TetraploidSNPMap

Software suite for linkage analysis and QTL mapping for autotetraploid populations using SNP dosage data

C. A. Hackett, B. Boskamp, A. Vogogias, K. Preedy, I. Milne

Introduction

TetraploidSNPMap is a development of a previous program, TetraploidMap for Windows, for calculating linkage maps in autotetraploid species. It is suitable for handling SNPs scored on two parents and the full-sib offspring of a cross between them. TetraploidSNPMap uses genetic marker data from new technologies that enable the dosage of SNP alleles to be estimated, rather than simply their presence or absence, and methods have been modified to handle large numbers of SNPs. The program has two modules, linkage analysis and QTL mapping.

The programs underlying the original TetraploidMap for Windows were developed by Z.W. Luo and C.A. Hackett with funding from the BBSRC under the initiative 'Genome Analysis of Agriculturally Important Traits' (GAIT). The Windows front end for that program, and many other improvements, were made by I. Milne. The programs for this extension have been developed by C.A. Hackett, B. Boskamp, A. Vogogias and K. Preedy, with advice from I. Milne. The theory for using SNP dosage data in autotetraploid mapping has been described in the following papers:

Hackett, C.A., McLean, K. and Bryan, G.J. (2013). Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PLoS ONE* 8, e63939.
(Allele dosage calling, linkage mapping and application to autotetraploid potato)

Hackett, C.A., Bradshaw, J.E. and Bryan, G.J. (2014). QTL mapping in autotetraploids using SNP dosage information. *Theoretical and Applied Genetics* 127: 1885-1904.
(QTL mapping theory and application to autotetraploid potato).

The older theory is described in the following papers:
Hackett, C.A., Bradshaw, J.E., Meyer, R.C., McNicol, J.W., Milbourne, D. *et al*, (1998). Linkage analysis in tetraploid species: a simulation study. *Genetical Research* 71: 143-154.
(Theory for dominant markers in simplex and duplex configurations)

Meyer, R.C., Milbourne, D., Hackett, C.A., Bradshaw, J.E., McNicol, J.W. *et al*., (1998). Linkage analysis in tetraploid potato and associations of markers with quantitative resistance to late blight (*Phytophthora infestans*). *Molecular and General Genetics* 259: 150-160.
(Theory for dominant markers extended to include double-simplex configurations, and applied to a potato population)

Luo, Z.W., Hackett, C.A., Bradshaw, J.E., McNicol, J.W. and Milbourne, D. (2000). Predicting parental genotypes and gene segregation for tetrasomic inheritance. *Theoretical and Applied Genetics* 100: 1067-1073.
(Theory to infer parental genotypes from parental and offspring phenotypes, and to test for double reduction)

Luo, Z.W., Hackett, C.A., Bradshaw, J.E., McNicol, J.W. and Milbourne, D. (2001). Construction of a genetic linkage map in tetraploid species using molecular markers. *Genetics* 157: 1369-1385.
(Theory for separating markers into linkage groups and calculating linkage maps)

Hackett, C.A., Pande, B. and Bryan, G.J. (2003). Constructing linkage maps in autotetraploid species using simulated annealing. *Theoretical and Applied Genetics* 106: 1107-1115.
(Theory of marker ordering using simulated annealing).

Hackett, C.A., Bradshaw, J.E. and McNicol, J.W. (2001). Interval mapping of QTLs in autotetraploid species. *Genetics* 159: 1819-1832.
(Methodology for QTL mapping)

The original software is described in
Hackett, C.A. and Luo, Z.W. (2003). TetraploidMap: construction of a linkage map in autotetraploid species. *Journal of Heredity* 94: 358-9.
Hackett C.A., Milne I., Bradshaw J.E. and Luo Z.W. (2007). TetraploidMap for Windows: linkage map construction and QTL mapping in autotetraploid species. *Journal of Heredity* 98: 727-729.

Applications of the original software have been published in:
Bradshaw, J.E., Pande, B., Bryan, G., Hackett, C.A., McLean, K. and Stewart, H.E. (2004). Interval mapping of quantitative trait loci for resistance to late blight (*Phytophthora infestans* (Mont.) de Bary), height and maturity in a tetraploid population of potato (*Solanum tuberosum* subsp. *tuberosum)*. *Genetics* 168: 983-995.
Bryan, G., Hackett, C.A., McLean, K., Pande, B., Purvis, A., Bradshaw, J.E. and Waugh, R. (2004). Genetical dissection of H3-mediated polygenic PCN resistance in a heterozygous autotetraploid potato population. *Molecular Breeding* 14: 105-116.
Bradshaw, J.E., Bryan, G.J., Hackett, C.A. McLean, K., Pande, B., Stewart, H.E. and Waugh, R. (2004). Dissection and analysis of quantitative disease resistance in tetraploid potato. *Euphytica* 137: 13-18.

Bradshaw J.E., Hackett C.A., Pande B., Waugh R. and Bryan G.J. (2008). QTL mapping of yield, agronomic and quality traits in tetraploid potato (*Solanum tuberosum* subsp. *tuberosum*) *Theoretical and Applied Genetics* 116: 193-211.

Note on datasets supplied with the software

Four datasets are supplied with the software. For the linkage analysis module, theta5332.loc contains dosage scores on 5332 SNPs from the Infinium 8303 potato SNP array, for a population consisting of the cultivar Stirling, the breeding line 12601ab1 and 190 offspring. Further details are available in Hackett et al. (2013). For the QTL mapping modules, there are files consisting of the linkage map of linkage group V for this cross from the above paper with marker phasing (Oct_LGV.map), the SNP data for this linkage group (Oct_LGV.SNPloc) and data on 12 phenotypic traits (demo.qua). The 12 traits are means of yield, plant height, size, shape, dry matter, frying colour at 4°C and 10°C, maturity, foliage blight, tuber blight, *Globodera pallida* cyst counts (on a square root scale) and flower colour, scored as 1 = white or 2 = blue. Further details of these traits can be found in Hackett et al. (2014), Bradshaw et al. (2004) and Bradshaw et al. (2008).
The four datasets (and this user manual) are installed in a folder named 'docdata' within the program installation folder, which is usually *C:\Program Files (x86)\TetraploidSNPmap 64bit*

System requirements

TetraploidSNPmap has been tested on 64-bit versions of Windows-7 and Windows-10, on PC's with at least 8GB ram. It requires the 64 bit Java Runtime Environment (JRE).

Module 1: Linkage Map Construction

Step 1. Data preparation

TetraploidSNPMap is designed to handle SNP dosage data. This can be obtained from different genetic technologies by different methods. Some details of methods to obtain dosages from Illumina theta scores are given in Appendix B.

1.1 Preparation of SNP dosage data file

The SNP dosage data is input as a text file, with a format where each row represents a SNP, and individuals are in different columns, as in Example 1.1 below.

Example 1.1 Small data file of 190 offspring and 50 SNPs, showing first five SNPs only.

```
190 50
solcap_snp_c1_1              3 3 3 9 4 3 4 3 3 2 2 4 4 3 3 3 3 2 3 2 2 4 3 4 3
3 2 3 3 4 2 3 2 4 4 3 3 2 3 2 3 2 3 3 3 2 3 3 4 2 3 2 4 3 3 3 4 2 3 3 2 4 3 4 3
4 2 4 3 4 3 4 3 3 3 2 3 3 2 3 2 3 2 3 2 3 2 3 2 2 3 3 3 3 3 4 4 3 4 3 3 3 4 3 3 2 2
2 3 4 2 3 3 3 4 4 3 3 2 4 3 3 2 3 4 2 4 3 3 4 3 3 3 4 3 3 4 2 3 3 2 2 3 3 3 2 2
3 3 2 3 4 4 2 3 2 3 3 3 2 4 2 2 3 3 3 3 3 4 4 3 2 4 3 2 3 4 4 2 2 2 3 3 3 4 2 3
3 3 3 2 3 4 3
solcap_snp_c1_1000           0 1 1 9 9 1 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 1 1 0 1
1 0 0 0 0 1 0 1 0 0 0 0 1 1 1 0 0 1 1 0 1 0 0 1 1 0 1 1 0 0 0 0 0 1 1 1 0 0 1 0
0 1 0 0 1 0 0 0 1 0 1 1 0 0 1 1 0 0 1 1 1 1 1 1 1 0 1 1 0 1 1 1 0 1 1 1 1 1 1 0
0 1 0 0 1 1 1 0 0 0 1 0 1 1 1 1 0 1 1 1 0 0 1 0 0 0 1 1 0 1 0 0 1 1 1 1 0 1 0 0
0 0 0 1 1 0 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1
0 1 1 0 0 0 1
solcap_snp_c1_10000          2 3 4 2 2 2 3 3 2 2 2 2 3 3 4 3 3 4 2 2 4 3 2 1 3
4 4 4 4 2 3 3 4 2 3 3 1 3 2 3 2 3 3 4 2 2 2 2 2 3 2 2 2 3 3 2 2 2 2 3 2 2 2 3 2
2 2 2 2 3 2 2 3 3 2 3 2 2 2 2 2 3 3 2 2 2 3 2 3 3 1 2 3 3 2 3 3 2 3 2 3 3 2 3 3
3 2 3 3 2 2 3 3 1 2 2 3 2 3 2 3 2 3 3 3 2 2 2 2 3 2 3 2 1 1 3 3 2 2 2 2 4 3 2 2
3 3 3 3 3 3 2 3 2 3 2 3 3 2 3 2 3 3 2 2 2 4 3 4 3 2 3 3 2 2 2 2 3 3 3 2 4 3 3 2
2 3 3 4 3 2 2
solcap_snp_c1_10001          4 2 2 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 4 2 3 3 4 3
3 3 3 2 3 3 3 3 3 2 3 4 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 2 3 3 3 3 3
4 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 2 3 3 3 3 3 3 3 3 3
3 3 2 3 3 3 3 3 4 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 2 3 3 3 3 3 2 3 3 3
2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 2 3 3 3 3 3 3 3 3 3 3 3 2 2 3 3
3 2 2 2 2 3 3
solcap_snp_c1_10011          1 2 3 1 1 1 2 2 1 1 1 1 2 1 2 2 2 3 1 1 3 2 1 0 1
2 2 2 2 1 2 2 2 1 2 2 0 2 1 2 1 2 2 3 1 1 1 1 1 2 1 1 1 2 2 1 1 1 1 2 1 1 1 2 1
1 1 1 1 2 1 1 2 2 1 2 1 1 1 1 1 2 2 1 1 1 2 1 2 2 1 1 2 2 1 2 2 1 2 1 2 2 1 2 2
2 1 3 2 1 1 2 2 0 1 2 2 2 2 1 2 1 2 2 2 1 1 1 2 2 2 2 1 0 0 2 2 1 1 1 1 3 2 1 1
2 2 1 2 2 2 1 2 1 2 1 2 2 1 2 1 2 2 1 1 1 3 2 3 2 1 2 2 1 1 2 1 2 2 2 1 3 2 2 1
1 2 2 3 2 1 1
etc
```

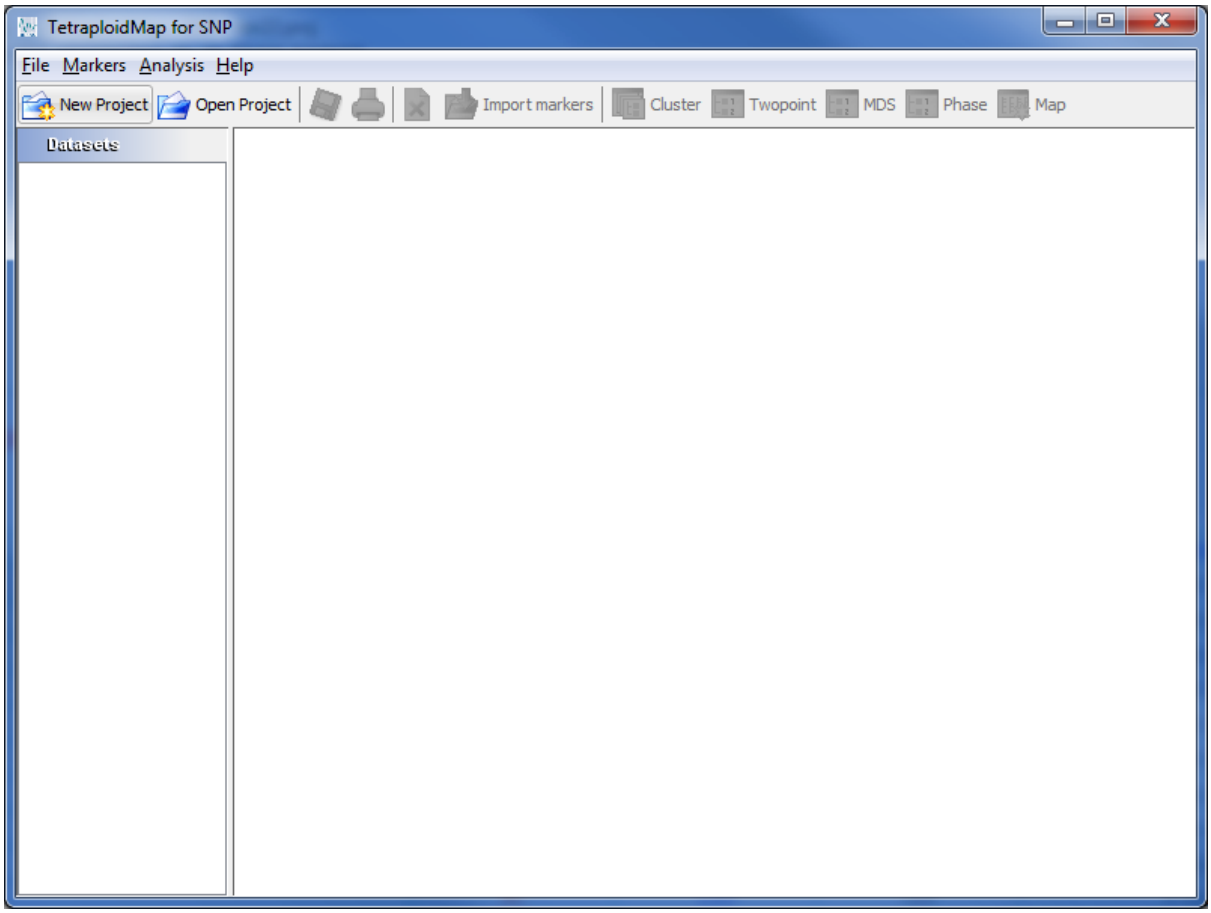
The first line of the file contains the number of offspring (not counting the two parents) and the number of loci in the file. This is followed by the data for each SNP locus.

This is laid out as the SNP name (maximum 20 characters), the dosage for each parent and then the dosage for each offspring. Dosages are coded as 0-4, representing genotypes AAAA, AAAB, AABB, ABBB and BBBB (ie a count of the number of B alleles) and 9 represents a missing value. If there are too many individuals for the data on one SNP to fit on a single line, the scores can be continued onto subsequent lines. Neither of the parental dosages can be missing. Note that when the chromosome of origin is known, it is useful to include this as part of the name.

If TetraploidSNPMap is unable to open a data file, this may be due to the lack of a carriage return after the last line of data in the file.

Step 2: Starting TetraploidSNPMap and creating a project

TetraploidSNPMap is started by starting the program and selecting the 'SNP' project option. It will start with a window:



The first step is to define a project by clicking on the New Project tab (or open an existing project with the Open Project tab). This will bring up a dialogue box where the user can select the directory containing the data, and define a project there. The default name is MyProject.proj. Once this has been created, the user is invited to import the SNP data using the Import markers tab.

