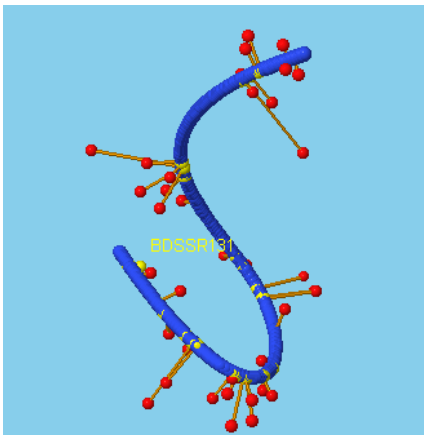
Documentation for THREaD Mapper Studio version 1.0

(Last updated April 25th 2010)



Release: 1.0 (30.12.2009)

Website address: http://cbr.jic.ac.uk/threadmapper/

Website name: THREaD Mapper Studio

Authors: Jitender Cheema[1], Noel Ellis[2] and Jo Dicks[1] (Email:jo.dicks@bbsrc.ac.uk)

1) Department of Computational and Systems Biology, John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH.
2) Department of Crop Genetics, John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH.

License:

Table of Contents

N.B. Users who wish to read just a quick start tutorial may find that skipping straight to section 15 proves useful.

1) Purpose of the website:
THREaD Mapper Studio version 1.0 is a web based tool that implements a novel, visual and interactive method for the estimation of genetic linkage maps from molecular markers. Particularly valuable features inherent to the method mean users can: a) visualise their maps in 3D, often allowing them to spot errors or problems with their maps, b) compare their maps to those genetic maps estimated under an alternative methods or to physical maps and, c) use a combination of algorithm- and user-based decisions in developing their maps.

2) Requirements:
THREaD Mapper Studio can be run within *JavaScript* and *Java Runtime Environment* enabled web browsers. Make sure you have the relevant Java plugin installed. We have tested the website extensively within the Internet Explorer 7, Internet Explorer 8, Firefox and Google Chrome browsers under Windows XP. We have seen a good performance with Google Chrome and Firefox on windows XP so we recommend these browsers to those wishing to use THREaD Mapper Studio extensively. We have used Firefox 3.5 as the reference browser during development and testing.

For each analysis which starts from the index or homepage, THREaD Mapper Studio assigns a ***session*** for that analysis and the session is stored on the server for a limited or fixed time, depending upon server load. Consequently, make sure ***you have cookies enabled*** otherwise you will repeatedly get *session expired errors*, as THREaD Mapper Studio relies heavily on sessions for server-browser interaction. If you encounter this error or the exception message *session expired*, please use your browser's Back button to go the previous step, or to keep going back until you no longer see that message. As each THREaD Mapper Studio analysis is divided into easy steps, parameter values may be changed during a session using your browser's Back, Reload and Forward buttons. This can be useful when encountering an exception, when wishing to undo an undesired setting or mouse gesture, or when deciding to repeat the latter stages of an analysis with new parameter values without restarting the analysis from the first step. Regarding browser reliability, from our experience we have found that Firefox provides reliable session tracking.

If you encounter an *internal error*, such as that seen below in Figure 1, please report it to us as quickly as possible as it will help us to resolve the underlying problem promptly. When an internal error is logged on the webserver, a ticket is issued. When informing us of errors, it would be most helpful if you would cut and paste the ticket line and email it to us at threadmapper@gmail.com

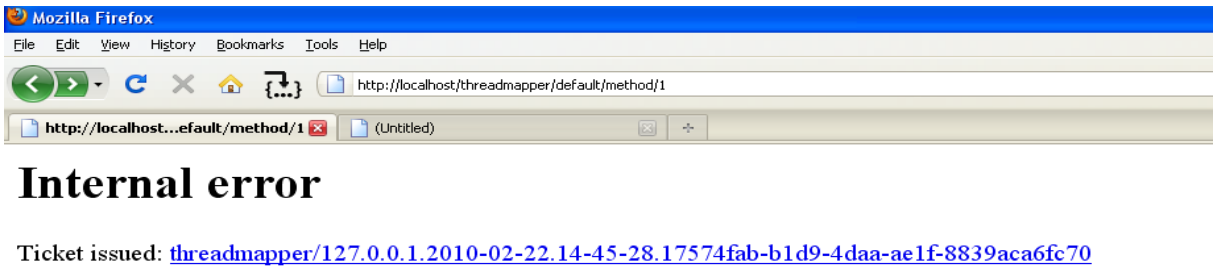


Figure 1: A THREaD Mapper Studio error and its associated ticket

We also recommend that users install their web browser's Adobe Flash Player plugin, in order to view the **Tutorial/Demo** screencasts mounted on the THREaD Mapper Studio home page. In the absence of such a plugin, users may encounter an error such as, "You either have JavaScript turned off or an old version of Macromedia's Flash Player. Get the latest Flash player." If you're receiving this error message when trying to play a tutorial video, it means that you *either* have **JavaScript turned off** in your browser, *or* you need to install **the latest version of the Adobe Flash Player**. While we will make every effort to keep our web application compatible with the various versions of Java available in common browsers, you might at times experience some difficulties. As before, we would welcome being informed of any such problems.

3) Input Genotype Files:

THREaD Mapper Studio version 1.0 currently accepts/supports four types of input file:

i) Locus file (.loc) format
ii) Raw file (.raw) format
iii) CSV/TSV format
iv) THREaD Mapper format v.1 (recommended)

i) Locus file (.loc) format:

This format has a compulsory **header part** followed by a series of *marker-genotype* lines with tab delimited marker names followed by their corresponding genotypes. Make sure the header includes the *nind* (number of individuals) and *nloc* (number of loci) lines, so that THREaD Mapper Studio can determine that this is a *loc formatted* file (see Figure 2 below for an example). Lines beginning with a semicolon (;) or exclamation mark (!) function as comment lines in these files.

sample.loc

```
name = sample_loc_file
popt = DH
nind = 12
nloc = 15

M1  A  A  A  A  -  A  A  A  A  A  A  A
M2  B  B  B  B  B  B  B  B  B  B  B  B
M3  A  A  B  B  B  B  B  B  B  -  B  B
M4  A  A  A  B  B  B  B  B  B  B  B  B
M5  A  A  B  B  B  A  A  A  B  B  B  B
M6  A  A  A  -  B  B  B  B  B  A  A  A
M7  A  A  -  A  A  A  A  A  A  B  B  B
M8  B  B  B  B  B  B  B  B  B  B  A  A
M9  A  A  A  B  B  A  A  B  B  A  A  A
M10 B  A  A  A  A  -  A  A  A  A  A  A
M11 B  B  B  A  A  A  A  A  A  A  A  A
M12 A  B  A  B  A  B  A  B  A  B  A  B
M13 A  B  B  B  B  B  B  B  B  B  A  A
M14 B  B  A  A  A  A  -  A  A  A  A  A
M15 B  A  B  A  B  A  B  -  -  A  B  A
```

Figure 2: A sample locus formatted (.loc) file with 12 individuals scored for 15 markers. This .loc formatted input file corresponds to the dataset in Figure 1 of Cheema and Dicks (2009) Briefings in Bioinformatics 10(6): 595-608.

ii) Raw file (.raw) format:

The .raw file format, as required by the MAPMAKER genetic mapping software, is a widely used format (see Figure 3 below for an example of this input file format). Briefly, the first line of a .raw data file should read something like:

data type poptype

where *poptype* is the appropriate experimental design, specified by symbols such as F2, RIL, DH, BC etc. The second line of the raw file should contain three numbers, separated by spaces, for example:

12 15 0

The first value stands for the number of individuals for which data are included in the file (in this case 12). The second value indicates the number of genetic loci or markers for which data are supplied (here 15). The third value indicates the number of quantitative traits in the data and is here set to zero. This second line also enables genotype symbol translations to be specified, as seen in the following example:

12 15 0 symbols 1=A 2=B 0=-

Here, the symbol **1** should be translated and interpreted as **A**, **2** as **B** and **0** as a *missing* genotype. After the first two lines, the .raw file should then present the locus data, in the following format. For each locus, the name of the marker preceded by an asterisk ("*"), should first be listed (NB: the asterisk symbol is compulsory and there must not be a space between the asterisk and the marker name), followed by one or more spaces or a tab, and then by the genotypic data for all individuals, in order, as in the following example:

*marker1 ABBA-ABA-BBB--BA

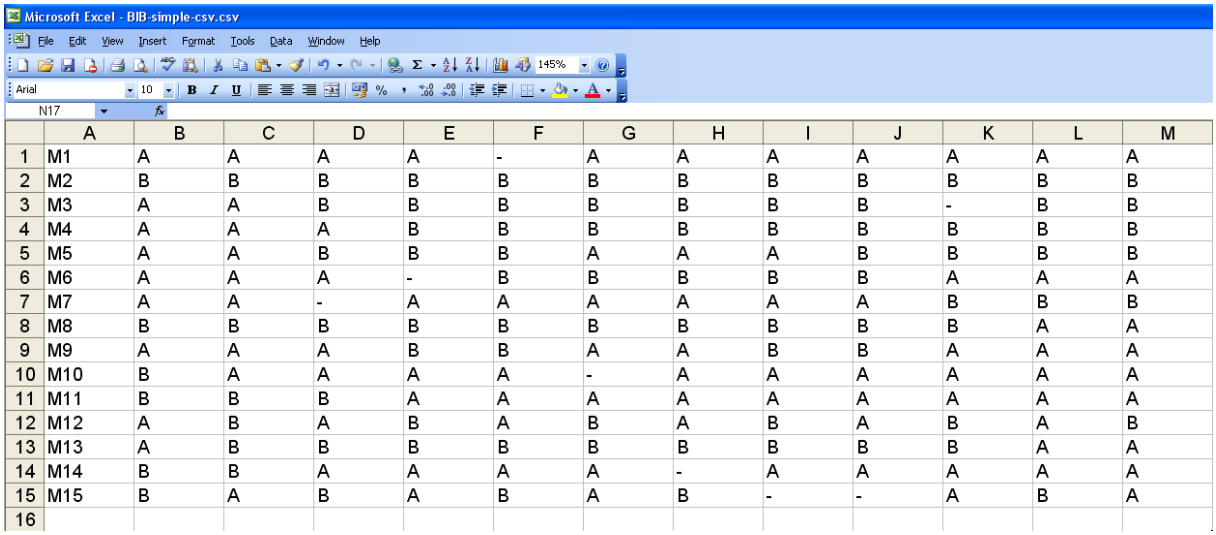
```
data type DH
12 15 0 symbols 1=A 2=B 0=-
*M1     111101111111
*M2     222222222222
*M3     112222222022
*M4     111222222222
*M5     112221112222
*M6     111022222111
*M7     110111111222
*M8     222222222211
*M9     111221122111
*M10    211110111111
*M11    222111111111
*M12    121212121212
*M13    122222222211
*M14    221111011111
*M15    212121200121
```

Figure 3: A sample raw formatted (.raw) file with a symbols directive in the second line. This .raw formatted input file again corresponds to the dataset in Figure 1 of Cheema and Dicks (2009) Briefings in Bioinformatics 10(6): 595-608.

Please note that THREaD Mapper Studio assumes that lines beginning with an asterisk are marker names. Therefore, any other information that begins with * should be removed from the file. Comment lines may be included by beginning them with the symbols "!", ";" or "#".

iii) CSV/TSV format:

One of the simplest input formats THREaD Mapper Studio accepts is a simple *comma separated value* (**CSV)** or *tab separated value* (**TSV**) file format (see Figure 4 for an example of a CSV file format). Both formats may be exported from Microsoft Excel spreadsheets. Such a file contains no header lines. On each row, the relevant marker name is followed by the corresponding genotypes, ordered according to the individuals (i.e. a column - after the first marker column - contains all the marker scores for a particular individual). If you have made such a file using a text editor, please make sure it loads correctly in Microsoft Excel and is properly delimited. Genotypes must be a single letter and belong to a set of valid symbols (see Section 4 below). Marker names should be one word and are not allowed to contain special character such as comma, tab, semi-colon, hash or exclamation mark. Marker names more than 30 characters long are not permitted.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	M1	A	A	A	A	-	A	A	A	A	A	A	A
2	M2	B	B	B	B	B	B	B	B	B	B	B	B
3	M3	A	A	B	B	B	B	B	B	B	-	B	B
4	M4	A	A	A	B	B	B	B	B	B	B	B	B
5	M5	A	A	B	B	B	A	A	A	B	B	B	B
6	M6	A	A	A	-	B	B	B	B	B	A	A	A
7	M7	A	A	-	A	A	A	A	A	A	B	B	B
8	M8	B	B	B	B	B	B	B	B	B	B	A	A
9	M9	A	A	A	B	B	A	A	B	B	A	A	A
10	M10	B	A	A	A	A	-	A	A	A	A	A	A
11	M11	B	B	B	A	A	A	A	A	A	A	A	A
12	M12	A	B	A	B	A	B	A	B	A	B	A	B
13	M13	A	B	B	B	B	B	B	B	B	B	A	A
14	M14	B	B	A	A	A	A	-	A	A	A	A	A
15	M15	B	A	B	A	B	A	B	-	-	A	B	A
16													

Figure 4: An example of a simple 15 × 12 CSV file, loaded in Microsoft Excel. Marker names are entered in the first column (A) and the individual genotypes from columns B to M. This .csv formatted input file again corresponds to the dataset in Figure 1 of Cheema and Dicks (2009) Briefings in Bioinformatics 10(6): 595-608.

