Using MareyMap

Delphine Charif Clément Rezvoy

13th May 2007

Following is a tutorial detailing the main tasks you will carry out while estimating local recombination rates using MareyMap. this tutorial assumes that you have already installed MareyMap and his dependencies as explained in Installing MareyMap[1].

1 Launching MareyMap

MareyMap is a set of R packages and therefore runs inside of R. The first thing you need to do is to launch R(under MacOS X and Windows you should have a menu entry for R, under linux just type `R` in a terminal window). Once R is started, type `library(MareyMapGUI)` at the R prompt (the `>` sign) and press `enter`. This will load MareyMap user interface as well as all the required packages. You can now start the graphical interface by typing `startMareyMapGUI()` and pressing `enter`.
The main window is composed of a menu bar and 5 different frames with which you can interact.

The graphical interface (Fig. 1) uses maps from the packages MareyBase by default. It is however possible to load a different collection using the *file* menu. After selecting a map in the list *Maps* ①. The selected map is displayed in the central part of the interface ②.

2 Map cleaning

Maps sometimes contain markers for which you may have valuable reason to believe that they have been misplaced (for instance if they break the monotonous growth of the Marey curve). clicking around a marker on the map will display information about this marker in the marker frame ④. If you unset the valid radio button, this marker will not be taken into account during the interpolations. You may as well completely delete the marker.

[1] http://pbil.univ-lyon1.fr/ rezvoy/mareymap/doc/Installing_MareyMap.pdf

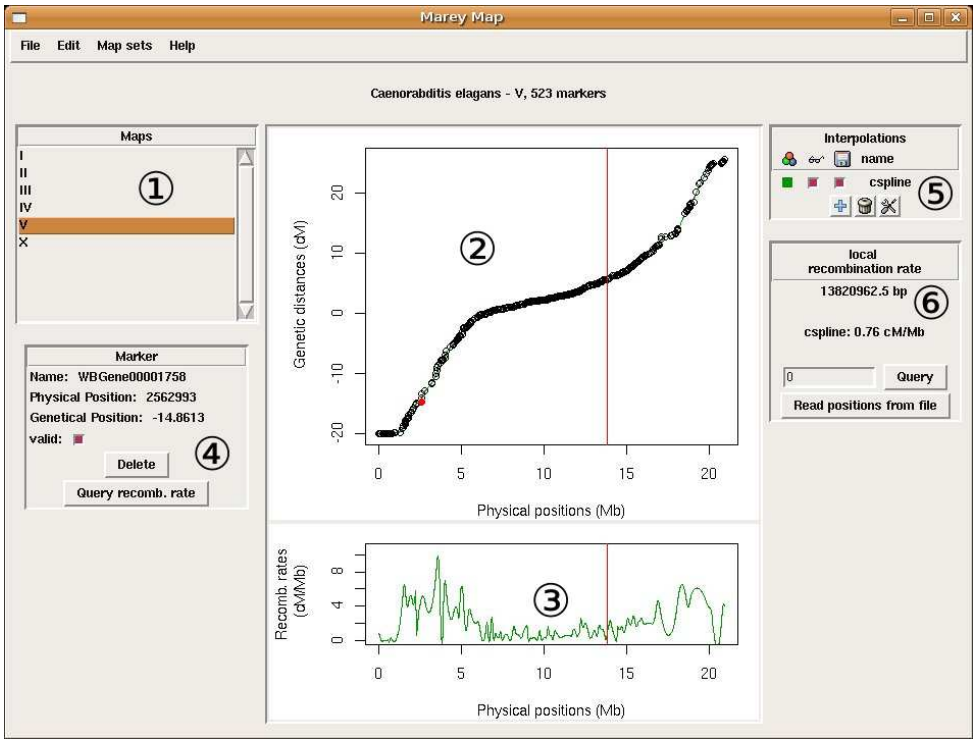

Figure 1: Detailed view of the main window.

Note the default map collection is reloaded by default each time you launch MareyMap, if you want to preserve you work, you will have to save your map to file.[2].

