Motivation:

Some existing filtering methods such as A/P, I/NI (FARMS), can’t be applied on every type of data, specially when PM/MM are not provided or when there are no multiple probes per gene: Agilent microarrays, RNA sequencing data, Affymetrix series matrix (CEL files not provided).

Data sets:

Lung data set (GSE8894) series matrix

Platinium spike (GSE21344) CEL files

Diabetis data (GSE5606) from CEL files

Ovarian tumors (GSE4122) from series matrix

Breast tumors (E-TABM-1389) on Agilent 244K-exonV3

CGH profiling of xenographs (E-TABM-992) on Agilent 244A

Gliomas (E-TABM-1107) on Agilent 4x44K + CGH.

Chiaretti ALL (Bioconductor)

TCGA RNASeq

Method

In order to locate the non-informative signal, both non-informative and informative probes are added as follow:

Non-informative probes can simply be added by randomly selecting row sand permuting all their values.

To add significant pseudo-probes, first and to mimic the real signal behavior, mean and variance were computed for each probe in order to generate M and V vectors, respectively.

Then, M was cut into intervals, *M1* to *Mn*, each of range 0.2, and each associated with its corresponding interval of variances *V1* to *Vn* such that:



Each added probe *g* can be defined as a mixture of 2 normal distributions, each with means µ1, µ2 randomly chosen in {*M1,…,Mn*}, and variances 1, 2 randomly chosen in the corresponding Vi.

To obtain a significant fold change, at least for some probes, µ1 only is chosen, and µ2 is computed as:



where *xg* is an increased/decreased coefficient assigned to gene *g* and randomly chosen between -0.6 and 1.5. The 2 marginal functions were then defined as:



To obtain a more realistic design, the signal of each pseudo-probe g was modified not for all the samples in group *k*, but only for a random proportion of them, chosen between 0.5 and 0.8.



Finally, Yg the signal of the pseudo-probe *g* is defined as:

