Improving Cancer Gene Expression Data Quality through a TCGA Data-Driven Evaluation of Identifier Filtering

Kevin K. McDade[1,2], Uma Chandran[1] and Roger S. Day[1]

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA. [2]Department of Science, The Pennsylvania State University, Shenango Campus, Sharon, PA, USA.

ABSTRACT: Data quality is a recognized problem for high-throughput genomics platforms, as evinced by the proliferation of methods attempting to filter out lower quality data points. Different filtering methods lead to discordant results, raising the question, which methods are best? Astonishingly, little computational support is offered to analysts to decide which filtering methods are optimal for the research question at hand. To evaluate them, we begin with a pair of expression data sets, transcriptomic and proteomic, on the same samples. The pair of data sets form a test-bed for the evaluation. Identifier mapping between the data sets creates a collection of feature pairs, with correlations calculated for each pair. To evaluate a filtering strategy, we estimate posterior probabilities for the correctness of probesets accepted by the method. An analyst can set expected utilities that represent the trade-off between the quality and quantity of accepted features. We tested nine published probeset filtering methods and combination strategies. We used two test-beds from cancer studies providing transcriptomic and proteomic data. For reasonable utility settings, the Jetset filtering method was optimal for probeset filtering on both test-beds, even though both assay platforms were different. Further intersection with a second filtering method was indicated on one test-bed but not the other.

KEYWORDS: data quality, transcriptomic, proteomic, identifier filtering, TCGA, correlation, microarray

CITATION: McDade et al. Improving Cancer Gene Expression Data Quality through a TCGA Data-Driven Evaluation of Identifier Filtering. *Cancer Informatics* 2015:14 149–161 doi: 10.4137/CIN.S33076.

TYPE: Original Research

RECEIVED: August 18, 2015. **RESUBMITTED:** October 26, 2015. **ACCEPTED FOR PUBLICATION:** November 03, 2015.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Eight peer reviewers contributed to the peer review report. Reviewers' reports totaled 2429 words, excluding any confidential comments to the academic editor.

FUNDING: Thank you to the National Library of Medicine (NLMT15LM007059) KKM 2009–2013 for funding. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: kkm5@pitt.edu

Introduction

Do commonly utilized methods to process raw data from the high-throughput genomic platforms differ much from each other? Does it matter which methods are utilized to process the data? Repositories of information, such as the Gene Expression Omnibus, cBioPortal, and The Cancer Genome Atlas (TCGA), contain hundreds of platforms and thousands of patient samples.[1–3] These platforms include measurement of gene expression, copy number variation, protein expression, and posttranslational modification. All of this information is available to users in levels of data, where, for most users, only the processed data are available. Some workflow options are ready-to-go: the data available are preprocessed, such as Affymetrix HU133 Plus 2.0 data, RNA-Seq data, methylation data, and many other data types in TCGA. Alternatively, many analysts prefer to start with raw data and apply a customized workflow consisting of their preferred sequence of processing steps. *Workflow options* include ready-to-go workflows, custom workflows, individual processing steps, or tuning parameters in a particular step. Any change in a workflow step or a parameter setting constitutes a new workflow option. To what extent do these choices affect the final data set to be analyzed? If the data sets differ substantially, will they differ in quality? If so, how can we tell which is best? Finally, will soundness of the scientific conclusions be harmed by suboptimal workflow choices and improved by better choices? Surprisingly, these questions are scarce in bioinformatics literature. Aside from the obvious benefit that the quality of analyses could be improved, there is the issue of comparing results from different studies. When two investigations report on comparable data sets, a third party may wish to compare or contrast the results, for example, for scientific validation. The choice of different workflows in the two studies generates a potential confounder in comparing them. Greater consensus on workflow choices would help alleviate this problem.

An example of a data setting burdened by a poor understanding of workflow option choices is the Affymetrix microarray. Affymetrix expression data are publicly available for more than 35,000 data sets, and is an immensely valuable resource for almost every type of cancer research.[1] However, there is no de facto standard of determining the gene expression values from raw data. Many processing and normalization options can yield values of gene expression on about 18,000 gene products from an ambiguous set of 54,675 probesets. A critical step in an Affymetrix workflow is to remove or filter poor quality probesets. This process

of removing bad measurement points has been defined previously as identifier filtering.[4]

Identifier filtering applied to Affymetrix chips presents an opportunity to evaluate workflow options concisely. We previously performed a comparison, not an evaluation, of identifier filtering. In identifier filtering, the user removes features (ie, probesets) judged to do a poor job, reflecting expression of their intended gene products. Table 1 outlines the implementation methods of identifier filtering tested here, including PLANdbAffy (PD), Jetset (J), AffyTag (AT), AffyGrade (AG), Masker (M), EnCode (E), and three methods derived from GeneAnnot (Geneannot specificity (GSPE), Geneannot sensitivity (GSEN), Geneannot quality (GQ)).[5–9] Table 1 also provides the abbreviations used in this article for each of the nine filtering strategies. These methods apply diverse criteria that consider nucleotide complementarity, probe design, and cross-hybridization of probe to off-target gene product.

This article presents a comprehensive evaluation of workflows consisting of probeset identifier filtering methods and their combinations. The methodology is previously published,[10] and this article is an application of the methodology to an important problem in bioinformatics practice. As a test-bed for this evaluation, it uses transcript expression data paired with protein expression data. However, the goal of this work is not specifically to guide analysis of paired data sets, but rather, a much broader goal, to provide guidance for feature filtering in transcript expression experiments.

Prior research by our laboratory group has documented disagreements among resources that map between identifiers for probesets and identifiers for proteins.[4] This previous work was implemented using a Bioconductor[11] package, IdMappingAnalysis.[12] We showed that the quality of the mappers could be compared based on real biological data.[4] Subsequent methodological work created a more general decision-theory-based approach and demonstrated how other workflow elements besides identifier mappers, including filtering methods and threshold choices, can also be evaluated.[10]

For a variety of reasons, previous investigators have examined correlations between data from pairs of expression platforms, for example, relating RNA-Seq to oligonucleotide data and relating oligonucleotide data to protein expression data.[13–16] A natural assumption is that greater transcript expression will lead to greater protein expression. There are, however, biological reasons that a particular mRNA species might have weak or no correlation with the expression of the correctly mapped protein.[10,15–19] The evaluation method applied here takes this into account, which is described below.

Methods

For reference, an overview of the methodology is shown in Figure 1.

Table 1. Identifier filtering methods and the scores utilized for filtering.

