

GGT 2.0

USER MANUAL

2007 EDITION

DEVELOPED BY RALPH VAN BERLOO
LABORATORY OF PLANT BREEDING
WAGENINGEN UNIVERSITY
THE NETHERLANDS

<http://www.dpw.wau.nl/pv/pub/ggt/>

www.plantbreeding.nl (in prep)

Contents:

Abstract	3
System Requirements	3
Input data	3
Building GGT files from locus and map data	4
Data Input/export using a spreadsheet	4
Alternative layout for spreadsheet input	6
Viewing the graphical genotypes	8
Statistics	9
Marker statistics	9
Image saving & printing, data export	9
Selection	9
Subset selection	10
Sorting of data	11
Editing of data	11
Advanced analyses	12
Advanced analyses 1: marker trait associations	12
Advanced analyses 2: Calculation of genetic distances	15
Advanced analyses 3: Marker linkage disequilibrium	17
Toolbar button descriptions	18
QTL data in GGT	18
Examples	18
Contact information	19
Acknowledgements	19
References	19
Links to external software	19
Appendices	20-24
-Cross pollinated (CP) and Doubled Haploid (DH) populations	20
-Figure 1: Schematic explanation of linkage group and individual viewing modes	21
-Figure 2: Example ggt drawing in linkage group mode	22
-Figure 3: Example ggt drawing in individual view mode	22
-Figure 4: Example ggt drawing in compound (show all) view mode	23
-Figure 5: Example of ggt drawing after data sorting	23

Abstract

A graphical representation of molecular marker data can be an important tool in the process of selection and evaluation of plant material. A computer program was developed that enables representation of molecular marker data by simple chromosome drawings in several ways. Commonly used marker file types that contain marker information serve as input for this program, which was named 'GGT' (an acronym of Graphical GenoTypes). Besides representation, GGT can also be used for a diverse range of selections and analyses.

System requirements and Installation

Although previously 16 bits versions of GGT existed, currently only GGT32, the 32 bit version of GGT is maintained. GGT32 runs on a computer running Windows. The installation will take about 5-10 MB of hard disk space. GGT comes as a self installing package, containing this manual, and example data files. The installer registers the .ggt data type with the GGT program. An uninstall option is added to the windows software menu.

Input data

GGT requires input in the form of GGT data files (or a spreadsheet, see further on). GGT data files are derived from two sources of data: A locus file, containing marker names and raw marker scores and a (linkage) map file, specifying marker positions on a linkage map.

> The locus file is a plain text file. It contains data on marker alleles using the Mapmaker/ Joinmap type of coding (Lander *et al.*, 1987; Stam, 1989).

```
name = listerset-5
popt = R18
nloc = 4
nind = 98

g3715
aabab bbba baabb abbba bbaba bbaba bbbbb bbbba abbbb bbbab
bbab abaab bbbbb abbaa abbab babba abb-a baaaa bbaba babab
w97
aabbb bbba baabb aab-a ababa bbaba bbbbb bbbba abbbb bbbba
bbab aabab bbbbb abaaa bbbbb babba bbbba baaaa bbaba babaa
w174
aabbb abba baabb aaa-a ababa bbabb bbbbb bbbba bbbbb bbbba
bbb abbab bbbbb bbaaa bbbbb bbbba bbbba baaaa bbbba babaa
w322
aaabb bbbb baaab aaa-a ababa bbabb bbbbb bbbba bbbbb bbaaa
bbb abbab bbbbb bbaaa bbbbb bbbba bbbba bbaba b-bba babaa
```

An example of a locus file is shown above.

> The map file is a plain text file. It lists, in two columns, markers and their map positions. GGT needs information on both the allelic composition of a locus as well as the map position.

```
chrom 5
g3715      0.0
w97        7.4
w174       12.8
w322       17.2
```

The map file corresponding to the previous shown locus file is shown above.

A GGT data file that combines this information can be constructed by merging a locus and a map file. This can be done by hand, but GGT also contains a convenient routine to perform this task: the 'Build GGT-file' option.

Starting with the 2005 edition, GGT is also equipped with a module to open GGT prepared data directly from a spreadsheet file like Microsoft Excel.

Building GGT files from locus and map data

Usually, marker data will be present in the form of a single locus file, containing raw marker data. This data is then used for mapping and subsequently split into several separate files, one for each chromosome/ linkage group. The 'Build GGT file' option 🍌 allows two ways of operation: (1) open both a map file for one chromosome and a corresponding locus file with the marker data on the same chromosome; merge the map data and the locus data and save the result as a GGT file. (2) Create a file with the combined maps of all (several) chromosomes. Markers belonging to a new chromosome must be preceded by a header, specifying the chromosome number (see example in the demo directory). This file is then opened, together with the locus file with 'raw' marker data (unsorted markers and marker data, all chromosomes lumped together) and merged by GGT. The latest versions of GGT also allow use of a single datafile, which may contain information on more than one linkage group. Compound map and loc files can be used for merging the data (provided they are properly formatted) and will yield a compound ggt datafile.

```
nchrom=2
nind=101
popt=ri8

name= 1
nloc=2
m488 0.0
B A A B B A - B A A A B A B B A - B A A A A B B B A A A B B B A A B B A A
B B A B B A B A A B B A A - B A B A A A A B B B A B B B B B B A B A B A A A
A B B B B B A A B B B A B B B B A A B A A A A A
g4715a 0.0
B A A B B A - B A A A B A B B A - B A A A A B B - A A A B B B A A B B
A A B B A B B B A A B B A A - B A B A A A A B B B A B B B B B B A B A B A
A A A B B B B B A A B - B A B B B B A A B A A A A A

name= 2
nloc=3
w51 0.0
B B B A A B - A B B A - B A A A B A - B B A A B A B B A A A B B B A B B A
B A A A B A A B B A A B B A A A B B A B B A B A B B A B A B B A B A A B B A A B
B B A A A A B B B B B A B B B A B B B A A B B
w204 3.3
A B B A A B - A B B B A B A A A B A - B B A A B A B B A B A B B B A B B A
B A A A B A A B B A B B B A A A B B A B B A B A B B A B A B B A B A A B B B A A
B B B A A A A B B B B B A B B B A A B B B A A B B
w122 4.0
A B B A A B - A B B B A B A A A B A - A B A A B A B B A B A B B B A B B A
B A A A B A A B B A B B B A A A B B A B B A B A B B A B A B B A B A A B B B A A
B B B A A A A B B B B B A B B B A A B B B A A B B
```

An example of the file structure of a compound, multi-group GGT file

Data input/ export using a spreadsheet

Starting with the 2005 edition, GGT was extended with a module to load and save GGT markerdata from/to spreadsheet files (eg Excel). This was done because many people use spreadsheets to store, arrange and manipulate markerdata. Data that is arranged in a proper format in a spreadsheet can now directly be read into GGT. Also options for editing of data in a spreadsheet-like environment within GGT have been added.

How to prepare your data in excel ?

Data should be formatted like the example shown below:

	A	B	C	D	E	F	G	H	I
1	alias		geno1	geno2	geno3	geno4	geno5		
2									
3	nchrom	2							
4	nind	5							
5	popt	F2							
6	:locus	pos	alleles						
7									
8	name=group1								
9	marker1	0	A	B	A	A	B		
10	marker2	25.8	A	U	A	H	B		
11	marker3	76.9	H	B	A	H	H		
12	marker4	101.3	U	B	H	A	A		
13									
14	name=group2								
15	marker21	15.5	H	B	A	H	A		
16	marker22	43.7	H	B	A	H	B		
17	marker23	81	A	U	A	H	B		
18									
19									
20									

Expected structure of this file:


- Top row contains a header row with the names of the genotypes, cell A1 contains the word 'alias'
- The number of chromosomes/ groups, and the population type are specified using, keywords **nchrom** and **popt** which should appear in the first column, while their value is given in the second column
- The first column contains the marker names
- The second column contains the marker positions
- Starting with the third column the actual marker data are listed
- The start of a new group is specified by the keyword '*name=*' followed by a group identifier in the first column
- Comments are allowed but should contain the ';' character as the very first character in the first column of the row that contains the comments

The import routine will determine by itself the number of loci per group and the number of genotypes, but manual specification using keywords **nind=** and **nloc=** (similar to textual compound ggt file) is allowed, but will be ignored.

To open an excel file with GGT data choose '*Load/Paste from Excel*' from the 'File' menu. In the window that opens click the '*Open Excel file*' button and specify the file. (alternatively, select the data in excel, select 'copy' in excel and click the button '*Paste Excel data*')

Hint: what to do if you have more than 250 genotypes/individuals?

The native GGT text format will allow this but also the GGT input spreadsheet allows more than 255 columns, unfortunately Excel does not!

