**Feature Selection****:** A combined background and variance filter was applied to the data matrix to identify the most variable genes using an in-house developed feature selection program. Firstly, a background filter was applied to remove genes with expression values too low to be distinguished from the background noise. A high threshold was used to remove a large number of probe sets and ensure these probe sets are highly expressed (threshold: α=0.001).

Background expression filter:

Expression values are defined as follows: *E*=median(log2 (*I*)) where:

E = Expression

*I* = Intensity

Expressions values must be above the user defined significance level α.

E>log2(*z*α σ*Bg*)

E = Expression

σBg = standard deviation of the background

*z*α= Quantile of the standard normal distribution (corresponding to α value).

α = Significance level

Background variance filter:

Expressions values must be above the threshold based on σBg.

Log2(varE) >2[log2(σBg) – *E –* log2 (log(2))]

varE = Expression variance

σBg = standard deviation of the background

E = Expression

Secondly, an intensity dependant variance filter was applied to the data matrix to remove probe sets with low variance across all samples (threshold: α =1.28 x 10-12).

This method filters variance to be above the expression dependent geometrical mean variance by a value specified by significance parameter α as follows:





varE = Expression variance

*z*α = Quantile of the standard normal distribution (corresponding to  value).

σ(varE) = standard deviation of the expression variance

*b*1 and *b*2 = coefficients of linear regression of log2(var*E*)

Expressions values must be above the user defined significance level α.

***Results – Feature selection***

Feature selection resulted in 1462 most variable probe sets.