

Pedimap

Software for the visualization of genetic and phenotypic data in pedigrees

© Plant Research International, 2004-2011
Roeland E. Voorrips (contact: roeland.voorrips@wur.nl)

What is Pedimap ?.....	1
Disclaimer	2
Setting up the program and license file	2
Introduction	2
Interface	3
Operation	5
Starting a new project	5
Views	5
Legend	5
Selecting individuals.....	6
Subpopulations	7
View Options.....	9
Layout.....	9
Individuals	10
IBD (Identity-by-Descent).....	12
Markers	13
Colors	14
Info	14
Default options	14
Projects	15
Data file	15
Header lines	16
Pedigree and phenotypic trait values.....	16
Linkage group data	17
Observed alleles sections	19
IBD probabilities section.....	19
Export.....	20
FlexQTL	20
Add semifounder parents	21
Impute missing founder alleles	22
Drawing limitation	22
Tips & tricks	23
Copy options between views	23
Copy views between subpopulations.....	23
Combine subpopulations.....	23
Names containing spaces, long names	23

What is Pedimap ?

Pedimap is software designed to graphically present genetic information in pedigrees. This information can include phenotypic data, observed marker alleles, calculated Identity by Descent (IBD) probabilities, and (to be added) QTL (Quantitative Trait Loci) allele probabilities. Pedimap is unique among other things in its capability to present some of this information in relation to genetic linkage maps, hence the name.

Pedimap is developed by Roeland Voorrips of the Plant Breeding group of Wageningen University and Research Center. It incorporates many ideas of Eric van de Weg of the same group.

Pedimap may be cited as follows:

Voorrips, R.E. (2007) Pedimap: Software for visualization of genetic and phenotypic data in pedigrees. Plant Research International, Wageningen, the Netherlands.

Disclaimer

This software is distributed “as is”. It has been designed and developed with the wishes of geneticists and breeders in mind and is believed by the author to work as intended, but neither the author nor Plant Research International or any body of which it is part makes any claim or offers any guarantee about its correctness or appropriateness or accepts any liability for material or immaterial damage resulting from the use of this software. Anyone using the software or results obtained by using the software does so at his or her own risk.

Setting up the program and license file

When you download Pedimap from the website you receive a zip-file containing the following files: Readme.txt, Pedimap.exe, PedimapManual.pdf, Example.dat and Example.pmp. You can save these files to any location on your machine, but a logical place would be a new folder “Pedimap” in the Program Files folder. The only files needed to run Pedimap are Pedimap.exe and PedimapManual.pdf. You can run Pedimap by double-clicking on Pedimap.exe in the Explorer, but it is more convenient to create a Start menu and/or desktop shortcut. Once you have run Pedimap for the first time, double-clicking on a Pedimap project file (*.pmp) will run Pedimap and open the project.

Pedimap will run with limitations if you do not have a license file (Pedimap.lic). A free personal license can be obtained from the Pedimap website. It will be mailed to you and must be copied into the same folder where Pedimap.exe is located. Pedimap must be closed and restarted to recognize the license.

If you have multiple versions of Pedimap on your system (not recommended!), be aware that each time one of the versions is started it associates the “.pmp” file extension with itself; i.e. double-clicking on a Pedimap project file will cause that Pedimap version to be started. Also, be sure that the Pedimap version started through a desktop or start menu shortcut is the one you want, and that the license file is in the same directory.

Introduction

Pedimap is a tool for visualizing data about individuals that are part of a pedigree. There are some important words here. First, Pedimap is a visualization tool. Pedimap does not perform statistical or (quantitative) genetic calculations; it just presents available data, including results obtained through analysis using other software, in a way that makes understanding these data easier.

Second, Pedimap presents data about individuals. These data can be divided into two classes: phenotypic and genotypic data. Phenotypic data are those that describe any trait of an individual, such as the color, taste, sex, size, resistance to particular diseases and so on. These can be expressed as numbers resulting from measurements, evaluations resulting in scores e.g. from 0 to 9, or a range of textual labels e.g. yellow, green, blue. It is also possible to add other information, such as the availability or location of a DNA sample, the year of birth etc that strictly speaking are not phenotypic data. Pedimap can handle all of these.

The other class of data about individuals are genotypic data. These consist of marker genotypes observed at specific loci that are arranged on a genetic linkage map, and can also include Identity-by-Descent probabilities calculated for a set of founder alleles at various positions on the linkage map. Both these data types: marker scores and IBD probabilities can be shown by Pedimap. Moreover, phenotypic and genotypic data can be shown in the same image, making it possible to show relations between the two.

Third, Pedimap is intended to work with pedigrees. Pedigrees can be very simple, just consisting of parents and some offspring. However they can also become quite complex, involving thousands of individuals over multiple generations, involving backcrosses with ancestors, sib matings and so on. Progeny can be generated by a cross of two parents, but also by selfing one parent, by vegetative propagation, mutation and doubled haploid induction. All these are handled by Pedimap. In contrast with some other pedigree software any individual is always shown only once in an image, even if it is a parent in multiple crosses. All the offspring of one cross or selfing is always shown as one group.

Because pedigrees can be large and complex it is often useful to view only a part of it. Pedimap offers the possibility to select individuals manually or automatically, based on family relationships with another individual (e.g. all its ancestors and/or offspring to some number of generations, along the paternal, maternal or both lines; its sibs; the other parents of its offspring). Pedimap lets you create projects where such subpopulations, sub-subpopulations

etc. are stored and can be quickly selected. For each subpopulation any number of views can be selected to show different phenotypic traits, markers from different linkage groups etc. Such views can be quickly copied from one subpopulation to another. Finally, subpopulations can be annotated, highlighted and merged.

New data tend to accumulate in breeding programmes, such as new individuals, new observed traits and markers. In Pedimap this is easily handled by importing a new data file into an existing project. All previous views are maintained, and subpopulations are adapted to the new pedigree; e.g. a subpopulation consisting of all the progeny to three generations of a particular individual is recalculated to include all new individuals with that specification as well.

All these facilities make Pedimap a versatile tool, but it may appear quite complex on the first introduction. In this manual we will try to guide you through the user interface and explain the structure of the data file.

Interface

The interface is set up around the concept of subpopulations. From the total population various subpopulations can be selected; for each subpopulation multiple views (graphical presentations) can be shown.

Below the menu and toolbar, the Pedimap screen is divided into three parts: the Main panel on the right and two smaller panels: the Population panel and the Information panel, on the left (Figure 1).

