*Published online 23 November 2011*

*Nucleic Acids Research, 2012, Vol. 40, No. 2* ***e15***
*doi:10.1093/nar/gkr1071*

# MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis

**Dongwan D. Kang[1], Etienne Sibille[2], Naftali Kaminski[3] and George C. Tseng[1,4,5,*]**

[1]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, [2]Department of Psychiatry, [3]Dorothy P. and Richard P. Simmons Center for ILD, Division of Pulmonary, Allergy and Critical Care Medicine, [4]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260 and [5]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA 15261, USA

Received July 16, 2011; Revised October 21, 2011; Accepted October 28, 2011

## ABSTRACT

**Genomic meta-analysis to combine relevant and homogeneous studies has been widely applied, but the quality control (QC) and objective inclusion/exclusion criteria have been largely overlooked. Currently, the inclusion/exclusion criteria mostly depend on *ad-hoc* expert opinion or naïve threshold by sample size or platform. There are pressing needs to develop a systematic QC methodology as the decision of study inclusion greatly impacts the final meta-analysis outcome. In this article, we propose six quantitative quality control measures, covering internal homogeneity of coexpression structure among studies, external consistency of coexpression pattern with pathway database, and accuracy and consistency of differentially expressed gene detection or enriched pathway identification. Each quality control index is defined as the minus log transformed *P* values from formal hypothesis testing. Principal component analysis biplots and a standardized mean rank are applied to assist visualization and decision. We applied the proposed method to 4 large-scale examples, combining 7 brain cancer, 9 prostate cancer, 8 idiopathic pulmonary fibrosis and 17 major depressive disorder studies, respectively. The identified problematic studies were further scrutinized for potential technical or biological causes of their lower quality to determine their exclusion from meta-analysis. The application and simulation results concluded a systematic quality assessment framework for genomic meta-analysis.**

## INTRODUCTION

Microarray gene-expression technology provides detailed and parallel expression profiles of tens of thousands genes. Since its introduction, it has led to tremendous amount of data that are accumulated in the public repositories such as NCBI Gene Expression Omnibus (1), EBI ArrayExpress (2) and Stanford Microarray Database (3). While investigators can retrieve individual datasets and potentially compare their own results to related studies, such analyses often seem to yield inconsistent results and are hampered by variable technical quality, heterogeneous cohorts, erroneous data annotation or problematic preprocessing (4–7).

Meta-analysis, successfully applied in traditional epidemiological and medical research, has been proposed in analysis of microarray data across studies. Meta-analysis methods for detecting differentially expression genes include Fisher's method (8,9), Stouffer's method (10), LASSO (11), random effects model (12,13), Bayesian methods (14,15), rank-based methods (16,17) and others (18,19) have been implemented. In addition to differentially expressed (DE) gene detection, a statistical framework for microarray pathway meta-analysis was also proposed (20). While these studies provided significant methodological insights they in general did not address the important question of dataset selection and in general subjective expert opinions or *ad hoc* criteria were used (21–26). One of the most important obstacles to successful meta-analysis is dataset quality (27). Inclusion of a poor quality or outlying study in the information integration can greatly dilute information contained, weaken statistical power or even distort final biological conclusions. To alleviate such potential pitfalls in meta-analysis (27), it is necessary to develop an objective inclusion/exclusion evaluation approach.

*To whom correspondence should be addressed. Tel: +1 412 6245318; Email: ctseng@pitt.edu

**Table 1.** Summary information and characteristics of six QC measures

| Types | Evaluation Criteria | External Pathway Knowledge Needed? | Clinical Outcome Needed? |
|---|---|---|---|
| IQC | Homogeneity of coexpression structure across studies | No | No |
| EQC | Consistency of coexpression information with pathway database | Yes | No |
| AQCg | Accuracy of biomarker detection | No | Yes |
| AQCp | Accuracy of enriched pathway detection | Yes | Yes |
| CQCg | Consistency of DE gene ranking | No | Yes |
| CQCp | Consistency of enriched pathway ranking | Yes | Yes |

In this article, we propose quantitative measures to assess the quality and consistency of microarray studies for meta-analysis. Specifically, we introduce six quality control (QC) measures (see Table 1 for a brief summary) and utilize principal component analysis (PCA) biplots and a standardized mean rank (SMR) summary score to assist identification of problematic studies. We then apply the proposed methods to four examples, each containing 7 brain cancer studies, 9 prostate cancer studies, 8 idiopathic pulmonary fibrosis (IPF) studies and 17 major depressive disorder (MDD) studies. We assess the impacts and effectiveness of the proposed inclusion/exclusion evaluation on the final meta-analysis results. Finally, we demonstrate the robustness and effectiveness of the proposed method by additional simulations. To our knowledge, this is the first systematic and objective quality assessment tool developed to decide inclusion/exclusion criteria for genomic meta-analysis.

