Supplementary Materials for Newton et al.

- Supplementary Tables 1-3
- Supplementary Figures 1-6

Supplementary Tables

Represent

Gene	1	2	3	⋯			G	
Selected	0	1	0	⋯	1	1	⋯	1	n
In category	1	1	1	⋯	1	0	⋯	0	m
In both		1			1				x

	Selected: yes	Selected: no	
C	x	$m-x$	m
not C	$n-x$	$G-n-m+x$	$G-m$
	n	$G-n$	G

Permute

Permuted	1	0	1	⋯	0	0	⋯	0	n
In category	1	1	1	⋯	1	0	⋯	0	m
In both	1		1						X

implies $X \sim$ Hypergeometric

Generalize

Gene score	s_1	s_2	s_3		⋯			s_G	
In category	0	1	0	⋯	1	0	⋯	0	m
Combined		s_2			s_g				x

permuting, $X/m \sim (\mu, \sigma^2)$

Table 1: [Supplementary Material] Random sets. The standard approach of Fisher's exact test involves cross classifying G genes by their presence in a functional category C (size m) and in a selected list of altered genes (size n). This is compactly displayed in a 2×2 table (upper right), though it is derived from a more extensive binary table (upper left) containing a row for each classifying factor and a column for each gene. A gene on both lists has 1's in both rows of its column; a third, summary row based on products indicates this feature. The hypergeometric distribution for X, the number of genes on both lists, corresponds to permuting one row and recomputing the summary row of the big table (middle left), which is equivalent to sampling a random set of m columns. Generalizing this concept (lower left), we can use quantitative gene-level scores in place of the selection indicators. The induced distribution on X is intractable but has a computable mean and variance.

Gene score	s_1	s_2	s_3			$\cdots$			s_G	
In C_1	0	1	0	$\cdots$	1	0	$\cdots$		0	m_1
In C_2	1	1	1	$\cdots$	0	0	$\cdots$		1	m_2
In C_3	1	0	0	$\cdots$	1	1	$\cdots$		0	m_3

Table 2: [Supplementary material] Multiple categories. A network of categories may be represented in an extended table, as indicated here, with a row for gene scores and then additional binary rows indicating membership of genes in each category. With all the lower rows fixed (to retain the overlap structure in the network), one randomizes by permuting the top row of gene scores, as in Table 1.

ID	Term	m	$\bar{X}$	Z	$Z/\sqrt{m}$
GO:0001772	immunological synapse	59	3.88	8.46	1.10
GO:0008332	low voltage-gated calcium channel activity	11	4.54	4.40	1.33
GO:0008113	protein-methionine-S-oxide reductase activity	10	4.45	4.11	1.30
GO:0004423	iduronate-2-sulfatase activity	12	4.10	4.08	1.18
GO:0019883	antigen presentation, endogenous antigen	48	3.90	7.68	1.11
GO:0000185	activation of MAPKKK activity	15	3.92	4.32	1.12
GO:0005031	tumor necrosis factor receptor activity	18	3.91	4.71	1.11
GO:0019911	structural constituent of myelin sheath	11	3.92	3.70	1.12
GO:0019903	protein phosphatase binding	11	3.88	3.65	1.10
GO:0003840	gamma-glutamyltransferase activity	13	4.18	4.34	1.20
GO:0042101	T cell receptor complex	11	4.55	4.42	1.33

Table 3: [Supplementary material] Continuing with the NPC example, shown are categories that are significant at the 5% level according to the *maxT* procedure, based on the enrichment score $Z/\sqrt{m}$ and using the average transformed Spearman correlation as the category statistic.

