Cancer Informatics

ORIGINAL RESEARCH

Cross-platform Comparison of Two Pancreatic Cancer Phenotypes

Robert B. Scharpf[1], Christine A. Iacobuzio-Donahue[1,2], Leslie Cope[1], Ingo Ruczinski[3], Elizabeth Garrett-Mayer[4], Sindhu Lakkur[2], Domenico Campagna[2] and Giovanni Parmigiani[5]

[1]Departments of Oncology, [2]Pathology, [3]Biostatistics, Johns Hopkins University, Baltimore, MD, USA. [4]Hollings Cancer Center, Medical University of South Carolina, Charleston, SC, USA. [5]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA.
Corresponding author email: rscharpf@jhsph.edu

Abstract: Model-based approaches for combining gene expression data from multiple high throughput platforms can be sensitive to technological artifacts when the number of samples in each platform is small. This paper proposes simple tools for quantifying concordance in a small study of pancreatic cancer cells lines with an emphasis on visualizations that uncover intra- and inter-platform variation. Using this approach, we identify several transcripts from the integrative analysis whose over-or under-expression in pancreatic cancer cell lines was validated by qPCR.

Keywords: microarrays, cross-platform, rank statistics, differential gene expression

Cancer Informatics 2010:9 257–264

doi: 10.4137/CIN.S5755

This article is available from http://www.la-press.com.

Introduction

The ability of high-throughput gene expression technologies to reproducibly capture differences between populations stratified by a clinical covariate, such as cancer metastasis, is difficult when the sample size is small. When increasing the sample size of a study is not possible due to limited resources, approaches that integrate information from otherwise similar studies, though possibly employing different high-throughput technologies, may be explored.

Combining data from multiple studies is often discussed in the context of its potential to increase the statistical power for detecting differentially expressed genes.[1] An additional advantage of a cross-study analysis is the potential to reduce spurious associations driven by an artifact in a single platform or study. Statistical approaches for integrating gene expression data from multiple high-throughput platforms include combining measures of statistical significance, such as *P*-values, calculated independently from the individual studies,[2] Bayesian models for the joint distribution of gene expression across studies,[3,4] and the derivation of a study-independent scale, such as posterior probabilities of differential expression,[5] to which standard single-study methods can be applied.[6] When the sample size in the individual studies is small, nonparametric approaches that use the rank of fold-changes in expression across a binary phenotype may provide additional robustness to technological artifacts and outliers.[7–9] This paper further explores the joint analysis of multiple platforms when the sample size in the individual studies is small.

The dataset used in this analysis consists of two primary (Capan2 and Panc1) and two metastatic (Capan1 and Hs766t) cell lines that were measured by 3 high-throughput technologies for gene expression: two-color cDNA arrays, Affymetrix oligonucleotide arrays, and serial analysis of gene expression (SAGE). The data was originally described elsewhere.[10] See[11] for a discussion of the technologies.[11] As late detection of pancreatic adenocarcinoma is a primary reason for its poor prognosis, the study of pancreatic cancer progression is a biologically relevant problem. Stage-specific genetic alterations can be useful markers for the more agressive phenotype.[12] The heterogeneity of the pancreatic cancer phenotype and the small number of samples motivated an approach that integrated information from the multiple technologies.

Methods

Preprocessing

Our analysis uses previously collected data from metastatic pancreatic cancer cell lines Capan1 (c1) and Hs766t (ht), and primary pancreatic cancer cell lines Capan2 (c2) and Panc1 (p1). See[10] for a description of the cell lines.[10] Inferences regarding differential expression depend crucially on appropriate pre-processing and normalization. Although we strongly prefer unprocessed data to processed data, this was not possible for the Affymetrix platform. We used Mas 5.0 normalized Affymetrix data without further processing. SAGE libraries were standardized to tags per 50,000. cDNA data was normalized by loess smoothing of M versus A scatterplots[13] without subtracting local estimates of background fluorescence.[14] Expression measures were transformed to the $\log_2$ scale and centered by the gene-specific means in each study.