## MATERIALS AND METHODS

### Objective quality control measures

*Internal quality control index*. In the first criterion, the internal homogeneity of coexpression structure among studies was evaluated as an internal quality control (IQC) index. IQC compared pair-wise differences among studies in an unsupervised manner (without any prior or external information other than the expression profile data) and the aim was to identify potentially inconsistent or outlier studies from quantified coexpression dissimilarity. We applied a concept of the correlation of correlations that was previously reported in the context of reproducibility analysis of gene coexpression patterns across studies, named as integrative correlation coefficients (32,33). We assumed $K$ studies to be combined. For a given study $k$, we defined $\rho_{kij} = cor(x_{ki}, x_{kj})$ as the Pearson correlation coefficient of gene-expression intensities between gene $i$ and gene $j$ in study $k$. The similarity between two studies $m$ and $n$ was defined as $r_{mn} = spcor\big((\rho_{mij}; 1 \le i \le j \le G), (\rho_{nij}; 1 \le i \le j \le G)\big)$, which was the Spearman's rank correlation of the pairwise correlation structure between study $m$ and $n$ ($G$ represents the total number of genes in the studies). The dissimilarity (or distance) between study $m$ and $n$ was defined as $d_{mn} = (1–r_{mn})/2$. For a given study $k$, we considered the set of distances from all other studies to the study $k$ (i.e. $\tilde{D}_k^* = \{d_{kn}\}_{1 \le n \le K; n \ne k}$) and the set of all pairwise distance that do not involve study $k$ (i.e. $\tilde{D}_k^{\#} = \{d_{mn}\}_{1 \le m \ne n \le K; m \ne k, n \ne k}$). When study $k$ was an outlying study that contained coexpression structure very different from all other studies, the distances in $\tilde{D}_k^*$ were generally much greater than those in $\tilde{D}_k^{\#}$. We assumed that the two sets of distances follow certain probability distributions: $\tilde{D}_k^* \sim \mathcal{F}_1$ and $\tilde{D}_k^{\#} \sim \mathcal{F}_2$. We performed a formal hypothesis testing based on $H_0 : \mathcal{F}_1 = \mathcal{F}_2$ **vs.** $H_a : \mathcal{F}_1 > \mathcal{F}_2$ and applied one-sided Wilcoxon rank-sum (a.k.a. Mann–Whitney U) test (34) to generate a $P$-value, $P_{IQC}(k)$. Figure 1A shows an example that study 1 has a very different coexpression structure from other three studies. When we compare $\tilde{D}_1^* = (d_{12}, d_{13}, d_{14})$ and $\tilde{D}_1^{\#}(d_{23}, d_{24}, d_{34})$ by Wilcoxon rank-sum test, we obtained a small $P$-value, $P_{IQC}(1)$ that rejects the null hypothesis.

The hypothesis testing described above gave a small $P$-value when study $k$ was an outlying study. We applied a reverse transformation $g(p)$ on $P_{IQC}(k)$ such that small $P$-values would be transformed to large pseudo $P$-values and vice versa. Consequently, large transformed $g(p)$ corresponded to an outlying study. The transformation was necessary for IQC to be consistent with the remaining five QC measures to be introduced later. We designed $g(p)$ as a monotone decreasing function and the statistical significance threshold 0.05 an invariant point. Specifically, we defined $g$ as $g(p) = 1 - \mathcal{F}_{D_2}(\mathcal{F}_{D_1}^{-1}(p))$, where $D_1 \sim \mathcal{N}(z_{.95}, 1)$ and $D_2 \sim \mathcal{N}(-z_{.95}, 1)$ (Figure 1B). For example, $g(0.05) = 0.05$, $g(0.5) = 0.0005$ and $g(0.01) = 0.17$. Finally, the IQC measure of study $k$ was defined as $IQC(k) = -log_{10}g\big(P_{IQC}(k)\big)$. We use log base 10 for all QC measures throughout this article. Small IQC indicated that the study had heterogeneous coexpression structure with other studies and was considered a candidate problematic study that should be excluded from meta-analysis.