*A solution can be to prepare your data in multiple sheets. Copy the first sheet into the spreadsheet. Add an additional column using the add column button . Next, select the data from the second sheet, position the cursor in the top cell of the added column in the GGT spreadsheet, right-click and select '**paste here**'. The additional data will be pasted in the right position. Repeat this if necessary.*

The data will appear in the GGT window like shown below

alias		geno1	geno2	geno3	geno4	geno5
nchrom	2					
nind	5					
popt	F2					
locus	pos	allele				
name=group1						
marker1	0	A	B	A	A	B
marker2	25.8	A	U	A	H	B
marker3	76.9	H	B	A	H	H
marker4	101.3	U	B	H	A	A
name=group2						
marker21	15.5	H	B	A	H	A
marker22	43.7	H	B	A	H	B
marker23	81	A	U	A	H	B

Alleles are shown in colored boxes according to the GGT settings. Data can be edited and changes are reflected immediately. Also data can still be pasted in place. E.g. if you want to overwrite data for the first marker with new data you have in excel this can be achieved as follows:

1. select the new data in excel
2. select copy
3. in the GGT select the first cell with data you want to replace and click in this cell using the right mousebutton
4. select 'paste here' from the options in the popup menu.

Current data will be overwritten without undo! so be cautious when using this option.

To transfer the data to the GGT main window click the button 'Import into GGT'



The grid view of the data remains available through the 'Export to Excel' option of the main GGT menu under – 'Export', or the toolbar button . Data can be saved in the required format in excel as well by using the "Save Excel File" button.

To keep a permanent copy of the data in the native ggt format, use the 'save as GGT file' option from the file menu.

Alternative layout

Since quite often marker data in a spreadsheet is arranged in such a way that for each marker the chromosome position is present in a separate column, starting with the 2007 edition GGT also supports this 'alternative layout'. An example is shown below. Key element that should be present if you want to import data arranged in this way is the phrase '**alternativelayout=true**', that should appear on the top of the file in the first column.

	A	B	C	D	E	F	G	H	I
1	:	chrom	pos	alleles					
2	alias			pl1	pl2	pl3	pl4		
3									
4	alternativelayout=true								
5									
6	name=alternative								
7	marker1_1	1	10.5	A	B	A	A		
8	marker1_2	1	30	H	H	A	H		
9	marker2_1	2	5	A	H	B	B		
10	marker2_2	2	17.3	A	B	B	B		
11	marker2_3	2	55.5	H	A	A	H		
12	marker3_1	3	0	H	A	H	H		
13	marker3_2	3	80	B	B	H	H		
14									
15									
16									

Recoding of data

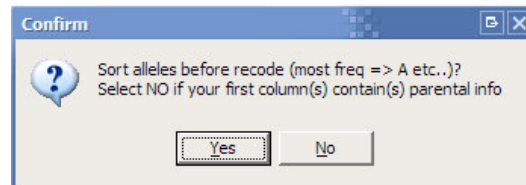
The native GGT coding style is based on single characters (this has historic roots in the locus format way of coding alleles, which formed the basis of GGT). However, when data are presented in the form of a spreadsheet the cell boundaries can be used to identify the separate alleles. Still the internally used allele coding was unchanged but an automatic translation of multiple character allele codes (e.g. SNP names, SSR repeat counts) to the GGT coding system is now possible. Use the

The figure below gives an example of the type of data that can be accepted. The left panel shows the original data as it was read into GGT spreadsheet module, while the right panel shows the same data after recoding.

alias		geno1	geno2	geno3	geno4	geno5
nchrom	2					
popl	F2					
locus	pos					
name=group1						
marker1	0	33	40	33	56	33
marker2	25.8	A	?	A	H	B
marker3	76.9	all1	all2	?	all2	all3
marker4	101.3	A	B	H	A	A
name=group2						
marker21	15.5	H	B	A	H	A
marker22	43.7	all1	all2	all3	all4	all2
marker23	81	A	?	A	H	B

Note that re-coding is not done randomly: by default the most frequent allele will be recoded as 'A', the 2nd frequent as 'B' etc.

It is also possible to present reference genotypes (e.g. population parents) in the first two columns with marker data and, by selecting NO on the popup question, force the allele present in the first column to be recoded as 'A' etc even if it is not the most frequent allele.



The recode-log (see the 'recode log' tab on the right) gives details on how the original data were translated. If there are any missing data, please specify the code used for this in the missing data box in this view BEFORE you click recode, otherwise the missing alleles are not recognized and get recoded as well. An example of the recode log is shown below.

marker1:	33=A	40=B	56=C	
marker2:	A=A	?=.	H=C	B=D
marker3:	all2=A	all1=B	?=.	all3=D
marker4:	A=A	B=B	H=C	
marker21:	H=A	A=B	B=C	
marker22:	all2=A	all1=B	all3=C	all4=D
marker23:	A=A	?=.	H=C	B=D


Missing data is coded as:
(leave blank to ignore)


?



Viewing the graphical genotypes

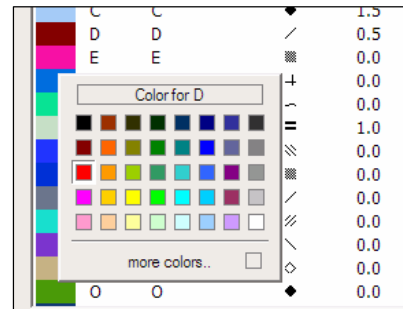
GGT data sets can be opened and viewed in GGT. The locus-allele data for several linkage groups can be represented in different ways. GGT allows a graphical representation of the data in 3 ways:

- i> arranged by linkage group
- ii> arranged by individual
- iii> arranged by linkage group but showing data on several linkage groups at once.

The differences between these two drawing modes are schematically drawn in Figure 1 and illustrated with examples in the appendix. The **'by linkage group'** viewing mode  (Figure 1A) is selected by default. This will show the first linkage group of all individuals next to each other. It is possible to display only a subset of the population.

The other mode for viewing graphical genotypes is **'by individual'**  (Figure 1B). This mode is useful to inspect several linkage groups of one particular plant or line together. Per individual, all linkage groups that have been opened are displayed next to each other, resulting in a partial or complete overview of a particular individual genome. The next individual can be viewed by clicking the 'up' button with the mouse or using the arrow keys on the keyboard.

Finally, the third drawing option **'show all'**  shows all available linkage groups (arranged similar to the 'linkage group mode', but rotated 90°) in one image. There are many user defined options for drawing the image. Size of graphical elements, fonts, colors, hatch patterns etc can all be modified from the 'settings' panel  **Settings**. For instance, to change the color of a specific allele select the allele-code from the list shown in the 'settings' panel after clicking the *'allele and general options'* tab, and click on the color. A popup box will appear allowing you to select a different color. (by clicking 'more colors' in this popup box even more color options will be shown. Similar changes to allele name, hatch pattern and numerical allelic value can be made by clicking on the appropriate area in the settings box.



You can customize the looks of the drawing by changing one of the many options.

- Discrete or gradual color changes; 3D looking bars or flat bars, Solid or hatched fill patterns.
- Change the bar width, space between bars, top margin, left margin
- Change fonts, label alignments
- Change the bar length multiplier, useful for very dense maps

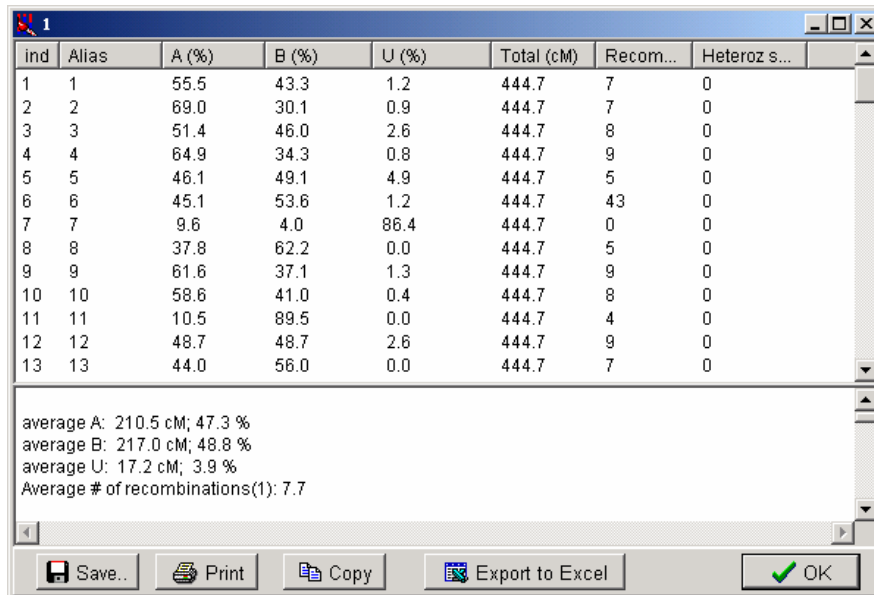
Note that the gradual color change and 3D-look options require the drawing of many more graphical elements, which slows down image creation.

