Convert Dosages to Genotypes

Author: Autumn Laughbaum, Golden Helix, Inc.

Overview

This script converts allelic dosage values to genotypes based on user-specified thresholds. The dosage data may be in Single- or Double-Dosage format and may have samples in the row labels or column headers. If the samples are in the column headers, the spreadsheet may contain map information and allele translation values.

This script is designed to run on a spreadsheet that contains dosage data and has already been imported into SVS. Thus, it is expected that the user import their dosage data using one of the many SVS import tools. In particular, the *Import Text* and *Import Third-Party* tools may be used for this purpose.

Recommended Directory Location

Save the script to the following directory:
*..**\Application Data\Golden Helix SVS\UserScripts\Spreadsheet\Edit\**

Note: The **Application Data** folder is a hidden folder on Windows operating systems and its location varies between XP and Vista. The easiest way to locate this directory on your computer is to open SVS and select **Tools >Open Folder > UserScripts Folder**. If saved to the proper folder, this script will be accessible from the spreadsheet **Edit** menu.

User Specified Options

First User Prompt:

- **Dosage format:**
  - Single-Dosage Format – Choose this option if each sample has only one row or column (depending on orientation) per marker. Generally in this case, the numeric values will range from 0 to 2.
  - Double-Dosage Format - Choose this option if each sample has two rows or columns per marker that contain the genotype dosages, p1 and p2, for A_A and A_B. The third genotype dosage (B_B) is calculated as 1-p1-p2. Dosages in this format generally range from 0 to 1.
    - The dosage columns/rows that correspond to the same sample must be adjacent. If the sample names do not match, the two names are concatenated with ‘-‘.
- **Samples are found in**
  - Row Labels – Choose this option if the row labels contain the sample names and the column headers contain the marker names.

    - Column Headers – Choose this option if the row labels contain the marker names and the column headers contain the sample names.
- **Numeric types to convert**
    - The spreadsheet is scanned for appropriate column types (Real, Integer and Binary) and the types and counts are displayed. Check the column types that contain the dosage data.
    - *Note*: If a dosage column were to contain only 0 and 1 the importer would detect it as binary. The same can be applied to columns containing 0,1 and 2 that are detected as integer. On the other hand, if the spreadsheet contains additional non-dosage columns (such as marker map data), you may want to uncheck some column types.

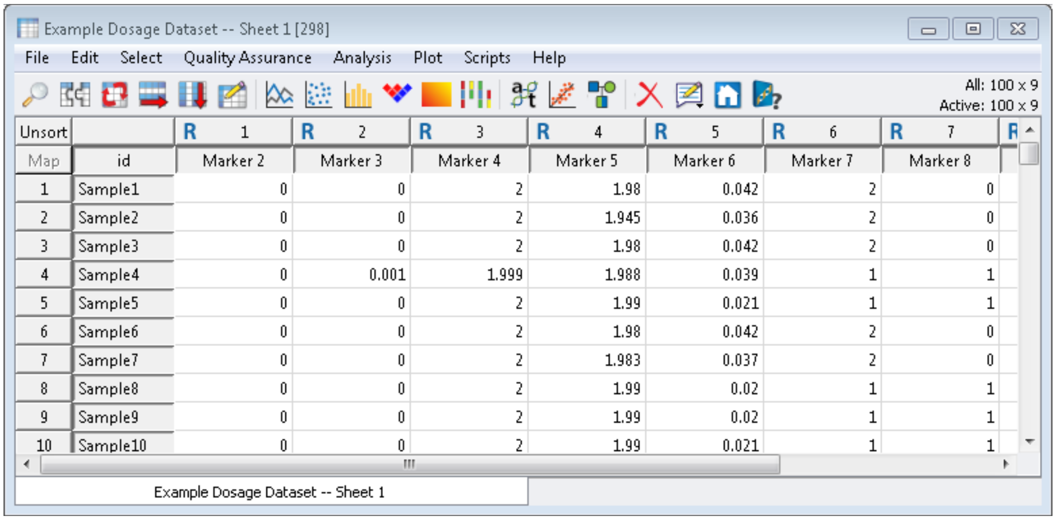
Second User Prompt:

