# TetraploidSNPMap: software for linkage analysis and QTL mapping in autotetraploid populations using SNP dosage data

Christine A Hackett1, Bram Boskamp1, Thanasis Vogogias12, Katharine F Preedy1, Iain Milne3

1 Biomathematics and Statistics Scotland, Dundee DD2 5DA, UK

2 Edinburgh Napier University, Edinburgh EH10 5DT, UK (current address)

3 Information & Computational Sciences, James Hutton Institute, Dundee DD2 5DA, UK

Christine.Hackett@bioss.ac.uk

Bram.Boskamp@bioss.ac.uk

T.Vogogias@napier.ac.uk

Katharine.Preedy@bioss.ac.uk

iain.milne@hutton.ac.uk

Running Title: TetraploidSNPMap software

## Abstract

An earlier software application of ours, TetraploidMap for Windows, enabled linkage analysis and quantitative trait locus (QTL) interval mapping to be carried out in an experimental cross of an autotetraploid species, using both dominant markers such as amplified fragment length polymorphisms (AFLPs) and codominant markers such as simple sequence repeats (SSRs). The size was limited to 800 markers, and QTL mapping was conducted for each parent separately due to the difficulties in obtaining a reliable consensus map for the two parents. Modern genotyping technologies now give rise to datasets of thousands of single nucleotide polymorphisms (SNPs), and these can be scored in autotetraploid species as SNP dosages, distinguishing among the heterozygotes AAAB, AABB and AABB, rather than simply using the presence or absence of an allele. The dosage data is more informative about recombination and leads to higher density linkage maps. The current program, TetraploidSNPMap, makes full use of the dosage data, and has new facilities for displaying the clustering of SNPs, rapid ordering of large numbers of SNPs using a multidimensional scaling analysis, and phase calling. It also has new routines for QTL mapping based on a hidden Markov model, which use the dosage data to model the effects of alleles from both parents simultaneously. A Windows-based interface facilitates data entry and exploration. It is distributed with a detailed user guide. TetraploidSNPMap is freely available from our GitHub repository.

## Features and Functionality

### Background

Linkage analysis and QTL mapping in experimental crosses of autopolyploid plant species are more complicated than in diploid species, and so methodology and software have developed more slowly. However, several important plant species are autotetraploid (including potato, alfalfa, some fruits such as blueberry and blackberry and various grass species. Our previous software, TetraploidMap for Windows (Hackett et al. 2007), carries out linkage analysis and QTL mapping in experimental populations of autotetraploids, using dominant markers such as amplified fragment length polymorphisms (AFLPs), scored as presence/absence of a single allele, and codominant markers such as simple sequence repeats (SSRs), scored as presence/absence of several alleles. TetraploidMap for Windows is, however, limited to handling a maximum of 800 markers overall, and for each linkage group, the ordering analysis limited to 50 markers as accurate ordering algorithms for this type of data became too slow with larger numbers. It was also found to be difficult to align maps from the two parents with confidence, so QTL mapping was carried out for each parent separately.

New genotyping technologies now allow generation of much larger sets of genotypic data, based on single nucleotide polymorphisms (SNPs). This work was motivated by the Illumina Infinium technology and the Infinium 8300 potato SNP array (Felcher et al. 2012). This platform gives measurements of allele intensities and different methods have been proposed to estimate dosage from these (Hackett et al. 2013; Voorrips et al. 2011; Serang et al. 2012). The theory for using the dosage data to cluster SNPs into linkage groups and to estimate recombination frequencies and LOD scores has been developed by Hackett et al. (2013). These authors ordered the SNPs using the commercial software JoinMap 4 (van Ooijen 2006) and its regression mapping approach on a pairwise dataset of recombination frequencies and LOD scores for each linkage group. They applied this approach to a full-sib potato population, with 190 offspring, from a cross between a processing clone 12601ab1 and the cultivar Stirling and derived linkage maps for each of the 12 potato chromosomes, with between 74 and 152 SNPs mapped and further SNPs allocated to approximate bins, so that overall 3839 SNPs were assigned to genetic locations.

Recently Preedy and Hackett (2016) have explored the use of multi-dimensional scaling as a rapid method for ordering genetic markers to form high-density linkage maps. This has been tested on various populations, including simulated tetraploid populations, and estimated maps have shown high correlations with the original marker order. This approach is sufficiently rapid to investigate the effect on the map of excluding problematic markers.

Hackett et al. (2013) also presented theory for QTL mapping using dosage data, and this was explored further by Hackett et al. (2014). The high density maps enable QTL mapping to be carried out for both parents together. The initial interval mapping fits a model of additive effects of all homologous chromosomes on a 1 cM grid of positions, using a regression approach weighted by the QTL genotype probability and tests for significance using a permutation test. A range of simple models can then be checked to see if they are a good fit to the data.

The TetraploidSNPMap software combines these routines into a user-friendly program. It has two modules, for linkage map estimation and QTL mapping.