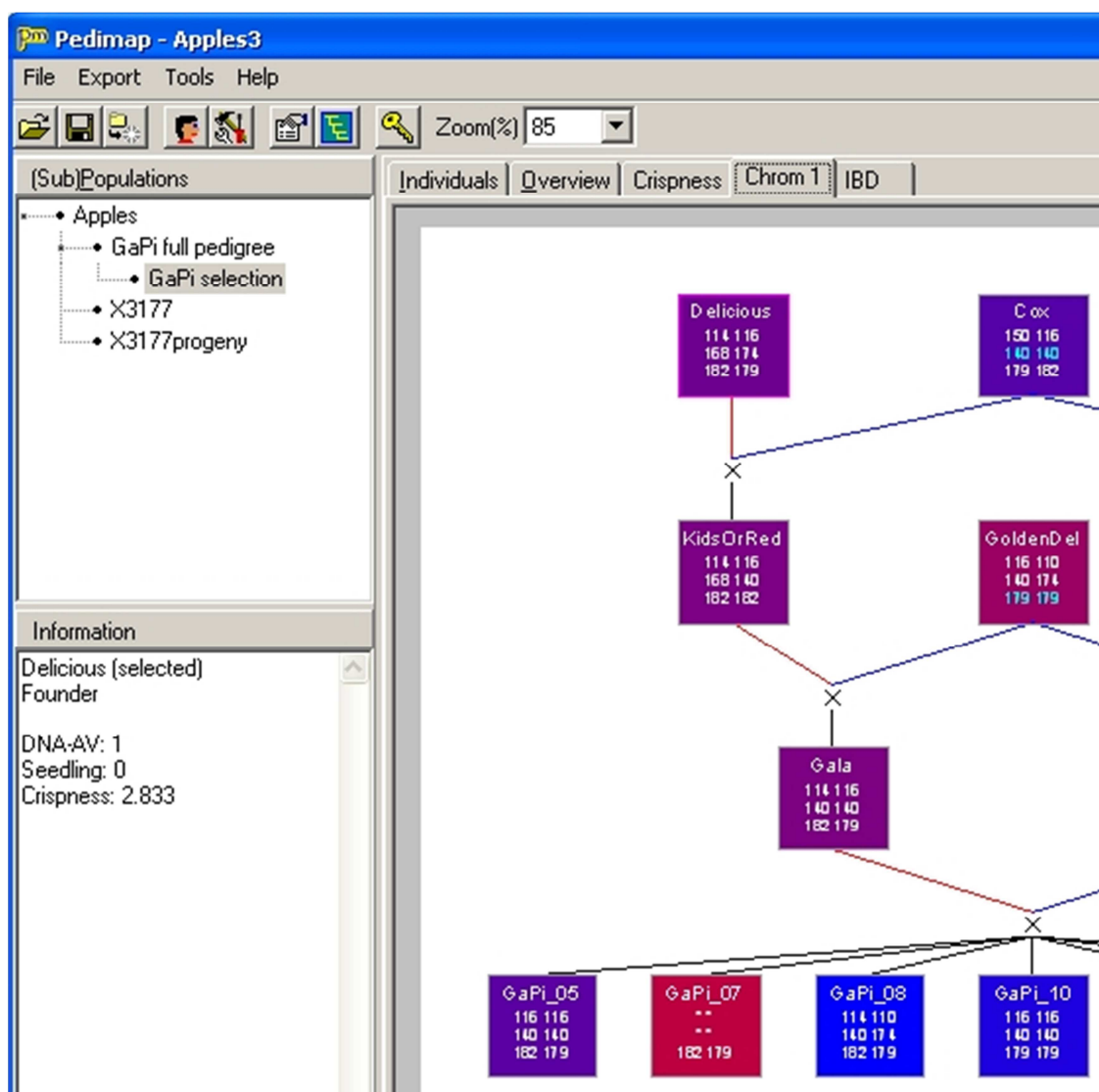


Figure 1. The user interface of Pedimap, with the Main panel on the right, the Population panel top left and the Information panel bottom left.

The Main panel on the right shows the current subpopulation of individuals; this panel has several pages that can be selected using the tabs on top.

The first page ("Individuals") shows a table of the individuals with their parents and phenotypic trait values (Figure 2). This table can be sorted on any column by clicking the caption.

The second page ("Overview") shows the pedigree structure, arranged with the generations from left to right (Figure 3). This is a compact view showing only the individual names. It is intended only for selecting subpopulations to inspect in more detail.

Individuals Overview DNA						
Nr	Name	Female	Male	DNA-AV	Seedling	Crispness
14	McIntosh	*	*	1	0	2.894
15	Melba	McIntosh	LivellRasp	1	0	2.461
16	NJ130	*	*	0	0	*
17	NorthSpy	*	*	1	0	3.712
18	Q53T136	*	*	1	0	2.446
19	Platt_Malba	*	*	0	0	*
20	PRI14-126	GoldenDel	F2_26829-2-2	1	0	*
21	PRI14-152	GoldenDel	F2_26829-2-2	1	0	*
22	PRI14-226	GoldenDel	F2_26829-2-2	0	0	*
23	PRI14-510	GoldenDel	F2_26829-2-2	0	0	*
24	PRI47-147	Jonathan	F2_26829-2-2	0	0	*
25	PRI610-2	McIntosh	PRI14-226	0	0	*
26	PRI612-1	Delicious	PRI14-126	1	0	2.984

Figure 2. The Individuals page contains a list of all individuals in the current subpopulation, with their parents and phenotypic data. It can be sorted for any column.

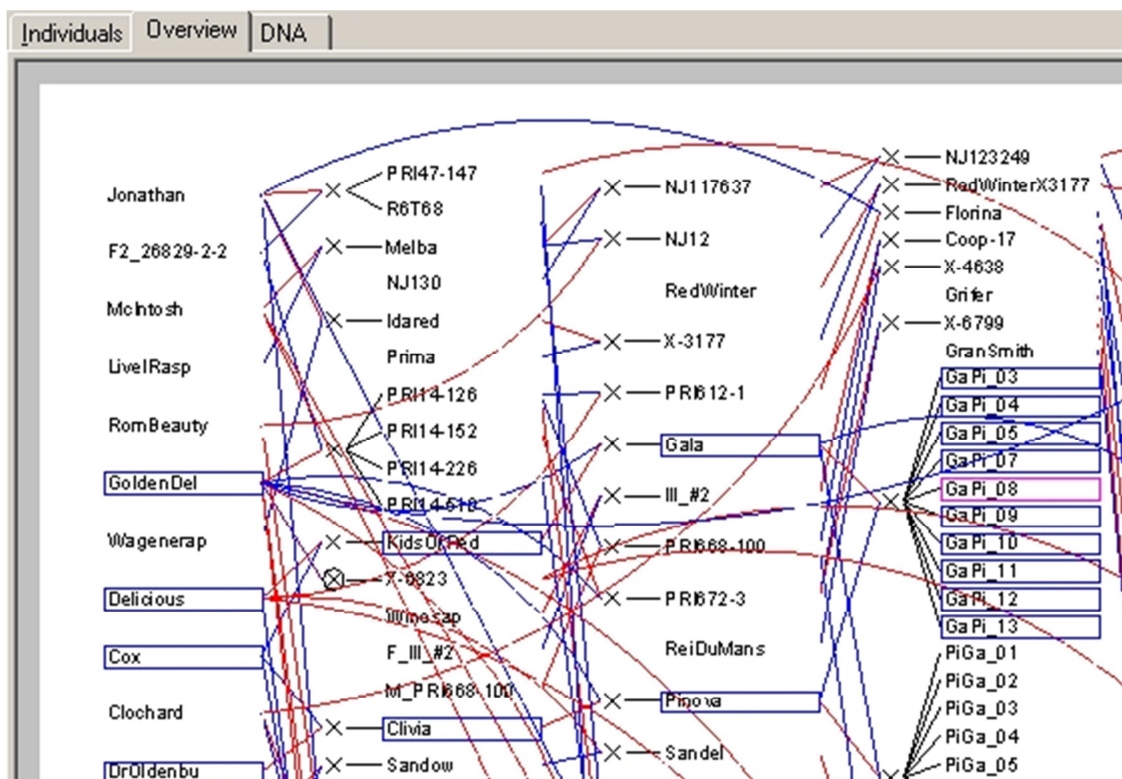


Figure 3. The Overview page shows a condensed view of the selected subpopulation with the generations arranged from left to right. Several individuals are selected (the sibs and ancestors of GaPi_08).


The other pages, if present, show different views of the pedigree structure, which can be customized using the Options dialog (see below).

To the left of the Main panel two smaller panels are shown. The Population panel at the top shows a tree-view of the population and all subpopulations that have been created (Figure 1). Here you can select one of these selections to be displayed in the main panel. After starting a new project only the full population is shown here.

The Information panel at the bottom left (Figure 1) shows information on the individual or cross under the mouse cursor in the pedigree views of the main panel.

Operation

Starting a new project

Start Pedimap and select File – Import or button  to import a data file. In this manual we select the file Example.dat in the download package. If the data are incorrect an error message will appear. In particular, if there are errors in the pedigree involving circular relations (e.g. two individuals being each other's ancestor) a file is exported listing the individuals involved in the loops.

When loading is completed and all is well, the Population panel shows the name of the population, and the Main panel contains two pages: the Individuals and the Overview page (Figure 4).

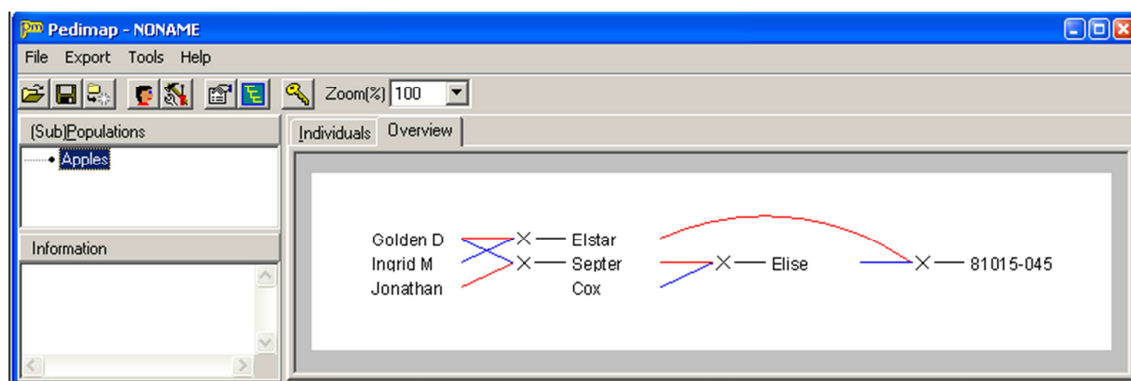



Figure 4. The Pedimap window after importing the Example.dat data file

For larger data sets loading can take considerable time; in that case the progress is indicated. However the first time a data file is loaded a binary file is created (in the same directory, with the same file name as the data file and extension .pmb); this results in a much faster loading of the data in subsequent sessions.

For pedigrees or subpopulations with more than a specified number of individuals (default: 1000), no Overview page is created to save system resources.


Views

In order to create extra, customizable views of the population or of any subpopulation, choose Tools – Create View or button . A new page (“View1”) will appear in the Main panel.