Supplementary Figures

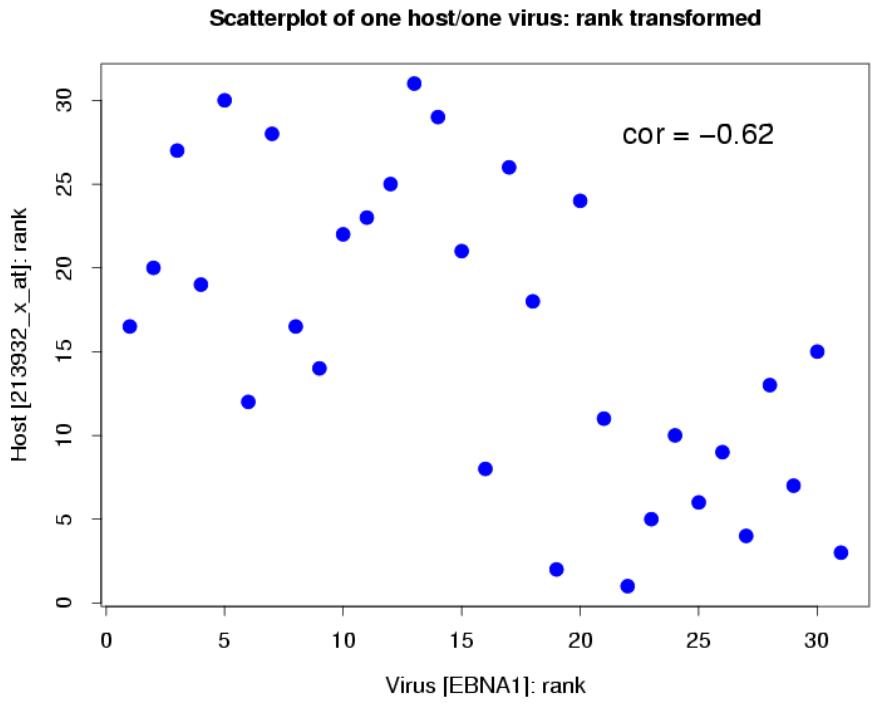

Figure 1: [Supplementary Material] Scatterplot of one host gene expression against the expression of the viral gene EBNA1, NPC example

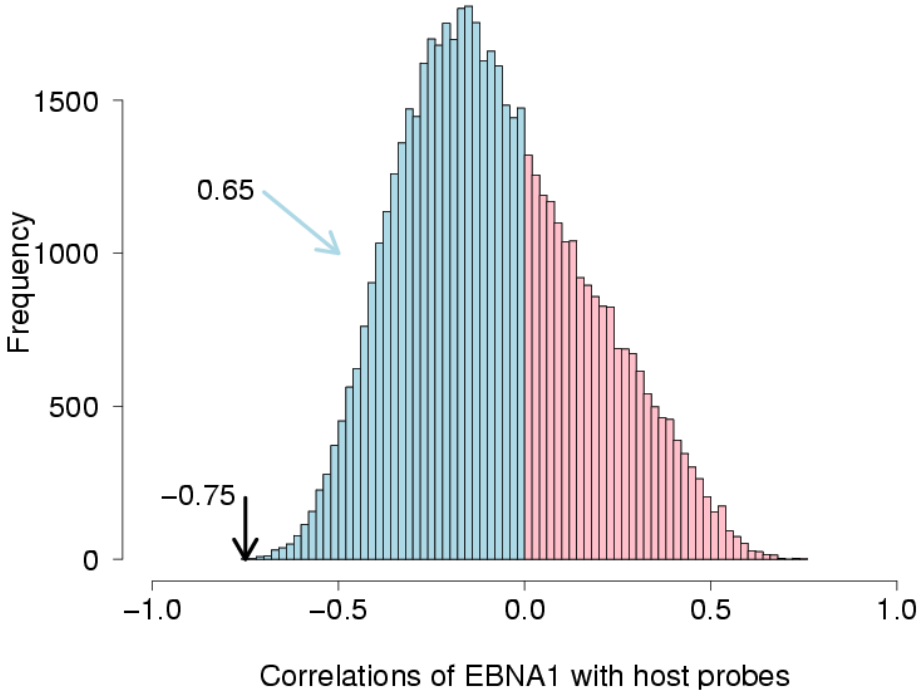

Figure 2: [Supplementary Material] Histogram showing all 54,675 host probe set Spearman correlations with EBNA1. Proportion negative and minimal correlation are highlighted, and shown to be significant if there is truly no host/virus association (next Figure). [This is a reproduction of Figure 3C from Sengupta *et al.* (2006), included here for easy reference.]