### Linkage Map Estimation Module

The steps to estimate the linkage map are the same as for TetraploidMap for Windows: analysis of single marker segregation, clustering into linkage groups, estimation of recombination frequency and LOD score between all pairs of SNPs within a linkage group, ordering based on pairwise data and inference of SNP phase.

The possible SNP genotypes are AAAA, AAAB, AABB, ABBB and BBBB and these are coded as 0, 1, 2, 3, 4 i.e. the number of ‘B’ alleles. SNP dosage data from the offspring and parents are read in from a plain text file, and the segregation ratios in the offspring are checked for compatibility with the genotype frequencies expected from those parents under a model of random chromosomal segregation (Hackett et al. 2013) using a chi-square test of goodness of fit. For example, if the parents are AAAB and AABB, the offspring are expected to be AAAA, AAAB, AABB, ABBB in a ratio 1:5:5:1 under random chromosomal segregation. For this configuration the offspring genotype BBBB could occur through double reduction and this is flagged for attention, as are offspring genotypes that cannot be obtained from the given parents. An initial cluster analysis of the simplex SNPs is also performed to identify which are linked in coupling. Alphabetical codes display this information at subsequent stages.

The SNPs that show compatibility with the expected ratios are then grouped by a hierarchical clustering algorithm using average-linkage clustering. A similarity measure derived from a chi-square test for independent segregation is used for calculating the pairwise distances in the dissimilarity matrix. The coding of the simplex SNPs aids interpretation. In addition to a static left-to-right representation of the dendrogram, a second visualisation component displays the tree using a space efficient radial layout. A top-down tree layout is also supported to highlight the size, the length and the shape of the branches that correspond to linkage groups.

The user can interact with the dendrogram and visualise feedback in real time. This is done using two dynamic sliders and the computer mouse is used for selecting elements, semantic zooming and panning. The first slider sets a similarity threshold at which clusters are merged and colour is used to encode the different clustering assignments. The second 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. Supplementary Figure S1 illustrates part of the radial layout. Full details of these graphics are described in Vogogias et al. (2016).

After partitioning into linkage groups, the SNPs in each group need to be ordered. A two-point analysis calculates the recombination frequency and LOD score for each pair of SNPs in each possible phase using an EM algorithm to maximise the likelihood, and identifies the phase with the highest likelihood among those with a recombination frequency less than 0.5. The recombination frequencies are converted to map distances using Haldane’s mapping function and these are then used to calculate the best order for the SNPs in the linkage group using weighted multi-dimensional scaling. This optimises a criterion known as the stress function, using a LOD2 weighting. Details are given by Preedy and Hackett (2016). The configuration of SNPs can be displayed in two or three dimensions to identify SNPs that are distant from the rest of the linkage group. A two-dimensional MDS plot is shown in Figure 1. The effect of excluding these outlying SNPs can be explored. The program then fits a principal curve through the configuration of the remaining SNPs to give the distances of the SNPs along the linkage map. This map can be displayed graphically and exported as a figure or text file.

Finally the phases of the ordered SNPs need to be inferred In order to proceed to a QTL analysis. In TetraploidMap for Windows this information had to be entered manually. In the current program as many phases as possible are determined using the simplex coupling SNPs as a framework. However the user is required to complete any outstanding phases from the details of the most likely phase for each pair. The information on SNP name, position and phase in each parent is saved as a text file ready for QTL mapping.

### QTL Mapping Module

The QTL analysis is run as a separate module, for each linkage group separately. It requires three input files, with the SNP data for that linkage group, the linkage map (including phase information) and the phenotypic traits. The SNP data is sorted into map order initially and any unmapped SNPs are omitted from the QTL analysis. Boxplots can be displayed to check for outliers in the trait data.

Interval mapping is run for all selected traits, and the results are displayed as a LOD profile along the chromosome, with summary statistics on the LOD score, percentage variation explained and QTL effects for each homologous chromosome. A permutation test can be run to test the significance of the QTL peak. This gives a chromosome-wide threshold: a genome-wide threshold can be obtained by using the same seed for the permutation when analysing each linkage group and then deriving the maximum LOD overall (outside TetraploidSNPMap). If there is a significant QTL, then simple models, such as a simplex, duplex or double-simplex QTL, are tested to see how well they fit to the estimated genotype means at the most likely QTL location. For a simplex QTL (Qqqq x qqqq), there are two possible genotypes qqqq or Qqqq, with separate means. However for duplex or double-simplex QTLs (QQqq x qqqq and Qqqq x Qqqq) there are three genotype means qqqq, Qqqq, QQqq and these can be modelled as an additive effect with means {m, m+a, m+2a}, or as separate groups with means {m1, m2, m3} or as a dominant effect with means {m, m, m+d}. All possible models are fitted and compared using the Schwarz Information Criterion (SIC, Schwarz 1978), and the best six models are displayed.

