Supplementary information: Artificial data

We constructed two artificial data sets starting from the expression measurements from the first 38 patients with acute leukemia (Golub et al., 1999). Golub et al. called this set of patients the training set (27 ALL and 11 AML).

RANDOMISED DATA

To destroy any true correlation of the gene expression profiles with the ALL-AML class distinction, we randomly and independently permuted the components of each gene expression vector in the training set from Golub et al., resulting in a data set expected not to contain genes with an actual correlation with the two conditions. After analysis, one can see in Figure 1 - as expected - that V_i reaches its constant level of approximately zero (so $N_1 \approx 0$, the null hypothesis is true for all the genes, so no genes are actually differentially expressed) starting from the first gene (t = 1), confirming that this data does not contain genes that, individually, contain real information about the difference between ALL and AML.

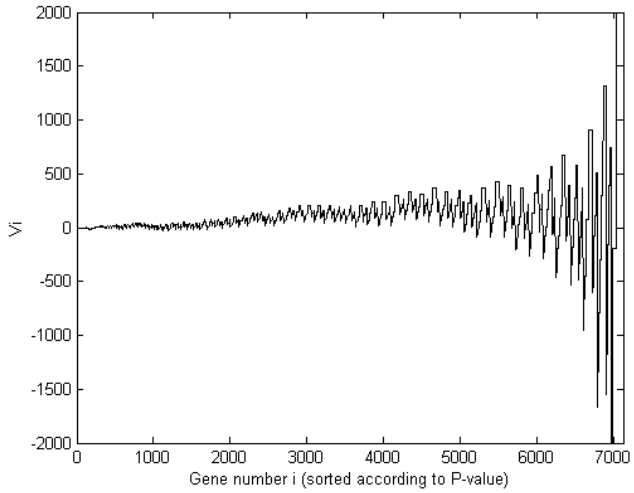

Fig 1. Plot of V_i versus the gene number i for the randomised training set of Golub et al. The constant level of V_i is approximately zero.

SIMULATED DATA

To construct an artificial data set we arbitrarily selected the gene expression profile from the non-randomised training set that, after sorting according to the p-values, was on the 1000th place (= g_{1000}). This gene had a p-value of 0.015 and therefore can, *on its own*, be considered as differentially expressed between ALL and AML. Consequently, we superposed noise to the components of this expression profile drawn from a uniform distribution in the range of $[-\sigma/4,\sigma/4]$, where σ was the standard deviation of the components of g_{1000} ($\sigma = 396$). This was repeated 1000 times and resulted in 1000 expression profiles (with p-value ranging from 0.00079 to 0.38), which are, by design, not accidentally correlated with the class distinction ALL-AML (for these genes the status of the null hypothesis is false and therefore are considered to be actually differentially expressed). Finally, we added these 1000 expression profiles to the 7129 profiles from the randomised training set (see previous section - for these genes the status of the null hypothesis is true), resulting in a data set with known values of $N = 8129$, $N_1 = 1000$ and $N_0 = 7129$. Note that the distribution of p-values in this data set (+/- uniformly distributed between 0 and 1 with a peak in the lower values) was similar to the distribution of p-values in all the real data sets we studied.

The results of the overall significance analysis can be inspected in Figure 2 (left hand side). V_i reaches a constant level of about 1009 (mean of V_i for i between 1800 and 3000, which is the estimated value for N_1) at the 1800th gene. Since, by design, we know the status of the null hypothesis for each gene (and therefore we know exactly which genes are actually differentially expressed and not), we can calculate the real value of the FDR and sensitivity at each level of rejection and compare this with the estimated FDR and sensitivity by our method. This is done in Figure 2 (right hand side). Note that the difference is minimal.

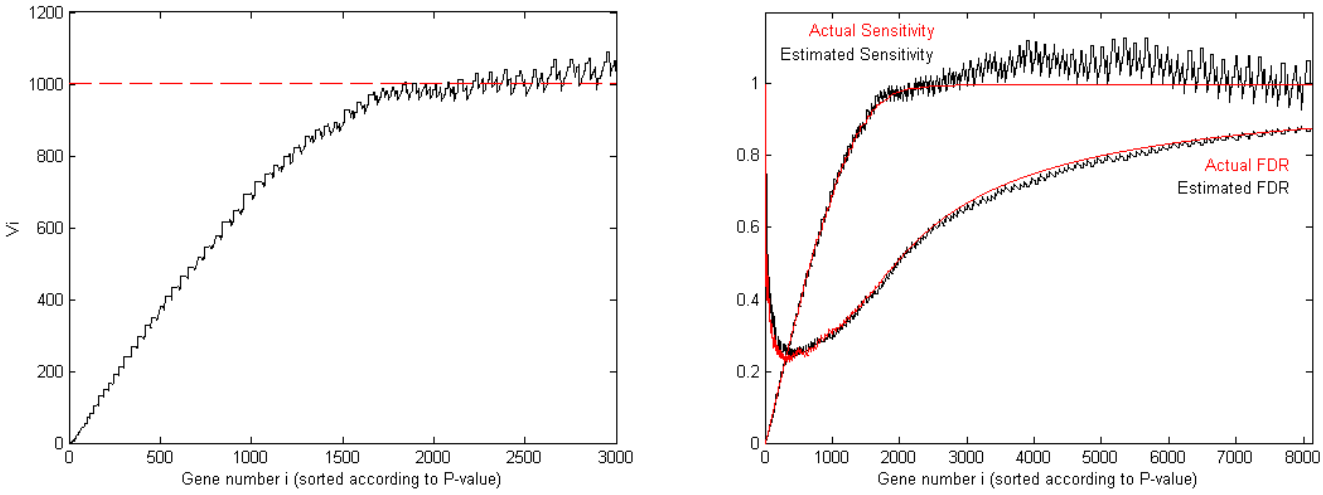

Fig 2. Analysis of the simulated data with known values for $N_1 = 1000$ and $N_0 = 7129$ **Left.** Plot of V_i versus the gene number i **Right.** Comparison between the real values for the FDR and sensitivity and the calculated or estimated values for the FDR and sensitivity. Note that, in this case, the calculated sensitivity, which is an estimate of the real sensitivity, does not always stay below is theoretical limit of one.

References

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.