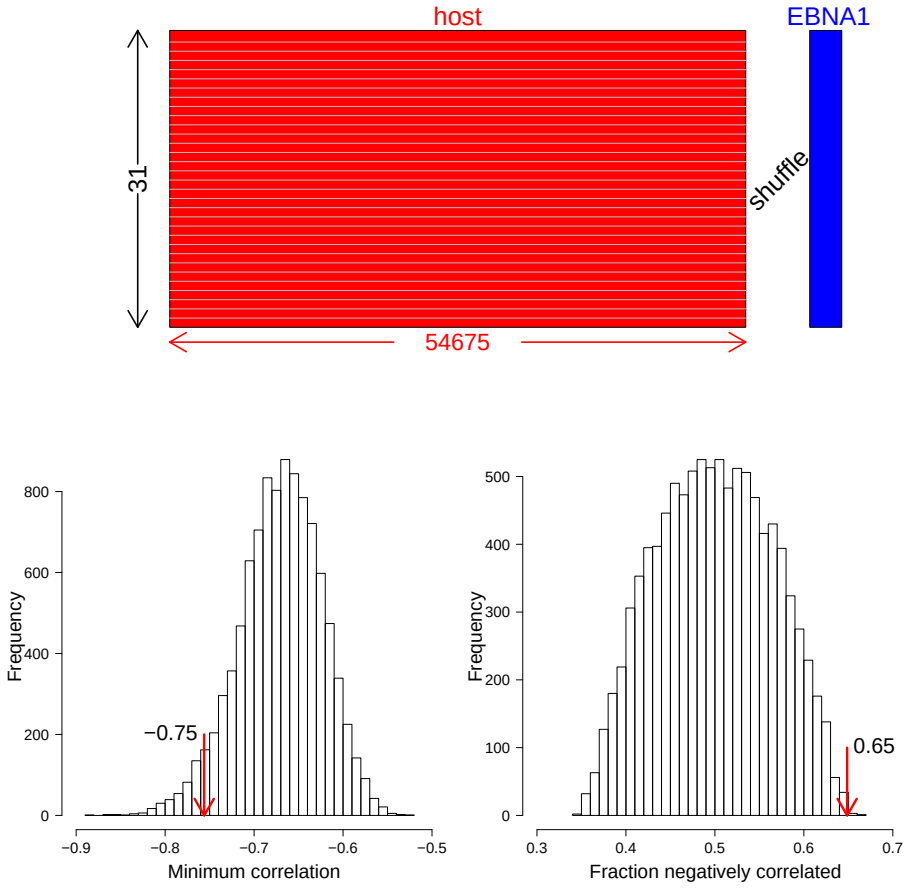

Figure 3: [Supplementary Material] Permutation analysis (10,000 replications randomly reassigning EBNA1 values to host microarrays, top) shows that the global association features from previous figure are significant: Minimum correlation p-value $= 0.04$; proportion negatively correlated p-value $= 6 \times 10^{-4}$.