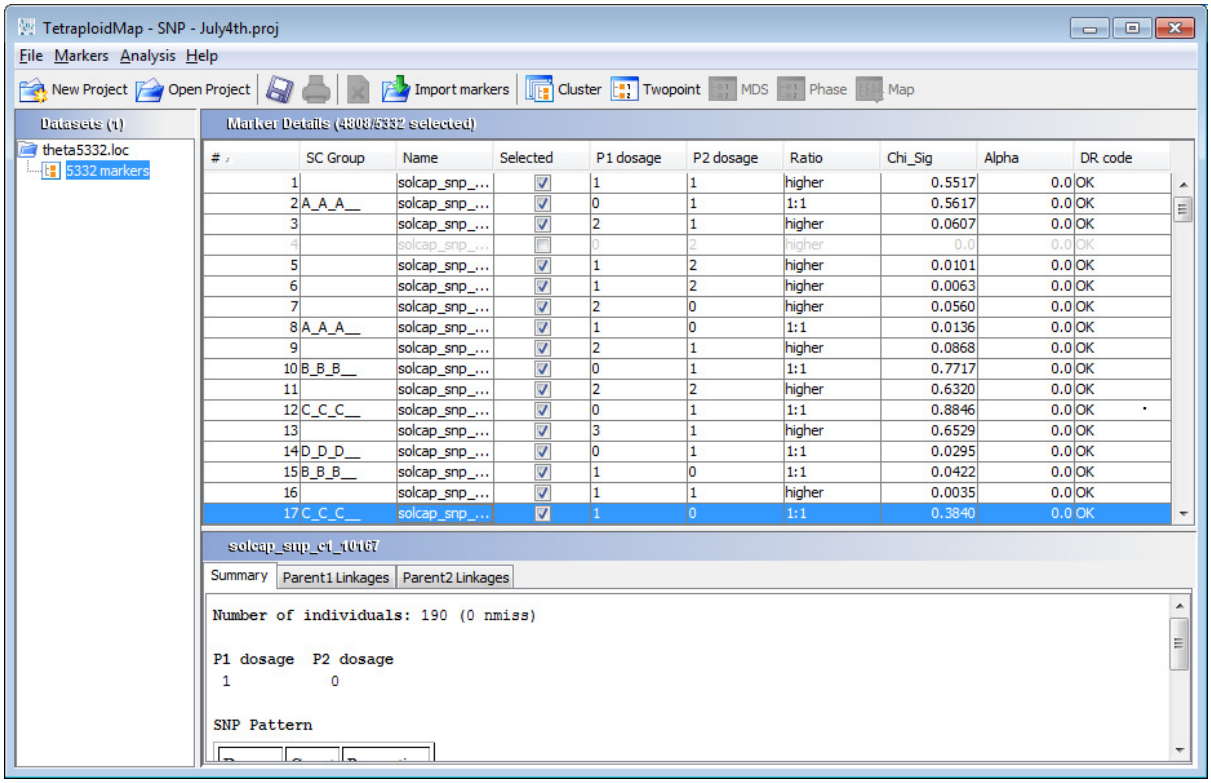
The tpmap folder

As TetraploidSNPMap is run, some files with extra output are written to a folder tpmap. These are temporary files and are overwritten at the next step. This folder may be placed in different local folders on different computers. For some analyses, supplementary information can be saved by accessing this folder and copying the files there to a permanent location. When this can be useful, we will refer to the tpmap folder and the necessary output file in the text below.

Step 3: Segregation analysis and preliminary grouping

Overview

Once the SNP data has been read successfully, TetraploidSNPMap automatically analyses each SNP in turn to test whether its segregation ratio is consistent with the parental dosages. SNPs with an offspring dosage that is either incompatible with the parental genotypes, or only compatible if double reduction occurs, are identified (eg for a simplex SNP, AAAB x AAAA or parental dosage (1,0), offspring with dosage 0 or 1 can be obtained by a simple model of random chromosomal segregation, dosage 2 can be obtained by double reduction but dosages of 3 or 4 are incompatible). If any incompatible scores are found, the SNP scoring must be corrected or that SNP excluded. At present TetraploidSNPMap uses a model of random chromosomal segregation and cannot include double reduction, so SNPs with double reduction products have to be excluded or the double reduction product replaced with the missing value code 9. For SNPs inferred as simplex, the program performs a preliminary cluster analysis to identify which SNPs are linked in coupling and shows this as a code before the SNP name. Chi-squared tests of independence are performed to give a first indication of relationships of these simplex SNPs to other well-behaved SNPs. The upper part of the output has the following form:



The columns show:

- the locus number
- 'SC Group' – a code indicating the simplex coupling group for SNPs inferred as simplex (see below).
- 'Name' - its name
- 'Selected' - whether it is currently selected or not (see Step 4)
- 'P1 dosage' and 'P2 dosage' - the phenotypes of the two parents
- 'Ratio' – 1:1 for simplex markers, or 'higher' for other dosages
- 'Chi_Sig' – the significance of a chi-square test for significant departures from the expected ratio (in the absence of double reduction).
- 'Alpha' – an estimate of the coefficient of double reduction, α.
- 'DR code' – a code indicating whether the offspring data is compatible with the parental dosages in the absence of double reduction (OK), is compatible with the parental dosages only if double reduction occurs (DR), is incompatible with the parental dosages (NP), is monomorphic (MO), has missing data for one or both parents (MP) or has zero code for both parents (00).

The results can be sorted by any column by clicking on the column name.

3.1 SNP summary information

Initially the 'Summary' tab is highlighted, and summary information about each SNP can be displayed in the lower section by clicking on its name in the upper section. A summary of the segregation is obtained by clicking once, and the full SNP data is obtained by clicking twice. For example, if we click once on solcap_snp_c1_10000 of *theta5332.loc* we obtain the following information:

Example 3.1 solcap_snp_c1_10000

Number of individuals: 190 (0 nmiss)

P1 dosage P2 dosage
2 1

SNP Pattern

Dosage	Count	Proportion
0	15	0.07895
1	80	0.42105
2	89	0.46842
3	6	0.03158

Test for double reduction
MLE of alpha: 0.0

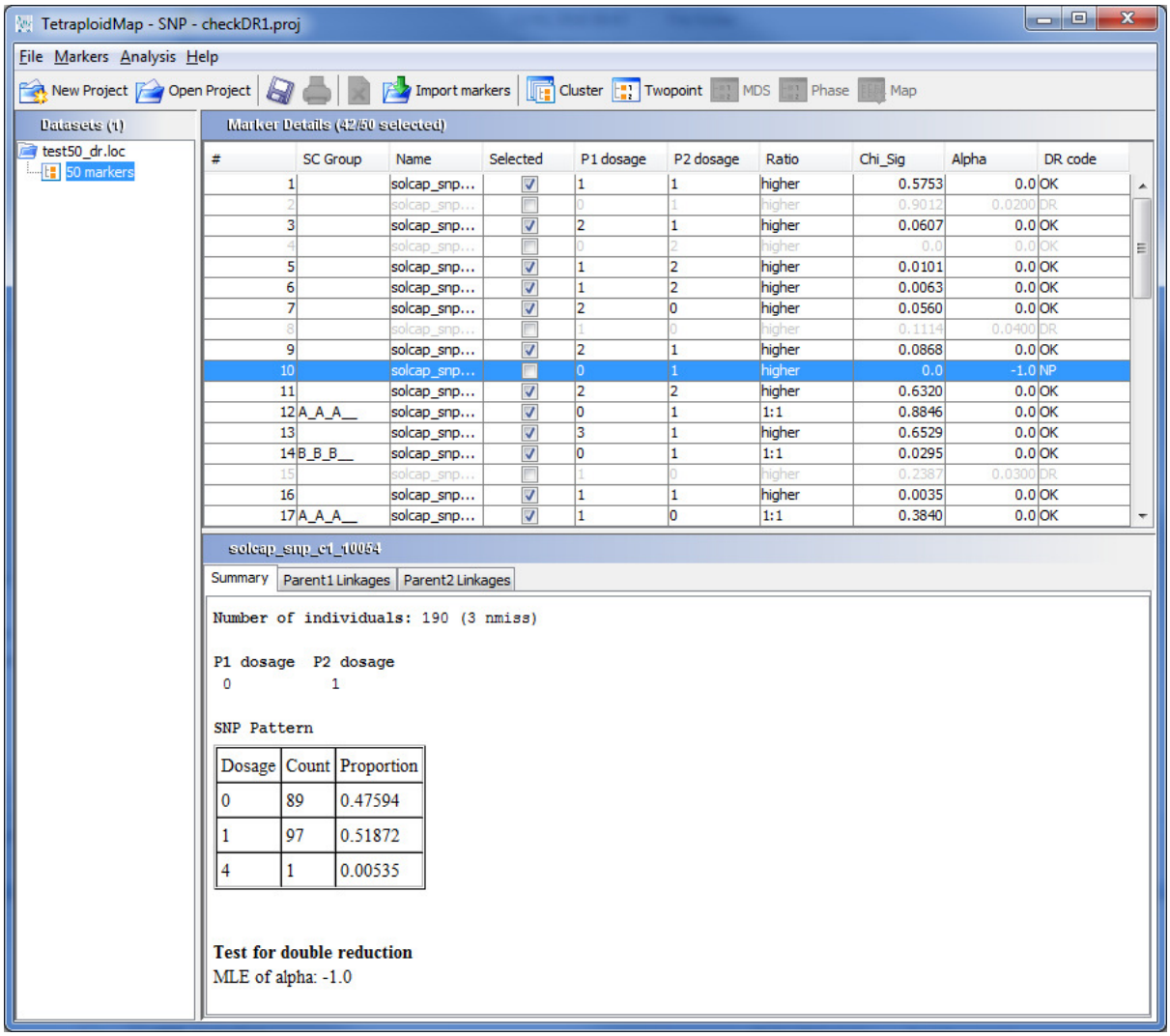
This SNP has dosage 2 in P1 and dosage 1 in P2, and is expected therefore to segregate in a 1:5:5:1 ratio. There are scores for 190 offspring, and no missing values. The chi-squared goodness of fit test for departures from this has p = 0.0607 (in the Chi_Sig column) and so there is little evidence for departures from the expected ratio.

3.2 Correcting problem markers

Markers with DR codes other than OK will be greyed out, and are not selectable at this point. The display below shows three greyed out markers labelled DR, and one NP. Summary statistics for the latter are shown at the bottom, showing that a code of 4 has been included by mistake in a simplex marker. (For markers with the NP code the double reduction coefficient alpha is always showed as -1). Double reduction genotypes and/or impossible genotypes can be replaced by missing values using the tab Markers > Fix DR/NP. If this is done, the DR code will change to FDR or FNP and the

markers can be selected for further use if it is now a good fit (Chi_Sig > 0.001). However **automatic fixing of the NP class of impossible genotypes should be used with caution, especially if it leads to a large number of missing values**. It is better to recheck impossible genotypes, correct them in the original file and restart the analysis.

Example 3.2 Problem markers



3.3 Preliminary grouping analysis

Coupling linkages between simplex SNPs are among the most reliable and informative linkages, and it is useful to know which SNPs are linked in coupling. Cluster analysis provides an easy way to determine this, as one common measure of similarity, the simple matching coefficient, is equivalent to the recombination frequency between pairs of simplex SNPs linked in coupling. The cluster analysis calls an externally developed routine, HCINFLU.F (Cheng and Milligan 1995) and more details of this are given in Appendix A (External routines used). The column in the output labelled

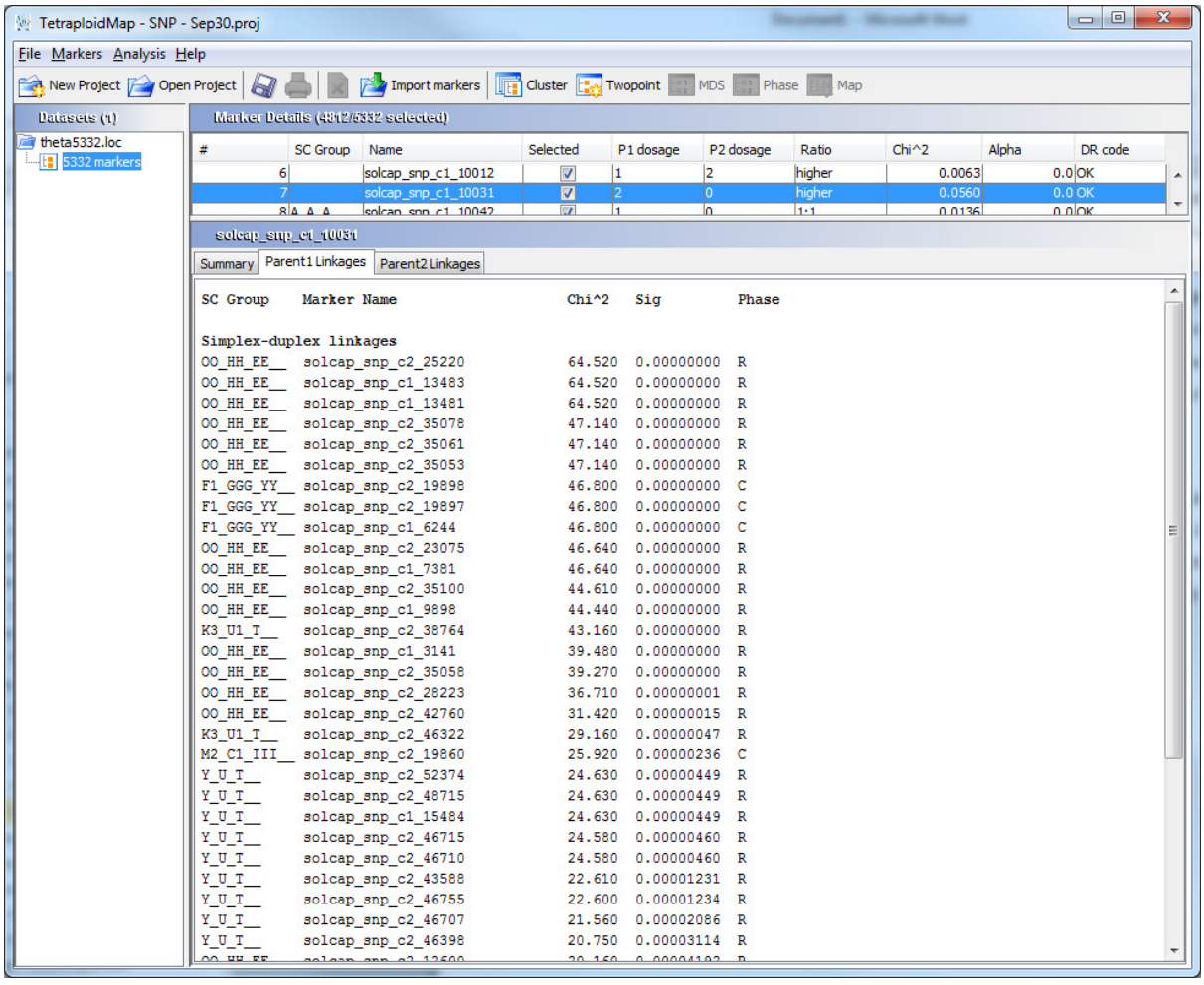
'SC group' summarises how each simplex SNP is clustered by a single linkage cluster analysis at various levels of the recombination frequency. Each code has the form

group at 0.10_group at 0.20_group at 0.30

For example, in the *theta5332.loc* file, loci solcap_snp_c2_25220 and solcap_snp_c1_13483 both have codes OO_HH_EE, while SNP solcap_snp_c1_13663 has code QQ_JJ_EE. The right hand side of the codes shows if the simplex SNPs are separated into groups so that each SNP has a recombination frequency ≤ 0.30 with at least one other, then these three SNPs are placed in the same group, coded 'EE'. If the threshold for separating the groups is decreased to 0.20, these first two SNPs remain together (in a group named 'HH'), while the last is in a separate group, coded 'JJ', as shown by the middle of the code. If the threshold is decreased to 0.10, these markers do not separate further.