FILTER METHOD SYMBOL	DESCRIPTION	DEVELOPER CRITERIA	IDENTIFIER FILTERING
AT[8,27]	Affytag–Pre-2004 Affymetrix annotation for the Affymetrix HGU133 Plus 2.0 array	Original annotation determined by mapping to UniGene and Locus Link. "_at is considered unique.	Filter al annotation tags that begin with "_[agirxsf]_at"
AG[8,27]	Affy Grade–Netaffx Transcript Assignment Pipeline	"A" grade is the highest grade where ≥9 probes match transcript sequence.	Filter grades not equal to A.
M[28]	Masker–National Cancer Institute alternative chip definition file (CDF) masking out probesets with poor target location	A CDF file which eliminates a probe when more than 2 nucleotides to not match the target as well as nonspecific probes	Filter any probeset that has no remaining probes on the mask
GSEN[9]	GeneAnnot Sensitivity	The fraction of the probes in a probeset that match Watson-Crick nucleotide base pairs in the nominal gene	Filter probesets with Geneannot Sensitivity <90%
GSPE[9]	GeneAnnot Specificity	Sum over the number of matching probes with lower weight to non-specific probes	Filter probesets with Geneannot Specificity ≤50%
GQ[9]	Geneannot Quality Score	A pipeline which confirms the probeset annotation with GeneCard data.	GQ = 1 is confirmed entirely with GeneCard data; Filter probesets with a GQ = [2–6]
E[29]	Encode–Encyclopedia of DNA elements	Protein coding genes are determined by human curation, RNA sequence and comparative genomics	Filter all probesets that map to a non-"Protein coding" target
PD[7]	PlandbAffy database	BLAT of target to the probe and evaluation of nucleotide mismatch or exon location	Filter all probesets with a proportion of "good" probes <30%
J[6]	Jetset Bioconductor package	Determines features such as robustness of the probe, coverage, as well as nucleotide alignment with the reference genome	Filter all except the highest-scoring probeset among those annotated for target gene

Data description. Two cancer data sets were utilized as test-beds for this evaluation of filter methods. The first is a data set of 91 endometrial cancer samples and 7 normal endometrium samples, studied with tandem mass spectrometry proteomic data and Affymetrix U133 2.0 Plus expression data from the Gynecologic Cancer Center of Excellence (GYN-COE).[4] The second is a data set of 401 ovarian serous cystadenocarcinoma samples with protein assay (reverse phase protein array (RPPA)) and Affymetrix U133A mRNA data from TCGA.[20,21] These data sets differ substantially in sample size, number of features, and platforms. Proteomic and mRNA feature identifiers are paired across platforms using the IdMappingRetrieval bioconductor package.[11,22] The ENvision mapping was selected based on the results from our previously published evaluation of identifier mapping resources.[4,23]

The endometrial cancer biomarker studies were performed by the GYN-COE.[4,24,25] The tissue samples were subjected to trypsin digest at the University of Pittsburgh. Tryptic peptide digests were separately analyzed in duplicate by Liquid chromatography-tandem mass spectrometry (LC-MS/MS) with an Linear Ion Trap - Fourier Transform (LTQ-FT) (ThermoFisher Scientific) and an Linear Ion Trap - Orbitrap (LTQ-Orbitrap) (ThermoFisher Scientific) mass spectrometer. The combined analyses yielded 12,288 distinct protein UniProt accessions across all samples and both instruments. The gene expression data were performed on the Affymetrix U133 2.0 Plus Array. For complete details of the microarray and proteomic studies, see Day et al.[4]

For the second test-bed, we turned to TCGA. TCGA has multiple levels of genomic, transcriptomic, somatic mutation, and protein expression data for many types of cancer data. The ovarian serous cystadenocarcinoma sample data set is especially useful here. The ovarian cancer data has 401 samples with various types of genomic, transcriptomic, and proteomic data. The data utilized here come from two platforms: the U133 A Affymetrix Array, with 22,277 probesets, and the RPPA studies on 68 proteins performed by MD Anderson Cancer Center.[2,20,26] The proteins selected for the RPPA studies were chosen for their cancer relevance. Using the IdMappingRetrieval Bioconductor package,[22] we obtained 151 probeset-to-protein pairs.

Filtering methods and strategies to be evaluated. Nine filtering methods were evaluated and compared, and they are listed in Table 1. AT removes probesets for which the Affymetrix identifier (ID) contains a qualifier, that is, the ID ends in "_[agirxsf]_at," reflecting original doubts concerning the correct and unique hybridization of the probes in each probeset, as documented by Affymetrix when the array was designed.[8,27] Although the identifier tags were initially

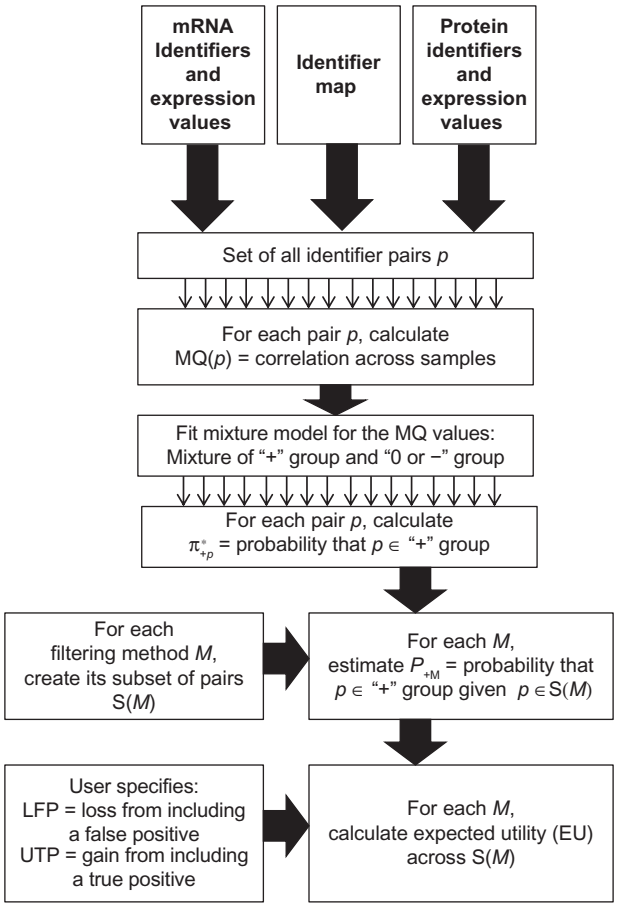


Figure 1. Identifier filtering evaluation flowchart. Flowchart summarizes the steps of evaluating identifier filtering methods.

used as the de facto quality measure, these tags had reliability problems. We include this tag-based probeset quality measure to verify that our quality assessment paradigm can detect the expected deficiency of performance in a superseded method relative to the more recent measures. AffyGrade (AG), provided by the NetAffx array annotation file, is a quality grade labeled as A, B, C, R, and others. Only probesets with A grade were accepted, since A grades represent at least nine matching probes to the target mRNA.[27] The National Cancer Institute (NCI) Masker[28] filter removes probesets omitted from the NCI masked chip description file (CDF). The Masker was produced by the NCI Laboratory of Population Genetics. The CDF file eliminates any probes that do not have at least 24 out of 25 nucleotides match the target GenBank transcript. In addition, it eliminates any nonspecific probes that map to a different chromosome, strand, or are part of a gene cluster that could cause cross-hybridization. The CDF file can be obtained from http://masker.nci.nih.gov/ev/.

