Manual for Euroformix v1:

Installation and running program:

1) Run R (>=3.0.1) in Windows, Linux or MAC (http://cran.r-project.org/).
2) Required packages to run GUI:
    a. gWidgetstcltk (depends on digest,tcltk)
    b. gWidgets
3) Other required packages:
    a. cubature
        i. Required for multivariate integration (marginalized LR).
4) Installation and run gammadnamix:
    a. install.packages("gammadnamix", repos="http://R-Forge.R-project.org")
    b. library(gammadnamix)
    c. euroformix()

GUI

Toolbar:

- File
    - **Set directory**: The user may select the working directory of the R-program.
    - **Open project**: The user may open an earlier project which is saved in a file on the form "projectname.Rdata".
    - **Save project**: The user may save the existing project into a file on the form "projectname".
        - Extension .Rdata is added automatically.
        - All data imported to the program and complete calculations are stored into a single project-file which may be open at any time in the program.
        - Saving a project makes:
            - Big reference databases are stored efficiently (the required space for the database is drastically reduced).
            - Time-consuming calculations are restored instantly (only required to be calculated ones).
- Frequencies
    - **Set size of frequency database**: User may specify number of samples used to create the population frequencies (N).
        - When new alleles from imported files are found, these are assigned to freq0.
            - If N=0 (default), freq0 is equal minimum observed frequency.
            - If N>0,), freq0='5/(2N)'.

        - New alleles are updated to the population frequencies when:
            - Importing a new reference database.
            - Calculating LR (found in evidence or reference profiles used in the calculation).
            - Frequencies are normalized for each of these two cases.
                - WARNING: Normalizing may be done twice if new alleles (not seen in population frequency table or reference database) are observed in the evidence/reference profile.
- Optimization
    - **Set number of random startpoints**: The user may set required number of independent random startpoints in the optimizer to ensure that the global maximum is attained for the Maximum Likelihood Estimator (MLE). Default is 3.
    - **Set variance of randomizer**: The user may set the variance parameter used for the random generation of startpoints used in optimizer. Default is 10.
- MCMC (Markov Chain Monte Carlo)
    - **Set number of samples**: The user may set the number of samples drawn from the posterior distribution of the parameters. Default is 10000.
    - **Set variance of randomizer**: The user may set the variance parameter scalar used in the 'Markov Chain Monte Carlo (MCMC) random walk Metropolis'. See vignette for details. Default is 10.
        - Note that this value should be tweaked such that acceptance rate of sampler are around 0.2.
- Integration
    - **Set relative error requirement**: The user may set the required estimated relative error used in the integration function adaptIntegrate {cubature}. See vignette for details. Default is 0.005.
    - **Set maximum of mu-parameter**: The user may set upper limit of mu-parameter (amount of DNA). See vignette for details. Default is 20000.
    - **Set maximum of sigma-parameter**: The user may set upper limit of sigma-parameter (coefficient of variation). See vignette for details. Default is 1.
- Deconvolution
    - **Set required summed probability**: The user may set required summed posterior genotype-probability which the deconvoluted list is ensured to contain. Default is 0.9999.
    - **Set max listsize:** The user may set maximum number of elements in the deconvoluted list. Default is 1000.
        - The greater max listsize, the more time-consuming (and memory consuming) the search-algorithm behind will be.
- Database search
    - **Set maximum view-elements**: The user may set maximum number of elements to show from the reference-database. Default is 10000.

        - The greater ‘value’, the more time-consuming will it become to show the table.
        - Note that the result table from the database search shows only the top ‘value’-ranked elements.