*External quality control index*. Compared to the unsupervised approach in IQC, the external quality control (EQC) criterion was supervised by external pathway information. Pathway knowledge (i.e. functional or coregulated gene sets) obtained from established databases (e.g. KEGG, GO, Biocarta and MSigDB) was applied to evaluate its consistency with a given study and subsequently to determine the study quality. We used a similar gene-pair correlation concept used in IQC and defined an

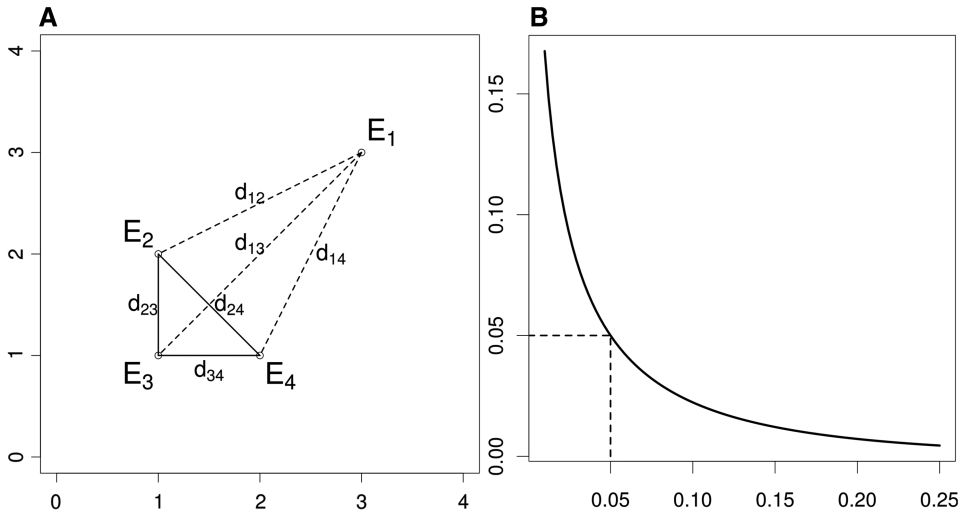


**Figure 1.** Example of IQC calculation and reverse transformation of $P_{IQC}$. (**A**) Three points (E2, E3 and E4) in the lower left represent homogeneous studies, and a point (E1) in the upper right is a heterogeneous study which has larger pair-wise distance to others. A heterogeneous study should have larger pair-wise distances with others. The IQC hypothesis setting compares ($d_{12}$,$d_{13}$,$d_{14}$) and ($d_{23}$,$d_{24}$,$d_{34}$) by Wilcoxon rank-sum test. (**B**) *X*- and *Y*-axes are *P*-values before and after applying the reverse transformation *g*. As a result, small *P*-values will be transformed to large pseudo *P*-values and vice versa.

association measure between study $k$ and a given pathway (gene set) $w$ by

$$t_k = t_k\left(\{\rho_{kij}\}_{1\le i,j\le|G_k|}; w\right) = \left(\frac{\sum_{i\neq j; i,j\in w}|\rho_{kij}|^l}{|w|\cdot(|w|-1)/2}\right)^{1/l} \Big/ \left(\frac{\sum_{1\le i<j\le|G_k|}|\rho_{kij}|^l}{|G_k|\cdot(|G_k|-1)/2}\right)^{1/l}$$