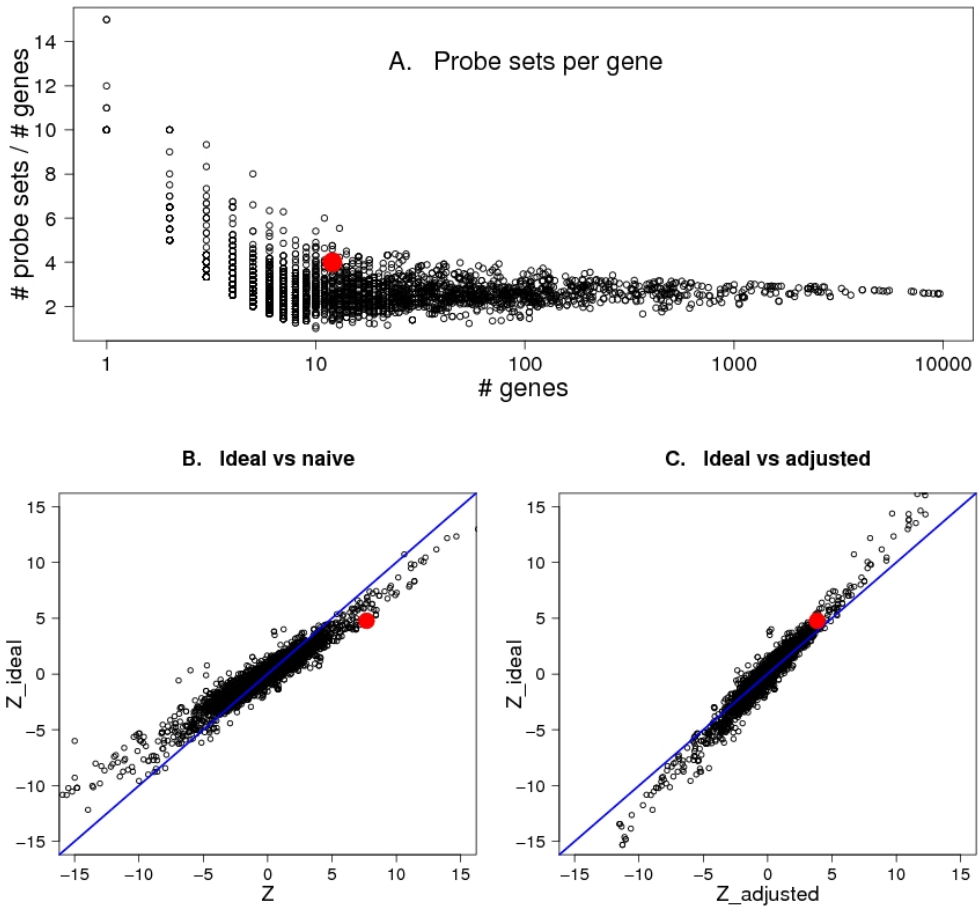

Figure 4: [Supplementary Material] On the number of probe sets per gene. Panel A shows the ratio of probe sets per gene on the Affymetrix hgu133plus2 microarray as this ratio is related to 2761 GO categories containing $m_p \geq 10$ probe sets. Various modified Z scores are considered; for illustration all are based on transformed Spearman correlations as the gene level score s_g. The naive approach is to ignore the probe set to gene ratio, giving Z. Ideally, we would work at the level of the genes themselves, by first collapsing the probe set data to the gene level, and then applying the enrichment machinery. This is computationally more challenging (we summarized by median), and gives Z_{ideal}. A much simpler adjustment uses $Z_{\text{adjust}} = Z\sqrt{m_g/m_p}\sqrt{(G-m_p)/(G-m_g)}$ to accommodate that the number of genes m_g in the category differs from the number of probes m_p. Note m_p was used in standardizing the category score in the first place. GO:0019883 is marked in red.