Importing data (page 2):

DATA IMPORT:

- **Common** for all files:
    - The extension (denotes file-type) of the file names does not matter. It may also have no extension at all.
    - All imported files must be either comma, semi-colon or tab-separated (‘,’,’;’,’\t’).
    - Required/optional headers (all are capital invariant):
        - “**sample**” is required header for sample(s) name(s).
            - The sample names are NOT capital invariant.
            - If more than one header name contains “**sample**”, it will select the header name which in addition contains “**name**” in the same string.
        - “**marker**” is required header for marker name(s).
            - Marker names are capital invariant.
            - If no header is found, the header containing “**loc**” will be used if found.
        - “**allele**” is required header(s) for allele-information.
            - This may be a vector (“alleleX1”,…,”allelleX10”) of any length denoting allele(s) to a given marker for a given sample. Here X1,…,X10 can be anything.
        - “**height**” optional header(s) for peak height-information.
            - This may be a vector (“heightX1”,…,”heightX10”) of any length denoting peak height to the corresponding allele(s) in “allele”. Here X1,…,X10 can be anything.
    - Note:
        - The imported data will use upper-letter of marker-names found in the file.
        - The user should check that the data are imported correctly!

- **Import population frequencies**:
    - Requires an own folder (population-folder) with **only** frequency-files.
    - File-format:
        - Filename:

                - The name of the filenames **needs** to be on the form "kit_population.ext", where ext can be any extensions (or be missing as well).
                - kit="kit-name" and population="population name"
                - The kit-name must be consistent with the short-name of the kit instrument. See ?plotEPG for more details.
            - File:
                - First column needs to be allele-information (header-name may be anything).
                - Other columns are frequency-information (header-name denotes the locus name (loci names are converted to capital letters)).
    - To import frequencies:
        - Push "**1) Select directory**" button to select the population-folder with the population frequency files.
        - Push "**2) Import from directory**" button to import the population frequency files from the selected folder.
            - It is possible to **add new files** into the selected population-folder **at any time** and push the button once again to include new information to the dropdown-list.
    - Selection of kit and population:
        - After importing the frequency-files (after pushed (**2**)), the user may select wanted kit and population from the two drop down lists at any time* (*not after a reference-database file has been imported).
            - This can be useful to see the EPG layout for different selected kits.

- **Import Evidence/Reference** sample:
    - **Multiple** evidence or reference profiles are **allowed** in each file.
    - In evidence files:
        - "height" header is required for any analysis (Deconvolution, 'LR calculation' and 'Database search').
    - In reference files:
        - "height" header is optional but will not be used further in any analysis.
    - Note:
        - The import function will not check:
            - That the length of allele and heights are equal long for a given locus.
        - Locus without allele-information (i.e. empty or dropped out), are **NOT** imported.

- **Import Database**:
    - Multiple database file may be imported (**must** be done one-at-the-time).
    - **Requires** that population frequencies are imported and selected.
        - Needed for decoding allele names.

        - WARNING: Population frequencies may not be changed again after database importing!
    - Note:
        - Same samples needs to be in same block (and ordered) but markers within sample can be different orders.
        - Some samples **may** have more/less markers than others (e.g. SGMplus profiles contra ESX17).
            - **Missing markers** for a sample are given with NA.
        - Only markers shared with selected population frequencies are imported.
        - Homozygote genotype may have an empty allele under 'Allele 2'.

VIEW DATA:

- **View frequencies** (for a selected population):
    - Creates a new window which shows the selected population frequencies in a table.
    - If any evidence profiles(s) are selected after evidence-import, the software makes a 'false positive probability' – plot for each selected profiles.
        - The plot shows the 'false positive probability' of random matching at least (2*n-4) up to 2*n allele matches (MAC) for a evidence profile. Here n is number of considered loci (which are both in evidence and population frequencies).
    - Note:
        - Only allele-information in evidence-profiles are used.
        - New alleles which are not found in the selected population are assumed to have allele-frequency 0.
        - Number of possible mismatches is set to 4. This cannot be changed (can make it optional in future version).

- **View evidence** (for selected evidence):
    - Prints imported alleles (and peak heights if any) for each selected evidence profile(s).
    - Plots EPG(s) for each selected evidence profile(s)
        - Requires that user have imported "Population frequencies".
        - The kit selected under '**Select kit**' denotes the EPG format.
        - Loci in evidence which are **inconsistent** with the ones in selected kit (or missing) are **not shown** in plot.
        - Evidence profiles without peak heights for corresponding alleles are given with peak height equal 1.
    - Note:
        - See ?plotEPG to see which kit-formats that are supported.
        - Reference profiles can be imported as evidence profiles and shown in a EPG.