3.4 Parent 1 linkages and Parent 2 linkages tabs

The clustering information is very useful to give a preliminary idea of where higher dosage SNPs belong. Consider solcap_snp_c1_10031, which is present in parent 1 only, with dosage 2 and whose segregation fits a 1:4:1 ratio, suggesting that it is duplex (AABB x AAAA). If we click the 'Parent 1 linkages' tab, and then this SNP name, the following output appears in the lower panel:



This shows which simplex SNPs are linked to solcap_snp_c1_10031 by a chi-squared test of independence, the chi-squared statistic and its significance, and the phase of the linkage. These are ordered by the significance of the association. (Note that the phase is based on an approximate test, and should be confirmed later). The significance of the chi-squared statistic is calculated using an external routine AS 170 (Narula and Desu 1981). More details of this are given in Appendix A (External routines used). We see that solcap_snp_c1_10031 is linked in repulsion to several simplex SNPs with code OO_HH_EE/QQ_JJ_EE, and also to SNPs with codes Y_U_T and K3_U1_T. It is also linked in coupling to markers with codes F1_GGG_YY, M2_C1_III and X_T_S, although some of these have lower significance. This suggests that these simplex SNPs lie on homologous chromosomes, two linked in repulsion with the duplex SNP and the others, possibly fragmented, in coupling.

The information shown depends on the type of SNP selected from the upper part of the display. If a simplex SNP is selected, associations are categorised as

- Simplex-duplex linkages (with phase shown)
- Simplex-double simplex linkages (for coupling linkages only)
- Other linkages to simplex markers (phase not shown)

Estimates of repulsion linkages between double-simplex (AAAB x AAAB) and simplex SNPs are very unreliable, and are not shown. Note that SNPs that do not fit well to one of the expected segregation ratios are excluded. The criterion for deciding that a SNP is duplex should be strict to avoid confusion between a duplex SNP and a double simplex SNP (although this is only likely if the parents have been mis-scored). As a consequence, slightly distorted duplex SNPs may be shown in the 'Other linkages' section rather than the 'Simplex-duplex linkages'. If a higher dose SNP is selected in the top section then only linkages to simplex SNPs are shown.

In general, an association is shown in this section if the chi-squared statistic is significant with $p < 0.001$. However when a large number of SNPs are being compared, some significant associations can arise by chance. We recommend caution in inferring linkages based solely on associations at the bottom of the list.

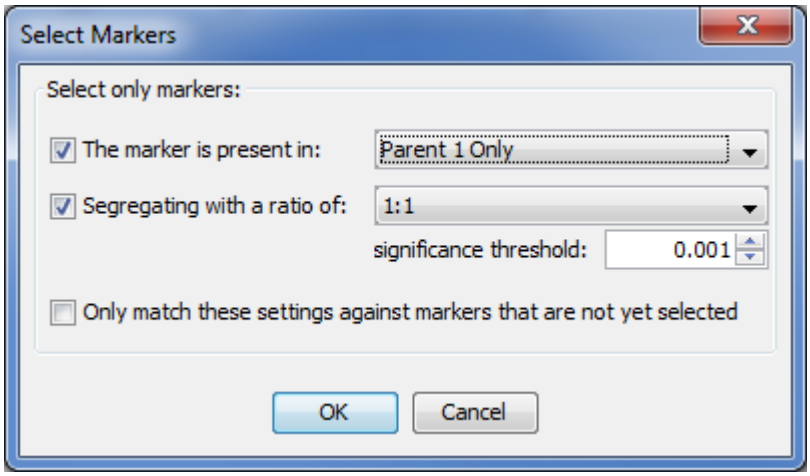
Note on the SC group codes:

The simplex coupling groups from the two parents are coded independently, so no relationship is implied between, say, group 'A' from parent 1 at a recombination frequency of 0.30 and group 'A' from parent 2.

Step 4 Selection of SNPs for further analysis

Once the SNP genotypes have been inferred, and SNPs with incompatibilities or odd segregation ratios have been re-checked and corrected if necessary, the next stage is to select SNPs to use in the linkage analysis. Initially all SNPs for which the Chi_Sig is greater than 0.001 are selected, shown by a tick in the 'Selected' column. A SNP can be excluded by clicking in this box to remove the tick. The program will not permit the selection of markers for which the Chi_Sig is less than 0.001.

In general the set of selected SNPs are a good first choice for creating a linkage map. However it may be useful on some occasions to choose different subsets, for example to create a map for one parent only, or to be more stringent regarding the degree of distortion at which SNPs are excluded. To select a different set of SNPs, it is easiest to use the Markers > Select Markers menu. This brings up the following menu box:



Consider selecting the SNPs from, say, Parent 1 only. The options above select all SNPs present in Parent 1 only, and segregating with a ratio that is not significantly different from 1:1 at a significance level $p < 0.001$ ie dosage (1,0). This selects 1018 of the 5332 SNPs in the theta5332.loc dataset. The other SNPs from this parent only ie dosage (2,0) can then be selected by changing the segregation ratio to 'higher', and checking the bottom box 'Only match these settings against markers that are not yet selected'. This selects further SNPs to give a total of 1438 selected in theta5332.loc. Our experience is that the mapping is more robust to distortions from the expected ratio in 1:1 SNPs than in higher dosage SNPs, and so the significance threshold for higher dosage SNPs can be changed to, say, 0.01 to eliminate more distorted SNPs if required.

We also recommend that SNPs with a (3,1) or a (1,3) configuration are excluded, as these generally have very low information about recombination. These have all been excluded in the analysis below.

Step 5 Cluster analysis of markers to identify linkage groups

The programme tests all pairs of markers to see whether they are segregating independently, using a χ^2 test of independence. The significances are then transformed into distances for the cluster analysis using

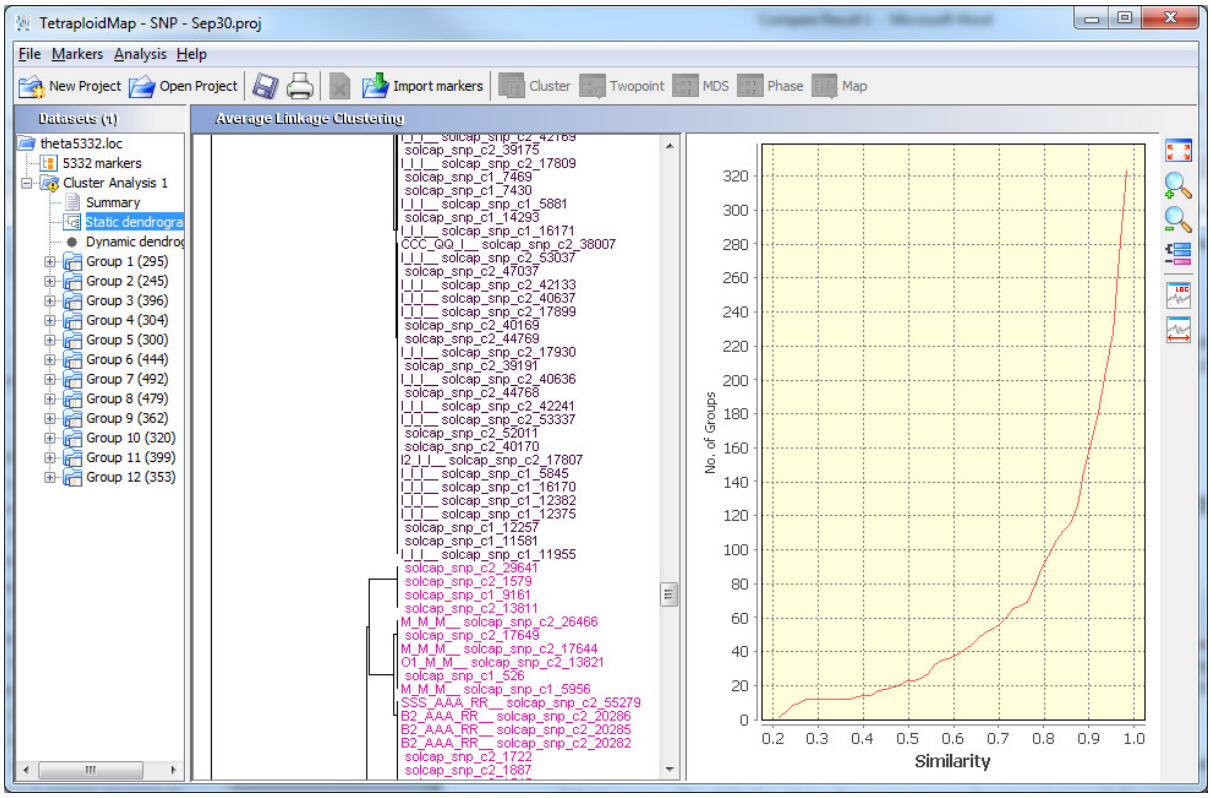
$$\text{distance} = 1 - 10^{-2*\text{significance}}$$

(This stretches the scale for significances < 0.05, which is the region of interest). The significance of the chi-squared statistic is calculated using an external routine AS 170 (Narula and Desu 1981) and the cluster analysis calls an externally developed routine, FastCluster (Müllner 2013). More details of these are given in Appendix A (External routines used). The distances are presented as dendrograms, and the user needs to examine the output to decide where it is appropriate to partition markers into linkage groups.

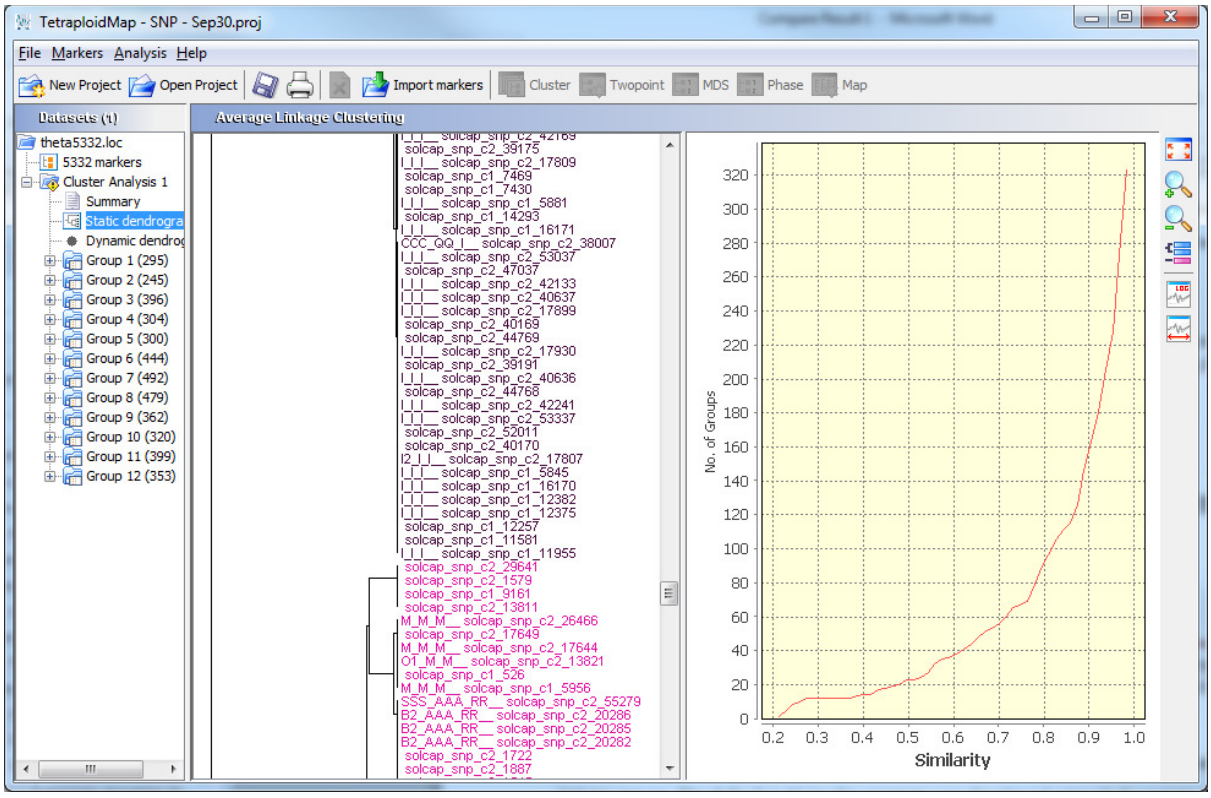
To cluster the selected markers, click on the Cluster tab. This will ask for an expected number of linkage groups. The default is 12 (as the program was developed primarily for potato data, with 12 sets of chromosomes). However the user is likely to run the Cluster stage more than once: for the first analysis of a set of data this has to be an educated guess based on the expected number of linkage groups; for subsequent analyses this should be based on inspection of the dendrograms.

The title 'Cluster Analysis 1' now appears in the left frame. Clicking on this shows subtitles 'Summary', 'Static Dendrogram', 'Dynamic Dendrogram' and 'Group 1' to 'Group 12'. Clicking on 'Static Dendrogram' shows a plot with a dendrogram on the left, and a plot of the number of linkage groups on the right, based on Average Linkage clustering (average distance between objects in the two clusters).