Common gene set

Following normalization, the studies were merged to produce a common set of features measured in all three platforms. Note that the unit of measure for the combined analysis need not be genes. For instance, one may map probe identifiers in each platform to exons using the sequence information of the probes, and then treat exon-level measures of expression as the unit to be compared across technologies.[15] As probes in an Affymetrix probe-set may map to more than one exon, one could pre-process and normalize probe-level intensities using redefined probe-sets.[16] Alternatively, one may map features in each platform to a Unigene cluster or refSeq identifiers. For this dataset, probe-level data was not available in the Affymetrix platform and the choice for cross-referencing annotations was limited. We therefore mapped probes (or probesets) in each platform to Unigene Cluster Identifiers (build 180) using the R package MergeMaid.[17] One-to-many mappings of probes to Unigene clusters were excluded and many-to-one mappings were averaged.

Gene filtering

As SAGE can in theory detect the mRNA transcripts for any gene, we only required membership in the Affymetrix and cDNA platforms. Specifically, any gene present in Affymetrix and cDNA that was not

detected by SAGE was assigned a count of zero in SAGE. For the cDNA and Affymetrix platforms, we excluded genes with very low levels of expression in 2 or more of the cell lines to limit the influence of low abundance genes in the combined analysis. While more aggressive filtering strategies can improve measures of cross-study correlation, inter-platform discordance can arise from technological as well as biological sources of variation (eg, probes from different platforms may hybridize to different regions of a gene that is alternatively transcribed). After the above filtering, 3117 genes remain and were used in the cross-study analysis of differential expression.

Concordance

Concordance of the log-fold changes across platforms were assessed using Spearman correlation coefficients, a rank-based alternative to the Pearson correlation coefficient, and Kappa statistics. Kappa statistics regard the high-throughput platforms as different observers of gene expression and can be used to quantify inter-observer agreement using qualitative measures of differential expression. Using a quantile of the fold-change distribution in each platform, the observed fold-changes were classi-fied as under-, none-, and over-expressed in each of the possible pairwise combinations of platforms. We used qualitative categories of under-expressed, over-expressed, and not differentially expressed, yielding a 3×3 table with elements on the diagonal corresponding to the number of genes that have the same qualitative category of differential expression in the two platforms. We adopted a weighted Kappa statistic that penalizes discordance of over versus under-expression.[18]

Results

Spearman correlations of the $\log_2$-transformed intensities of cDNA and Affymetrix ranged from 0.21–0.57 (Fig. 1). Correlations of cDNA and Affymetrix intensities to standardized SAGE counts ranged from 0.0–0.32, likely reflecting the greater dissimilarity in the technologies.