- **Threshold Values**:
    - If data is in single-dosage format, the user must specify three ranges for A_A, A_B, and B_B, such that if the dosage value for a cell falls within the range, the genotype is called.
        - If the dosage value is not within any range, a missing genotype is called.
    - If the data is in double-dosage format, the user must specify one threshold value, such that if one of p1, p2 or p3 is larger than that value, the corresponding genotype is called.
        - Note that if two of the three dosages are larger than that value, precedence is given in the order of p1 > p2 > p3.
    - *Note*: In both cases, <= and >= are used to compare dosage values against threshold values.
- Additional options for spreadsheets that have samples in columns:
    - **Create Marker Map from columns** – if the spreadsheet contains chromosome and position information in columns, a marker map can be created and applied to the converted spreadsheet.
        - The column containing the marker names must be categorical and can also be the row labels.
        - The column containing the chromosome must be categorical or integer and can also be the row labels.
        - The column containing the position must be integer.
    - **Translate Alleles** – if the spreadsheet contains allele translation values for A and B (or A1 and A2), the user can optionally code the genotypes with the translated alleles.

Using the Script

1. Open a spreadsheet containing dosage data in Real, Integer and/or Binary columns.
2. Select **Edit > Convert Dosages to Genotypes.**

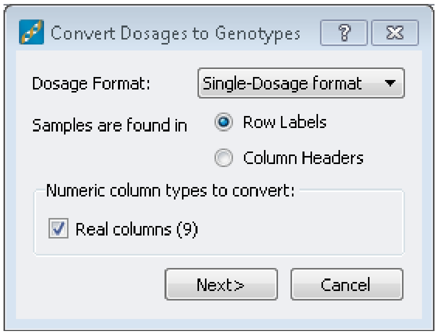
3. **CASE 1**: Samples in Row Labels, Single-Dosage Format (see Figure 1).


Example Dosage Dataset -- Sheet 1 [298]
File Edit Select Quality Assurance Analysis Plot Scripts Help
All: 100 x 9
Active: 100 x 9
Example Dosage Dataset -- Sheet 1

Unsort		R 1	R 2	R 3	R 4	R 5	R 6	R 7
Map	id	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6	Marker 7	Marker 8
1	Sample1	0	0	2	1.98	0.042	2	0
2	Sample2	0	0	2	1.945	0.036	2	0
3	Sample3	0	0	2	1.98	0.042	2	0
4	Sample4	0	0.001	1.999	1.988	0.039	1	1
5	Sample5	0	0	2	1.99	0.021	1	1
6	Sample6	0	0	2	1.98	0.042	2	0
7	Sample7	0	0	2	1.983	0.037	2	0
8	Sample8	0	0	2	1.99	0.02	1	1
9	Sample9	0	0	2	1.99	0.02	1	1
10	Sample10	0	0	2	1.99	0.021	1	1

Figure 1. Dosage data in Single-Dosage format with samples in row labels

a. In the first dialog, leave **Spreadsheet is in Single-Dosage format (one column/row per sample)** checked.
b. Also leave **Row Labels** selected after **Samples are found in**.
c. Since this spreadsheet contains only real columns, this is the only numeric type available. Leave it checked and click **Next>.**



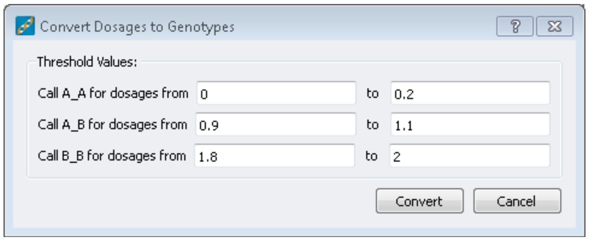
Dosage Format: Single-Dosage format

Samples are found in ● Row Labels ○ Column Headers

Numeric column types to convert:

☑ Real columns (9)

d. In the next prompt, specify minimum and maximum threshold values for A_A, A_B and B_B. Click **Convert**.