Important remark:

When the data indicate that a crossover event has occurred in between two loci, the *exact location* of the crossover is unknown. GGT uses the most likely position for the crossover, exactly in the middle of the two markers, to change the color presentation. The user should be aware of this! Caution should be taken when interpreting crossover events, especially crossovers occurring between markers that are positioned at large distances from each other.

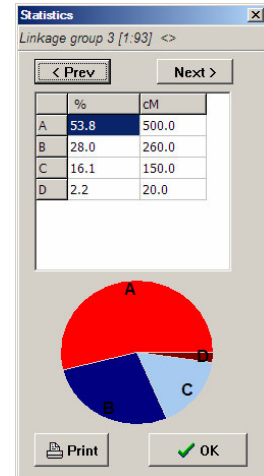
Statistics

The estimated proportion of the genome that was derived from different sources is calculated. This can be viewed by clicking the 'Stats' button, or selecting 'Statistics' from the menu. A detailed list of individual (plant/line) statistics is available through the 'Statistics' menu. In the *global statistics* window there are also estimates of the number of recombinations, the number of heterozygous fragments (useful in backcross programs) and lists of questionable allele changes (so called *singletons*)



ind	Alias	A (%)	B (%)	U (%)	Total (cM)	Recom...	Heteroz s...
1	1	55.5	43.3	1.2	444.7	7	0
2	2	69.0	30.1	0.9	444.7	7	0
3	3	51.4	46.0	2.6	444.7	8	0
4	4	64.9	34.3	0.8	444.7	9	0
5	5	46.1	49.1	4.9	444.7	5	0
6	6	45.1	53.6	1.2	444.7	43	0
7	7	9.6	4.0	86.4	444.7	0	0
8	8	37.8	62.2	0.0	444.7	5	0
9	9	61.6	37.1	1.3	444.7	9	0
10	10	58.6	41.0	0.4	444.7	8	0
11	11	10.5	89.5	0.0	444.7	4	0
12	12	48.7	48.7	2.6	444.7	9	0
13	13	44.0	56.0	0.0	444.7	7	0

average A: 210.5 cM; 47.3 %
average B: 217.0 cM; 48.8 %
average U: 17.2 cM; 3.9 %
Average # of recombinations(1): 7.7



Overview of individual statistics, clicking the column headers will sort the data using the selected parameter as a key

Marker statistics

The option 'marker statistics' can be selected from the 'statistics' menu entry. This will open a new window with on the right a tabular summary, for each marker, of the number of alleles present. Results can be exported to excel. Only for F2 populations coded as 'A' 'H' and 'B' an extra column is reported: expected segregations are calculated for F2 data and markers deviating from expectations will be indicated by an asterix.

On the left of the marker statistics window a treeview of all markers is shown. When a marker is selected, a pie-chart of the allelic composition of the marker will be drawn.

Image Saving & Printing, data export

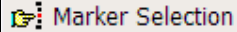
GGT produces hi-resolution enhanced metafile drawings, which can be saved to disk or printed from within GGT. Alternatively the drawings can be saved as Jpeg files. Tip: Saving and then importing of a saved image often produces better results than copy-paste.

GGT files of specific individuals can be converted to Mapchart format and opened in the Mapchart drawing package (Voorrips, 2002). The locus data can also be exported to MEGA format (www.megasoftware.net). MEGA needs to be tricked to accept locus data as it is aimed at sequence and protein data, but by pretending alleles are amino acids MEGA will accept the data and allow e.g. hierarchical clustering.

Selection

With GGT it is possible to specify a desired (marker) genotype for markers that may reside on different linkage groups, and select among the genotypes those genotypes that comply with the demands. The result will be a filtered view, in which only the graphical genotypes of individuals that match the specified criteria are drawn. Also a textual summary of criteria and results is presented. In this way a population can be screened very efficiently for regions of interest.

Setting selection criteria

To set selection criteria visit the *marker selection* tab . You will see a list with all markers for all groups. Positioning your mouse above a marker for a second will select that marker (it gets a blue color). Use the **keyboard** to enter a selection allele code. (note: input is case sensitive, use *SHIFT* to obtain capitals). E.G. Pressing 'A' will force selection for genotypes that have the 'A' allele at the selected marker. To remind you on the available options the alleles that are present for each marker in the population are shown in the last column.

It is also possible to specify exclusive selection criteria. Use the *CTRL* key while pressing a key will specify 'Select all but the specified allele'. This is indicated with a '!' character (! is often used as the symbol for *NOT*) preceding the selected character. Also note the change of icons in the first column. To undo the selection for a marker press the asterisk (*) character.

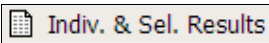
The example below illustrates this: selection criteria were set to select individuals with the 'B' allele for marker *g17288 that also do NOT have the 'A' allele for marker *g6842 (so 'B', 'H' etc are all included).

<input type="checkbox"/>	*g4514	7.8	A11	BAAAABBBBABABAAAAUBBBABBBBAUBBAAAABBBBBUBBBBBABBBBUABBBUBBAAAA
<input type="checkbox"/>	*m323	12.7	A11	BAAAABUBBBABABAAAAABBBBAABABAABBAABBBBUBBBBBABBBBUABBBUABBBAB
<input checked="" type="checkbox"/>	*g17288	12.7	B	BAAAABBBBABABAAAAABBBBAABABAABBAABBBBUBBBBBABBBBUABBBBAABBBAB
<input type="checkbox"/>	*w277	13.7	A11	BAAAABUBBBABABAAAAUBBAABABAABBAABBBBUBBBBBABBBBUABBBBAABBBAB
<input type="checkbox"/>	*w54	20.0	A11	BAAAABUBBBABABABABUBAABABAABBBABBBBUBBBBBABBAABBAABBBBAABBBAB
<input type="checkbox"/>	*m220	20.5	A11	BAAAABUBBBABABABABABABAABBAABBBBUBBBBBABBBBUABBBUABBBUABBBAB
<input type="checkbox"/>	*ez	27.3	A11	BAAAABUBBBABABABABUBAABABAABBBABBBBUBBBBBUBBAABBAABBBBAABBBB
<input checked="" type="checkbox"/>	*g6842	30.7	!A	BAAAABUBBBABABABABABABBAABBBBUBBBBBUBBBBUABBBBUABBBUBBBBAUABBBB
<input type="checkbox"/>	*m251	38.7	A11	BAAAABUBBBABABBBBBAABABAABBBBUBBBBBABBBBUABBBUABBBUBBBBAUABBBB
<input type="checkbox"/>	*w122	39.5	A11	BAAAABUBBBABABBBBBAABABAABBBBUBBBBBABBBBUABBBUABBBUBBBBAUABBBB

Clicking the Apply button will activate the selection criteria and list the results, also the GGT drawings will show only the selected genotypes (those that comply with the set criteria). Historically "U" was used for 'unknown', therefore 'U' individuals will still be included even is a selection is set to include only a specific allele.

A scenario where the select option may be useful could be when dealing with a backcross population. Criteria can be set such that the genome at regions of interest (possibly on multiple chromosomes simultaneously) are of the donor type. Inspection of the individuals that meet the selection criteria can simplify selection of the most promising ones, i.e. those with a high amount of recipient background.

Viewing selection results

The *Indiv & Sel results* tabsheet  does not only display a summary of selection criteria and results, but also allows adjustment of the selected subset, by selecting or unselecting individual genotypes, simply by adding or removing the checkmarks shown before each individual.

Selection of a subset


The option to select a (core) subset attempts to select a subset of the available genotypes of user-defined size in such a way that the total diversity, remaining in the genotypes that make up the subset, is as high as possible. The best set will be found by an exhaustive search but this is computationally not feasible. For instance a core subset of 25 genotypes out of a larger set of 100 would mean testing

of $\binom{100}{25} = 2.4 * 10^{23}$ possible combinations. We therefore apply an iterative optimization procedure

that was based upon the procedure introduced by Bataillon (1996). A brief explanation of the procedure: When selecting a core subset of size K from a larger base set of size N, aiming to maximize remaining diversity in set K the procedure will start with a random set of size K, then temporarily remove the K members (one at a time) and evaluate the diversity of the so obtained K sets of size K-1. The best set is retained. Next one by one the N-K+1 remaining genotypes are added to the set (and removed again) so N-K+1 sets of size K are tested. The genotype that, on addition, has yielded the highest increase in the diversity parameter is retained. Next one by one each of the K

members of the new set are removed and re-added etc. This procedure is iterated until the same genotype is first removed and then added again, or until a predefined number of iterations has passed. This procedure has the advantage that it is relatively fast, but it will usually not yield the optimal core set. Most likely it will, however, report a near optimal set that still has practical relevance. In GGT this whole procedure is performed a number of times (default=100) and statistics on the diversity score of each set, and the number of times it was selected are reported. Also a table is presented of all genotypes involved in selection that lists how often each genotype was involved in the final subset of a selection run. This may provide useful information to identify important/essential genotypes within the larger set.