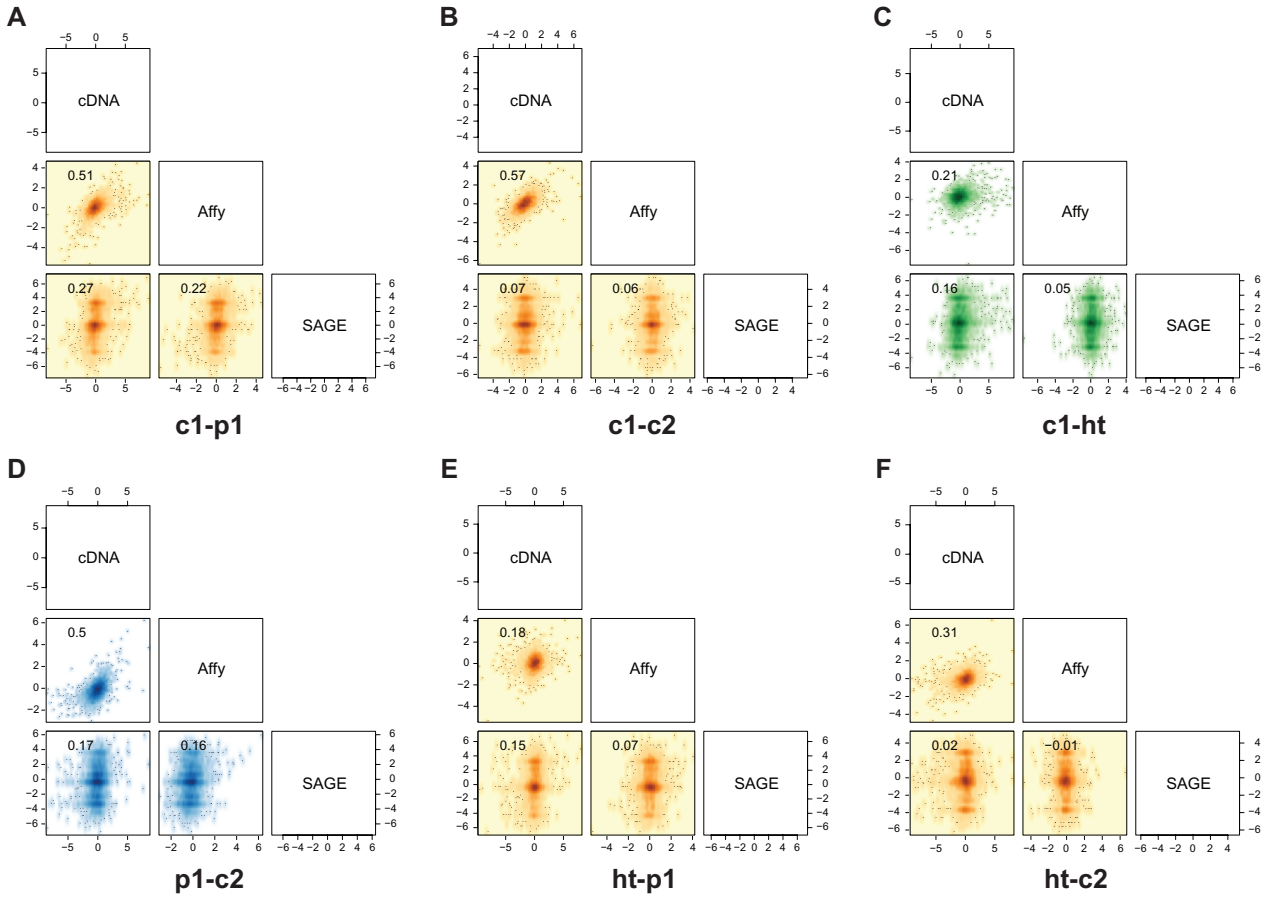

Figure 1. For each platform, we calculated all 6 pairwise combinations of fold changes in expression between the four cell lines. Panels are color-coded to indicate whether the $\log_2$ fold change was between primary-primary (blue), metastatic-primary (yellow background), or metastatic-metastatic (green) cell lines.

As a potentially more robust alternative to the Spearman correlation coefficient, we also assessed inter-platform concordance using qualitative categories of differential expression using a weighted Kappa statistic. We avoided characterizing agreement using a single arbitrary threshold by plotting the weighted Kappa over a range of thresholds based on quantiles of the fold-change distribution (Fig. 2). Kappa-statistics estimated from absolute fold-changes can be used to relax the assumption that the percentage of over- and under-expressed genes are the same, but were qualitatively similar to the Kappa plots in Figure 2 in the pancreatic cancer dataset (not shown). Again, the two intensity-based platforms (cDNA and Affymetrix) have the highest inter-platform agreement (Kappa $>$ 0.4). Together with the Spearman correlation coefficients, the small Kappa in several of the pancreatic cancer cell line comparisons reflects (i) our decision to minimize data filtering prior to the combined analysis, (ii) the absence of technological replicates for quality control measures, (iii) the biological heterogeneity of the pancreatic cancer cell lines through passage, (iv) different laboratories performing the experiments (batch effects), and (v) the non-overlapping technologies. To the extent that each platform is measuring a similar biological process, concordant findings in multiple platforms may reduce the occurrence of spurious single-study associations. With this view, we explore rank-based approaches for prioritizing a gene list and provide visualizations that make the cross-platform variability in the ranks transparent.

Rank-based approaches for cross-study analysis of differential expression in high-throughput microarray platforms have been proposed by others.[19,8] An implementation of these approaches is available in the R package RankProd. For this dataset, we ranked the average fold changes for the metastatic to primary cancer comparisons, and summarized the study-specific ranks by the arithmetic mean. By contrast, RankProd computes the geometric mean of the ranked fold changes.[19,8] An advantage of using a geometric mean (instead of an arithmetic mean) is that the ranking will be more robust to *unusual* observations. In our dataset, there are four possible pairwise comparisons within a study and a given cell line is represented in half of the possible comparisons. In the absence of a better gene-specific measure of *unusual,* the effect of the arithmetic mean is that genes with higher variance in the ranks within and across studies will tend