Example 5.1a Group numbers on natural scale



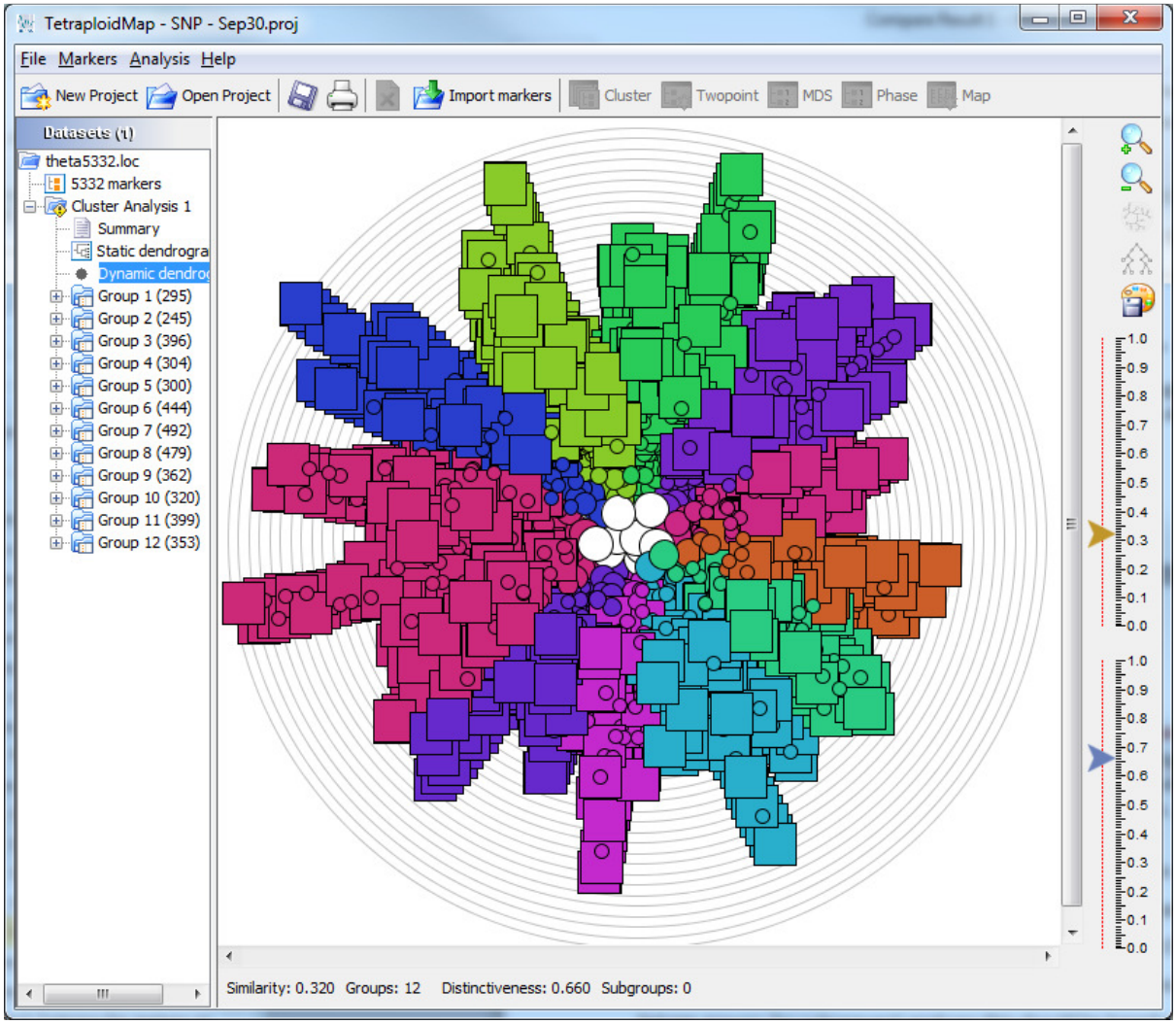
Example 5.1b Group numbers on log scale



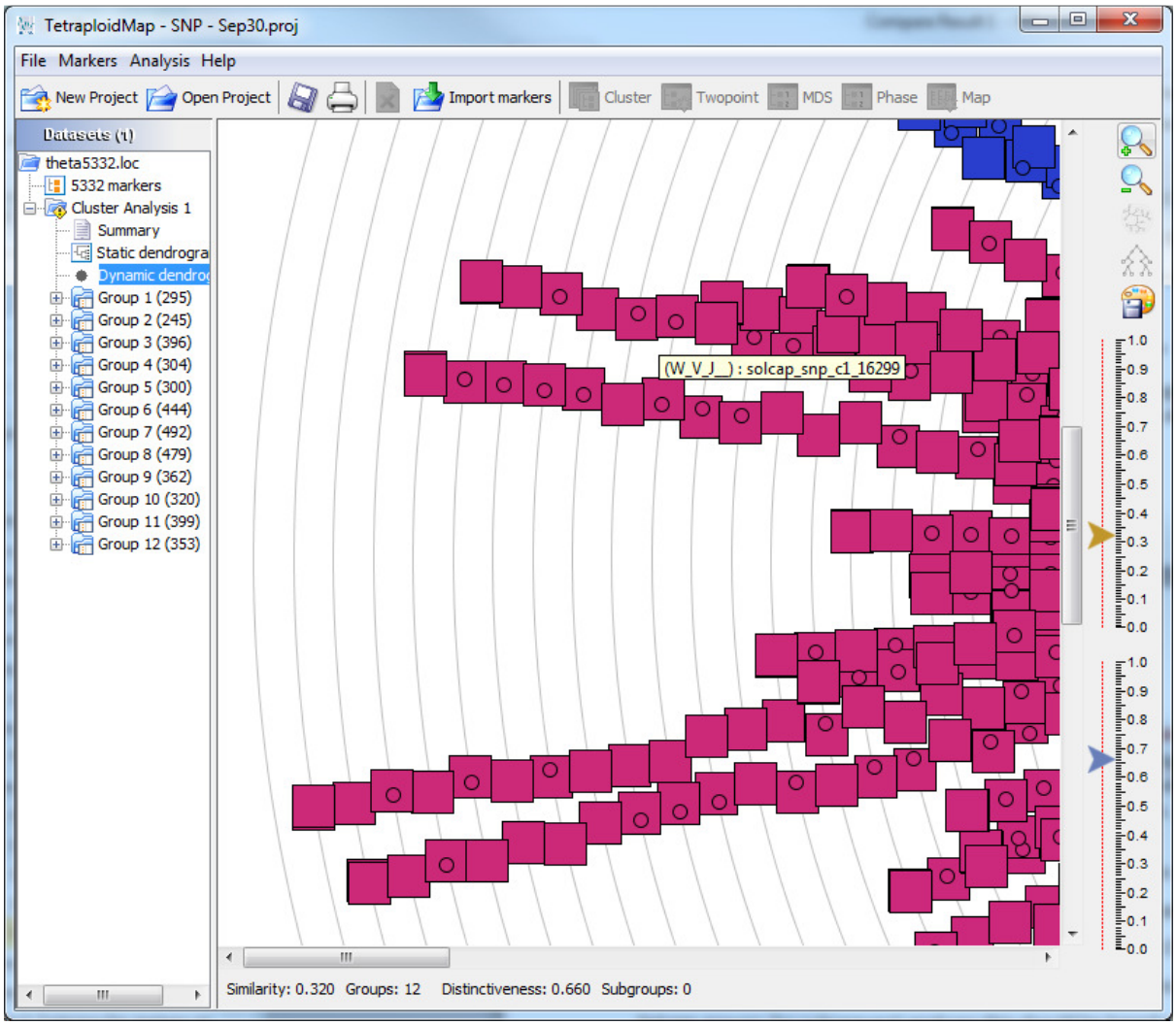
In this case the plot on the right shows a constant number of groups over a wide range of similarities. This may be clearer on the $\log_{10}$ scale, using the second button from the bottom on the right side. One useful way to inspect the dendrograms is to move the mouse from side to side along the Similarity (horizontal) axis. A vertical line is drawn and if the mouse is clicked at any position, the number of clusters with that level of similarity is shown. The marker names are then shown in different colours, depending on the linkage group. The SC Group codes are also shown on the dendrogram to help interpretation – for instance, we see at a glance that several markers coded B_B_B are close together in the dendrogram.

In addition to the static dendrogram, the 'Dynamic Dendrogram' tab has alternative tools to visualise the clustering interactively. The JUNG (Java Universal Network/Graph framework has been used here (O'Madadhain et al. 2003).The dendrogram can be viewed either as a linear view or as a radial view, using the third and fourth buttons on the right of the display. There are two sliders on the right for changing the display. The upper slider changes the similarity level at which the markers are clustered into groups; the lower slider sets a distinctiveness threshold used to identify "weak-edges" which indicate potential nested clusters and outliers. Weak edges are shown as red lines and distinct branches are shown in black. Markers are presented by squares, while circles show the different similarity levels. The smaller the circle at the node between two markers, the higher the similarity between them. Full details of these graphics are described in Vogogias et al. (2016).

Example 5.2a Radial layout, with 12 clusters indicated



Example 5.2b. Enlargement of the left side of the radial plot. Hovering the mouse over a marker (square) identifies it, while hovering the mouse over a circle shows the similarity.



Having decided on a final number of linkage groups, the first cluster analysis can be deleted by clicking on 'Cluster Analysis 1' in the left frame and then either clicking on the red cross in the toolbar, or by using the Analysis – Remove Results menu. While it is not essential to remove the results, it avoids confusion at later stages. The cluster analysis can then be rerun using the final number of linkage groups. This separates the markers into groups, which should represent the linkage groups. However it is not unusual for markers to be placed in an incorrect linkage group at this stage, and the next part of the analysis (section 6) is helpful in determining that.

For the SNPs from the theta5332.loc data, the cluster analysis with 12 linkage groups separates the SNPs into groups of 295, 245, 396, 304, 300, 444, 492, 479, 362, 320, 399 and 353 SNPs. Inspection of the markers by clicking on the group number shows information about the segregation of the markers within that group.

It is useful to inspect each group of marker codes carefully to see how many simplex coupling groups are present before proceeding to order them. If there are more than four groups with completely different codes, this is an indication that the markers have been partitioned incorrectly and this can be examined further by rerunning the cluster analysis for that group only to examine it in detail. Examining the duplex-simplex associations may reveal discrepancies with the grouping by cluster analysis.

A marker can be excluded from the ordering analysis by unselecting it in the 'Selected' column. A marker can also be moved from one group to another by using either Markers > Move Markers to group or CTRL-double click on its name. If a group is changed by moving a marker into or out of it, then it is important to delete any old analyses and rerun them with the new group.

Step 6 Ordering the markers within a linkage group

6.1 Twopoint analysis

The marker ordering step is run for the markers from each linkage group in turn. It first runs a two point analysis, using the Twopoint tab, which calculates the recombination frequency and LOD score for each pair of markers in each possible phase, and infers the most likely phase as the one with the highest likelihood among those with a recombination frequency < 0.5. This twopoint data is then used to calculate the best order for the markers in the linkage group. The code for sorting vectors is taken from Brainerd et al. (1996), and further details are given in Appendix A (External routines used).

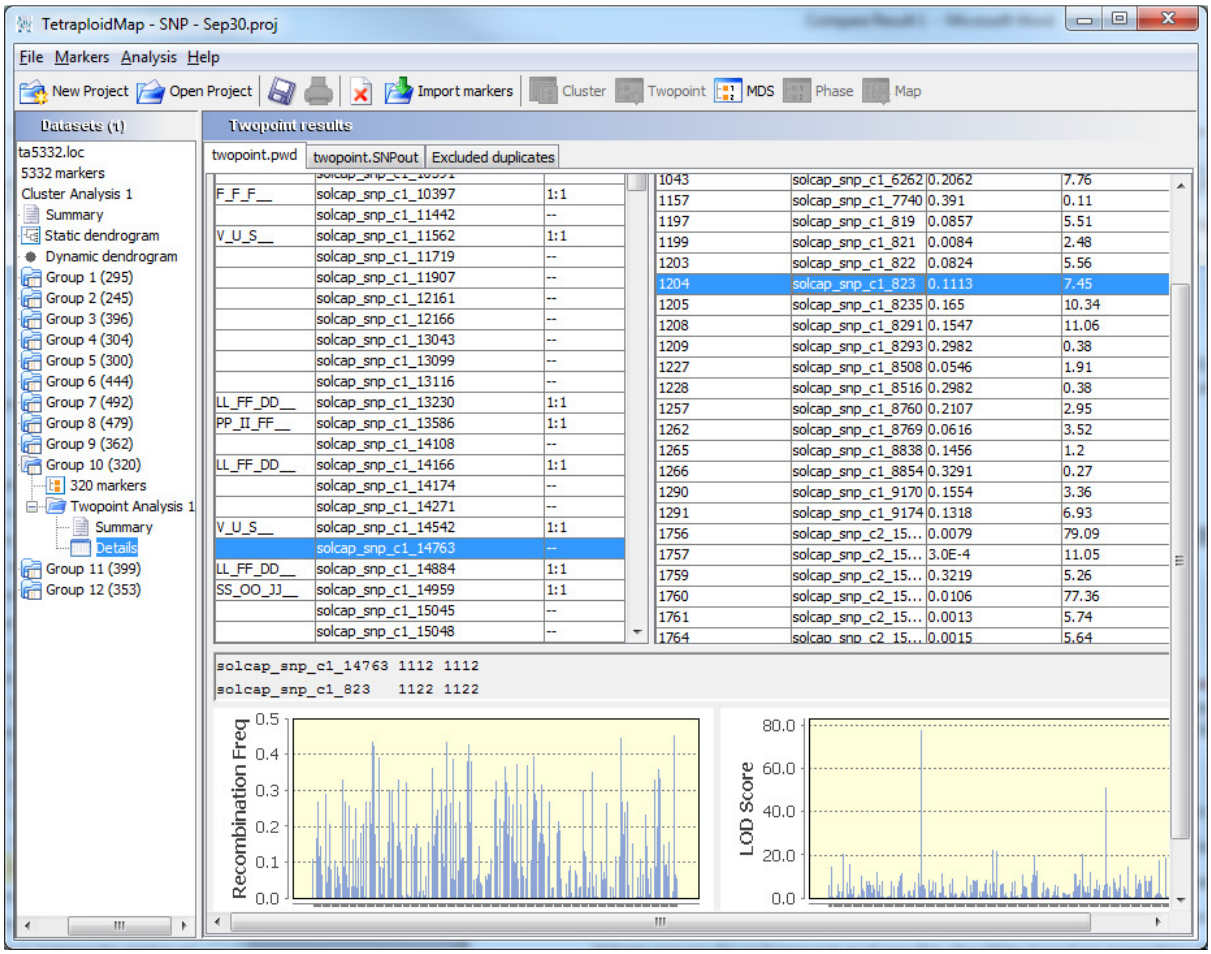
The twopoint analysis will ask if the user wants to remove duplicate and near-duplicate SNPs, which are ones with at most two differences from another locus. The later marker in the file is removed, so it may be helpful to have markers sorted by decreasing quality scores if these are available from the technology, or by increasing number of missing values. It also gives the option to load the full output from the twopoint analysis into the TetraploidMap browser – if this is not needed then the analysis is faster.

The twopoint.pwd tab in the Twopoint Analysis Details lists the unordered markers and gives details of recombination frequencies and LOD scores among all pairs. When a SNP is selected in the left pane, a list of its recombination frequency and LOD score with each other marker is listed in the right pane, and plotted at the bottom. When a second SNP is selected in the right pane then their relative

phases are shown between the upper and lower panes. For example the box below shows a double simplex SNP solcap_snp_c1_14763 in the left panel, linked with recombination frequency 0.1113 and LOD 7.45 to double duplex SNP solcap_snp_823 in the right panel, with the B ('2') allele from the double-simplex SNP linked in coupling to one of the B alleles of the double duplex SNP for each parent.

Full details of the phase comparison analysis are available in the twopoint.SNPout tab, and a list of excluded markers is shown in the Excluded duplicates tab.

Example 6.1. Twopoint output for group 10, showing linkages to highlighted marker solcap_snp_c1_14763



6.2 Ordering using multidimensional scaling

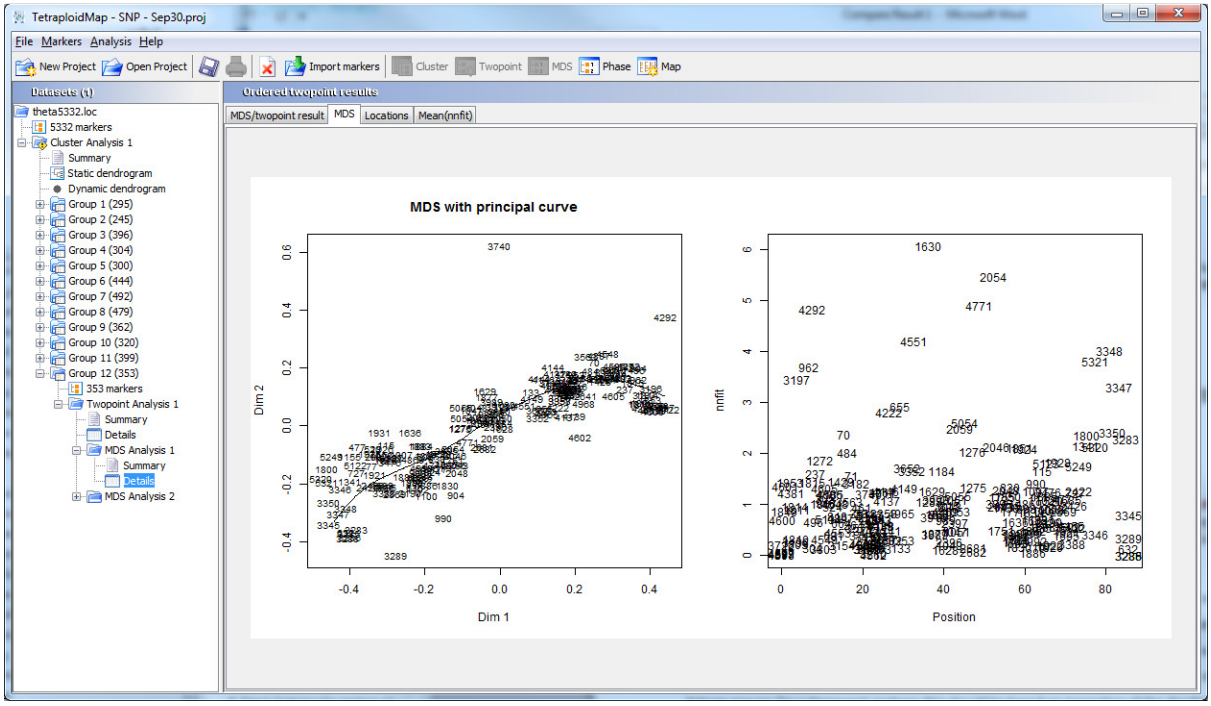
To order a large number of markers rapidly, we have developed an approach based on multi-dimensional scaling. This step of the analysis is programmed in R (R Core Team 2014), using routines in the smacof package (de Leeuw and Mair 2009) and the princurve package (Hastie and Weingessel 2013). Details are given in Preedy and Hackett (2016). The multi-dimensional scaling (MDS) analysis

calculates a configuration in two dimensions by optimising a criterion known as the stress function, which is a function of the recombination frequencies and the squared LOD scores for all pairs of markers, and is similar to the weighed least squares criterion used by the regression mapping approach of the JoinMap software (Van Ooijen 2006). Ideally the configuration will resemble a continuous curve; in practice there are likely to be some outlying markers away from the curve that should be omitted (by unchecking them in the Selected list). This procedure is rerun to obtain a satisfactory ordering. To obtain the final ordering a principal curve is fitted through the marker configuration and the order of markers is given by this.

