Supplementary Information

Microarray proteomic analysis discriminate tumorigenic mouse ovarian surface epithelial cells of divergent aggressive potential.

Ulises Urzúa[1], Lionel Best [2] and David J. Munroe[2].

[1] Laboratorio de Genómica Aplicada, ICBM-Facultad de Medicina, Universidad de Chile, Independencia 1027, Santiago, Chile; [2]Laboratory of Molecular Technology, SAIC-Frederick, Inc., NCI at Frederick, 915 Tollhouse Avenue, Suite 211, Frederick, MD 21701.

Design of antibody microarray experiments

A "common reference" experimental design was used in this work (Supplementary Figure 1). Test protein samples were pooled from 3 independent 90% confluent cultures while reference sample was a protein extract obtained from a single whole newborn male mouse. Repeated dye swap assays were done for the two pooled protein samples extracted from MOSE cells displaying extreme aggressiveness *in vivo*. The advantages and limitations of sample pooling in proteomics studies are currently debatable. A recent report by Diz et al. [1] concludes that the protein level in a pool should equal the mean level of biological replicates or individuals making up the pool. The measured biological variance in a pooled sample decreases as a function of 1/r, where r is the number of replicates contributing equally to the pool. Reduced biological variance may thus result in a reduction of statistical power [2]. In transcriptional profiling studies, biological averaging may not hold especially for very small designs and biological replicates are advantageous over technical replicates [3]. Karp and Lilley also assessed the pooling problem on mouse and human brain proteomic data obtained with DIGE [4]. These authors concluded that biological averaging by means of pooling is valid for mouse brain tissue without any bias but the same does not hold for human brain tissue. Moreover, a power of 0.95 with 0.01 confidence in a 2-class test in mouse can be achieved with 11 gels comparing 5 vs 6 (or 6 vs 5) pools, each pool composed of 3 individuals. In contrast, for the same pool size, 13 vs 13 pools are required to attain equivalent statistical threshold but running over twice assays (26 gels) in humans [4]. Thus, the source of samples to be pooled must also be considered. It is reasonable that variability among human tissue biopsies can be higher than that of inbred mice or parallel cultures of an established cell line. Often, pooling constitutes a reliable approach when cost and/or sample amount impose limitations to research.

Data analysis workflow

The sequence of steps for data processing and mining is summarized in both Supplementary Figure 2 and Supplementary Table 1. Step 1 was lowess normalization performed with the DNMAD tool, which uses the R language and is based on the BioConductor package Limma with modifications [5]. Normalization minimizes technical variations in protein expression levels of the two samples co-incubated on the microarray, so that actual biological differences can be identified. The original algorithm used in DNMAD was described by Yang et al [6] and corrects intensity and spatially dependent bias within a single slide. In addition, adjusts the scales of percentile $\log_2$ ratio distributions across multiple slides. Based on $\log_2$ intensity ratios (M) and on the average $\log_2$ intensity in both channels (A), the method applies a robust local regression according the print-tips of a microarray thus computing a different normalization factor for each spot on the microarray. Since the Panorama microarrays are not whole genome, some of the print-tip loess assumptions were not met [5], and thus global loess was used in this work. The second data processing step was the removal of inconsistent dye-swap ratios. A simple logical step was applied to normalized data in an Excel spreadsheet (see Supplementary Table 1). To date, the dye-bias issue has been poorly addressed in 2-colors microarray-format proteomic studies. Dye swapping is essential to counteract gene-specific dye bias in DNA microarray experiments [7]. For DIGE experiments, an offset/scale normalization of spatial background variation is needed particularly for low intensity spots [8]. In general, removal of any systematic bias is recommended prior to statistical tests. The next step in data processing was done again in an Excel spreadsheet and consisted of the transposition of repeated data within a single slide to generate an additional column, i.e. 2 values in 2-rows per protein were resolved to 4 values in a single row per protein. This step generated a 157-protein dataset that was subjected to statistical 2-class analysis.