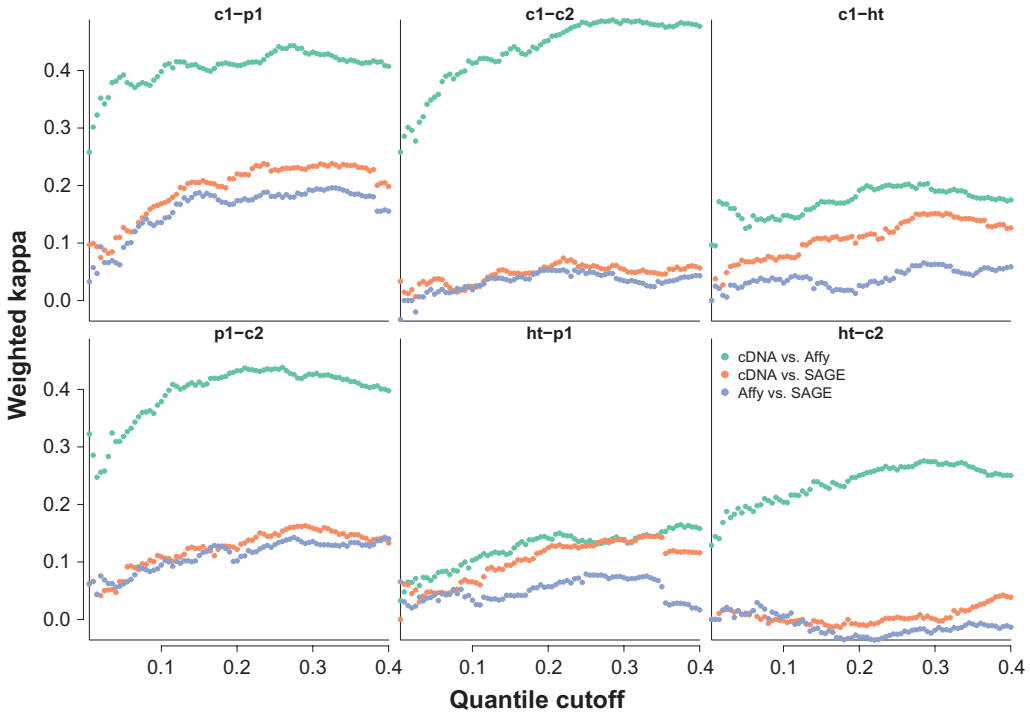

Figure 2. Genes were classified as under-expressed (fold-change $<$ q), not differentially expressed (q $<$ fold-change $\leq$ 1 $-$ q), or over-expressed (fold-change $>$ 1 $-$ q), where q is a quantile of the fold-change distribution. On the vertical axis is a weighted kappa-statistics that penalizes discordance (over-expressed in platform 1, under-expressed in platform 2).

to be positioned further down the list. See[19] for a more detailed discussion of the when an arithmetic mean may be preferable.[19] Overall, we expect that the two approaches will be qualitatively similar for the set of genes with low variance in the rankings.

While the rank of the average rank can be used to prioritize genes that show on average large fold changes in expression, such a statistic hides the variability in the ranks. We adopted the range of the ranks for each gene as a measure of spread. Of interest are genes that are consistently over (under)-expressed in metastatic relative to primary cell-lines as reflected by a small range of ranks and a large (small) average rank. In order to obtain a null distribution, we permuted the vector of ranked average fold changes in each study independently of the other studies and recomputed the range of ranks and the rank of the average ranks. Repeating the permutation a large number of times, we obtained a null distribution for the range of ranks for each rank of the average rank. Figure 3 plots the observed range (y-axis) against the rank of the average rank (x- axis) as blue points. The background is shaded by the density of the null distribution for the range of ranks, where lighter shades of gray denote more densely plotted regions of the null. Boxplots of the null distribution at the far left and far right of this plot can be useful for magnifying the lowest and highest average ranks, respectively (Fig. 4). In addition to plotting the null distribution for the range of ranks, we flagged genes for which the log fold change between the 2 primary cell lines or the 2 metastatic cell lines exceeded 3 in one or more of the platforms. Again, a motivation for the flag is that the average fold-change can hide the underlying variability from which the ranks were estimated.