This new view shows the population using the default options (Figure 5). The view can be customized using the Options dialog (see next chapter). Repeat the Create View command to create as many views as required.

By right-clicking on the selected View tab (except Individuals or Overview) a pop-up menu appears. This menu allows to delete or rename the View, to copy the current view options or to paste view options that were copied earlier.

Legend

The legend is a small window that can be shown by choosing Tools – Show legend or button . It contains information on the current view, including (depending on the selected options) the trait colors, the IBD colors, the linkage group map or the selected markers.

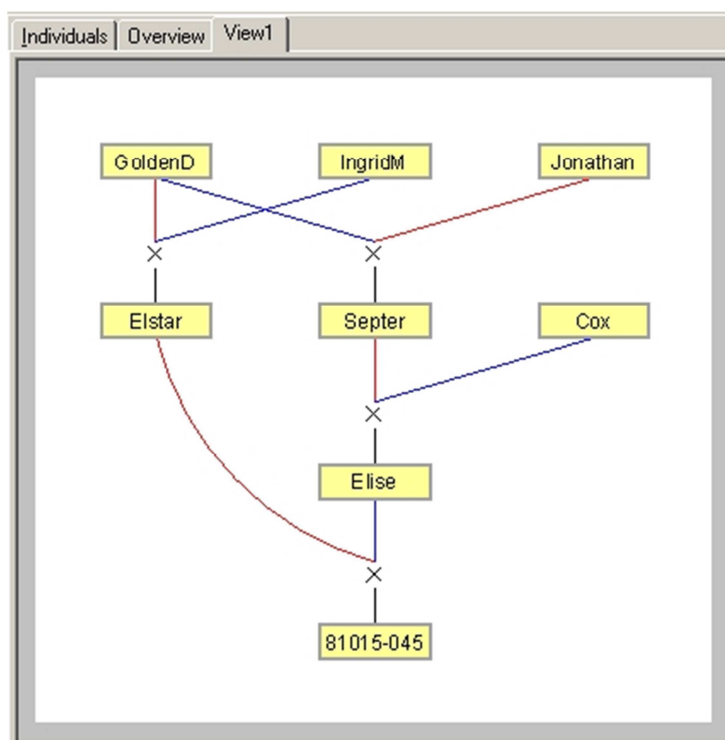



Figure 5. The default view generated when Tools – Create view is selected.

Selecting individuals

Individuals can be selected by clicking on them in any of the views and/or in the Individuals table. The selection is maintained when changing between views and/or table. Selected individuals are shown with a blue border in the pedigree views. The currently focused individual is shown with a purple or red border, depending on whether it is part of the selection or not.

Figure 6. The Select relatives dialog.


Simple clicking on an individual selects only that individual and de-selects all previously selected individuals; clicking while holding down the Control key only selects or deselects that individual while leaving the remainder of the selection unchanged.


Further, individuals can be selected or de-selected based on their family relationship (ancestor, progeny, sib) with the focused individual. For this, choose Tools – Select relatives or button . This shows the Select relatives dialog (Figure 6).

The Select relatives dialog can be used to extend or reduce the current selection of individuals based on pedigree relations with the currently focused individual. At the top a choice is offered to Select additional individuals to be included in the current selection (it does not matter if some of them are already in the selection), or to Deselect some of the individuals in the current selection. Then the criteria for selecting or deselecting individuals can be specified. The number of ancestor and progeny generations from the currently focused individual are chosen, and the tracing of the ancestors and progeny can be done through the male or female line or through both lines.

The “Include progeny parents” checkbox allows to not only select the direct descendants, but for each descendant also the parent that is not part of the line from the focused individual (sometimes this is referred to as the “cold side” of the family). Finally the full sibs can be selected.

After pressing the OK button the specified individuals are added to or removed from the current selection. The resulting selection can again be modified manually (by Control-clicking individuals) or automatically by again invoking the Select relatives dialog.

Finally a new subpopulation with the selected individuals can be created by choosing Tools – Create subpop or button .

In our example, we click on Septer and then on button  to show the Select relatives dialog. In the dialog we set both Ancestor and Progeny generations to 1, trace through both lines, check the Include progeny parents and clear the Include sibs boxes, and press OK. The resulting selection is shown in Figure 7. Note that you can switch between the pages Individuals, Overview and View1 and see the same selection highlighted in each case.

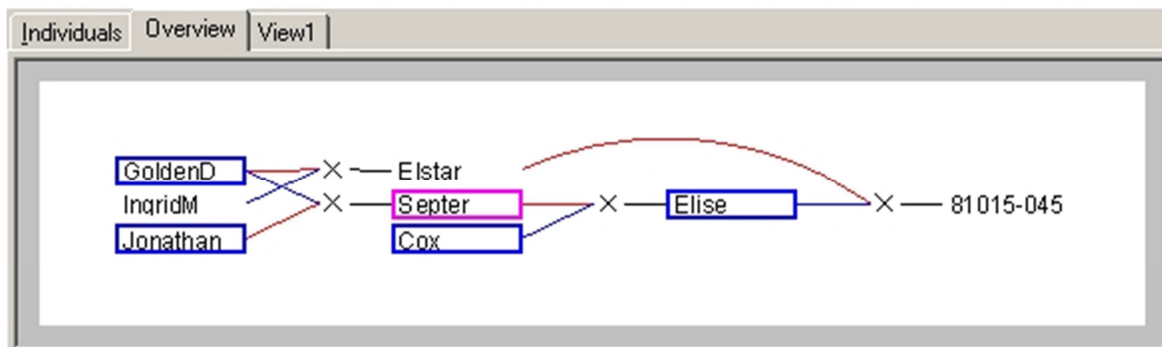




Figure 7. A selection highlighted in the Overview page.

Subpopulations

In the Population panel the current (sub-)population is highlighted. A different subpopulation can be selected by a left-click with the mouse. Using the Control key it is possible to select multiple subpopulations where the last clicked will be the “current” population; this is useful for combining individuals from different subpopulations or to delete multiple subpopulations in a single operation.

A new subpopulation containing only the selected individuals can be created by choosing Tools – Create Subpop or button . A new entry appears in the Population panel as a branch from the previous (sub)population, and in the Main panel an Individuals and an Overview page are created for the new subpopulation (the Overview is suppressed if the subpopulation contains more than a specified number (by default, 1000) of individuals). Creating views and selections in subpopulations are done exactly as described above for the full population; from a subpopulation again subpopulations can be created.

In our example, with the selection from the preceding section still highlighted, press . In the population panel a new selection Apples-1 is shown, and on the Main panel the Overview of this subpopulation appears.

The name of a (sub-)population can be changed by selecting it in the Population panel and clicking again on the name. In our example we enter the name “Septer relatives”.

By right-clicking on the name a pop-up menu appears with the following options:

- Delete all selected subpopulations (and all the Views and derived subpopulations)
- Info: display and change information on the subpopulation. This allow also to set a “flag” shown as an exclamation mark in front of the subpopulation name.
- Mark the (sub-) population as a source from which Views can be copied to other subpopulations
- Add the Views from the current source population. In this way it is easy to create different subpopulations with the same set of Views.
- Order the subpopulations of the current (sub-)population in alphabetical order or in original order (the order in which they were created)
- Export the contents of the (sub-)population for import by FlexQTL
- Combine subpopulations: create a new subpopulation directly under the full population, containing all individuals present in the selected subpopulations

In our example, we right-click the “Septer relatives” subpopulation and select Info. This brings up the Info dialog (Figure 8). The top of this panel shows the selection history: the criteria used (in the previous section) to select the individuals for this subpopulation. The “Base individual” (Septer) is the individual that was focused when the Select relatives dialog was activated. We see than one ancestor generation and one progeny generation was chosen, and that the “progeny parents” were included.

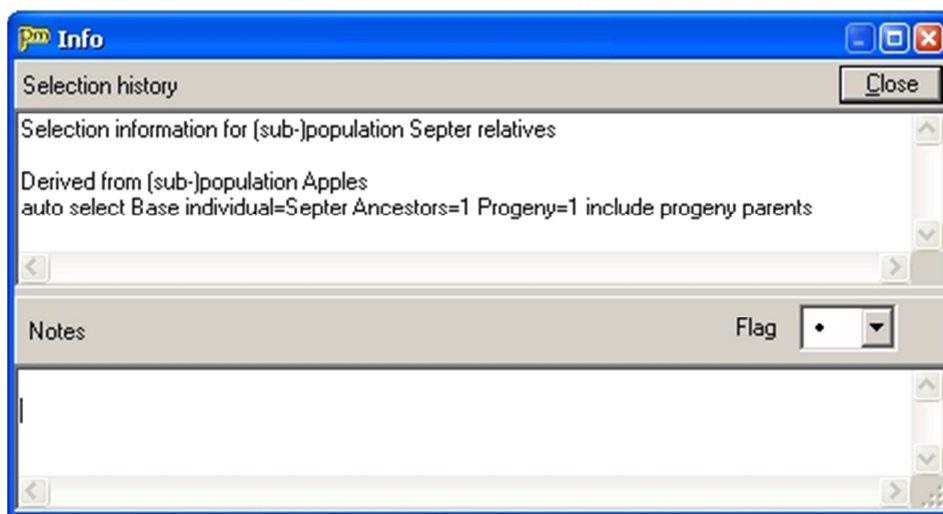


Figure 8. The Info dialog for subpopulation Septer relatives.

