BIOMÉRIEUX

BIONUMERICS®

version 8 - PLUGINS

SNP calling plugin

Contents

NOTES

SUPPORT BY APPLIED MATHS, A BIOMÉRIEUX COMPANY

While the best efforts have been made in preparing this manuscript, no liability is assumed by the authors with respect to the use of the information provided.

Applied Maths, a bioMérieux company, will provide support to research laboratories in developing new and highly specialized applications, as well as to diagnostic laboratories where speed, efficiency and continuity are of primary importance. Our software thanks its current status for a part to the response of many customers worldwide. Please contact us if you have any problems or questions concerning the use of BIONUMERICS®, or suggestions for improvement, refinement or extension of the software to your specific applications:

Applied Maths NV
Keistraat 120
9830 Sint-Martens-Latem
Belgium
PHONE: +32 9 2222 100
FAX: +32 9 2222 102
E-MAIL: BE-DAU-INFO@biomerieux.com
URL: https://www.bionumerics.com

Applied Maths, Inc.
11940 Jollyville Road, Suite 115N
Austin, Texas 78759
U.S.A.
PHONE: +1 512-482-9700
FAX: +1 512-482-9708
E-MAIL: US-DAU-INFO@biomerieux.com

LIMITATIONS ON USE

The BIONUMERICS® software, its plugin tools and their accompanying guides are subject to the terms and conditions outlined in the License Agreement. The support, entitlement to upgrades and the right to use the software automatically terminate if the user fails to comply with any of the statements of the License Agreement. No part of this guide may be reproduced by any means without prior written permission of the authors.

BIONUMERICS® uses following third-party software tools and libraries:

- Python 3.8 release from the Python Software Foundation, https://www.python.org/
- Xerces library for XML input and output from the Apache Software Foundation, https://xerces.apache.org/
- NCBI toolkit version 2.11.0, https://www.ncbi.nlm.nih.gov/BLAST/
- SRA Toolkit, https://ncbi.github.io/sra-tools/
- Boost c++ libraries, https://www.boost.org/
- Samtools for interacting with SAM / BAM files, https://www.htslib.org/download/
- 7-Zip (7za.exe), https://www.7-zip.org/
- Zlib library, https://zlib.net/
- Pigz for parallel gzip compression, https://zlib.net/pigz/
- Cairo 2D graphics library version 1.12.14, https://cairographics.org/
- Crypto++ library version 5.5.2, https://www.cryptopp.com/
- OpenSSL library, https://www.openssl.org/
- libSVM library for Support Vector Machines, https://www.csie.ntu.edu.tw/~cjlin/libsvm/
- SQLite version 3.7.17, https://www.sqlite.org/
- pymzML Python module version 2.4.7, https://github.com/pymzml/pymzML
- NumPy Python library version 1.19.1, https://www.numpy.org/
- BioPython Python library version 1.78, https://www.biopython.org/
- pyodbc Python module version 4.0.30, https://pypi.org/project/pyodbc/
- jinja2 Python library version 2.11.2, https://pypi.org/project/Jinja2/
- MarkupSafe Python library version 1.1.1, https://pypi.org/project/MarkupSafe/
- regex Python library version 2.5.91, https://pypi.org/project/regex/
- Chromium Embedded Framework, https://bitbucket.org/chromiumembedded/cef/wiki/Home
- SPAdes genome assembler version 3.15.3, https://bioinf.spbau.ru/spades *
- SKESA version 2.3.0, https://github.com/ncbi/SKESA/releases
- Unicycler version 0.5.0, https://github.com/rrwick/Unicycler/releases *
- Velvet for Windows, source code can be downloaded from https://www.bionumerics.com/download/open-source
- Bowtie2 version 2.2.5 (https://bowtie-bio.sourceforge.net/bowtie2/index.shtml)*
- SNAP version 2.0.0, https://www.microsoft.com/en-us/research/project/snap/
- RAxML version 8.2.11, https://github.com/stamatak/standard-RAxML/releases

- FastTree version 2.1.10, https://www.microbesonline.org/fasttree/
- CFSAN SNP pipeline version 2.2.0, https://github.com/CFSAN-Biostatistics/snp-pipeline *
- Prokka version 1.14.5, https://github.com/tseemann/prokka *
- sourmash version 4.1.0, https://github.com/dib-lab/sourmash **
- SeqSero2 for Windows, source code can be downloaded from https://www.bionumerics.com/download/open-source
- Fastp version 0.22.0, https://github.com/OpenGene/fastp

*: On Calculation Engine only **: See license conditions below

Sourmash license conditions:

Chapter 1

Starting and setting up BIONUMERICS

1.1 Introduction

This guide is designed as a tutorial for the *SNP calling plugin* of BIONUMERICS. This plugin facilitates data analysis for single nucleotide polymorphism (SNP) genotyping using TaqMan® probes or similar technology in BIONUMERICS.

The *SNP calling plugin* is supported in all BIONUMERICS configurations.

1.2 Startup program

Make sure the latest version of BIONUMERICS is installed (https://www.bionumerics.com/download/software). The installation manual can be downloaded from https://www.bionumerics.com/download/manuals.

When BIONUMERICS is launched from the Windows start panel or when the BIONUMERICS shortcut ( ) on your computer's desktop is double-clicked, the **Startup program** is run. This program shows the *BIONUMERICS Startup* window (see Figure 1.1).

A new BIONUMERICS database is created from the Startup program by pressing the button.

An existing database is opened in BIONUMERICS with or by simply double-clicking on a database name in the list.

1.3 Creating a new database

3.1 Press the button in the BIONUMERICS *BIONUMERICS Startup* window to enter the *New database* wizard.

3.2 Enter a name for the database, and press <***Next***>.

A new dialog box pops up, prompting for the type of database (see Figure 1.2).

3.3 Leave the default option selected and press <***Next***>.

A new dialog box pops up, prompting for the database engine (see Figure 1.3).

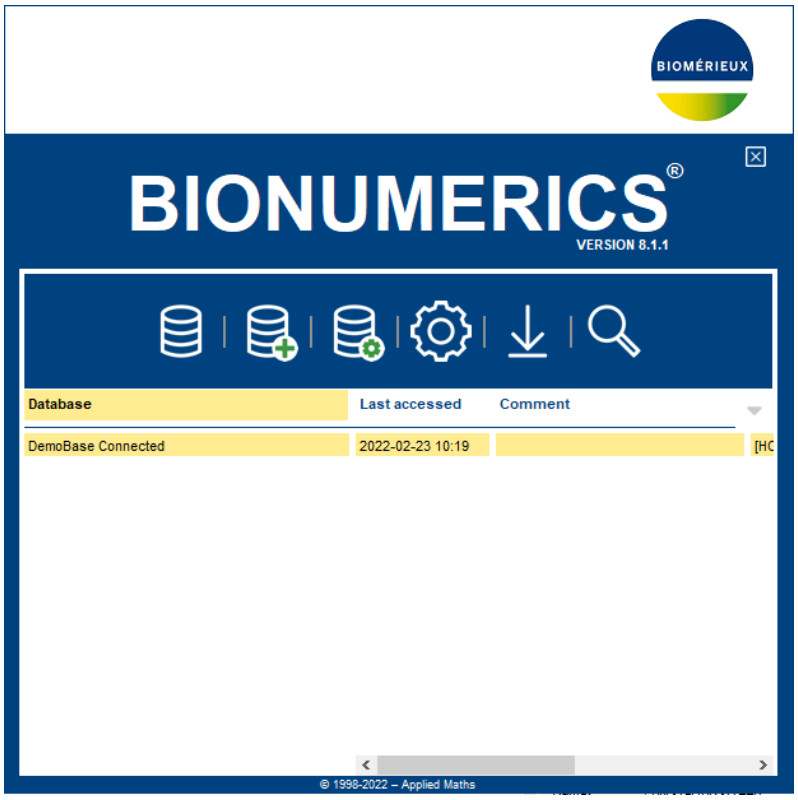

Figure 1.1: The *BIONUMERICS Startup* window.

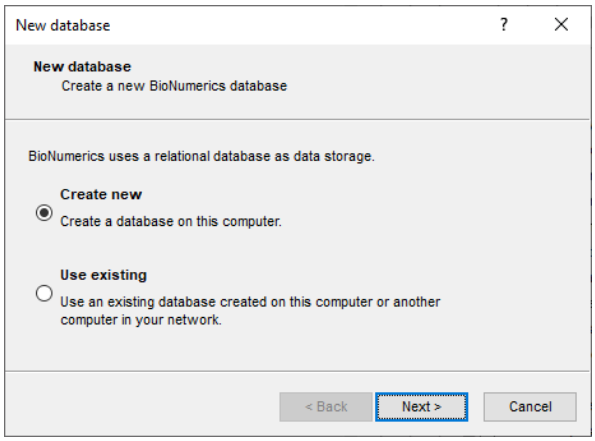

Figure 1.2: The *New database* wizard page.

3.4 Leave the default option selected and press <***Finish***> to complete the setup of the new database.

1.4 Installing the SNP calling plugin

The *Plugins and Scripts* dialog box can be called from the *Main* window by selecting ***File*** > ***Install / remove plugins...*** ( ) (see Figure 1.4).

When a particular plugin is selected from the list of plugins, a short description appears in the right panel.

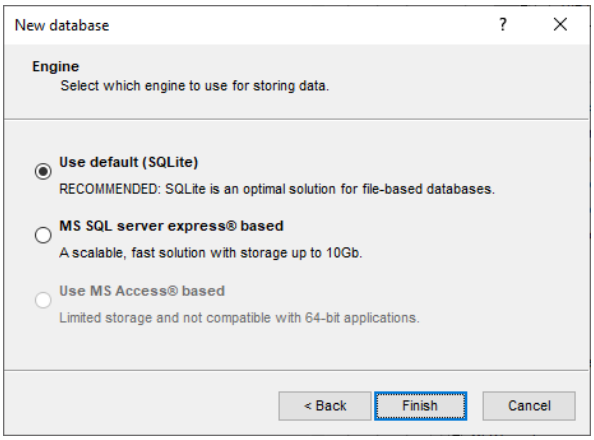

Figure 1.3: The *Engine* wizard page.

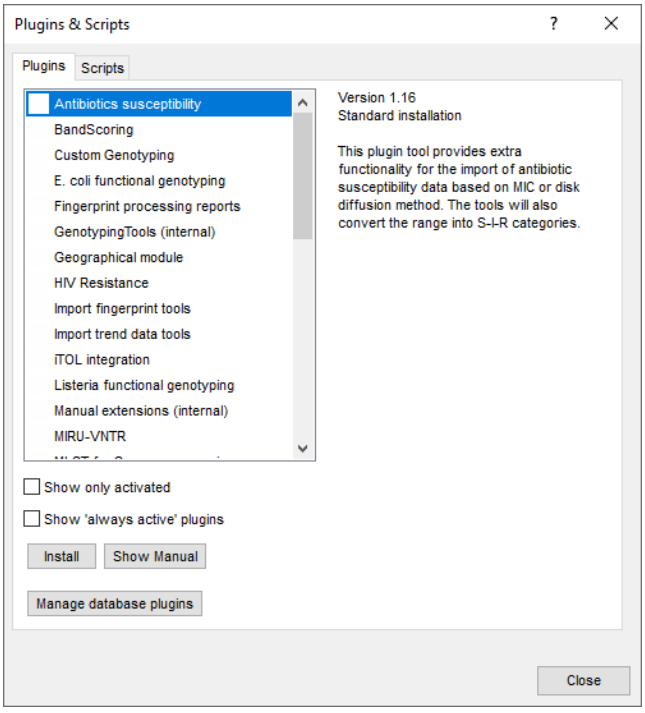

Figure 1.4: The *Plugins and Scripts* dialog box.

A selected plugin can be installed with the <***Install***> button. The software will ask for confirmation before installation. Some plugins are only supported in specific BIONUMERICS configurations. If the plugin is not supported by your BIONUMERICS configuration, it cannot be installed and an error message will be generated.

Once a plugin is installed, it is marked with a green V-sign. It can be removed again with the <***Uninstall***> button.

If the selected plugin is documented, pressing <***Show Manual***> will open its manual in the *Help* window.

4.1 Select the *SNP calling plugin* from the list and press the <***Install***> button.

The installation of the *SNP calling plugin* requires administrator privileges.

4.2 The program will ask to confirm the installation of the plugin. Press <***Yes***> to confirm the installation.

The *License string* dialog box pops up (see Figure 1.5). The *SNP calling plugin* can only be installed and activated with a valid *license number*, which needs to be purchased from Applied Maths. Make sure you have the License Number ready.

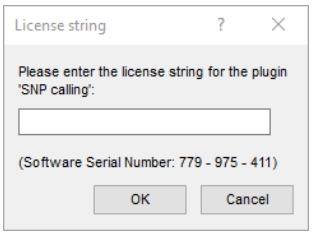

Figure 1.5: The *License string* dialog box.

The *License string* dialog box prompts for a license string that is compatible with the ***Software Serial Number*** listed in the dialog box.

4.3 Enter the six digits license string and press <***OK***>.

Before the plugin can be installed, the settings listed in the *SNP calling install* dialog box (see Figure 1.6) need to be specified. Please note that these settings cannot be changed afterwards!

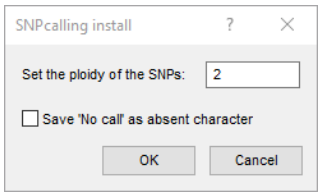

Figure 1.6: The *SNP calling install* dialog box.

The *ploidy* of the SNPs in the current database should be set. Ploidy is defined as the number of complete sets of chromosomes in a biological cell. For diploid organisms such as humans, this should be set to “2” (unless the SNPs are located on sex-determining chromosomes for males), for triploid organisms (e.g. most apple cultivars), this should be “3”, etc.. From the ploidy, the theoretically possible call states can be derived: ”XX”, ”XY” and ”YY” for diploids, ”XXX”, ”XXY”, ”XYY” and ”YYY” for triploids, etc..

The option ***Save ”No call” as absent character*** is important when comparing samples based on their SNP calls (see 1.5). With this option checked, a SNP experiment that was performed but could not be called unambiguously, will be saved as an absent character value in the **SNP_call** experiment. In BIONUMERICS, absent character values are never penalized when calculating pairwise similarity values. Therefore, a ”No call” will match any possible genotype in this case when calculating a cluster analysis, performing an identification, etc.. With the ***Save ”No call” as absent character*** option unchecked (default), a ”No call” will be saved as zero, and will only match with other ”No calls”.

4.4 For diploid organisms, leave the default value of “2” and press <***OK***> to proceed with the installation.

4.5 A message box pops up, confirming the installation of the plugin. Press <***OK***>.

4.6 Press <***Close***> to close the *Plugins and Scripts* dialog box and to continue to the *Main* window.

4.7 Close and reopen the database to activate the features of the *SNP calling plugin*.

The software warns that some additional tables, which are needed for the SNP calling functionality, will be created in the database. This will require administrator privileges on the connected relational database.

4.8 Press <***Yes***> update the relational database table structure.

The *SNP calling plugin* installs an additional *SNP files panel* in the *Main* window and a number of extra buttons and menu items (see Figure 1.7). In addition, a character experiment type, **SNP_call** is created after installation of the *SNP calling plugin*. See 4.1 for more information about the data contained in this character type.