where $\rho_{kij}$ was the Pearson correlation coefficient of gene $i$ and gene $j$ in study $k$ as defined in IQC, the numerator was the $l$-norm average of absolute pairwise correlation in pathway $w$, the denominator was the corresponding $l$-norm average in the background genome $G_k$, and $|w|$ and $|G_k|$ were the number of genes in the pathway $w$ and study $k$. If pathway $w$ was relevant to disease status or experimental perturbation, we expected that the l-norm average among the pathway in the numerator would be much larger than that among genome background in the denominator and $t_k$ should be significantly $>1$. In this association measure, we disregarded the sign of correlation coefficients and used $l$-norm to inflate differential impact of high and low correlations in the measure. We use $l = 2$ throughout the article to down-weight medium to low correlation coefficients and to give higher relative weight to large correlation coefficients (e.g. $0.8^2 = 0.64$ and $0.3^2 = 0.09$). We set up hypothesis testing $H_0$:$t_k = 1$ vs. $H_a$:$t_k > 1$ and applied Monte-Carlo permutation analysis to obtain the empirical null distribution of the test statistic $t_k$ (35,36). Specifically, we randomly sampled from $G_k$ a random pathway $w^{(b)}$ of equal size (i.e. $|w^{(b)}| = |w|$) in the b$^{th}$ simulation, calculated the corresponding $t_k^{(b)}$ and repeated for $B$ times ($b = 1, \ldots, B$). The resulting $P$-value of the test was calculated as $P_{EQC}(k; w) = (\sum_{b=1}^{B} I(t_k^{(b)} > t_k)+1)/(B+1)$, where $I(\bullet)$ was an indicator function. We adopted a conservative procedure to add 1 to both denominator and numerator in $P$-value calculation, considering the observed statistics was one of the simulated cases (36). The EQC measure was then defined as $EQC(k;w) = -logP_{EQC}(k;w)$. Similar to IQC, small $EQC(k;w)$ indicated that the study had low association with pathway $w$ in terms of gene pairwise correlation structure and was thus considered a candidate of problematic study.

We further extended the EQC measure above to a set of pathways. We assumed that $M$ pathways ($W = \{w_m, 1 \le m \le M\}$) were available and a significant portion of them had high association measure with study $k$. We defined a Fisher's score by $S_k = -2\sum_{m=1}^{M} logP_{EQC}(k; w_m)$ to aggregate the association measures of $M$ pathways. If the pathways were independent, the $S$ score followed a chi-squared distribution with degree of freedom $2M$ under the null hypothesis. However, since the biological pathways always have hierarchical structure and high overlapping, we performed permutation analysis for $B$ times to obtain simulated $S_k^{(b)}$. The resulting $P$-values was calculated as $P_{EQC}(k;W) = (\sum_{b=1}^{B} I(S_k^{(b)} > S_k)+1)/(B+1)$ and the EQC measure was similarly defined as $EQC(k;W) = -logP_{EQC}(k;W)$.

Comparing IQC and EQC, we note that EQC relied on a good selection of pathway set $W$ and the evaluation of one study was independent from other studies. IQC, on the other hand, was a relative measure that depended on other studies under consideration but did not require external biological information.

*Accuracy quality control (AQCg and AQCp) and consistency quality control (CQCg and CQCp) indexes.* For the third and fourth criteria, we proposed an accuracy quality control (AQC) and a consistency quality control (CQC) criteria that were aimed at quantifying the reproducibility (accuracy or consistency) of DE genes (or pathways) detected in an individual study compared to those detected by meta-analysis from all other studies. For AQCg of study $k$, the identified DE

gene list from meta-analysis excluding study $k$ (using Student's $t$-test for each individual study and Fisher's method to combine with Benjamini-Hochberg correction under FDR = $r$%) was served as a gold standard. The DE gene list detected by study $k$ (using Student's $t$-test with Benjamini-Hochberg procedure under FDR = $r$%) was then compared to the gold standard to generate a $2 \times 2$ table. One-sided Fisher's exact test was used to determine the association (reproducibility) of DE gene list identified by meta-analysis and that identified by study $k$ ($H_0$: the two gene lists have no association. versus $H_a$: the two gene lists have association). The $P$-value for study $k$ was calculated from hypergeometric distribution:

$$P_{AQCg}(k; r) = \sum_{t=t_k}^{min(T^{(k)}, T^{(-k)})} \frac{\binom{T^{(k)}}{t}\binom{G_k - T^{(k)}}{T^{(-k)} - t}}{\binom{G_k}{T^{(-k)}}}$$