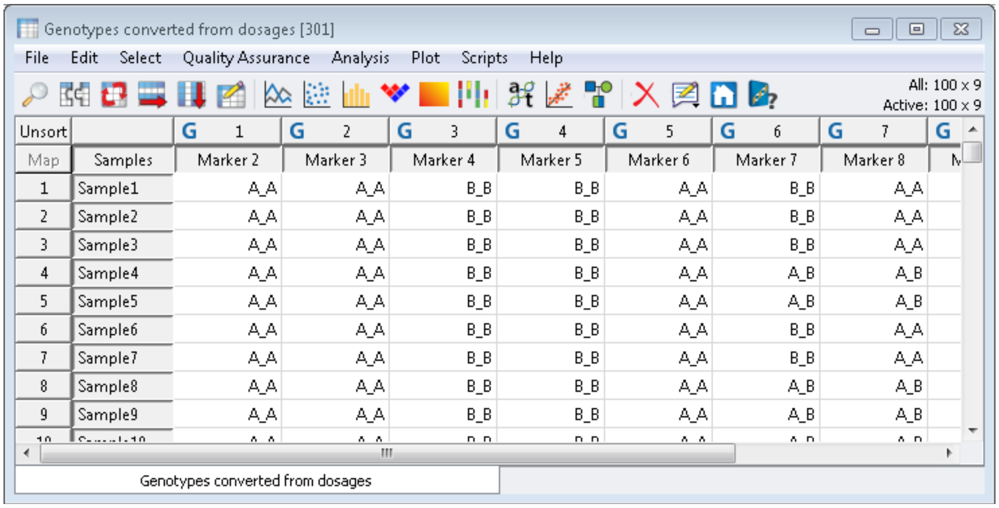
Threshold Values:

Call A_A for dosages from 0 to 0.2

Call A_B for dosages from 0.9 to 1.1

Call B_B for dosages from 1.8 to 2

e. The resulting spreadsheet contains the converted genotypes and is a child of the original spreadsheet.



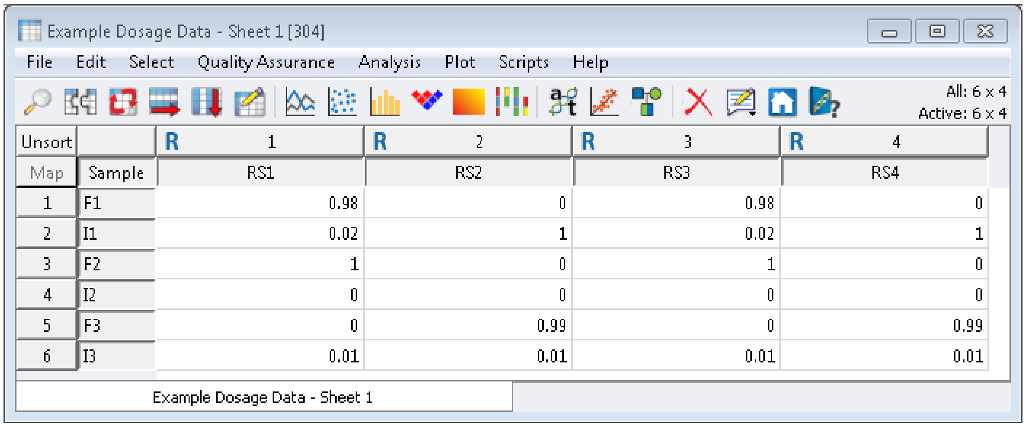
Genotypes converted from dosages [301]