The MDS analysis is run from the Details tab of the Twopoint Analysis, by clicking on the MDS tab. There is the option to use a two-dimensional or three-dimensional configuration. Each produces an MDS Analysis tab, with further Details within that. The Details button shows four tabs, MDS/twopoint result, MDS, Locations and mean(nnfit). The MDS/twopoint result tab is similar to that in the twopoint analysis, but the markers are now ordered. The diagnostic plots are shown in the MDS tab.

The figure below shows the two-dimensional MDS plots for group 12 (353 SNPs), excluding duplicates to leave 205 SNPs. The left panel shows the marker configuration in two dimensions, while the right panel plots the nearest-neighbour fit (nnfit) of each marker (the sum of the absolute difference between the observed and estimated distance between that marker and the nearest informative neighbours on either side) against the map position.

Example 6.2a 2d MDS analysis, first round



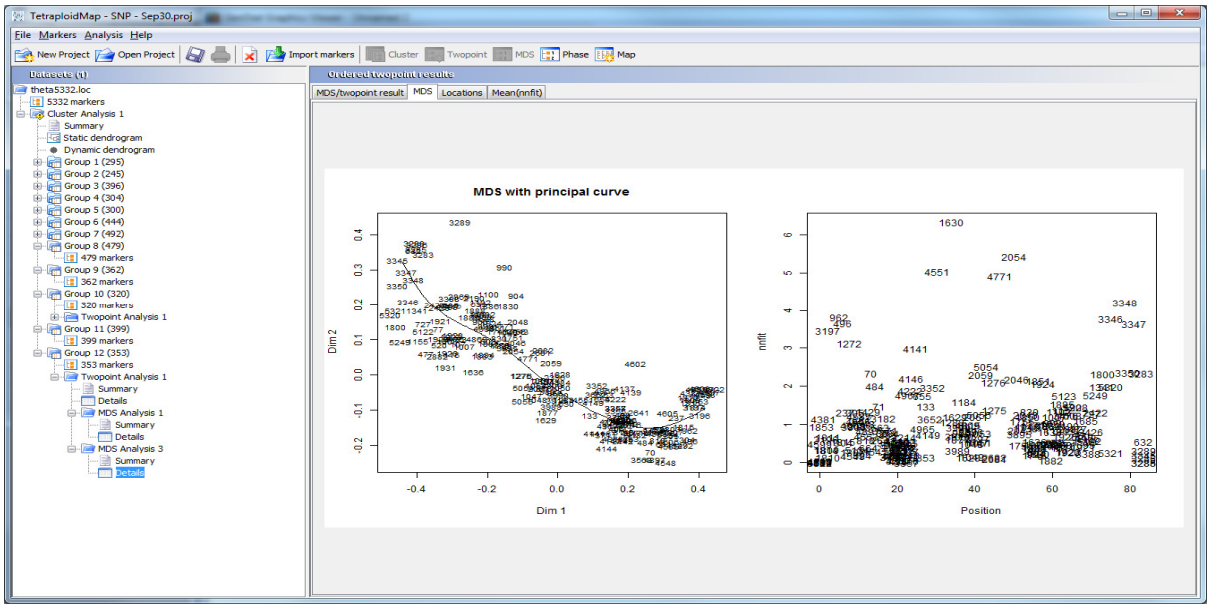
The markers 3740 and 4292 are particularly outlying in the MDS plot. (It is always subjective which points to exclude as outliers, and more than one round of the MDS analysis may be necessary.)
The Locations tab lists the identifier of each marker, the name, the map position and the nnfit, and can be used to identify the names of the outliers. This identifies 3740 = c2_43735 and 4292 = c2_49745. The mean nnfit for this configuration is 1.017.

Inspecting the information about these SNPs in the Group tab shows that neither marker is significantly distorted, which is a common reason for outliers. C2_43735 shows strong associations with simplex markers in group 8, and should be moved there – it has been misplaced by the cluster analysis. The marker c2_49742 is a double-simplex marker, but with few strong linkages detected to other markers from the first parent. Inspection of the original theta scores showed that this was a difficult marker to genotype, and should be dropped. Another reason for outliers may be that they are part of an isolated pair or small group, only linked to each other and not to the rest of the group. These can be revealed by inspection of the twopoint output and the segregation significance.

It is also worth inspecting SNPs with the highest nnfit statistic, although in this case no values are unusually high. (SNPs near the end of the maps can also have high values as they are linked to a small number of other markers, without being erroneous in any way.)

These problem markers were all excluded by unchecking them in the Group 12 tab, and the MDS was run again. (There is no need to repeat the calculation of the twopoint information). The resulting plots are better, with a smaller range on the y axis in each case. The mean nnfit has reduced to 0.988.

Example 6.2b 2d MDS analysis, second round



When the 3d MDS analysis was run on this group (omitting the same outliers), then mean nnfit dropped further to 0.970. No further outliers were apparent. Comparison of the two maps (outside TetraploidSNPMap showed a small rearrangement of markers at the bottom of the chromosome.

The maps can be displayed graphically by clicking the Map tab in the top toolbar. This can be inverted, saved as a .png image or saved as a text file using the buttons on the right side of the Map display.

The selected markers can be exported in a separate file for each chromosome for QTL mapping from the Group 12 (markers) tab by using File > Save Selected.

Note that the MDS analysis cannot be run if there are markers that are nulliplex in both parents, but no markers that have dosage greater than zero in both parents (bridging markers). In the absence of such bridging markers, separate maps for each parent must be created.

Step 7 Inferring phase

Once a satisfactory map has been obtained, the final step of the linkage mapping is the phase analysis. This identifies the alleles that lie on each of the homologous chromosomes of the two parents. The simplex SNPs form a framework for the linkage groups, ideally identifying all four homologous chromosomes for each parent, with the homologue with most simplex SNPs for the first parent being defined as h1, the homologue with the second most simplex SNPs for the first parent being defined as h2, the homologue with the most simplex SNPs for the second parent being defined as h5 etc. If there are fewer than four simplex SNPs linked in coupling then that group is not used as part of the framework. Other SNPs are placed relative to this framework. The phase analysis checks the most likely phase of each non-simplex SNP with respect to each simplex SNP on the framework and designates alleles as in coupling or repulsion relative to each if the LOD score is greater than 5.0. For example, a duplex marker 1122 x 1111 might show a coupling linkage with a simplex marker on h1, a repulsion linkage with a simplex marker on h2, a coupling linkage with a simplex marker on h3 and a non-significant linkage with coupling on h4. This would be coded as 212a. The 'a' shows that this is not a proven linkage, but that there are as many proven linkages as the dosage. If a duplex marker has a coupling linkage with a simplex marker on h1, a repulsion linkage with a simplex marker on h2, and non-significant linkages with simplex markers on h3 and h4 (or if there are no simplex markers on h3 and h4), this is shown as phase 21--.

Any SNP with an unknown or incomplete phase after this step has to have its phase assigned manually, using SNP pairs with as high a LOD as possible and also a high LOD difference between the most likely and the second most likely phase.

The phase analysis is run from the Details tab of the MDS analysis, by clicking on the Phase button. This produces a list of markers in map order, labelled Phase Analysis 1. After information on SC group, name, ratio and map position, columns P1 and P2 carry estimated phases as far as possible. P1d and P2d are the parental dosages, to aid checking.

Some manual editing with a separate text editor will almost always be needed on the phase information, depending on the proportion of simplex markers. A file in the right format can be exported from the Phase Analysis 1 tab using File > Save Selected and saving the output as a text file. This file has the correct column spacing for use as a map file in the QTL analysis, so it is important to maintain this when editing. (The editor Notepad++ is a convenient editor that maintains column

spacing and allows selection of rectangular areas using Alt-Select.) The ‘a’s in the phase codes generally can be replaced globally with a 1, but it is recommended to do this after other phases have been entered. Before exporting the data, the map can be inverted (by clicking on the Location tab) or the homologous chromosomes for each parent can be reordered (by clicking on the P1 or P2 tabs).

In editing the phase information, we recommend starting in the middle of the chromosome and working out to each end in turn, as there may be less information here. Information for filling in the phases can be found in the MDS Analysis/Details tab, under MDS/twopoint results. More detailed information can be found in the .SNPpwd file in the tpmap folder, which can be read into Excel and filtered to look at the markers of interest.

For group 1, the simplex markers with codes ZZ_Y_X, MM_GG_X and CC_Y_X are designated as homologue h1, R3_AAA_UU and the smaller linked sets SSS_AAA_UU and YYY_AAA_UU form h2, FFF_RR_MM and V1_RRR_MM form h3 and WW_NN_JJ forms h4. (The shared second and third parts of the codes show that these markers are linked in coupling with recombination frequency between 0.1 and 0.2.) For parent 2, the codes are HH_DD_AA (h5), KK_GG_DD (h6), V3_J1_OOO and M2_J1_OOO (h7). There are no simplex markers in a sufficiently large cluster on h8 to be included as part of the simplex framework.

Consider the incomplete section below, from group 1:

Example 7.1a Initial output from phase program

```
Solcap_snp_c1_11319         47.01  1122 1-2-    2    2
solcap_snp_c2_54992         47.30  -121 1111    2    0
solcap_snp_c2_37932         47.85  a112 1111    1    0
solcap_snp_c2_15535         49.16  1111 ----    0    1
solcap_snp_c2_54993         49.18  1111 ----    0    1
solcap_snp_c2_15528         50.42  a2a1 2-1-    1    2
solcap_snp_c2_15606         50.49  --2- aa2a    2    1
solcap_snp_c2_15536         51.02  1111 ----    0    1
solcap_snp_c2_15530         51.06  a2a1 2-1-    1    2
solcap_snp_c1_10315         51.14  1111 122a    0    2
solcap_snp_c2_15605         51.81  ---- --1-    1    2
solcap_snp_c1_4990          53.82  a2a1 1111    1    0
```

solcap_snp_c2_54992 at 47.3cM has dosage 2 in parent 1, and it is already filled in that there is a '2' allele on h3 while h2 and h4 are both allele '1'. The phase of this marker is completed as 2121 x 1111.

Inspection of the MDS/twopoint results shows that this marker is closely linked to solcap_snp_c2_15606, with recombination frequency 0.0648 and LOD 13.66. This is a duplex x simplex marker and the alleles are both in coupling in parent 1, so the phase of solcap_snp_c2_15606 is completed as 2121 x 1121.

Solcap_snp_c2_15535, solcap_snp_c2_15536 and solcap_snp_c2_54993 are simplex markers from P2 linked in coupling on a currently unidentified chromosome. These are linked to some markers in duplex configurations from P2, but in repulsion with all located markers. We will provisionally designate this as h8. The duplex-duplex marker solcap_snp_c1_11319 has one allele linked in coupling with these, so its phase is completed as 1122 x 1122. The simplex-duplex markers solcap_snp_c2_15528 and solcap_snp_c2_15530 are closely linked in coupling with each other and in repulsion with solcap_snp_c2_15535, solcap_snp_c2_15536, solcap_snp_c2_54993 and solcap_snp_c1_11319, and so their phases are both completed as 1211 x 2211. Finally solcap_snp_c2_15605 shows a weaker linkage in coupling with solcap_snp_c2_15528 and solcap_snp_c2_15530 and so its phase is also completed as 1211 x 2211.

This is the completed section, with the 'a's replaced by 1s. The final two dosage columns are also deleted before the map is ready for use with the QTL mapping program.

Example 7.1b Completed phase section

```
solcap_snp_c1_11319         47.01  1122 1122
solcap_snp_c2_54992         47.30  2121 1111
solcap_snp_c2_37932         47.85  1112 1111
solcap_snp_c2_15535         49.16  1111 1112
solcap_snp_c2_54993         49.18  1111 1112
solcap_snp_c2_15528         50.42  1211 2211
solcap_snp_c2_15606         50.49  2121 1121
solcap_snp_c2_15536         51.02  1111 1112
solcap_snp_c2_15530         51.06  1211 2211
solcap_snp_c1_10315         51.14  1111 122a
solcap_snp_c2_15605         51.81  1211 2211
solcap_snp_c1_4990          53.82  1211 1111
```

If it is unclear about the best phase for a small number of markers, these can be dropped from the map initially and mapped onto it using their theta scores.

The manual editing of the phase to complete the map is a lengthy procedure. However if the ordering of the map is subsequently changed, there is no need to restart the phasing from the beginning. The phased markers can be reordered for an alternative map. Mapping the theta scores (or other quantitative measures) from which the marker dosages were derived provides an excellent check on the map and the inference of phase. More details about this can be found in Appendix B. We strongly recommend that these scores are mapped before starting the analysis of other phenotypic traits.

Module 2: QTL interval mapping

The QTL interval mapping module is accessed by starting the program again and choosing the SNP-QTL option.

Step 1: Preparation of data files for QTL interval mapping

The QTL mapping analysis is conducted for each linkage group separately. Three data files are needed: trait data, genotype data and map data.

1.1 Trait data

The trait data is placed in a text file, usually with extension .qua. The first line of the file should contain the number of traits, and the second line shows the names of the traits. The trait data is then given, with each trait in a separate column, preceded by a numerical identifier for each offspring. Missing values can be coded as -99.0.

Example 1.1 Beginning of trait file *demo.qua*, with 12 traits.

```
12
     myield    mht   msize  mshape   mdm  mfry4  mfry10  mmat mfb4   mtb%   sqpcn   fc
  4  10.525  52.89   4.217   4.531 22.48   3.154 5.569  4.977  1.5   82.5    3.038  1
  5  10.551  56.62   4.632   4.935 19.95   3.753 6.474  3.442  2.5   85      3.441  2
  6   7.812  56.99   4.515   3.807 19.95   2.993 6.693  5.976  1     50.71   2.323  1
  8   8.403  58.61   4.304   5.115 21.19   2.716 5.912  5.334  6.5   28.36   3.651  2
 10   6.496  45.62   3.294   3.960 23.41   5.365 7.311  2.705  3.5   97.5    2.39   2
 11   9.672  43.66   4.430   3.831 22.64   4.122 6.556  4.025  5.5   68.66   4.997  -99.0
 12   9.687  49.84   4.159   3.980 22.88   4.625 7.075  2.657  2.5   90      3.645  2
 15   9.882  52.08   4.069   4.747 23.32   3.800 6.121  2.966  7     70      3.341  1
etc
```

1.2 SNP genotype data

The SNP genotype data is in the same format as the original data. During the linkage analysis process, separate files should have been created with the SNP genotype data for each linkage group, and these are best used here. If markers in this file are not included on the map, they will be ignored for this analysis.