- **View reference** (for selected reference):

- Prints imported genotypes for each selected reference profile(s).
- If any evidence profiles(s) are selected after evidence-import, the software counts number of matching alleles (MAC) for each loci of the selected reference profiles, for each selected evidences.
    - MAC = number of alleles for the reference which are included in the evidence.

- **View database** (for selected database):
    - Creates a new window (for each selected database) which shows the genotypes for every reference in the database.
        - "NA" means that the genotype of a reference was missing.
    - If any evidence profiles(s) are selected after evidence-import, the software counts number of matching alleles (MAC) for all references in the database against each of the selected evidences. The results are shown in a MAC-ranked table in a new window (for each selected database).
        - MAC = number of alleles for the reference which are included in the evidence.
        - The summed MAC over all selected evidence are used to rank table.
        - nLocs is number of reference-loci which has been used to evaluate the MAC.
    - Note:
        - Max number of elements to view in a database can be changed under "Database search" in toolbar.

INTERPRETATIONS:

- **Generate sample**:
    - Generates alleles using the population frequencies and draws peak heights for a specified hypothesis using the continuous model as described in the vignette.
    - Requires: Imported population frequencies.
    - Feature: Simulates allele-dropout, drop-in (with a peak height model) and stutter.

- **Deconvolution**:
    - Deconvolution ranks the most probable combined genotype profiles given a specified hypothesis and the Maximum Likelihood Estimates of the parameters in the continuous model as given in the vignette.
    - Requires: Imported population frequencies and selection of at least one evidence profile with peak height information. References are optional to condition on in the hypothesis.
    - Feature: Model may handle replicates, allele drop-in, drop-out and stutter. A non-flat prior for the stutter-ratio can be applied.

- **LR calculation**:
    - LR calculation does weight-of-evidence by comparing the Likelihood Ratio (LR) between the specified hypotheses Hp (prosecution) and Hd (defence) using the continuous model as given in the vignette.
    - Modules:
        - 1) Maximum Likelihood Estimation; Optimizes model parameters.
        - 2) Marginalized integration; Integrates out the model-parameters.
    - Requires: Imported population frequencies, at least one evidence profile with peak height information and at least one reference profile (suspect) to weight evidence for. Additional reference profiles are optional to condition on in the hypotheses.
    - Feature: Model may handle replicates, allele drop-in, drop-out, stutter and fst-correction. A non-flat prior for the stutter-ratio can be applied.

- **Database search:**
    - Does weight-of-evidence by comparing the Likelihood Ratio (LR) between the specified hypotheses Hj (reference j in database) and Hd (defence) using the continuous model as given in the vignette.
    - Modules:
        - 1) Maximum Likelihood Estimation; Optimizes model parameters.
    - Requires: Imported population frequencies, at least one evidence profile with peak height information and at least one reference-database. Reference profiles are optional to condition on in the hypotheses.
    - Feature: Model may handle replicates, allele drop-in, drop-out, stutter and fst-correction. A non-flat prior for the stutter-ratio can be applied.

Model specification (page 3):

MODEL SPECIFICATION

- **Evidence(s)**:
    - Shows selected evidence(s) from 'Import data'.
    - All interpretations support multiple replicates.
    - All replicates are assumed to have same parameter sets (see vignette for details).

- **Contributors under Hp (case: Deconvolution):**
    - Not considered, since Deconvolution only considers the model under Hd.

- **Contributors under Hp (case: 'LR calculation' or 'Database search'):**
    - User may condition on selected references (from 'Import data') in the hypothesis Hp.
    - #unknowns under Hp:
        - Denotes number of unknown contributors under the prosecution hypothesis Hp.

        - Must be an non-negative integer

    - Special case for **'Database search':**
        - The contributor from the reference-database is already included in the hypothesis Hp.