We test three filters utilizing Geneannot,[9] a database of gene expression annotations and quality that evaluates the Affymetrix probesets on the following criteria. For each of the probesets on the Affymetrix chip, sensitivity, specificity, and an overall quality score is determined. Sensitivity is defined as the fraction of the probes in a probeset that match Watson–Crick nucleotide base pairs in the nominal gene. This classification is labeled as GSEN. The next classification is labeled as GSPE and is a sum over the number of matching probes with lower weight placed upon nonspecific probes. Thresholds defining GSEN and GSPE were, respectively, sensitivity metric ≥ 0.9 and specificity metric ≥ 0.5, each chosen by maximizing expected utility. Finally, the GQ measure is determined from the ordinal rank assigned by Geneannot to demonstrate the confirmation of the probeset to mRNA match. A score of 1 is reported to be the best, which demonstrates that the probes were confirmed using the GeneCards data via EntrezGene or Ensembl. The worst score is 6, which is defined as probesets where the only information available is original NetAffx annotation.[27] For the purposes of this study, GQ accepts only probesets with a score of 1. Our EnCode (E) filter utilizes the EnCode[29] project's determination of protein-coding status of the target sequence location in the genome to remove probesets of non-coding targets. The files are available at http://encodeproject.org/ENCODE/downloads.html. The GENCODE version 12 annotation files were used to determine gene status from human genome build 37. The gene status is classified as protein coding, transcribed pseudogene, untranscribed pseudogene, lincRNA, not identified by GENECODE, etc.[29] Only probesets with the protein coding Ensembl code were accepted. The Ensembl codes were matched to the UniProt accession codes present in our analysis.

The PD filter uses the PD[7] database, which uses the probeset sequence and the BLAT database to align probe nucleotide sequences to the target and assign to each probe a grade reflecting alignment mismatches, alignment to other sequences risking cross-hybridization, and intronic versus exonic location. The PD filter was defined to accept a probeset if 30% of the probes within the probeset were classified as perfect exonic, noncross-hybridization matches. To set the threshold, we maximized expected utility, as described in Day and McDade.[10]

The Jetset (J) filter uses the Jetset[6] assessment, which not only considers nucleotide complementarity across the probesets but also considers splice isoform coverage and transcript degradation. In addition, Jetset (J) will score each probeset of a target gene and select the best probeset (of currently defined probesets) for each gene on the chip. Therefore, Jetset (J) is a stringent eliminator of probesets.

Methodology for identifier filtering evaluation. The identifier filtering evaluation of probesets uses a previously published methodology for comparison of bioinformatics workflow options to determine the evaluation metric. The steps in this application to identifier filtering are summarized in Figure 1. For more details, see Day and McDade.[10] The method requires the following inputs:

- A large number of *biological samples* from a biological repository, such as TCGA, or a private collection of biological samples.
- Two high-throughput platforms each with a feature list of identifiers; the two platforms must be on the same *biological samples*.
- A planned set of workflow options to compare.
- An identifier map, which connects the pairs of data across the platforms (ie, transcript to protein).
- A *model quality score* for each pair p, designated MQ_p. The MQ scores are treated independently for modeling the mixture distribution. In applications thus far, this score is a correlation coefficient.
- For each method M, the set of pairs accepted or produced by this method is designated as $S(M)$.

In the current application, each pair is an mRNA transcript feature paired with a protein feature linked through the ENvision identifier mapping resource. Membership of a pair p in the set $S(M)$ means that the method M claims that the transcript feature in p should be included for any data analysis. The two platforms, respectively, assess the two processes: gene expression and protein expression.

The *model quality* score in this application is the correlation of the two measurements across the biological samples. We consider the probability density of the correlation values for all pairs produced by the method M (Fig. 2, black line). This density is modeled as a mixture with the following components:

- "+": The transcript feature and the protein feature are correctly identified, and they are truly *biologically*

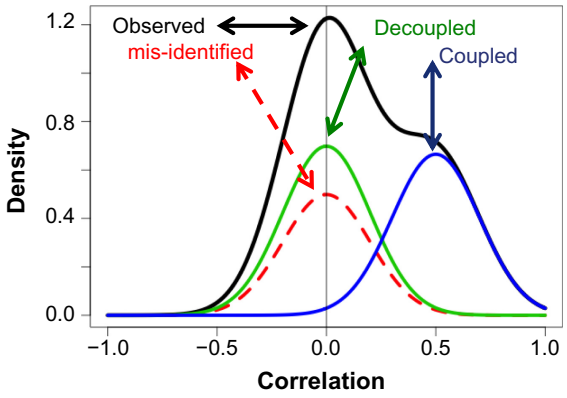
coupled. This means that a pair in this component is correctly mapped between transcript and protein identifier, and transcript abundance and protein abundance are monotonically related. The blue line in Figure 2 represents the "+" component.

- "0": The transcript feature and the protein feature are correctly mapped but *biologically decoupled*. This means that the expected monotonic relationship between a transcript and a protein is not observed. There are many biological reasons for decoupling, including RNA interference by microRNA, post-translational processing, and any other mechanism causing the protein abundance to fail to reflect the transcript abundance. The green line in Figure 2 represents the distribution of correlations for decoupled pairs, which we refer to as the "0" component.
- "x": An incorrect mRNA/protein pair relationship was assigned. The red dashed line in Figure 2 represents the distribution of correlations for misidentified pairs, which we refer to as the "x" distribution. Pairs included in this distribution should not be assigned. These assignments may be due to incorrect actions on the part of the identifier mapping or the workflow in general.

Estimation of the "+" posterior probability. We would like to identify the features in either "+" or "0" for inclusion. However, the data cannot distinguish between the "0" and "x" groups. Under mild assumptions, the method with the highest posterior probability for "+" is also the method with the highest posterior probability for "+" or "0." We refer to the combined "0" and "x" groups as the "-" group. Even though groups "0" and "x" cannot be distinguished, basing the relative performance of workflow methods on the mixture distributions from the observed correlations is likely to yield the correct decision; the argument for this statement is previously reported.[10]

Let $G(p)$ be the component, whether "+", "0", or "x", to which the pair p belongs to. A better workflow option should do a better job at excluding incorrectly mapped pairs (ie, those with $G(p)$ = "x"). Increasing the probability that $G(p)$ = "+" should reduce the $G(p)$ = "x" component. Let the proportion of pairs in group g be $\Pr(G(p)=g)=\pi_g$ for $g \in \{"+","0","x"\}$. The mixture model provides the opportunity to estimate $\Pr(G(p)$ = "+") for each pair p. This probability provides the metric we need to evaluate workflow options.

We now assume that the true correlations for all the pairs in group g are distributed as a mixture of normal distributions with mean φ_g and variance V_g. There is also measurement error, so the correlation of each pair p in group g is normally distributed with marginal mean φ_g and marginal variance $\tau_{gp}=V_g+\sigma_p^2$, where σ_p^2 is the measurement error variance specific to pair p. We estimate σ_p^2 by the bootstrap method, as described in Day and McDade.[10] To estimate the probability of a pair belonging to the "+" group, we use an expectation conditional maximization (ECM) algorithm to determine the following parameters: (1) the prior probability π_+ of belonging to the "+" group, (2) the within-group



ID pair group	*True correlation*	Description	Pair counts for method *M* (*A* or *B*)
Observed	Mixture	All observed pairs	n_M
+: coupled	>0	True biological coupling	n_{M+}
0: decoupled	=0	Biological decoupling	n_{M0}
x:mis-identified	=0	Mapping error, etc.	x_{M0}

Figure 2. Hypothetical mixture components for correlation. Reproduced from Day RS, McDade KK. A decision theory paradigm for evaluating identifier mapping and filtering methods using data integration. *BMC Bioinformatics*. 2013;14(1):223, under the terms of the Creative Commons Attribution License.

