SNP-SCALE Excel Macro: Read Me File

Michael Mann, Gavin Hinten, and Jon Slate. University of Sheffield, 2006

LICENSE AND DISTRIBUTION

The SNP-SCALE excel macro is freeware and can be used by anyone free of charge. Re-distribution without the authors’ permission is forbidden.

WARRANTIES

The program comes with no warranties whatsoever.

CITATION

If you use the SNP-SCALE excel macro please cite it as:


BACKGROUND

The SNP-SCALE excel macro has been designed and tested on data obtained from SNP markers run on an ABI 3730 automated sequencer, with allele fluorescence intensities (peak heights) obtained using GeneMapper software. However the macro should work equally well for any equivalent system that quantifies peak fluorescence intensities. The macro is designed to assist with semi-automated analysis of genotype data obtained using the SNP-SCALE genotyping system (Hinten et al., in review). The macro produces a “Data” sheet summarising your data, constructs three cluster plots (Cartesian, polar, and peak intensity) from your data, and automatically assigns genotypes to each sample. The limits of each cluster group, as well as failed samples, can be manually adjusted for the polar cluster plot and will apply to all three cluster graphs. Output from the peak intensity plot can be formatted and used as input data for the ClusterA program (Lovmar et al., 2005; http://www.medsci.uu.se/molmed/snpgenotyping/software.htm). ClusterA (can be used to calculate silhouette scores, which provide an objective measure of cluster quality. Lovmar et al. (2005), recommend that silhouette scores >0.65 indicate high quality genotype assignment.
**INPUT**

In Sheet one your data should be arranged in three columns, with column titles on row one.

Column 1: **Sample Name**
Column 2: **Fam Ht.** The peak height of the Fam-labelled allele (x-allele) in Relative Fluorescence Units (RFUs)
Column 3: **Hex Ht.** The peak height of the Hex-labelled allele (y-allele) in RFUs.

The sheet also contains three limit values for producing polar plots and automatically assigning genotypes. These limits can be manually adjusted by the user to set boundaries for genotype clusters. The limits are:

Limit 1: **The Distance score** (see *Chart 2: Polar Cluster Plots*) below which genotypes are not scored (Fail) (Default: $\log_{10}$ RFU=2).
Limit 2: **The angle above** which genotypes are assigned as homozygous for allele 2 (Hex) (Default: 1.4 radians).
Limit 3: **The angle below** which genotypes are assigned as homozygous for allele 1 (Fam) (Default: 0.3 radians)

Individuals with a genotype with angle between limit 2 (1.4) and limit 3 (0.3) are assigned as heterozygotes.

**RUNNING THE MACRO**

Place the cursor in cell A1, then go to Tools > Macro > Macros > Process Data and click “Run”.

The Macro creates four new sheets: a Data sheet and three cluster graphs. If you wish to re-run the macro, for example after removing ambiguous samples or adjusting polar plot cluster boundaries, you must first delete the Data sheet and the three graph sheets before rerunning the macro, or it will crash.

*** DATA ****

This sheet contains 9 columns, sorted according to genotype calls.

Column 1: The **name** of each sample

Column 2: **The peak height** of the Fam allele (in RFUs)

Column 3: **The peak height** of the Hex allele (in RFUs)

Columns 4 & 5: The **polar co-ordinates** of each sample, with distance in column 4 and angle in column 5.

Columns 6 & 7 are **allele calls** where 0 is unscored and 1 and 2 are the two alternative alleles (1=Fam, 2=Hex).

Columns 8&9: The **peak intensity** co-ordinates of each sample, with **relative peak height** in column 8 and **absolute peak height** in column 9.
*** Chart 1: Cartesian Cluster Plot ****

A cartesian (x,y) plot of each genotype where Fam peak heights (in RFUs) are on the x-axis and Hex peak heights are on the Y axis.
Genotype assignments are indicated for each individual.
Note how each genotype forms a discrete cluster, but clusters converge on the origin

*** Chart 2: Polar Cluster Plot ***

A polar plot of the Fam and Hex peak heights, expressed as (Log$_{10}$D,r):

where D (distance of the Cartesian (x,y) co-ordinate from the origin) is the square root of ($x^2 + y^2$), and r (angle of the Cartesian (x,y) co-ordinate with respect to the x-axis) is the arctangent of ($y/x$), measured in radians.

Polar plots should produce three horizontal clusters around $y = 0$ (Fam homozygote), $y = \pi/4$ (heterozygote), and $y = \pi/2$ (Hex homozygote) radians. Genotypes should form discrete clusters that do not converge. Failed samples are indicated by a red cross.

*** Chart 3: Peak Intensity Cluster Plot ***

A peak intensity plot of Fam and Hex peaks, expressed as (PH$_{R}$, PH$_{A}$):

where $PH_R$ = Fam Peak Height/ (Fam Peak Height + Hex Peak Height)
$PH_A$ = Log$_{10}$(Fam Peak Height + Hex Peak Height)

Peak intensity plots should produce three vertical clusters around $x = 0$ (Hex homozygote), $x = 0.5$ (Heterozygote), and $x = 1$ (Fam Homozygote). Genotypes form discrete clusters that do not converge. Failed samples are indicated by a red cross.

CHECKING AND ADJUSTING POLAR PLOT CLUSTER LIMITS

We recommend that the macro is first run using the default values for cluster limits. The user can then manually inspect the polar cluster plot and decide whether these limits need adjusting, or if outlier samples should be removed and the macro re-run. Before re-running the macro, you must first delete the Data sheet and the three charts, or the macro will crash. The macro can then be re-run as above with the modified parameters.

PREPARING DATA FOR INPUT INTO THE ClusterA PROGRAM

The output data for either the polar cluster plot or the peak intensity cluster plot can be formatted and inputted into the ClusterA program (Lovmar et al., 2005) to calculate silhouette scores, which provide an objective assessment of genotype quality. For either cluster plot we recommend you use the one-dimensional analysis method because the distance score (D) for the polar plot and the absolute peak
intensity score \((PH_{i})\) for the peak intensity plot are used simply to identify and remove failed or weakly amplifying samples. They provide no genotype information.

The one-dimensional input file for the ClusterA program requires three columns of data with no column titles, formatted as TAB delimited text. The input file should be prepared in a separate worksheet. Column one contains the name of the dataset; column two contains the name of the cluster each sample belongs to; and column three contains the datapoint information for that sample (sample angle for the polar cluster plot, or sample relative peak intensity for the peak intensity cluster plot).

REFERENCES