- **Contributors under Hd (same for all cases):**
    - User may condition on selected references (from 'Import data') in the hypothesis Hd.
    - #unknowns under Hd:
        - Denotes number of unknown contributors under the prosecution hypothesis Hd.
        - Must be an non-negative integer

- **Parameters**:
    - Q-assignation: {checked,unchecked}
        - If checked, all alleles **not** present in the evidence are considered as allele "99". Its frequency will be given as the sum of the frequencies for all the "non-present" alleles.
        - If unchecked, the original alleles in the population are used as before.
    - 'Detection threshold': [0,->)
        - The threshold of allele peak heights of whether an allele is present in the evidence or not.
            - WARNING: If peak heights in evidence are lower than the specified threshold, the corresponding alleles below threshold are not removed automatically. Hence the user should carefully select a threshold lower than observed peak heights.
    - 'Stutter ratio': [0,1]
        - Stutter ratio is a constant parameter "**xi**" which denotes the proportion of peak heights from allele 'a' which is added to allele 'a-1'. See vignette for more details.
            - If allele 22 with peak height y_22 is contributed by a contributor and allele 23 did not have a observed peak height, then the stutter contribution to allele 21 from allele 22 will be (**xi*** y_22).
    - 'Probability of drop-in': [0,1]
        - Assumed probability of a random allele drop-in to the evidence at a given locus. See vignette for more details.
            - Can be any allele from the population.
    - fst-correction: [0,1]
        - Assumed co-ancestry parameter assigned in the genotype probability for each contributor in the hypotheses. See vignette for more details.
    - 'Dropin peak height hyperparam': [0,1]

        - Assumed hyper-parameter to model the peak height of the dropped in allele caused by a 'random allele drop-in' if 'Probability of drop-in'>0.
            - See vignette for more details.
    - 'Stutter prior function': {textstring}
        - The user may specify an assumed prior function to the 'stutter ratio' parameter **xi**. By default this is flat non-informative.

- **Database(s) to search (case: 'Database search')**
    - Lists the selected imported reference-databases to do the database search for.

DATA SELECTION

- **Select/unselect loci**:
    - The user may select or unselect loci for each selected evidence(s) and reference(s) from "Import data"-page.
    - Note:
        - If a locus has been unselected for any of the evidence(s) or reference(s), the unselected locus will not be evaluated at all.

- **Missing data**:
    - Data with missing allele or peak heights (for evidence) in any of the loci will automatically be deselected (inactivated) such that the corresponding loci will be unavailable to evaluate.

- **New alleles**:
    - If new alleles (does not exist in the population frequency table) occurs in the imported evidence or reference profile, the new alleles are assigned as 'freq0'. 'freq0' is equal minimum observed frequency in population if N=0, or 'freq0'=5/(2N) where N is size of imported frequency database under "Frequencies" in Toolbar. The frequencies are after normalized.

CALCULATIONS

- **'Maximum Likelihood Estimation'**
    - Optimizes the Likelihood of the unknown parameters given the assumed model so they attain maximum values for the specified hypothesis Hd (and Hp in case of "LR calculation").
        - The optimizer should return a global maximum. However, it may sometimes just return a local maximum. Number of startpoints should be increased to ensure that the optimizer finds the global maximum of the Likelihood function. This can be changed under "Optimization" in Toolbar.

- **'Marginalized LR' (case 'LR calculation')**:
    - Instead of optimizing the Likelihood of the unknown parameters, a **multivariate integration** over the unknown parameters are applied both under hypothesis Hp and Hd.
    - The accuracy of the integral depends on the specified 'relative error requirement' (see vignette for details). This can be changed under "Integration" in Toolbar.
    - In the output, also the relative error of the LR is given in brackets.
    - The integral requires that an **upper boundary** for the parameters mu (amount of DNA) and sigma (coefficient of variation) is specified. As default these are 20000 and 1, respectively. These values may be changed under "Integration" in Toolbar. See vignette for details.