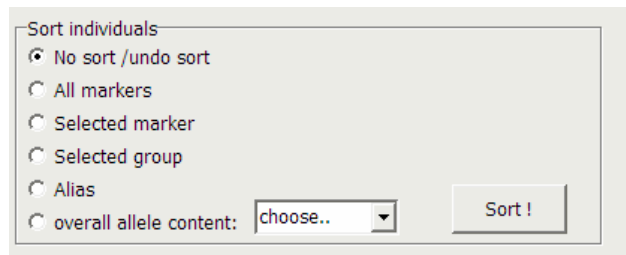
A diversity test statistic is needed to decide on the value of selected sets, GGT presents three options for this statistic: (1) the total number of alleles present, which is a useful measure to avoid allele fixation. Another option is (2) the Garza-Williamson (G-W) index (adapted, see Excoffier, 2006) which is also a measure for the number of alleles, but relative to the allele range for each marker, so alleles with a low frequency will obtain a higher weight. The final measure is (3) an unnamed diversity index, which is basically reversely proportional to the sum of all relative allele frequencies.

A recent addition to this module is the option to 'fix' certain genotypes. By selecting genotypes from the list and clicking the lock button  these genotypes will be selected directly and remain in the selection regardless of the test statistic. Optimization of diversity statistics will proceed as usual, but the procedure is restrained to keep the fixed genotypes in the selected set. This can be useful if there are certain genotypes (containing valuable resistance gene, or some QTL etc) that need to be included for certain and you want to select an optimal set by addition of those genotypes that supplement this fixed prior selection.

Sorting of data

Individuals and their associated marker-data can be re-arranged based on a number of sorting criteria, like marker content, alias names and even trait values (these first need to be entered into GGT separately, see further on in the section on Associations).



When sorting on marker content: all loci, all loci within a linkage group or the alleles of a specific marker can be used as a sorting key. E.g. it is possible to select from the treeview a marker that is linked to a gene or QTL of interest, and use this as a key for sorting. As a result all individuals that are of the same type for that marker (and contain or lack the gene) will be grouped together, also in the GGT drawings.



Other sort options are: revert to original data order (No sort), sort by genotype names/Alias sort by allele content, sort by allele content or sort the genotypes according to a trait value, supplied separately (see under Association).

Editing of Data

The Data Edit & Sort can also be used to edit the data in GGT. But the 2006 version of GGT introduced a spreadsheet type dataview and edit option that may be easier to use.

The button 'Show & Edit data in gridview'  **Show & edit data in gridview** will bring up the data arranged in a spreadsheet-like structure, ready for editing. Also the menu option 'Excel-style edit' from the 'edit' menu or the toolbar button  can be used to access this structure.

Note that after editing and clicking 'Import into GGT' you essentially erase all current data and replace this with the data present in the spreadsheet.

Advanced Analyses...

Starting with the fall 2005 edition, GGT options were extended considerably. Association studies of trait data on marker data, calculation of genetic similarities between genotypes and the plotting of linkage disequilibrium (LD) against map distance are all new features that should still be considered experimental. Most of these features were untested for (and will probably not work) cross pollinated datatypes.

As these features have not yet matured completely also this section of the manual is expected to be updated and extended in future.

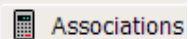
Numerical allele values

In order to perform calculations with marker data numerical values need to be assigned. For commonly used alleles default values are set, but it is possible to provide the numerical values for marker-alleles yourself, by shifting to the 'settings' tabsheet, and clicking on the 'numval' column. Choices for numerical values could be related to the amount of donor genome associated with an allele code, or simply a 0/1 binary value setting. For instance alleles 'A', 'B' and 'H' have default values 2, 0 and 1. This is useful when these alleles are connected with gel intensity scores where 'A' indicates homozygous presence, B homozygous absence and H indicates a heterozygous present marker. Use -99 as the value for missing data. These will then be ignored in any calculations.

Pasting numerical marker values

It is possible to replace the inferred numerical values for markers by values from a spreadsheet. This is useful when, for instance, you have true gel intensity data available, and would like to perform intensity-trait association or LD analysis. Pasting marker-data is possible when the column headers contain genotype labels identical to the original labels (and identical to the labels used in the trait data) and the row headers contain marker-names that are identical to the locus names.

Advanced Analyses 1: marker-trait associations



Association analysis of marker-trait associations is commonly used in sets of related genotypes that are not constructed as a mapping population, meaning that classical QTL mapping methods are not directly applicable. Still, exploratory analyses can be performed to investigate possible association between trait observations and (conserved) marker regions. This process is called association analysis/mapping.

Input of trait data by Copy/Paste from spreadsheet

To perform this analysis, in addition to the marker data, also trait data needs to be presented to GGT. Input of trait data can be performed by selecting the 'Associations' tabsheet. Next, from the bottom set of sheets, select the sheet 'Trait data'. By clicking the button 'Paste trait data' data that was previously selected and copied from an (excel) spreadsheet will be pasted into GGT. The selected data must contain, in the first row, the names of the genotypes identical to the names (alias) specified for the marker data. The data must also contain the trait names in the first column. An example of a properly setup excel sheet is given below. Note that the order of the genotypes does *not* have to be identical to the order used in the marker data.

	A	B	C	D	E	F	G
1		geno1	geno2	geno3	geno4	geno5	
2	trait1	20	30	15	25	33	
3	trait2	1.1	1.5	1.4	1.3	1.3	
4	trait3	-0.4	-0.3	-0.3		-0.4	
5							
6							
7							
8							
9							
10							
11							
12							

	geno1	geno2	geno3	geno4	geno5
trait1	20.0	30.0	15.0	25.0	33.0
trait2	1.1	1.5	1.4	1.3	1.3
trait3	-0.4	-0.3	-0.3		-0.4

After pasting the data it will be available to GGT as shown on the right (only section shown).

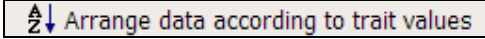
Inspecting the trait data.

The 'Trait stats' tabsheet is used to select, after analysis, which traits should be plotted in the association plot. It also allows inspection of the trait data. Clicking on a trait name in the left panel will plot a histogram of the trait data, and allow visual inspection of the trait distribution.

Sorting on trait data

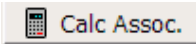
The trait data can now also be used as a key for sorting the genotypes! This can be used to study putative correlations between certain traits and marker segregation patterns.

Simply click the trait you want to use as key for sorting (it should become highlighted) and click the

'Arrange data according to trait values' button  to sort the data according to the selected trait. Use the 'Data Edit & Sort' tab and select 'no sort' to undo this sorting.

Association Analysis

Once trait data and marker data are available to GGT, click the button 'Calc Associations'

 to start the calculation of marker-trait associations. Statistics that will be calculated are the squared correlation coefficients between marker data and trait values (R^2 values) and the associated probabilities of the correlation values. As these probabilities become very small, the ^{-10}LOG values are reported. A value of $^{-10}\log(p)=3$ thus indicates a correlation probability value of 0.001.

The calculated association statistics are reported in the tabsheets ' $^{-10}\text{LOG}(p)$ ' and ' R^2 '

The color coding of the cells is adjusted to easily spot the larger association statistic values. An example of the type of display of the results is shown below:



traits	trait4	trait5	trait6	trait7	trait8	trait9	trait10	trait11	trait12	trait13	trait14	trait15	trait16	trait17	trait18	trait19	trait20
3	0.3	0.1	0.2	0.4	0.0	0.0	0.2	0.2	0.4	0.4	0.0	0.0	0.3	0.0	0.0	0.4	0.0
3	0.4	0.1	0.3	0.5	0.0	0.0	0.1	0.2	0.6	0.4	0.0	0.0	0.3	0.0	0.0	0.4	0.0
1	0.2	0.0	0.1	0.2	0.0	0.0	0.1	0.2	0.2	0.1	0.0	0.0	0.0	0.1	0.0	0.1	0.0
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.4	0.1	0.2	0.4	0.0	0.0	0.2	0.2	0.6	0.5	0.0	0.0	0.3	0.0	0.0	0.5	0.0
0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1
0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.2	0.1	0.1	0.3	0.0	0.0	0.1	0.2	0.4	0.3	0.0	0.0	0.2	0.0	0.0	0.2	0.0
2	0.3	0.1	0.3	0.4	0.0	0.0	0.2	0.2	0.6	0.5	0.0	0.0	0.3	0.0	0.0	0.4	0.0
3	0.3	0.1	0.2	0.5	0.0	0.0	0.2	0.2	0.6	0.5	0.0	0.0	0.3	0.0	0.0	0.4	0.0
3	0.4	0.1	0.3	0.5	0.0	0.0	0.3	0.3	0.7	0.6	0.0	0.0	0.4	0.0	0.0	0.5	0.0
3	0.3	0.1	0.3	0.4	0.0	0.0	0.2	0.2	0.7	0.5	0.0	0.0	0.3	0.0	0.0	0.4	0.0
3	0.4	0.1	0.3	0.6	0.0	0.0	0.3	0.3	0.7	0.6	0.0	0.0	0.4	0.0	0.0	0.6	0.0
3	0.4	0.1	0.3	0.6	0.0	0.0	0.3	0.3	0.7	0.6	0.0	0.0	0.4	0.0	0.0	0.6	0.0