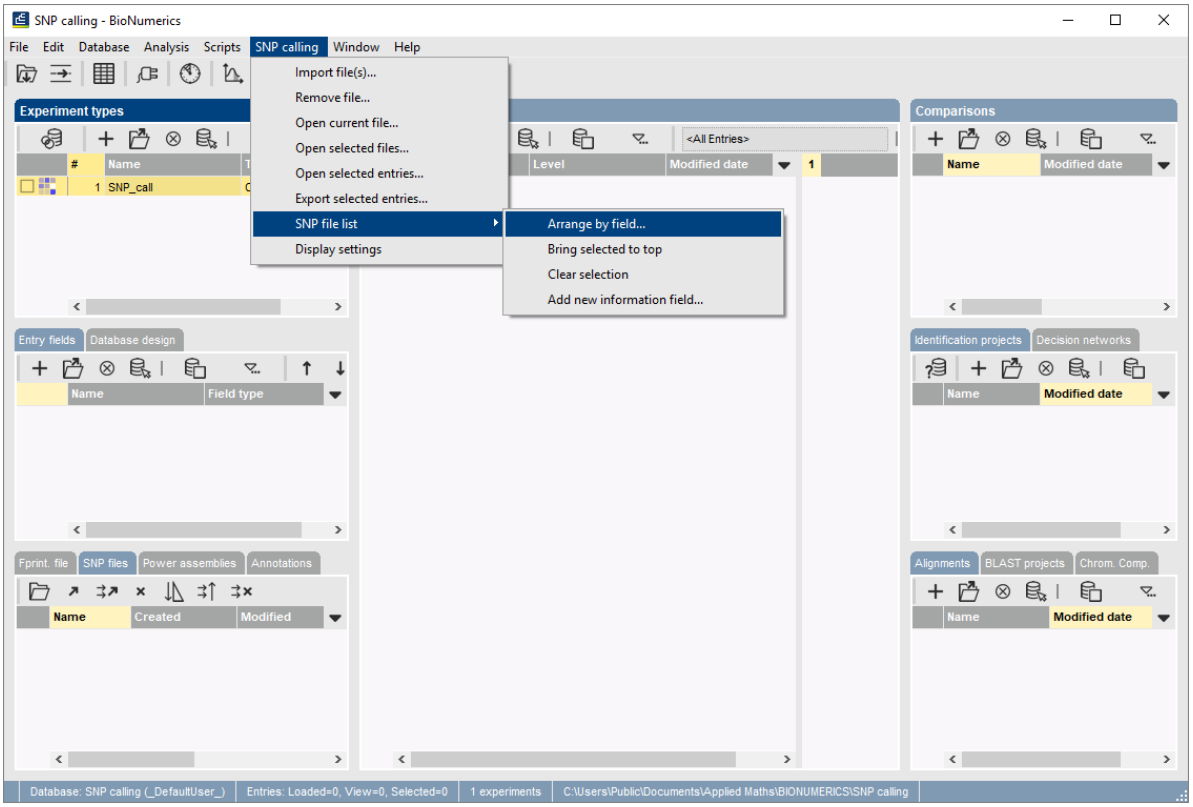

Figure 1.7: The *Main* window after installation of the plugin.

1.5 SNP display settings

The colors in which the different SNP calls are displayed, can be set by the user.

5.1 Select ***SNP calling*** > ***Display settings*** to call the *Display settings* dialog box.

Possible calls are ***No call*** (when the genotype could not be unambiguously determined), ***NTC*** (for No Template Control) and all possible genotypes as predicted by the ploidy. Figure 1.8 displays this dialog box for diploid organisms, hence ***XX***, ***XY*** and ***YY*** are the possible genotypes. Please note that more genotypes will be listed when a higher ploidy was set during installation (see 1.4). In the *Display settings* dialog box, the colors for the different calls can be set. Pressing the <***Change***> button for any call pops up the *Color* dialog box.

Any desired color can be picked from this dialog using (a combination of) any of the methods below:

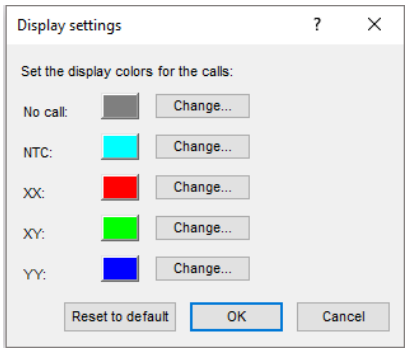

Figure 1.8: The *Display settings* dialog box.

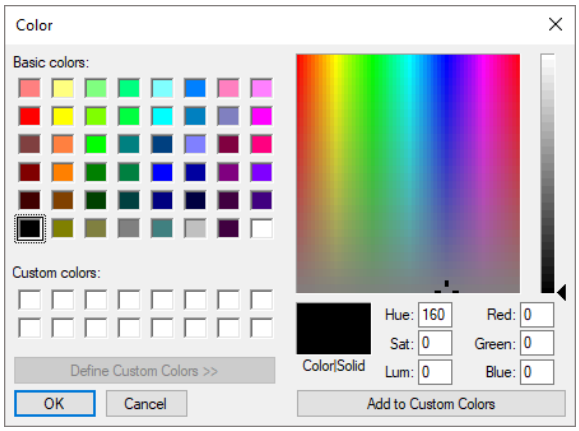

Figure 1.9: The *Color* dialog box.

- By clicking any of the ***Basic colors*** in the upper left-hand side of the dialog box.
- By clicking any of the ***Custom colors*** (if defined) in the lower left-hand side of the dialog box.
- By entering ***Red***, ***Green*** and ***Blue*** values in the corresponding text boxes.
- By entering ***Hue***, Saturation (***Sat***) and Luminosity (***Lum***) values directly in the corresponding text boxes.
- By picking a point in the hue-saturation plot on the right-hand side of the dialog box and selecting a luminosity value using the slider on the far right.

When a color is selected, it can be added to the ***Custom colors*** by clicking on one of the custom color cells and pressing the <***Add to Custom Colors***> button.

Press <***OK***> to use the selected color in the *Color* dialog box. Conversely, press <***Cancel***> to keep the original color.

Chapter 2

Importing SNP data

2.1 SNP genotyping: an introduction

In SNP genotyping using TaqMan® technology, the 5' nuclease activity of Taq polymerase is used to generate a fluorescent signal during PCR. One pair of TaqMan® probes and one pair of PCR primers is needed per interrogated SNP. The assay uses two TaqMan® probes that differ in sequence only at the SNP site, with one probe complementary to the wild-type allele and the other to the variant allele. A 5' reporter dye and a 3' quencher dye are covalently linked to the wild-type or variant allele probes. When the probes are intact, fluorescence is quenched because of the physical proximity of the reporter and quencher dyes. This phenomenon is known as Förster resonance energy transfer (FRET). In the PCR annealing step, the TaqMan® probes hybridize to the targeted SNP site. During PCR extension, the TaqMan probe is degraded by the 5' nuclease activity of the Taq polymerase, leading to an increase in the characteristic fluorescence of the reporter dye. Degradation only occurs for the perfectly hybridized probe, mismatched probes are displaced. At the end of the PCR reaction, the fluorescent signal for the two reporter dyes is measured. The ratio of the signals will be indicative for the genotype of the sample (Figure 2.1).

In most assays, the fluorescent signals of the two reporter dyes are normalized by a the signal of a third dye (e.g. ROX), of which the intensity is proportional to the template DNA concentration and the extent of the PCR reaction. Furthermore, as for nearly all PCR-based assays, it is common to include one or a few No Template Control (NTC) reactions in each run or plate. Typically, the reporter dye signals after PCR are visualized in a plot.

A number of other commercially available genotyping systems using the FRET principle (e.g. Invader®, Molecular Beacons®, Scorpion® and other probe technologies) can be analyzed and visualized in the same way as TaqMan® probes using the *SNP calling plugin*.

Before discussing the *SNP calling plugin* functionality in detail, it might be useful to explain the terminology used in the plugin and its manual:

- A **SNP** (or single nucleotide polymorphism in full) is a nucleotide position on a chromosome for which allelic variants exist. It is considered a genetic marker if it has been linked to a certain phenotypic trait.
- With **SNP file** we refer to a result file, generated by the software of a SNP genotyping system, that contains results for a batch of reactions, performed e.g. in the same plate or run. The results in this file are fluorescence measurements and optionally also calls.
- **SNP experiment** is used to denote a sample/SNP combination in a SNP file, i.e. a sample that was genotyped for a certain SNP.

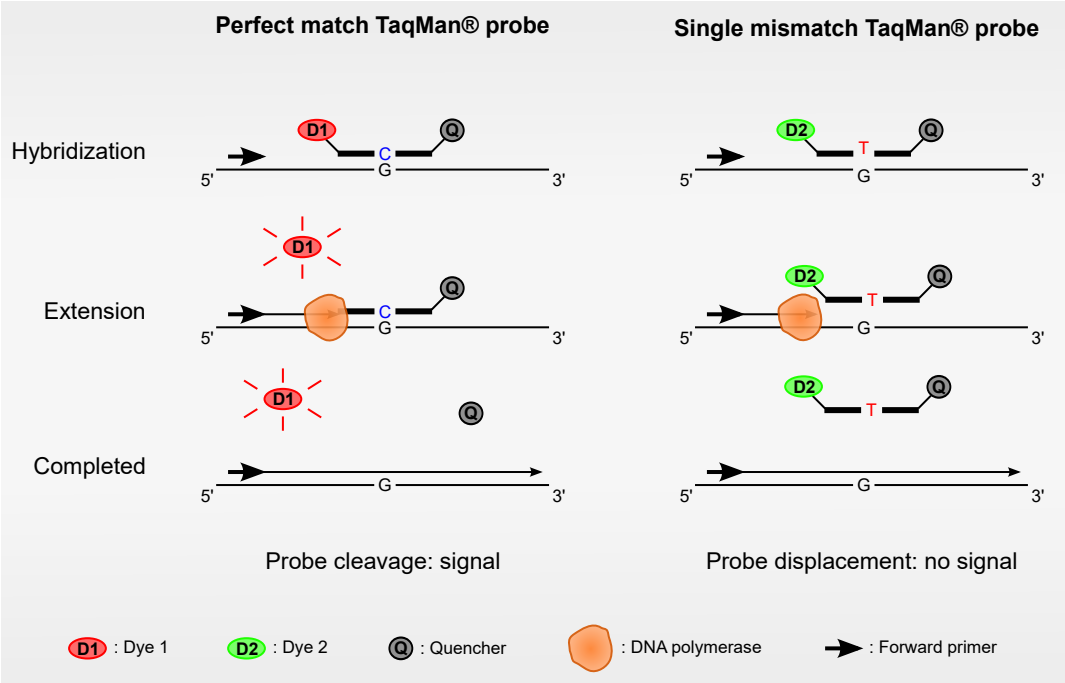

Figure 2.1: Principle of SNP genotyping using TaqMan® technology.

- A **SNP call** is the assignment of a SNP experiment to a certain genotype.

2.2 SNP calling work flow in BIONUMERICS

The *SNP calling plugin* is very flexible and accommodates for different work flows for SNP genotyping data analysis:

1. Perform an auto calling in other software and import the calls and their corresponding confidence values in BIONUMERICS.
2. Perform an auto calling in BIONUMERICS during import.
3. Import data as ”No call” and perform an auto calling in the *SNP calling* window.

The first option is very convenient if a vast amount of ”historic” SNP data is available, i.e. SNP data that was called in other software and subjected to a visual check. In this case, time can be saved by importing the existing data including their calls. Option 2 makes use of the advanced auto call algorithm in the *SNP calling plugin* (see 2.12 for more information) and would probably be the preferred work flow with newly generated data. While option 3 typically can be useful in an implementation phase to evaluate various auto call settings, this would certainly not be the preferred work flow in a routine analysis. In any of the above scenarios, it will be possible for a user to manually override an automatic call (see 3.3). The three options will be illustrated in the tutorials that deal with the different data formats (see further).

2.3 Compatible data formats

Currently, SNP data in seven different file formats can be imported using the *SNP calling plugin*:

- **Applied Biosystems 7900HT**: *.TXT files as generated by Applied Biosystems' 7900HT Fast Real-Time PCR System, containing processed 384-well plate reads.
- **Applied Biosystems ViiA 7**: *.TXT files as generated by Applied Biosystems' ViiA 7 Real-Time PCR System, containing processed 384-well plate reads.
- **BMG LabTech**: *.DAT files generated by BMG LABTECH CLARIOstar 384-well micro plate readers, containing raw fluorescence values in a micro plate layout.
- **BMG LabTech (Dual Em.)**: *.DAT files generated by PHERAstar BMG LABTECH 384-well micro plate readers, containing raw fluorescence values in a micro plate layout.
- **Douglas Scientific**: *.TXT files generated by the Douglas Scientific Array Tape system, in 384-well format.
- **Fluidigm**: *.CSV files as generated by the Fluidigm Dynamic Array, containing up to 9,216 data points (96 samples tested against 96 SNPs).
- **KBiosciences**: *.CSV genotyping reports of KBiosciences (LGC).
- **Tecan Safire/Infinite**: *.TXT files generated by Tecan Safire or Tecan Infinite micro plate readers, in 96-well and 384-well micro plate formats.

However, other data formats that you may have could be used for this type of analysis and the *SNP calling plugin* could be adapted to import these files as well. Please contact Applied Maths with any inquiries you may have regarding data formats.

For most supported data formats, one or a few example files are available from our website. They can be downloaded from the download page on the BIONUMERICS website (https://www.bionumerics.com/download/sample-data, click on "SNP genotyping data").

2.4 Importing Applied Biosystems 7900HT files

2.4.1 Example data

Five example files (AB_Dr35_Xr40-100209, AB_Lc62_01-091019, AB_Lc62_02-091214, AB_Lc62_03-100215, and AB_Tt60-100208) will be used to illustrate the import of Applied Biosystems 7900HT files. The files are included with the example data, which can be downloaded from the Applied Maths download page (https://www.bionumerics.com/download/sample-data, click on "SNP genotyping data").

As mentioned earlier (see 2.2), two different work flows will be illustrated here: Firstly importing Applied Biosystems SNP files with an auto calling in BIONUMERICS based on the fluorescence signals provided in these files. In the second tutorial, the same files will be imported, this time taking over the calls made by the Applied Biosystems software.

2.4.2 Import of SNP files with auto calling in BIONUMERICS

4.1 In the *Main* window, select ***SNP calling*** > ***Import file(s)*** or press the 🗁 button in the *SNP files panel*.

This action starts the import wizard (see Figure 2.2).

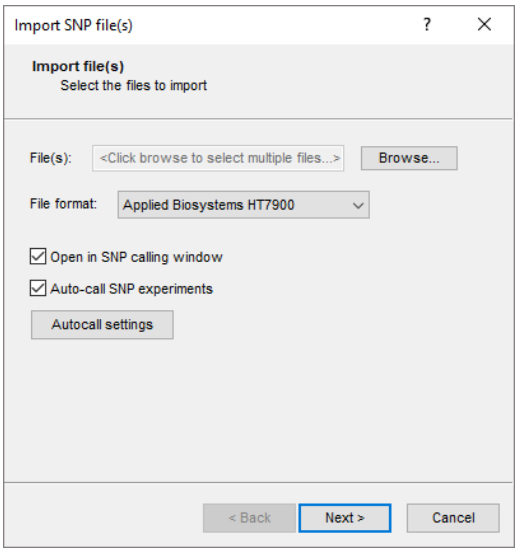

Figure 2.2: The *Import files page*.

Pressing the <***Browse***> button allows you to select the SNP files that you want to import and which are located on your computer, external drive or on a network location. Note that you can import multiple SNP files at once. All selected SNP files should have the same format.

Different SNP data file formats are supported (see 2.3) and can be selected from the ***File format*** drop-down list. The file format that corresponds to the selected files should be picked from the list.

When ***Open in SNP calling window*** is checked, imported SNP experiments will be displayed in the *SNP calling* window (see 3.1) after the import completes.

Check ***Auto-call SNP experiments*** to let BIONUMERICS perform an automatic SNP calling during import, according to the Auto call settings. If ***Auto-call SNP experiments*** is unchecked, the calling performed by other software (if present) will be taken over. If the SNP file does not contain call information, all SNP experiments will be imported as ”no call”.