All: 100 x 9
Active: 100 x 9

Unsort		G 1	G 2	G 3	G 4	G 5	G 6	G 7
Map	Samples	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6	Marker 7	Marker 8
1	Sample1	A_A	A_A	B_B	B_B	A_A	B_B	A_A
2	Sample2	A_A	A_A	B_B	B_B	A_A	B_B	A_A
3	Sample3	A_A	A_A	B_B	B_B	A_A	B_B	A_A
4	Sample4	A_A	A_A	B_B	B_B	A_A	A_B	A_B
5	Sample5	A_A	A_A	B_B	B_B	A_A	A_B	A_B
6	Sample6	A_A	A_A	B_B	B_B	A_A	B_B	A_A
7	Sample7	A_A	A_A	B_B	B_B	A_A	B_B	A_A
8	Sample8	A_A	A_A	B_B	B_B	A_A	A_B	A_B
9	Sample9	A_A	A_A	B_B	B_B	A_A	A_B	A_B

Genotypes converted from dosages

4. **CASE 2**: Samples in Row Labels, Double-Dosage Format



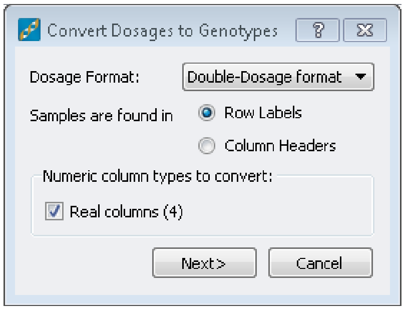
Example Dosage Data - Sheet 1 [304]

All: 6 x 4
Active: 6 x 4

Unsort		R 1	R 2	R 3	R 4
Map	Sample	RS1	RS2	RS3	RS4
1	F1	0.98	0	0.98	0
2	I1	0.02	1	0.02	1
3	F2	1	0	1	0
4	I2	0	0	0	0
5	F3	0	0.99	0	0.99
6	I3	0.01	0.01	0.01	0.01

Example Dosage Data - Sheet 1

a. In the first dialog, uncheck **Spreadsheet is in Single-Dosage format (one column/row per sample)**.
b. Leave **Row Labels** selected after **Samples are found in**.
c. Since this spreadsheet contains only real columns, this is the only numeric type available. Leave it checked and click **Next>.**



Convert Dosages to Genotypes

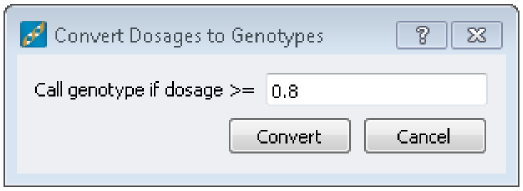
Dosage Format: Double-Dosage format

Samples are found in ● Row Labels
○ Column Headers

Numeric column types to convert:

☑ Real columns (4)

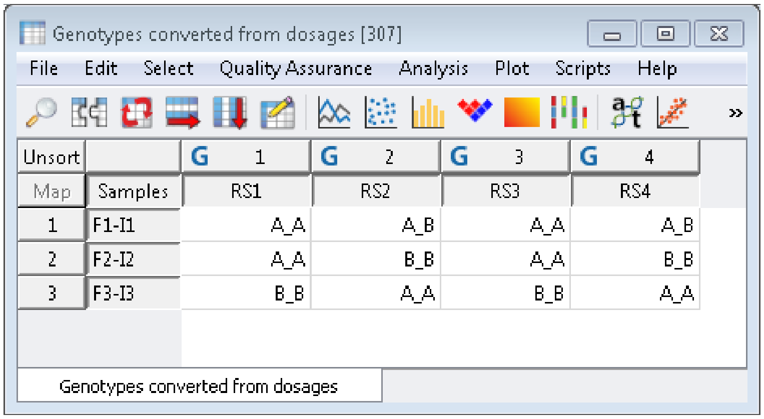
d. In the next prompt, specify the minimum dosage value for a called genotype. Click **Convert**.



Convert Dosages to Genotypes

Call genotype if dosage >= 0.8

e. The resulting spreadsheet contains the converted genotypes and is a child of the original spreadsheet.