False discovery rate statistics

A way to deal with significance thresholds while performing large numbers of tests, as is done in association analysis, is to estimate the *false discovery rate* (for explanation see Storey, 2002 and related papers). An FDR threshold for the p-values of individual association tests is calculated by GGT. This threshold is based on the empirical data and gives a less conservative way to filter out putative interesting markers that show marker-trait associations. The distribution of the p-values of the individual association tests plays a key role in these estimates. By clicking on a column in the p-value datasheet plots are shown that allow visual evaluation of the p-value distribution.

Association plots

The results can also be presented in a graphical way – for one trait at a time. By clicking a column in the $-10\text{LOG}(p)$ sheet this column, and the trait shown in the column are selected. Clicking the button

'Plot Assoc'  **Plot Assoc.** or the corresponding toolbar button  will create a plot that shows association plotted along the chromosome bars.

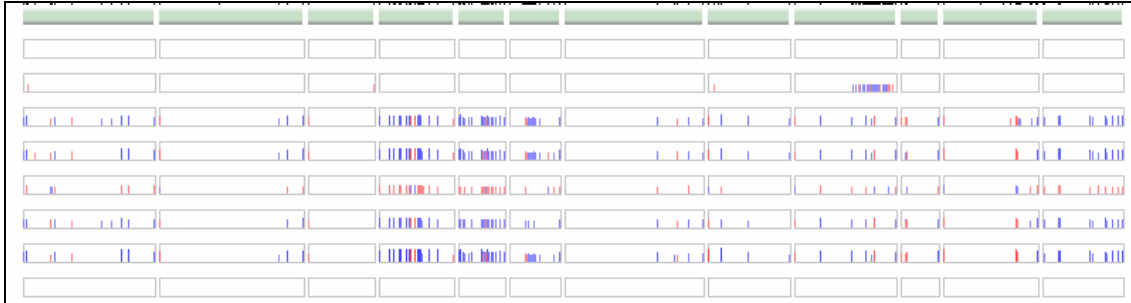
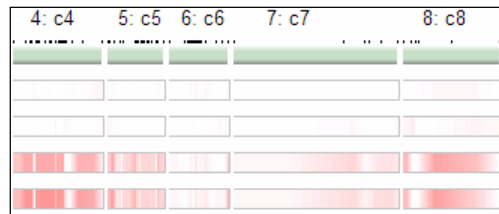


Figure above: An example of the overview association plot along the chromosome bars: Markers with an FDR value above the threshold (adjustable in settings) are shown, the height of the bar and color intensity are proportional with the amount of association. Positive associations are shown in red, negative associations in blue. Although the sign of the association between neutral markers and traits is in itself not relevant, these colors can be used to compare associations for different traits in the same genomic regions. If one trait is negatively associated with a marker while another trait is positively associated with the same markers, this is an indication that these traits are negatively correlated (at that position).

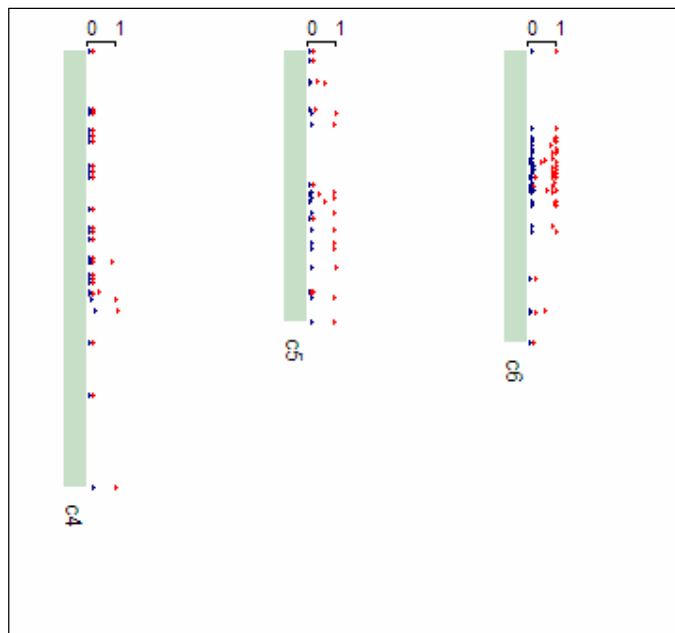
An alternative representation that can be selected from the settings – general options is the QTL heatplot, which plots the observed R^2 values as a color intensity along the chromosomes and interpolates these intensities in between observations.

Although visually appealing, caution should be taken when interpreting these plots in the case of a poor marker density, as the interpolation can give rise to biases. An example of a section of such a plot is shown on the right.



Detail plots

Clicking on a bar for a specific trait will plot a detail plot for the selected trait, revealing for each marker position the observed R^2 value (red dots) and the corresponding q-value ($1-q$ is plotted in blue). An example of this plot is shown on the right.



Starting with the fall 2006 edition a second option for association analysis was implemented, aimed at analysis of multi-allelic markers. This approach, which is essentially a one-way anova for all allele classes, uses a third party statistical component and should be considered 'preliminary' It also does not provide yet all options for visualization.

Advanced Analyses 2: Calculation of genetic distances

The availability of marker data allows comparisons of genotypes for these marker data. An overall analysis of the relatedness of all genotypes in the dataset can be performed by calculating the genetic distance for each pair of genotypes. Several measures for estimating the genetic distance (based on marker data) exist. GGT offers a choice between the simple matching coefficients, (the number of shared alleles as a proportion of all alleles), the Jaccard distance or the Euclidean distance (the square root of the sum of all squared differences between alleles). The Euclidean distance is often used for quantitative data, and may be considered somewhat artificial for re-coded marker data.

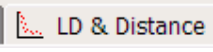
The differences between these well known distance measures are illustrated in a numerical example below. Assume we have scored two markers for eight individuals, and these scores have resulted in these data:


	Ind 1	Ind 2	Ind 3	Ind 4	Ind 5	Ind 6	Ind 7	Ind 8
Marker 1	0	0	1	2	-	0	1	2
Marker 2	0	1	1	3	0	-	1	2

Then the different distance measures can be calculated as follows:

- Allele sharing/ simple matching distance: 4 individuals with shared alleles (Ind 1,3,7 and 8) out of 6 (8-2 with missing data) observations: $\text{Similarity}_{\text{SM}} = 4/6 = 0.667$; $\text{Distance}_{\text{SM}} = 1 - 0.667 = 0.333$
- Jaccard distance: 3 individuals with shared non null alleles (Ind 3,7,8) out of 5 (8-2 with missing data and 1 with shared null allele) observations: $\text{Similarity}_{\text{Jac}} = 3/5 = 0.4$; $\text{Distance}_{\text{Jac}} = 1 - 0.4 = 0.6$
- Euclidean distance: only take the 6 pairs of data without missing data. $\text{Distance}_{\text{EUCLID}} = \sqrt{[(0-0)^2 + (0-1)^2 + (1-1)^2 + (2-3)^2 + (1-1)^2 + (2-2)^2]} = \sqrt{(0+1+0+1+0+0)} = \sqrt{2} = 1.41$

Note that for allele sharing distance any allele code can be used, GGT will check for matching alleles, so this measure can also be used for multi-allelic datasets.

The tabsheet '*LD & Distance*'  gives access to the LD & Distance calculation

functions, select the sheet  from the bottom row to perform distance calculations.

Clicking the button 'Calculate Distances (after selecting the desired measure for similarity) will calculate all pairwise distances and show the results in a half diallel matrix.

The distance matrix can be saved as an excel file, but also as a textfile ready for opening by MEGA as distance matrix file, or a NEXUS textfile that can be opened by, for instance, SPLITSTREE 4 (<http://www-ab.informatik.uni-tuebingen.de/software/splitstree4/>).

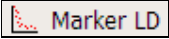
The MEGA package (www.megasoftware.net) can further be used to create hierarchical and other cluster diagrams, based on the distance matrix calculated by GGT, while the SPLITSTREE package is a versatile package for visualization and analysis of molecular evolutionary data.