Pressing <***Auto call settings***> will open the *Autocall settings* dialog box (see 2.12 for more information).

At least one SNP file should be selected before one can proceed to the next page of the wizard. Therefore, we will use the Applied Biosystems 7900HT files to illustrate the different steps of this wizard.

4.2 Press the <***Browse***> button, select the five example SNP files with the ”AB_” prefix and press <***Open***>.

4.3 Make sure that ”Applied Biosystems HT7900” is selected as ***File format***. Leaving ***Open in SNP calling window*** and ***Auto-call SNP experiments*** checked, press the <***Next***> button.

This displays the *Parse file IDs page* of the import wizard (see Figure 2.3).

Using a ***File ID parsing string***, a File ID can be parsed from the SNP file name. A preview of the parse operation can be obtained by pressing the <***Preview***> button.

4.4 For the example Applied Biosystems 7900HT files, enter “[FILEID]-*” as ***File ID parsing string***, to parse the redundant date suffix from the file names.

4.5 Press <***Next***> to proceed.

This displays the *Database entries page* of the import wizard (see Figure 2.4).

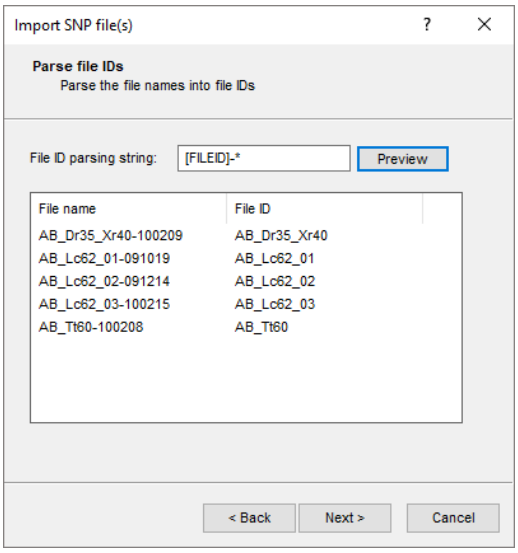

Figure 2.3: The *Parse file IDs page*.

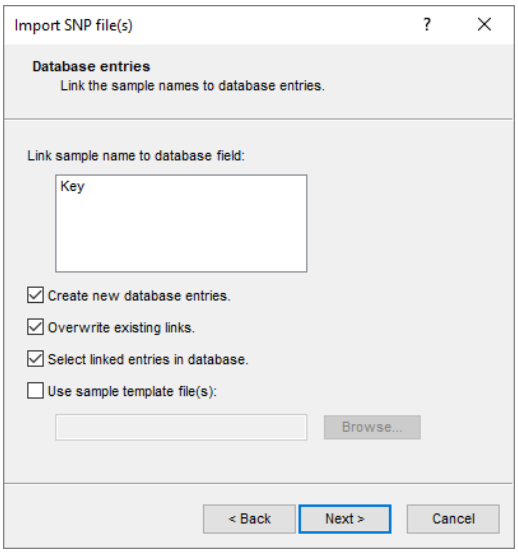

Figure 2.4: The *Database entries page*.

The sample names, present in the SNP file(s), can be linked to any available database information field by selecting it from the list.

If ***Create new database entries*** is checked, the plugin is allowed to create new entries in the database. This option could be disabled in case all imported SNP experiments should link to already existing database entries, e.g. when a sample sheet (see 2.13) was imported prior to the SNP files import.

With the option ***Overwrite existing links*** checked, existing links to database entries will be overwritten with the new SNP experiments.

If ***Select linked entries in database*** is checked, entries linked to SNP experiments, originating from this import batch of SNP files, will be selected in the database.

An optional sample template file can be specified when ***Use sample template file*** is checked. See 2.6 for an illustration how to use such a template file.

The file path that can be entered or browsed for when ***Use sample template file*** is checked, accepts the token [FILEID] as parsed from the sample file name in the previous step. This allows the use of multiple template files (one for each sample file) during import of a batch.

4.6 Since 'Key' currently is the only information field available, select "Key" as the information field to link the SNP experiment to.

When an information field different from 'Key' is selected as link field, an arbitrary unique key will be automatically generated.

4.7 Leave all other settings checked and press <***Next***>.

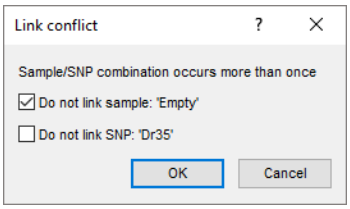

Figure 2.5: The *Link conflict* dialog box for a non-unique sample/SNP combination.

Before the import actually starts, all SNP experiments from the imported SNP files will be scanned for *link conflicts*. In case a sample name - SNP combination is not unique, the *Link conflict* dialog box pops up, from which the user can choose not to link the SNP experiment to a database entry and/or not to link a SNP to a character of the **SNP_call** character type experiment (see 4.1 for more information about this character type).

4.8 In the example files, the "Empty" samples (no template controls) should not be linked to database entries, so check only ***Do not link sample: "Empty"*** and press <***OK***>.

The selected SNP files are now imported, which may take a while, depending on the number of files selected and their file size. During import, a status bar is displayed in the bottom left corner of the *Main* window. When the import is complete, the SNP files are shown in the *SNP files panel*, entries are added to the database and a colored dot is displayed in the *Experiment presence* panel for the character type experiment **SNP_call**.

Any problems that might have occurred for sample/SNP combinations are listed in the *Autocall warnings* dialog box.

Possible error messages are:

- No NTC samples present.
- Some SNP samples fall in the NTC cluster.
- Some NTC samples fall in a SNP calling cluster.
- Too few sample points (n) to perform an auto call: With n the number of samples. Minimum five samples are required.
- n samples below confidence cutoff: With n the number of samples.
- n samples below ROX cutoff: With n the number of samples.
- Found only x cluster(s): With x a lower value than the expected number of clusters.

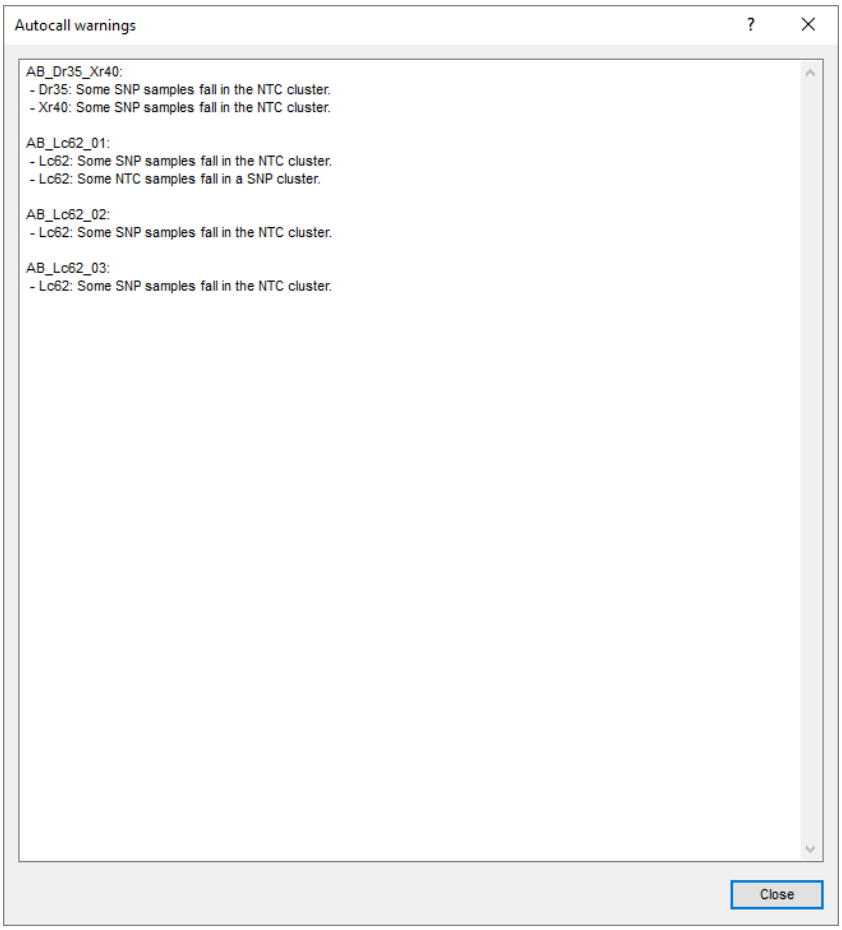

Figure 2.6: The *Autocall warnings* dialog box.

If ***Open in SNP calling window*** was checked in the first step of the import wizard and no problems were encountered during the import, the imported SNP experiments are plotted in the *SNP calling* window (see 3.1).

In case warnings were generated during the import, the *Autocall warnings* dialog box should be closed before the *SNP calling* window appears. SNPs that had an auto call warning will be selected.

2.4.3 Import of SNP files without auto calling in BIONUMERICS

Since the example Applied Biosystems 7900HT files contain call information, the calls made by the Applied Biosystems software and their corresponding confidence values can be taken over by BIONUMERICS.

4.9 In the *Main* window, select ***SNP calling*** > ***Import file(s)*** or press the button in the *SNP files panel*.

4.10 Press the <***Browse***> button, select the five example SNP files with the ”AB_” prefix and press <***Open***>.

4.11 Make sure that ”Applied Biosystems HT7900” is selected as ***File format***.

4.12 This time, uncheck ***Auto-call SNP experiments*** and press the <***Next***> button.

4.13 In the *Parse file IDs page* of the *Import SNP file(s)* wizard, enter e.g. “[FILEID]” as ***File ID parsing string***, so that unique file IDs will be generated in case you already performed an import with auto calling.

4.14 Press <***Next***> to proceed.

4.15 In the *Database entries page* of the *Import SNP file(s)* wizard, make sure ***Overwrite existing links*** is selected and press the <***Next***> button.

The *SNP call mappings dialog box* will appear (see Figure 2.7).

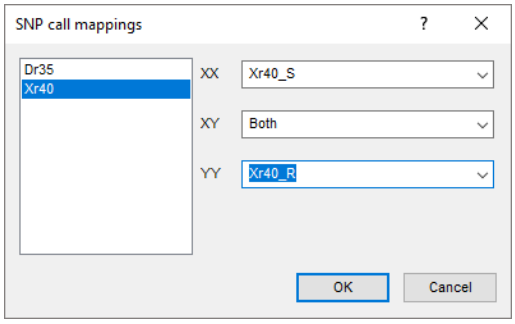

Figure 2.7: The *SNP call mappings* dialog box.

For all SNPs from the imported SNP files, a mapping should be provided between each possible genotype (***XX***, ***XY*** and ***YY*** for diploids) and a descriptive name for this genotype. The descriptive name is extracted from the SNP file content and made available from the drop-down lists. Alternatively, the name can be typed directly in the text boxes. Once the mapping is made, the *SNP call mappings* dialog box will not display again, unless a different descriptive name is found in a new SNP file. The mapping provided will be stored as a character information field in the **SNP_call** *Character type* window (see 4.1).

4.16 For SNPs **DR35** and **Xr40**, map ***XX*** to the ”Sensitive” phenotype (indicated with the ”_s” suffix), ***XY*** to ”Both” and ***YY*** to the ”Resistant” phenotype (indicated with the ”_r” suffix).

4.17 In the *Link conflict* dialog box that pops up, check only ***Do not link sample: ”Empty”*** and press <***OK***>.

4.18 For SNP **Lc62**, map ***XX*** to ”Lc62:A”, ***XY*** to ”Both” and ***YY*** to ”Lc62:G”.

4.19 For SNP **Tt60**, map ***XX*** to ”Tt60_s”, ***XY*** to ”Both” and ***YY*** to ”Tt60_r”.

The imported SNP experiments will be plotted in the *SNP calling* window (see 3.1).

2.5 Importing Applied Biosystems ViiA 7 files

Importing data generated by Applied Biosystems ViiA7 system follows exactly the same steps as for the 7900HT system from the same manufacturer (see 2.4).

2.6 Importing BMG Labtech files

The example files `BMG_Dr45`, `BMG_Tt35a`, and `BMG_Tt35b` will be used to demonstrate the import of BMG LABTECH PHERAstar files. These files are included with the example data, which can be

downloaded from the Applied Maths download page (https://www.bionumerics.com/download/sample-data, click on ”SNP genotyping data”).

6.1 In the *Main* window, select ***SNP calling*** > ***Import file(s)*** or press the button in the *SNP files panel* to call the import wizard.

6.2 Browse for the three BMG LABTECH example files and press <***Open***>.

6.3 Make sure to select ”BMG LabTech (Dual Em.)” as ***File format***, check ***Open in SNP calling window*** and ***Auto-call SNP experiments*** and press the <***Next***> button.

6.4 Enter e.g. “[FILEID]” as ***File ID parsing string*** and press <***Next***>.

6.5 Select the ’Key’ field to link the sample information to.

The BMG LABTECH file format (*.DAT files) only contains well positions and raw fluorescence data, but no sample or SNP information. Therefore, a *sample template file* is needed to provide the latter ”meta” information. BMG LABTECH files can only be imported without sample template file if all wells from the same plate (corresponding to a single SNP file) are analyzed for the same SNP. Since this is the case for the example BMG LABTECH files, we can use them to illustrate the import without sample template file.

6.6 Make sure the option ***Use sample template file*** is unchecked and press <***Next***>.

The *Link SNP file* dialog box appears (see Figure 2.8).

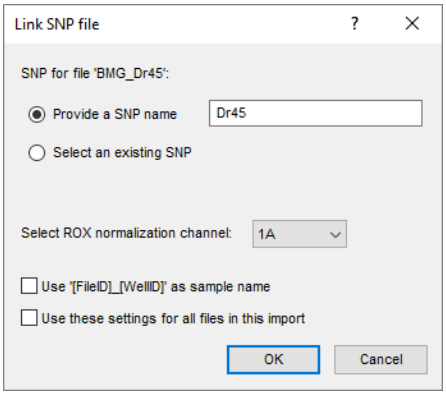

Figure 2.8: The *Link SNP file* dialog box.

In case the SNP is not yet present in the database (i.e. no data for this SNP was imported previously), check ***Provide a SNP name*** and enter the name of the SNP in the corresponding text box. In case the SNP is already present, you can check ***Select an existing SNP*** and pick the SNP from the list that appears.

From the ***Select ROX normalization channel*** drop-down list, one of the two available ROX channels (”1A” or ”2A”) can be selected to normalize the VIC and FAM channels.

If ***Use ”[FileID]_[WellID]” as sample name*** is checked, a unique sample name will be generated using the File ID, separated with a ”_” (underscore character) from the Well ID. When this option is unchecked, only the Well ID will be used as sample name.

If ***Use these settings for all files in this import*** is checked, the above settings will be applied for all selected SNP files. With this option unchecked, a *Link SNP file* dialog box will pop up for every SNP file in this import.

6.7 With ***Provide a SNP name*** checked, enter “Dr45” in the corresponding text box.

6.8 Select channel ”1A” as normalization channel.

6.9 Uncheck ***Use ”[FileID]_[WellID]” as sample name*** and uncheck ***Use these settings for all files in this import***, since the two remaining SNP files contain data for a different SNP.

6.10 Press <***OK***>.

This will pop up a *Link SNP file* dialog box for file `BMG_Tt35a`.

6.11 Provide “Tt35” as SNP name and select channel ”1A” again as normalization channel.

6.12 Leave ***Use ”[FileID]_[WellID]” as sample name*** unchecked, but this time check ***Use these settings for all files in this import***, because the last SNP file also has data for the **Tt35** SNP.

6.13 Press <***OK***> to import the SNP files.

6.14 Press the <***Close***> button of the *Autocall warnings* dialog box.

The imported SNP experiments will be plotted in the *SNP calling* window (see 3.1).

When importing SNP files as described above, the results of the auto calling might be incorrect since no NTC information is available.