We selected eight genes for qPCR validation from the top 100 over- or under-expressed using the criteria that the range of ranks was generally below the 25th percentile of the null and whose whose biological characteristics were of interest to collaborators with expert knowledge of pancreatic cancer. Among the under-expressed genes includes NME4, a membrane protein that shares homology with the putative metastasis suppressor gene NME1,[20] and TALDO1, an enzyme that helps protect cellular integrity from oxygen intermediates. Included in the over-expressed list are NSDHL, a protein involved in cholesterol synthesis, ATAD2, and CEACAM5. ATAD2 and CEACAM5 are both known to be up-regulated in cancer cells.[21,22] Figure 5 plots the fold changes measured by qPCR as a fourth platform together with the high-throughput fold changes. The direction of the average qPCR fold change (up-regulated or down-regulated) is consistent with the high-throughput platforms for 8/10 of the genes validated by qPCR, or 35/40 of the pairwise combinations. While the $\log_2$ fold-changes for Affymetrix and cDNA are often uncorrelated with SAGE, the direction of differential expression in SAGE is generally concordant for the set of qPCR-validated genes.

Among the qPCR-validated genes that were under-expressed in metastatic relative to primary

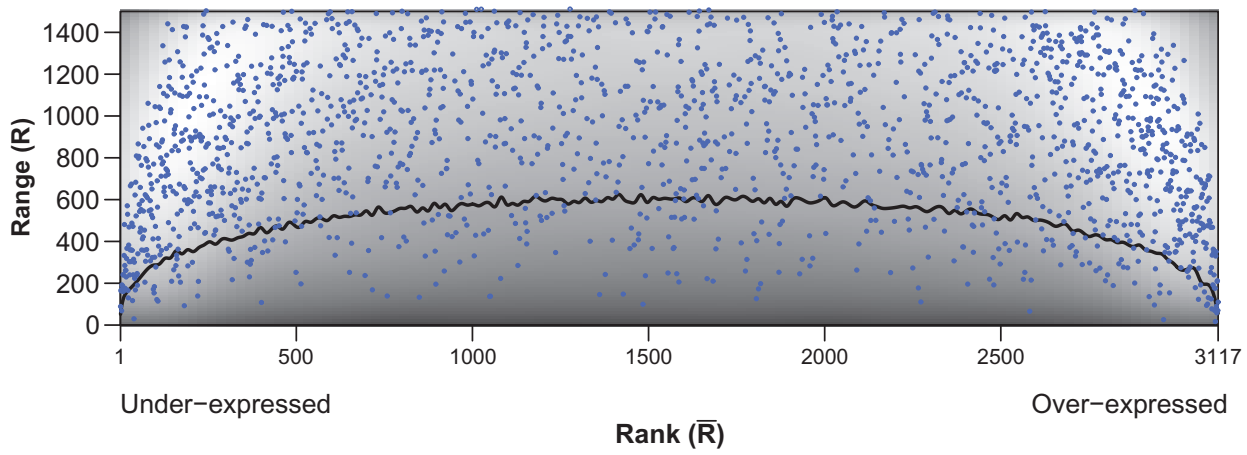

Figure 3. Plotted are the ranks of the average platform-specific ranks of the average fold-change in expression comparing metastatic to primary cell lines (x-axis) versus the range of the within-platform ranks (y-axis). The null distribution was obtained from 1000 permutations of the gene labels within each study. For each permutation, the range of ranks were ordered by the average rank. The shaded background depicts the density of the null distribution of the range of ranks at each rank of the average rank, with lighter shades denoting more densely plotted regions of the null. The black line is the 0.05 quantile of the permutation distribution.