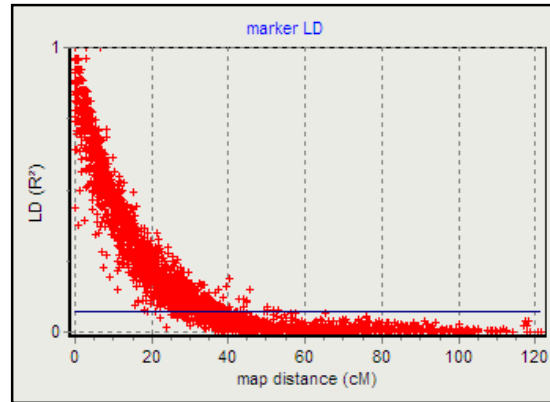
Recently the Neighbor-joining (single linkage) & UPGMA (average linkage) clustering algorithms were implemented to allow direct cluster analysis in GGT, resulting in Newick style output files which can directly be visualised by many tree drawing programs (like MEGA, treeview etc).

Advanced Analyses 3: Marker Linkage Disequilibrium

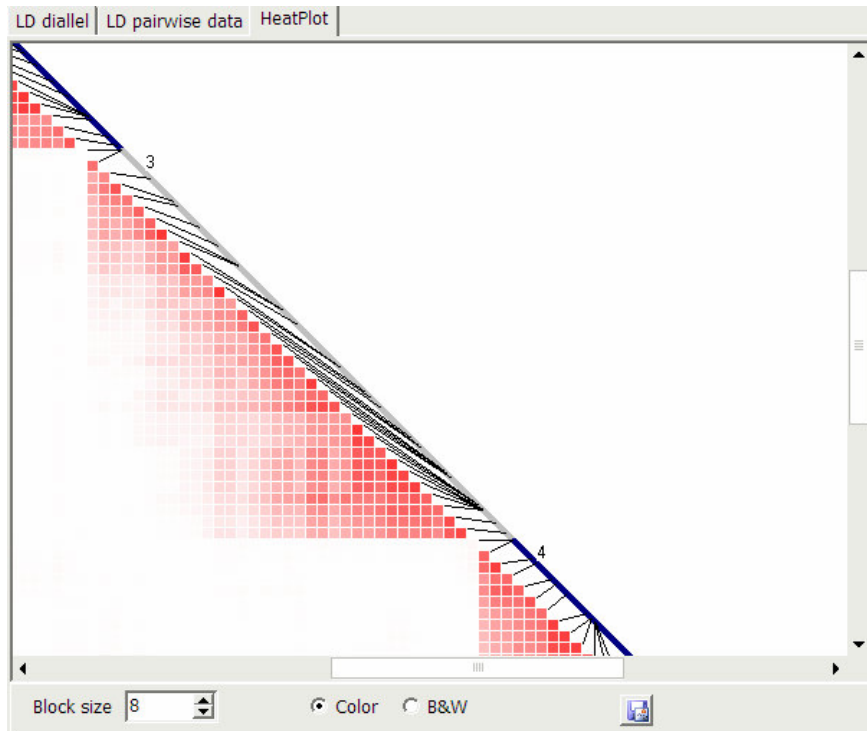
Especially when dealing with non-mapping population data it may be of interest to have a look at marker behaviour, especially of closely linked markers. It is assumed that repeated meioses and recombination will result in markers that are in Hardy-Weinberg or Linkage Equilibrium with regard to each other. However, markers that are very closely linked may remain linked if insufficient recombination events have passed. This is called Linkage Disequilibrium (LD; See Gaut and Long, 2003 for a nice review), and several measures to observe this phenomenon were defined. It is expected that Linkage Disequilibrium is reversely related to genomic (and genetic) distance.

GGT offers a few simple measures of linkage disequilibrium: the simple and squared correlation between markers, the probability of this correlation, Lewontin's D' value as well as the χ^2 value (Zhao et al., 2005).

The tabsheet 'Marker LD'  offers access to the LD and haplotype module. Select a group and your LD statistic and click the button 'Calculate marker LD stats for group>' to calculate all pairwise marker LD statistical values. Or calculate all to calculate LD for all groups. Automatically a plot is created that plots for each marker pair the observed LD value against the marker map distance. A decay of LD with map distance is expected, in general. Pairwise LD values are also retained and can be exported for use in external software, and a heatmap of r^2 values is available as well.



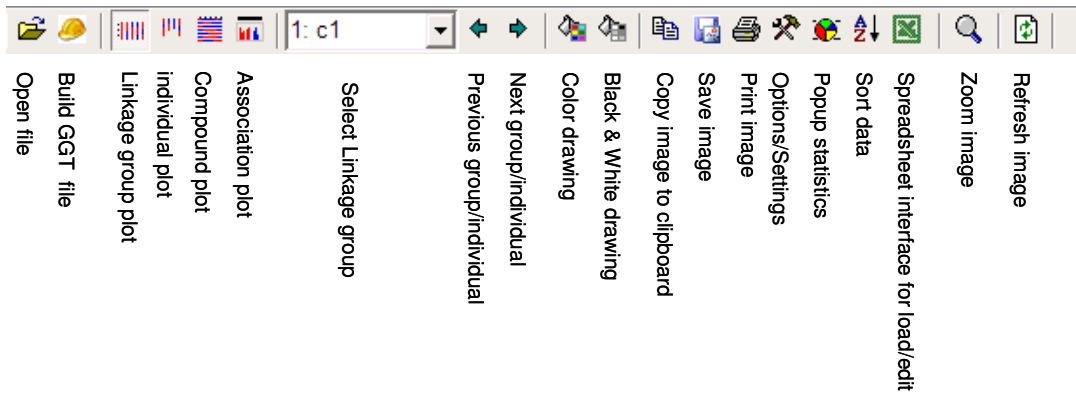
Example of a heat plot



Explanation of commands

Most frequently used commands are available from the main toolbar, the menu supplements the toolbar functions with commands for exporting, detailed statistics etc.

The functions of the toolbar items are explained below:



QTL data

The 2004 program revision added the option to use up to 80 different codes for alleles, as well as the inclusion of QTL data. QTL data are specified as marker data, but contain two extra positions to indicate the start and end of the supporting confidence interval.

An example of the specification of a QTL is shown below

```
name=1
nloc=4
m488 0.0
BAABBAUBAAABABBAUBAAAABBBAAABBAABBAABBABBABABAABBAUBABAAAAABBBABBBBBBABABAAA
ABBBBBBAABBBABBBBAABAAAA
g4715a 0.0
BAABBAUBAAABABBAUBAAAABBBUAAAABBAABBAABBABBABABAABBAUBABAAAAABBBABBBBBBABABAAA
ABBBBBBAABUBABBBBAABAAAA
FT_qt11 3.0/0.1/8.0
BAABBAUBAAABABBAUBAAAABBBUAAAABBAABBAABBABBABABAABBAUBABAAAAABBBABBBBBBABABAAA
ABBBBBBAABUBABBBBAABAAAA
w372 0.0
BAABBAUBAAABABBAUBUAAAABBBAAABBAABBAABBABBABABAABBAABBBABAAAAABBBABBBBBBABABAAA
ABBBBBBAABBBABBBBAABAAAA
```

Examples

Some examples of graphical genotype drawings are presented in Figures 2 to 5. Figure 2 shows the 'linkage group view' of a graphical genotype. Figure 3 shows in 'individual view' a graphical representation of the complete genome of an *Arabidopsis* line. Figure 4 shows (a section of) the third drawing option that draws whole genomes as horizontal lines. Figure 5 shows a simple application of the sorting possibilities in GGT.

Troubleshooting

See the frequently asked questions on the GGT website for the most up to date information on frequently asked questions and answers, GGT updates and planned development.

Contact information

Dr. Ir. Ralph van Berloo
Laboratory of Plant Breeding,
Wageningen University
PO Box 386 6700 AJ Wageningen, The Netherlands
email: Ralph.vanBerloo@wur.nl - ggt@berloo.com

Acknowledgements:

I want to thank many GGT users for useful suggestions and their help with debugging the software.

References:

Bataillon TM., David JL, and Schoen DJ (1996) Neutral genetic markers and conservation genetics: simulated germplasm collections. *Genetics*, 144, 409-417

Excoffier, L (2006) Arlequin 3 user manual: <http://cmpg.unibe.ch/software/arlequin3/>

Gaut BS and Long AD (2003) The Lowdown on Linkage Disequilibrium *The Plant Cell* 15:1502 - 1506

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newberg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1:174-181

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121:185-199

Stam P (1993) Construction of integrated genetic maps by means of a new computer package: JoinMap. *The Plant Journal* 3:739-744

Storey J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479-498.