Importing BMG LABTECH-formatted SNP files using a *sample template file* is the most versatile way to link well positions to sample and SNP names. It has the additional advantage that the type of sample (NTC or Unknown) can be specified. The sample template file should be a tab-delimited text file with headers, containing at least a *Well ID* (of the format ”row-column”, with ”row” and ”column” the exact row and column designations as they appear in the SNP file), *SNP name* and a *Sample name*. An example of a sample template file is included with the example SNP data.

We will use example BMG LABTECH files and the `SampleTemplateBMG.TXT` file to illustrate the import steps when using a sample template file. First, we will need to delete the previously imported SNP files.

6.15 In the *SNP files panel* of the *Main* window, click on **BMG_Dr45** to make this the active (= highlighted) SNP file and select ***SNP calling*** > ***Remove file*** or press the × button in the toolbar of the *SNP files panel*.

6.16 Confirm the delete action.

The SNP file and associated character values are removed from the database.

6.17 Repeat Instruction 6.15 to Instruction 6.16 to remove the two other SNP files (**BMG_Tt35a** and **BMG_Tt35b**).

Now that the SNP files are removed from the database, we will re-import them using a sample template file.

6.18 Repeat Instruction 6.1 to Instruction 6.5.

6.19 Check ***Use sample template file*** and browse for the `SampleTemplateBMG.TXT` file from the example data.

6.20 Press <***Next***> to proceed to the next step of the import wizard (see Figure 2.9).

When the option ***Use sample template file*** is checked in the previous step, the *Link template file page* is displayed. The columns in the selected template file need to linked and the normalization channel needs to be specified.

6.21 Click on the arrow in the Template field column next to ”File ID”: a drop-down list is displayed that contains all fields from the template file. Select ”File” to link this field to ”File ID”.

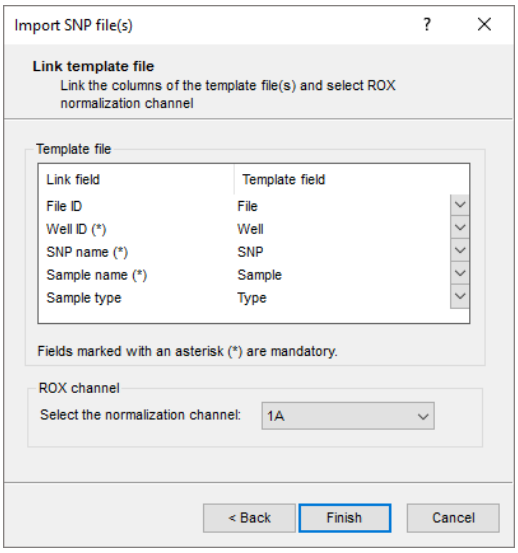

Figure 2.9: The *Link template file page*.

6.22 Repeat the above action to link ”Well ID” to ”Well, ”SNP name” to ”SNP”, ”Sample name” to ”Sample”, and ”Sample type” to ”Type”.

From the drop-down list next to ***Select the normalization channel***, one of the two ROX channels (”1A” or ”2A”) can be selected that will be used to normalize the VIC and FAM channels.

6.23 Leave ”1A” selected as normalization channel and press <***Next***>.

6.24 In the *Link conflict dialog box* that pops up, check only ***Do not link sample: ”NTC”*** and press <***OK***>.

6.25 Press the <***Close***> button of the *Auto call warnings dialog box*.

The imported SNP experiments will be plotted in the *SNP calling* window (see 3.1).

2.7 Importing Douglas Scientific files

Importing text files generated by the Douglas Scientific Array Tape system is very similar to importing BMG LABTECH files (see 2.6). Since these files only contain well positions and raw fluorescence data, an additional *sample template file* is needed to provide sample and SNP information. As opposed to BMG LABTECH PHERAstar files, only one ROX channel is available for normalization.

2.8 Importing Fluidigm files

The `Fluidigm_run1.CSV` file will be used to demonstrate the import of Fluidigm files. This file is included with the example data, which can be downloaded from the Applied Maths download page (`https://www.bionumerics.com/download/sample-data`, click on ”SNP genotyping data”).

8.1 In the *Main* window, select ***SNP calling*** > ***Import file(s)*** or press the button in the *SNP files panel* to call the *Import SNP file(s)* wizard.

8.2 Browse for the `Fluidigm_run1.CSV` example file and press <***Open***>.

8.3 Make sure to select ”Fluidigm” as ***File format***, check ***Open in SNP calling window*** and ***Auto-call SNP experiments*** and press the <***Next***> button.

8.4 Enter e.g. “[FILEID]” as ***File ID parsing string*** and press <***Next***>.

8.5 Select the ’Key’ field to link the sample information to.

8.6 Leave all other options in the *Database entries page* checked and press <***Next***>.

8.7 For the example file, link conflicts occur: Sample ”Empty” and SNP ”Empty” need not be linked and both options in the *Link conflict dialog box* can therefore be checked.

The 9,216 SNP experiments from the example Fluidigm file (96 samples $\times$ 96 SNPs), are now being imported.

8.8 When the import has finished, a number of warnings are listed in the *Auto call warnings dialog box*. Press <***Close***> to proceed to the *SNP calling* window (see 3.1).

Similar as described for the import of Applied Biosystems 7900HT files, the calls made by the Fluidigm software and their corresponding confidence values can be imported by unchecking ***Auto-call SNP experiments*** in the *Import files page* of the *Import SNP file(s)* wizard. However, since Fluidigm files do not contain information about the call mode, all calls will be imported as ”Automatic”.

2.9 Importing Tecan Safire/Infinite files

The example files `Safire_96` and `Safire_384` will be used to demonstrate the import of Tecan Safire files. These files are included with the example data, which can be downloaded from the Applied Maths download page (`https://www.bionumerics.com/download/sample-data`, click on ”SNP genotyping data”).

9.1 In the *Main* window, select ***SNP calling*** > ***Import file(s)*** or press the button in the *SNP files panel* to call the *Import SNP file(s)* wizard.

9.2 Browse for the two Tecan Safire example files and press <***Open***>.

9.3 Make sure to select ”Tecan Safire/Infinite” as ***File format***, check ***Open in SNP calling window*** and ***Auto-call SNP experiments*** and press the <***Next***> button.

9.4 Enter e.g. “[FILEID]” as ***File ID parsing string*** and press <***Next***>.

9.5 Select the ’Key’ field to link the sample information to.

The Tecan Safire or Infinite file format (*.TXT files) always contains well positions and raw fluorescence data. It can *optionally* contain sample information as well. SNP information should be provided via a *sample template file* or the marker name can be indicated during import. The latter option is only applicable when all wells from the same plate (corresponding to a single SNP file) are analyzed for the same SNP marker. Since this is the case for the example Tecan Safire files, we can use them to illustrate the import without sample template file.

9.6 Uncheck ***Use sample template file*** and press <***Next***>.

The *Link SNP file* dialog box appears (see Figure 2.10).

In case the SNP is not yet present in the database (i.e. no data for this SNP marker was imported previously), check ***Provide a SNP name*** and enter the name of the SNP in the corresponding text

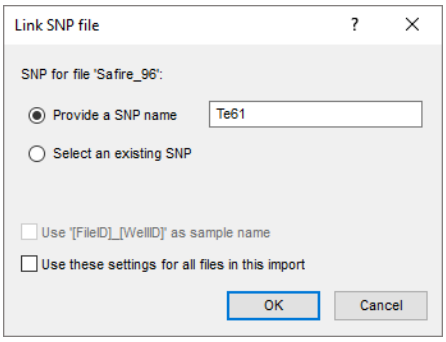

Figure 2.10: The *Link SNP file* dialog box.

box. In case the SNP is already present, you can check ***Select an existing SNP*** and pick the SNP from the list that appears.

If ***Use "[FileID]_[WellID]" as sample name*** is checked, a unique sample name will be generated using the File ID, separated with a "_" (underscore character) from the Well ID. When this option is unchecked, only the Well ID will be used as sample name.

If ***Use these settings for all files in this import*** is checked, the above settings will be applied for all selected SNP files. With this option unchecked, a *Link SNP file dialog box* will pop up for every SNP file in this import.

9.7 With ***Provide a SNP name*** checked, enter "Te61" in the corresponding text box for the `Safire_96` file.

9.8 Uncheck ***Use "[FileID]_[WellID]" as sample name*** and uncheck ***Use these settings for all files in this import***, since the other SNP file contains data for a different SNP.

9.9 Press <***OK***>. This will pop up the *Link SNP file* dialog box for file `Safire_384`.

9.10 Provide "Wr42" as SNP name and press <***OK***> to import the SNP files.

9.11 Press the <***Close***> button of the *Autocall warnings* dialog box.

The imported SNP experiments will be plotted in the *SNP calling* window (see 3.1).

When importing SNP files as described above, the results of the auto calling might be incorrect since no NTC information is available.

Importing Tecan Safire or Infinite SNP files using a *sample template file* is the most versatile way to link well positions to sample and SNP names. It has the additional advantage that the type of sample (NTC or Unknown) can be specified. The sample template file should be a tab-delimited text file with headers, containing at least a *Well ID* (of the format "row-column", with "row" and "column" the exact row and column designations as they appear in the SNP file), *SNP name* and a *Sample name*. A sample template file is included with the example SNP data.

We will use example Tecan Safire files and the `SampleTemplateSafire.TXT` file to illustrate the import steps when using a sample template file. First, we will need to delete the previously imported SNP files.

9.12 In the *SNP files panel* of the *Main* window, click on **Safire_96** to make this the active (= highlighted) SNP file and select ***SNP calling*** > ***Remove file*** or press the × button in the toolbar of the *SNP files panel*.

9.13 Confirm the delete action. The SNP file and associated character values are removed from the database.

9.14 Repeat Instruction 9.12 to Instruction 9.13 to remove the **Safire_384** SNP file.

Now that the SNP files are removed from the database, we will re-import them using a sample template file.

9.15 Repeat Instruction 9.1 to Instruction 9.5.

9.16 Check ***Use sample template file*** and browse for the `SampleTemplateSafire.TXT` file from the example data.

9.17 Press <***Next***> to proceed to the next step of the wizard (see Figure 2.11).

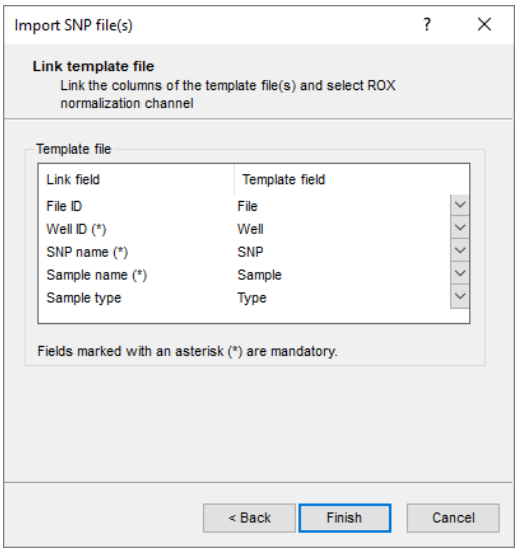

Figure 2.11: The *Link template file page*.

9.18 Click on the arrow in the Template field column next to ”File ID”: a drop-down list is displayed that contains all fields from the template file. Select ”File” to link this field to ”File ID”.

9.19 Repeat the above action to link ”Well ID” to ”Well, ”SNP name” to ”SNP”, ”Sample name” to ”Sample”, and ”Sample type” to ”Type”.

9.20 In the *Link conflict* dialog box that pops up, check only ***Do not link sample: ”NTC”*** and press <***OK***>.

9.21 Press the <***Close***> button of the *Autocall warnings* dialog box.

The imported SNP experiments will be plotted in the *SNP calling* window (see 3.1).

2.10 Warning messages after import

During import of SNP files in any of the supported data formats, following checks will be performed to ensure consistency between the BIONUMERICS database and the SNP files being imported:

- Are the mappings of the calls in the SNP file the same as the default mappings stored with the **SNP_call** character experiment type (in character info fields ’XX’, ’XY’ and ’YY’)?
- In case data is already present in the BIONUMERICS database (as **SNP_call** values) for key / marker combinations that occur in the SNP file, are the actual calls the same?

When discrepancies are detected, a warning message will appear at the end of the import action (e.g. ”Some SNP mappings have been changed. An overview file will be opened.”) and an overview of the differences is reported in an `export.csv` file, which will be opened in MS Excel by default.

Please note that, if a previously generated `export.csv` is locked for editing in MS Excel, it will not be possible for the *SNP calling plugin* to generate the `export.csv` again.

2.11 Working with SNP files

All imported SNP files are listed in the *SNP files panel* in the *Main* window. The *SNP files panel* is a grid panel that contains by default following information fields:

- ’Name’: The SNP file ID, as parsed from the SNP file name (max. 255 characters).
- ’Created’: The date and time that the SNP file was imported in the database.
- ’Modified’: The date and time that any call from the SNP file was last modified.
- ’Run date’: The date and time that the SNP file was run. This information is imported from the original SNP file, if available.
- ’Plate layout’: Either ”Tecan 96” or ”Tecan 384” (see Figure 2.12 for an explanation) and automatically filled in when importing Tecan Safire/Infinite files. The plate layout holds the information about how four 96-well plates are combined into one 384-well plate.

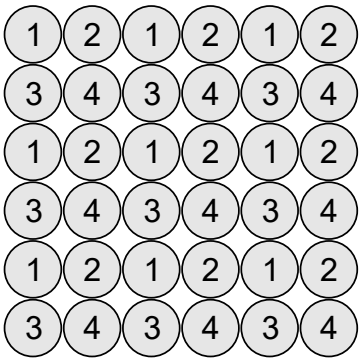

Figure 2.12: The Tecan 384 plate format explained. The circles depict the wells in a 384-well micro plate, the numbers correspond to the original 96-well plates.

11.1 Additional information fields can be added with ***SNP calling*** > ***SNP file list*** > ***Add new information field***. This pops up the *Add SNP file information field* dialog box (see Figure 2.13).

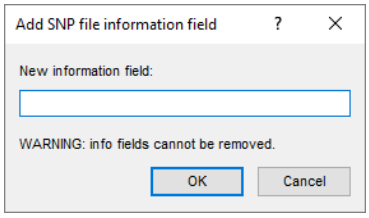

Figure 2.13: The *Add SNP file information field* dialog box.

The *Add SNP file information field* dialog box prompts for the new information field name.

11.2 SNP files can be sorted according to information in any of the fields with ***SNP calling*** > ***SNP file list*** > ***Arrange by field*** or the ⇅ button in the *SNP files panel*. This pops up the *Arrange by field* dialog box (see Figure 2.14).

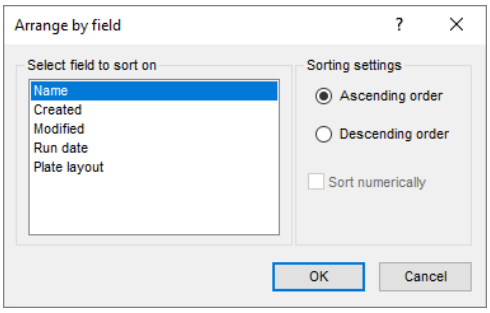

Figure 2.14: The *Arrange by field* dialog box.

The active field will be automatically highlighted in the ***Select field to sort on*** list, but a different sort field can be selected as well. Rows can be sorted in ***Ascending order*** or in ***Descending order***. If the field contains numerical data, the option ***Sort numerically*** will be available and checked by default.

Pressing <***OK***> will sort the grid panel in the desired order.

As for all grid panels, fields to be displayed in the *SNP files panel* can be set by clicking the column properties button ▾ and selecting ***Set active fields*** from the menu that appears. For additional display options of grid panels, see the Reference manual, Chapter Database objects.