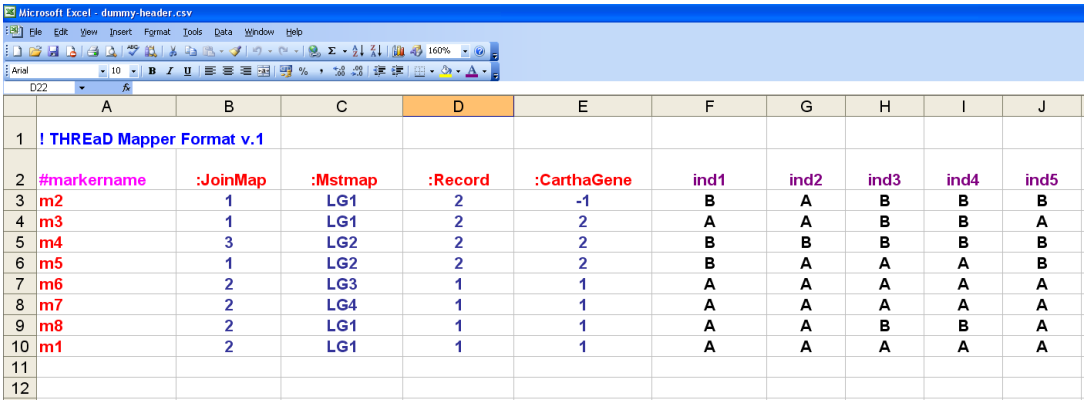
iv) THREaD Mapper format v.1:

This format (see Figures 5 and 6 below for examples) is similar to the simple **CSV/TSV** format described above but with additional support for named column attributes. Such support is achieved by using a *column header line* that provides information about the structure of the file and helps THREaD Mapper to locate the markers, their names and other user-specified attributes. Indeed a key feature of this file format is the ability to characterise markers according to shared attributes, known as "Attribute Groupings". Such a grouping could be based on any category but we have found it useful to denote a) marker type, b) physical map number (e.g. via BLASTing the DNA sequence of each marker against the genome sequence of the same or different species as that for which the genetic map is being developed), and c) genetic map number obtained from another genetic map estimation tool such as JoinMap (we recommend using two or more tools in the estimation of genetic maps as the comparison

highlights similarities and inconsistencies between the estimates that can be highly useful in developing a map).

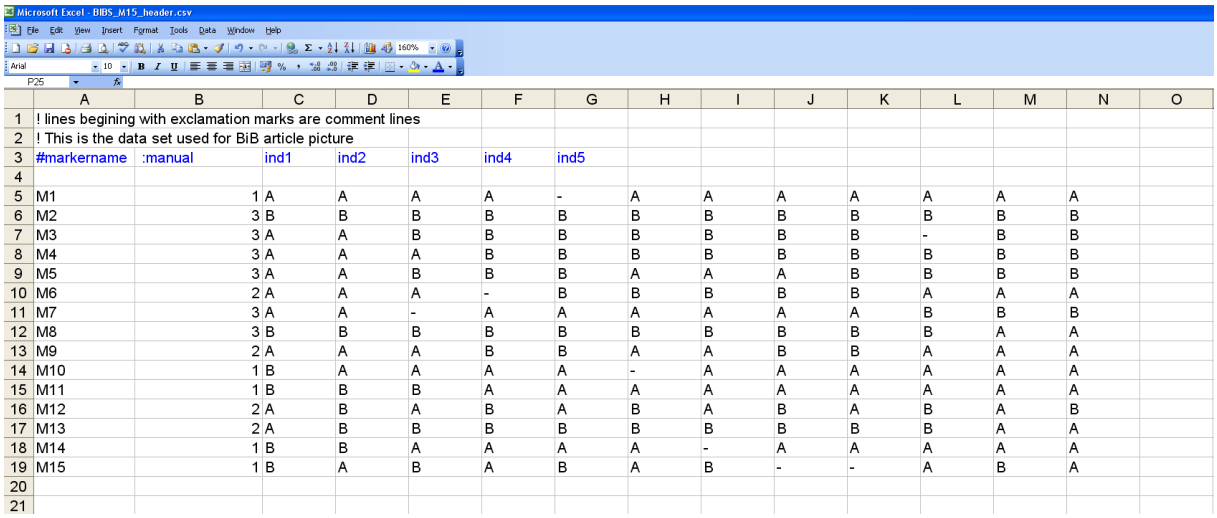
The first column of the input file is a compulsory column that consists of marker names (marker names should be one "word" that includes no special characters such as commas, colons and exclamation markers). The next series of columns are the optional attribute columns. Although there is no defined limit to the number of attribute columns that can be specified, we advise you to keep it below ten. Integer values for attributes are preferred as a convention but alternative text values such as LG1, LG2 etc. are also possible. We recommend using the symbols "0" or "-1" to denote an unknown attribute. Make sure that attributes, like marker names, do not include special characters. The attribute columns are followed by the individual genotype columns. Currently permitted genotype symbols are within the following set: {A, B, C, D, H, -, M}. See the next section on "Genotype Codes" for the recommended use of these symbols. In addition, the THREaD Mapper format v.1 input files may contain zero or more comment lines (rows) at the beginning of the file. All comment lines must begin with the reserved symbol "!".

The above column structure is referenced by a *column header line*. A line in a THREaD Mapper format v.1 file is interpreted as a *column header line* if it begins with the symbol "#" and contains strings beginning with ":". Please use the column header "**#markername**" for the single marker column and the column headers "**:<myattribute>**" (where <**myattribute**> is replaced by the name of a particular attribute) for each attribute column. Users may find it helpful to use text column headers to reference individual genotype columns, though these headers are not compulsory within THREaD Mapper format v.1 files and may be left blank.



	A	B	C	D	E	F	G	H	I	J
1	! THREaD Mapper Format v.1									
2	#markername	:JoinMap	:Mstmap	:Record	:CarthaGene	ind1	ind2	ind3	ind4	ind5
3	m2	1	LG1	2	-1	B	A	B	B	B
4	m3	1	LG1	2	2	A	A	B	B	A
5	m4	3	LG2	2	2	B	B	B	B	B
6	m5	1	LG2	2	2	B	A	A	A	B
7	m6	2	LG3	1	1	A	A	A	A	A
8	m7	2	LG4	1	1	A	A	A	A	A
9	m8	2	LG1	1	1	A	A	B	B	A
10	m1	2	LG1	1	1	A	A	A	A	A
11										
12										

Figure 5: A dummy dataset formatted in THREaD Mapper format v.1 and loaded in Microsoft Excel. The figure is coloured for the sake of clarity to show the various sections of the input file. Line 1 is a comment line, of which there may be zero or more. The first column, denoted by the reserved column header '#markername' (shown here in pink), has the names of markers from m2 to m1. The file has four attribute columns, with the corresponding column headers each beginning with the reserved symbol ":". The four columns represent the linkage groupings achieved using the JoinMap, MSTMap, Record and CarthaGene software tools respectively. The five individuals are referenced by the column headers ind1, ind2, ind3, ind4 and ind5 (show here in violet), though these headers are optional.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	! lines begining with exclamation marks are comment lines														
2	! This is the data set used for BiB article picture														
3	#markername	:manual	ind1	ind2	ind3	ind4	ind5								
4															
5	M1	1	A	A	A	A	-	A	A	A	A	A	A	A	
6	M2	3	B	B	B	B	B	B	B	B	B	B	B	B	
7	M3	3	A	A	B	B	B	B	B	B	B	-	B	B	
8	M4	3	A	A	A	B	B	B	B	B	B	B	B	B	
9	M5	3	A	A	B	B	B	A	A	A	B	B	B	B	
10	M6	2	A	A	A	-	B	B	B	B	B	A	A	A	
11	M7	3	A	A	-	A	A	A	A	A	A	B	B	B	
12	M8	3	B	B	B	B	B	B	B	B	B	B	A	A	
13	M9	2	A	A	A	B	B	A	A	B	B	A	A	A	
14	M10	1	B	A	A	A	A	-	A	A	A	A	A	A	
15	M11	1	B	B	B	A	A	A	A	A	A	A	A	A	
16	M12	2	A	B	A	B	A	B	A	B	A	B	A	B	
17	M13	2	A	B	B	B	B	B	B	B	B	B	A	A	
18	M14	1	B	B	A	A	A	A	-	A	A	A	A	A	
19	M15	1	B	A	B	A	B	A	B	-	-	A	B	A	
20															
21															

Figure 6: The THREaD Mapper format v.1 input file for the dataset in Figure 1 of Cheema and Dicks (2009) Briefings in Bioinformatics 10(6): 595-608 (complete with the linkage group designations of the 15 markers in the second column, an attribute column).

4) Genotype Codes:

The recommended use of the genotype symbols is related to the experimental design of the mapping experiment. The current version of THREaD Mapper Studio supports the following four population types: Doubled Haploid (DH), Backcross (BC), F2, and RIL (Recombinant Inbred Line).

(i) For DH we recommend the use of symbols {A, B, -}:

'A' : (homozygote for parental genotype A)
'B' : (homozygote for parental genotype B)
'-' : (missing data)

(ii) For BC we recommend the use of symbols {A, H, -}

'A' : (homozygote for parental genotype A)
'H' : (heterozygote carrying both alleles A and B)
'-' : (missing data)

(iii) For F2 we recommend the use of symbols {A, B, H, C, D, -}

'A' : (homozygote for parental genotype A)
'B' : (homozygote for parental genotype B)
'H' : (heterozygote carrying both alleles A and B)
'C' : (not a homozygote for allele A, either BB or AB genotype)
'D' : (not a homozygote for allele B, either AA or AB genotype)
'-' : (missing data)

(iv) For RIL we recommend the use of symbols {A, B, -} Note: We currently support advanced RILs only.

'A' : (homozygote for parental genotype A)
'B' : (homozygote for parental genotype B)
'H' : (heterozygote carrying both alleles A and B)
'-' : (missing data)

5) Loading an Input Genotype File into THREaD Mapper Studio:

Figure 7 below shows the upload form which is used to load a new Genotype File onto the THREaD Mapper Studio webserver for analysis. Users must provide four items in order to load a file successfully. Firstly, they must specify a location on their computer for their

Genotype File. Secondly, they must enter an item called *Title*, which accepts a short description of the analysis/file. The title provided must be unique, or an error message such as "value already in database" or "Please enter a title" will appear.

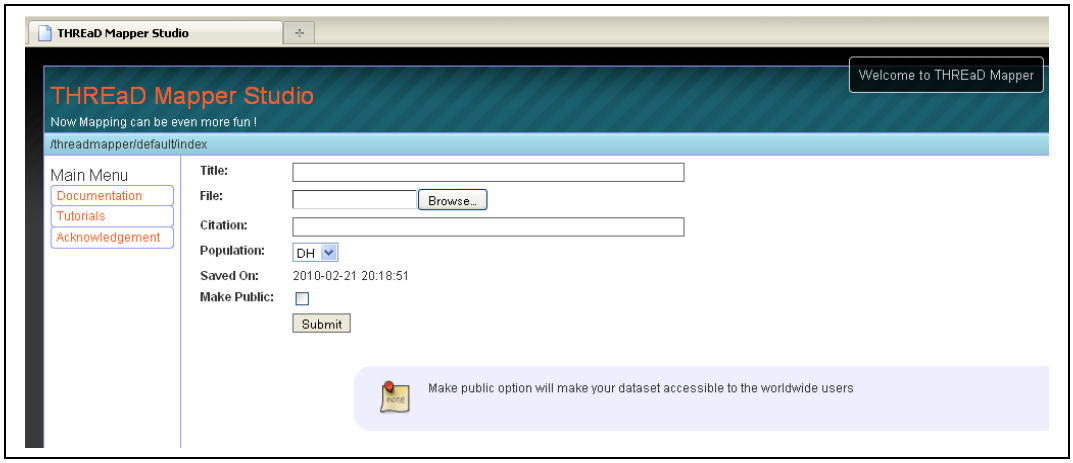


Figure 7: A screenshot of the Genotype File submission form on the THREaD Mapper Studio homepage.

Thirdly, users should provide a Citation or reference for the data set. Figure 8 below shows a typical citation associated to a deposited dataset. If a user chooses to make their dataset available to all THREaD Mapper Studio users, by checking the "Make Public" tickbox seen above in Figure 7, the citation will help users to learn more about the data set. In such a case, the dataset will be appended to the section "List of datasets deposited on the server" to be found on the THREaD Mapper Studio homepage and may then be selected for further analysis by means of single click. **Please note:** if you click the checkbox in order to make a dataset public, you are agreeing that the dataset may be seen by users worldwide. Finally, a user must specify the experimental design from which the dataset is derived. As noted above, the current options are BC, DH, F2 and RIL.

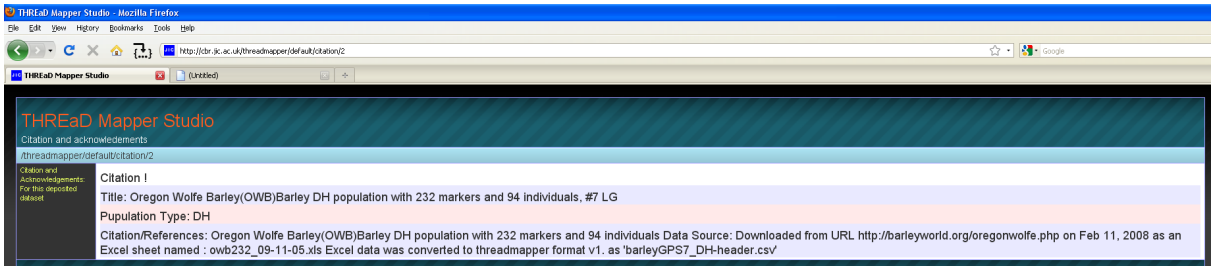


Figure 8: A citation added to an inputted barley genotype input file.

Once the title, citation, population type and file location are entered into the upload form, clicking on the *Submit* button will cause a basic sanity check to be performed on the Input Genotype File. If any of the input form fields are not filled in correctly, a user will see validation errors next to the unfilled or non-validated fields (red coloured boxes, as seen in Figure 9 below). The sanity check also involves making sure the size of the file is less than 1MB, that it is formatted according to one of the four permitted formats (also end of line characters should be intact depending if saved in Windows/Mac/Linux), that the number of columns is consistent for each row and that the characters within the Genotype columns are valid symbols. If an input file does not pass the sanity check, an error message is displayed

along with the first line number of the file at which an error occurs. Please make sure that all marker names are unique. If they are not, you may encounter an error such as "*Duplicate marker names not allowed: found 3 markers with same name ['w456']*". This would simply mean that marker name w456 appeared three times in the input locus file.

Figure 9: An input file form with upload errors denoted by red boxes

An input file which passes the sanity check is uploaded onto the webserver. If the Make Public tickbox has been checked, it is appended to the List of datasets deposited on the server, as shown in Figure 10 below. The item on this list named "BiB data set with Header Attributes" is the dataset shown above in Figure 6. Any file on this list may be selected for further analysis by simply clicking on "Select" to the right of the file description. A "Download" link is also available for each file, so that it may be downloaded and its format examined in detail. Furthermore, a "Citation" link provides further information about a relevant data set and a time stamp and population type provide information about the time at which the dataset was uploaded onto the server and the experimental design under which the dataset was generated, respectively.