## Implementation and Availability

TetraploidSNPMap is written in Java (version 8), which provides a user interface to the underlying analysis routines. Most of these are in Fortran 90, apart from the routines for the MDS analysis, and cluster analysis, which are written in R. A full user guide, a Windows installer, test datasets, as well as the source code are all available from our GitHub repository. The installation includes a small R distribution in order to run the analysis routines written in R. The software is made available under GNU General Public License (version 3).

The software uses OpenMP multithreading to optimise speed for multi-core workstations in several Fortran analysis routines. OpenMP is used in the calculation of similarity matrices for the preliminary grouping of simplex markers as well as in the main clustering routine and also in the twopoint ordering routine. Multi-process parallelism is used for the QTL random trait permutation routine, which relies on a non-thread-safe module. The main cluster analysis for all selected markers uses the R implementation of fastCluster (Müllner 2013).

TetraploidSNPMap has several steps where it calls Fortran routines that have been made generally available online by researchers. These are listed in Table 1. Full details of the adaptations and where in TetraploidSNPMap these programs have been used are given in the user guide.

## Examples of Use

TetraploidSNPMap is supplied with example data sets for each module. For the linkage analysis module, the dataset consists of 5332 segregating SNPs scored on the parents and 190 offspring of the potato cross Stirling × 12601ab1, which were analysed by Hackett et al. (2013). For the QTL mapping module, there are three datasets consisting of the linkage map with phasing for linkage group V from the same cross, the SNP data for that linkage group, and a set of trait data for eight traits analysed by Hackett et al. (2014). The user guide shows screen shots of each main stage of the analyses of these datasets.

## Conclusions

TetraploidSNPMap implements the methodology developed by Hackett et al. (2013, 2014) to help genetics researchers carry out linkage analysis and QTL mapping in full-sib populations of autotetraploid species. It is suitable for analysing genotype data from any technology that is informative about allele dosage. Its development was motivated by research in potato, but we anticipate that it will be a useful tool for a wider range of species.

## Source code and download links

TetraploidSNPMap, its documentation, source code and windows binary installation files can be downloaded from our GitHub repository:

https://github.com/BiomathematicsAndStatisticsScotland/TetraploidSNPMap

## Funding

This work was supported by the Scottish Government’s Rural and Environment Science and Analytical Services Division (RESAS). The original TetraploidMap software was developed with funding from the UK Biotechnology and Biological Science Research Council, grant 213/GAT09080.

## Acknowledgements

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.

## Figure legends

**Figure 1.** The MDS configuration for the markers in group 12 from the cluster analysis. Markers 4292, 3289 and 990 are slight outliers and the effects of omitting these should be explored.

**Figure S1.** Detail of a radial layout of a dynamic dendrogram to explore similarity among the markers. Markers are represented by squares and similarities by circles, with a decreasing diameter illustrating higher similarity. The weakest edges are shown in red.

## References

Brainerd WS, Goldberg CH, Adams JC (1996) *Programmer’s Guide to Fortran 90*, Third Edition. Springer.

Cheng R and Milligan GW (1995) [Hierarchical-clustering algorithms with influence detection.](http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=1&SID=U1JMQmHXjVAAc7ccVHN&page=1&doc=3) *Educational and Psychological Measurement*  55: 237-244.

Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP, et al. (2012) Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS ONE* 7: e36347 doi: 10.1371/journal.pone.0036347.

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

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

Hackett CA, Bradshaw JE, Bryan GJ (2014) QTL mapping in autotetraploids using SNP dosage information. *Theoretical and Applied Genetics* 127: 1885-1904.

Hackett CA, McLean K, Bryan GJ (2013) Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PLoS ONE* 8: e63939.

Hackett CA, Milne I, Bradshaw JE, Luo ZW (2007) TetraploidMap for Windows: linkage map construction and QTL mapping in autotetraploid species. *Journal of Heredity* 98: 727-729.

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

Müllner D (2013) fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python, *Journal of Statistical Software* 53: 1–18.

Narula SC and Desu MM (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 , 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](http://www.R-project.org/).

Schwarz GE (1978) Estimating the dimension of a model*. Annals of Statistics* 6: 461-464.

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

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

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

**Table 1.** External routines used by TetraploidSNPMap

|  |  |  |
| --- | --- | --- |
| Task | Routine | Web site and reference |
| Cluster analysis | HCINFLU.F | <http://www.pitt.edu/~csna/Milligan/readme.html>; Cheng and Milligan (1995) |
| Chi-squared probabilities | Applied Statistics routines AS170, AS147, AS245 | ftp://sunsite.univie.ac.at/mirrors/lib.stat.cmu.edu/apstat/.index.html; Narula and Desu (1981), Lau (1980), Macleod (1989) |
| Weighted linear regression | LSQ.f90, AS274 | <http://jblevins.org/mirror/amiller/>; Miller (1992, 2002) |
| Random permutations |  | Green (1963, 1977) |
| Calculation of quantiles | NUR.f90 | http://www.it.uu.se/edu/course/homepage/algpar1/ht02/ Quicksort\_examples/nur.f90. |
| Sorting vectors |  | Brainerd et al. (1996) |

Figure 1



Supplementary Figure S1