11.3 The active (= highlighted) SNP file can be displayed in the *SNP calling* window with ***SNP calling*** > ***Open current file*** or by pressing the ↗ button in the toolbar of the *SNP files panel*. Double-clicking a SNP file has the same effect.

SNP files can also be selected to open a number of files in the same *SNP calling* window. After a SNP file import, all SNP files that were imported in that batch are selected by default.

11.4 To select individual SNP files, use **Ctrl+click** (holding the **Ctrl**-key on the keyboard while clicking the mouse). Repeating this action unselects the SNP file again.

11.5 A range of SNP files can be selected by clicking on the first SNP file and, while holding the **Shift**-button, clicking the last SNP file in the range.

11.6 Selected SNP files can be brought to the top of the list with ***SNP calling*** > ***SNP file list*** > ***Bring selected to top*** or the ⇡↑ button.

11.7 All selected SNP files can be opened together in a single *SNP calling* window with ***SNP calling*** > ***Open selected files*** or by pressing the ⇡↗ button in the toolbar of the *SNP files panel*.

This feature is very useful, e.g. if you want to compare a recent batch against a number of older batches to examine for possible degradation of the FRET probes.

11.8 All selected SNP files can be unselected at once with ***SNP calling*** > ***SNP file list*** > ***Clear selection*** or by pressing the ⇡× button in the toolbar of the *SNP files panel*.

11.9 To remove a SNP file from the database, click on the SNP file to activate it and select ***SNP calling*** > ***Remove file*** or press the × button from the toolbar of the *SNP files panel*. You will be asked to confirm this operation.

Ultimate flexibility is provided by the option to plot any selection of database entries in the *SNP*

calling window:

11.10 In the *Database entries* panel, select the entries that you wish to plot in a single *SNP calling* window, using any of the available manual or automatic selection options.

11.11 Select ***SNP calling*** $>$ ***Open selected entries*** to open the SNPs of the selected database entries in the *SNP calling* window.

The functionality of the *SNP calling* window is discussed in 3.1.

2.12 Auto call settings

Auto calling of SNP experiments can be performed during import of SNP files or afterwards in the *SNP calling* window (see 2.2 for an overview of supported work flows). As a consequence, the auto call settings can be accessed in two different ways:

- By pressing $<$***Auto call settings***$>$ in the first page of the *Import SNP file(s)* wizard (see 2.4).
- By selecting ***Plot*** $>$ ***Auto call settings*** in the *SNP calling* window (see 3.2).

Either action opens the *Autocall settings* dialog box, as displayed in Figure 2.15.

General tab:

The *General tab* contains the more ”basic” auto call settings, such as the number of clusters one expects and some quality parameters.

A ***Minimum confidence value*** can be specified as a percentage, below which the calls are automatically set as ”No Call”. This confidence value is related to the probability that the point does not belong to the cluster it is assigned to, based on Hotelling's T^2 test [1].

The intensity of the ROX signal is a measure for the quality of the PCR reaction. Therefore, a ***Minimum ROX channel value*** can be specified, below which reactions should be considered unreliable and hence are set as ”No Call”. Because the ROX signal is used to normalize the reporter dye signals, a low ROX signal will result in dots that appear scattered on the plot.

The ***Expected number of clusters*** should be a number between 1 and the theoretical maximum number for a given ploidy (see 1.4). When fewer clusters are expected than predicted from ploidy, it is assumed that at least one homozygous cluster is present and that all other clusters are contiguous.

With ***Warning if no NTC samples present*** is checked, a warning will be listed in the *Autocall warnings* dialog box in case a SNP file contains no blank (No Template Control or NTC) samples for a given SNP. When ***Calculate confidence for NTC samples*** is unchecked, confidence values for NTCs will be set as 100. Checking this option will allow confidence values to be calculated for the NTC samples, but calculations will take slightly longer.

A cluster calling can be performed on the data as such or after normalization against the NTC values by checking ***Perform NTC normalization first***. The NTC normalized signal S_{norm} is calculated from the raw signal S and the average NTC value $\bar{S}_{NTC}$ as

$$S_{norm} = \frac{S}{\bar{S}_{NTC}} - 1$$

When ***Normalize per file*** is checked, the average NTC value is calculated per SNP file. When ***Normalize per plate*** is checked, average NTC values per quadrant (i.e. original 96-well plate that is combined into a 384-well plate) will be used.

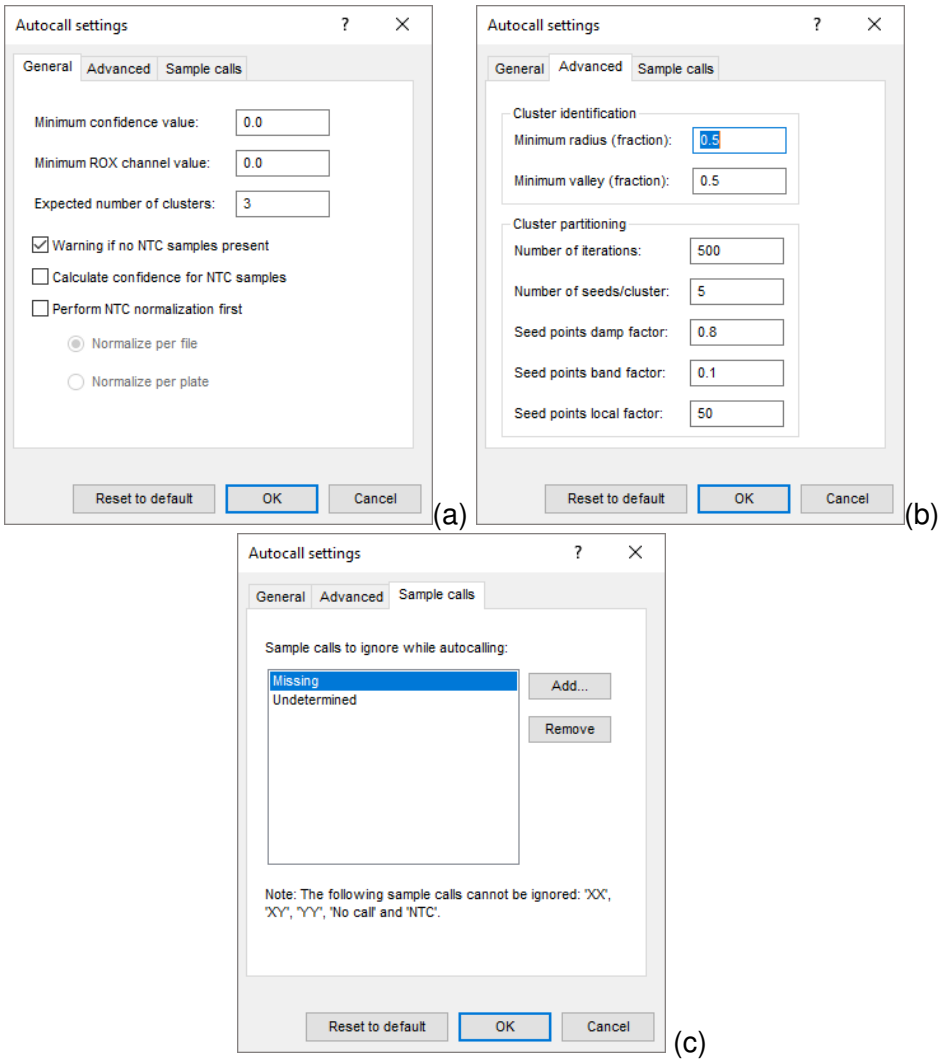

Figure 2.15: The *Autocall settings* dialog box: General settings (a), Advanced settings (b) and Sample calls (c).

Advanced tab:

The *Advanced tab* lists a number of parameters that are specific for the automatic cluster calling algorithm. The auto call is based on a partitioning that takes multiple seed points as input and is therefore a two-step process.

In the first step or ***Cluster identification*** step, the position (angle) of the seed points is determined. Using polar coordinates, an *angle distribution curve* is constructed for all points above a certain ***Minimum radius***, which can be specified as a fraction of the mean radius of all points. On the angle distribution curve, peaks are detected. A ***Minimum valley*** (expressed as a fraction of the height of the smallest peak) is imposed in order to filter out sub-clusters. The higher this parameter, the less stringent the filter will be and the more peaks will be retained. With the peak positions known, the first and last observed peak are set as the two homozygous predicted peaks and the heterozygous predicted peaks are then distributed uniformly between them. The predicted positions of the clusters are then aligned to the observed peaks, using a simple scoring function, to find the angles at which clusters occur. This step will filter out any small sub-groups or stray points.

The second step or ***Cluster partitioning*** step performs a partitioning in the X-Y coordinates based on the seed positions obtained in the first step. The ***Number of seeds per cluster*** can be entered.

The specified number of seed points will be distributed evenly between the first and last quartile radii of sample points for each cluster. If NTC samples are present, an extra seed point is set at the median position of the NTC sample points.

Each sample point is assigned to the closest seed point (in Euclidean distance). Then the position of the seed points is iteratively adjusted. The ***Number of iterations*** can be set. A new seed point position is calculated based upon the position of the points assigned to it, taking into account a ***Seed points local factor*** to weigh the contribution of the sample points to the new seed position by their distance to it (the more distant, the smaller the contribution). To avoid seed points jumping wildly over the plot, a ***Seed points damping factor*** can be set. The damping allows the seed point to travel from its old position in the direction of its new position, but only by a fraction of the distance to it. Additionally, sets of seed points are linked together by a ***Seed points band factor*** that pulls them back toward each other, ensuring that seed points representing the same cluster do not end up dispersed over the plot area.

Finally, each point is assigned to the cluster, containing the seed point it was assigned to. NTC samples that are assigned to a call cluster are set to ”No call”; samples that are assigned to the NTC cluster are also set to ”No call” and a warning is generated.

Sample calls tab:

In the *Sample calls tab*, sample calls (made by the software of the hardware manufacturer) can be specified that should be ignored by the BIONUMERICS auto calling algorithm. Initially, the list of sample calls will be empty. Press <***Add...***> to add a new sample call or <***Remove...***> to remove the highlighted sample call. The default sample types ”XX”, ”XY”, ”YY”, ”NTC” and ”Unknown” always need to be called and hence cannot be added to this list.

When an auto calling is performed in BIONUMERICS and samples are encountered with one of the calls specified here, these samples are not used as seed points by the algorithm and maintain their original call in the *SNP calling* window (see 3.1). In the **SNP_call** character type (see 4.1), these samples are called as ”No call”.

Display colors for the sample types can be set in the *Plot settings* dialog box (see 3.1).

2.13 Linking with sample information

The sample name from the SNP file (or from the sample template file) is a unique identifier that typically does not contain much descriptive information about the sample. Additional sample information can be imported in a BIONUMERICS database in several ways:

1. Importing information from a flat text file or from MS Excel.
2. Importing information from an ODBC-compatible database.
3. Connecting BIONUMERICS to an external database that contains the sample information.

Option 1 and 2 can be achieved using the *Import data* wizard. In case you want to proceed with option 3, more information about the configuration of a connected database and mapping of table names is explained in the Reference manual, Chapter The BIONUMERICS relational database.

We will illustrate here the first option and will import sample information from an external Excel file. This file is called `SampleInfo.xls` and can be found with the example data.

13.1 Select ***File*** > ***Import...*** (, **Ctrl+I**) to call the *Import data* wizard.

13.2 Press the <***Browse***> button, navigate for the `SampleInfo.xls` file and press <***Open***>.

13.3 With the ***Import fields (Excel file)*** option highlighted, press <***Finish***>.

13.4 Select ***Sheet1*** from the ***Table*** list and press <***Next***>.

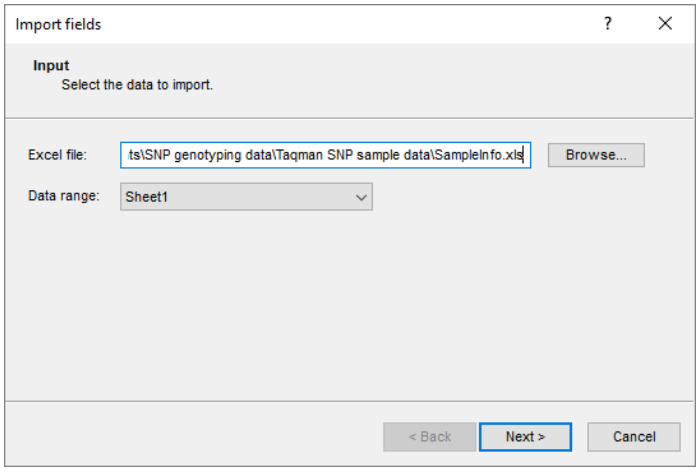

Figure 2.16: Select the Excel file and sheet name.

The way the entry information should be imported from the selected file into the database needs to be specified with an import template (see Figure 2.17).

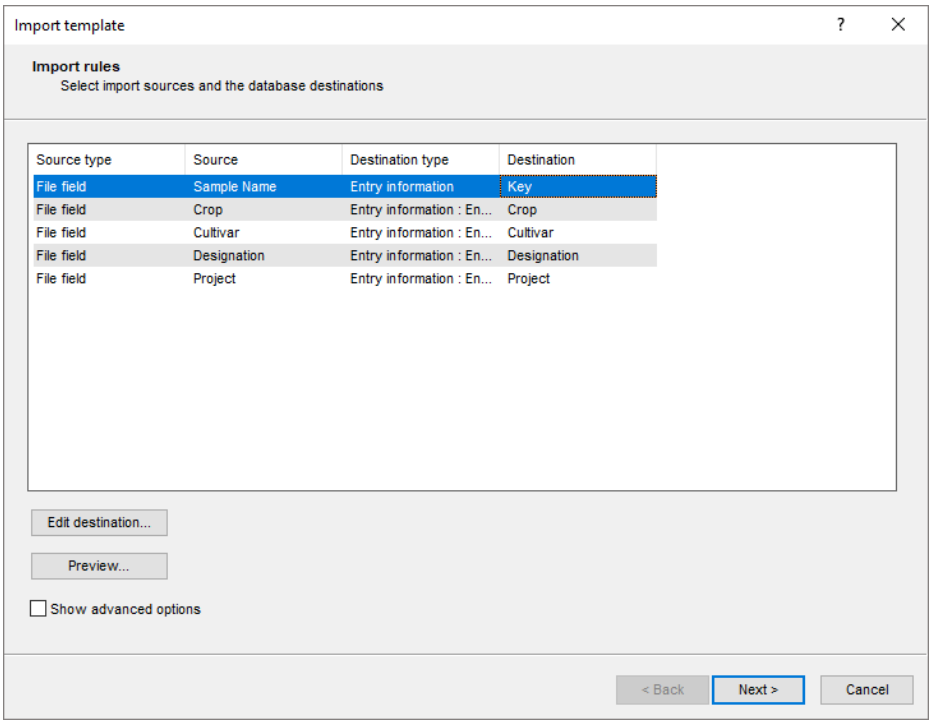

Figure 2.17: Define a new import template.

Each column in the selected sheet of the Excel file corresponds to a row entry in the grid panel (column 1 in the file corresponds to row 1 in the grid, column 2 corresponds to row 2, etc.). The text ***File field*** is specified in the ***Source type*** column and the column names are displayed in the ***Source*** column (see Figure 2.17).

13.5 Select the first row entry in the grid, press the <***Edit destination***> button and select the BIONUMERICS ***Key*** field from the list. Press <***OK***>.

The grid is updated (see Figure 2.17).

13.6 Highlight the four other external fields in the grid panel using the **Shift**-key. Press the <***Edit destination***> button and select the ***Entry info field*** option from the list. Press <***OK***>.

13.7 Press <***OK***> once more to accept the default suggested names and press <***Yes***> to confirm.

The grid is updated (see Figure 2.17).

