JoinMap ® 5 – New features

This document describes the new features of JoinMap version 5 that were added in comparison to version 4.1. The enhancements are intended as improvements for working with very large sets of markers.

Many features are not yet available in the current early release edition of JoinMap v5. They will be made available in future releases during the next months.

These are the new features available in the current early release edition of JoinMap v5:

1. 64-Bit
The v5 executable program is a 64-bit MS-Windows application. This means that it can have access to more than 4 GB memory (ie RAM) of the computer (4 GB is a 32-bit limitation). Obviously, the program requires that it runs under a 64-bit version of the MS-Windows operating system and that the computer has more than 4 GB of RAM; an amount of 16 GB of memory is recommended. The access to more memory means in practice that computations can take place in memory without having to store intermediate results on the hard drive, resulting in higher speeds.

2. Database driven
The various data within the JoinMap v5 projects are stored in databases. Viewing these data in tables, tree views and as plain text is made in such a way that only the currently visible part is retrieved from the databases. In contrast to this, in the previous versions of JoinMap always the entire data were retrieved from file, which can make the program very slow if the file sizes become extremely large (which is especially the case when dealing with pairwise data). Using a database system greatly improves the responsiveness of the user interface with large datasets; with smaller datasets the responsiveness is slightly slower than with previous JoinMap versions due to the database system overhead. The database system used is the embedded SQL database engine called SQLite, which does not require any database server installation or maintenance.

3. Parallelization
Some calculations were enhanced to be able to run in parallel. Modern CPUs have often multiple computation cores. JoinMap v5 makes use of all available cores, thus the more cores the computer has, the faster the computations: the speed scales linearly with the number of cores, except for a small amount of overhead. It is difficult or sometimes even impossible to change algorithms towards a parallel approach. In JoinMap v5 the following algorithms run in parallel: the calculation of (a) the locus similarities and (b) the individual similarities in population nodes, (c) the determination of the groupings in population nodes, (d) the computation of the recombination frequencies in group nodes. A parallel algorithm is only really useful if all the data are accessible in memory (ie RAM). In practice, this may be problematic for the determination of the groupings of very large datasets, because the amount of pairwise information scales quadratically with the number of markers. Therefore, these computations are made in such a way in v5 that they dynamically switch to regular serial (vs parallel) computations using temporary files on the hard drive if it turns out insufficient memory is available. It is good to realize that the speed of the hard drive will be the limiting factor in dealing with large datasets: the enormous amount of results of computations on all pairs of markers must be stored in huge database files.

4. Identical loci
The production of a reliable high resolution linkage map does not only require many markers, it also requires a population of sufficient size containing the necessary segregation information. Unfortunately this latter aspect is not always the case. In such instances, there will be a large amount of redundancy in the marker data, which should lead to many markers being identical. In most computations JoinMap v5 determines which markers have an identical segregation pattern and will perform the requested computation only for the representative marker of any set of identical markers. Subsequently, it will present the same results for the identicals next to their representative. Detecting and removing identical markers in the population nodes, as advised in previous versions of JoinMap, is therefore not necessary in v5.
5. **Multipoint recombination frequency estimation**
For all population types, except type CP, the time consuming Gibbs sampling (which is a so-called Monte Carlo Expectation-Maximization (EM) algorithm) to determine the multipoint recombination frequencies in the maximum likelihood mapping procedure is replaced by a much faster true EM algorithm (implemented as a so-called forward–backward algorithm). For population type CP this could not be achieved due to the complexity of the likelihood.

6. **Batch computations**
Identical computations can often be done in batches in v5. For this, nodes of the same type in the navigation panel must be marked by right-clicking. When subsequently the calculations for a certain selected tabsheet are requested, the same calculations will be done for all marked nodes. This also applies to the calculation of maps with group nodes. Double-clicking on a node will mark all nodes of the same type with the same parent node, eg all group nodes of a grouping node, or all population nodes. Double-clicking in the navigation panel but not on any node will un-mark all nodes.

7. **Archive**
Any JoinMap project may grow to a situation where the navigation tree contains very many nodes. At some point, certain nodes may be regarded redundant, at least for the time being. JoinMap v5 offers the possibility to move entire tree branches to be stored under an archive node. There, the data remain available for viewing and even for computations, while at the same time the more essential part of the navigation tree remains more clearly arranged. Archived branches can be returned to the regular project tree if needed. The two relevant functions reside in the File menu.