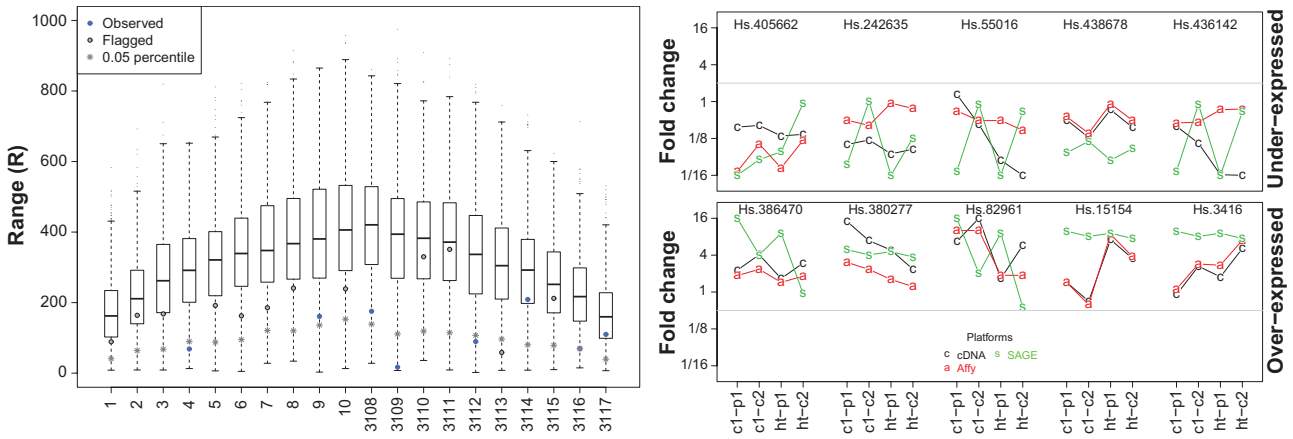

Figure 4. Left: Boxplots of the null distribution of the range of ranks for the top 10 under-expressed and over-expressed genes across. The observed range is plotted as a circle. Flagged circles indicate a within-class (primary or metastatic) fold change exceeding 3 in one or more of the platform. Right: The observed fold changes for the top genes with the smallest (under-expressed in metastatic) and largest (over-expressed in metastatic) rank of the average rank.

cell lines, the qPCR-estimated fold changes are overall correlated to the high-throughput measures of fold-changes, yet for a few of the comparisons the direction of fold change is discordant. In particular, the fold change estimated by qPCR for the ht-p1 comparison is greater than 2 for both the TALDO1 and PAM genes, whereas the fold change is slightly below 1 in each of the high-throughput platforms for these genes (Fig. 5). To determine whether alternative splicing could be a contributing factor for the apparent discordance between the high-throughput platforms and qPCR, we mapped the probe sequences in each of the four platforms to exons and used Aceview to assess whether known isoforms could account for differences.[23] The exon mappings for 3 genes in which discordance could plausibly arise from alternative splicing are displayed in Figure 6. For instance, qPCR primers for TALDO1 in the initial experiment (q1) map to the 4–5 exon junction. The 4–5 exon junction is not spanned by the high-throughput platforms and is absent in some transcripts on Aceview (not shown). Repeating the qPCR-validation (q2) with different primers, we found that the ht-p1 and ht-ct comparisons for TALDO1 were more consistent with the high-throughput platforms, yet fold-changes in expression remained discordant in others (6).

Discussion

This paper explores several approaches to assess concordance of differential gene expression measured from 2 primary and 2 metastatic pancreatic cancer cell lines by 3 high-throughput platforms, with the goal of prioritizing a gene list for validation by qPCR. Pairwise scatter plots of the fold-changes in expression highlight the challenges of this analysis, with near-zero correlations observed between the intensity-based arrays (Affymetrix and cDNA) and SAGE. As qualitative categories of differential expression can be less sensitive to technological differences than quantitative measures of fold change, Kappa statistics plotted as a function of the threshold used to classify differential expression can provide a more robust assessment of concordance. For the pancreatic cancer analysis, Spearman correlation coefficients and Kappa statistics indicate moderate concordance of the Affymetrix and cDNA platforms, but low concordance with SAGE. As opposed to dropping SAGE from the analysis, we adopted an approach whereby genes prioritized by the rank of the average platform-specific ranks could be visualized along with the spread of the observed ranks. Plotted against a background estimated from a null that assumes that the the ranks were independent across platforms, one can identify genes near the top and bottom of the list that are ranked more consistently than one would expect under the null. Such a visualization could also be applied to alternative rank-based schemes for prioritizing gene lists, and alternative measures of rank variability. Overall, the fold changes in expression as measured by the high-throughput platforms for 10 genes near