13.8 Press <***Next***> and <***Finish***>.

13.9 Specify a template name, and optionally add a description. Press <***OK***>.

The import template is added to the list and is automatically selected.

13.10 Press <***Next***> and press <***Finish***>.

The additional sample information from the `SampleInfo.xls` file is now imported in BIONUMERICS. For more information about the functionality of the import routines, we refer to the BIONUMERICS manual.

Chapter 3

Visualizing and calling SNP experiments

3.1 The SNP calling window

The *SNP calling* window will be displayed automatically after an import of SNP files if the option ***Open in SNP calling window*** was checked in the import wizard (see Figure 2.2). Alternatively, the window can be opened by any of several available commands (see 2.11).

To illustrate the functionality of this window, we will display some of the previously imported Applied Biosystems 7900HT files (see 2.4). Proceed as follows:

1.1 In the *SNP files panel* of the *Main* window, use **Ctrl+click** to select the ”AB_Dr35_Xr40-100209”, ”AB_Lc62_01-091019”, ”AB_Lc62_02-091214”,”AB_Lc62_03-100215”, and ”AB_Tt60-100208” SNP files.

1.2 Select ***SNP calling*** > ***Open selected files*** or press the button from toolbar in the *SNP files panel*. This will display the *SNP calling* window as in Figure 3.1.

The *SNP calling* window consists of four dockable panels: the *SNPs panel*, *SNP files panel*, *Plot panel*, and *Samples panel*. The configuration of this window can be modified by docking the panels at different relative positions. To restore the default configuration of the *SNP calling* window, select ***Window*** > ***Restore default configuration***. For a detailed description of configuring windows that contain dockable panels, we refer the Reference manual, Chapter The BIONUMERICS user interface.

The *SNPs panel* is a grid panel that lists all SNPs from the selected SNP file(s) or entries. For each SNP, the ”Name” and the descriptive genotype names are displayed by default. The same information is displayed (and can be modified if necessary) in the *Character type* window of the **SNP_call** experiment (see 4.1). Information fields can be displayed or hidden by clicking on the column properties button on the right-hand side of the panel header and selecting ***Set active fields***. SNPs can be sorted according to any of the active fields with ***SNPs*** > ***Arrange by field*** or the button in the toolbar of the *SNPs panel*. This calls the *Arrange by field dialog box* as described under 2.11. From the column properties button, additional functionality that is common to all grid panels is available. For a description of this functionality, we refer to the BIONUMERICS manual.

The *SNP files panel* lists all SNP files that contain the active (i.e., highlighted) SNP from the *SNPs panel*. Fields that are displayed by default are ’Name’ (the SNP file ID), ’Created’, ’Modified’ and ’Run date’, i.e. the same fields as displayed in the *SNP files panel* in the *Main* window. Similar as described of the *SNPs panel*, fields can be hidden or displayed and SNP files can be sorted

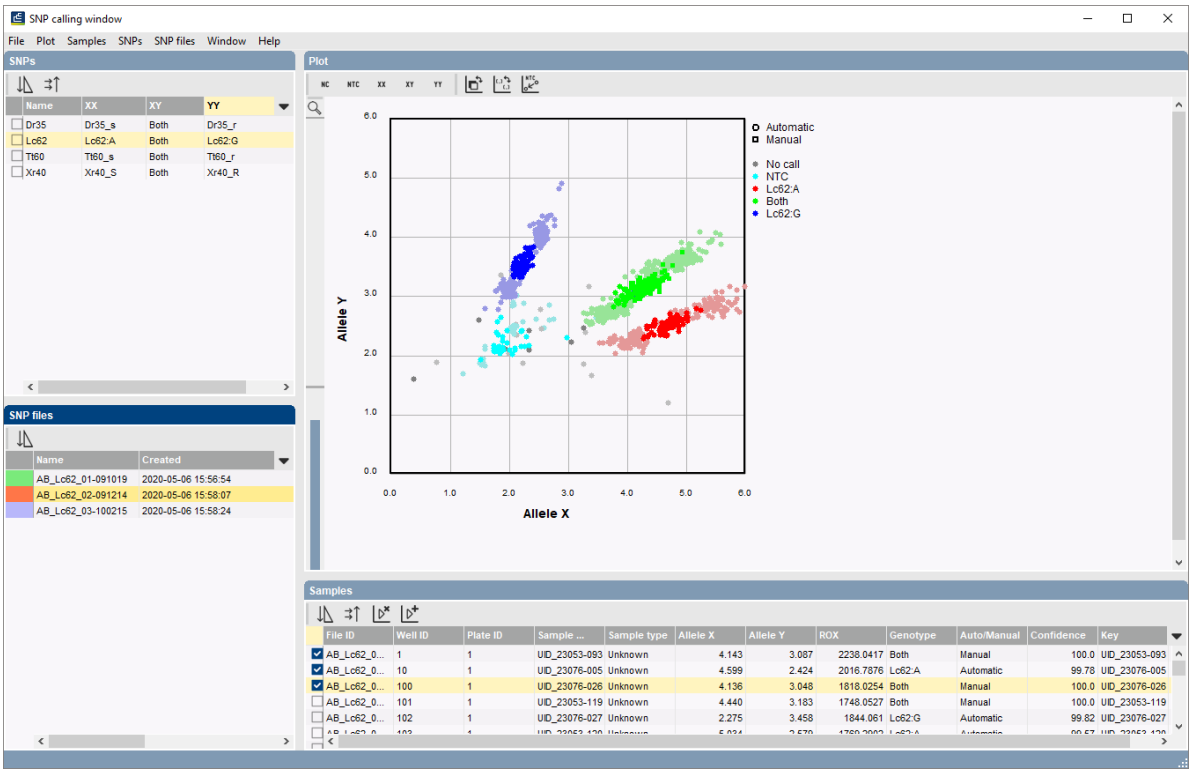

Figure 3.1: The *SNP calling* window in default configuration.

according to any active field.

In the *Plot panel*, experiments for the active SNP in the *SNPs panel* and the active SNP file(s) in the *SNP files panel* are plotted according to one of three coordinate systems: Cartesian, Polar and Contrast. On the right-hand side of the plot, a convenient legend is shown.

Only experiments from a single SNP can be displayed in the same plot.

SNP experiments from SNP files that are not active, i.e. not shown highlighted in the *SNP files panel*, are plotted as faintly colored (desaturated) dots in the *Plot panel*. These ”inactive” SNP experiments cannot be selected, nor can their color be modified (see below). This feature is very useful to display SNP experiments from a certain SNP file against a ”background” of SNP experiments from other SNP files (see Figure 3.1 for an illustration).

To select different coordinates, select ***Plot*** > ***Coordinates*** and the desired coordinates (***Cartesian***, ***Polar***,or ***Contrast***). Alternatively, press the button repeatedly to toggle through the available coordinate systems.

Other plot settings are available via ***Plot*** > ***Plot settings***. This action pops up the *Plot settings* dialog box as shown in Figure 3.2.

Plot axes tab:

In this dialog, the coordinate system can be specified (***Cartesian***, ***Polar***, and ***Contrast***).

When ***Use fixed axes*** is checked, the ***X range*** and ***Y range*** can be entered as minimum and maximum values in the corresponding text boxes. With the ***Use fixed axes*** option unchecked, the software will determine the X and Y range automatically.

Call colors tab:

This tab is only visible if ignored sample calls are defined (see 2.12). For each sample call as de-

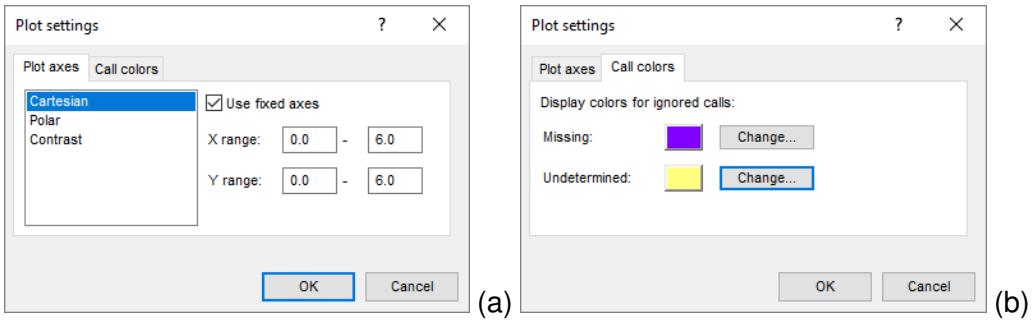

Figure 3.2: The *Plot settings* dialog box: *Plot axes tab* (a) and *Call colors tab* (b).

fined in the *Autocall settings* dialog box, a display color can be specified. Pressing the <***Change***> button next to a sample type opens the *Color* dialog box, similar as for the SNP display settings (see 1.5).

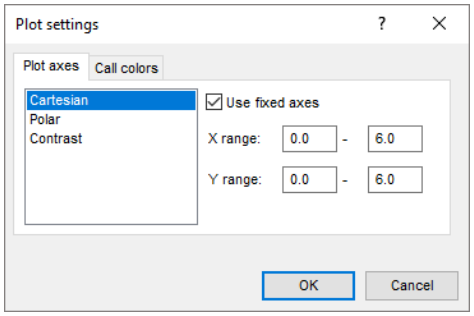

Figure 3.3: The *Plot settings* dialog box.

In the *Plot settings* dialog box, the coordinate system can be specified (***Cartesian***, ***Polar***, and ***Contrast***).

When ***Use fixed axes*** is checked, the ***X range*** and ***Y range*** can be entered as minimum and maximum values in the corresponding text boxes. With the ***Use fixed axes*** option unchecked, the software will determine the X and Y ranges automatically.

To switch between different ways of normalization, select ***Plot*** > ***Normalization*** and the desired normalization (***No normalization***, ***NTC normalization (per file)***, and ***NTC normalization (per plate)***). Alternatively, press the button repeatedly to toggle through the available normalizations. See 2.12 for the formula how the NTC normalization is calculated.

It is possible to zoom in or out on the plot, using the zoom slider. The default position of the zoom slider corresponds to an auto-fit of the plot into the panel.

A reference panel can be displayed in the plot background for each SNP marker to check for call consistency (see 4.2).

The dots on the plot, representing SNP experiments, can be colored according to call, file or confidence values.

1.3 Select ***Plot*** > ***Colors*** > ***Call*** to color the dots according to the call that was made for the SNP experiment it represents. The colors specified in the *Display settings* dialog box (see 1.5) will be used.

1.4 Select ***Plot*** > ***Colors*** > ***Confidence*** to display the confidence of the calls using the ”Quality” color map, as can be specified in the *Preferences* window. For more information about the preferences, we refer to the Reference manual, Chapter The BIONUMERICS user interface.

1.5 Select ***Plot*** > ***Colors*** > ***File*** to color the dots according to the SNP file they originate from. The colors for the comparison groups will be used. These can be set in the *Comparison* window via ***Groups*** > ***Edit group colors***.

1.6 Alternative to the procedure outlined above, press the button repeatedly to toggle through the available color modes (call, confidence or file).

The legend will explain the symbols and coloring used.

When the *SNP calling* window is closed, the current display settings will be saved. The next time that the *SNP calling* window is opened in the same database, the saved settings will be applied. Saved display settings include general settings, such as window and panel sizes, relative panel locations, etc., but also the coordinate system and the kind of coloring used in the plot.

The content of the *Plot panel* can be exported to the Windows clipboard with ***File*** > ***Copy to clipboard***. From there, it can be pasted in any application that is compatible with the Windows metafile format, e.g. Word or PowerPoint.

Individual SNP experiments can be selected from the plot:

1.7 Hold the **Ctrl**-key on the keyboard and click on a dot with the mouse to select it.

1.8 Click and drag with the mouse to select a group of dots using the lasso tool.

The selected SNP experiments are also selected in the *Samples panel* (see below).

1.9 To unselect all dots, simply click on a random position within the plot.

1.10 To select all SNP experiments in the plot at once, select ***Samples*** > ***Select all***.

When a different SNP is selected from the *SNPs panel*, the plot will be updated with other SNP experiments, but the selection will be retained on the entry level: entries that were selected for one SNP will be selected for the other SNP.

The *Samples panel* shows all SNP experiments for the active SNP in the *SNPs panel* and the active SNP file(s) in the *SNP files panel*. Following fields are displayed by default in the *Samples panel*:

- 'File ID': The SNP file ID, as parsed from the SNP file name.
- 'Well ID': A number or letter-number combination (e.g. "A1") that uniquely identifies the well position within a plate.
- 'Plate ID': The original 96-well plate (numbered 1 to 4) from which the well on the 384-well plate originates. This information is only available for the Tecan Safire/Infinite format and is dynamically calculated from the Well ID and Plate layout (see 2.11).
- 'Sample name': The name of the sample, as imported from the SNP file or sample definition file (the latter only for the BMG LABTECH format).
- 'Sample type': The type of sample, currently either "NTC" for a blank (= No Template Control) or "Unknown" for a sample that is being genotyped.
- 'Allele X': The normalized fluorescence signal for allele X.
- 'Allele Y': The normalized fluorescence signal for allele Y.
- 'ROX': The raw ROX fluorescence signal (if available).
- 'Genotype': Descriptive name of the call.

- 'Auto/Manual': Whether the call was done automatically or manually.
- 'Confidence': The confidence value of the call.
- 'Key': The key of the database entry to which the SNP experiment is linked (empty when not linked).

Fields can be displayed or hidden by clicking on the column properties button on the right-hand side of the panel header and selecting ***Set active fields***. SNP experiments can be sorted according to any of the active fields with ***Samples*** > ***Arrange by field*** or the button in the toolbar of the *Samples panel*. This calls the *Arrange by field* dialog box as described under 2.11. From the column properties button, additional functionality that is common to all grid panels is available. For a description of this functionality, we refer to the BIONUMERICS manual.

In the *Samples panel*, SNP experiments can be selected:

1.11 Use **Ctrl+click** (holding the **Ctrl**-key on the keyboard while clicking the mouse) to select an individual SNP experiment. Repeating this action unselects the SNP experiment again.

1.12 A range of SNP experiments can be selected by clicking on the first SNP experiment and, while holding the **Shift**-button, clicking the last SNP experiment in the range.

Selected SNP experiments in the *Samples panel* will be selected in the *Plot panel* as well (see above).

1.13 To bring selected SNP experiments to the top of the list, select ***Samples*** > ***Bring selected to top*** or press the button in the toolbar of the *Samples panel*.

From a selection of SNP experiments, a selection of database entries (= samples) that are linked to these SNP experiments, can be created:

1.14 Select ***Samples*** > ***Select linked entries (selection)***.

The linked entries are now selected in the database - indicated as - and can be easily retrieved by choosing ***Edit*** > ***Views*** > ***Switch to ¡Selected Objects*** > ***view*** (**Ctrl+Shift+S**).

By default, all samples are plotted in the *Plot panel*. However, using ***Samples*** > ***Remove selection from plot*** a selection (e.g. some outliers) can be removed from the plot. To plot all samples again, first use ***Samples*** > ***Select all*** and then ***Samples*** > ***Include selection in plot***.

As an exercise, we will use the functionality described above to find the calls for the **Dr35** SNP that have confidence values lower than 95 and to indicate these dots on the plot:

1.15 In the *SNPs panel*, highlight the **Dr35** SNP by clicking it. The SNP experiments from the **Dr35** SNP will now be displayed in the *Plot panel*.

1.16 Optionally, select ***Plot*** > ***Colors*** > ***Confidence*** or press the button to indicate the confidence values as colors on the plot.

1.17 In the *Samples panel*, click on the 'Confidence' field header to highlight this field.

1.18 Press the button to sort the SNP experiments according to increasing confidence values.

1.19 Click on the first SNP experiment in the list and, while holding the **Shift**-button on the keyboard, click on the last SNP experiment with a confidence lower than 95.