1.3 Map data

The map file is a text file, usually with extension .map. The first line of the .map file contains the number of mapped markers. The rest of the file contains the markers, in map order, and details of

position and phase. The format of each line is name, position phase code for parent 1 (4 digits), phase code for parent 2 (4 digits). The phase codes are obtained from the phase analysis that forms the last part of the linkage analysis, and some editing is usually necessary to complete these codes. The phase codes are described in detail there. At present the column spacing needs to be exactly as in the example, with the decimal point of the position in column 31 and the phase codes starting in columns 36 and 41. (The output files produced by the phase tab of the linkage module has the correct column spacing).

Example 1.1 Beginning of map file with phase information

```
119
c2_23643                        0.00  1111 1121
c2_23741                        0.49  1221 2111
c2_33545                        1.21  2111 1122
c2_23735                        2.15  1111 1121
c2_23669                        2.52  1221 2111
c2_23728                        3.43  1221 2121
c2_23829                        3.98  1221 2111
c2_23780                        4.01  1111 1211
c2_23832                        4.31  1111 1121
c2_23828                        4.36  1221 2111
c2_23740                        4.73  1221 2111
c2_23834                        4.93  1112 1211
etc
```

Step 2: Starting TetraploidSNPMap and creating a QTL project

The QTL mapping module of TetraploidSNPMap is started by starting the program and then choosing the SNP-QTL option.

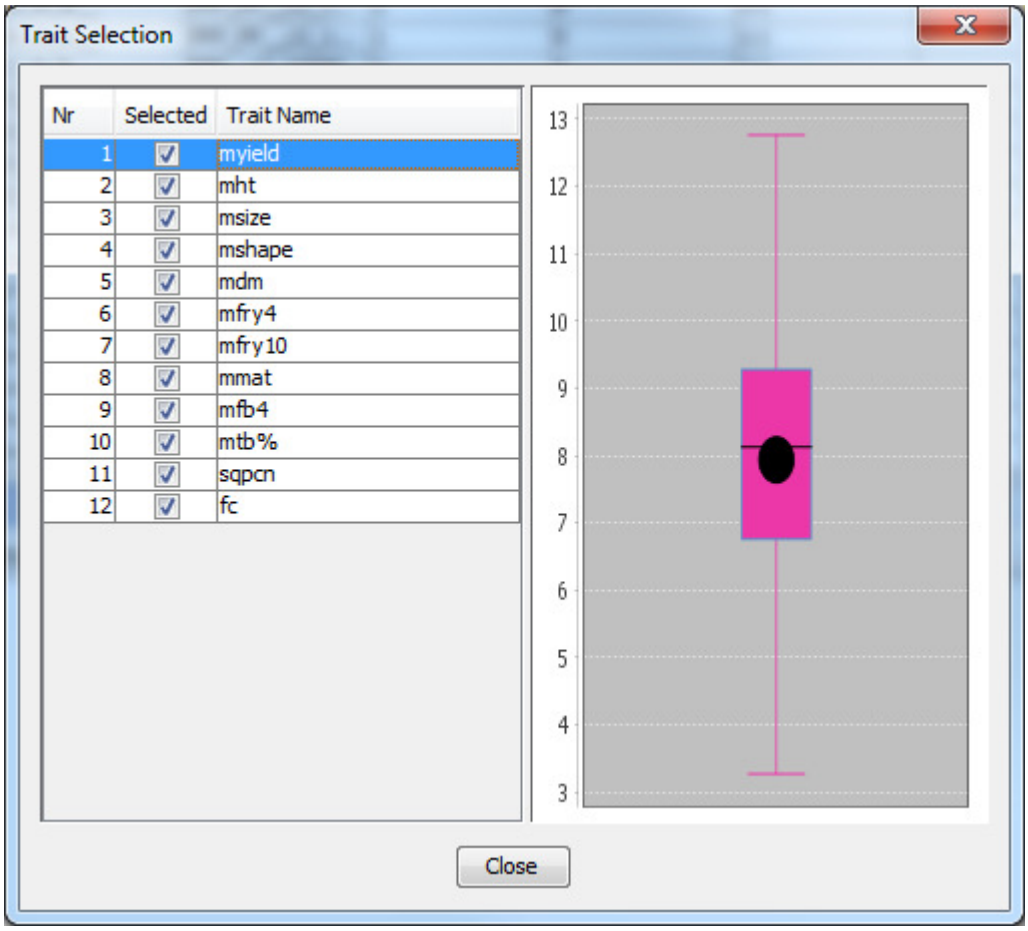
The first step is to define a project by clicking on the New Project tab (or open an existing project with the Open Project tab). This will bring up a dialogue box where the user can select the directory containing the data, and define a project there. The default name is MyProject.proj. It is helpful to choose a project name that identifies this as a QTL project rather than a linkage mapping project. Once this has been created, the next step is to read in the three files described above.

The files should be imported in the following order:

1) Marker genotype data, using the Import markers tab or File > Import > SNP dataset.
   This reads in the SNP data and displays segregation ratios, linkages etc as for the linkage analysis

2) Trait data, using the menu Trait data > Associate trait data. There is an option to use all traits in the file (default) or to exclude some at this stage. Clicking on a trait name will show a boxplot of the distribution. The central line shows the median of the trait distribution, the length of the box shows the inter-quartile range and lines go to the maximum and minimum values. The circle is centred at the trait mean.

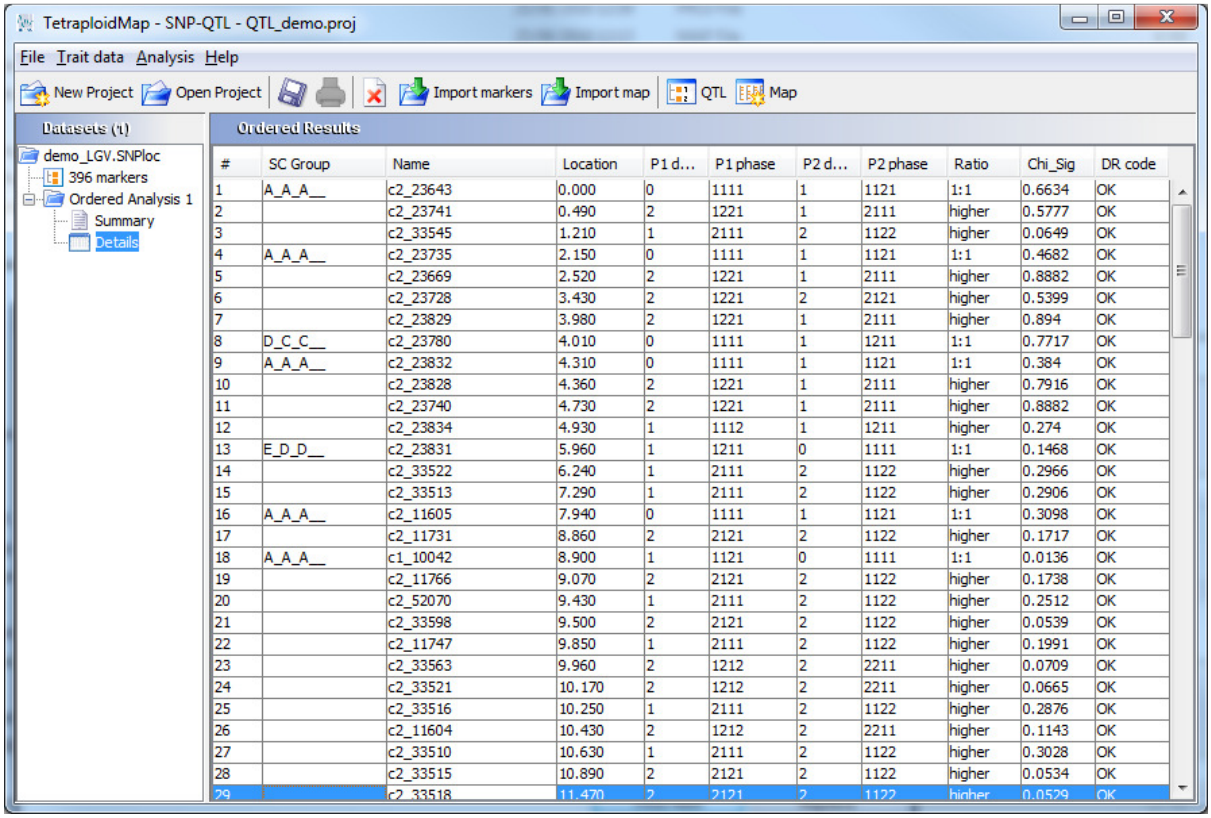
Example 2.1 List of traits with boxplot of myield



3) Map file, using the Import map tab or File > Import > map file

Once the three files have been imported the markers are sorted into map order, omitting any that are not placed on the map.

Example 2.2 Markers sorted into map order



#	SC Group	Name	Location	P1 d...	P1 phase	P2 d...	P2 phase	Ratio	Chi_Sig	DR code
1	A_A_A__	c2_23643	0.000	0	1111	1	1121	1:1	0.6634	OK
2		c2_23741	0.490	2	1221	1	2111	higher	0.5777	OK
3		c2_33545	1.210	1	2111	2	1122	higher	0.0649	OK
4	A_A_A__	c2_23735	2.150	0	1111	1	1121	1:1	0.4682	OK
5		c2_23669	2.520	2	1221	1	2111	higher	0.8882	OK
6		c2_23728	3.430	2	1221	2	2121	higher	0.5399	OK
7		c2_23829	3.980	2	1221	1	2111	higher	0.894	OK
8	D_C_C__	c2_23780	4.010	0	1111	1	1211	1:1	0.7717	OK
9	A_A_A__	c2_23832	4.310	0	1111	1	1121	1:1	0.384	OK
10		c2_23828	4.360	2	1221	1	2111	higher	0.7916	OK
11		c2_23740	4.730	2	1221	1	2111	higher	0.8882	OK
12		c2_23834	4.930	1	1112	1	1211	higher	0.274	OK
13	E_D_D__	c2_23831	5.960	1	1211	0	1111	1:1	0.1468	OK
14		c2_33522	6.240	1	2111	2	1122	higher	0.2966	OK
15		c2_33513	7.290	1	2111	2	1122	higher	0.2906	OK
16	A_A_A__	c2_11605	7.940	0	1111	1	1121	1:1	0.3098	OK
17		c2_11731	8.860	2	2121	2	1122	higher	0.1717	OK
18	A_A_A__	c1_10042	8.900	1	1121	0	1111	1:1	0.0136	OK
19		c2_11766	9.070	2	2121	2	1122	higher	0.1738	OK
20		c2_52070	9.430	1	2111	2	1122	higher	0.2512	OK
21		c2_33598	9.500	2	2121	2	1122	higher	0.0539	OK
22		c2_11747	9.850	1	2111	2	1122	higher	0.1991	OK
23		c2_33563	9.960	2	1212	2	2211	higher	0.0709	OK
24		c2_33521	10.170	2	1212	2	2211	higher	0.0665	OK
25		c2_33516	10.250	1	2111	2	1122	higher	0.2876	OK
26		c2_11604	10.430	2	1212	2	2211	higher	0.1143	OK
27		c2_33510	10.630	1	2111	2	1122	higher	0.3028	OK
28		c2_33515	10.890	2	2121	2	1122	higher	0.0534	OK
29		c2_33518	11.470	2	2121	2	1122	higher	0.0529	OK

If the phase and parental dosage have any errors or incompatibilities there will be an error message, and the markers will be shown in red. These must be fixed before any further analysis is attempted.

The output of the ordering contains the columns

Order

SC Group

Name

Location (cM)

P1 dosage

P1 phase, coded as 1 or 2 (see details in Phase section of the linkage analysis)

P2 dosage

P2 phase

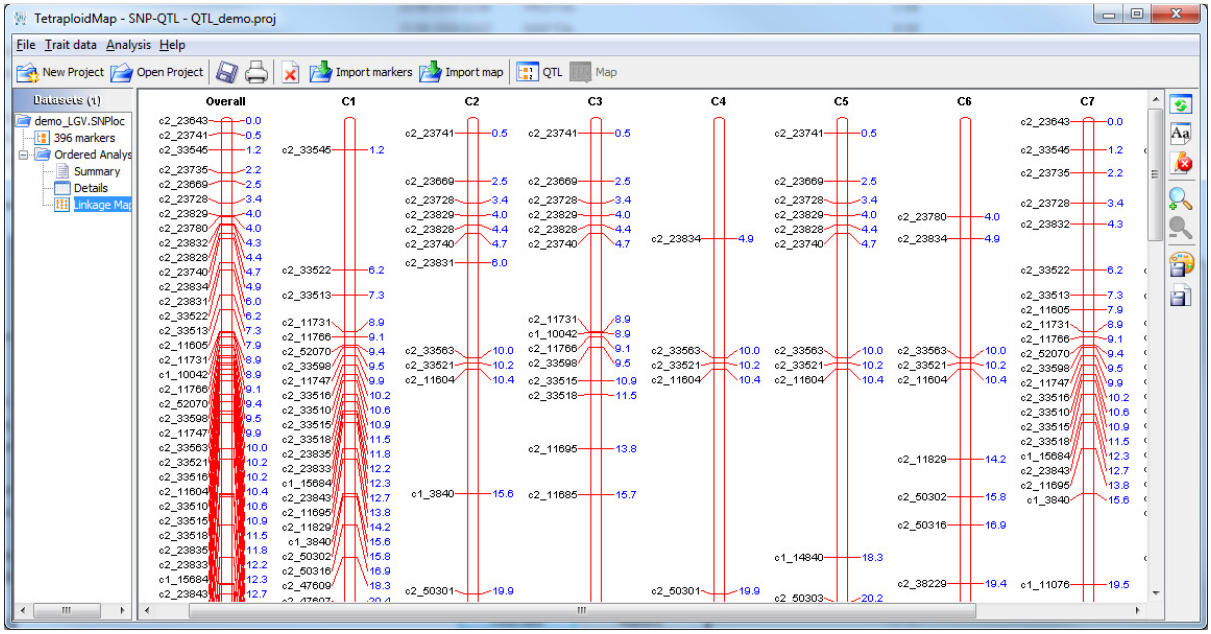
Ratio (1:1 or higher)

Chi_Sig, the significance of the departure from the expected ratio

DR code (as for the linkage analysis)

The map can be display graphically by clicking on the Map tab.