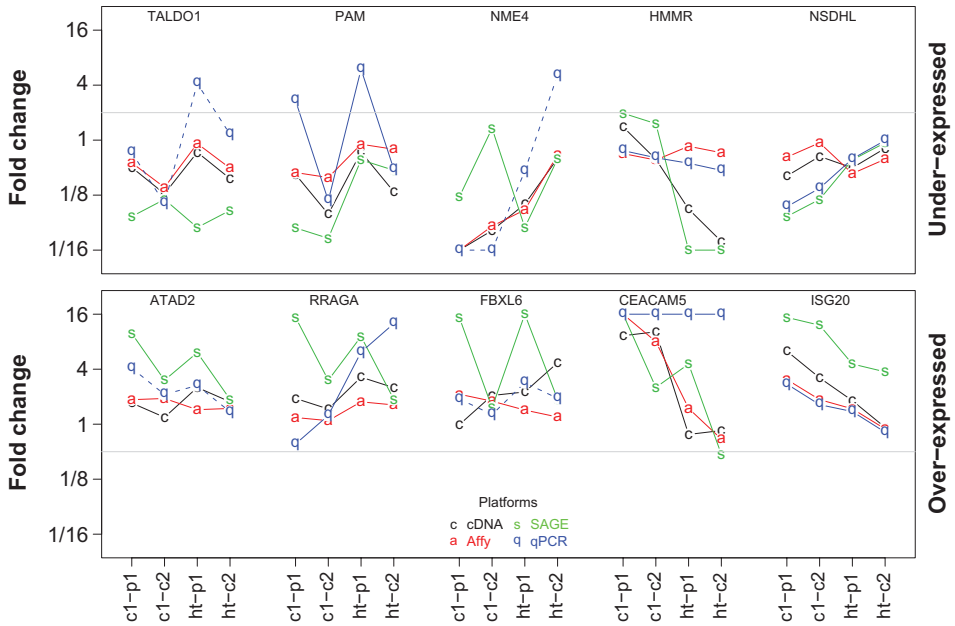

Figure 5. The observed fold changes of 10 genes selected for validation by qPCR. Fold changes less than 1/16 or greater than 16 were thresholded.

the top (under-expressed in metastatic) and bottom (over-expressed in metastatic) of the list were in agreement with the fold changes measured by qPCR. While batch- and technological artifacts unrelated to the sequence-characteristics of the probes in the individual platforms are likely to account for much of the cross-platform discordance, hypotheses regarding biological mechanisms for discordance such as alternative splicing can be explored using the approaches discussed here.

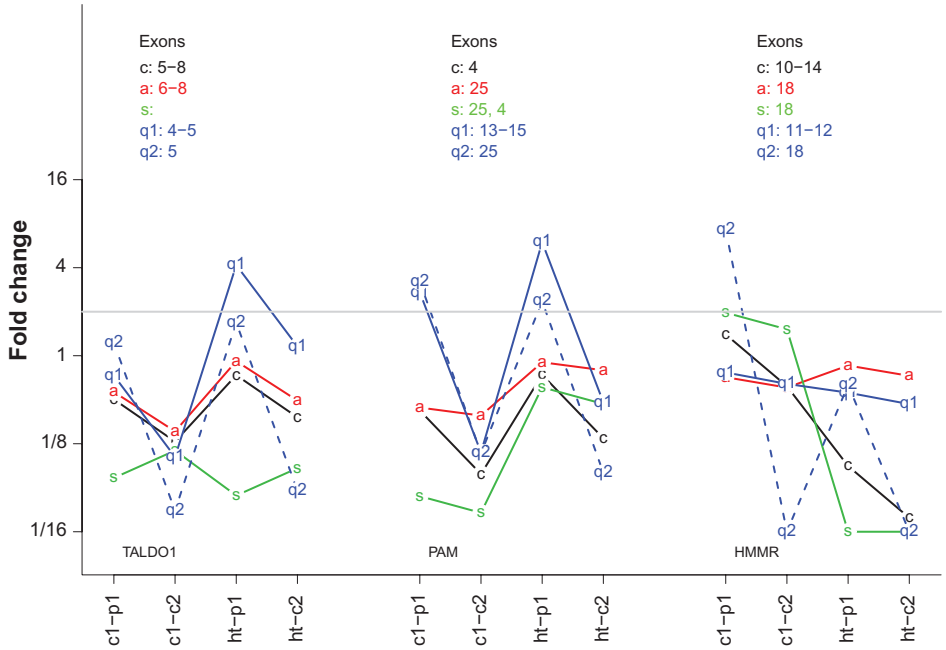

Figure 6. Probes for the high-throughput platforms and primers from the initial (q1) and repeated qPCR (q2) experiments were each mapped to exons (top). For TALDO1, the fold changes measured by qPCR were more consistent with the high throughput platforms using a primer that does not span the exon 4-5 junction. For PAM, the pattern of expression is similar in both qPCR experiments, regardless of whether the primer spanned exons 13-15 or 25. For the HMMR gene, we observe substantial heterogeneity in the SAGE platform with probes for exon 18. While we observe similar levels of heterogeneity in the repeated qPCR experiment with primers for exon 18, the fold changes were only moderately correlated.