Notes: Observed (black): marginal density of correlations. Misidentified (red, dotted): density of correlations where features are either misidentified or incorrectly mapped. Decoupled (green): density of correlations of pairs correctly mapped but biologically uncorrelated (discordant). Coupled (blue): density of correlations of pairs correctly mapped and biologically coupled.

true variance V_+ of the correlations in "+" group, and (3) the within-group true variance $V_- = V_0 + V_x$. Here *true* signifies without sampling error. This is possible since we are able to constrain the variance of the "0" and "x" groups to 0. For a complete description of the ECM algorithm, see Additional File 1 of Day and McDade.[10]

Having determined the maximum likelihood estimates of the parameters, we can now calculate for each pair p the posterior probability of belonging to the "+" group by defining the following:

$$\pi^*_{+p} = \Pr(G(p) = + \mid MQ_p \text{ and parameter estimates})$$

$$\pi^*_{-p} = 1 - \pi^*_{+p} = \pi^*_{xp} + \pi^*_{0p}$$

This calculation provides the posterior probability of estimation of pair p belonging to the "+" component, given the correlation MQ_p and its sampling variance σ^2_p, from bootstrap sampling. To convert this variance into the variance of the posterior probability, the approximation in the delta method is used. This consists of multiplying the variance of the correlation times the square of the derivative of the posterior probability as a function of the correlation.

$$v^*_{+p} = \text{var}\left(\pi^*_{+p}\right) \cong \text{var}\left(MQ_p\right) \times \left(\frac{d\pi^*_{+p}}{dMQ_p}\right)^2.$$

The expression for the derivative is presented in an appendix.

A weighted mean of the "+" proportion provides an expected proportion of "+" group pairs for a given identifier filtering method. The weighted mean is estimated using the posterior probabilities of each pair and the variances of these posterior probabilities.

$$P_{+M} = \frac{\sum_{p \in S(M)} \pi^*_{+p} \left(v^*_{+p}\right)^{-1}}{\sum_{p \in S(M)} \left(v^*_{+p}\right)^{-1}}$$

This quantity provides the basis for comparing the methods, $M \in \{M_1,\ldots,M_k\}$.

Expected utility model. It is important to consider that different analysts have different analysis goals. One method may include a pair or a feature, while another excludes it. The pair will be either a *true positive* of the first method or a *true negative* of the second method. The relative value of including a true positive versus excluding a false positive will be different for different scientific goals. Utility values can express these valuations. We utilize the Bayesian decision principle of maximizing expected utility. This principle is useful for selecting a single filtering method, choosing a threshold for a method (Geneannot, PD), or selecting a Boolean composite filtering strategy (described in the next section).

For this study, we set the following values:

- L_{FP} = the loss associated with a false positive, ie, 1.
- U_{TP} = the utility of including a true positive, ie, 2.

We explored sensitivity of the comparisons between methods to these three values, and found that the comparisons are relatively insensitive (data not shown).

The Bayesian expected loss calculation is as follows:

$$EU = U_{TP} P_{+M} - L_{FP} P_{-M}$$

This is the mean expected utility (MEU). As an alternative, the analyst may choose to use total expected utility (TEU), which simply is the product of the number of methods compared and MEU.

Composite filtering strategies. Boolean conjunction (intersection; "and") and disjunction (union; "or") operators can create composite filtering strategies, which are easily evaluated. An analyst may consider whether the union or intersection of two or more filtering methods is worth the extra effort. Given a current strategy, for each so-far-unused method, one can automatically construct and evaluate the strategies formed by conjoining this method to the current strategy via conjunction or disjunction. A forward selection assesses the expected utility for each of these conjoined strategies and chooses the one with the highest expected utility. This is referred to as *greedy* selection because it takes the apparent best step, in sequence. In contrast is the exhaustive search of every Boolean combination of the methods, which in principle could find better strategies, but is impractical.

Results

Odd ratios demonstrating the disagreement between filtering methods. The nine filtering methods are far from redundant. Many analysts who use one of the filtering methods listed in Table 1 might expect only minimal differences in the probesets retrieved and retained. Instead, the nine filtering method strategies do not demonstrate similar probeset decisions. Table 2 compares the classifications of each pair of methods.

Each table entry is the odds ratio from the 2 × 2 table cross-classifying probesets as either filtered or retained by the two methods. The odds ratio is the product of the agreements divided by the product of the disagreements. An odds ratio of 1.0 indicates that the two classifications are providing independent information; an odds ratio much larger than 1 indicates redundant information, and an odds ratio much smaller than 1 indicates contradictory information. For example, the odds ratio of 5.9 comparing Jetset to EnCode in Table 2A indicates considerable redundant

Table 2. Odds ratio chart for probeset filtering.

For abbreviations, see Table 1. The table cell entries are the odds ratios assessing the degree of association of each pair of filtering methods. Each table entry is the odds ratio from the 2×2 table cross-classifying probesets as either filtered or retained by the two methods. The odds ratio is the product of the agreements divided by the product of the disagreements. For details regarding the interpretation of the odds ratios, see text.

PANEL A: ALL PROBESETS ON AFFYMETRIX HGU133 PLUS 2.0 ARRAY									
Filter	**J**	**E**	**PD**	**GSEN**	**GSPE**	**GQ**	**AT**	**AG**	**M**
J	–	5.9	3.7	3.1	24.7	29.3	1.2	20.3	0.35
E		–	11.4	29.3	32.5	46.4	0.3	14.1	0.60
PD			–	6.8	6.8	6.4	0.5	4.7	1.07
GSEN				–	50.0	1103.0	0.2	301.0	0.45
GSPE					–	760.0	0.3	53.2	0.79
GQ						–	0.3	50.1	0.69
AT							–	0.3	0.93
AG								–	1.85
M									–
PANEL B: THE 887 PROBESETS FROM THE ID PAIRS IN THE ENDOMETRIAL SAMPLE									
Filter	**J**	**E**	**PD**	**GSEN**	**GSPE**	**GQ**	**AT**	**AG**	**M**
J	–	1.87	2.34	0.96	2.62	2.73	4.91	2.81	5.05
E		–	4.00	7.34	4.24	21.60	0.68	16.20	1.49
PD			–	1.23	35.90	2.13	2.19	1.45	1.46
GSEN				–	0.74	Inf	0.42	70.80	1.26
GSPE					–	Inf	5.23	2.08	1.29
GQ						–	0.03	127.00	1.32
AT							–	0.31	2.77
AG								–	0.90
M									–

information. The odds of a probeset being excluded by EnCode are 5.9 times greater if the probeset is also excluded by Jetset versus if it is included by Jetset. In contrast, the odds ratio of 1.07 comparing PD to Masker indicates nearly independent information: knowing whether Masker includes a probeset says almost nothing about whether PD does. More remarkable still is the odds ratio of 0.35 for Jetset and Masker, which indicates that knowing that Masker includes a probeset considerably *decreases* the odds that Jetset includes it; Jetset and Masker provide contradictory information. (One might hope that they usefully complement each other. The analysis of Boolean combinations will address this hope). For details about odds ratios, see Szumilas.[30]

Fitted mixture models for the correlations. For each of the test-beds (endometrial and ovarian), a correlation mixture model was fitted to all feature pairs as described in the Methods section. Figure 3 shows the fitted mixture components for the two test-beds. They appear considerably different. Nevertheless, as we will see, the two mixture models lead to similar comparative evaluations of the filtering methods, suggesting that the best-practices conclusions we are seeking may have general application.