3 Interpolations methods

To add an interpolation to a map, click onto the sign in the **interpolations** frame ⑤ and select an interpolation method from the list. Once the interpolation method has been computed the result is displayed in the bottom of the display frame ③. MareyMap currently provides three interpolation methods *Sliding Windows*, *Loess* and *Cubic Splines*.

3.1 *Sliding window*

This methods estimates the local recombination rates by carrying out linear regressions inside windows of a given physical size. You may adjust the size of the windows (parameter **size**), the distance between two successive windows (parameter **shift**), as well the minimum number of marker per window to validate the interpolation (paramater

[2]see section 5 *saving/loading/exporting maps*

thereshold).

3.2 *Loess*

Loess (or lowess for LOcally WEighted Scatterplot Smoothing) estimates the recombination rates by locally adjusting a polynomial curve (1^{st} or 2^{nd} degree). The size of the window is defined as a percentage of the total number of markers and therefore can adapts to the variation of the density of markers across the map.Inside of a given window each marker is attributed a weight depending on how far they are from the center of the window. the parameters $\hat{\beta}$ of the curves are those that minimize the mean squared deviation between the data and the curve:

$$Q = \sum_{i=1}^{n} \omega_i [y_i - f(x_i, \hat{\beta})]^2$$

Where (x_i, y_i) are the observed data and ω_i is the weight of each marker calculated by:

$$\omega(u) = (1 - u^3)^3$$

with:

$$u = \frac{|x_0 - x_i|}{max_N(x_0)|x_0 - x_i|}$$

For this method, you can select the degree of the fitted curve (parameter **degree**) and the size of the window (parameter **span**). The span parameter is the percentage of the total number of points to take into account to calculate the local polynomial on the neighborhood of a marker. Span controls the degree of smoothing. It is constant over the entire range of predictor values. However, a constant value will not be optimal if either the error variance or the curvature of the underlying function f varies over the range of x.

This method is directly based on the function *loess*. to get more information about this method you can type `?loess` at the R prompt and press `enter`.

3.3 *Cubic splines*

A cubic smoothing spline behaves approximately like a kernel smoother, but it arises as the function $\hat{f}$ that minimizes the penalized residual sum of squares given by:

$$PRSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(t))^2 dt$$

3.3.1 spar

λ is the smoothing parameter, corresponding to the span in loess. A different λ can be specified using the `spar` argument. It is not intuitively obvious what a "good" choice of λ might be. In general, you should let R estimates the smoothing parameter either by locally or generalized cross-validation.

3.3.2 degree of freedom

Controls the amount of smoothing by setting the degree of freedom, which corresponds to the trace of the smoothing matrix. It is not intuitively obvious how to choose a value for this parameter and is often more convinient to rely on spar or cross-validation.

3.3.3 cross validation

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \hat{f}_\lambda^{-i}(x_i))$$

Here $\hat{f}_\lambda^{-i}(x_i)$ is the leave-one-out smooth at x_i, that is, it is constructed using all the data except for (x_i,y_i) and then the resultant least squares line is evaluated at x_i. CV is calculated for different values of λ and the λ that minimize this criterion is choosen. The "generalized" cross-validation method should be used when there are duplicated points in 'x'.

This method is directly based on the function *smooth.spline*. to get more information about this method you can type `?smooth.spline` at the R prompt and press `enter`.

You can update the parameters of your interpolations by clicking on the  icon in the **interpolations** frame and remove them by clicking on the  icon. Interpolations can be either visible or invisible depending on their toggle button state. the  toggle button indicates wether the interpolation should be kept or not when saved to text file.

4 Queries

Once interpolations have been defined on the map, you can query local recombination rates using the frame **local recombination rate ⑥**. There are 3 different ways of using this frame. First of all, you can query the local recombination rate on a physical position of the currently displayed map. the position must be specified in base pair (*ex.* 31564623) but can also be expressed using **Mb** or **Kb** (*ex.* $31Mb$, $564Kb$ or even $31Mb + 564Kb + 623$).