Figure 10: Screenshot of the List of datasets item on the THREaD Mapper Studio homepage

6) Basic statistics for a Genotype File:

Once a Genotype File has been loaded successfully onto the THREaD Mapper Studio webserver, the user is presented with a page showing basic statistics for the file (see Figure 11). The ***Left side panel*** contains a spreadsheet-like display containing **Genotype primary statistics**. At the top of the panel, the numbers of markers and individuals within the dataset

are displayed, along with three option buttons ("Add a Marker", "Genotype" and "Remove Marker").

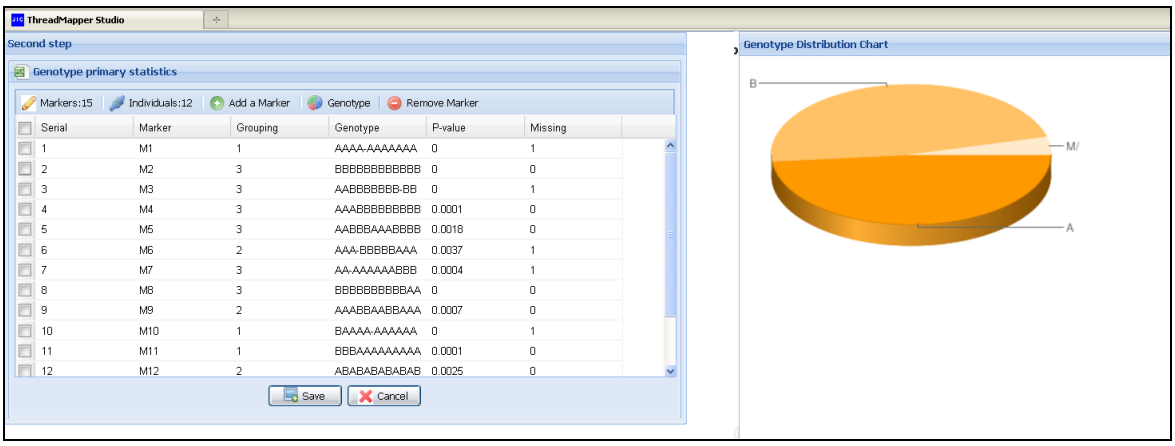


Figure 11: Screenshot of a typical **Genotype primary statistics** panel

Note that the number of markers shown in the ***Left Side Panel*** represents only the unique markers found within the Input Genotype File. By unique markers we mean that if three markers have same score for all the individuals we treat them as a single marker. For example, imagine markers ***m1a***, ***m1b***, ***m1c*** all have an identical associated raw genotype score, say BBBBB. Only one (the first in the list) is retained and the others are deleted. In this example, **m1a** would be retained and **m1b** and **m1c** would be deleted, as shown in Figure 12 below. In particular note the ***Duplicate Marker Panel*** at the bottom of the webpage, which indicates the identities of duplicate markers deleted by this requirement.

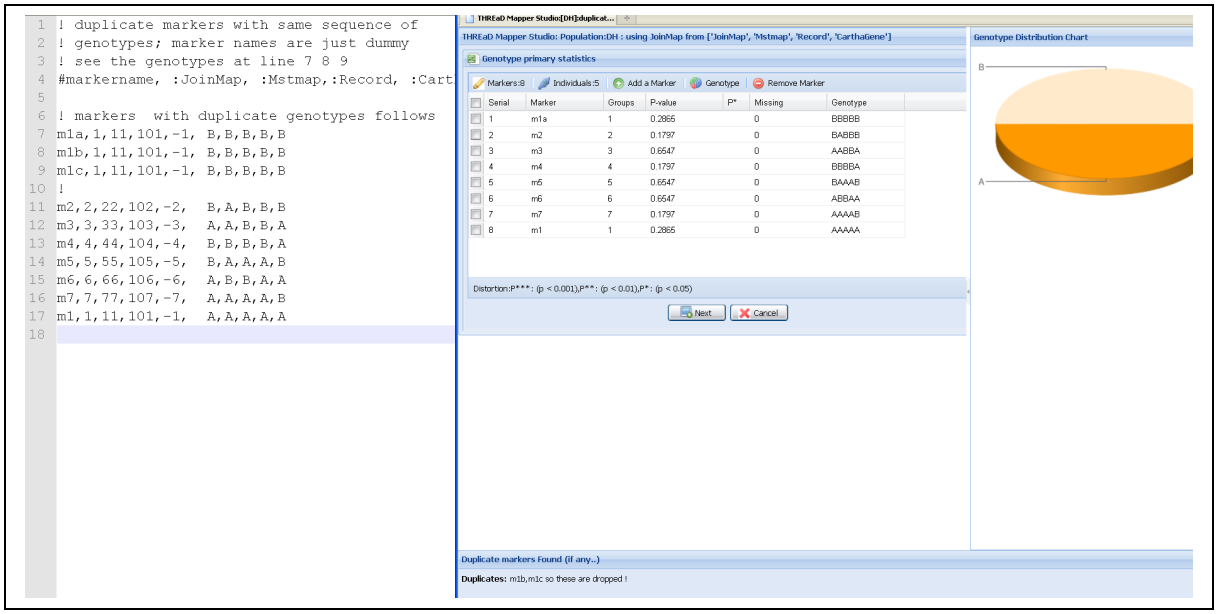


Figure 12: A THREaD Mapper formatted v.1 input file successfully loaded onto the webserver exhibits three duplicate genotype signatures (left). The duplicate markers detected are dropped from the analysis (right).

The *Add a Marker* option button is disabled in the current version of THREaD Mapper Studio. The *Genotype* option button causes a pop-up window to display a **Genotype**

Distribution Chart, as described below. The *Remove Marker* option button removes from the spreadsheet any markers that have been selected using the tick boxes to the left of the marker name. The spreadsheet display itself contains columns for marker selection, marker name, attribute grouping (as described above), marker scores, P-value for segregation distortion, and number of missing marker scores. Each of the columns may be sorted according to the values contained within it, by clicking on the corresponding column title. This can be useful, for example, for excluding markers that show strong segregation distortion or contain a high number of missing scores. Simply click on the column of interest, select markers with undesirable properties by clicking on the corresponding checkboxes, and remove them from the analysis by choosing the *Remove Marker* option button. Note that the *Remove Marker* option can remove several markers at once.

The ***Right side panel*** contains a **Genotype Distribution Chart** for the dataset. This displays a Pie Chart for the genotype frequencies found within the dataset. Once the user is content to continue the analysis, clicking on the *Save* button at the bottom of the **Genotype primary statistics** panel will take the analysis to the next step, that of choosing a scoring scheme for the analysis.

7) **Selecting a Scoring scheme for a Genotype File:**
The analysis carried out by the THREaD Mapper algorithms requires "distances" to be calculated between all pairs of markers. These distances are calculated according to scoring schemes. These scoring schemes are available depending upon the population type/ experimental cross in question. In the current version of THREaD Mapper Studio, we allow several scoring schemes for each of the following population types: DH, BC, F2 and RIL.

i) DH population type
For an analysis of a DH dataset, seven scoring schemes are available, as shown in Figure 13: Hamming, Expected number of crossovers/meiosis, r-hat, P-val, Haldane distance, Kosambi distance and Carter and Falconer generic distance.

a) Hamming scoring scheme
Hamming, the simplest of the scoring schemes, considers the level of similarity/difference between two marker scores. Imagine we have two markers each scored in 12 individuals:

```
M1='AAAABBAA-A--'
M2='AB-ABAAA-AB-'
```

The Hamming distance between a pair of markers is simply the proportion of individuals whose scores differ between them, here coloured in red (i.e. **4/12=0.333** in this example).

b) Expected number of crossovers/meiosis scoring scheme
This scoring scheme is the number of scores N multiplied by the observed proportion of recombinants (ignoring individuals with one or more missing scores). In the example above this distance is **12 × (2/8) = 3**. Note that this distance is the r-hat distance below multiplied by N.

c) r-hat(DH) scoring scheme
This scoring scheme is related to Hamming but like the distance in b) above only considers the differences between two marker scores where neither of the scores are missing values (i.e. only those scores coloured magenta below). Here, this is simply **2/8 =0.25.**

```
M1='AAAABBAA-A--'
M2='AB-ABAAA-AB-'
```

d) P-val scoring scheme

This scoring scheme is calculated using the P-values associated with the Chi-square for segregation distortion, as seen in the **Genotype Distribution Chart** seen in section 6 above. The distance is defined as the absolute difference between the P-values of the two markers.

e) Haldane scoring scheme

The Haldane centiMorgan distance *d* is calculated according to the widely used Haldane mapping function and is related to **r-hat(DH)** as follows:

$$d = -50\ln(1-2\hat{r})$$

f) Kosambi scoring scheme

The Kosambi centiMorgan distance *d* is calculated according to the widely used Kosambi mapping function and is related to **r-hat(DH)** as follows:

$$d = 25\ln\left(\frac{1+2\hat{r}}{1-2\hat{r}}\right)$$

g) Carter and Falconer scoring scheme

Carter and Falconer (1951) developed a family of mapping functions, with variable levels of chiasma interference accounted for by a parameter γ (where $0< \gamma <1.0$). The Haldane and Kosambi mapping functions are special cases of this family with $\gamma = 0$ and $\gamma = 0.5$ respectively. This generic mapping function is related to **r-hat(DH)** as follows:

$$d = \int_0^{\hat{r}} \frac{1}{1-(2\hat{r})^{\alpha}} \qquad \text{where } \alpha = \left(\frac{1}{1-\gamma}\right)$$

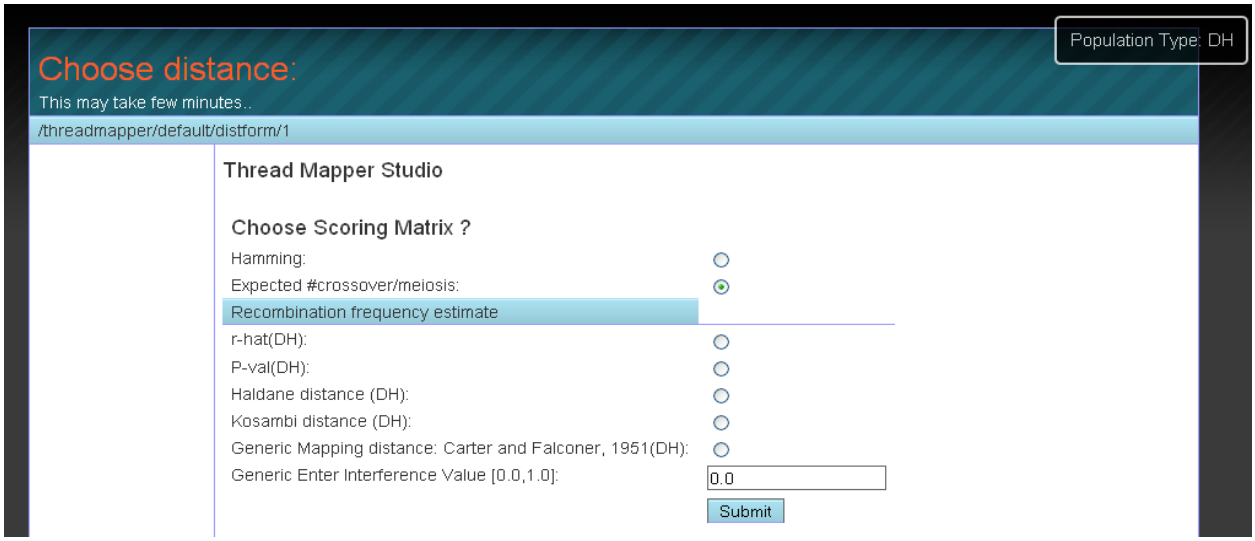


Figure 13: Screenshot of the scoring scheme options available for DH populations.

ii) BC population type

For an analysis of a BC dataset, seven scoring schemes are available as for DH datasets, as shown in Figure 14: Hamming, Expected number of crossovers/meiosis, r-hat, P-val, Haldane distance, Kosambi distance and Carter and Falconer generic distance. All calculations proceed in a similar fashion to those for the DH population, with the DH genotype symbol B substituted by the BC genotype symbol H.

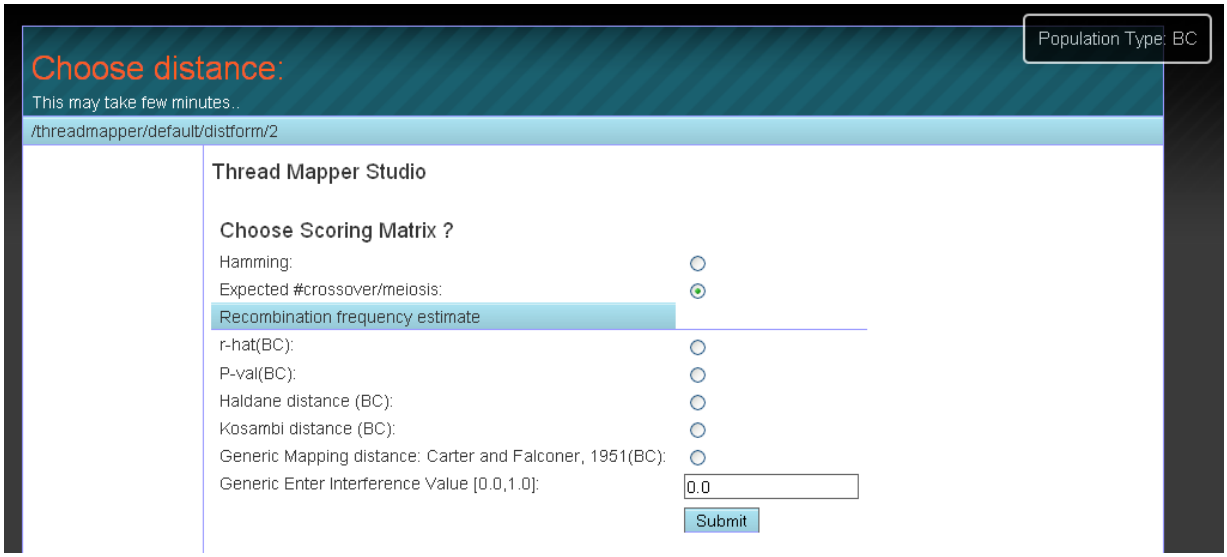


Figure 14: Screenshot of the scoring scheme options available for BC populations.

iii) F2 population type

For an analysis of an F2 dataset, nine scoring schemes are available, as shown in Figure 15: Hamming, PIP, PnP, Expected number of crossovers/meiosis, r-hat, P-val, Haldane distance, Kosambi distance and Carter and Falconer generic distance.