Each mixture distribution has two components: one centered at 0 and the other with mean >0. The right-most component, labeled "+", corresponds to the pairs where both features are correctly identified and mapped and also biologically coupled through the translation process, protein synthesis. For each pair, we calculated the posterior probability for belonging to the "+" component. Summing or averaging across the pairs accepted by a filtering method, we calculated the expected utility for that method.

Expected utility. The purpose of probeset filtering is to remove incorrectly identified or ineffectively designed probesets without removing too many correct probesets. Some investigators may want to apply stringent filtering criteria, for example, to reduce multiple comparisons penalties and false discoveries, while others would be more concerned with missing a true discovery. For purposes of illustration, we fix a utility of a true positive to 2 and a loss of a false positive to 1 (see the Methods section). This implies that an investigator would wish to include a true positive at the cost of including a false positive feature, but not at the cost of including three false-positive features, with indifference if the cost is two false positives. The different quantity–quality priorities of investigators are represented by two ways of combining expected utilities:

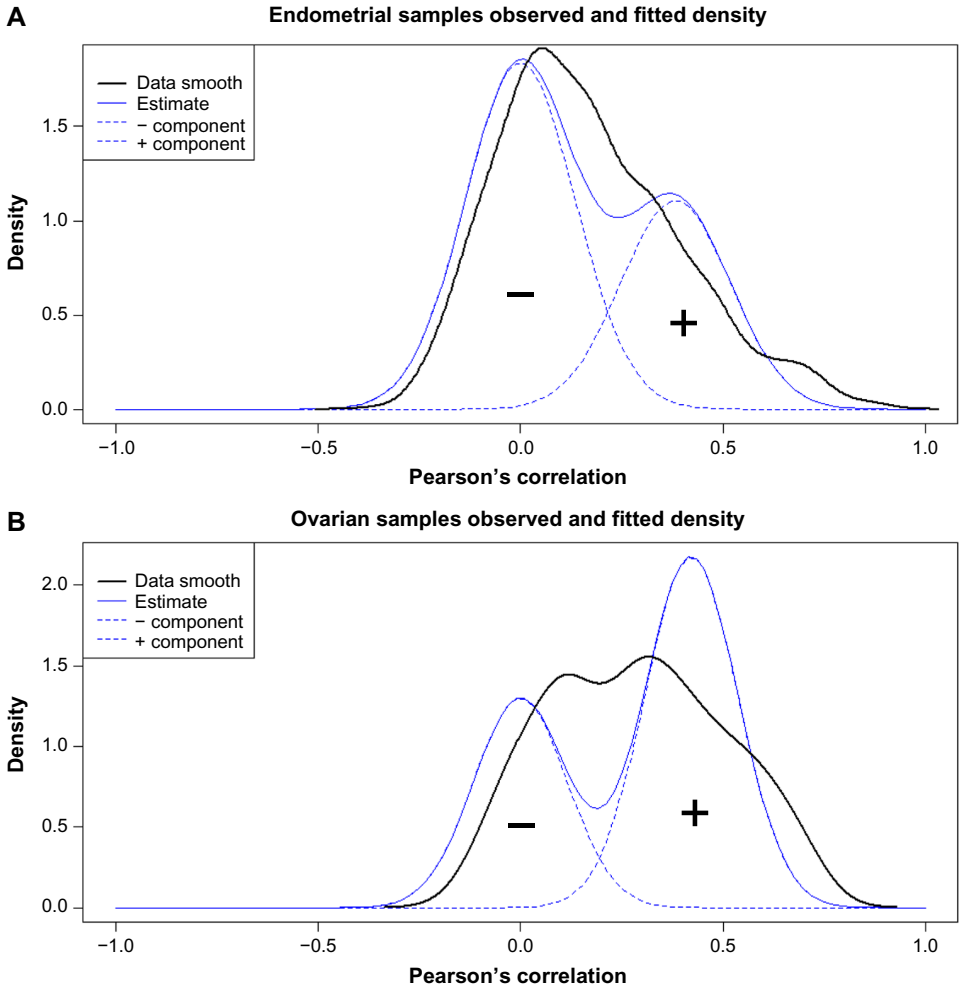


Figure 3. Observed and fitted density distributions for probeset filtering example: (**A**) GYN-COE endometrial experiment and (**B**) TCGA ovarian experiment.

Notes: In each panel, the horizontal axis represents the Pearson's correlation for pairs of mRNA expression and protein expression features (887 pairs in panel A, 151 pairs in panel B). The solid black line *data smooth* is a nonparametric estimate of the probability density of observed correlations. The solid blue line is a mixture distribution estimate of the probability density of the true correlations, determined from the generalized expectation maximization algorithm, which deconvolves the error term with individual variances for each correlation. Note that when this estimate is convolved with the error variances, the result matches the observed data smooth closely (data not shown). The dotted lines are the mixture components. The mixture component labeled "–" is interpreted as incorrect or decoupled feature pairs ($P = 0.624$ in endometrial samples and $P = 0.373$ in ovarian samples). The mean is constrained to 0 (see the Methods section). The mixture component labeled "+" is interpreted as correct and coupled feature pairs ($P = 0.376$ in endometrial samples and $P = 0.627$ in the ovarian samples).

the TEU and the MEU. An analyst choosing TEU wants as many features as possible, perhaps driven by the need to feed some systems biology algorithm. An analyst choosing MEU is more concerned with the quality of the resulting data set. Summary figures demonstrate the greedy forward selection for the endometrial and ovarian data sets (Fig. 4). For each of the filters applied in Figure 4A, there is a removal of poor quality probesets with a gain in TEU. Figure 4B illustrates the MEU over a series of filters; each successive filter reduces probesets with an increase in expected utility.

Figure 5 demonstrates a more detailed picture with each circle data point representing a set of probesets obtained from the application of an identifier filtering method. The two paths represent using TEU or MEU as the metric for the greedy forward selection. For the endometrial data, Figure 5A plots the estimated proportion of true coupled (quality) data versus the number of remaining pairs of data (quantity). The point at the upper left corresponds to including all 887 features pairs obtained with no filtering. The proportion of "+" pairs is only 0.30, which implies that the TEU is −81.9 and the MEU −0.0923. The conclusion is that, without filtering, one should not analyze these data. The labeled points correspond to reduced feature sets created by a single filtering method. The paths correspond to successive application of filters selected by a greedy forward selection of intersections and unions.