Next we see a drop-down box “Flag”. Here we can select another symbol (several colored exclamation marks) to highlight this subpopulation in the Population panel. Finally there is a Notes panel, where any notes on this subpopulation can be entered.

We select a red exclamation mark as Flag and enter as notes: “Immediate relatives of Septer”, and click the Close button. In the Population panel now the Septer relatives node is highlighted (Figure 9).

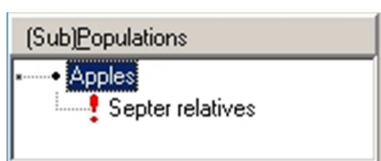


Figure 9. The Population panel with the “Septer relatives” node highlighted.

View Options

Various options determine how the data are displayed graphically. These options can be set through the Options dialog (select Tools – Options from the main menu). The options dialog is organized into several pages:

Layout

On this page the general appearance of the pedigree is specified (Figure 10).

The first option is to show the pedigree oriented with the generations from top to bottom or from left to right.

The second option is whether cross symbols are shown, and what their size is. With cross symbols, one or two lines connect the parent(s) to the cross symbol and the cross symbol is connected to each of the progeny. Without cross symbols, each progeny is connected directly to its parent(s).

Further the spacing between the generations and between individuals within the generations is specified here.

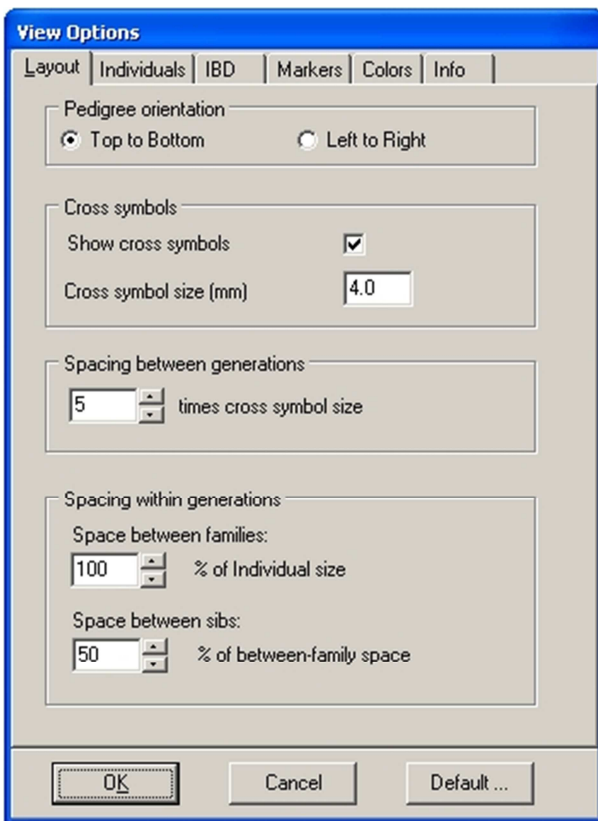


Figure 10. The View Options dialog with the Layout page selected.

Individuals

Here is specified how individuals are shown (colors, font used for name), and which extra information is shown in each individual (Figure 11).

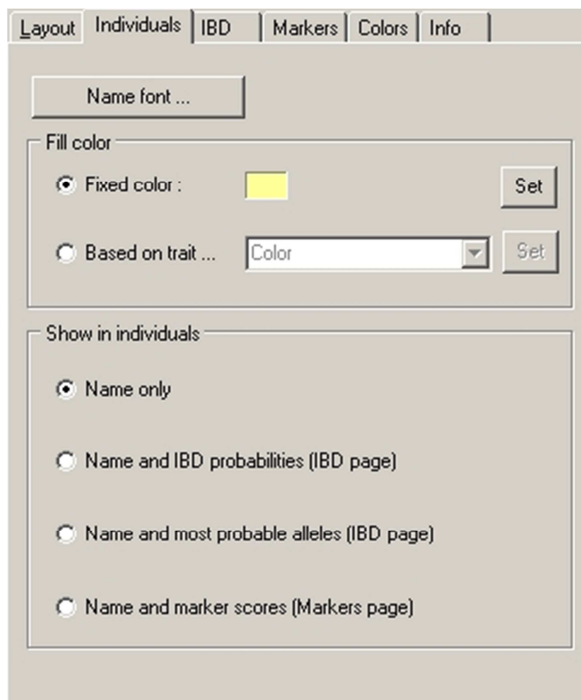
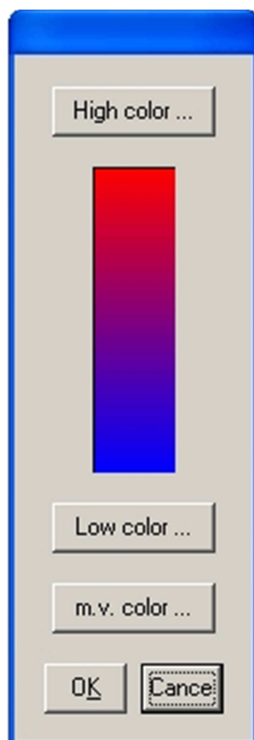


Figure 11. The Individuals page of the View Options dialog.

Depending on the information available in the data file, some of the options on this page may be disabled. Figure 11 shows the default individual options, which result in an image as shown in Figure 5.





The Name font button calls up a further Font dialog. The next option is the Fill color of the individuals. This can be the same color for all individuals (to be selected by clicking the Set button), or it can be based on the individual's value for any of the phenotypic traits. If this option is chosen the colors for the trait values may be specified by clicking the Set button. This calls up a further dialog where either the color for the minimum and maximum value can be set (for continuous variables) or where a color for each possible trait value can be set (for discrete variables). (A trait is considered discrete if all non-missing values are single characters or if at least one value is not a number. For example, traits with values A...F, or 1...9, or green-yellow-red are discrete, but traits with values 1...50 or -2...2 or 2.5...7.4 are treated as continuous.)

Finally the contents of the individuals are specified: names only, or names with other information. In those cases further options must be set on any of the next pages, which are automatically shown when the corresponding option on this page is clicked.

Figure 12. The color selection dialog for continuous traits.

In our example we shall create two Views for the “Septer relatives” subpopulation, one where the individuals are colored according to their length and one where they are colored according to the color property.

In the Population panel select the Septer relatives node and press  to create a new View. This view will represent the lengths, so we rename it accordingly: right-click on the tab (it says “View1”), select Rename view from the pop-up menu and enter Length as the new name of the view. Next click  to show the View Options dialog, and on the Individuals page of the dialog set the Fill color to Based on trait. In the drop-down box select trait Length. Click on the Set button next to it; this will display the color selection dialog for continuous traits (Figure 12).

In this dialog three colors can be specified by clicking the buttons for High color, Low color and “m.v. color” (the color used for individuals with a missing value for the trait). In this case we accept the default settings so we click Cancel and in the View Options dialog we click OK.

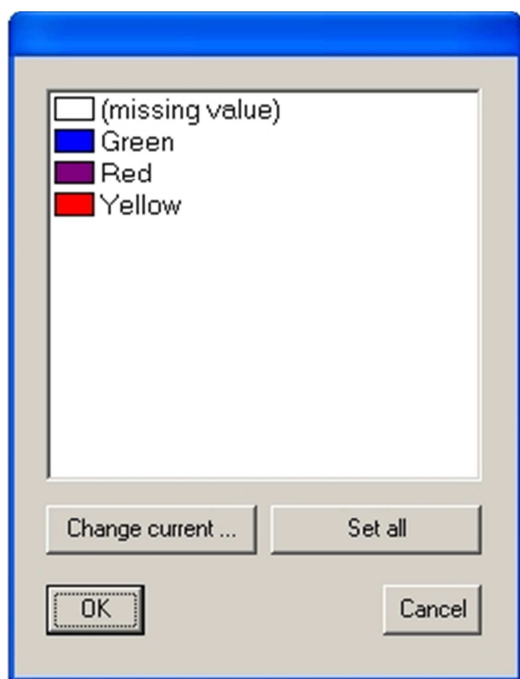


Figure 13. The color selection dialog for discrete traits.

Next we again create a view, which we rename to Color, and we show the View Options dialog. On the Individuals page we again set the Fill color to Based on trait, and in the trait drop-down box we select Color. We click the Set button and see the color selection dialog for discrete traits (Figure 13).