Three tests were applied to generate a consensus list of differentially expressed proteins. These tests were performed with the T-rex tool, part of the GEPAS server (go to http://www.gepas.org/, mouse over “Tools” and click on “Differential expression”) [9]. The t-test calculates a conventional *p*-value, as well as FWER (family wise error rate) and FDR (false discovery rate) based adjusted *p*-values to cope with the multiple comparisons issue. A *q*-value that describes the minimum FDR at which the test may be assumed significant is also calculated by the 2-classes

tests in T-rex. Regarding our data, those protein levels statistically significant judged by a FDR based $p \leq 0.05$ in at least 2 tests (see Figure 2) were subsequently analyzed functionally (see Table 1 and Figure 3). Thirty-eight antibodies corresponding to 31 unique proteins met this criterion. Of these, 11 proteins were common to the 3 tests. The SAM (Significance analysis of microarrays) test produced the longest list. This algorithm assigns a score to each gene value based on its change relative to the standard deviation of repeated measurements for that value. If the gene score is higher than a fixed threshold, then it is significant [10]. Regarding multiple test control, SAM actually does not calculate a FWER but estimates a particular FDR by permutations assuming that all null hypotheses are true. Therefore, SAM usually performs less stringent tests than other methods including Bonferroni and the step-down correction by Westfall and Young [10]. Accordingly, in the present study SAM identified 21 unique proteins that not appeared in the other 2 analyses. In contrast, no unique proteins arose exclusively from the Clear or t-test (see Figure 2A). The Clear method combines two parallel tests: a fold change, z-test that uses a single estimated variance pooled across all genes, and a χ^2 test to address variability [11]. This program generates a graphical display describing four qualitative categories of statistically significant proteins, namely differentially and not differentially expressed, each with high or low variability. The Clear test represents an improvement of the t-test since bypasses small fold changes with low variance [11]. The third statistical test applied was a t-test, which is a standard method to compare the means between 2 classes. Supplementary Figure 3 shows the graphical output of this test from T-rex. T-statistics and FDR based *p* values are shown to the left side of color-coded standardized values. Finally, the 31 differentially expressed proteins were subjected to functional interpretation using the tools DAVID, WebGestalt and SNOW [12-14].

Supplementary references

[1] Diz AP, Truebano M, Skibinski DO. The consequences of sample pooling in proteomics: an empirical study. Electrophoresis. 2009 Sep;30(17):2967-75.

[2] Karp NA, Spencer M, Lindsay H, O'Dell K, Lilley KS. Impact of replicate types on proteomic expression analysis. J Proteome Res. 2005 Sep-Oct;4(5):1867-71.

[3] Kendziorski C, Irizarry RA, Chen KS, Haag JD, Gould MN. On the utility of pooling biological samples in microarray experiments. Proc Natl Acad Sci U S A. 2005 Mar 22;102(12): 4252-7.

[4] Karp NA, Lilley KS. Investigating sample pooling strategies for DIGE experiments to address biological variability. Proteomics 2009 Jan;9(2):388-97.

[5] Vaquerizas JM, Dopazo J, Díaz-Uriarte R. DNMAD: web-based diagnosis and normalization for microarray data. Bioinformatics. 2004 Dec 12;20(18):3656-8.

[6] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002 Feb 15;30(4):e15.

[7] Martin-Magniette ML, Aubert J, Cabannes E, Daudin JJ. Evaluation of the gene-specific dye bias in cDNA microarray experiments. Bioinformatics. 2005 May 1;21(9):1995-2000.

[8] Kreil DP, Karp NA, Lilley KS. DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results. Bioinformatics. 2004 Sep 1;20(13):2026-34.

[9] Tárraga J, Medina I, Carbonell J, Huerta-Cepas J, Minguez P, Alloza E, Al-Shahrour F, Vegas-Azcárate S, Goetz S, Escobar P, Garcia-Garcia F, Conesa A, Montaner D, Dopazo J. GEPAS, a web-based tool for microarray data analysis and interpretation. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W308-14.

[10] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001 Apr 24;98(9):5116-21. Epub 2001 Apr 17. Erratum in: Proc Natl Acad Sci U S A 2001 Aug 28;98(18):10515.