This action will select all SNP experiments with a confidence lower than 95, in the *Samples panel* as well as in the *Plot panel* (using blue borders around the dots). The calls with a low confidence can now be re-evaluated if deemed necessary (see 3.3).

3.2 Auto calling SNP experiments in the SNP calling window

In addition to performing an auto call while importing SNP files, SNP experiments can also be auto called afterwards in the *SNP calling* window. The auto call settings are discussed in 2.12 and can be accessed from the *SNP calling* window with ***Plot*** > ***Auto call settings***.

To perform an auto call for the currently active SNP and SNP file(s), select ***Plot*** > ***Auto call*** in the *SNP calling* window. All calls (automatic as well as manual calls) and the associated confidence values will be updated. However, the modified call information initially only exists in the *SNP calling* window. To actually save the updated call information in the database, you should select ***File*** > ***Save changed calls***. In case you try to close the *SNP calling* window when modified calls are not saved yet, the software will prompt with the question whether or not to save the call changes.

The auto call algorithm might give unexpected results when only a subset of a SNP file is displayed in the *SNP calling* window via the command ***SNP calling*** > ***Open selected entries*** (see 2.11).

We will illustrate auto calling of SNP experiments in the *SNP calling* window on the previously imported BMG LABTECH files (see 2.6). Proceed as follows:

2.1 In the *SNP files panel* of the *Main* window, select all SNP files with the ”BMG_” prefix by clicking on the first SNP file and, while holding the **Shift**-key, clicking on the last BMG LABTECH SNP file.

2.2 Select ***SNP calling*** > ***Open selected files*** or press the button from the toolbar in the *SNP files panel* to plot the BMG LABTECH SNP files in the *SNP calling* window.

These SNP experiments originate from a backcross analysis, so we only expect the genotype of the recurrent parent (here XX) or heterozygotes (XY). During import, some calls were incorrectly made, because the number of expected clusters was not specified. We will correct for this now.

2.3 Select ***Plot*** > ***Auto call settings*** to call the *Autocall settings* dialog box (see Figure 2.15).

2.4 In the *General tab*, enter “2” for ***Expected number of clusters*** and press <***OK***>.

2.5 With the **Dr45** SNP active, select ***Plot*** > ***Auto call***.

2.6 Click on the **Tt35** SNP. Make sure that both SNP files that contain this SNP are active and select ***Plot*** > ***Auto call*** again.

The SNP experiments should now be correctly called.

2.7 To save the calls to the database, select ***File*** > ***Save changed calls***.

3.3 Manually overriding auto calls

Even a sophisticated auto call algorithm sometimes calls SNP experiments differently than an experienced analyst would. Therefore, the *SNP calling plugin* allows manual overruling of calls. The mode of a call (”Automatic” or ”Manual”) is saved with the SNP experiment. For a manual call, the associated confidence value is always set to 100.

Using the lasso tool, select a group of SNP experiments for which you want to change the call.

3.1 To call a group of doubtful SNP experiments as ”No call”, select ***Plot*** > ***Change call (selection)*** > ***No call*** or press the button in the toolbar of the *Plot panel*.

3.2 A no template control (NTC) that was not indicated as such in the SNP file can be called with ***Plot*** > ***Change call (selection)*** > ***NTC*** or by pressing the NTC button.

3.3 To call the selected SNP experiments as "XX", select ***Plot*** > ***Change call (selection)*** > ***XX*** or press the XX button in the toolbar of the *Plot panel*.

3.4 To call the selected SNP experiments as "XY", select ***Plot*** > ***Change call (selection)*** > ***XY*** or press the XY button in the toolbar of the *Plot panel*.

3.5 To call the selected SNP experiments as "YY", select ***Plot*** > ***Change call (selection)*** > ***YY*** or press the YY button in the toolbar of the *Plot panel*.

3.6 Alternative to the above, select ***Plot*** > ***Change call (selection)*** > ***Change call*** to pop up the *Change call* dialog box (see Figure 3.4).

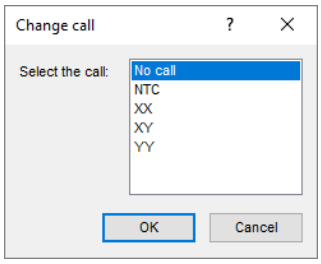

Figure 3.4: The *Change call* dialog box for a diploid organism.

The ***Select the call*** list displays all theoretically possible calls, based on the ploidy specified (see 1.4). The call that is appropriate for the selected SNP experiments can be selected from this list.

In case of polyploid SNPs, the commands from Instruction 3.3 to Instruction 3.5 will always display the *Change call* dialog box, from which the appropriate call can then be selected.

Chapter 4

Analysis and reports

4.1 The SNP_call character type experiment

Call data for SNP experiments are stored in the dedicated ”SNP experiment” object type and in the **SNP_call** character type. This makes the data amenable to different kinds of analyses in BIONUMERICS, such as cluster analysis, identification, statistical tests, etc.

The information that is stored in the **SNP_call** character type is displayed in its *Character type* window.

1.1 In the *Experiment types* panel of the *Main* window, double-click on the **SNP_call** character type to pop up the *Character type* window for this experiment (see Figure 4.1).

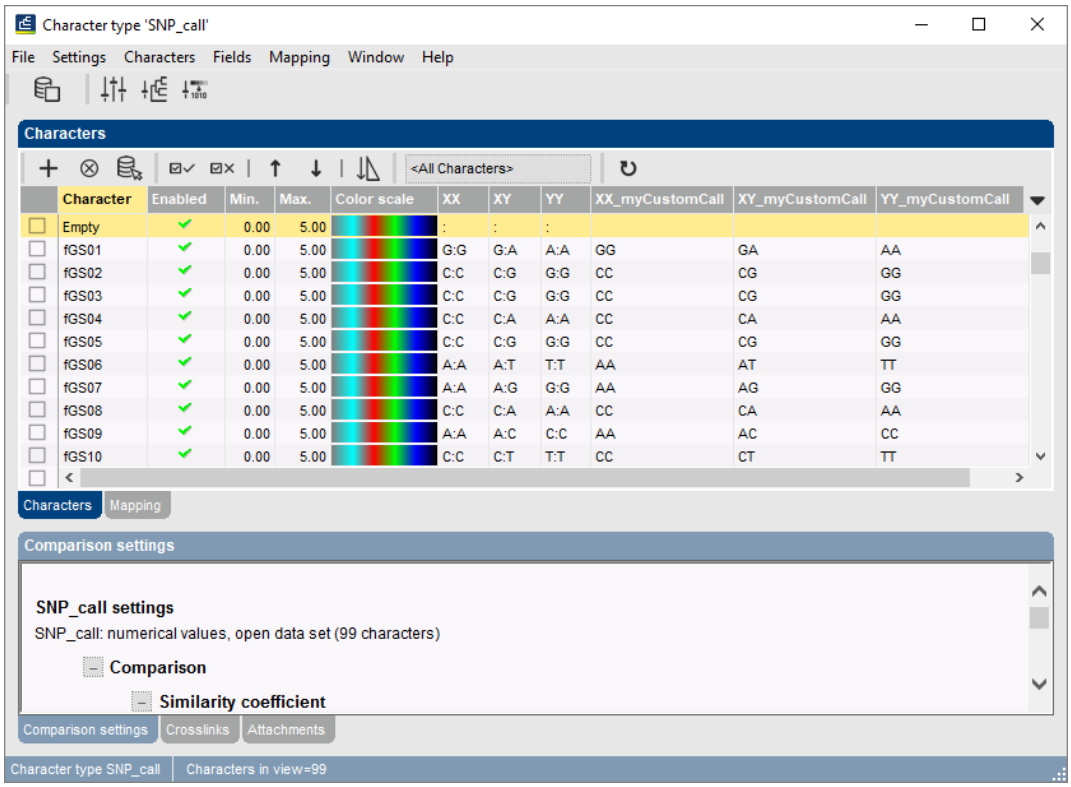

Figure 4.1: The *Character type* window for the **SNP_call** character type experiment (*Characters* panel displayed).

Each SNP is listed as a character of the **SNP_call** character type experiment. The minimum

character value 'Min.' is always zero, while the maximum character value 'Max.' depends on the ploidy that was specified for the database (see 1.4). The color scale corresponds to the colors that were specified in the SNP display settings (see 1.5).

All possible genotypes, as predicted by the ploidy, are saved as character information fields and the default descriptive name of the call is stored herein. For example, Figure 4.1 shows three columns that correspond to all genotypes for a diploid organism, i.e.'XX', 'XY' and 'YY'.

Alternative call names can be stored in additional character information fields. In this case, a character information field should be created via ***Fields*** > ***Add new field...*** for each possible genotype named after the genotype, followed by an underscore character "_" and a name for the vocabulary. For example, Figure 4.1 shows three columns ('XX_myCustomCall', 'XY_myCustomCall' and 'YY_myCustomCall') that contain the alternative call names for vocabulary "myCustomCall". Such alternative call vocabularies can be selected during export of SNP data (see 4.3).

Character information fields can be displayed or hidden by clicking on the column properties button ▾ on the right-hand side of the panel header and selecting ***Set active fields***.

In the *Mapping* panel, the character mapping is displayed.

With the exception of entering alternative descriptive genotype names in the character information fields, there is no reason to alter any settings of the **SNP_call** character type experiment directly, or to change its character values.

4.2 Specifying reference panels

For each SNP marker, a *reference panel* can be specified to check call consistency in the *SNP calling* window. A reference panel consists of one or more previously analyzed SNP data files of which the calls are known to be correct. Each time data from the marker are plotted in the *SNP calling* window, the corresponding reference panel is loaded along and displayed as read-only data in the background. Reference panels are very useful to evaluate automatic calls for consistency, especially in data files where not all genotypes are represented (e.g. YY and XY present, but not XX).

To specify reference panels, open the *Character type* window for the **SNP_call** character type (see 4.1).

The names of reference SNP files should be stored in character information fields starting with "refsnpfile" (case insensitive, "RefSnpFile" or "REFSNPFILE" will work as well):

2.1 Select ***Fields*** > ***Add new field...***, enter e.g. "RefSnpFile_1" and press <***OK***>.

If reference panels consist of multiple SNP data files, additional fields (e.g. "RefSnpFile_2", "RefSnpFile_3", . . .) can be created in the same way.

2.2 For each marker, enter the name of a reference SNP file in the newly created information field.

A list of all imported SNP files can be obtained by clicking the column properties button (▾) on the right hand side in the information fields header of the *SNP files panel* and selecting ***Save content to file***.

In the *SNP calling* window, reference SNP files are added with the prefix "[REF]" and their calls cannot be changed.

4.3 SNP calling reports

A customizable SNP calling report in text format can be generated for any selected entries.

3.1 In the *Database entries* panel of the *Main* window, select a number of entries for which you want to generate a SNP calling report.

Since SNP experiments are objects in a BIONUMERICS database, *object queries* can be used to query SNP experiments for certain features and select all entries linked to them. For more information about object queries, see the Reference manual, Chapter Database objects.

3.2 Select ***SNP calling*** > ***Export selected entries*** to pop up the *Export settings* dialog box (see Figure 4.2).

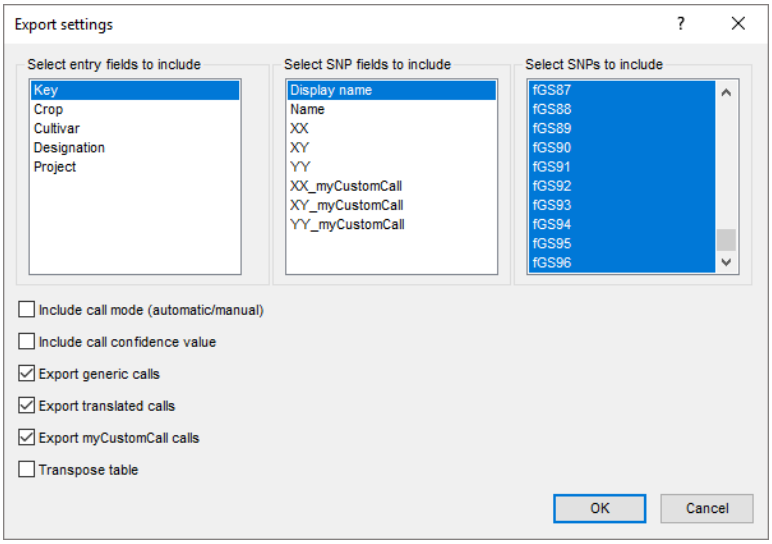

Figure 4.2: The *Export settings* dialog box.

In the top part of the dialog box, three lists are displayed:

- From the list on the left, highlight the database information fields that you wish to include in the report.
- From the middle list, select the SNP fields (character information fields in the **SNP_call** experiment type) that should be exported.
- From the list on the right, select the SNP(s) for which the call information should be reported. By default, only those SNPs for which the selected entries actually contain information are highlighted.

Checking ***Include call mode (auto/manual)*** will export an additional column for every SNP, holding the mode of the call (either ”Automatic” or ”Manual”).

Checking ***Include call confidence value*** will export an additional column for every SNP, holding the confidence value of the call.

When ***Export generic calls*** is checked, the call is exported as the ”XX”, ”XY”, ”YY”, etc. genotype designation.

When ***Export translated calls*** is checked, the call is reported as the descriptive name (see 2.4) and the generic calls are only used when no descriptive names are specified for the SNP.

An additional option (such as ***Export myCustomCall calls*** in Figure 4.2) becomes available for each alternative call vocabulary specified in the **SNP_call** experiment type) (see 4.1).

By default, the samples will be organized in rows and the SNP markers in columns in the exported file. With ***Transpose table*** checked, a transposed table will be exported, i.e. each column will represent a sample and each row a SNP.

Pressing <***OK***> will export a CSV file containing all specified information as `export.csv` to the database folder. The file will open automatically in the default editor for CSV files. On most systems, this will correspond to MS Excel.

4.4 Creating graphs from SNP calling data

A wide variety of graphs and plots, ranging in complexity from very simple to extremely sophisticated, can be created based on SNP calling data. The way it works is that the *SNP calling* window exposes a number of data sets to the *Charts and statistics* window in BIONUMERICS. These data sets are live data, meaning that, if something changes in the *SNP calling* window (e.g. the selection state or calls), this change is immediately reflected in the graphs. The *Charts and statistics* window furthermore offers the option to store the ”recipe” to create a graph (chart template), so that complex graphs can be re-created on other, similar data in just a few mouse clicks. For more information about all options available in the *Charts and statistics* window, we refer to the Reference manual, Part Charts.

To create a chart based on data that is summarized over the SNP markers, select ***SNPs*** > ***Chart and statistics***. This calls the *Create chart* dialog box (see Figure 4.3). By default, there is one pre-defined chart present to create an allelic frequency plot (”<Create SNP call histogram>”). The option ”<Create from SNP summary>” exposes the same data set, but without the histogram.

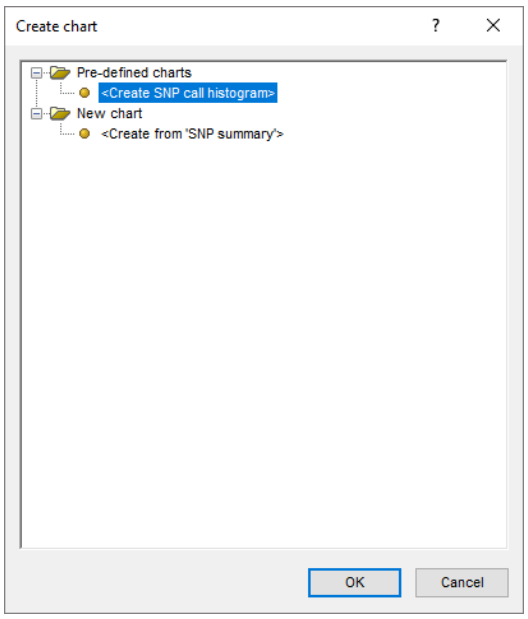

Figure 4.3: The *Create chart* dialog box.