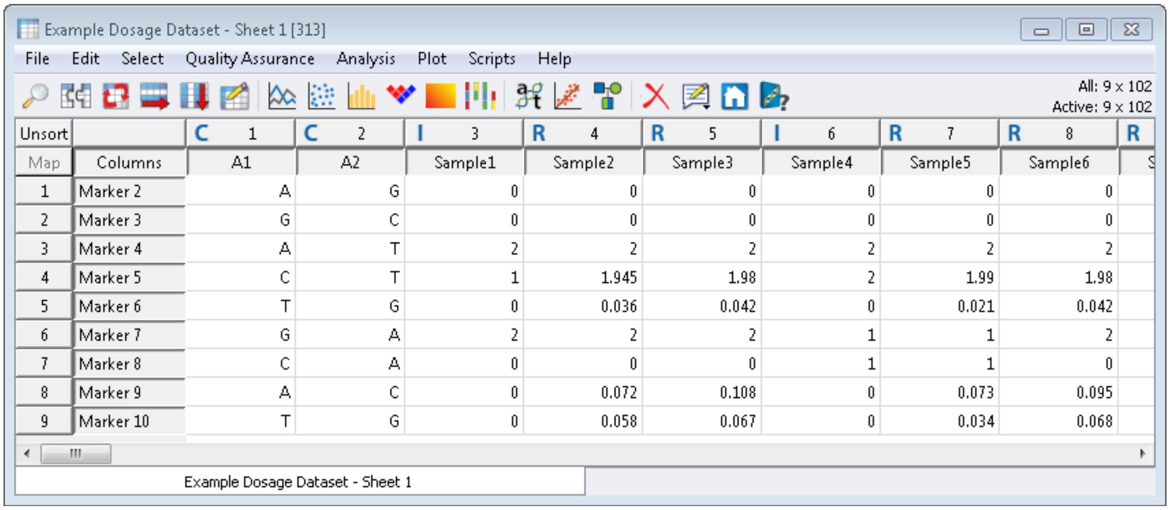


Genotypes converted from dosages [307]

Unsort		G 1	G 2	G 3	G 4
Map	Samples	RS1	RS2	RS3	RS4
1	F1-I1	A_A	A_B	A_A	A_B
2	F2-I2	A_A	B_B	A_A	B_B
3	F3-I3	B_B	A_A	B_B	A_A

Genotypes converted from dosages

5. **CASE 3**: Samples in Column Headers, Single-Dosage Format.



Example Dosage Dataset - Sheet 1 [313]

All: 9 x 102
Active: 9 x 102

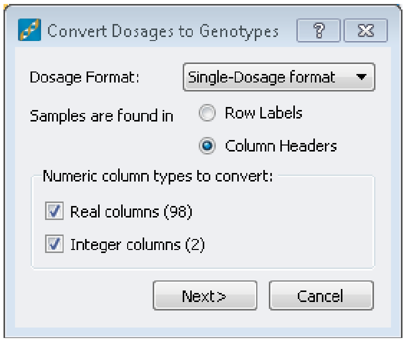
Unsort		C 1	C 2	I 3	R 4	R 5	I 6	R 7	R 8
Map	Columns	A1	A2	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
1	Marker 2	A	G	0	0	0	0	0	0
2	Marker 3	G	C	0	0	0	0	0	0
3	Marker 4	A	T	2	2	2	2	2	2
4	Marker 5	C	T	1	1.945	1.98	2	1.99	1.98
5	Marker 6	T	G	0	0.036	0.042	0	0.021	0.042
6	Marker 7	G	A	2	2	2	1	1	2
7	Marker 8	C	A	0	0	0	1	1	0
8	Marker 9	A	C	0	0.072	0.108	0	0.073	0.095
9	Marker 10	T	G	0	0.058	0.067	0	0.034	0.068

Example Dosage Dataset - Sheet 1

a. In the first dialog, leave **Spreadsheet is in Single-Dosage format (one column/row per sample)** checked.

b. Select **Column Headers** after **Samples are found in**.

c. This spreadsheet contains integer and real columns that contain dosage data. Leave both types checked and click **Next>.**