[11] Valls J, Grau M, Solé X, Hernández P, Montaner D, Dopazo J, Peinado MA, Capellá G, Moreno V, Pujana MA. CLEAR-test: combining inference for differential expression and variability in microarray data analysis. J Biomed Inform. 2008 Feb;41(1):33-45.

[12] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.

[13] Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W741-8.

[14] Minguez P, Götz S, Montaner D, Al-Shahrour F, Dopazo J. SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W109-14.

Supplementary table 1 – Tools used in data processing and analysis [a]

Step (tool)[a]	Statistical method or calculation [reference]	
1) Normalization (DNMAD)	$\log_2 R/G \rightarrow \log_2 R/G - c(A) = \log_2 R/[k(A)G]$ [b]	[5]
2) Dye-swap consistency (logical test in Excel)	$\log_2 R/G > 0$, $\log_2 G/R < 0$ in dye-swap replicates	
3) Resolution of repeats (data handling in Excel)	(2 rows × 2 columns) $\log_2$R/G → (1 row × 4 columns) $\log_2$R/G	
4) Two-class differential expression (T-rex)	a) SAM: $d(i) = [\, x_1(i) - x_U(i) \,] / s(i) + s_o$ [c]	[6]
	b) Clear: combined z-test and a χ test	[7]
	c) t-test: $t = [X_1 - X_2] / [S_{X1X2}\,(2/n)^{1/2}]$ [d]	
5) Data mining and interpretation	a) DAVID: Classification and annotation. [e]	[12]
	b) WebGestalt: Classification and annotation. [f]	[13]
	c) SNOW: Protein interaction network. [g]	[14]

[a] Numbers in the sequence of data analysis steps as appear in Supplementary Figure 1.
[b] The term $c(A)$ is the lowess fit to the MA-plot. This plot represents (R,G) data, where $M = \log_2 R/G$ and $A = \log_2 \sqrt{(R \times G)}$.
[c] The term $d(i)$ corresponds to the relative difference in gene expression. The terms $x_1(i)$ and $x_U(i)$ describe the average expression levels of gene (i) in the two conditions I and U, respectively. The term $s(i)$ is the standard deviation of repeated expression measurements while s_o=3.3 is a constant to minimize the coefficient of variation.
[d] The terms X_1 and X_2 are the means in two conditions 1 and 2. S is the estimated variance and n, the number of replicates or observations per condition.
[e] Major features are fuzzy clustering, enrichment as a geometric mean of P-values, Fischer test, multiple test correction.
[f] Major features include hypergeometric and Fischer's exact test, directed acyclic GO graphs; pathways maps; multiple test correction, GRIF and Pubmed tables.
[g] The minimal connected network (MCN) is composed of 4 major elements: node (ν), rC_B (relative betweeness centrality), C (clustering coefficient of a protein list) and hubs, which are highly connected nodes.

Supplementary table 2 – Statistics of GO and KEGG analyses [a]

Cell	Term or Pathway [b]	Enrichment	p-value
IG10	Renal cell carcinoma	4.42	0.0056
	Positive regulation of transcription	3.48	0.0022
	Cell projection	3.42	0.0073
	ErbB signaling	2.73	0.0105
	Transcription regulator activity	2.62	0.0066
	Focal adhesion	2.46	0.0100
	T-cell receptor signaling	2.45	0.0586
	DNA binding	2.03	0.0493
	MAPK signaling	1.86	0.0748
IF5	Nuclear chromatin	14.22	0.0446
	DNA packaging	9.27	0.0135
	p53 transcription [b]	6.80	0.0052
	Organelle organization	2.34	0.0437
	Cell cycle	2.03	0.0691

[a] The IG10 and the IF5 genelists were analyzed with the tool WebGestalt (http://bioinfo.vanderbilt.edu/wg_gsat/) using an hypergeometric test. Terms and pathways are ranked according to their enrichment scores.

[b] Not a GO term. Corresponds to the TransFac ID 1559: V$P53_02.