- **'Generate sample' (case 'Generate sample')**:
    - A dataset will be randomly simulated under the specified model under "Model specification".
    - Reference profiles may be imported and selected as assumed known in the hypothesis.
    - Detection threshold, stutter ratio, probability of drop-in and drop-in peak height hyperparam may all be used in the simulation (**fst** and **stutter prior function** are not used).
    - The unknown contributor profiles under the hypothesis will be randomly generated using the selected population frequencies.
    - The simulated peak heights of the evidence in the dataset are entirely based on the continuous model for assumed values of the model-parameters (**mu,sigma,xi,mx)**. Default these are given as **mu**=1000, **sigma**=0.15, **xi**=0.1, **mx**=(C:1)/ sum(C:1), where C is number of contributors.

MLE fit (page 4):

MAXIMUM LIKELIHOOD ESTIMATES UNDER Hd (and Hp for case: '**LR calculation**')

- **Confidence Interval of parameters**:
    - param: The unknown parameters in the model (see vignette).
    - qq2.5 and qq97.5: Denotes the 95% normal approximated confidence interval of the unknown parameters in the model (see vignette).

- mode: The optimized[1] parameters in the model which attains a maximum point of the likelihood function.

- **Maximum Likelihood value**:
  - log10lik and Lik: The ten-logged and the original value of the Likelihood value attained from the optimization[1].
  - Laplace P(E): The Laplace approximated integral (marginalized LR).
    - A normal-based approximation of the integral where parameters in model are integrated out (see vignette).

- **Further Action**:
  - MCMC simulation:
    - Performs 'Markov Chain Monte Carlo (MCMC) random walk Metropolis' samples under the desired hypothesis.
      - Uses the mode and the covariance matrix attained from the optimization. See vignette for details.
    - The first column in the output shows the estimated posterior distributions for each of the unknown parameters in the model.
    - The second column in the output monitors the parameter samples in the simulation.
    - A message-box is shown with the acceptance rate of sampler. This is calculated as number of accepted samples divided by number of proposed samples. Ideally this should be around 0.2 to ensure that parameter space has been fully explored. This number will change when the variance of the randomizer is changed.
    - User may change the number of required samples in the simulation and the variance of the randomizer under "MCMC" in Toolbar.
    - The purpose of the MCMC simulation is to use it as an exploratory tool to see:
      - That the optimizer has found the global maximum.
      - The shape of the posterior distribution of the parameters.

  - Deconvolution:
    - Performs "Deconvolution" under the desired hypothesis. (See Deconvolution (page 5) for details.

  - Model validation: NOT IMPLEMENTED YET

FURTHER EVALUATION

- **Optimize model more**:

[1] This may be only a local maximum point, not the global maximum which will be the Maximum Likelihood Estimate. Increase number of start points under "Optimization" in Toolbar to ensure a global maximum.

    - The optimization procedure can be run again with the same specifications as selected in "Model specification" to ensure that a global maximum is attained.
    - It is recommended to do this and check that the optimized Likelihood value is not changed (increased) further.
    - Note: The optimization with greatest Likelihood value will always be shown in this page.
- **Database search (case: 'Database search'):**
    - A database search with the specified continuous model will be applied. (See Database search (page 6) for details.
- **Marginalized LR (case 'LR calculation')**
    - See CALCULATIONS under section "Model specification (page 3)".

SAVE RESULTS TO FILE

- **All results**:
    - The confidence interval of the parameters and the likelihood values will be printed to file for all hypotheses on page.
- **Only LR results**: **(case 'LR calculation')**
    - The LR calculated values shown in WEIGHT-OF-EVIDENCE will be printed to file.

WEIGHT-OF-EVIDENCE **(case 'LR calculation')**

- Description:
    - The 'LR calculation' weights the likelihood of the two specified hypothesis Hp and Hd specified in page 3.
    - The 'LR calculation' is based on the continuous model and may hence handle allele drop-in, drop-out and stutter.

- **Maximum likelihood based:**
    - LR: 'Likelihood value under optimization under Hp' divided by 'Likelihood value under optimization under Hd'
    - log10: The ten-logged value of LR.
- **LR for each loci:**
    - The LR for each loci separately (under the parameter-optimization under Hp and parameter-optimization under Hd). See vignette for detail.
- **Laplace approximation based:**
    - LR: The ratio of the normal-based integral-approximations under the optimizations of the hypotheses Hp and Hd. See vignette for detail.