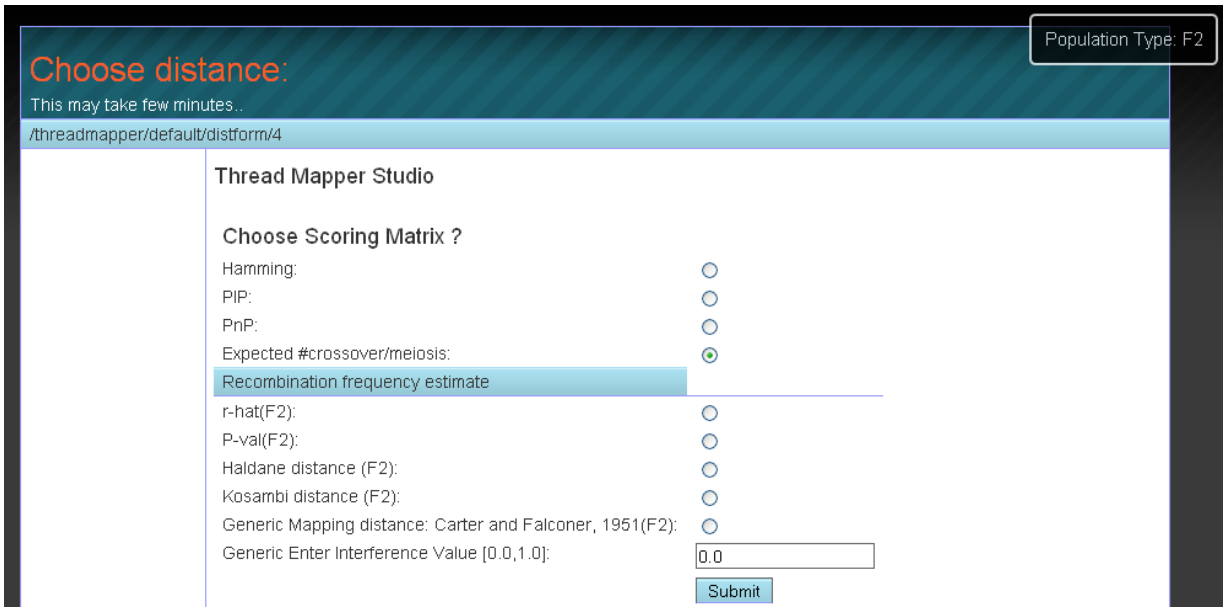


Figure 15: Screenshot of the scoring scheme options available for F2 populations.

Pairwise marker distances under the seven scoring schemes shared between F2 and DH are calculated in a similar fashion to the DH population type. The main difference comes in the calculation of r-hat. In THREaD Mapper Studio we follow the methodology used within the AntMap genetic mapping tool (http://cse.naro.affrc.go.jp/iwatah/antmap/). We begin by estimating whether each marker is more likely to be co-dominant or dominant. We divide the permitted F2 genotype symbols {A, B, H, C, D, -} into three overlapping sets: CODOM = {A, B, H, -}, DOMINANT Parent 1 = {B, D, -} and DOMINANT Parent 2 = {A, C, -}. If, for example, marker M1 has been scored as ABHBBBBABABH-HH within a set of 15 individuals then it is predicted to be a co-dominant marker as its scores are only consistent with the CODOM symbol set. Next, we invoke different sets of equations, based on whether we are comparing similar or different types of markers and whether the scores are in coupling or repulsion phase, to find the expected recombination fraction. We have adapted the AntMap equations (written in the Java programming language) to make new Python routines for THREaD Mapper Studio. These equations estimate the recombination fraction using a bisection search, varying r within the range [0.0, 1.0] to find the root of the equation.

Two scoring schemes, PIP and PnP, are used only for the F2 and RIL population types. They were added by us and we have yet to verify their relationships to the other scoring schemes. They are probabilistic in nature, essentially calculating the probability that two genotype scores are different depending on whether we account for the fact we do not know the phase of our datasets (PIP – Probabilistic including Phase) or whether we ignore phase (PnP – Probabilistic no Phase). See Figure 16 below for the PIP and PnP distance matrices.

a)

PIP				
	A	**B**	**H**	**-**
A	**0**	**1**	**0.5**	**0.5**
B	**1**	**0**	**0.5**	**0.5**
H	**0.5**	**0.5**	**0.5**	**0.5**
-	**0.5**	**0.5**	**0.5**	**0.5**

b)

PnP				
	A	**B**	**H**	**-**
A	**0**	**1**	**0.5**	**0.5**
B	**1**	**0**	**0.5**	**0.5**
H	**0.5**	**0.5**	**0**	**0.33**
-	**0.5**	**0.5**	**0.33**	**4/9**

Figure 16: The THREaD Mapper Studio PIP and PnP distance matrices for F2 and RIL population types

iv) RIL population type

For an analysis of a RIL dataset, nine scoring schemes are available, as shown in Figure 17: Hamming, PIP, PnP, Expected number of crossovers/meiosis, r-hat, P-val, Haldane distance, Kosambi distance and Carter and Falconer generic distance.

Again, the main difference in distance calculations with the previous population types comes in calculating r-hat (RIL). Here, the following formula is used:

$$\hat{r} = \left(\frac{R}{2(1 - R)} \right)$$

where R is the proportion of recombinant scores (ignoring individuals with one or more missing scores). In our previous example:

```
M1='AAAABBAA-A--'
M2='AB-ABAAA-AB-'
```

R = 2/8 = 0.25 and $\hat{r}$ =0.167

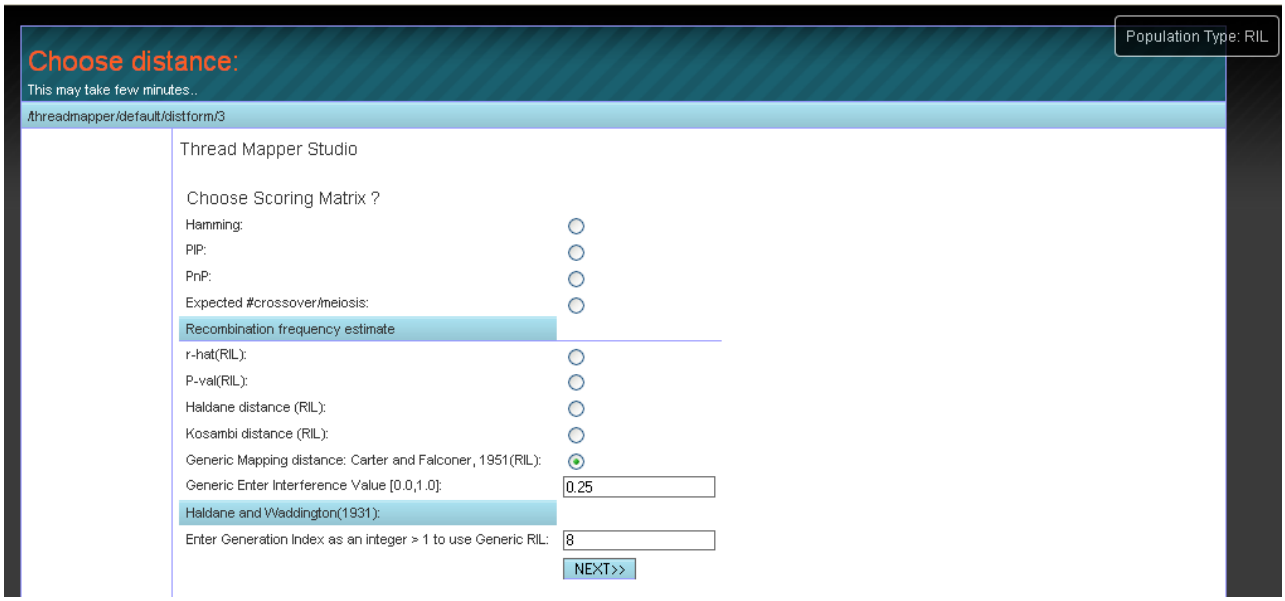


Figure 17: Typical scoring scheme selection menu for an eighth generation RIL population (RIL-8) where a recombination fraction is calculated according to the Haldane and Waddington (1931) equations. The resulting estimate is then used as input to the Carter and Falconer (1951) generic mapping function (here with a user defined degree of chiasma interference $\gamma = 0.25$), producing pairwise marker distances in the centiMorgan scale.

Additionally for RILs, we also provide the Haldane and Waddington (1931) method for estimation of recombination fractions for a RIL population at a given generation index. By generation index we mean F1, F2, F3,..., Fn. So for an F8 population, its generation index is 8 which should be entered as an integer into the Generation Index box as shown in Figure 17 above. This method by default is disabled using a Generation Index of -1. In implementing

this method, we have used the Haldane and Waddington equations, as shown in the scheme provided within Supplementary Text S1 of Wu et al. (PLOS Genet 4(10) 2008), and we have adapted the corresponding C++ class from the MSTMap genetic mapping tool (http://138.23.191.145/mstmap/) into Python code within THREaD Mapper Studio. This code uses a simple approach to estimating the recombination fraction, breaking the interval [0,0.5] into small subintervals and deploying an objective function so as to minimize the sum of square errors between the expected (using Haldane and Waddington's equations) and the observed fractions. The resulting recombination fractions may then be used as input to the Carter and Falconer generic mapping function instead of the usual r-hat values.

v) Choosing a scoring scheme

In Figures 13, 14, 15 and 17 above we have seen THREaD Mapper Studio menus for population types DH, BC, F2 and RIL respectively that enable users to choose a scoring scheme for their dataset. Each page has a default scheme already selected. However, users can choose a different scheme if they so wish by clicking on the relevant radio button (and possibly entering one or more numeric values into appropriate fields). Once a scoring scheme has been selected, clicking on the *Submit* button at the bottom of the page takes the user to the next step, displaying a heatmap for the inter-marker distance matrix.

8) Displaying a heatmap:

Using the scoring schemes described in the previous section, distances between the genotypes of each pair of markers are calculated, resulting in a symmetrical $m \times m$ matrix of inter-marker distances. In this step, this matrix is displayed as a heatmap where each distance is represented by a colour, with large distances tending towards red and small distances tending towards blue. So, for a dataset with markers ordered according to map order we should see a strong blue diagonal from top left to bottom right. Such a figure may be useful for more experienced users in interpreting the linkage group structure within a dataset. See Figure 18 below for a heatmap of the BiB sample dataset in Figure 6, scored according to the Hamming scoring scheme.

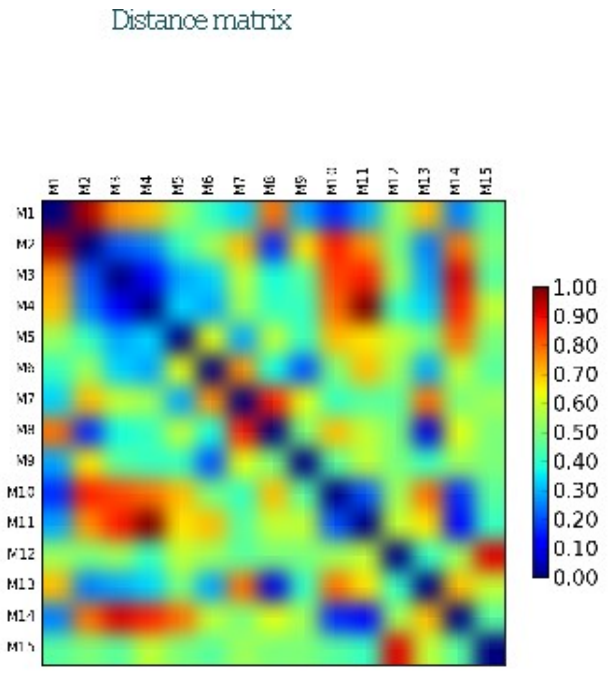


Figure 18: A heatmap of the BiB sample dataset, using the Hamming scoring matrix.

In addition, if you have already analysed your dataset with another genetic mapping tool it may be useful to order your markers according to those results, to see if the linkage group structure appears consistent with that ordering. See Figure 19 for an example of this type of analysis. Here, we see the strong blue diagonal noted above.

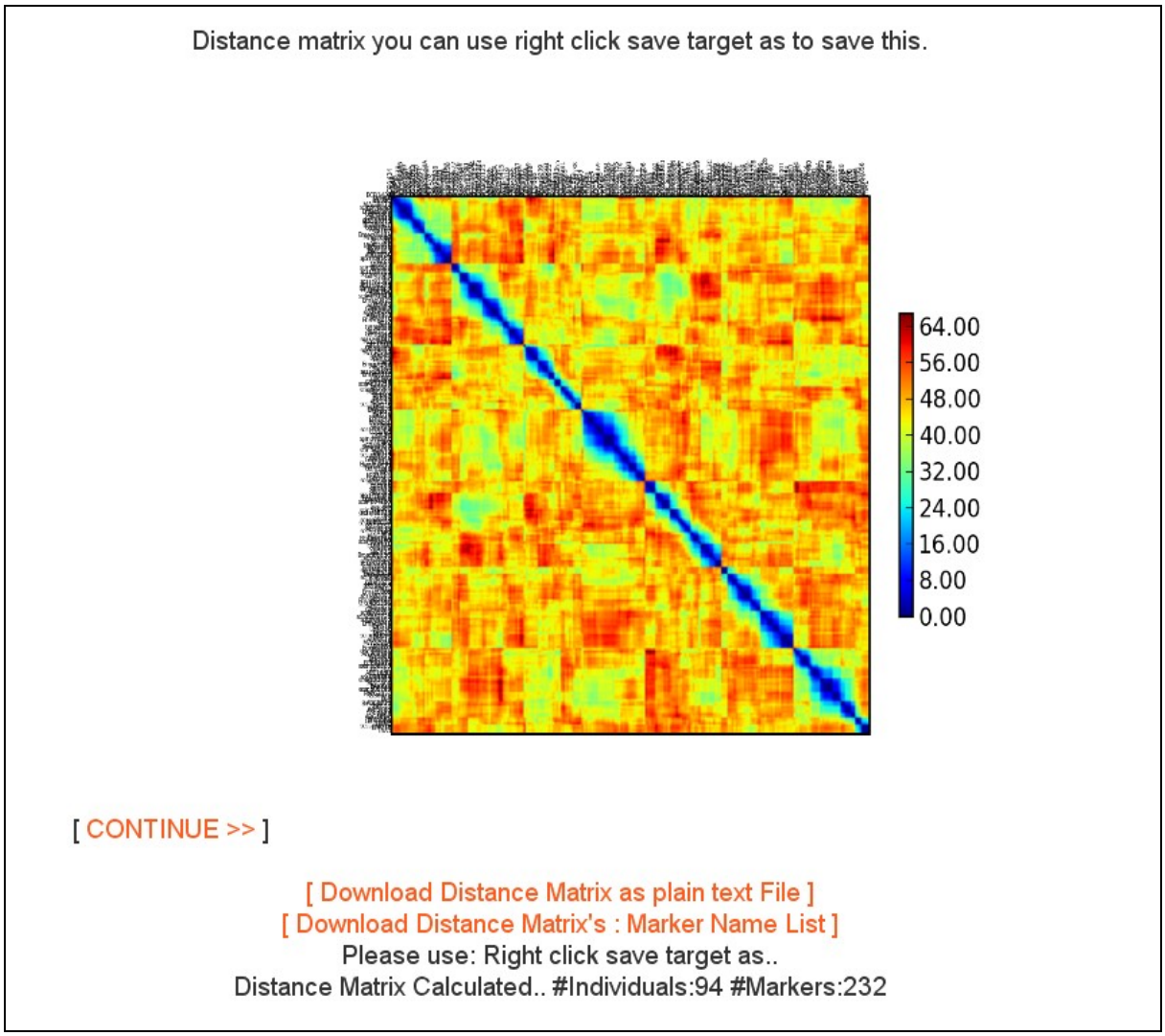


Figure 19: Markers in the seven barley linkage groups as predicted by the MSTMap genetic mapping tool are ordered according to the results of this earlier analysis. The resulting heatmap plot is coloured using Expected number of crossovers per meiosis units.