Acknowledgments

This work was supported by grant 5T32ES012871 from the U. S. National Institute of Environmental Health Sciences, grant DMS034211 from the NSF, and grant 5P30 CA06973-44 from the NIH.

Disclosure

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*. 2003;19–1:I84–90.
2. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. 0008-5472 (Print) Journal Article Meta-Analysis. *Cancer Res*. 2002;62:4427–33.
3. Conlon Erin, Song Joon, Liu Jun. Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*. 2006; 7:247.
4. Scharpf Robert B, Tjelmeland Håkon, Parmigiani Giovanni, Nobel Andrew. A Bayesian model for cross-study differential gene expression. *JASA*. 2009;104:1295–310.
5. Parmigiani G. Measuring uncertainty in complex decision analysis models. *Stat Methods Med Res*. 2002;11:513–37.
6. Shen Ronglai, Ghosh Debashis, Chinnaiyan Arul. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics*. 2004;5:94.
7. Breitling Rainer, Herzyk Pawel. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J Bioinform Comput Biol*. 2005;3:1171–89.
8. Hong Fangxin, Breitling Rainer, McEntee Connor W, Wittner Ben S, Nemhauser Jennifer L, Chory Joanne. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*. 2006;22:2825–7.
9. Elo Laura L, Filen Sanna, Lahesmaa Riitta, Aittokallio Tero. Reproducibility-Optimized Test Statistic for Ranking Genes in Microarray Studies IEEE/ ACM Trans. *Comput Biol Bioinformatics*. 2008;5:423–31.
10. Iacobuzio-Donahue CA, Ashfaq R, Maitra A, et al. Highly Expressed Genes in Pancreatic Ductal Adenocarcinomas: A Comprehensive Characterization and Comparison of the Transcription Profiles Obtained from Three Major Technologies. *Cancer Research*. 2003;63:8614–22.
11. Speed TP ed. Statistical Analysis of Gene Expression Microarray Data. London: Chapman and Hall 2003.
12. Hruban RH, Goggins M, Parsons J, Kern SE. Progression model for pancreatic cancer. *Clin Cancer Res*. 2000;6:2969–72.
13. Yang Yee Hwa, Dudoit Sandrine, Luu Percy, et al. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*. 2002;30:e15.
14. Scharpf Robert B, Iacobuzio-Donahue Christine A, Sneddon Julie B, Parmigiani Giovanni. When should one subtract background fluorescence in 2-color microarrays? *Biostatistics*. 2007;8:695–707.
15. Carter Scott L, Eklund Aron C, Mecham Brigham H, Kohane Isaac S, Szallasi Zoltan. Redefini-tion of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics*. 2005;6:107.
16. Dai Manhong, Wang Pinglang, Boyd Andrew D, et al. Evolving gene/ transcript definitions signifi-cantly alter the interpretation of GeneChip data. *Nucleic Acids Res*. 2005;33:e175.
17. Cope Leslie, Zhong Xiaogang, Garrett Elizabeth, Parmigiani Giovanni. MergeMaid: R tools for merging and cross-study validation of gene expression data. *Stat Appl Genet Mol Biol*. 2004;3:Article29.
18. Everitt BS. Moments of statistics kappa and weighted kappa. *The British Journal of Mathematical and Statistical Psychology*. 1968;21:97–103.
19. Breitling Rainer, Armengaud Patrick, Amtmann Anna, Herzyk Pawel. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett*. 2004; 573:83–92.
20. MacDonald NJ, Freije JM, Stracke ML, Manrow RE. Steeg PS. Site-directed mutagenesis of nm23-H1. Mutation of proline 96 or serine 120 abrogates its motility inhibitory activity upon transfection into human breast carcinoma cells. *J Biol Chem*. 1996;271:25107–16.
21. Zimmer R, Thomas P. Mutations in the carcinoembryonic antigen gene in colorectal cancer patients: implications on liver metastasis. *Cancer Res*. 2001;61:2822–6.
22. Zou June X, Revenko Alexey S, Li Li B, Gemo Abigael T, Chen Hong-Wu. ANCCA, an estrogen-regulated AAA+ ATPase coactivator for ERalpha, is required for coregulator occupancy and chromatin modification. *Proc Natl Acad Sci U S A*. 2007;104:18067–72.
23. Thierry-Mieg Danielle, Thierry-Mieg Jean. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol*. 2006; 7 Suppl 1:S12.1–1214.