Jetset filtering provided the best single-method strategy for both TEU and MEU criteria (label = J). It is notable that Jetset was optimal even for TEU despite removing roughly

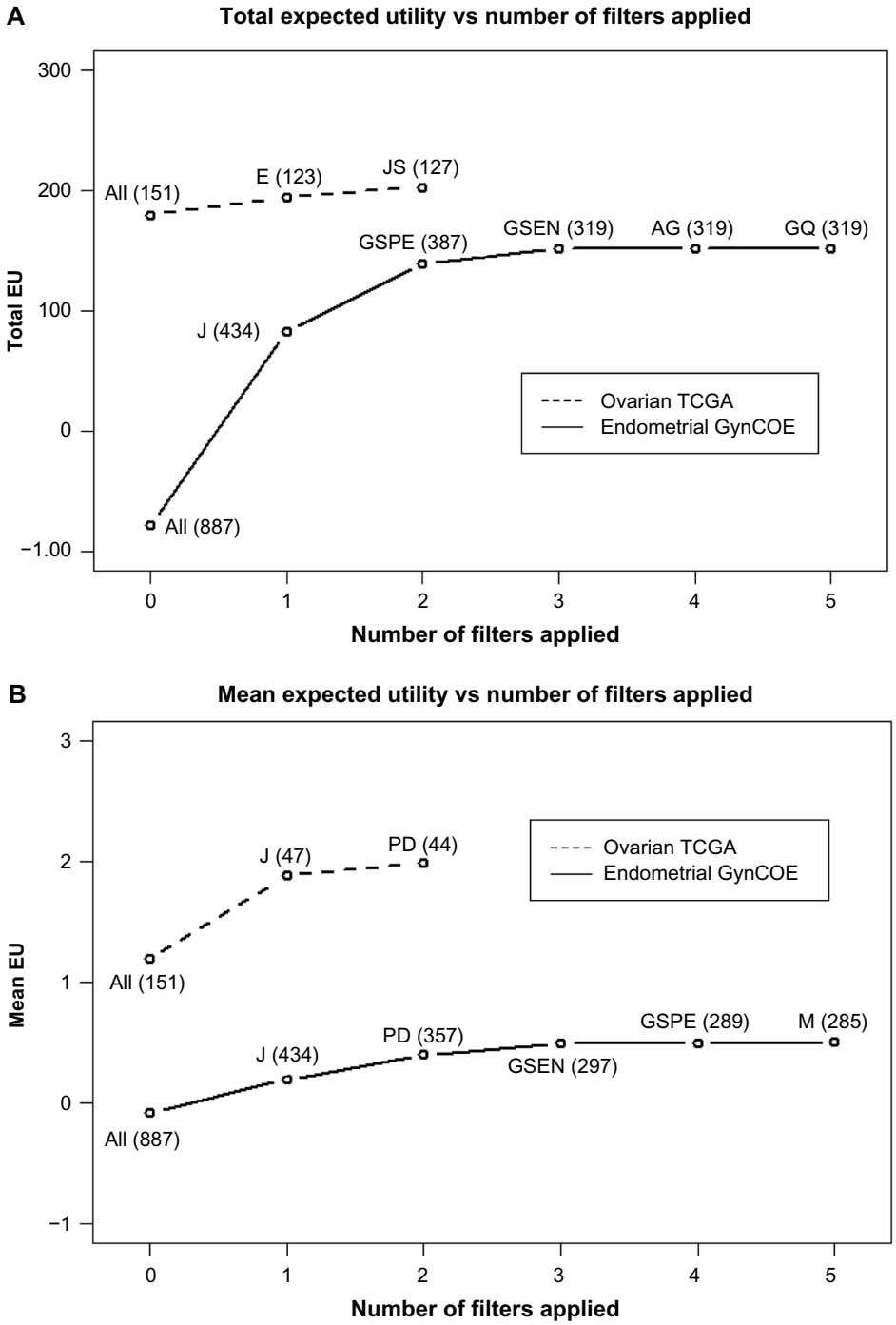


Figure 4. Greedy forward selection for probeset filtering example.

Notes: Starting with all probesets, the filters are applied to each cancer type using a greedy forward selection. The numbers of probesets are shown above each data point. The filter number represents the next filter intersection in the greedy forward selection, a union or an intersection. (**A**) The TEU is the greedy forward selection criterion. (**B**) The MEU is the greedy forward selection criterion.

half of the probesets (from 887 to 434; 51.1% probesets removed). For Jetset, TEU = 80.3 and MEU = 0.185, both in the positive zone, suggesting at least that after filtering, a data set is of sufficient quality to deserve analysis. Figure 5 shows the subsequent improvements by greedy selection of higher order Boolean combinations for the TEU (Panel A) and MEU (Panel B) criteria. Intersecting Jetset with GSPE was the best next step for the TEU criterion (filtering away 56.1% of the probesets), 138.9 TEU; intersecting with PD was the best next step for MEU criterion (filtering away 59.8%), 0.3868 MEU. Further selection did not improve either criterion noticeably (maximum TEU 148.5, maximum MEU 0.4864).

In the endometrial data set, the estimated proportion of true coupled (Pr("+")) is 0.303 with 887 mRNA–protein pairs. The endometrial greedy forward selection shows a very similar path, and in fact after one greedy node, both greedy search modes

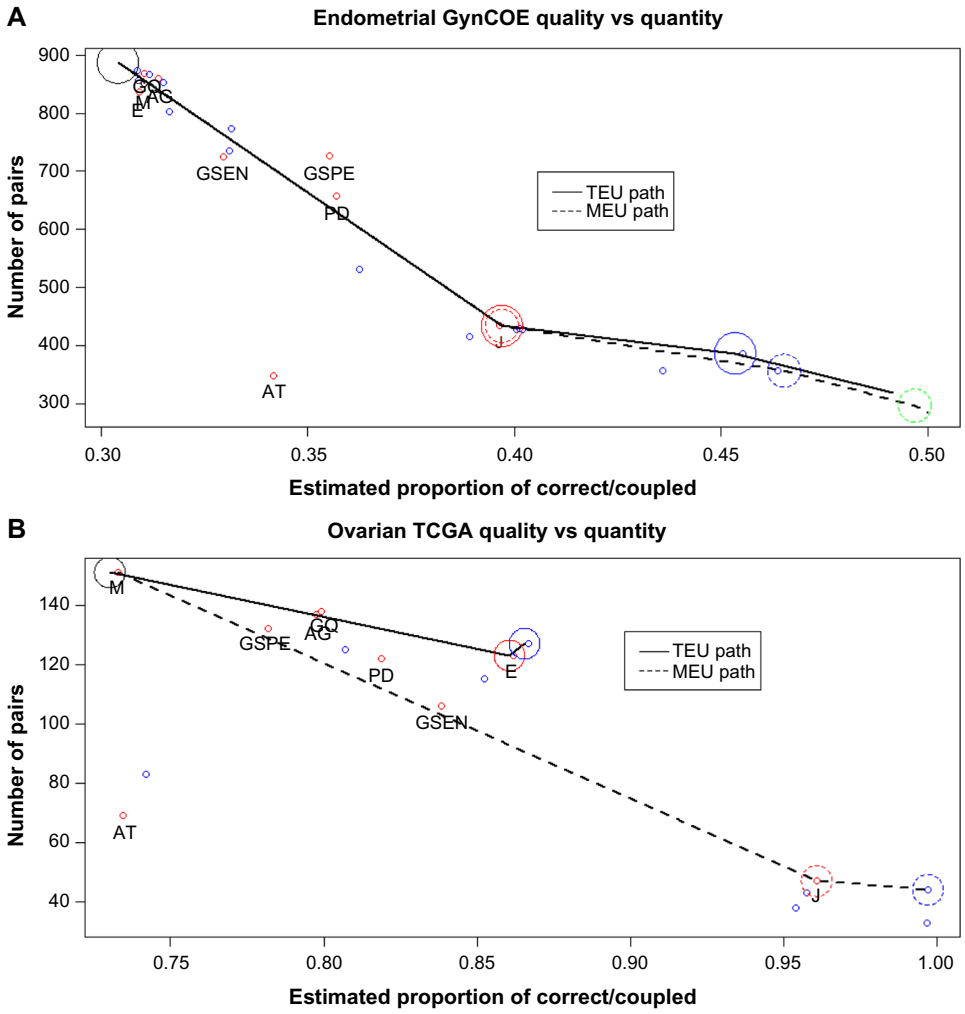


Figure 5. Quality versus quantity for probeset filtering example.