In the data file, color is coded by a few textual labels: Green, Red and Yellow. The default colors are not appropriate so we adjust them. Click the second line (Green) and click Change current. In the color dialog select a green color and click OK. Do the same with the lines for Red and Yellow, press OK and in the View Options dialog again OK. The Main panel should now look as in Figure 14.

Moving the mouse over the different individuals you will see that the Color property shown in the Information panel (lower left of the Pedimap window) matches the fill color of the individuals.

The Set all button in the color selection dialog has no function in this case. However with discrete traits that are coded with a single character (e.g. 1...9) it is possible to set the lowest and highest value color and to fill in the intermediate value colors similarly to the continuous color case using the Set all button.

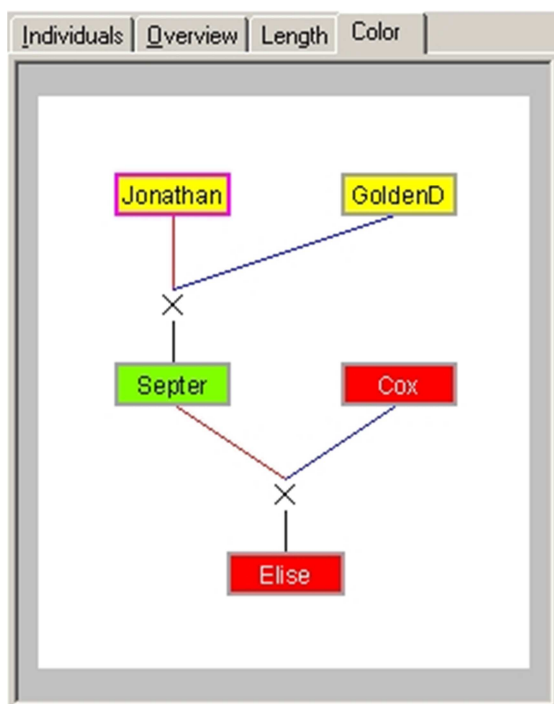


Figure 14. The Main panel with the Color view of subpopulation Septer relatives displayed.

IBD (Identity-by-Descent)

If on the Individuals page the options are selected to show the IBD probabilities or the most probable alleles, the IBD page must be used to set the relevant options (Figure 15).

Figure 15. The IBD page of the View Options dialog.

If the IBD probabilities are to be shown the options at the top and bottom of this page are relevant. The IBD probabilities are depicted in the form of one rectangle for both haplotypes (copies) of a selected linkage group (or more than two haplotypes in case of polyploid data

sets). At the top of this page the linkage group can be selected and the size of the rectangles is set. The display of IBD probabilities as rectangles for each haplotype is explained below.

At the bottom of the page the colors for each founder allele can be set. There are three buttons that set the colors of all founder alleles simultaneously. "Homoz. founders" assigns the same color to all haplotypes in a given founder and different colors between founders. "Heteroz. founders" assigns a different color to each founder, and within founders the different haplotypes are assigned lighter and darker shades of the same color (this is the default). "All same color" assigns one color to all haplotypes of all founders; this is useful if one is interested in the inheritance of just a few founder haplotypes that can then be given different colors. Assigning a new color to a specific haplotype is done by selecting the founder haplotype in the drop-down box at the bottom and then clicking the Set button.

If the most probable allele (based on the IBD probabilities) is to be shown for specific markers the Probability rectangles part of the page is not relevant. As before the linkage group is selected at the top. The Allele font button allows to specify the font in which the allele names are shown. The Select markers button shows a dialog where the markers for which the most probable alleles are to be shown can be selected (the list can be sorted by marker name or by position by clicking the column captions). The probability thresholds specify the thresholds that separate the three presentations of the allele names: bold, normal or italic. As before, the colors used for the different founder alleles can be set in the bottom part of the page.

Markers

If on the Individuals page the option to show marker scores is selected, the Markers page must be used to specify how they should be shown (Figure 16).

Order markers: ☐ by Name ☒ by Position

Marker	Linkage Group	cM
<input type="checkbox"/> SSR1	A	0.0
<input type="checkbox"/> SSR2	A	20.0
<input type="checkbox"/> SSR3	A	29.0
<input type="checkbox"/> SSR10	B	0.0
<input type="checkbox"/> SSR103	C	0.0

Check All Clear All

Marker font ...

Set marker code color: 0

Change ...

Figure 16. The Markers page of the View Options dialog.

In contrast to the IBD presentations where only one linkage group at the time can be viewed, the marker scores can be viewed for any selection of markers. The list can be sorted by marker name or position (linkage group and position within linkage group) using the selection buttons at the top. The Marker font button allows to specify the font in which the marker alleles are shown.

Further it is possible to assign a color code (from 0 up to any desired value) to each marker score in the data file. The Set marker code color drop-down list and Change button allow to set the color for each of the color codes occurring in the data file. By default, different colors are specified for codes from 0 to 9; all codes higher than 9 are shown in gray. However all code colors can be redefined here.

Color codes for marker scores can be used for many purposes. One that is used in the interaction between Pedimap and the FlexQTL software (see introduction) is that FlexQTL

assigns colors based on inferred inconsistencies in the data. This is very helpful during data validation, where FlexQTL determines inconsistencies and other problems that can then be quickly visualized by Pedimap.

Colors

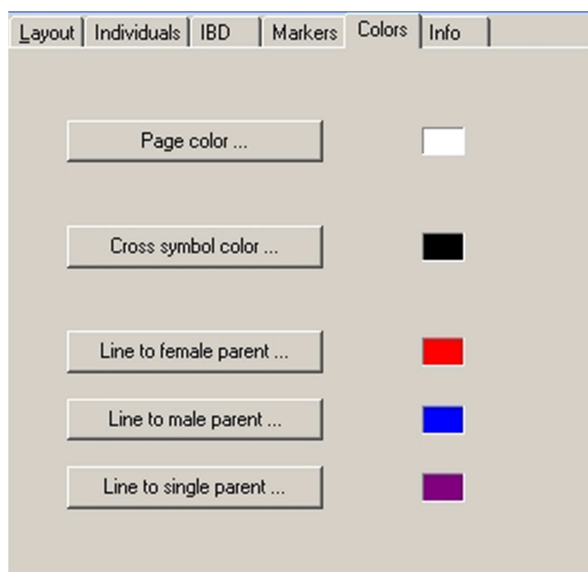


Figure 17. The Colors page of the View Options dialog.

This page allows to set the colors of various chart elements. Most are self-explanatory. In case of a cross the two parents can have connecting lines of different colors (by default red for the female and blue for the male parent). Non-cross types of procreation (selfing, vegetative propagation, mutation and doubled-haploid induction) involve only one single parent for which the color by default is purple.

Info

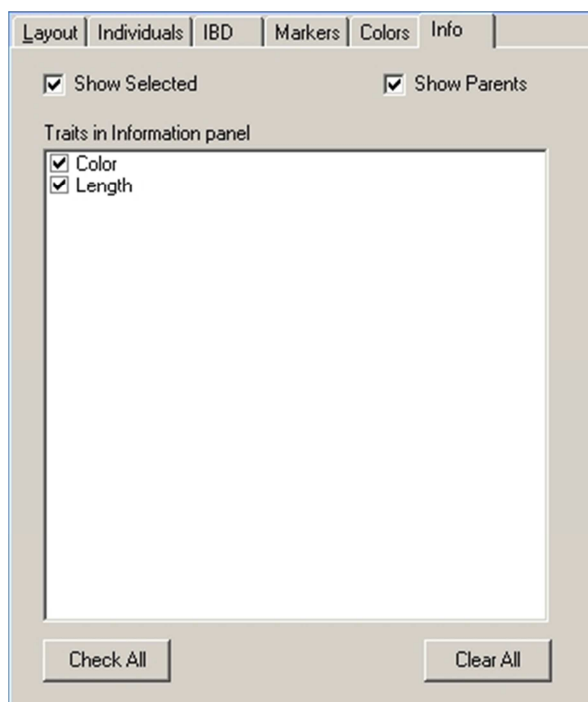


Figure 18. The Info page of the Options dialog.

On the Info page of the Options dialog the information is specified that appears in the Information panel when the mouse is positioned over an individual in the Main panel (Figure 18).

The Information panel (at the lower left of the Pedimap window) shows the name of the individual, and optionally whether it is part of the current selection, its parents, and selected traits. By default all the information is shown. Note that parents are listed here even if they are not part of the current subpopulation.

Default options

At the bottom of the View Options dialog (Figure 10) there is a button Defaults. This button displays a dialog (Figure 19) that can be used to save the options of the current View as default options (meaning that newly created Views will have these same options), to replace the options of the current View with the pre-set options (the default options when Pedimap was first installed) or to replace the options of the current View with the previously saved default options.

Some issues with default options: (1) the saved options do not include project-specific data, such as the markers or traits which are to be displayed. And (2), the lengths of some lists of color values (IBD colors, marker code colors etc) is project specific. When working in a project with a different number of items, the saved default lists are ignored or only used partially.