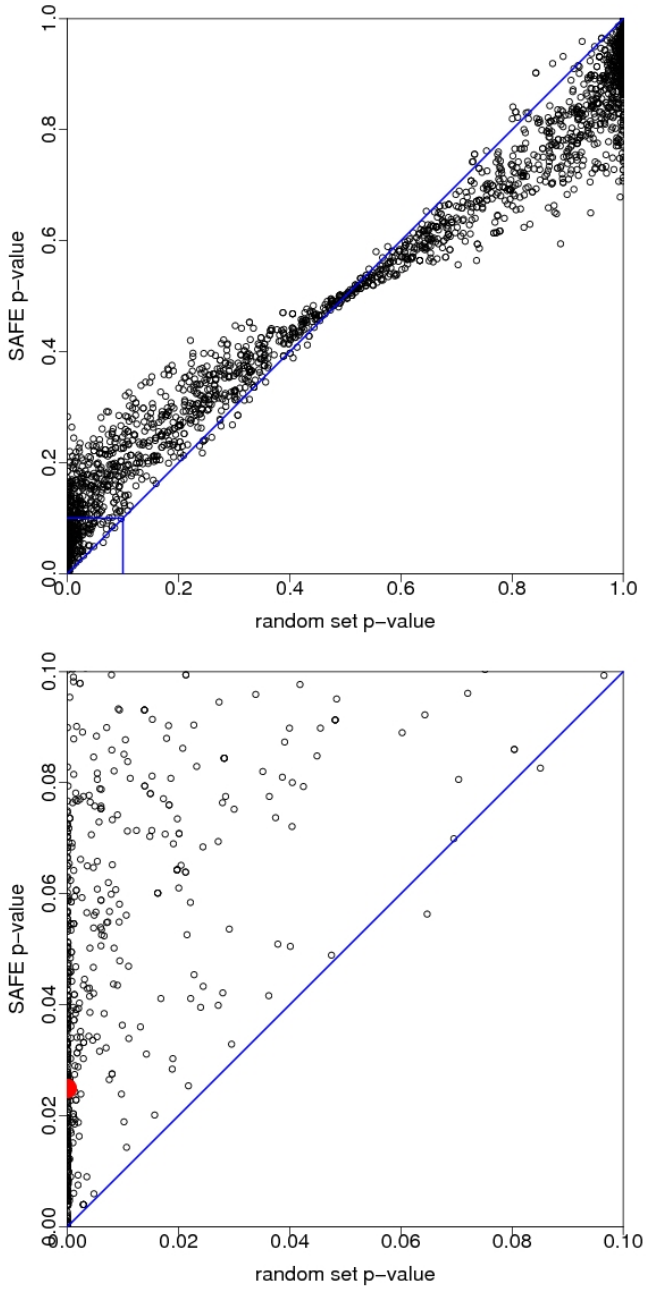

Figure 5: [Supplementary Material] Comparison with SAFE/GSEA: SAFE was applied to all 2761 GO categories that contain $m \geq 10$ probe sets, and resulting $p-$values are plotted against the $p-$values obtained by converting random-set Z scores into normal probabilities. In both cases the category statistic is the mean rank of probe-set statistics, so the differences are attributable only to the calibration used. The lower panel highlights small $p-$values from the upper panel, and shows the results for GO:0019883 in red.

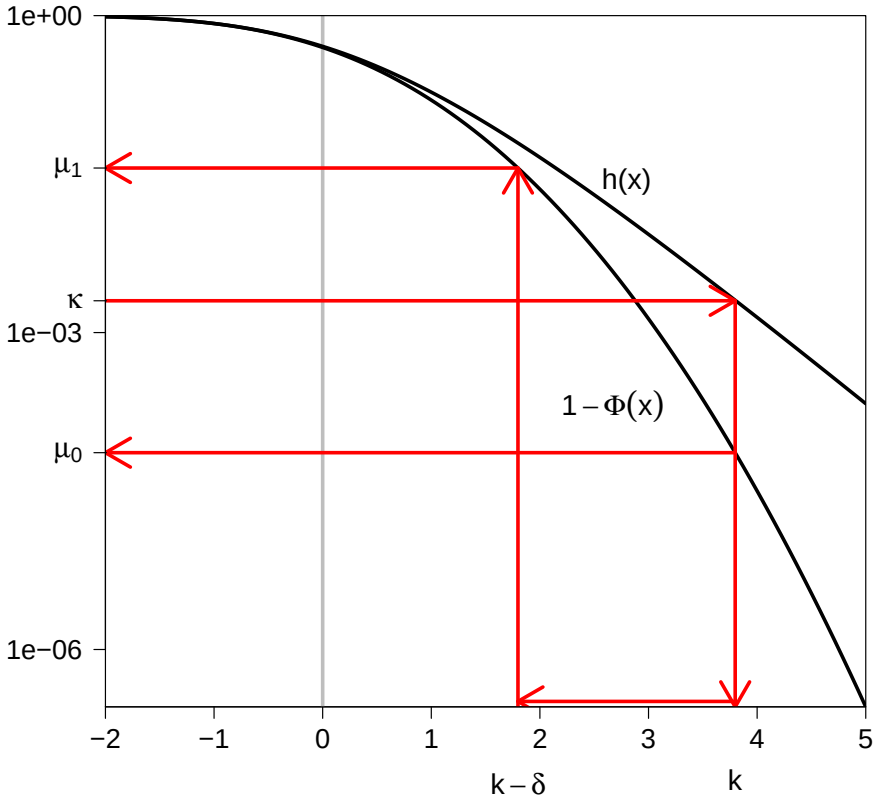

Figure 6: [Supplementary Material] Theoretical components. Given a gene effect $\delta > 0$, the function $h(x) = [1 - \Phi(x - \delta)] \,/\, [1 - \Phi(x)]$ is key in finding the threshhold for an FDR controled selection method. Starting at κ on the vertical axis, we map back to k where $h(k) = \kappa$. Then we get mean values μ_0 and μ_1 from these threshholds.