Example 2.3 Display of linkage maps



This shows the overall map, and how the markers are placed on the separate homologous chromosome. Buttons on the right side of the display give options to zoom in and out, to invert the map, to remove the overall chromosome and to save the image either as a graphical .png file, or as a text file that can be used with other map drawing software such as MapChart (Voorrips 2002)

Step 3: QTL interval mapping

To run the QTL mapping analysis, click on the QTL tab. The program runs the QTL interval mapping analysis for each selected trait, considering markers from both parents. The program first estimates the probabilities of each possible QTL genotype from the parental phases and the parental and offspring dosages, using a Hidden Markov Model (HMM). This is shown as Phase 1 of the calculation in the Progress bar. It then interpolates QTL probabilities at a 1 cM grid of positions along each chromosome. The analysis then models the trait as an additive function of the QTL allele effect on each of the eight homologous chromosomes for each position on the grid. The additive model has the form:

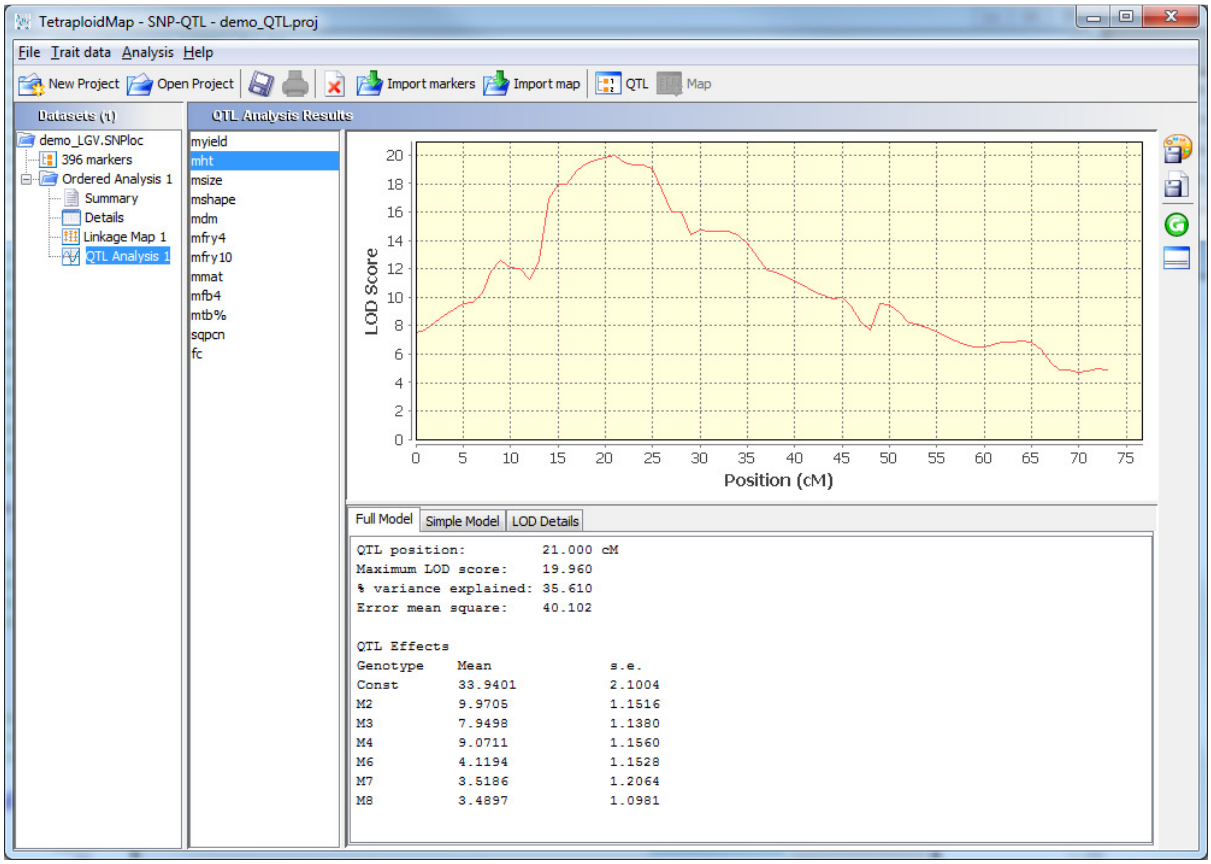
$$Y_G = \mu + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_6 X_6 + \alpha_7 X_7 + \alpha_8 X_8$$

where X_2 is an indicator (0/1) variable for chromosome 2 etc. There are constraints on this model as $X_1+X_2+X_3+X_4 = 2$ and $X_5+X_6+X_7+X_8 = 2$: therefore the coefficients express differences from the trait value of an individual with chromosomes 1 and 5. For a detailed discussion of the model see Hackett et al. (2013, 2014).

This model is fitted by regression of the trait values on the QTL genotypes, weighted by the genotype probabilities from the HMM. The linear regression uses the lsq.f90 module of Alan Miller, which is an upgraded version of Applied Statistics algorithm AS 274 (Miller 1992, 2002). Further details are given in Appendix A.

When the analysis has finished a panel with a list of trait names is produced, and by clicking on a trait name the output is displayed.

Example 3.1 QTL profile for height on linkage group V



This shows the LOD profile for mean height (mht) along linkage group V. Information about the maximum of the LOD profile is summarised in the lower panel, under the Full Model tab. Here the maximum LOD is 19.96, at a position 21cM along the chromosome. The QTL effects are presented as an overall constant and estimated effects for each chromosome, as in the additive model above, with their associated standard errors. For this trait all coefficients are positive and significantly different from zero, and those relating to chromosomes from the first parent (M2, M3 and M4) are larger than those from the second parent. The top two buttons on the right side enable the LOD profile to be exported either as a graphics (.png) file or as a text file.

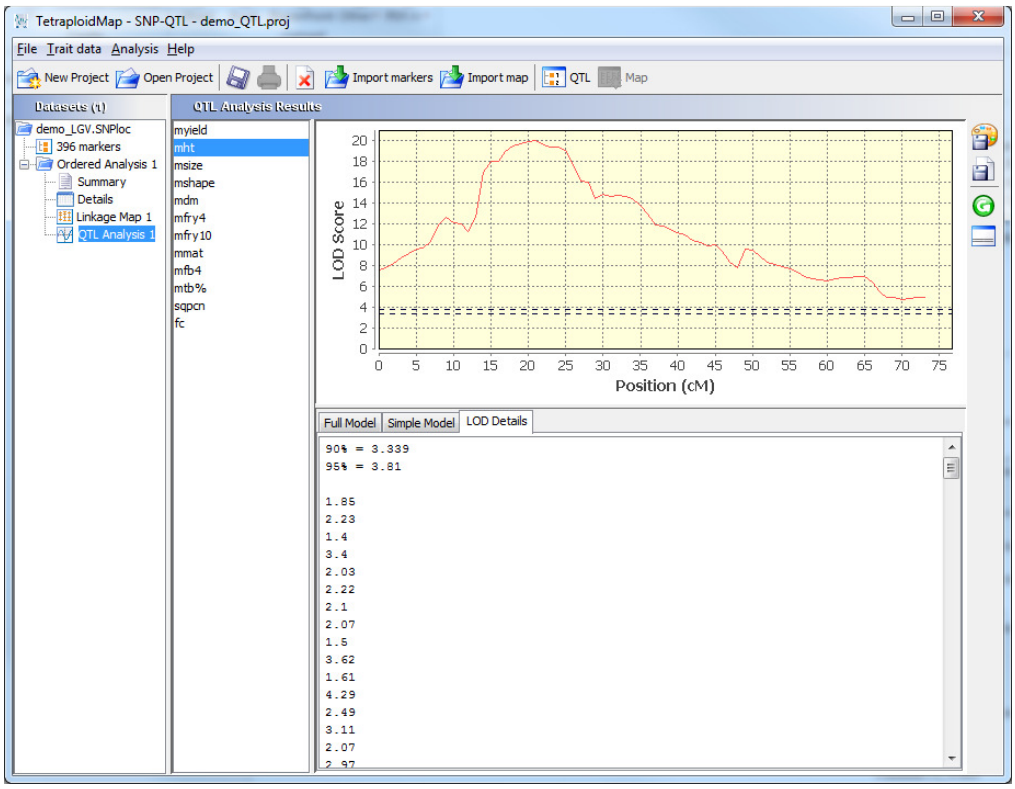
Step 4: Run permutation test to derive threshold

If there appears to be a significant QTL, this can be assessed using a permutation test. This is called using the bottom button on the right side of the QTL analysis display.

The program will ask for a number of permutations (between 100 and 1000), and a seed to start the randomisation. If the user wants to obtain a genome-wide permutation threshold, rather than a chromosome-wide permutation threshold, we recommend using the same seed for each chromosome so that the same permuted values are analysed for each chromosome. Then the maximum LOD for each trait across all the chromosomes can be derived and the desired genome-wide percentage point can be estimated (outside TetraploidSNPMap). The routine for random permutations is derived from code by Green (1963, 1977) and the code for sorting vectors is taken from Brainerd et al. (1996). The code for estimating percentage points nur.f90 is from the web site of Alan Miller. Further details of all these are given in Appendix A (External routines used).

The 90% and 95% permutation thresholds are shown as blue horizontal lines in the upper panel, and are listed under the LOD Details tab. The maximum LOD scores from each permutation are also given in the LOD Details tab: these can be copied and saved to a plain text file or spreadsheet if required.

Example 4.1 QTL profile for height on linkage group V, with thresholds from permutation test



Step 5: Run test for simpler models

The last step is to test whether a simpler model, for example a simplex QTL Qqqq x qqqq, is a good model to the data. This analyses the trait means for each of the 36 possible genotypes at the location with the highest LOD score, which are estimated from the probabilities of each genotype there. **It is important to note that the variance explained in this section (Adj R^2) is based on these 36 means and not on the same units as for the Full Model tab, and the figures should not be compared with each other.**

The simple models are in various series:

1. Simplex models eg H1

   This tests for a simplex allele in one of the parents ie H1 is for a QTL genotype on homologue 1 of parent 1: Qqqq x qqqq. H5 etc is used for a simplex allele for parent 2: qqqq x Qqqq

2. Duplex as a codominant variate eg V12

   This tests for a duplex model with additive effects of the Q allele. V12 represents a QTL genotype QQqq x qqqq with the QTL genotypes qqqq, qqqQ, qqQQ having means m, m+a, m+2a respectively. Note that some combinations are equivalent (eg V12 and V34) so only one of these is tested.

3. Duplex as a codominant factor eg F12

   As for 2, but the QTL effects are not assumed to be additive. All three classes have different means m1, m2, m3.

4. Dominant duplex allele eg D12

   As for 2 and 3, but this assumes Q is dominant and tests for two genotype categories, qqqq and Q-qq.

5. Double-simplex as a codominant variate eg S15

   S15 represents a double-simplex genotype Qqqq x Qqqq. The QTL genotypes and means are additive, as for 2 above.

6. Double-simplex as a codominant factor eg FS15

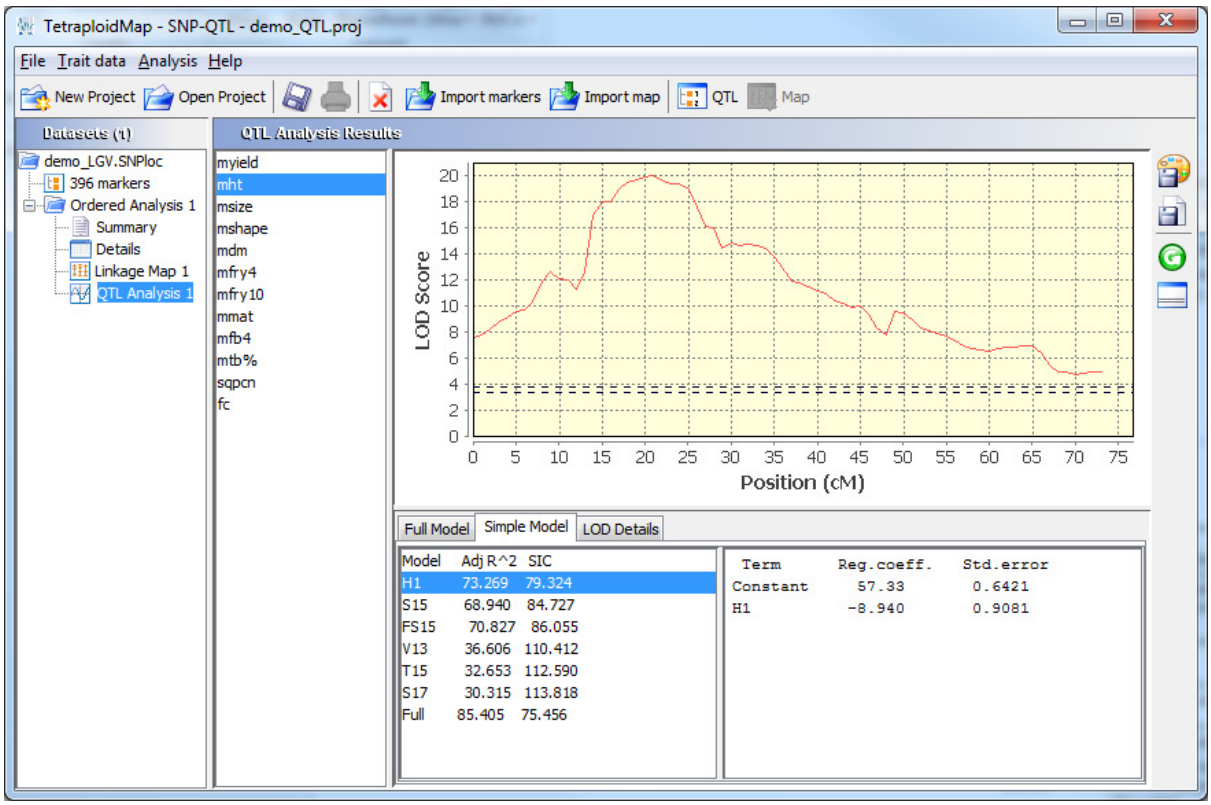
This tests for a double-simplex genotype but with three means m1, m2, m3 as for 3 above.

7. Dominant double-simplex eg T15

This tests for a double-simplex genotype as for 5 and 6 but assumes Q is dominant and tests for two genotype categories, qqqq and Q-qq, as for 4 above.

The simple models program is called using the second button from the bottom on the right side of the QTL display. The results are shown under the Simple Model tab. The left side lists the six best simple models, ordering by the Schwarz information criterion (SIC), and the last row, labelled Full, shows the corresponding figures for the Full additive model.

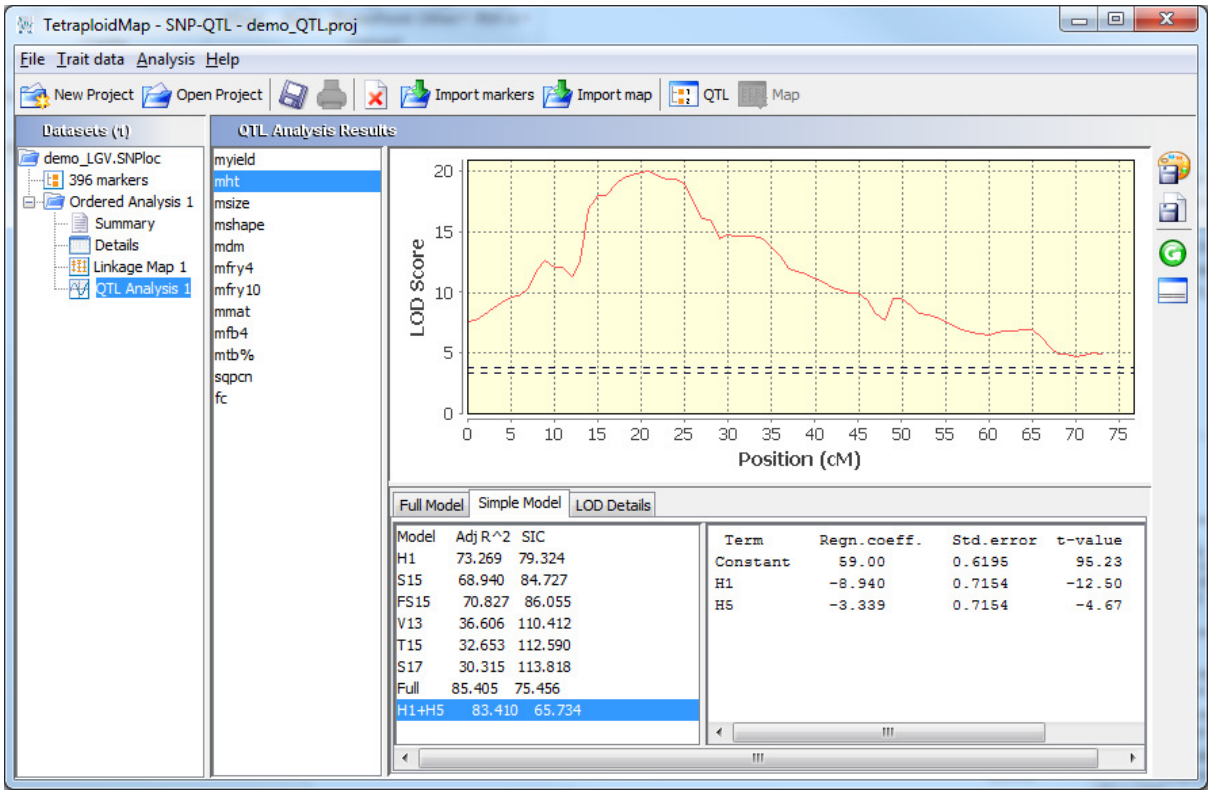
Example 5.1a Simpler models for height on linkage group V



By clicking on a model, its parameters are shown on the right side. For height, the best simple model is H1, and has SIC=79.3, greater than that of the Full model (ie a worse fit than the Full model). H1 has a coefficient of -8.9, implying that lines with the H1 allele are on average 8.9cm shorter than those without H1. However an inspection of the other strong models and the coefficients of the Full model in the Full model tab suggests that both H1 and H5 are important.