where $G_k$ was the total number of genes in study $k$, $T^{(k)}$ was the number of DE genes detected by study $k$, $T^{(-k)}$ was the number of DE genes detected by meta-analysis excluding study $k$ and $t_k$ was the number of DE genes detected both by study $k$ and by meta-analysis excluding study $k$ (see the $2 \times 2$ table in Table 2). The AQCg score was defined as $AQCg(k;r) = -logP_{AQCg}(k;r)$. We normally used FDR threshold r% = 5% but could relax it to 10 or 20 when the data had weak signal. Large AQCg measure for a given study $k$ indicated that DE genes produced by study $k$ were reproducible compared to DE genes detected by meta-analysis excluding study $k$. We extended AQCg to AQCp where DE genes in the AQCg definition were replaced by enriched pathways. The pathway enrichment could be obtained by simple Fisher's exact test under certain DE gene threshold or other methods in the literature [e.g. GSEA (37) or GSA (38)]. In this article, we used Kolmogorov–Smirnov test under FDR = 5% threshold to obtain enriched pathways.

In contrast to evaluating DE gene lists from a hard threshold in AQCg, we also applied an alternative of CQC measure by evaluating the consistency of differential expression ranking from single study analysis and meta-analysis. Specifically, ranks of differential expression evidence of study $k$ were first calculated by Student's $t$-test and defined as $R_g^{(k)}$ for gene $g$ and study $k$. From meta-analysis (using Fisher's method) excluding study $k$, the ranks of differential expression evidences were denoted as $R_g^{(-k)}$. The Spearman rank correlation between two rank vectors was defined as $\rho_k = spcor((R_g^{(k)}; 1 \le g \le G_k), (R_g^{(-k)}; 1 \le g \le G_k)) = 1 - 6 \cdot \sum_{g=1}^{G_k}(R_g^{(k)} - R_g^{(-k)})^2 / G_k(G_k^2 - 1)$. To test $H_0$:$\rho_k = 0$ vs. $H_a$:$\rho_k > 0$, we approximated that $t = \rho_k \cdot \sqrt{G_k - 2/1 - \rho_k^2}$ followed a Student's t distribution with $G_k - 2$ degree of freedom under null hypothesis (39). The resulting $P$-value was calculated as $P_{\mathrm{CQCg}}(k) = 1 - \mathcal{F}_{G_k-2}\left(\rho_k \cdot \sqrt{G_k - 2/1 - \rho_k^2}\right)$, where $\mathcal{F}_{G_k-2}$ represented the cumulative distribution function (cdf) of Student's $t$-distribution with $G_k - 2$ degree of freedom. The CQCg score was defined as $CQCg(k) = -logP_{CQCg}(k)$. Having a large CQCg measure for a given study $k$ indicated that DE evidence produced by study $k$ was consistent with DE evidence generated by meta-analysis excluding study $k$. We similarly extended CQCg to CQCp where DE evidence and gene ranking in the CQCg definition were replaced by enriched pathways.

**Table 2.** Contingency table for AQC inference

| | DE genes detected by meta-analysis excluding study k | | sum |
|---|---|---|---|
| | yes | no | |
| DE genes detected by study k | | | |
| yes | $t_k$ | $T^{(k)} - t_k$ | $T^{(k)}$ |
| no | $T^{(-k)} - t_k$ | $G_k$-$T^{(-k)} - T^{(k)} + t_k$ | $G_k - T^{(k)}$ |
| sum | $T^{(-k)}$ | $G_k - T^{(-k)}$ | $G_k$ |

### Visualization and summarization to assist decision

We applied PCA biplots (40) to assist the visualization and decision for inclusion or exclusion of studies in meta-analysis. A PCA biplot is a popular technique to show both observations and relative positions of variables in two dimensions so that the performance of each observation can be interpreted by each variable intuitively. In this article, each microarray study was projected from 6D QC measures to a 2D PC subspace. The direction of each quality control measure was juxtaposed on top of the 2D subspace using arrows. Specifically, the coordinates of each quality criterion were determined by its correlation to the two driving PCs. The origin of the biplot was taken as the statistical threshold with Bonferroni correction [i.e. projected from −log(0.05/#studies) in each of the QC measure dimensions], suggesting that studies located in the opposite area of arrows were candidate outlier studies. The scale of each QC measure was standardized before PCA to avoid dominance of a particular QC measure due to scale problem. In addition to biplot visualization, we also defined a quantitative summary score by calculating the ranks of each QC measure among all studies and then computed a SMR of each study: (mean rank of all QC measures/# of studies). By definition, $0 < \mathrm{SMR} \le 1$ and large SMR represented a likely problematic study.