You can also specify several position on the currently displayed map at once by separating them by **:** (*ex.* $31Mb : 12287456 : 44Kb + 564$). In that case the results are displayed in a separate window and can be saved to text file. Finally this frame also allow batch queries where position are read in from a text file (see fig. 2 for example). to use this feature you can either enter the path to the file you want to process or click on **read positions from file** and select the file using the file selector dialog. the input file must be a text file containing at least a column **chr** and a column **phys** indicating respectively the map and the physical position of each query. This file can also contain a column **spc** if your query spans over several species (if this column is not present all the queries are carried out on the current species. Any other column will be ignored by the program. Such files can be easily created using spreadsheet softwares.

```
set map phys
"Homo sapiens (mean)" "Chromosome 03" 10mb
"Homo sapiens (mean)" "Chromosome 03" 50mb
"Homo sapiens (mean)" "Chromosome 03" 100mb
"Homo sapiens (mean)" "Chromosome 03" 150mb
"Homo sapiens (male)" "Chromosome 03" 10mb
"Homo sapiens (male)" "Chromosome 03" 50mb
"Homo sapiens (male)" "Chromosome 03" 100mb
"Homo sapiens (male)" "Chromosome 03" 150mb
"Homo sapiens (female)" "Chromosome 03" 10mb
"Homo sapiens (female)" "Chromosome 03" 50mb
"Homo sapiens (female)" "Chromosome 03" 100mb
"Homo sapiens (female)" "Chromosome 03" 150mb
```

Figure 2: example query file.

5 saving/loading/exporting maps

Upon starting, MareyMap will reload the default map collection. Any changes made to the map can be saved to **R data** files or to text files in order to preserve your work. **R data** is the standard binary file format of **R** it allows you to save directly the R objects with no transformations.

Saving to text file will create a file containing a line per marker with columns: **spc** for the species name, **chr** giving the name of the chromosome (or name of the map), **phys** giving the physical position of the marker, **gen** giving the genetic position of the marker, and the column **valid** indicating if the marker is valid or not. The file also contain a column

per interpolation with the local recombination rate calculated for each marker. Call order to recreate the interpolations are also saved as comments at the beginning of the file (see fig. 3 for example).
Maps can also be graphically exported as jpeg,png,pdf or eps.

```
"set" "map" "mkr" "phys" "gen" "vld" "default name"
"Caenorhabditis elegans" "III" "WBGene00003670" 13482.5 -27.2631 TRUE 1.74
"Caenorhabditis elegans" "III" "WBGene00003931" 91912.5 -26.9858 TRUE 1.33
"Caenorhabditis elegans" "III" "WBGene00006895" 148578.5 -26.9829 TRUE 0.85
"Caenorhabditis elegans" "III" "WBGene00006894" 160182 -26.9799 TRUE 0.73
"Caenorhabditis elegans" "III" "WBGene00001963" 186286 -26.9769 TRUE 0.48
"Caenorhabditis elegans" "III" "WBGene00003510" 187571.5 -26.9573 TRUE 0.47
"Caenorhabditis elegans" "III" "WBGene00003134" 478878.5 -26.9381 TRUE 0.98
"Caenorhabditis elegans" "III" "WBGene00006781" 496804 -26.9298 TRUE 1.21
"Caenorhabditis elegans" "III" "WBGene00002269" 599935.5 -26.7291 TRUE 2.54
"Caenorhabditis elegans" "III" "WBGene00001412" 633324.5 -26.5299 TRUE 2.95
"Caenorhabditis elegans" "III" "WBGene00001709" 678078.5 -26.3941 TRUE 3.49
"Caenorhabditis elegans" "III" "WBGene00001573" 739817.5 -26.2617 TRUE 4.2
"Caenorhabditis elegans" "III" "WBGene00004212" 767591 -26.1457 TRUE 4.56
"Caenorhabditis elegans" "III" "WBGene00004922" 789889 -26.0233 TRUE 4.91
"Caenorhabditis elegans" "III" "WBGene00000903" 812789 -25.8937 TRUE 5.33
```

Figure 3: excerpt from a map input/output text file.