8. **Marking regions**
Marking rows in the tables for functions on the loci or individuals in these rows is not done the standard MS-Windows way, but more easily by marking the top and the bottom row by (a) right-clicking, (b) pressing the keyboard space bar or (c) the F8 key: the first right-click (or key stroke) on a row or in the text sets both marks; any next right-click (or key stroke) replaces the second mark. Pressing the Ctrl+A key combination marks the entire text or set of rows. The marked region will be highlighted. Cancelling the marking can be done (a) by pressing the Esc key, (b) by double-clicking anywhere in the table or text and (c) by right-clicking the row where you started the marking. In the Dataset tabsheet (see below) the two marks correspond to top-left and bottom-right cells instead of top and bottom rows. In tabsheets with plain text the marks correspond to the first character and the last character of a contiguous part in the text. In editable plain text tabsheets the marking cannot be done with the space bar as this is a character that may be entered into the text; instead you can use the F8 key.

The functions acting on marked regions, which are now enabled, are (a) **Exclude marked items of Individual Genot. Freq.** and **Locus Genot. Freq.** tabsheets, (b) **Set X2-Test Classification for Marked Loci of the Locus Genot. Freq.** tabsheet of a population or map node, (c) checking or un-checking a range of checkboxes, (d) copying and exporting tables and text, (e) moving loci in a grouping node, (f) **Exclude Marked Loci in Group Node** of various tabsheets of a map node, and (g) **Flip Marked Genotypes** of a dataset node.

9. **Dataset**
The dataset node functionality is renewed to be able to accommodate thousands of loci. For instance, a set of data of 50,000 markers for 100 individuals copied from an MS-Excel spreadsheet can be pasted into a JoinMap Dataset tabsheet, which takes less than half a minute. Within the Dataset tabsheet copying, cutting and pasting can be done with the regular key combinations and with the corresponding toolbar buttons. Marking regions for copying and cutting is done as described above. The checking of the coding with the Dataset menu function **Check for Coding Errors** will generate a report that is available under the blue i-button of the tool bar. If there are coding errors, it will also indicate the first and the last error detected. The Dataset menu function **Flip Marked Genotypes** will recode genotypes in the opposite way, in case the coding had been done erroneously, eg the code A was used instead of B and vv. The recoding depends on the population type.

10. **Project reconstruction**
A major effort was made for JoinMap to be fault tolerant with project files. In the hopefully very rare occasion that the project cannot be properly opened or appears corrupted, you may have JoinMap attempt to reconstruct the project database (Project.sqlite in the project directory). You can do this by renaming the original project database to eg Project.sqlite.bak and next open the project in JoinMap. The program will attempt to reconstruct the project database as good as possible. If the reconstructed project appears better than that from the original project database, you may decide to continue with this reconstructed version, and otherwise you close the project and replace the new project database by the .bak version in order to continue with the original version.

JoinMap ® 5 – New Features – p.2
11. Further
Having large datasets will make that computations, administrative work and database actions may take a while. In order to give feedback to the program user, messages are given on the status bar and often progress is indicated with a growing progress bar. The program maintains a history of the last 50 messages that were shown on the status bar (as long as the program is active); this message history can be accessed by right-clicking on the status bar. Some database actions cannot be predicted for their duration, so that the standard progress bar growing to 100% cannot be used. To give feedback that the program really is busy in such cases, the progress bar area will show sequences of ‘>’ symbols.