Clicking on the simple model tab for a second time enables the user to try simple additive effects of just two chromosomes. If we select 1 and 5 here, a better model is obtained:

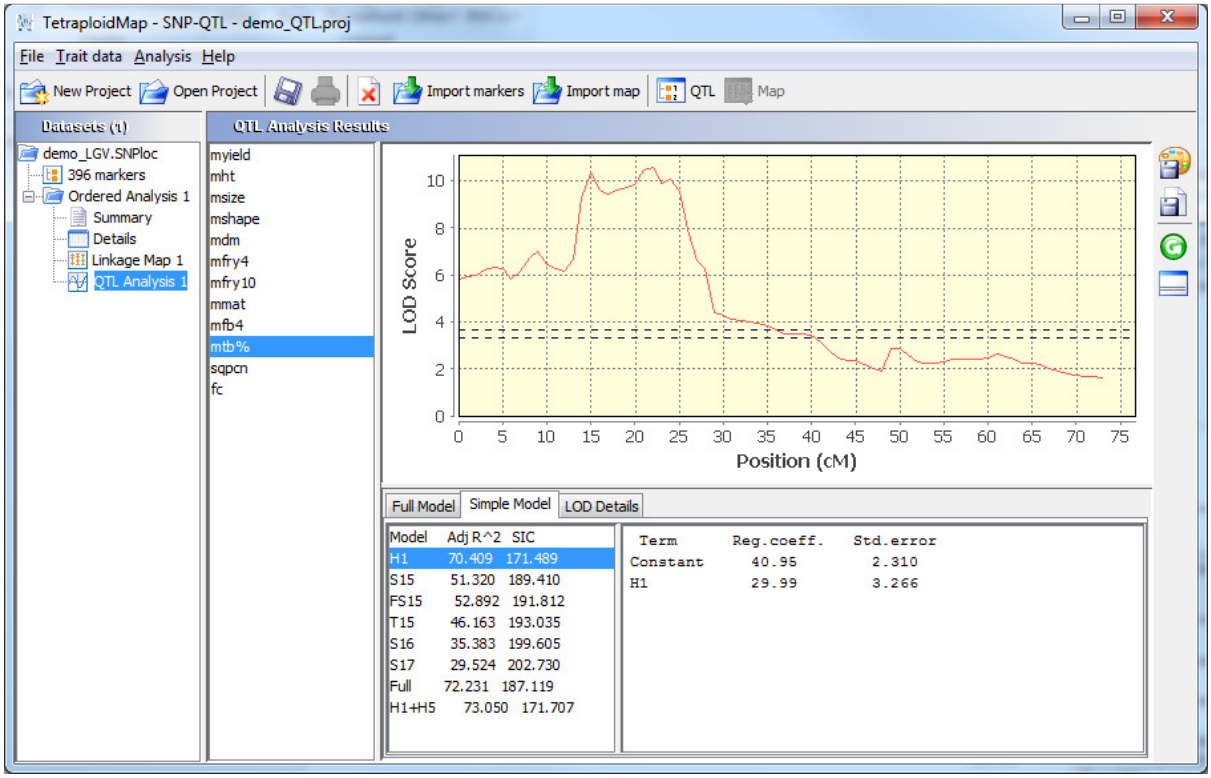
Example 5.1b Further simple models for height on linkage group V



The SIC for model H1+H5 has decreased to 65.7, which is now an improvement on the Full model, and the variance explained for the 36 means is 83.4%, close to that for the full model of 85.4%. These simple additive models can be run more than once.

A second example QTL on this linkage group is for tuber blight and foliage blight. Both map to a similar position, 21-22 cM, but for these the simple model H1 (of a simplex QTL on h1) is clearly better than the full model and adding an effect of h5 does not cause a significant improvement.

Example 5.2 Simple models for tuber blight % on linkage group V



For some traits, there will be no clear simple model and it best to use the Full model.

Acknowledgements

The authors would like to acknowledge the financial support of the Scottish Government Rural and Environment Science and Analytical Services Division (RESAS). The data distributed with the paper were collected by John Bradshaw, Glenn Bryan and colleagues at the James Hutton Institute. The first version of this software was tested at a workshop organised by David Douches, Joe Coombs and their colleagues at Michigan State University in 2014. The final version has been tested by Alicia Massa and Norma Constanza Manrique-Carpintero of Michigan State University, Washington da Silva and Jaebum Park of Cornell University and Molla Mengist of Teagasc, Carlow, Ireland. We are very grateful to all the above for their helpful feedback.

Appendix A: External routines used

Cluster analysis within initial simplex grouping and within phase analysis

The Fortran code used for cluster analysis is adapted from the routine HCINFLU.F by Glenn Milligan, revised by Richard Cheng, of Ohio State University. Details can currently be found on the web site http://www.pitt.edu/~csna/Milligan/readme.html (checked August 2016) and in their publication Cheng and Milligan (1995) and references within that.

We adapted this code to avoid production of graphical displays and calculation of their influence measure.

This routine is used in the following parts of TetraploidMap:

Initial processing, to cluster the simplex coupling markers

Phase analysis, to recluster selected markers

Cluster analysis of all selected markers

The cluster analysis of all selected markers into linkage groups uses the R fastCluster package by Daniel Müllner. Details can be found on his website http://danifold.net/fastcluster.html (checked September 2016) and in the publication Müllner (2013).

Chi-squared probability calculations

The Fortran code for chi-squared probability calculations is adapted from the Applied Statistics routine AS 170 (Narula and Desu 1981), obtained from the StatLib site ftp://sunsite.univie.ac.at/mirrors/lib.stat.cmu.edu/apstat/.index.html hosted by the Department of Statistics at Carnegie Mellon University. As currently shown this routine calls further routines AS 147 (Lau 1980) and AS 245 (Macleod 1989).

We adapted this code so that it only calculated central chi-squared probabilities.

This routine is used in the following parts of TetraploidSNPMap:

Initial processing, to assess the significance of associations between pairs of markers

Clustering, to assess the significance of associations between pairs of markers

Twopoint analysis, to assess the goodness of fit of the recombination model to the frequency of pairs of markers.

Sorting routine

The code for sorting vectors is taken from Brainerd et al. (1996), using the swapping routine on page 98 and the sorting routine on page 153. The latter was modified to return a vector of the permutation used.

This routine is used in the following parts of TetraploidSNPMap:

Twopoint analysis, to identify the phase with the maximum likelihood

Phase analysis, identify the largest simplex coupling groups

QTL permutation test, to sort the LOD scores

QTL simple model analysis, to identify the models with the lowest BIC

Linear regression

The weighted linear regression analysis is carried out using the lsq.f90 module of Alan Miller, which is an upgraded version of Applied Statistics algorithm AS 274 (Miller 1992, 2002). This can be found on the web site http://jblevins.org/mirror/amiller/ (checked August 2016).

This routine is used in the following parts of TetraploidSNPMap:

QTL interval mapping

QTL permutation test

Testing for simple models

Random permutations

Random permutations of trait data for the permutation test are performed using the routine of Green (1963, 1977), rewritten into Fortran 90.

This routine is used in the following parts of TetraploidSNPMap:

QTL permutation test

Calculation of percentage points

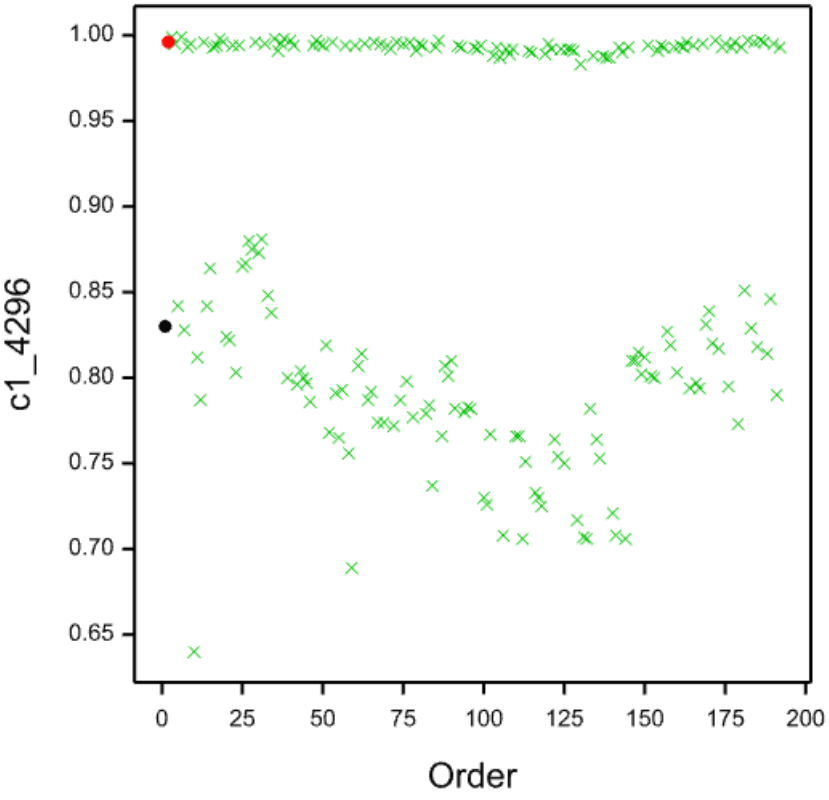
The calculation of percentage points uses code extracted from the statistics subroutine of the program nur.f90 by Alan Miller. This can be found on the web site http://www.it.uu.se/edu/course/homepage/algpar1/ht02/Quicksort_examples/nur.f90 (checked February 2016).

This is used for deriving LOD thresholds in the permutation test program.

Appendix B: Dosage estimation

Hackett et al. (2013) give details of the clustering procedure used to obtain genotype dosages from the theta scores output by Genome Studio. This analysis assumes random chromosomal segregation, so that any offspring showing a double-reduction phenotype would be grouped with the closest dosage obtainable without double reduction. Other authors have also worked on dosage estimation in tetraploid species, using more flexible models. Further developments in this area are likely, both in terms of sequencing technology and methods for estimating dosages. We have therefore chosen to start the analysis in TetraploidSNPMap once the dosages have been called. Two publically available programs for this are FitTetra (Voorrips et al. 2011) and SuperMASSA (Serang et al. 2012; Mollinari and Serang 2015).

A feature of the potato dataset described in Hackett et al. (2013) was that plots of the theta scores against order of the samples on the plates could show a spatial trend for some markers. For example:



The parents are shown as red and block circles. This resembles a simplex marker, with two categories, but the trend will make this difficult to genotype. We recommend that graphical displays such as the above are used to check the theta scores before estimating the dosages, and that the theta scores are mapped as quantitative traits as a check on the map before further QTL mapping.

In order to do this, the theta scores should be prepared as for any other file of trait data, with an initial column of offspring numbers and the theta scores for that chromosome in columns, with their names. In the output, the significant coefficients of the QTL model should correspond to the phasing of the marker. For example, a duplex SNP on homologues h3 and h4 of parent 1, and so with phase 1122 x 1111, will have significant coefficients for M3 and M4, and all other coefficients should be close to zero. If the theta scores of the parents are close to 0.5 and 0.0, the coefficients of M3 and M4 will be positive, while if the theta scores are close to 0.5 and 1.0, the coefficients of M3 and M4 will be negative. Because of the parameterisation, a simplex allele on h1 (or h5) will appear as significant coefficients of a similar size for all of M2, M3 and M4 (or M6, M7 and M8).

For inspecting the results for a large number of traits, it may be more convenient to use the output file (.out) in the tpmap folder, where the QTL positions and parameter estimates can be found and copied into Excel spreadsheets etc for exploration.

References

Brainerd, W.S., Goldberg, C.H. and Adams, J.C. (1996). *Programmer's Guide to Fortran 90*, Third Edition. Springer.

Cheng, R. and Milligan, G.W. (1995). Hierarchical-clustering algorithms with influence detection. *Educational and Psychological Measurement* 55, Issue 2: 237-244.

de Leeuw, J. and Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software* 31: 1-30.

Green, B.F. (1963). *Digital computers in research.* New York: McGraw-Hill.

Green, B.F. (1977). FORTRAN subroutines for random sampling without replacement. *Behavior Research Methods & Instrumentation* 9: 559.

Hackett, C.A., Mclean, K. and Bryan, G.J. (2013). Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PLoS ONE* 8, e63939.

Hackett, C.A., Bradshaw, J.E. and Bryan, G.J. (2014). QTL mapping in autotetraploids using SNP dosage information. *Theoretical and Applied Genetics* 127: 1885-1904.

Hastie, T. and Weingessel, A. (2013). princurve: Fits a Principal Curve in Arbitrary Dimension. R package version 1.1-12. http://CRAN.R-project.org/package=princurve

Lau, C. (1980). Algorithm 147: A simple series for the incomplete gamma integral. *Journal of the Royal Statistics Society. Series C (Applied Statistics)* 29: 113-114.

Macleod, A.J. (1981). Algorithm AS 245: A robust and reliable algorithm for the logarithm of the gamma function. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 38: 397-402.

Miller, A.J. (1992). Algorithm AS 274: Least squares routines to supplement those of Gentleman. . *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41: 458-478.

Miller, A.J. (2002). *Subset selection in regression*, Second Edition. Chapman & Hall / CRC Press.

Mollinari, M. and Serang, O. (2015). Quantitative SNP genotyping of polyploids with MassARRAY and other platforms. *Plant Genotyping: Methods and Protocols*, pp 215-241.

Müllner D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python, *Journal of Statistical Software* 53: no. 9, 1–18, URL http://www.jstatsoft.org/v53/i09/

Narula, S.C. and Desu, M.M. (1981). Algorithm 170: Computation of probability and non-centrality parameter of a non-central chi-squared distribution. *Journal of the Royal Statistics Society. Series C (Applied Statistics)* 30: 349-352.

O'Madadhain, J., Fisher, D., White, S. and Boey, Y. (2003). The JUNG (Java Universal Network/Graph) framework. *University of California, Irvine, California.*

Preedy KF and Hackett CA (2016) A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. *Theoretical and Applied Genetics* DOI 10.1007/s00122-016-2761-8.

R Core Team (2014) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org

Serang, O., Mollinari, M. and Garcia, A.A.F. (2012). Efficient exact maximum a posterior computation for Bayesian SNP genotyping in polyploids. *PLoS ONE* 7: e30906.

Van Ooijen, J.W. (2006). *JoinMap ® 4; Software for the calculation of genetic linkage maps in experimental populations.* Wageningen; Netherlands: Kyazma B.V

Vogogias A, Kennedy J, Archambault D, Smith VA, Currant H (2016) MLCut: exploring multi-level cuts in dendrograms for biological data. In *EG UK Computer Graphics and Visual Computing* (eds Cagatay Turkay and Tao Ruan Wan). The Eurographics Association DOI 10.2312/cgvc.20161288.

Voorrips, R.E., Gort, G. and Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* 12: 172.

Voorrips, R.E. (2002). MapChart: Software for the graphical presentation of linkage maps and QTLs. *The Journal of Heredity* 93: 77-78.