Tanksley SD (1993) Mapping polygenes. *Annu Rev Genet* 27:205-233

Van Berloo R (1999): GGT: software for the display of graphical genotypes. *J Hered* 90: 328-329

Voorrips RE (2002) MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs *J Hered* 93: 77-78.

Young ND, Tanksley SD, 1989. Restriction fragment length polymorphism maps and the concept of graphical genotypes. *Theor Appl Genet* 77:95-101

Zhao H, Nettleton D, Soller M, Dekkers JCM (2005) Evaluation of linkage disequilibrium between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genet Res Camb* 86: 77-87

External software links:

Mapchart: <http://www.biometris.wur.nl/uk/Software/MapChart/>

MEGA: <http://www.megasoftware.net>

Splitstree 4: <http://www-ab.informatik.uni-tuebingen.de/software/splitstree4/>

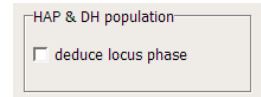
Appendix: Cross-pollinated & DH populations

CP populations are also supported by GGT, but require some extra care.

The main difficulty with CP populations is the segregation. In a population derived from two heterozygous parents the origin of an observed allele is not always clear.

It is good practice, when dealing with CP populations, to infer the segregation from the segregation pattern. Also information on linkage phase of neighbouring markers provides important information. GGT uses a routine to estimate the most probable phase information from the data. When dealing with CP populations, we need to be aware that we are looking at the result of two independent meioses, one paternal and one maternal meiosis. The different alleles originating from the different parental progenitors should be identified as such when possible. GGT will recode the data whenever it can resolve the parental origin of a locus, however this is also dependent on the linkage phase, and only possible in a few cases. For example in an <abxaa> segregation allelecode "ab" must have received the 'b' from the mother parent. In the case of <aaxab>, allelecode "ab" has received the 'b' allele from the father. Now it depends on the linkage phase between these loci how they should be represented.

DH populations are often scored without reference to parental genotypes, simply scoring marker presence and absence. When this is the case the linkage phase between markers becomes important to reconstruct parental origin. GGT is able to infer the most likely linkage phase, and draw the GGT image taking linkage phase into account. In Settings – General options check the box next to 'deduce locus phase' if you want GGT to infer linkage phase from the data.



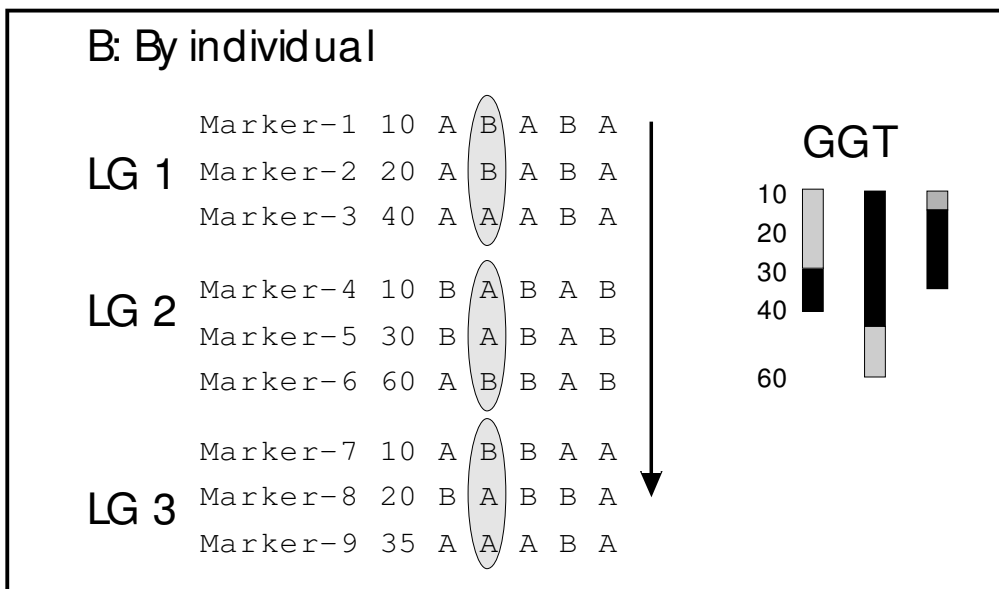
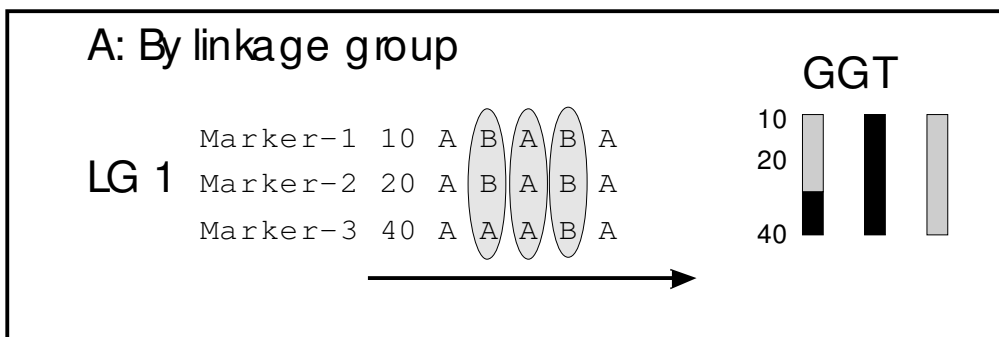


Figure 1: Two viewing modes for displaying graphical representations of molecular marker data by GGT. Left the original data is shown and on the right the resulting Graphical Genotype drawing is shown.

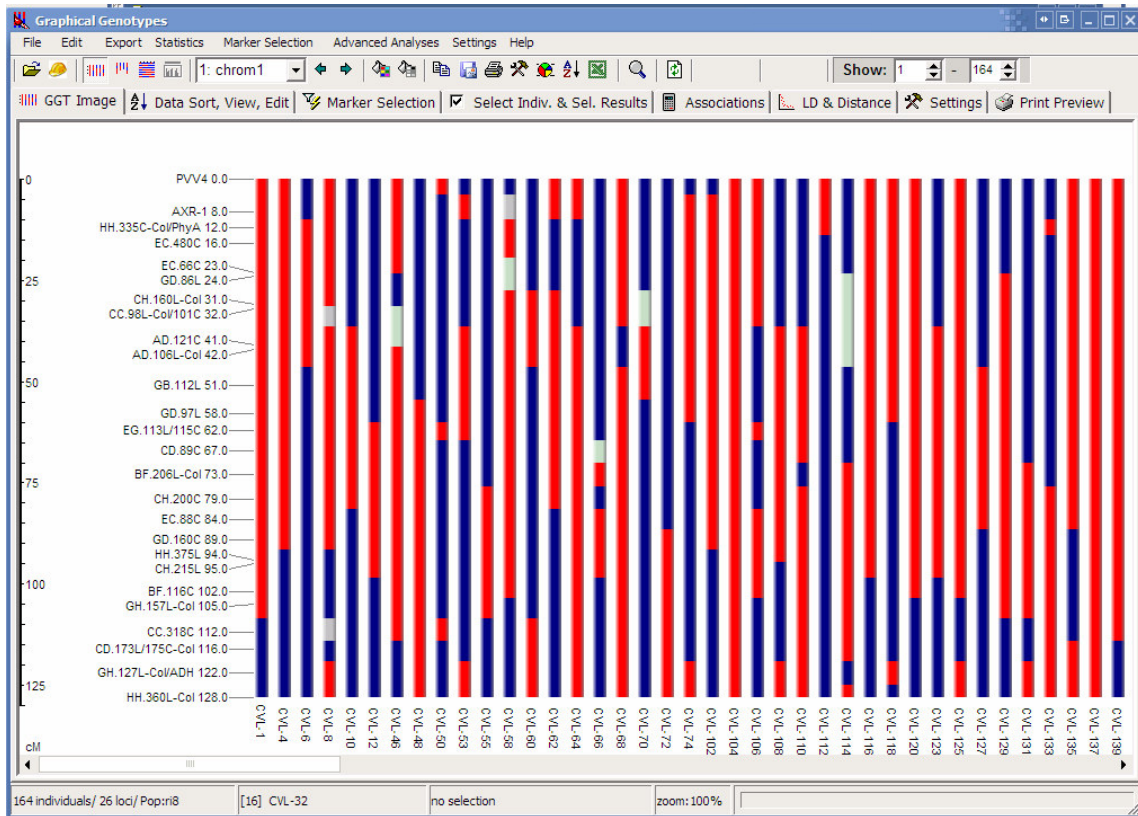


Figure 2: Example of a graphical genotype in 'Linkage group view' mode.