12. Added functions and modifications
(3 June 2017)
  Toggling a checkbox in tables (e.g. in the Exclude column of the Loci tabsheet) can be done with the Enter key, but only after having clicked once in the checkbox column.
(14 July 2017)
  The Grouping menu function Move Marked Loci.
(11 August 2017)
  Copying, cutting and pasting marked regions in plain text tabsheets.
(25 September 2017)
  • Map charts.
  • Copying marked regions in the Further Information window.
(7 October 2017)
  Export Map tabsheets as map-files.
(27 October 2017)
  • Export Map tabsheets from multiple marked map nodes as one joint map-file (e.g. for use with MapQTL). Unlike the normal export function, here all records will be exported of each Map tabsheet (i.e. not just the marked records).
  • Export Data tabsheets from multiple marked nodes (population, group or map nodes) as one joint loc-file (e.g. for use with MapQTL). Unlike the normal export function, here all records will be exported of each Data tabsheet (i.e. not just the marked records), except the data of the loci and individuals indicated as excluded on the Loci and Individuals tabsheets (if present), respectively.
(6 November 2017)
  The Colorize function for graphical genotypes takes the linkage phases (if available) into account for population types CP, DH and HAP.
(20 December 2017)
  • The map building approach of the ML mapping algorithm was adjusted in order to produce more reliable results for high density maps. Spatial sampling is used for gradually building the map. After the sampling threshold of 0 is reached, the map building proceeds with adding batches of random loci instead of simply adding all remaining loci in one go. The batches are of fixed size (default 50 loci), which can be adjusted in the Calculation Options. Three of the map optimization (simulated annealing) parameters are now specific to each spatial sampling level (now called S1 to S5) and to the building with batches after the sampling threshold of 0 is reached (called T0). This is meant to be able to speed up the first map building levels. The three parameters are: (1) the number of optimization rounds per sample, (2) the chain length with constant acceptance probability and (3) the stop criterion (stop after # chains without improvement).
  • The history of the status bar is increased from 25 to 50 messages.
  • Two parameters used in the true EM approach with the multipoint recombination frequency estimation are adjustable in the Calculation Options: the maximum number of iterations and the recombination frequency tolerance. (The iterations stop when either the maximum is reached or when the relative change in the sum of rec. frequencies of adjacent segments is less than the tolerance.) Generally, these parameters do not need to be changed.
  • When applying the Invert Map function of a map node, then on the Plausible Positions tabsheet the map was inverted, however the position counts were not shown at their inverted positions and a recalculation would be required. This is now working as expected.
(8 January 2018)
  In the ML mapping algorithm corrections were made for two errors that were present since the first pre-release. (1) For population type BCbxFy the expected recombination counts, the map distances and possibly the map order were estimated erroneously. (2) The N.N. Stress was computed erroneously for a few population types: BC1 (only if coded with H's and B's), BCbxFy, DH and HAP.
(21 February 2018)
  • Map nodes resulting directly from mapping contain various tabsheets with diagnostics about the loci, such as the N.N. Fit and Stress. In these tabsheets, rows can be marked and subsequently the Map menu function Exclude Marked Loci in Group Node can be applied. In the map node's grandparental group node this will check the Exclude checkboxes of the loci that correspond to the marked rows.
• Marked regions in the Dataset tabsheet and in editable plain text tabsheets (eg Notes) can be deleted with key combination Ctrl+Del and with an option of the Edit menu.

• A new menu function Flip Marked Genotypes was added for the dataset node. Sometimes genotypes may have been coded erroneously, eg the A was used instead of the B and vice versa; this may result in recombination frequency estimates larger than 0.5, showing up in a group node's Suspect Linkages tabsheet. This new flip function will allow easy correction of such genotype data in the dataset node. The function will replace genotype codes in the marked region in their opposite codes, in effect exchanging the (grand)parents of segregating population.

  The type of replacement depends on the population type. In a BC1, the A's will be replaced by H's and vv if there are A's in the marked range, while if there are B's in the marked range then B's will be replaced by H's and vv. For population types DH, HAP, HAP1, DH1 and DH2 the A's will be replaced by B's and vv. For population types BCpxFy, Rlx, IMxFy and F2 the same happens, and in addition C's will be replaced by D's and vv. For population type CP the following genotype mutual replacements take place: ad ↔ bc, ef ↔ eg, ll ↔ nn, lm ↔ np, while in addition the segregation types <lmxll> and <nnxnp> will be replaced by each other as well as their corresponding phase and classification types. (Remark: the other CP genotype codes are symmetrical with respect to the parents and do not need to be changed.)

  In all cases the replacement is case-insensitive and also the genotype code must be a correct one, incorrect codes are left unchanged. Therefore, it is advised to apply (and pass succesfully) the Check for Coding Errors function in advance to the flip.

• Another function was added to the Dataset menu: Create Dataset from Data Tabsheet. This function creates a new dataset node and fills it with the genotype data of a marked population, group or map node that has a Data tabsheet with genotype data. The data of loci and individuals that are checked as excluded on the Loci and Individuals tabsheets, if present, will not be transfered to the new dataset node.

• The color of marked regions in the tables is changed to red. (Marked regions in plain text remains the MS-Windows default, usually blue.)

• The system of marking nodes in the navigation panel is improved.