Once a heatmap has been viewed by the user, clicking on the *CONTINUE >>* button will move the analysis on to the next step, choosing parameters for linkage group and ensemble analysis. The heatmap webpage also contains clickable buttons that allow users to download the distance matrix and the list of markers for their dataset, both in text file format.

9) Linkage group and Ensemble parameter choice

Within this step users are asked to specify whether their dataset consists of a single or multiple linkage groups. This choice is made by clicking on the relevant radio button. Users who are not sure whether their dataset consists of one or more linkage groups should choose multiple linkage groups. Users are also asked whether they wish to generate an Ensemble for the remainder of the analysis (see Figure 20 below for a screenshot of this webpage). The

ensemble procedure perturbs the dataset in a carefully controlled manner to see how robust the dataset (and the inferences made from it) is to local error. Later in a THREaD Mapper analysis, markers in a dataset are connected by a Minimum Spanning Tree (MST) in three dimensional space. Though the MST captures the general connectivity of a data set, it is a sparse representation of a dataset and as a result is sensitive to noise. The presence of local noise may cause changes to the MST structure and the Ensemble procedure attempts to see if this is a serious issue for a particular dataset. In such a case, an Ensemble will be easily seen as a graphical structure that encapsulates alternative MSTs, particularly at the local level (i.e. there may be more than one possibility of the order of connectivity in certain regions of the graph).

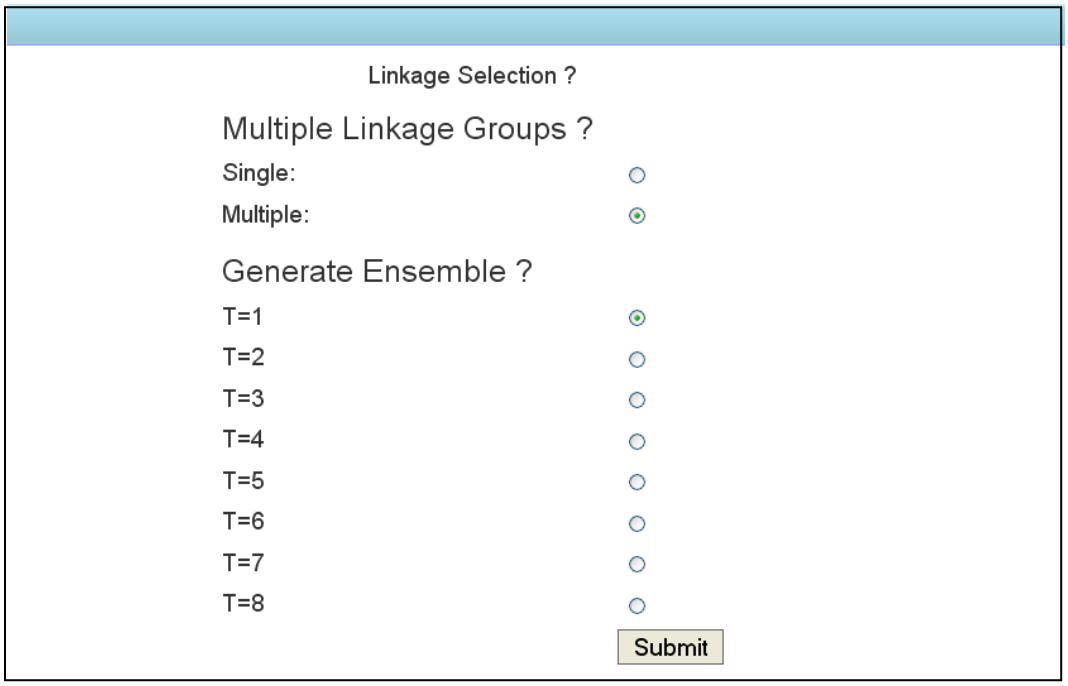


Figure 20: The linkage group and Ensemble parameter choice webpage

A growing value of the ensemble parameter T corresponds to a stronger perturbation. Users not wishing to carry out an ensemble analysis can simply choose the default value of T=1 (no embedding). We advise users to choose T=1 for at least the first pass analysis of their dataset. The analysis may then be repeated, in order to check the robustness of the data set to noise, choosing progressively higher values of T on each pass (e.g. use your web browser's Back button for this procedure). We have found that T = 2 to 4 will usually provide a good indication of the strength and effects of noise. For more details on the Ensemble procedure, please refer to the article by Carreira-Perpiñán and Zemel (2005). Once a user has chosen their preferred parameter values, clicking on the Submit button will take them to the next step in the analysis, Thresholding.

10) Thresholding: automatic or user-led linkage group clustering

Thresholding is an optional analytical step enabling a user to use computational methods, or alternatively their own opinion, to cluster their markers into distinct putative linkage groups prior to further analysis (NB: These linkage groups will not be fixed but may be altered at a later stage in the analysis). Users are presented with an interactive plot which is used to inform the clustering procedure. Earlier we mentioned that all markers in the dataset could be connected using an MST graphical structure. The Y-axis of the plot shows the inter-marker

distance according to the chosen scoring scheme (see Section 7 above) of an edge of this MST (T=1) or alternatively the MST ensemble (T>1). On the X-axis of the plot, the edges in the MST are sorted according to their length, with shortest edges to the left of the plot and longest edges to the right. In Figure 21 below, we see such a plot for a dataset consisting of 232 markers from 7 linkage groups. The MST connecting these markers consists of 231 edges and we see the various sizes of these edges within this plot.

The basic idea of Thresholding is that this plot captures the inherent cluster structure in the dataset. Where we see an obvious jump in the distance values (i.e. between edges 224 and 225 in Figure 21, towards the righthand side of the plot, corresponding to a jump in edge distance from 23.25 to 27.89), this may correspond to a move from intra-linkage group distances to inter-linkage group distances in the MST. The horizontal red line on the plot corresponds to the Y-value at which to separate marker pairs, so all inter-marker edges within the MST that are larger than this value will be severed, resulting in a number of distinct graphs. If the threshold value is very low, we may produce a completely unconnected graph and if the value is very high, we may end up with too many linkage groups. Consequently, we need good ways of choosing the Thresholding value.

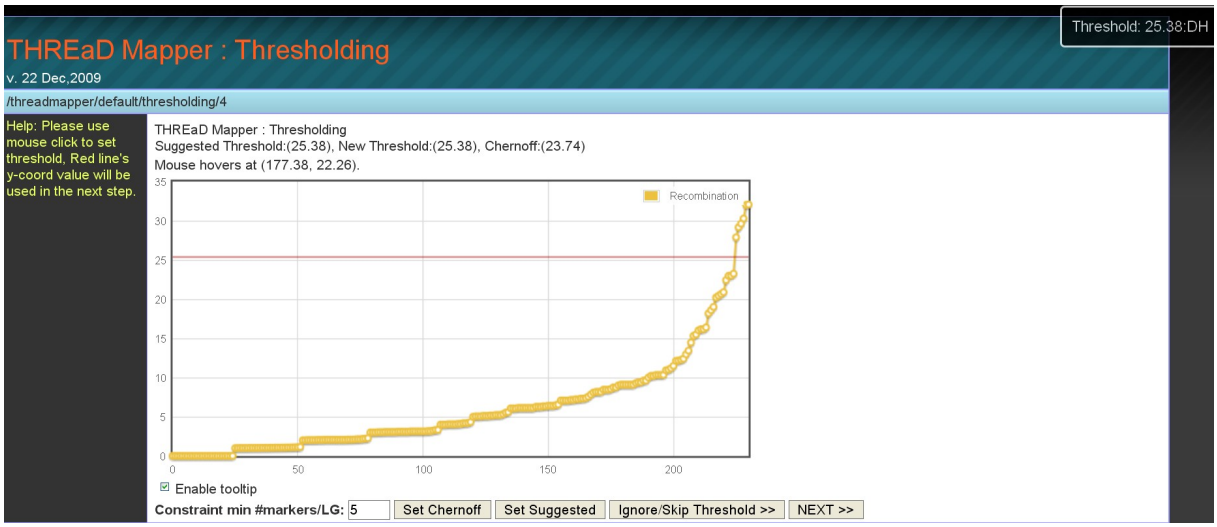


Figure 21: The Thresholding webpage, showing a plot of a dataset consisting of 232 markers in 7 linkage groups

There are three ways to select this Y-value within THREaD Mapper Studio. Firstly, users may choose the computational Chernoff Bound value by clicking on the *Set Chernoff* button (please note this is available for DH and BC population types only). The Chernoff Bound was first used for this purpose within the MSTMap genetic mapping software and is theoretically based. Alternatively, users may choose an empirically based value by clicking on the *Set Suggested* button. This method uses a simple heuristic. If the chosen scoring scheme is the expected number of crossover/meiosis, the threshold is set at 0.27 times the number of individuals (with the added constraint of five markers per linkage group). For other scoring schemes, the threshold is set corresponding to the largest edge size jump between the 75th and 85th percentiles of the edge sizes, again constrained by at least five markers per linkage group. We have also tried various graph clustering and partitioning algorithms for Y-value choice in addition to the community detection algorithm. However, due to execution time constraints they are currently not included with this version of the THREaD Mapper Studio webserver.

We may release a version including these options as a separate command line tool in the future.

Users may also choose to set their own Thresholding value. This can be done by clicking inside the plot at the preferred Y-value. In practice, we have found visual cut-off choice suitable for many datasets and users are encouraged to use their visual intuition. Additionally, users can constrain the Thresholding rule so that severing of edges larger than the chosen Y-value are not made if they result in a linkage group with fewer than a fixed number of markers. This is achieved by entering the chosen minimum number into the "Constraint min #markers/LG" box. Once Thresholding option has been chosen, clicking on the *Next* >> button moves the user onto the next step, that of viewing the marker membership of the distinct putative linkage groups.

Users may also wish to skip the Thresholding step by clicking on the *Ignore/Skip Threshold* >> button (which also moves the user onto the next step). We recommend choosing this option on the first pass analysis of a new dataset, followed by a second pass Thresholding analysis (using the Back button of the web browser). The Ignore/Skip option is also appropriate when working with a single linkage group.

11) Viewing the linkage group marker membership

Once the Thresholding analysis is complete, users are presented with the marker membership of the current linkage group clustering, in a simple spreadsheet-like format as shown in Figure 22 below. If the Thresholding step had been skipped (perhaps because the dataset consisted of a single linkage group) all markers will belong to a single linkage group. Each linkage group is numbered and the number of markers within it is indicated. The plus/expand symbol to the left of the Group tag can be clicked to view the names of the constituent markers, along with their linkage group number and their individual scores. Within a linkage group, markers may be sorted by name. A link at the bottom of the webpage is provided to enable users to download the current linkage group clustering, in a CSV file format. For some web browsers, users may need to use the save target option (using a right mouse click) to download the file.

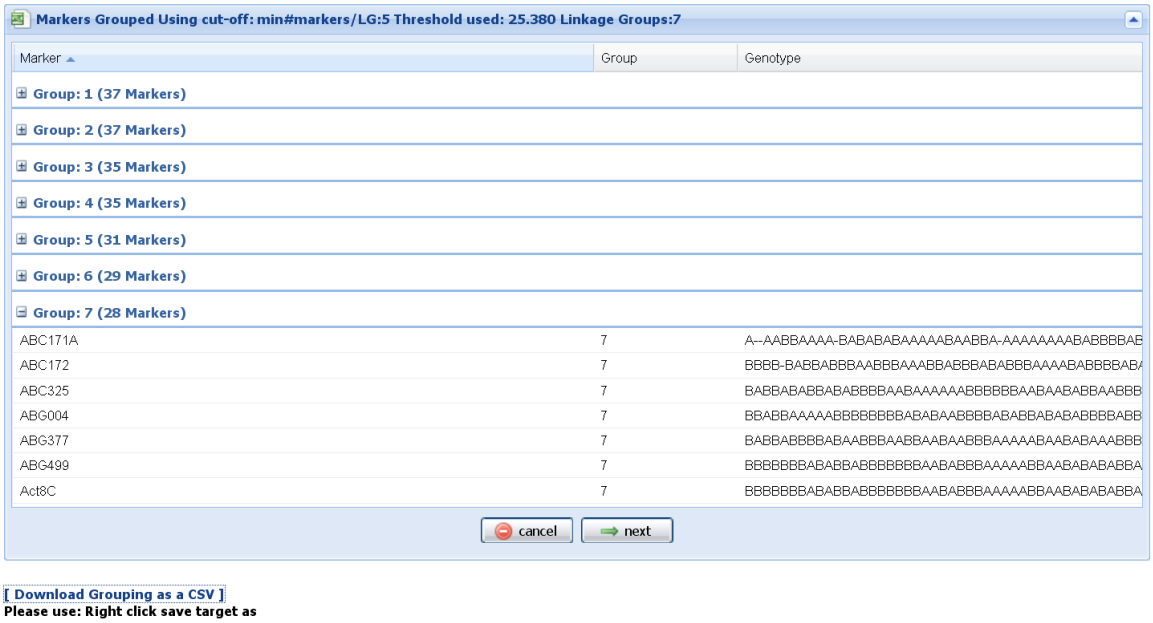


Figure 22: Marker membership of the 7 linkage groups generated via a Thresholding analysis of the dataset in Figure 21, using the Set Suggested method

12) Choosing an Embedding Method:
The embedding step of a THREaD Mapper Studio analysis transforms the inter-marker distance matrix (see section 7 above), projecting marker points into three dimensional space in such a way that allows the linkage group and ordering information to be visualised and inferred. The current version of THREaD Mapper Studio provides four embedding methods, as can be seen in Figure 23.