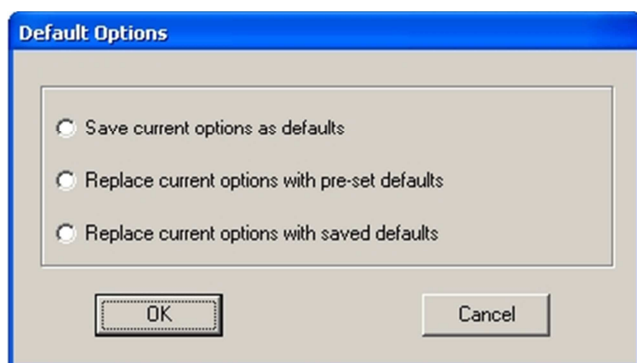


Figure 19. The Default Options dialog.

The default options are saved in a file Options.ini in the Pedimap program folder. A simple way to revert from the last saved default options to the original default options is to delete this file. Note that it is not possible to save default options if you have no write access to the program folder.

Projects

The complete structure of population, subpopulations and views are saved as a Pedimap project file, with extension “.pmp”. This file does not contain the data but instead refers to the data file. Therefore, if the data file is replaced or changed, the project changes accordingly. This offers a possibility to add new data (new individuals, markers or traits) or to correct scores. If new individuals were added to the pedigree, these will also be included in the appropriate subpopulations. For instance, if a subpopulation was derived by automatic selection of all progeny of individual X, and the new data file contains extra offspring of X, these will also be included in the subpopulation.

In some cases Pedimap will not be able to completely re-create the project if the data file is changed, if individuals, traits, linkage groups or markers were removed from the data file. For example, if a subpopulation was originally created by selecting all ancestors of individual Y, and Y doesn't occur any more in the new data file, Pedimap won't create this subpopulation when re-opening the project. Similarly, if a View refers to a particular phenotypic trait, linkage group or marker which doesn't occur any more, the corresponding View Options are reset to their default settings.

It is also possible to replace the data while a project is open, by Importing a new data file. Pedimap will warn you that you are about to replace the data in your current project and will ask for confirmation. Note that if the new data file contains errors you won't be able to restore the original project, so it is a good idea to save your project before importing a new data file.

The project file (*.pmp) can be viewed with Notepad or any other text editor. Usually there is no need to view or edit the project file directly. However if you move the data file to another folder you can edit the second line of the project file to add the correct path to the filename. The same can also (and preferably) be accomplished by importing the data file from its new location. The only reason to edit the project file instead would be to avoid re-parsing of the data file which might take some time if the data file is very large.

Please note that the format of the project file is very strict, so don't make other changes, such as adding or deleting empty lines, spaces etc. In any case it is a good idea to keep a copy of the original project file.

Data file

All data needed to create a new Pedimap project are imported from a data file. At present this is an ordinary text file, which must be created according to a specific format.

Note that in the following “string” means any sequence of characters, excluding blanks (spaces, tabs etc). In order to include spaces in a string (e.g. in individual, trait or locus names) it must be enclosed in “double quotes”.

Keywords can be entered in any case; individual, locus, allele and trait names are case sensitive.

Numbers with decimals (phenotypic trait values, locus positions and IBD probabilities) may be given with a decimal point or a decimal comma, irrespective of Window’s regional settings. All other numbers in the file are integers and must be given without decimal point or comma. Do not use thousands-separators in numbers.

Comments can be entered in the data file but are ignored by Pedimap. Comments start with a semicolon (“;”) followed by a space. Comment lines (lines where the first non-blank character is a semicolon) and blank lines can be inserted anywhere in the data file. Comments can also be added after the end of the data on a line, but not between data. Apart from giving information on the data, such as the date, references to experiments or source files, comments can also be used to (temporarily) delete individuals or other information. However removing information may lead to inconsistencies such as missing parents etc.

Header lines

The file starts with a few header lines. Depending on the other information in the file, some or all of the header lines may be omitted.

If present they contain a keyword (not case sensitive), “=”, a value (spaces are ignored)

The header lines that are present may appear in any order:

1. POPULATION = population name ; if not present, the filename is used.
2. UNKNOWN = <strings, separated by spaces> ; if not present, “-” is assumed. All strings listed here are interpreted as missing or unknown values when reading the data file; only the first string is used in the Individuals list and Info panel. Note that here the strings may not be enclosed in “double quotes” and may not contain spaces.
3. NULLHOMOZ = <character> ; if not present, “\$” is assumed. Used to represent the situation where the second marker allele is either a null allele or equal to the other allele
4. CONFIRMEDNULL = <string> ; if not present, “null” is assumed. Used to represent confirmed null alleles
5. NALLELES = <number> ; if not present, 0 is assumed. This is the number of founder alleles (equal for all loci). NALLELES must be >0 if Founder alleles and IBD allele probabilities are given, else if present NALLELES must be 0 (or omitted).
6. PLOIDY = <number> ; if not present, 2 (diploid) is assumed

All header lines may be omitted if the default values are used and no IBD information is present.

Pedigree and phenotypic trait values

The Pedigree section starts with a line on which the first word is “PEDIGREE”; the rest of this line is ignored. Then follows a caption line starting with the words “NAME”, “FEMALE” and “Male” (or vice versa), followed by the names of the phenotypic traits, all separated by blanks (spaces or tabs). For compatibility with earlier versions, instead of FEMALE and MALE also MOTHER and FATHER or PARENT1 and PARENT2 may be used. FEMALE and MALE may also be reversed (in plant genetics it is customary to start with the female parent, while in animal and human genetics usually the male parent is listed first).

The pedigree is specified with one individual per line. Each line contains at least three entries: the names of the individual and its first and second parent. If one or both parents are unknown, an UNKNOWN string is given (default -, see Header lines section). If a name contains spaces, it must be enclosed in “double quotes”.

Several types of uniparental descent are recognized:

- Selfing: second parent is identical to first, or second parent replaced by keyword *SELF
 - Doubled-haploid induction: second parent replaced by keyword *DH
 - Mutant: second parent replaced by keyword *MUT
 - Vegetative propagation, clone: second parent replaced by keyword *VP
- (note the asterisk at the start of each of these keywords). After the second parent (or keyword) the values of the traits are listed, again with the UNKNOWN string (by default: -) for missing values. Phenotypic values can be any data: single characters, strings, or numbers. Each line must contain the same number of phenotypic data (including missing values) as listed in the caption line. All items on the line are separated by spaces and/or tabs.

The pedigree may be given in any order; Pedimap will sort the pedigree such that individuals always appear after their parents. (If the pedigree is already given in this order no sorting takes place.) During the sorting, the order of sibs in full-sib families is conserved. The individuals are numbered on the Individuals page in sorted order.

Common errors in pedigree data

- Names containing spaces are not enclosed in “double quotes”
- Spelling of name of the same individual varies. Note that this spelling must match exactly, including upper / lower case of all letters
- The same individual is listed more than once
- Individuals occurring as parents are not listed
- Incorrect parents specified: if a circular reference occurs (e.g. two individuals being each other's ancestor) the pedigree cannot be sorted

Table 1. The header lines and pedigree in Example.dat

```
; Example of a Pedimap data file. All keywords are shown in UPPERCASE

POPULATION = Apples          ; if not specified, the file name is used
UNKNOWN      = - *           ; both - and * are accepted as input but
                               ; only the first is used for display
                               ; if UNKNOWN not specified, - is used
NULLHOMOZ    = $             ; this is the default
NALLELES     = 8              ; for 8 founder-alleles per locus the IBD
                               ; probabilities are given
PLOIDY       = 2              ; this is the default

PEDIGREE
NAME          FEMALE      MALE      Color      Length
81015-045     Elstar       Elise      Green      12.3
Cox           -           -          Red        18.0
Elise         Septer       Cox       Red        23.9
Elstar        GoldenD      IngridM   Red        17.6
GoldenD       -           -          Yellow     20.1
IngridM       -           -          Red        11.4
Jonathan      -           -          Yellow     14.8
Septer        Jonathan     GoldenD   Green      22.0
```

Note that except UNKNOWN and NALLELES all header lines could have been omitted as they specify the default values. Also note that the pedigree is not sorted (not all parents occur before their offspring). When importing this file the order on the Individuals page will be different. Also note that while the names are those of apple varieties, their pedigree relations and phenotypic and genotypic data in this example are fictitious.

Linkage group data

After the Pedigree section, a sequence of linkage group sections may follow. A linkage group section starts with a line with the keyword LINKAGEGROUP followed by the linkage group name. Each linkage group section contains one Map section, followed by Locus sections and/or an IBDPOSITIONS statement. The Linkage group section lists the general information about loci and IBD positions in the linkage group; data for individuals are listed in the Observed allele sections and the IBD probabilities sections that appear after the last Linkage group section.

Map section

The map section is always required and starts with a line with the keyword MAP. The following lines contain the locus name followed by the locus position (usually in centimorgan) along the linkage group. The loci must be listed in ascending order of position.