Notes: Points are plotted for filtering strategies constructed from filtering methods by Boolean operators. A Level 1 strategy (red) is a single filtering method. A Level 2 strategy (blue) is the intersection or union of two Level 1 strategies and so forth. The lines connect the best strategies (circled) at each Boolean complexity level according to greedy forward selection. A Level 3 strategy (green) is the intersection of three filtering strategies.

Abbreviations: J, Jetset; GQ, Geneannot quality; GSPE, Geneannot specificity; GSEN, Geneannot sensitivity; M, Masker; PD, Plandbaffy; AG, Affymetrix Grade; AT, Affymetrix Tag, E, EnCode.

find Jetset as the methodology of option, increasing the Pr("+") from 0.303 to 0.396. The optimal set for TEU is Pr("+") = 0.496, while the MEU finds a set with Pr("+") = 0.503.

In the ovarian cancer data set (Fig. 5B), Jetset filtering again provided the best single-method strategy for MEU criteria. Jetset reduced the number of probesets even more severely, from 151 to 47 (78.9% probesets filtered away) for the MEU selection criterion. The benefit in terms of the quality was quite dramatic, but the cost in terms of pair reduction actually decreased the TEU from 290 to 131. Figure 5B shows the subsequent improvements by greedy selection of higher order Boolean combinations. Intersecting Jetset with EnCode was the best next step for the TEU criterion (filtering away 18.5% of the probesets), 1.58 TEU; taking a union with Jetset after the EnCode intersection restored four probesets and increased the TEU very slightly to 1.60. No further union or intersection provided any improvement.

In the ovarian serous carcinoma TCGA data set, the Affymetrix to reverse phase protein assay data provide 151 pairs at a high Pr("+"), 0.733. Unlike the endometrial data, the TEU and the MEU provide two different paths to best practice of probeset filtering. The MEU path chooses the Jetset filter method by throwing away all but the 47 pairs in the Jetset optimization with a Pr("+") of 0.961. After two levels, the MEU maximizes to a Pr("+") of 0.997 and eliminated all but 44 pairs. The TEU favors quantity by keeping 123 pairs and a Pr("+") of 0.862. The TEU actually adds back in the union of the Jetset of four probesets to bring the total pairs to 127 and a Pr("+") of 0.867. Whether filtering with two methods rather than one is worth the extra effort is, of course, the judgment of the analyst.

Discussion

The goal of this work is to provide guidance for choosing a probeset filtering strategy in transcript expression experiments.

It is not to guide analysis of paired data sets. The usefulness of linking with the protein abundance data is specifically to help evaluate and compare different filtering methods.

Many investigators assume that the filtering methods available are close enough that the choice of filtering method will have a negligent effect on the overall results of an analysis. The odds comparing the methods pairwise demonstrate that some of the methods differ considerably.

The two test-beds both selected Jetset as the best single-method strategy for the MEU criterion. This happens despite the fact that both the mRNA expression platform and the proteomics platform are different between the two test-beds and the range of correlations is quite different as well. This result provides encouragement that the evaluation methodology applied here can produce best-practices conclusions that can be useful for external microarray data sets.

Jetset eliminates as many as 80% of probesets. This may seem extreme to a user of microarrays but considering that there are roughly three times the probesets as there are protein-coding genes in the human genome, if (as Jetset does) a method selects only the best probeset for each gene, then a minimum of two-third of the probesets must be eliminated. When the goal of the biological research requires more features than a strict filtering method like Jetset would allow, then Jetset would not be used. Our method reflects this by granting the user goal-specific utility values that will penalize false negatives more stringently. If increasing this penalty leads to an unacceptable number of false positives, then the research goal cannot be achieved, and it is best that the investigator know this.

In the ovarian data set, an investigator leery of discarding such a large proportion of probesets would be attracted to using EnCode, guided by our TEU criterion. In both test-beds, EnCode removes few probesets, but in the ovarian test-bed, the probesets removed are of especially poor quality. This may be related to the fact that the mass spectrometry platform in the endometrial test-bed is not designed for accurate quantification. In contrast, the RPPA platform uses selected validated antibodies so that one source of poor correlations is greatly reduced. Since RPPA data are a ligand-based local protein expression assay, the sensitivity for an individual protein is much higher than the LC-MS/MS data. This method is sensitive to correlation of mRNA expression to protein, and the RPPA data have a more reliable protein measurement at low protein expression.

Conclusions

Many investigators use publicly available data, such as the TCGA data warehouse, to unlock discoveries at the genome, transcript, and protein levels of cancer biology. Previously, in merging and analyzing data from an expression data set and proteomic data on the same samples, our team found startling differences in the identifier mapping services. We developed a principled, data-grounded method to evaluate and compare these services. This method has broad generalizability to evaluating many kinds of data pipeline choices and strategies, including identifier filtering methods and read filtering methods to remove erroneous or poor quality features, and tuning parameter settings in pipelines. We are developing a new package that will support much wider applications to all kinds of workflow options. This package will include the decision theory component as well.

In conclusion, the evaluation methodology applied here has some major virtues. First, the identifier filtering decisions are based on real, not simulated, data. In addition, these results can be subject to replication independently on multiple test-beds. Lastly, the identifier filtering methodology is responsive to the needs of investigators through the decision theory framework, which helps an investigator decide how much data to filter away based on mRNA to protein correlation.

Acknowledgments

The results shown here are in whole or part based upon data generated by the TCGA Research Network (http://cancergenome.nih.gov/). Thank you to Amyjo McDade for contribution to English copy edit.

Author Contributions

Conceived and designed the experiments: KKM, UC, RSD. Analyzed the data: KKM. Wrote the first draft of the manuscript: KKM. Contributed to the writing of the manuscript and graphics: KKM, RSD. Agreed with manuscript results and conclusions: KKM, UC, RSD. Jointly developed the structure and arguments for the paper: KKM, UC, RSD. Made critical revisions and approved the final version: KKM, UC, RSD. All authors reviewed and approved the final manuscript.

REFERENCES

1. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002; 30(1):207–10.
2. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2(5):401–4.
3. Akbani R, Ng PK, Werner HM, et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat Commun*. 2014;5:3887.
4. Day RS, McDade KK, Chandran UR, et al. Identifier mapping performance for integrating transcriptomics and proteomics experimental results. *BMC Bioinformatics*. 2011;12(1):213.
5. The Encode Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*. 2011;9(4):e1001046.
6. Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC. Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics*. 2011;12(1):474.
7. Nurtdinov RN, Vasiliev MO, Ershova AS, Lossev IS, Karyagina AS. PLAN-dbAffy: probe-level annotation database for Affymetrix expression microarrays. *Nucleic Acids Res*. 2010;38(Database issue):D726–30.
8. Sleuthing with the affymetrix netaffx ™ website: identifing and examing probe sets and their genomic context. *Affymetrix Genechip IVT Array White Paper Collection*. Available at: http://www.affymetrix.com/support/technical/whitepapers/Sleuthing_NetAffx_whitepaper.pdf. 2003.
9. Ferrari F, Bortoluzzi S, Coppe A, et al. Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics*. 2007;8:446.
10. Day RS, McDade KK. A decision theory paradigm for evaluating identifier mapping and filtering methods using data integration. *BMC Bioinformatics*. 2013; 14(1):223.
11. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004; 5(10):R80.