For single linkage groups the most appropriate method is multidimensional scaling (MDS). For a given distance matrix of inter-marker distances, MDS finds a corresponding set of low dimensional points (here the marker points in 3D space) that possess similar interpoint distances. The Isomap method is similar to MDS, instead applying the MDS methodology to a matrix of shortest paths or geodesic distances (see Tenenbaum, Silva and Langford, 2000). Isomap is renowned for its distance preserving properties so if your dataset derives from a single linkage group, you should embed it using MDS or Isomap.

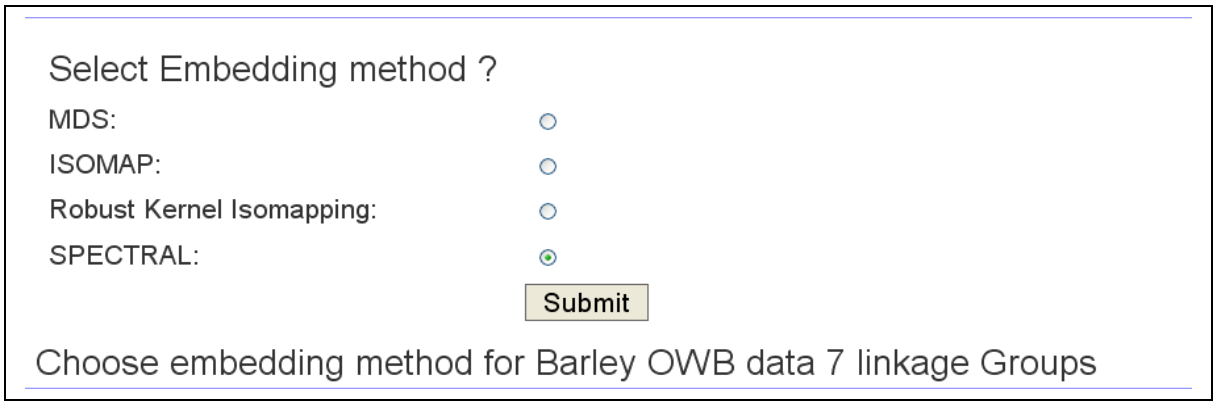


Figure 23: An example of the Embedding webpage, showing the four currently available methods

The primary method for multiple linkage groups is Spectral embedding (see Schoelkopf, 1998). The particular Spectral embedding algorithm we have deployed within THREaD Mapper Studio uses the power method to find small eigenvalues and eigenvectors of the Laplacian graph, then using those eigenvectors as coordinates for the nodes (Koren, 2005). This method is based on the eigenvectors of the affinity matrix that extract segmentation information from the top three eigenvectors. We have found this method to give a good performance for most datasets of this type. As it considers not only inter-marker distances but also the connectivity between markers (as opposed to MDS which uses only the inter-marker distances) it is a good method for drawing out a dataset's linkage group structure. See Figure 24 below for an example of a Spectral embedding. Users should avoid using Spectral embedding in the presence of one or more singleton markers (this can sometime be seen in the Thresholding plot as a large jump between the penultimate and last edges). The Robust Kernel Isomap (see Choi and Choi, 2007) is a recent addition in the field of kernel methods. It has been shown to possess better generalization properties and robustness against noise than other methods but has the disadvantage that it can be slow to run. Consequently, we recommend its use for smaller datasets only.

As with most THREaD Mapper Studio steps, users can use their web browser's Back and Forward buttons to change webpage settings. Here, this allows them to see the differences in the geometry of their dataset when analysed with different embedding methods. If a user is

unsure which method is most appropriate for their dataset then the SPECTRAL option should be used. An analytical option is selected by clicking on the radio button to the right of the desired analysis. Once this has been selected, clicking on the *Submit* button at the bottom of the page will take the user to the main analysis step.

Figure 24: A Spectral embedding of the barley 7 linkage group dataset seen in Figures 21 and 22 within THREaD Mapper Studio. The 7 linkage groups are clearly distinct.

13) The THREaD Mapper Studio Embedding Frame

i) Layout of the Embedding Frame

The user is presented with the **THREaD Mapper Studio Embedding Frame**, which is divided into five regions: a) the ***Top Toolbar***, b) the ***Left Side Marker Grouping Panel***, c) the ***Bottom Controls Toolbar and Attribute Groupings Panel***, d) the ***Right Side Zoom Control Bar and Connections Panel***, and e) the ***Central 3D Display***. Each of the regions may be resized by a mouse drag action upon the white boundaries between adjacent regions. Users can also collapse the bottom, left and right regions completely to widen their view of the central panel, the ***Central 3D Display***.

ii) The Central 3D Display

The ***Central 3D Display*** shows an embedding of the entire marker dataset. The display consists of a graph, with markers represented by spheres and relationships between "close" markers represented by lines. This display, which has a sky blue background, is zoomable and spinnable. To spin the display, simply click and hold down the left mouse button while moving it within the display. To zoom the display, use the mouse wheel to increase or decrease the size of the marker graph (alternatively, depress the Alt button on your keyboard while holding down the left mouse button and moving the mouse within the display). To move the position of the marker graph within the display, hold down the right mouse button and move the mouse within the display whilst depressing the Ctrl button on your keyboard. To find out the name of a particular marker within the ***Central 3D Display***, hold the mouse

over a marker of interest and its name should appear both next to it and on the extreme right hand side of the ***Top Toolbar***.

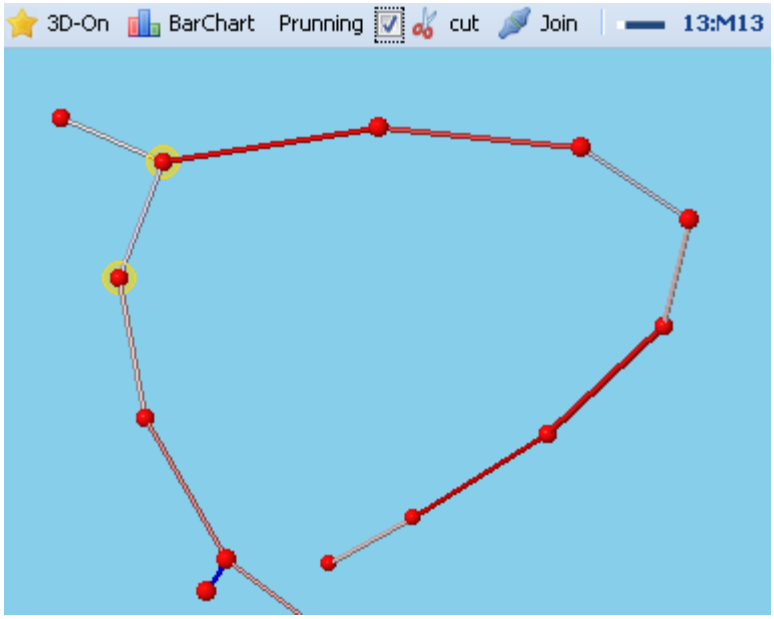


Figure 25: We first tick the *Pruning* checkbox, second use the mouse to select two adjacent markers (highlighted in yellow), and third use the *Cut* button to sever the connection between the two selected markers.

iii) The Top Toolbar

Let us now go through each of the regions of the **THREaD Mapper Studio Embedding Frame** in more detail, describing the effect of various options on the ***Central 3D Display.*** The ***Top Toolbar*** contains several important buttons for analysis of the marker graph. As noted above, the last (or current) marker in the ***Central 3D Display*** identified via a mouseover action is named at the right hand side of the ***Top Toolbar***. The *Pruning*, *Cut* and *Join* buttons may be used to sever and join connections between pairs of markers (thereby splitting or merging linkage groups). To cut a connection between two markers (i.e. to create an additional linkage group) first tick the Pruning checkbox by clicking on it so that THREaD Mapper Studio knows an edit is about to be attempted. Then, select a pair of markers within the ***Central 3D Display*** that are joined by a connection. You will see each of the markers surrounded by a yellow 'halo' once they have been selected (as seen in Figure 25 above).

Note that the ***Zoom Control Toolbar*** or the zoom functions within the ***Central 3D Display*** may be used if the markers are difficult to visualise. Next, click on the *Cut* button and finally untick the *Pruning* checkbox to turn off editing. A similar process, instead using the *Join* button, can used to make a connection between two unconnected markers (i.e. merging two distinct linkage groups), as shown in Figure 26 below.

The *Join* function can be useful to include a distant singleton marker within a linkage group or to remedy an erroneously made cut. The *3D-On* button is currently disabled. The *Update* button is used to save changes that have been made within the **Embedding Frame**, such as the result of separating the marker graph into distinct linkage groups. Note that the *Update* button by itself does not involve any reanalysis (i.e. re-embedding) of the dataset according to the current marker content and linkage group division.

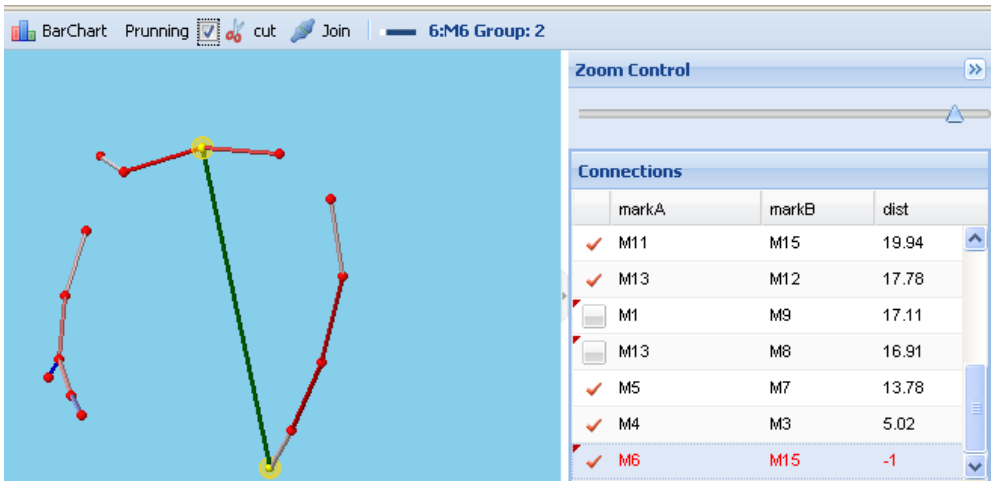


Figure 26: A new join is made between M6 and M15. Note that these two markers are not adjacent. This join in thereby registered as an 'unhappy' join and is displayed in red font in the ***Right Side Connections Panel***.

A new option recently added to the **Top Toolbar** is the *Re-embed* checkbox, which allows users to carry out a new embedding of their dataset following a major change to the data (for example, once markers have been assigned to different linkage groups). Note that this option should be chosen rarely due to its high computational expense. In order to use the *Re-embed* option, users should tick the *Re-embed* checkbox and should then click on the *Update* button. This will re-embed the dataset using the current choice of embedding algorithm, so if your current embedding has been carried out using MDS then re-embedding will also use MDS. When using Spectral embedding, you must be careful that you are not choosing to re-embed your dataset in the presence of a newly singleton marker, as the embedding may fail. Options here are to use your web browser's Back button in order to change the embedding method or to delete the singleton marker from the analysis.

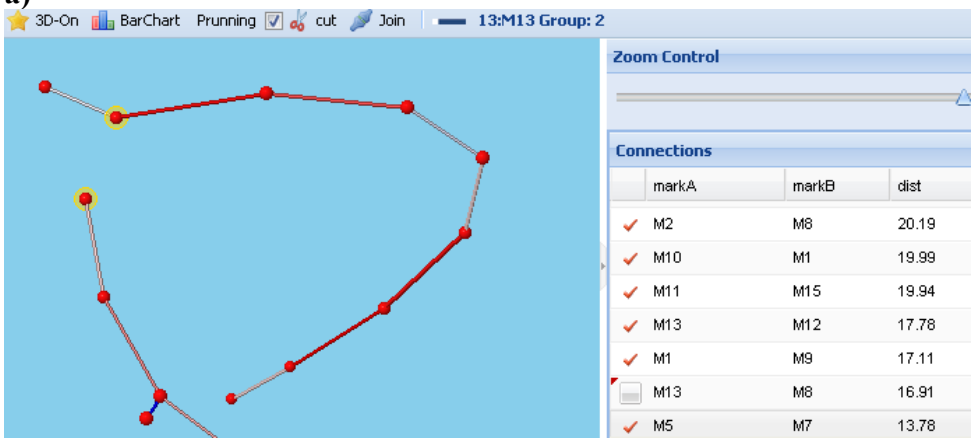
<u>iv) The Right Side Zoom Control Bar</u>

The ***Right Side Zoom Control Bar*** consists of a sliding control that adjusts the size of the spherical markers in the ***Central 3D Display***. Upon sliding this control to the left or right (using a click, slide and release action of the mouse) the marker sizes are reduced or increased respectively.