Note that our visualization and summarization tools were not meant for an automated recommendation for inclusion/exclusion decision. In the examples we explored, there were roughly three categories in the QC results: definite exclusion cases with poor quality, definite inclusion cases with good quality and borderline cases. Definite exclusion cases were often on the opposite side of arrows in the PCA biplots and had large SMR scores. These studies were strongly suggested to be excluded from meta-analysis. On the other hand, definite inclusion cases were on the same side of arrows in the PCA biplots and had small SMR scores. They were clearly of good quality that should be included. Borderline studies happened to be in between the two extreme cases. Although an automated

**Table 3.** Summary information of studies used in four examples

| Author | Year | Platform | Sample Size | Source |
|---|---|---|---|---|
| Brain cancer studies | | | | |
| Freije *et al.* (44) | 2004 | HG-U133A,B | 85 | GSE4412 |
| Phillips *et al.* (45) | 2006 | HG-U133A,B | 100 | GSE4271 |
| Sun *et al.* (46) | 2006 | HG-U133 Plus 2 | 100 | GSE4290 |
| Yamanaka *et al.* (47) | 2006 | Agilent | 29 | GSE4381 |
| Petalidis *et al.* (48) | 2008 | HG-U133A | 58 | GSE1993 |
| Gravendeel *et al.* (49) | 2009 | HG-U133 Plus 2 | 175 | GSE16011 |
| Paugh(50) | 2010 | HG-U133 Plus 2 | 42 | GSE19578 |
| Prostate cancer studies | | | | |
| Dhanasekaran *et al.* (51) | 2001 | cDNA | 28 | www.pathology.med.umich.edu |
| Welsh *et al.* (52) | 2001 | HG-U95A | 34 | public.gnf.org/cancer/prostate/ |
| Singh *et al.* (53) | 2002 | HG-U95Av2 | 102 | www.broad.mit.edu/ |
| Lapointe *et al.* (54) | 2004 | cDNA | 103 | GSE3933 |
| Yu *et al.* (55) | 2004 | HG-U95Av2 | 146 | GSE6919 |
| Varambally *et al.* (56) | 2005 | HG-U133 Plus 2 | 13 | GSE3325 |
| Nanni *et al.* (57) | 2006 | HG-U133A | 30 | GSE3868 |
| Tomlins *et al.* (58) | 2006 | cDNA | 57 | GSE6099 |
| Wallace *et al.* (59) | 2008 | HG-U133A2 | 89 | GSE6956 |
| IPF Studies | | | | |
| Pardo *et al.* (60) | 2005 | Codelink | 24 | GSE2052 |
| Yang *et al.* (61) | 2007 | Agilent 43 K | 29 | GSE5774 |
| Larsson *et al.* (62) | 2008 | HG-U133 Plus 2 | 12 | GSE11196 |
| Vuga *et al.* (63) | 2009 | Codelink | 7 | GSE10921 |
| Konishi *et al.* (64) | 2009 | Agilent 4x44K | 38 | GSE10667 |
| Emblom *et al.* (65) | 2010 | cDNA | 58 | GSE17978 |
| KangA | 2011 | Agilent 4x44K | 63 | Dr Kaminski |
| KangB | 2011 | Agilent 8x60K | 96 | Dr Kaminski |
| MDD studies | | | | |
| MD1_AMY | 2009 | HG-U133 Plus 2 | 28 | Dr Sibille |
| MD3_AMY | 2009 | HumanHT-12 | 42 | Dr Sibille |
| MD1_ACC | 2009 | HG-U133 Plus 2 | 32 | Dr Sibille |
| MD3_ACC | 2009 | HumanHT-12 | 44 | Dr Sibille |
| MD2_ACC_M | 2010 | HG-U133 Plus 2 | 18 | Dr Sibille |
| MD2_ACC_F | 2010 | HG-U133 Plus 2 | 26 | Dr Sibille |
| MD2_DLPFC_M | 2010 | HG-U133 Plus 2 | 28 | Dr Sibille |
| MD2_DLPFC_F | 2010 | HG-U133 Plus 2 | 32 | Dr Sibille |
| NY_DLPFC_M | 2004 | HG-U133A | 26 | Dr Sibille |
| NY_oFC_M | 2004 | HG-U133A | 24 | Dr Sibille |
| Feinberg | - | HG-U95Av2 | 27 | www.stanleygenomics.org |
| KatoB | 2004 | HG-U95Av2 | 26 | www.stanleygenomics.org |
| Kemether | - | HG-U133p | 24 | www.stanleygenomics.org |
| AlartC | - | HG-U133A | 22 | www.stanleygenomics.org |
| SklarA | - | HG-U95Av2 | 23 | www.stanleygenomics.org |
| SklarB | - | HG-U95Av2 | 23 | www.stanleygenomics.org |
| Sokolov | - | HG-U95A | 26 | www.stanleygenomics.org |