Locus section

For each locus specified in the MAP section, a Locus section must be present. A Locus section starts with a line with the keyword LOCUS followed by the locus name. This name must be identical (upper/lower case relevant) to the name used in the MAP section.

The Locus section contains one or two statements depending on the type of data present.

Allele names statement

The allele names section consists of a single line with the keyword ALLELENAMES, followed by the allowed allele names, separated by blanks. This list may not include UNKNOWN or NULLHOMOZ strings (see Header section). Preferably also the CONFIRMEDNULL string should not occur, but for compatibility with earlier version this is tolerated at present. The Allele names section is always required. All allele names mentioned in the Founder alleles statement and in the Alleles section (see below) must appear in the Allele names section.

Founder alleles statement

The Founder alleles statement consists of a single line with the keyword FOUNDERALLELES, followed by exactly NALLELES allele names. Unknowns are allowed, and the same allele name may occur several times. This section should be present at all loci or at none of the loci, and is allowed only if NALLELES>0. If present, this information allows to display the most probable alleles present at the locus in each individual. Note that this is only possible if the data file also contains IBD probabilities (see IBD probabilities section, below). The Founder allele names must be given in the same order as the IBD probabilities.

For every Locus a corresponding Observed alleles section must be present in the data file (see below)

IBD positions statement

After the last Locus section in the linkage group an IBD positions statement may be present. This consists of a single line starting with the keyword IBDPOSITIONS, followed by one or more positions in ascending order. This statement gives all positions for which IBD probabilities are listed (see below).

Table 2. The first Linkage group section in Example.dat

```
LINKAGEGROUP A

MAP
SSR1                0
SSR2                20
SSR3                29

LOCUS      SSR1

ALLELENAMES 180  182  188  194  202 ; these are the only allowed allele names
                                   ; for this locus, each may occur only
                                   ; once in this list

FOUNDERALLELES 180 188 180 194 182 202 180 194 ; these are the allele names
                                                ; corresponding to the 8 founder alleles
                                                ; for this locus
                                                ; note that some alleles (180 and 194)
                                                ; occur more than once

LOCUS      SSR2

ALLELENAMES 143  149  177

FOUNDERALLELES 149 149 149 143 149 177 143 143

LOCUS      SSR3

ALLELENAMES 219  221  225  235  237

FOUNDERALLELES 235 221 225 219 221 237 219 219

IBDPOSITIONS 0 20 29 ; the IBD positions correspond here to the locus positions
                   ; but this is not necessary
```

After the end of the last LINKAGEGROUP section the individual-specific data are listed. First for every locus of every linkage group an Observed alleles section appears (in linkage group – map order). After the last Observed alleles section the IBD probabilities sections are listed (if present), again in linkage group – map position order.

Observed alleles sections

For each Locus mentioned in the Linkage group sections, an Observed alleles section must be present. Each Observed alleles section starts with a line with the keyword ALLELES followed by the name of the locus (this must be identical to the name given in the MAP and LOCUS sections above, upper/lower case sensitive). The next lines specify each for one individual the observed alleles for this locus. Each line starts with the name of the individual (which must be the same as that in the PEDIGREE section, upper/lower case relevant), followed by the alleles. Two formats are accepted: Either just the allele names are given (2 per individual in the diploid case), or each allele name is followed by a color code (0..250; 0..9 have predefined colors assigned) which can be used to display the alleles in different colors (e.g. to highlight incorrect scores). In the first format all alleles are assigned the color code 0. Allele names may be replaced with an UNKNOWN string for missing data, or with the NULLHOMOZ string if appropriate.

Individuals may be listed in any order, and individuals not listed are assigned missing allele scores.

Table 3. The first Observed alleles section in Example.dat

```
ALLELES SSR1 ; LG A
GoldenD      180 0 188 0
IngridM      180 0 194 0
Jonathan     182 0 202 0
Septer       202 0 188 0
Cox          180 0 194 0
Elise        202 0 180 0
Elstar       180 2 $ 2 ; either homozygous or null allele;
                ; highlighted with color code 2
81015-045    180 0 202 0
```

Note that we list the linkage group as a comment next to the locus name. This is not necessary but helps when making corrections. Also note that here we use the format that includes a color code for each allele score; most color codes are 0 (the default) but for individual Elstar color code 2 is used to highlight the ambiguity between homozygosity or a null allele.

IBD probabilities section

For each IBD position mentioned in the IBDPOSITIONS statements in the linkage group sections, an IBD probabilities section must be present. Each IBD probabilities section starts with a line with the keyword IBDPOSITION and the position (in cM). The next lines specify the calculated probabilities for each founder allele for this locus, for each individual. The individuals may be listed in any order, but all individuals from the pedigree must be present. Each line start with the name of the individual (which must be the same as that in the PEDIGREE section, upper/lower case relevant). Next follow NALLELES probabilities (0..1, sum=1) for each founder allele in the first homologue, then the NALLELES values for the second homologue (and so on, in case of higher ploidy values). Missing data are not allowed in the IBD probabilities sections.

Table 4. The first IBD probabilities section in Example.dat

```
IBDPOSITION 0 ; LG A
;
;founders:      maternal IBD probabilities      paternal IBD probabilities
;allele number: GoldenD IngridM Jonathan Cox      GoldenD IngridM Jonathan Cox
1 2 3 4 5 6 7 8      1 2 3 4 5 6 7 8

GoldenD      1.0 0 0 0 0 0 0 0      0 1.0 0 0 0 0 0 0
IngridM      0 0 1.0 0 0 0 0 0      0 0 0 1.0 0 0 0 0
Jonathan     0 0 0 0 1.0 0 0 0      0 0 0 0 0 1.0 0 0
Septer       0 0 0 0 0 1.0 0 0      0 1.0 0 0 0 0 0 0
Cox          0 0 0 0 0 0 1.0 0      0 0 0 0 0 0 0 1.0
Elise        0 0 0 0 0 1.0 0 0      0 0 0 0 0 0 1.0 0
Elstar       1.0 0 0 0 0 0 0 0      0 0 1.0 0 0 0 0 0
81015-045    0.22 0 0.78 0 0 0 0 0      0 0 0 0 0 1.0 0 0
```

Note that again we list the linkage group name as a comment next to the IBD position. Also note that the headers of the IBD tables are not necessary (they are comments) but very helpful to interpret these tables.

The interpretation of these tables is as follows. There are four founders in this small pedigree (GoldenD, IngridM, Jonathan and Cox). Each founder contributes two alleles at each locus (two haplotypes per linkage group), i.e. 8 founder alleles are segregating in the pedigree. Each individual (listed at the left margin) inherits two gametes, one from its female and one from its male parent. These are represented by the left (maternal) and right (paternal) half of the table.

For founders such as GoldenD by definition the probability that they have their own founder alleles is 1.0. Therefore in the first line there is a 1.0 probability that the maternal allele is allele 1 and the paternal allele is allele 2. For non-founders such as 81015-045 (last line) there can be uncertainties about which founder alleles they have inherited. For example, on the last line the probability that the maternally inherited allele is founder allele 1 is 0.22 and for founder allele 2 it is 0.78. In this small pedigree all other probabilities are either 0 or 1.

Limited data sets

A data file can contain less information than listed above. It is possible to have only the pedigree itself, with or without phenotypic data. In that case no linkage group sections are required at all. It is also possible to list observed alleles but not give IBD probabilities. In that case, the NALLELES header line is omitted (or NALLELES = 0), the linkage group sections may not contain FOUNDERALLELES and IBDPOSITIONS statements and no IBD probabilities sections occur. Also, it is possible to list only the IBD probabilities and no observed alleles. In that case, Linkage group sections contain no MAP and LOCUS sections but only an IBDPOSITIONS statement, and there are no Observed alleles sections.

Export

Both charts and lists of individuals can be exported, either by copying to the clipboard or by saving to a file. This is done by selecting Export from the main menu. The list of individuals is exported in the current sort order.

Further it is possible to export all the data (excluding IBD probabilities) to a data file that can be read by FlexQTL. This is discussed in the next chapter. This file is also easy to import into Excel, allowing for data modification.

FlexQTL

FlexQTL (www.flexqtl.nl) is a computer package developed by Marco Bink of the Biometris group of Wageningen University and Research Center. FlexQTL is a statistical package developed to perform QTL mapping and Identity-by Descent (IBD) calculations in pedigreed populations. Pedimap and FlexQTL are designed to work together. Data can be exported from Pedimap in a format that is accepted by FlexQTL. Using Export on the main menu, either the full population can be exported, or only the individuals of the current subpopulation; in that case, parents of individuals not in the subpopulation are listed as unknown. Data that would interfere with reading by FlexQTL are exported on lines starting with a semicolon; these lines are ignored by FlexQTL. In order to provide full compatibility with FlexQTL, which does not accept "semi-founders" (i.e. individuals with only one known parent), Pedimap will add extra ancestors from the pedigree as required until no semi-founders remain. This may be unsuccessful if the full population includes semi-founders (an error message appears and a list of exported semi-founders is printed at the bottom of the export file in that case). Further, FlexQTL requires at least two loci per linkage group; therefore for linkage groups with only one locus an extra "dummy"-locus is exported with a map position at 50 cM from the true locus and with only missing marker scores.