4.1 Select ”<Create SNP call histogram>” and press <***OK***>. This creates a histogram as in Figure 4.4.

The appearance of the histogram can still be modified if needed.

To use the currently displayed samples (i.e. individual SNP experiments or data points in the plot)

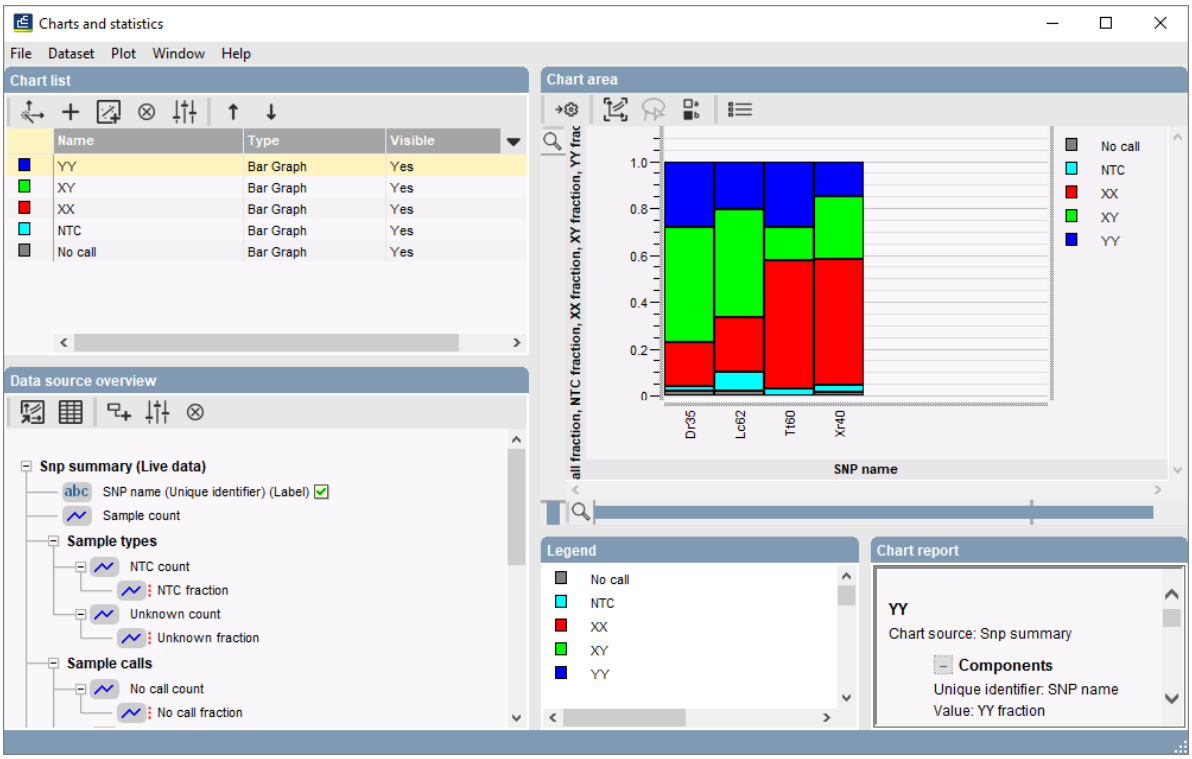

Figure 4.4: The *Charts and statistics* window, with an allelic frequency histogram displayed.

as a data set, select ***Samples*** > ***Chart and statistics***. From the *Create chart* dialog box that appears, select "<Create from "Samples">" and press <***OK***> (see Figure 4.5).

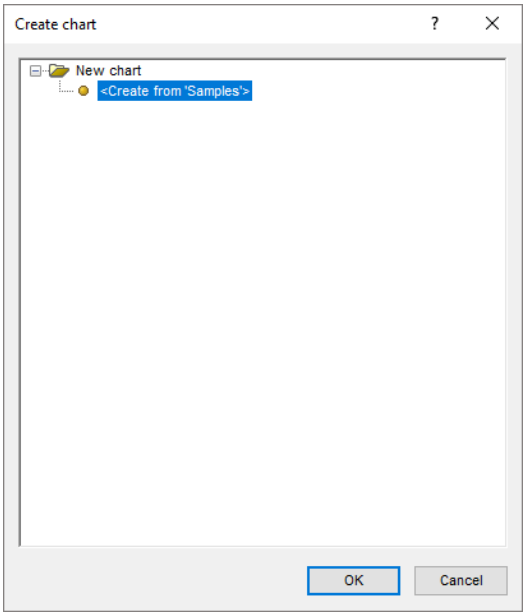

Figure 4.5: The *Create chart* dialog box.

This displays the *Charts and statistics* window again, from which graphs can be created based on the samples. See Figure 4.6 for a few examples of possible graphs.

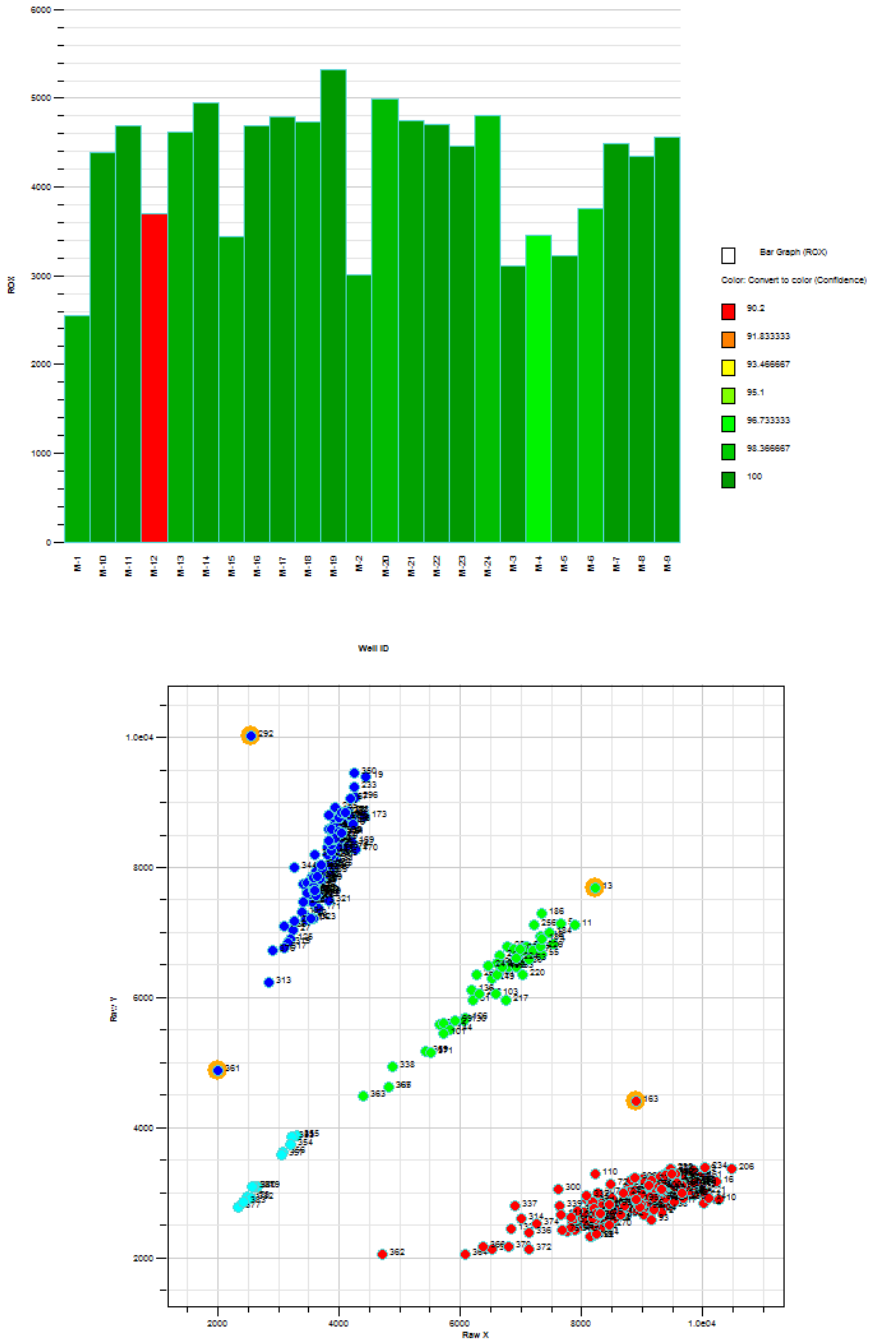

Figure 4.6: A few examples of the numerous graphs that can be generated from the ”Samples” data set in the *Charts and statistics* window.

4.5 Analyzing SNP data in comparisons

Since all SNP calling data are saved as character values in the **SNP_call** character type experiment (see 4.1), the latter experiment can be used for cluster analysis, identification, statistical tests, etc.. Although this section is not intended to give a comprehensive overview of all analysis possibilities in BIONUMERICS, we will highlight a few useful features in the context of SNP data analysis.

In BIONUMERICS, the *Comparison* window is the main analysis window. The Fluidigm samples, which were imported in 2.8, will be used to illustrate some of the analysis features available in this

window. Additional sample information will be available if you went through the tutorial described in 2.13. Since a comparison is always created based on selected entries in the database, we first need to select all entries that are linked to SNP experiments from the Fluidigm_run1 SNP file.

5.1 Click in *Database entries* panel and select ***Edit*** > ***Find object in list...*** (, **Ctrl+Shift+F**). This pops up the *Find* dialog box.

5.2 In the *Find* dialog box, enter “Carrot” as search string and press the <***Select all***> button.

If the additional sample information was imported (see 2.13), the above action should select all 94 entries from the example Fluidigm file.

5.3 Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** (+) to create a new comparison for the selected entries.

5.4 To display the SNP calls using the colors as specified in the SNP display settings (see 1.5), click the eye button ( ) left of the **SNP_call** experiment, listed in the *Experiments* panel.

To perform a cluster analysis on the SNP calls, which will cluster similar entries together, proceed as follows:

5.5 Select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***.

5.6 Click on ***Categorical (values)*** as ***Similarity coefficient*** and press <***Next***>.

5.7 Leave the settings in the *Page 2* wizard page at their default values and press <***Next***>.

This displays a dendrogram for the entries, which were clustered according to similarity of the calls.

Using composite data sets, it is possible to perform a *transversal* cluster analysis, based on similarities among entries *and* characters (see the Reference manual, Chapter Cluster analysis of composite data sets for more information).

A question that often pops up in a backcross analysis is to find offspring with the highest similarity to the recurrent parent. This can quickly be achieved as follows:

5.8 Click on the recurrent parent in the breeding experiment. In the example data, this is the sample that has ”Parent1” in the ’Designation’ information field.

5.9 Select ***Edit*** > ***Arrange entries*** > ***Arrange entries by similarity*** ( ).

5.10 Answer <***Yes***> to the question ”Do you want to use the existing similarity matrix for ”SNP_call”?”.

The samples are now arranged according to decreasing similarity with the recurrent parent. The *Similarities* panel shows the actual similarity values.

Another useful feature is the ability to sort samples according to the call for a certain SNP.

5.11 Highlight the SNP you want to arrange according to call by clicking on its name in the header of the *Experiment data* panel, e.g. **fGS12**.

5.12 Select ***Characters*** > ***Sort by character value*** ( ).

The samples are now arranged according to their call for SNP **fGS12** in the order ”no call”, ”NTC”, ”XX”, ”XY” and ”YY”.

Contingency tables, one of the available statistics tools, can be used to quickly search for samples with a desired call combination of two SNPs.

5.13 Select ***Statistics*** > ***Chart and statistics...*** (**F7**).

5.14 In the *Create chart* dialog box, select ***Comparison entries*** and press <***OK***>.

5.15 Select the ***Contingency table*** as chart type and press <***Next***>.

5.16 In the final step, select two mapping characters (e.g. ”fGS01” and ”fGS12”) from the lists and press <***Finish***> to display a contingency table in the *Charts and statistics* window.

5.17 In the contingency table, use **Ctrl+click** to select the single entry that has a ”YY” genotype for both **fGS01** and **fGS12**.

The entry is now selected in the database and can be brought to the top of the list with ***Edit*** > ***Arrange entries*** > ***Bring selected entries to top*** (, **Ctrl+T**).

Other useful features in the context of SNP data analysis, not linked to the *Comparison* window, include:

- When sufficient well-characterized samples are available, identification projects (see the Reference manual, Chapter Identification projects) can be used to identify new samples, e.g. in the context of trueness-to-type analysis.
- Advanced queries or decision networks could be used to select samples that fulfill certain call criteria for a number of SNPs. For example, Figure 4.7 shows an advanced query and a decision network that select all samples in the database that have a resistant genotype for SNPs **Dr35**, **Xr40** and **Tt60**.

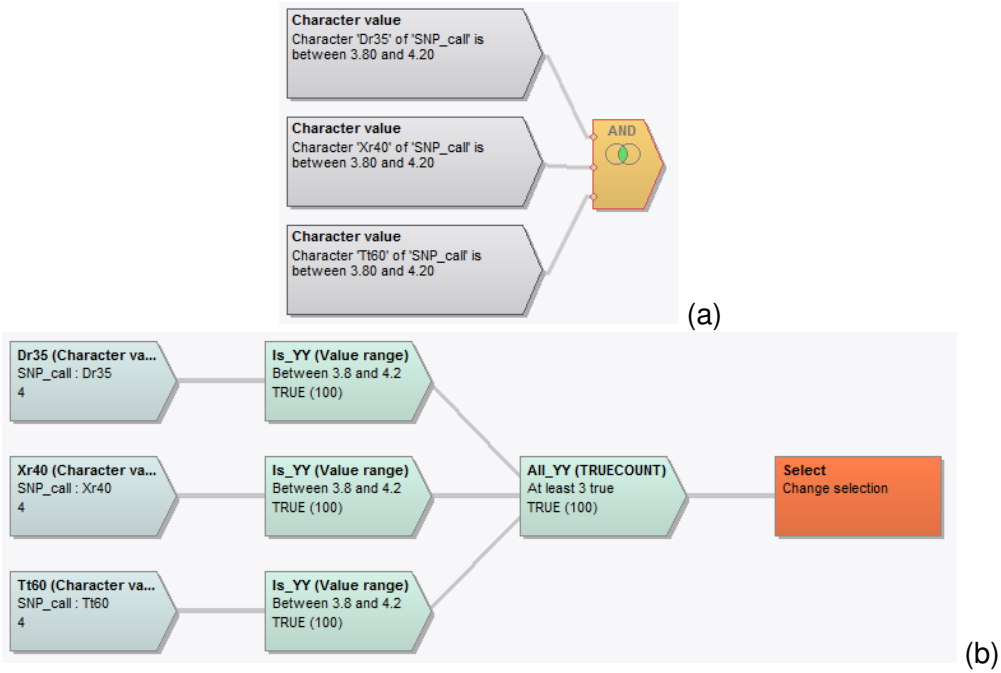

Figure 4.7: Examples of an advanced query (a) and a decision network (b) for automatically selecting samples that fulfill certain call criteria for specific SNPs.

The use of decision networks is supported in the **BIONUMERICS-MALDI**, **BIONUMERICS-SEQ** and **BIONUMERICS-SUITE**.

For a detailed description of the numerous analysis possibilities in the BIONUMERICS software, we refer to the BIONUMERICS reference manual.

Bibliography

[1] J.D. Jobson. *Applied multivariate data analysis: regression and experimental design*, volume 1. Springer, 1991.

bioMérieux S.A. • 69280 Marcy l'Étoile • France • Tel.: + 33 (0)4 78 87 20 00 • Fax: +33 (0)4 78 87 20 90
www.biomerieux.com