0.12  0.00  0.72  0.16  0.00:  0.025
32:   0.11  0.00  0.89  0.00  0.00:  0.505
.
.
560:  0.99  0.01  0.00  0.00  0.00:  0.83
561:  0.49  0.00  0.00  0.51  0.00:  0
562:  0.37  0.00  0.04  0.59  0.00:  0.01
.
.
.
```

Installation

BAPS is currently available for Windows 2000/XP/Vista/Win7, Mac OS X and Linux operating systems.

The executable program can freely be downloaded from http://www.helsinki.fi/bsg/software/BAPS/. To use the software, a runtime component available at the website (different versions for the three operating systems) has to be installed first. The runtime component can be installed anywhere in the operating system, EXCEPT under the Matlab path, if Matlab is installed on your computer. After the installation of the runtime component is finished, unzip the BAPS package to any folder, and the program is ready for use (except for Mac OS X and Linux, see further installation details on the BAPS website). Notice that in Mac OS X you can only have a single version of the runtime component installed at a time.

References

Cheng L, Connor TR, Aanensen DM, Spratt BG, Corander J (2011) Bayesian semi-supervised classification of bacterial samples using MLST databases. BMC Bioinformatics, 12:302.

Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. (2013) Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Molecular Biology and Evolution, 2013, doi: 10.1093/molbev/mst028.

Corander J, Marttinen, P. (2006). Bayesian identification of admixture events using multi-locus molecular markers. Molecular Ecology, 15, 2833-2843.

Corander, J., Marttinen, P. and Mäntyniemi, S. (2006). Bayesian identification of stock mixtures from molecular marker data. Fishery Bulletin, 104, 550-558.

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. (2008a). Bayesian spatial modelling of genetic population structure. Computational Statistics, 23, 111-129.

Corander J, Marttinen P, Sirén J, Tang J. (2008b). Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics, 9:539.

Kass R, Raftery AE (1995) Bayes factors. J Amer Stat Assoc 90: 773-795.

Tang J, Hanage WP, Fraser C, Corander J. (2009). Identifying currents in the gene pool for bacterial populations using an integrative approach. PLoS Computational Biology, 5(8): e1000455.

Rob J. L. Willems, Janetta Top, Willem van Schaik, Helen Leavis, Marc Bonten, Jukka Sirén, William P Hanage and Jukka Corander (2012) Restricted gene flow among hospital subpopulations of *Enterococcus faecium*. *mBio,* 3, e00151-12.