Convert Dosages to Genotypes

Dosage Format: Single-Dosage format

Samples are found in ○ Row Labels ● Column Headers

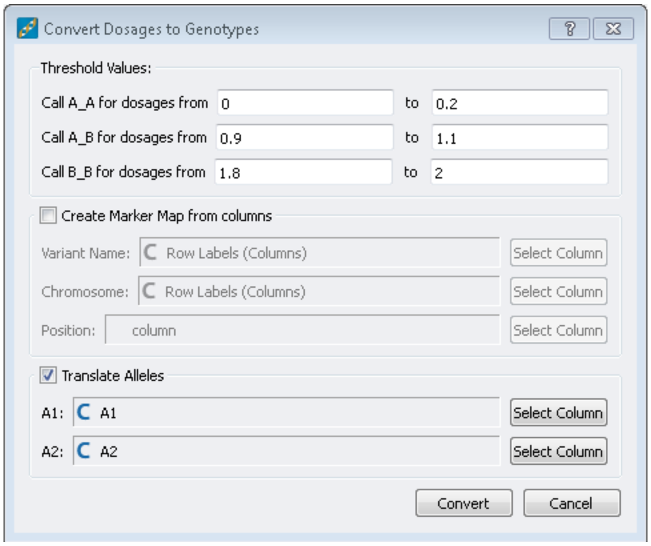
Numeric column types to convert:

☑ Real columns (98)

☑ Integer columns (2)

d. In the next prompt, specify the minimum and maximum dosage values for each genotype.

e. This spreadsheet also contains allele translation columns for A and B. Check **Translate Alleles** and choose the appropriate columns (this may be auto-detected and checked in your own spreadsheet).



Convert Dosages to Genotypes

Threshold Values:

Call A_A for dosages from 0 to 0.2

Call A_B for dosages from 0.9 to 1.1

Call B_B for dosages from 1.8 to 2

□ Create Marker Map from columns

Variant Name: Row Labels (Columns)

Chromosome: Row Labels (Columns)

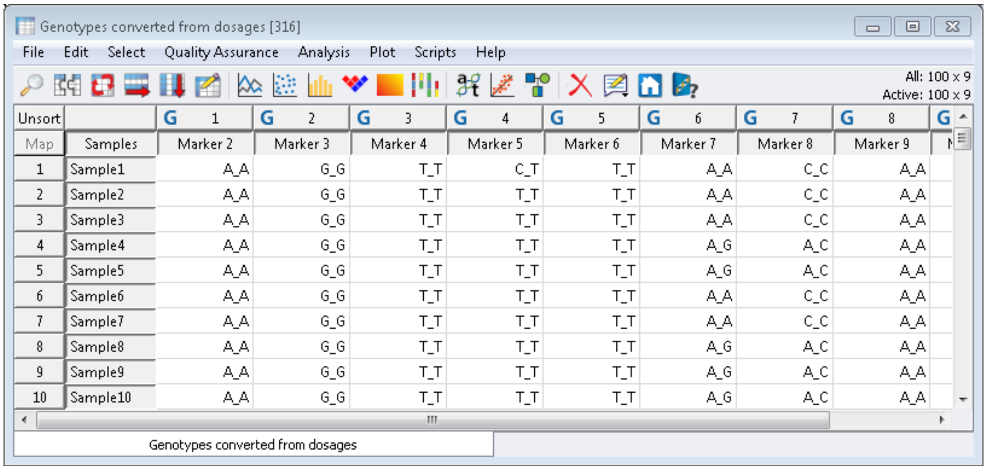
Position: column

☑ Translate Alleles

A1: A1

A2: A2

f. The resulting spreadsheet contains the converted genotypes (with translated alleles) and is a child of the original spreadsheet.