<u>v) The Right Side Connections Panel</u>

The ***Right Side Connections Panel*** shows a table of pairwise marker connections within the ***Central 3D Display***. There are four columns within the table. The first column consists of tickboxes for each pairwise connection, with a box ticked if the corresponding connection is currently active (i.e. it hasn't been severed using a *Cut* event). The second and third columns indicate the pair of markers joined by the connection. The final column shows the distance between the pair of markers. This panel is used mostly to show the current state of the ***Central 3D Display*** (see Figure 27 below). However, it can also be used to edit it. For example, when a dataset undergoing analysis is very large, it may be difficult to perform a *Cut* event. In such cases, it may be simpler to sever a connection by unticking the relevant box in the ***Right Side Connections Panel***.

a)



b)

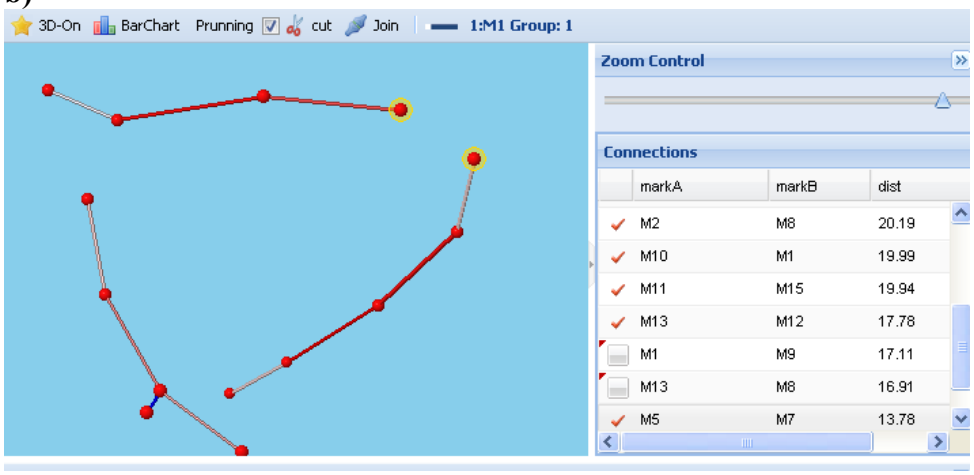


Figure 27: In a), a cut is made between markers M13 and M8 (highlighted in yellow). The event is registered within the ***Right Side Connections Panel***, as we see the corresponding connection is now unticked. In b), a further cut is made between markers M1 and M9, which is again registered within the ***Right Side Connections Panel***.

vi) The Left Side Marker Grouping Panel

The ***Left Side Marker Grouping Panel*** provides a summary of the linkage group information via a very intuitive scheme, where each folder represents a single linkage group and the marker membership of a linkage group can be seen by clicking on the *plus symbol* next to each folder icon (or collapsed by clicking on the *minus symbol*). Users can also expand/collapse all the linkage groups at once using the blue coloured *plus* or *minus symbol* icons next to the panel heading. Markers may be moved between linkage groups simply by dragging and dropping them from one group to another. For example, Figure 28 below describes a scenario where one marker (M4) is moved from linkage group (LG:3) to linkage group (LG:2) by an drag drop event.

Let us explore the additional functionalities associated with this panel. The Add icon (green) beneath the panel heading will add an entirely new, empty linkage group. Users may then add markers to it by drag-dropping from other linkage group folders, as described above. Figure 29 below shows an example of this process, with the new linkage group named GP300. Users may rename this newly created linkage group to a more appropriate name by a single click and typing action. In Figure 29 we have named the new linkage group as 'NEW4'. Note that this action provides the third of three ways in which to split a linkage group into two (the

others being via the *Cut* option in the ***Top Toolbar*** and by unticking a connection in the ***Right Side Connections Panel***), as described above.

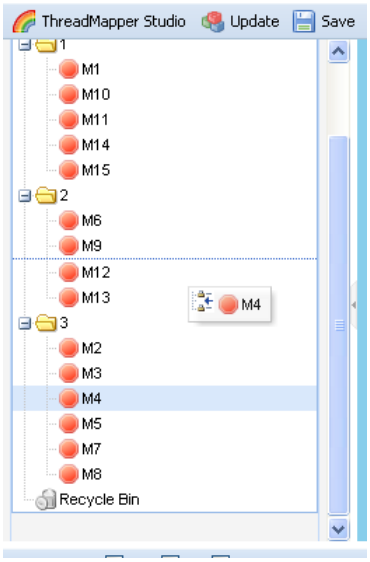


Figure 28: The process of moving a marker from one linkage group to another using a drag-drop event within the ***Left Side Marker Grouping Panel***. Marker M4 from linkage group 3 is moved to linkage group 2 by dragging it and then dropping it either into the target folder icon or anywhere between the area bounded by markers M6 and M13.

a) **b)**

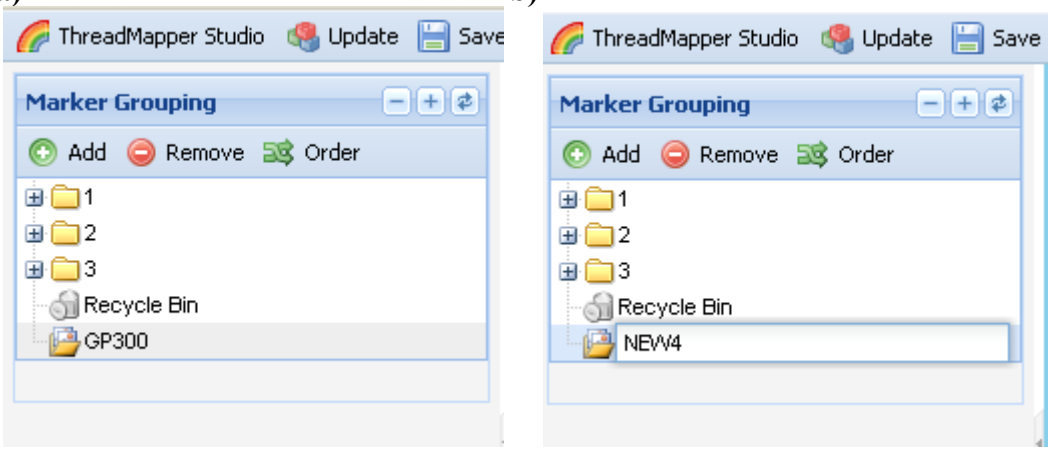


Figure 29: a) Adding and b) renaming a new linkage group.

In addition to creating a new linkage group, an entire linkage group, including all the markers contained within it, can be removed by selecting it and then clicking the *Remove* icon (red) button beneath the panel heading. Finally, the *Order* icon (green) button, again beneath the panel heading, is used to order a selected linkage group once it is finalised, as described below.

vii) The Bottom Controls Toolbar

The ***Bottom Controls Toolbar*** contains simple options that control the ***Central 3D Display***. Clicking the *Axes* box toggles on or off the 3D axes corresponding to the co-ordinate system

in which the markers are plotted. The *Spin* box toggles on or off an automated spin of the marker graph. The *Marker* and *Label* boxes toggle on or off the text display of marker input order (i.e. within the Genotype File) or marker name adjacent to the marker spheres. The *MST* box toggles on or off the Minimum Spanning Tree connecting the marker points. This Toolbar also contains a *Colour GPs* button which, when clicked, allows users to toggle on and off the marker colouring via Attribute Grouping facilities in the ***Bottom Attribute Groupings Panel***. The ***Bottom Controls Toolbar*** also contains buttons called *Back* and *Next* to navigate through multiple Attribute Groupings within the ***Bottom Attribute Groupings Panel***, should these be available.

viii) The Bottom Attribute Groupings Panel

In the ***Bottom Attribute Groupings Panel***, users will find the Attribute Grouping information contained within the original dataset. As noted in section 3, multiple Attribute Groupings may be provided within an input file. Users may move through these various options using the *Next* and *Back* buttons on the ***Bottom Attribute Groupings Panel***. Expand a chosen Attribute Grouping by clicking on the + sign to the left of the Attribute Grouping name. To the left of each group within an Attribute Grouping, a checkbox will be found. Clicking on a checkbox causes the relevant marker spheres in the ***Central 3D Display*** to be coloured according to their Attribute Grouping designation. By clicking on the *Colour GPs* button in the ***Bottom Controls Toolbar***, the Attribute Grouping colouring will be toggled on and off.

14) Ordering markers within a Linkage Group:

We saw in section 13vi above that a linkage group could be selected for ordering by clicking on the *Order* button within the ***Left Side Marker Grouping Panel***, once a chosen linkage group had been selected, as shown in Figure 30 below. If the linkage group consists of three or more markers, they will be ordered using a novel algorithmic approach called nonlinear geodesic smoothing. Here, a trendline is "threaded" through the markers plotted in 3D space.

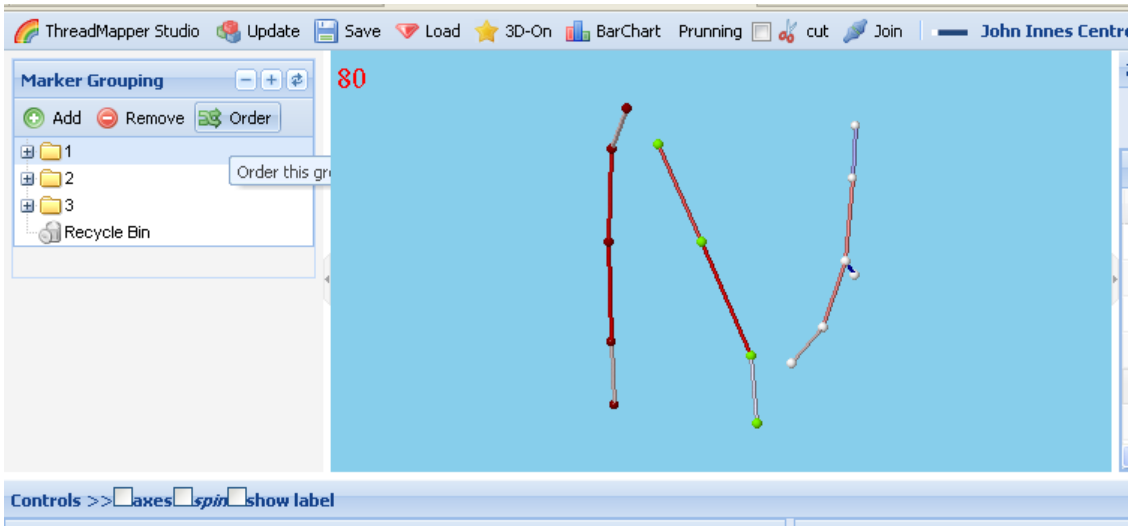


Figure 30: Linkage group 1 (here, white-coloured markers) of the BiB sample Dataset is selected and the *Order* button in the ***Left Side Marker Grouping Panel*** is pressed.

Once a linkage group has been ordered, the screen looks similar to the **Embedding Frame** in the previous step (this screen is also seen when choosing the MDS method of analysis – see section 12). However, some functionality is no longer available (e.g. *Pruning*, *Cut* and *Join* buttons on the ***Top Toolbar***) and some new functionality has been added. Three significant

new functions are the ***Right Side Ordering Panel*** (which replaces the Right Side Connections Panel), the *Heatplot* button on the ***Top Toolbar***, and the download and additional display options on the ***Bottom Controls Toolbar.***

i) The Right Side Ordering Panel

This panel presents a simple table of four columns for the ordered linkage group. In the first column, the marker names are given, in marker order along the linkage group. In the second column, the shortest distance (residual) of the relevant marker from the trendline is given. In the third column the distance of the marker along the linkage group is given and in the fourth, the corresponding distance in centimorgans is given.

ii) The Heatplot button

The *Heatplot* button in the ***Top Toolbar*** presents two heatplots to be compared. The first heatplot visualises the inter-marker distance matrix of the unordered markers within the linkage group undergoing ordering. The second heatplot shows the same matrix, but this time for ordered markers. This function can be useful in providing a visual verification of the ordering process.

iii) The Bottom Controls Toolbar

This Toolbar provides several options, both for downloading the linkage group results in various formats and for changing the display of the ***Central 3D Panel***. Two download options are available. The *Ordered CSV: Link* option allows users to save, via right mouse click (and subsequent file location and name specification) the ordered linkage group in CSV format. The first column of this file is the marker name, followed by columns specifying the final marker order, the input marker order, the residual (distance from the trendline), the geodesic and centimorgan distances, and finally the marker scores. This file can then be used for further data exploration, for example in third party plotting software. The *Line Plot PDF: Link* function allows a PDF file with a familiar linear genetic map graphic, as shown in Figure 31 below, to be downloaded.

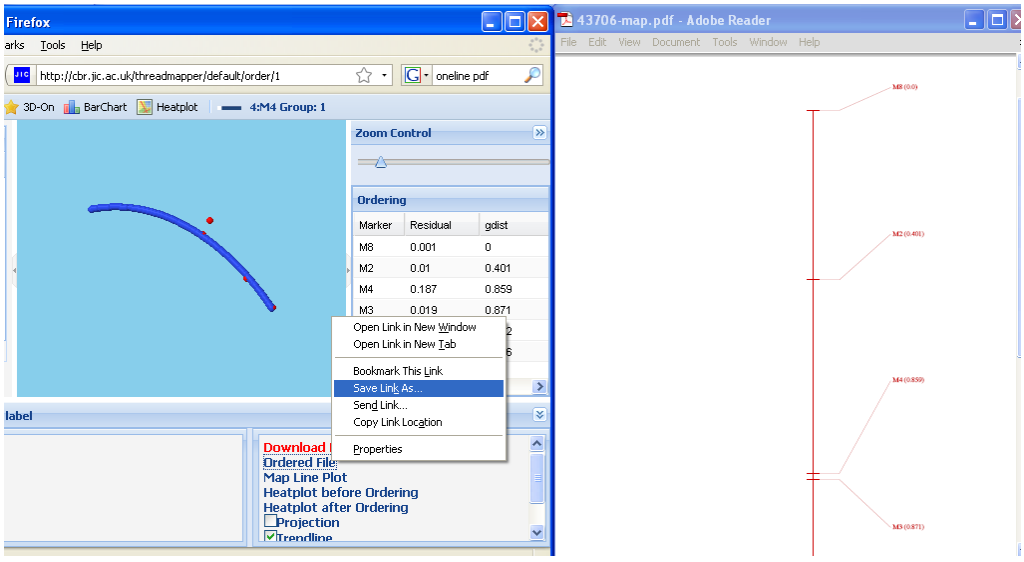


Figure 31: The *Map Line Plot* function can be used to save, as a PDF formatted file, a visualisation of the genetic map estimate for the selected linkage group.