Deconvolution (page 5):

- Description:
    - Deconvolution is applied for a given specification of the continuous model under a specific hypothesis. It conditions on the optimized parameter (i.e.

the MLE under page 4 MLE fit) to resolve a ranked list of the **posterior probabilities** of the combined genotype-profiles (see vignette for details).
    - The deconvolution is based on the continuous model and may hence handle allele drop-in, drop-out and stutter.

- **Table**:
    - The columns in the table show the resolved genotype for each contributor in the specified hypothesis (per locus).
    - The combined profiles are ranked due to their **posterior probabilities**.
    - Note:
        - Having only sub-optimized parameters will not give the most likely genotypes.
        - Q-assignation is recommended to use since dropped out alleles are equally threated and assigned as "99".
    - Size of table:
        - The length of the table ensures that the sum of the **posterior probabilities** are at least 0.9999.
        - Maximum size of table is default 10000.
        - These two quantities can be changed under "Deconvolution" at the Toolbar.
- **Save table:**
    - The full table will be exported to a tabulator-separated text-file.

Database search (page 6):

- Description:
    - The 'Database search' is very similar as the 'LR calculation' with the only difference in that each individual in the reference-database is assumed as a contributor in the hypothesis Hp and hence give a LR-value for each individual (LRj). The resulting table will rank the individuals due to LRj.
    - The 'Database search' is based on the continuous model and may hence handle allele drop-in, drop-out and stutter.

- **Table**:
    - The table shows the ranked individuals in the database due to their continuous LR values (**cont.LR**).
    - 'Reference name' is name of individuals given in the reference-database.
    - MAC (Matching allele counter) is number of alleles in the reference-profile which matches the evidence(s).
    - nLocs.LR is number of unique loci which the reference is evaluated with.
    - nLocs.MAC is number of loci in the reference-profile which are used to calculate the MAC.

        - Note: Some references in the database may be registered with fewer loci than the evidence has. This should be pointed out.

    - Note:
        - Maximum number of elements to view a 'Database search' result table is 10000. This can be changed under "Database search" at the Toolbar.
        - Putting fst>0 may be very time-consuming and therefore not recommended in a database search.

- **Save table:**
    - The full table will be exported to a tabulator-separated text-file.

Generate data (page 1):

- Description:
    - Generates alleles using the population frequencies and simulates peak heights for a specified hypothesis using the continuous model as described in the vignette.
    - The generation may simulate allele-dropout, drop-in (with a peak height model) and stutter.

- **Parameters**:
    - **mu** - amount of DNA
    - **sigma** - coefficient of variance
    - **xi** – stutter ratio
    - **mx=(mx1,…, mxC)** – mixture proportion for contributor 1,..,C.
        - Note: **mx** will be normalized if it's not already.

- **Edit**:
    - Loci: Loci name of the population frequency used to generate the dataset.
    - Evidence: The allele information is given in the left column while the peak height information is given in the right column. Each element **needs to be** separated with ",".
    - Reference: The alleles of the true contributors to the generate evidence is sequentially shown in each column.
    - All the loci names, evidence-allele and heights and reference-alleles may be edited before storing.

- **Import/Export**:
    - **Save data**:
        - Stores the generated (and possible edited) evidence- or reference-profile to a file.
        - Extension .csv added automatically.
    - **Load data:**

        - Loads profiles from file into the selected entries (evidence or reference).
        - If any locus is missing from the loaded evidence or reference file, the edit-cell will be empty.
        - The order of the loci in the file does not matter.

- **Further action:**
    - Generate again**:** Make a new simulation using the selected values of the parameters under **Parameters**.
    - Plot EPG: Plots the generated (and possible edited) evidence in a EPG plot.
        - It will use the "kit" selected under "Import Data"-page.
        - See ?plotEPG to see which kit-formats that are supported in the EPG.

To be implemented:

- Label the alleles of the selected references to the EPG-plot.