Genotypes converted from dosages [316]

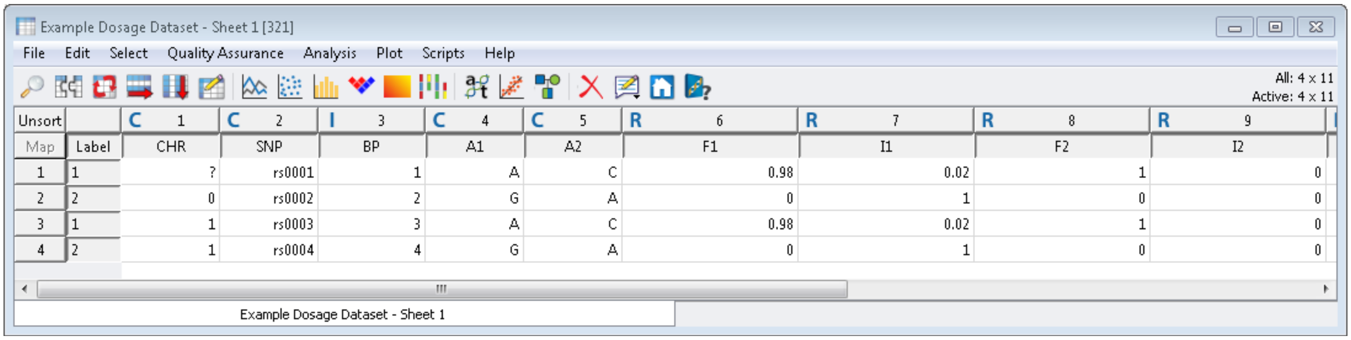
File Edit Select Quality Assurance Analysis Plot Scripts Help

All: 100 x 9
Active: 100 x 9

Unsort		G 1	G 2	G 3	G 4	G 5	G 6	G 7	G 8	G
Map	Samples	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6	Marker 7	Marker 8	Marker 9	M
1	Sample1	A_A	G_G	T_T	C_T	T_T	A_A	C_C	A_A	
2	Sample2	A_A	G_G	T_T	T_T	T_T	A_A	C_C	A_A	
3	Sample3	A_A	G_G	T_T	T_T	T_T	A_A	C_C	A_A	
4	Sample4	A_A	G_G	T_T	T_T	T_T	A_G	A_C	A_A	
5	Sample5	A_A	G_G	T_T	T_T	T_T	A_G	A_C	A_A	
6	Sample6	A_A	G_G	T_T	T_T	T_T	A_A	C_C	A_A	
7	Sample7	A_A	G_G	T_T	T_T	T_T	A_A	C_C	A_A	
8	Sample8	A_A	G_G	T_T	T_T	T_T	A_G	A_C	A_A	
9	Sample9	A_A	G_G	T_T	T_T	T_T	A_G	A_C	A_A	
10	Sample10	A_A	G_G	T_T	T_T	T_T	A_G	A_C	A_A	

Genotypes converted from dosages

6. **CASE 4**: Samples in Column Headers, Double-Dosage Format.



Example Dosage Dataset - Sheet 1 [321]

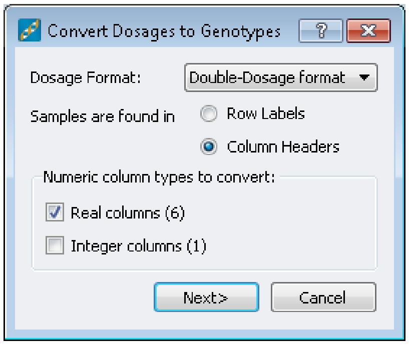
File Edit Select Quality Assurance Analysis Plot Scripts Help

All: 4 x 11
Active: 4 x 11

Unsort		C 1	C 2	I 3	C 4	C 5	R 6	R 7	R 8	R 9
Map	Label	CHR	SNP	BP	A1	A2	F1	I1	F2	I2
1	1	?	rs0001	1	A	C	0.98	0.02	1	0
2	2	0	rs0002	2	G	A	0	1	0	0
3	1	1	rs0003	3	A	C	0.98	0.02	1	0
4	2	1	rs0004	4	G	A	0	1	0	0

Example Dosage Dataset - Sheet 1

a. In the first dialog, uncheck **Spreadsheet is in Single-Dosage format (one column/row per sample)**.
b. Select **Column Headers** after **Samples are found in**.
c. This spreadsheet has real columns that contain dosage data and an integer column with position. Since you do not want to convert position, uncheck **Integer columns (1)** click **Next>.**