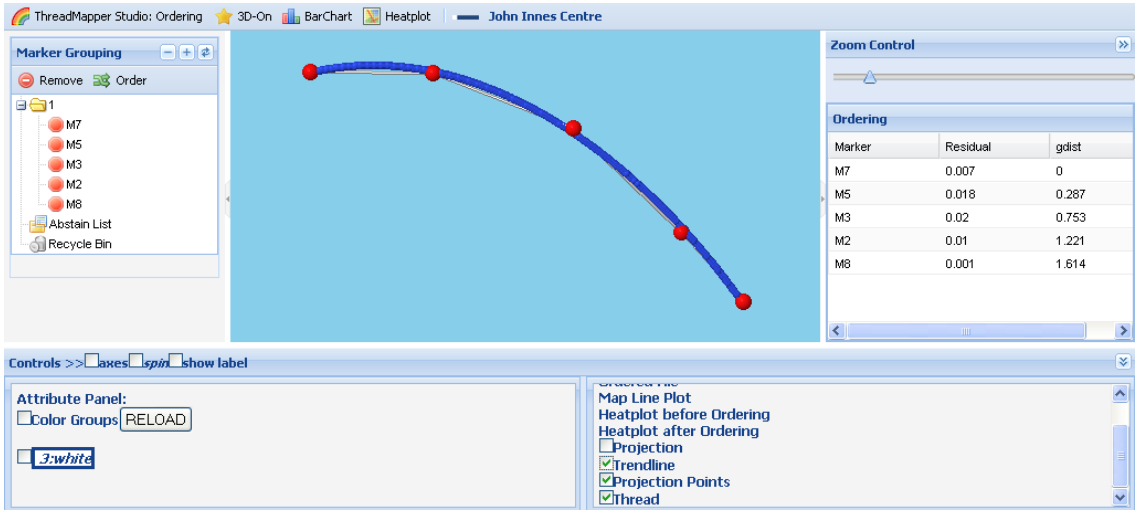


Figure 32: Toggle functions in the ***Bottom Controls Toobar*** change the appearance of the ***Central 3D Display***.

Other options within this toolbar change the appearance of the ordered linkage group within the ***Central 3D Display***, as indicated in Figure 32 above. These controls are check boxes meant to be toggled on and off (with a ticked box being on and an unticked boxed being off). However, for some browsers we have found that these states appear inverted (i.e. with a ticked box being off and an unticked box being on) and we are currently investigating solutions to this problem. The *Axes*, *Spin*, *Marker* and *Label* options function as before. In addition, the *Trendline* function toggles on and off the thick blue trendline running through the midst of the marker points. The *Projection* function toggles on and off the "projection lines" (i.e. shortest routes in 3D space) from the markers to the trendline. The *Projection Points* function toggles on and off the points on the trendline onto which the markers are projected (these projected points provide the distances in the final genetic map estimation). The *Threading* function toggles on and off the shortest lines between adjacent markers. The *Back*, *Next*, and *Colour GPs* buttons function as before with the grouping within the ***Bottom Attribute Groupings Panel***.

<u>iv) Ordering several linkage groups</u>

Once a linkage group has been ordered, further linkage groups may be ordered, one at a time. For each new linkage group, users should simply click the Back button of their web browser. They should then select the new linkage group to be analysed and proceed as before.

15) An example analysis: the BiB dataset

Now that we have looked at the functionality of the THREaD Mapper Studio website, let us illustrate its use via a simple example. We will examine a basic analysis of the BiB dataset shown in Figure 6.

a) Select the dataset. This can be done quickly and easily by going to the **THREaD Mapper Studio** homepage, going to the **List of Datasets** section and clicking on *Select* to the right of the "BiBs data set with Header Attributes" file description.

b) On the Basic Statistics page, click on the *Save* button to select the dataset without removing any markers (remember, markers with high levels of segregation distortion or missing scores can be removed via the checkboxes and the *Remove Marker* button). The

Genotype Distribution Chart shows that the number of missing scores is quite small for this dataset.
c) Select the Hamming scoring matrix for inter-marker distance calculation by clicking on the radio button to the right of its text description and then click on the *Submit* button.

d) Examine the Heatplot of the dataset (which consists of 3 linkage groups) and click on the *CONTINUE>>* button.

e) At the linkage group and ensemble choice webpage, we will need to choose the default method of multiple linkage groups as the BiB dataset consists of 3 linkage groups. In addition, we do not wish to generate an Ensemble MST for this analysis, so we will choose the default of T=1. To continue to the next step, click on the *Submit* button.

f) The Thresholding plot shows three major jumps. On this occasion, we will not use the plot to cluster the dataset into distinct linkage groups. It is good practice, at least in the early stages of any analysis, to skip this feature so that the uncut structure of the dataset may be seen in its entirety. Once this structure is understood, a user may wish to use the Thresholding tool. Click on the *Ignore/Skip Threshold >>* button to move to the next step without clustering the markers.

g) As we have not clustered the dataset into distinct linkage groups at the Thresholding step, the Linkage Group spreadsheet shows that all 15 markers belong to a single linkage group. Click on the *next* button to move to the next step of the analysis.

h) As the BiB dataset consists of 3 linkage groups, we wish to carry out a SPECTRAL analysis upon it, as this is suitable for the analysis of many linkage groups simultaneously. A SPECTRAL analysis uses the connectivity within the inter-marker distance matrix to tease apart the different linkage groups. As a SPECTRAL analysis is the default embedding method, we do not need to click on any of the radio buttons. Proceed to the next step by clicking on the *Submit* button.

i) You should now be presented with the **THREaD Mapper Embedding Frame** for this dataset. The first thing that should be done is to use the ***Right Side Zoom Control Bar*** to resize the markers in the ***Central 3D Display*** to a comfortable viewing size. You should then be able to see clearly that the display consists of a graph, with markers represented by spheres and relationships between "close" markers represented by lines. Figure 33 below shows a **THREaD Mapper Embedding Frame** for the BiB sample Dataset, following a SPECTRAL analysis.

j) In the ***Bottom Attribute Groupings Panel***, you will find the Attribute Group information contained within the original dataset (note that this information corresponds to the known linkage groups and that it need not have been contained within the original dataset - we added it for greater understanding of the dataset and the analytical process). By clicking on the *Next* button on the extreme right of the ***Bottom Controls Toolbar***, the Attribute Grouping information (here called manual, consisting of the 3 known linkage groups) may be seen. Expand the manual Attribute Grouping by clicking on the + sign to the left of the group name. The three groups (coloured white, dark red and lawn green respectively) can be marked for display within the ***Central 3D Display*** by clicking the checkboxes to the left of each group. This cause the marker spheres in the ***Central 3D Display*** to be coloured according to their linkage group designation. Clicking on the *Colour GPs* box on the ***Bottom Controls Toolbar***,

causes the marker Attribute Group colouring to be toggled on and off. By using this feature, it becomes apparent that the major "bends" within the multi-linkage group dataset are where distinct linkage groups meet. Experienced users will find they are able to spot and explore these bends relatively quickly and accurately.

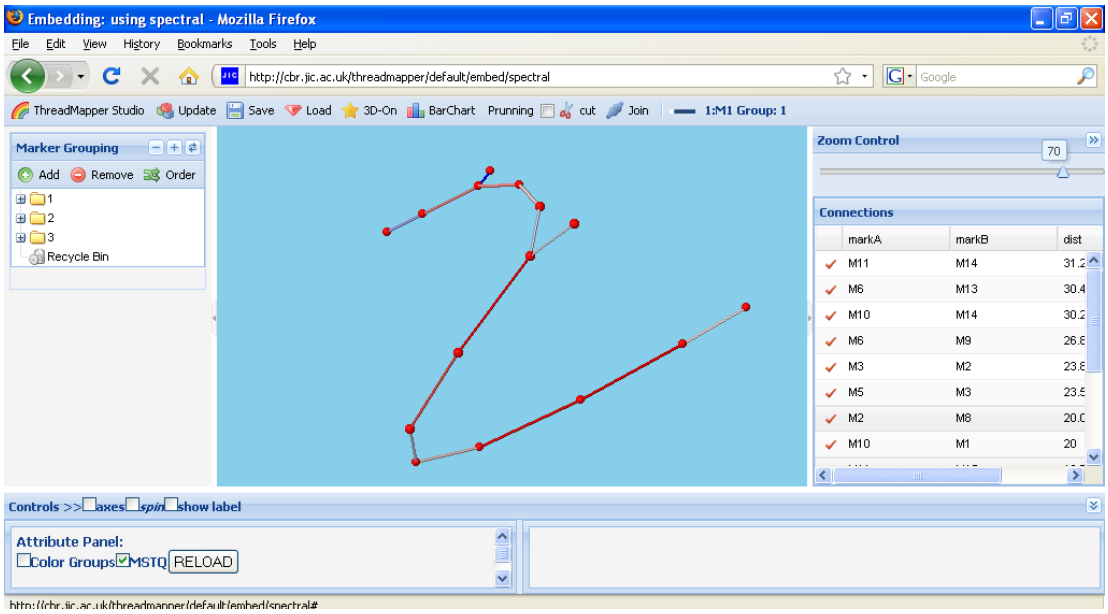


Figure 33: The **THREaD Mapper Embedding Frame** for the BiB dataset.

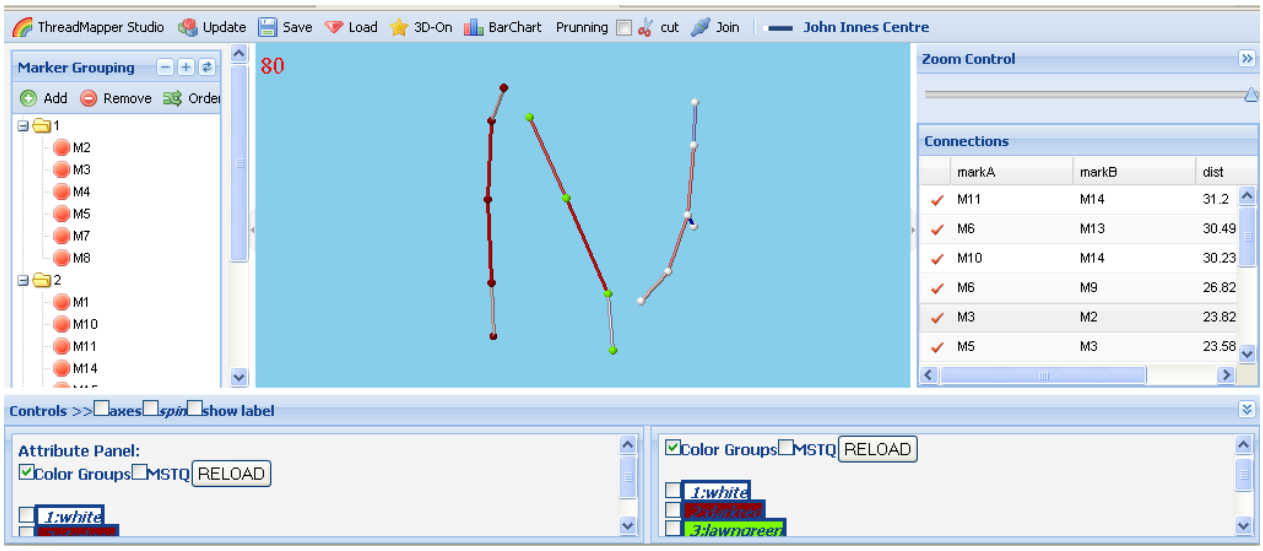


Figure 34: The BiB dataset has been cut into 3 distinct linkage groups (each of which is now analogous to an Attribute Group).

k) Now we will cut the dataset into three linkage groups using the *Pruning* and *Cut* options in the ***Top Toolbar*** and using the ***Right Side Connections Panel***. First, we will separate the white-coloured and dark-red coloured groups. Click on the *Pruning* checkbox so that it is ticked. Next, select the adjacent M1 (white) and M9 (lawn green) markers, that bound the connection between these two groups. You will notice that they are surrounded by a yellow 'halo' once selected. Next click on the *Cut* button and the two groups become disjointed. Click on the *Update* button in the ***Top Toolbar*** to save this event. Note that the information

displayed in the ***Left Side Marker Grouping Panel*** and in the ***Right Side Connections Panel*** both change upon this action and that the markers revert to a single colour. Click the *Colour Groups* checkbox in the ***Bottom Attribute Groups Panel*** again to bring back the colour groupings. Second, we will separate the dark red-coloured and the lawn green-coloured groups. In the ***Right Side Connections Panel*** find the connection row between markers M13 (lawn green) and M8 (dark red) that bound the connection between the two groups. Click on the tickbox at the left side of the row so that it becomes unticked. Then click on the *Update* button to save the result of this event. We have now separated our dataset into its constituent 3 linkage groups. Figure 34 above shows the result of this process, with 3 distinct groups.

l) Select linkage group 1 (white), consisting of 6 markers, in the ***Left Side Marker Grouping Panel*** and click on the *Order* button within the same panel. Save the resulting map co-ordinates as a CSV file by right clicking on the *Ordered CSV: Link* button within the ***Bottom Download Panel***, and save a PDF graphic of the ordered genetic map by right clicking on the adjacent *Line Plot PDF: Link* button.

m) Repeat step l) above for linkage groups 2 and 3. To do this first click on the Back button of your web browser to return to the previous main ***Embedding Frame***. The analysis is now complete.

16) References

i) Carter, T. C. and Falconer, D. S. (1951) Stocks for detecting linkage in the mouse and the theory of design. *J. Genet.* 50: 307–323.

ii) Tenenbaum, J. B., De Silva, V. and Langford, J. C. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290 (5500): 2319-2323.

iii) Carreira-Perpiñán, M. Á. and Zemel, R. S. (2005) Proximity graphs for clustering and manifold learning. *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pp. 225-232.

iv) Koren, Y. (2004) Joint EUROGRAPHICS - IEEE TCVG Symposium on Visualization

v) Deussen, O., Hansen, C., Keim, D. A. and Saupe, D. (Editors) Graph Drawing by Subspace Optimization

vi) Carreira-Perpinan, M. A. and Zemel, R. S. (2005) Proximity graphs for clustering and manifold learning. In L. K. Saul, Y. Weiss and L. Bottou (Eds.), Advances in neural information processing systems 17. Cambridge, MA: MIT Press

vii) Economou, G., Pothos, V. and Ifantis, A. (2004) Geodesic distance and MST based image segmentation. In XII European Signal Processing Conference (EUSIPCO 2004), pages 941–944,

viii) Extracting Dynamics from Static Cancer Expression Data(2008) EEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) archive Volume 5, Issue 2 (April 2008) pp 172-182, ISSN:1545-5963.

ix) Koren, Y. (2005) Drawing graphs by eigenvectors: theory and practice Computers & Mathematics with Applications Volume 49, Issues 11-12, June 2005, Pages 1867-1888.