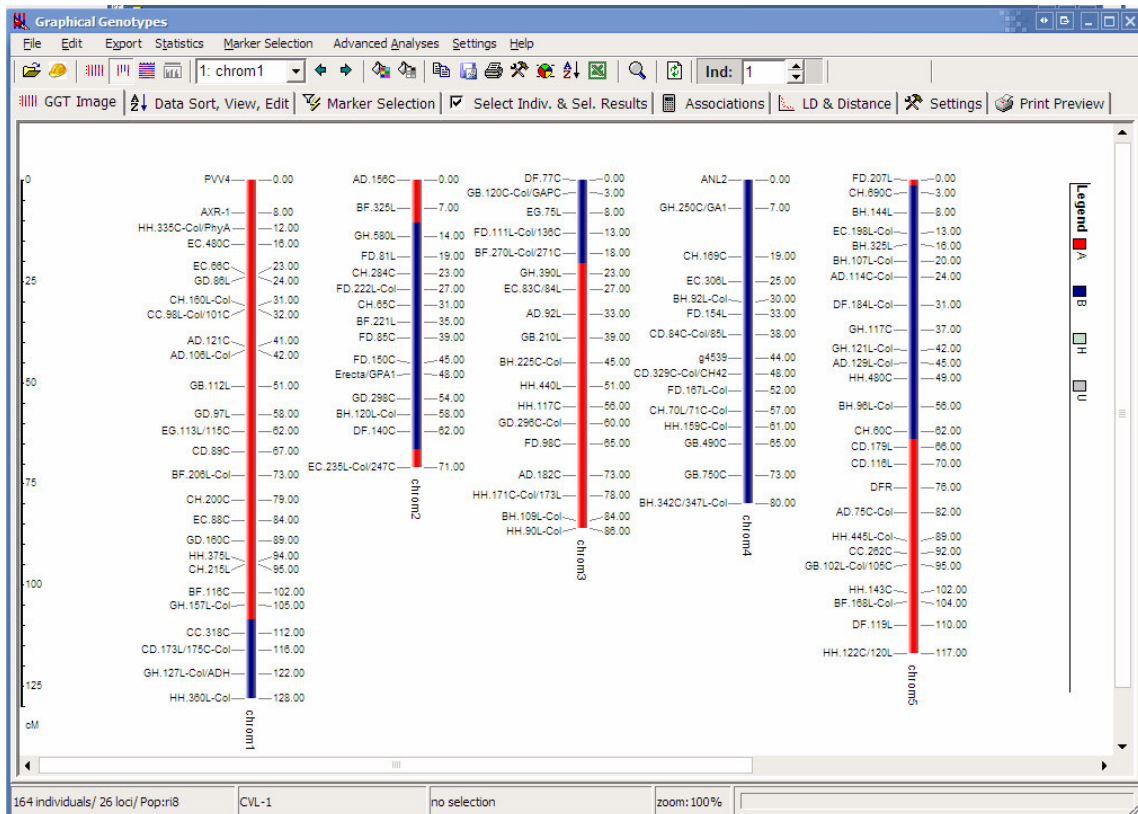


Figure 3: Example of a graphical genotype in 'Individual view' mode.

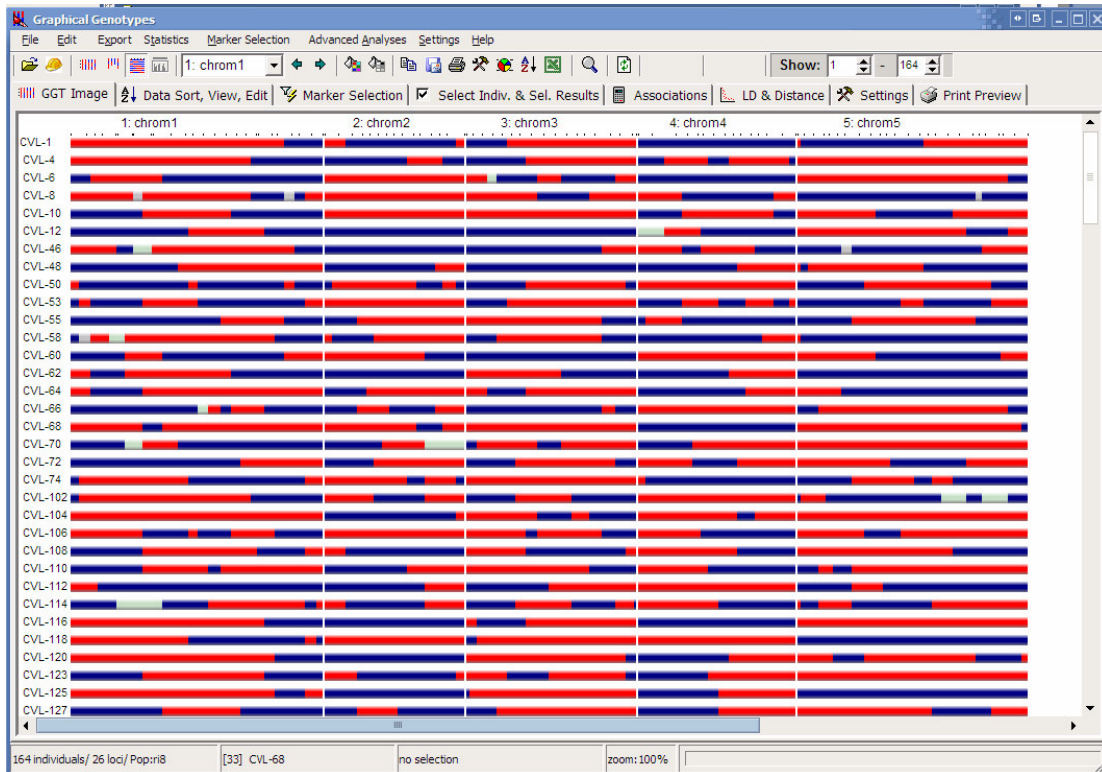


Figure 4: Section of a zoomed out GGT image in the “draw all” modus. This will give a complete overview of the data on all individual for all linkage groups.

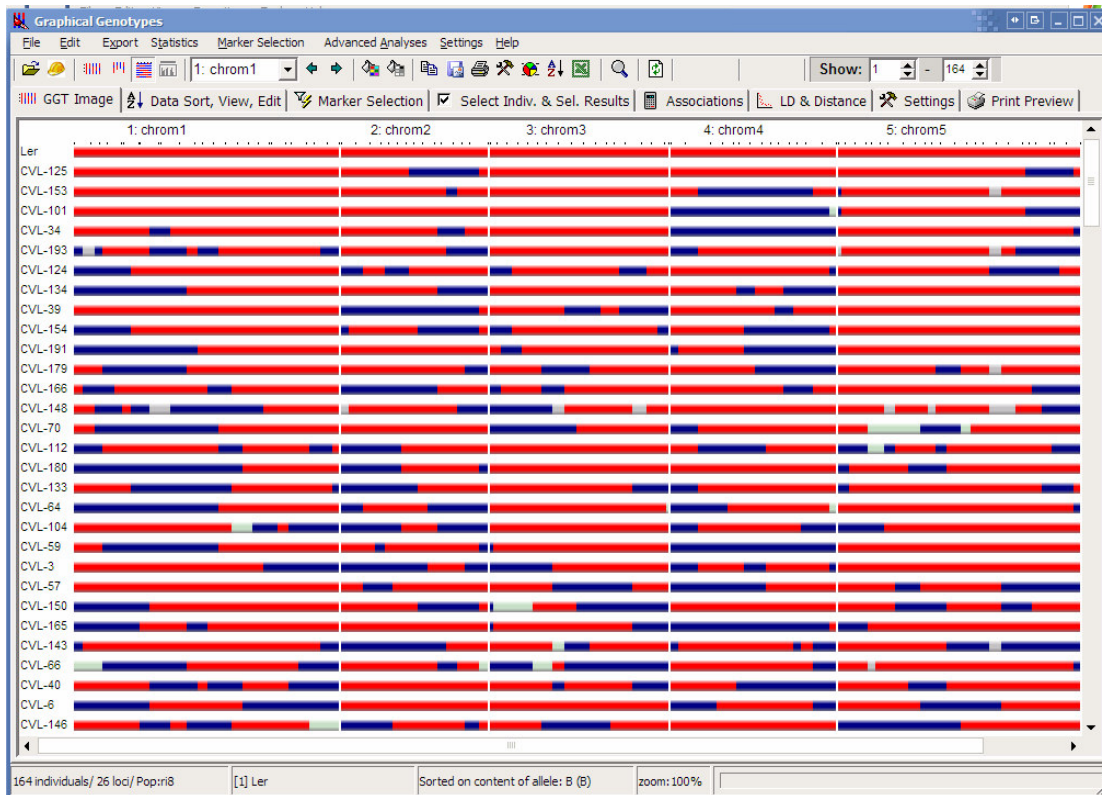


Figure 5: The same data as in fig 4 after sorting for allele content (sort was on the red B-allele, descending).