Convert Dosages to Genotypes

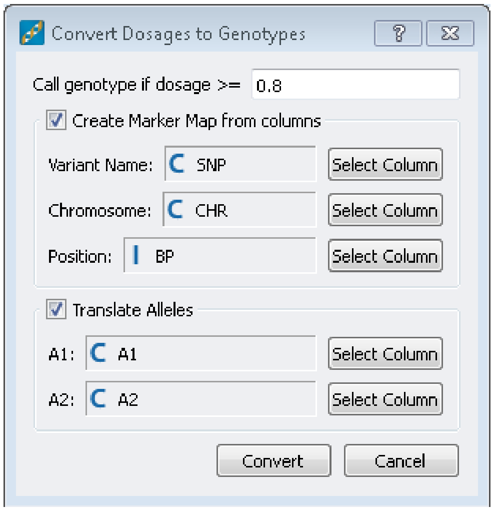
Dosage Format: Double-Dosage format

Samples are found in ○ Row Labels
● Column Headers

Numeric column types to convert:
☑ Real columns (6)
☐ Integer columns (1)

d. In the next prompt, specify the minimum dosage value for a called genotype.
e. This spreadsheet contains marker map information so check Create Marker Map from columns and choose the appropriate columns (these may be auto-detected and checked in your own spreadsheet).

f. This spreadsheet also contains allele translation columns for A and B. Check **Translate Alleles** and choose the appropriate columns.



Convert Dosages to Genotypes

Call genotype if dosage >= 0.8

☑ Create Marker Map from columns

Variant Name: C SNP

Chromosome: C CHR

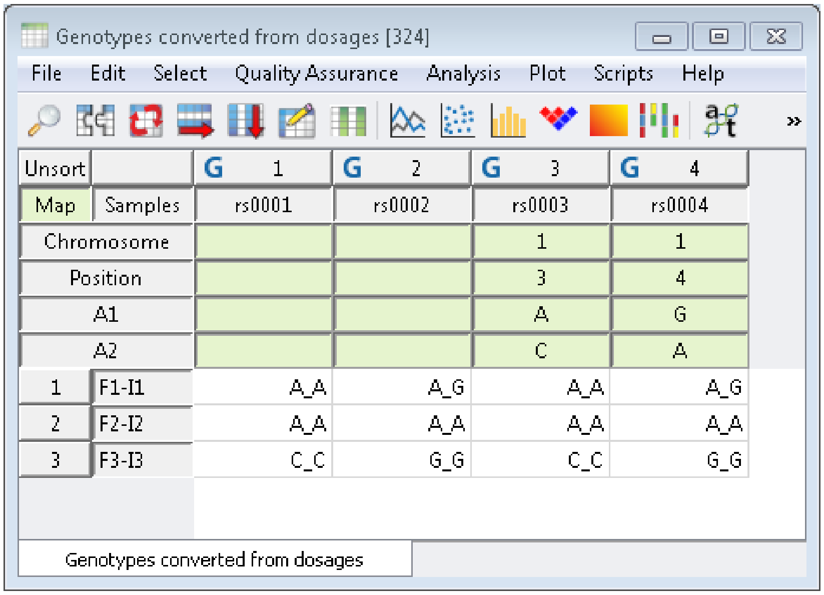
Position: I BP

☑ Translate Alleles

A1: C A1

A2: C A2

g. The resulting marker-mapped spreadsheet contains the converted genotypes (with translated alleles) and is a child of the original spreadsheet.



Genotypes converted from dosages [324]

Unsort		G 1	G 2	G 3	G 4
Map	Samples	rs0001	rs0002	rs0003	rs0004
Chromosome				1	1
Position				3	4
A1				A	G
A2				C	A
1	F1-I1	A_A	A_G	A_A	A_G
2	F2-I2	A_A	A_A	A_A	A_A
3	F3-I3	C_C	G_G	C_C	G_G

Genotypes converted from dosages