Besides the data in the export file, FlexQTL also requires a linkage map. This map is printed in the export file on lines starting with a semicolon; these lines must be copied and saved as a separate text file, with the semicolons removed. Further, in order to let FlexQTL report the trait names these can be copied from the column captions in the export file and added to the flexqtl.par file using keywords namesN and namesT (see FlexQTL documentation).

Conversely, FlexQTL can export data in a Pedimap data file format. This data file is always named "Pedimap.ped". FlexQTL can check the marker scores for inconsistencies (inheritance patterns that are impossible given the pedigree) and correct or delete these, and it can fill in missing marker scores based on data of relatives. When exporting data in Pedimap format, FlexQTL adds various color codes to the marker scores to indicate how it has handled them.

Table 5 lists the color codes currently (June 2010) used by FlexQTL to signal various issues; however this is subject to change so consult the FlexQTL documentation for an up-to-date list of codes. Also, keep in mind that the solutions found by FlexQTL are not necessarily correct; for instance a signaled error might actually be caused by a wrong marker score in a different, closely related individual. Generally it is a good idea to check the data when FlexQTL reports any inconsistencies instead of immediately accepting the corrections made by FlexQTL.

Table 5. Color codes for marker scores used by FlexQTL

FlexQTL-code	Explanation
0	(OC) observed & consistent
1	(MA) missing & augmented
2	(OEA) observed & error & augmented
3	(pnuA) putative null-allele & augmented
4	(pnuM) putative null-allele & missing
5	(IM) incomplete & missing
6	(OpE) observed & putative error
7	(OE) observed & error
8	(pnuE) putative null-allele & error
9	(ME) missing & error

Further FlexQTL can calculate IBD scores and export them in Pedimap format. Data files containing IBD scores for sizable populations can be large (several hundred MB) and initially take a long time to parse. However Pedimap stores most of the information in binary format in a file with the same name as the data file and extension “.pmb”, which reduces the loading time in subsequent sessions considerably.

Add semifounder parents

While Pedimap has no problems with “semifounders” (individuals that have only one known parent), these are not accepted by FlexQTL. Therefore Pedimap has an option to add the missing parents of all semifounders. The new individuals (all of them are founders with only one offspring) get a name derived from the semifounder whose parent they are, with prefix sfM_ for a mother, sfF_ for a father and sfP_ for a parent if the female and male parents are not specified in the pedigree. The phenotypic trait values and all marker scores for the new individuals are unknown (but see next section). **Warning:** if IBD probabilities were present in the original file these will all be deleted from the new file, because missing IBD data are not accepted.



Figure 20. The Add semifounder parents dialog.

This tool can be reached by selecting Tools – Add semifounder parents from the main menu. This opens a dialog (Figure 20) where the name of the new data file can be entered (a suggestion is given) and a choice is presented to immediately open the new data file in the current project.

Impute missing founder alleles

The new founder individuals added by the “Add semifounder parents” tool (see previous section) are created with all marker information missing. However it is often possible to infer which alleles can be present in these founders, based on the marker information of the offspring itself, its other parent and the ancestors of that other parent. If for a certain marker locus only one allele can be inherited from the founder it is useful to assign a heterozygous genotype, where one allele is the known allele and the other is an allele that does not occur in the population; this will improve subsequent IBD estimates.

Pedimap can fill in the missing marker information for these individuals (actually for all founders with only one offspring; this includes all founders added by the “Add semifounder parents” tool). From the main menu select Tools – Impute missing founder alleles. This opens a dialog (Figure 21) where name of the new data file can be specified and a choice if given to immediately open the new data file in the current project. Further a name can be specified for the non-transmitted alleles in the founders; as mentioned above this is preferably an allele name that does not occur at any locus in the population. Also the color code for highlighting the imputed alleles can be specified. The default setting is 1, corresponding to the code used by FlexQTL for “Missing and augmented” (Tables 5).

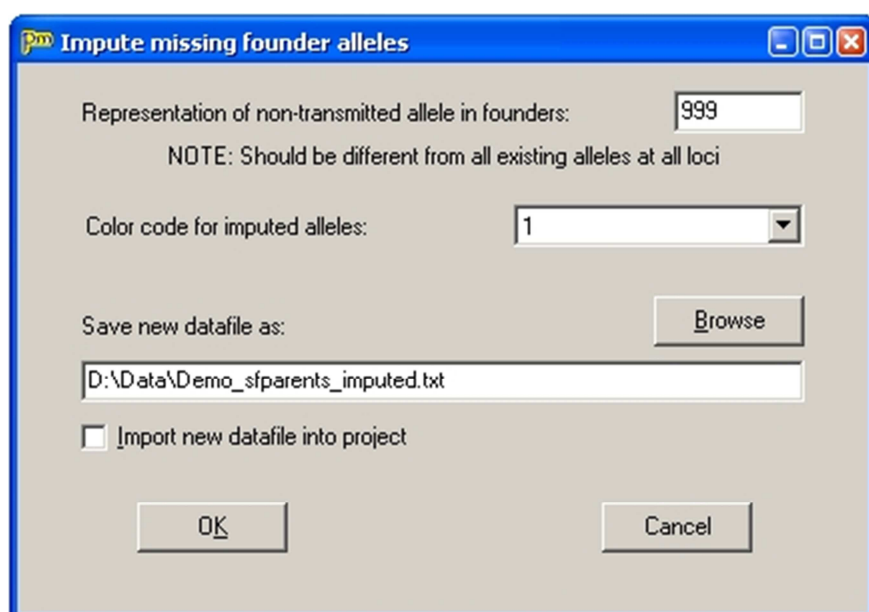


Figure 21. The Impute missing founder alleles dialog.

Besides a new data file Pedimap also generates a “Pedimap imputation.log” file. This file mentions all founders investigated and lists any “impossible” genotype scores found in their offspring, given the genotypes of its ancestry through the other parent.

Drawing limitation

There is a limitation to the size of the (sub)populations for which Views (including the Overview) are drawn. This limit is by default 1000 individuals. Images with such large numbers of individuals are not very helpful. Zooming out to view the whole image makes the data unreadable and when zooming in the image becomes so large that it is hard to get the whole pattern. Also, drawing images with so much information can take a lot of time and may freeze the program or the whole machine because system resources get exhausted. When working with larger pedigrees, initial selections have to be made on the Individuals page. As usual the automatic selection of individuals based on pedigree relations can be used. For subpopulations with less than 1000 individuals an Overview will be generated and further Views can be created.

The limit can be changed using the command-line parameter /DL=<number> (DL stands for “Drawing Limit”). For example, a desktop shortcut to Pedimap could specify as Target: “C:\Program Files\Pedimap.exe /DL=1500” to allow drawing up to 1500 individuals.

Tips & tricks

Copy options between views

Sometimes we want to create a new View similar to an existing one. Then it is easier to copy all the options and only change a few ones, than to start all over again.

In such a case first create a new View. Then select the view from which the options must be copied; this view can belong to the same or in a different subpopulation. Right-click on its tab and select “Copy View options”. Next, select the new view, right-click its tab and select “Paste View options”. Now the two views are identical. Finally some changes can be made to the new View's options using the View Options dialog.

Copy views between subpopulations

Suppose we have a series of views for a subpopulation A, and we want to copy that whole series to another subpopulation B. First, in the Population panel, right-click on subpopulation A and select “Set as Views source”. Next, right-click on subpopulation B and select “Add Views from source”. You will see two all views from subpopulation A, with the same names and options, added to those already present in subpopulation B.

Combine subpopulations

Suppose we want to combine all the individuals in two subpopulations A and B (more than two subpopulations is also possible) into one new subpopulation C.

In the Population panel, select one of the two subpopulations, then add the other subpopulation to the selection (Ctrl – left mouse button). On one of the selected subpopulations, right-click and select “Combine subpops”.

Now a new subpopulation “Combination-1” is created containing all individuals that occur in subpopulations A or B or in both. In the Info dialog (right-click on the new subpopulation and select Info) the selection history just lists all individuals, as no automatic selection of relatives was used to create this node, but a Note is added mentioning the source subpopulations.

Names containing spaces, long names

In Pedimap, names of individuals, linkage groups, loci and alleles, and the levels of phenotypic traits may contain spaces. If they do, they have to be enclosed in “double quotes” at every occurrence in the data file. Note that FlexQTL and most other software does not allow names containing spaces, so avoid using them except in projects that do not have to exchange data with other software.

As all individuals are drawn to the same size, the length of the longest name of any individual determines the size of all. So it is best to use names of more or less equal length and to avoid very long names. The same is true for the names of alleles.