12. Lisovich A, Day RS. The IdMappingAnalysis Package in Bioconductor: Critically Comparing Identifier Maps Retrieved from Bioinformatics Annotation Resources. Version 1.2.1 Bioconductor Release 211; 2012:1–18. Available at: http://www.bioconductor.org/packages/release/bioc/html/IdMappingAnalysis.html.
13. Iorns E, Lord CJ, Grigoriadis A, et al. Integrated functional, gene expression and genomic analysis for the identification of cancer targets. *PLoS One*. 2009;4(4):e5120.
14. Irmler M, Hartl D, Schmidt T, et al. An approach to handling and interpretation of ambiguous data in transcriptome and proteome comparisons. *Proteomics*. 2008;8(6):1165–9.
15. Fu X, Fu N, Guo S, et al. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*. 2009;10:161.
16. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509–17.
17. Iorio MV, Croce CM. MicroRNAs in cancer: small molecules with a huge impact. *J Clin Oncol*. 2009;27(34):5848–56.
18. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. *Nature*. 2008;455(7209):64–71.
19. Chen G. Discordant protein and mRNA expression in lung adenocarcinomas. *Mol Cell Proteomics*. 2002;1(4):304–13.
20. Gao J. *Correlating Protein Phosphorylation with Genomic Alterations in Cancer RPPA: Reverse Phase Protein Arrays. TCGA Meet*. Available at: https://www.genome.gov/Multimedia/Slides/TCGA1/TCGA1_Gao.pdf. 2011.
21. The Cancer Genome Atlas Consortium. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474(7353):609–15.
22. Lisovich A, Day RS. *IdMappingRetrieval: Id Mapping Data Retrieval*; 2013. Available at: http://www.bioconductor.org/packages/2.12/bioc/manuals/IdMappingRetrieval/man/IdMappingRetrieval.pdf.
23. Kahlem P, Birney E. ENFIN a network to enhance integrative systems biology. *Ann N Y Acad Sci*. 2007;1115(0):23–31.
24. Maxwell GL, Hood BL, Day R, et al. Gynecologic oncology proteomic analysis of stage I endometrial cancer tissue: identification of proteins associated with oxidative processes and inflammation. *Gynecol Oncol*. 2011;121(3):586–94.
25. Risinger JI, Allard J, Chandran U, et al. Gene expression analysis of early stage endometrial cancers reveals unique transcripts associated with grade and histology but not depth of invasion. *Front Oncol*. 2013;3:1–10.
26. Tibes R, Qiu Y, Lu Y, et al. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol Cancer Ther*. 2006;5(10):2512–21.
27. Transcript Assignment for NetAffx Annotations. *Affymetrix Genechip IVT Array White Paper Collection*; 2006:1–9.
28. Zhang J, Finney R, Beutow K. Custom Chip Definition Files (CDF) for Unified Gene Expression Analysis with Affymetrix Chips; 2005. Available at: http://masker.nci.nih.gov/ev/.
29. Harrow J, Denoeud F, Frankish A, et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol*. 2006;7(suppl 1): S4.1–9.
30. Szumilas M. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*. 2010;19(3):227–9.

Appendix: The Derivative of the Variance of the Posterior Probability

We suppose that the correlations MQ_p are from an approximate mixture of normal distributions with means φ_+ and φ_-, and variances V_+ and V_-. The mean of the zero-correlation component, φ_-, is fixed at zero, and the other parameters are fitted with an ECM (expectation-conditional-maximization) method. The posterior probability that pair p belongs to group "+" is obtained from Bayes Theorem in its formulation in terms of prior and posterior odds.

$$\frac{\pi^*_{+p}}{\pi^*_{-p}} \doteq \frac{\pi_+}{\pi_-} \frac{\exp\left(-(MQ_p-\varphi_+)^2/2(V_+ + \sigma_p^2)\right)/\sqrt{V_+ + \sigma_p^2}}{\exp\left(-(MQ_p-\varphi_-)^2/2(V_- + \sigma_p^2)\right)/\sqrt{V_- + \sigma_p^2}}$$

Where σ_p^2, the sampling variance for pair p is estimated from the bootstrap.

To apply the delta method to the variance of the posterior probability, we apply the differential operator D to the posterior odds:

$$\begin{aligned}
&D\left(\frac{\pi^*_{+p}}{\pi^*_{-p}}\right)\\
&= D\left(\frac{\pi_+}{\pi_-}\frac{\exp\left(-(MQ_p-\varphi_+)^2/2(V_+ + \sigma_p^2)\right)/\sqrt{V_+ + \sigma_p^2}}{\exp\left(-(MQ_p-\varphi_-)^2/2(V_- + \sigma_p^2)\right)/\sqrt{V_- + \sigma_p^2}}\right)\\
&= \frac{\pi^*_{+p}}{\pi^*_{-p}} D\left(-\left(MQ_p-\varphi^{*2}_+\right)/2\left(V^*_+ + \sigma_p^2\right) + \left(MQ_p-\varphi_-\right)^2/2\left(V_- + \sigma_p^2\right)\right)\\
&= \frac{\pi^*_{+p}}{\pi^*_{-p}} D\left(MQ_p\right)\left(\frac{MQ_p-\varphi_-}{V_- + \sigma_p^2} - \frac{MQ_p-\varphi_+}{V_+ + \sigma_p^2}\right)
\end{aligned}$$

But also,

$$D\left(\frac{\pi^*_{+p}}{\pi^*_{-p}}\right) = \frac{\pi^*_{-p}D\pi^*_{+p} - \pi^*_{+p}D\pi^*_{-p}}{\left(\pi^*_{-p}\right)^2} = \frac{1}{\left(\pi^*_{-p}\right)^2} D\pi^*_{+p}.$$

Therefore

$$\begin{aligned}
\frac{d\pi^*_{+p}}{dMQ_p} &\doteq \left(\hat{\pi}^*_{-p}\right)^2 \frac{\hat{\pi}^*_{+p}}{\hat{\pi}^*_{-p}}\left(\frac{MQ_p-\hat{\varphi}_-}{\hat{V}_- + \sigma_p^2} - \frac{MQ_p-\hat{\varphi}_+}{\hat{V}_+ + \sigma_p^2}\right)\\
&= \hat{\pi}^*_{+p}\ \hat{\pi}^*_{-p}\left(\frac{MQ_p}{\hat{V}_- + \sigma_p^2} - \frac{MQ_p-\hat{\varphi}_+}{\hat{V}_+ + \sigma_p^2}\right)
\end{aligned}$$

We have confirmed this computation against the numerical derivative in R.

It is worth noting that, if σ_p^2 is very large, then changing the correlation changes the posterior probability very little. The slope is small. So ironically, a large measurement variance for the correlation means a small variance in the posterior probability, but in those cases, the posterior probability is near the prior probability. Again, this is confirmed by R code.