quantitative decision looks desirable, it is not practical in general. One should seek additional qualitative evidences (such as sample size, platform or other experimental conditions) for the causes of poor quality in both definite exclusion or borderline studies.

## Application, implementation and simulation in real datasets

We evaluated our proposed method in four examples: brain cancer (seven studies), prostate cancer (nine studies), IPF (eight studies) and MDD (17 studies). Details of these studies were listed in Table 3. Most microarray data sets were collected from public repositories such as NCBI Gene Expression Omnibus (1) and EBI ArrayExpress (2), or web pages directed from the original articles. Several non-published data sets were obtained from labs of coauthors of this article (Dr Kaminski and Dr Sibille). Most data sets were preprocessed and normalized by original authors. When raw data of Affymetrix platform were available, RMA (41) was applied for preprocessing. To obtain a robust result, we applied a gene filtering procedure in each study level, which removed 40% of non-expressed genes based on mean intensities and 40% of non-informative genes based on variance. Gene matching across studies was done by matching official gene symbols using Bioconductor packages. When multiple probes matched to one gene symbol, the probeset with the largest inter-quartile range (IQR) was selected.

In EQC evaluation, external pathways were needed for calculating EQC measures. We only considered pathways that have at least five genes in each study. Conceptually, using pathways relevant to the disease or experimental perturbation would generate better EQC evaluation. For

cancer studies, we chose to use GSEA Biocarta v3.0 pathways (37) since the pathways were cancer specific. A total of 217 Biocarta pathways were used in the brain cancer example. For prostate cancer studies, the overall data quality and information seemed to be weaker than brain cancer studies and we chose only the top 50 pathways among the 217 pathways for better performance (top pathways were identified by combining *P*-values using Fisher's method). For MDD studies, 99 pathways were selected from GSEA MSigDB v3.0 by keyword search using a list of MDD relevant terms: GABA, INSULIN, DIABETES, IMMUNE, THYROID, ESTROGEN, DEPRESSION, AGING, ALZHEIMERS, PARKINSONS and HUNTINGTONS. For IPF studies, we chose top 50 pathways out of all 6769 number of GSEA MSigDB v3.0 pathways (similar to prostate cancer application, top pathways were identified using Fisher's method). For AQCp and CQCp measures, pathway database was also needed to generate enriched pathways before evaluation. Since exhaustive pathway enrichment analysis is usually preferred, we used all MSigDB c2 v3.0 pathways for both AQCp and CQCp in all four examples.

We performed 100 000 simulations in the permutation analysis of Fisher scores in EQC measure and thus the largest range of EQC measure is limited to 5 (that corresponds to $P = 1E{-}5$). For AQC measures, we applied two-sample Student's *t*-test and Kolmogorov–Smirnov test for AQCg and AQCp, respectively. All *P*-values were adjusted by Benjamini–Hochberg procedure (42) to control FDR at the level of 0.05 unless otherwise specified. Fisher's method (sum of minus log-transformed *P*-values) was used for meta-analysis in both AQC and CQC evaluation when performing meta-analysis of all studies except for the study *k*. In AQC measures, MDD was found a weak signal example that generated only very few DE genes or pathways that made AQC measure invalid or unstable. We chose a more liberal cutoff (unadjusted $P < 0.05$) to avoid the issue.

To assess the validity and performance of our proposed method, we performed downstream analysis to assess its impact on DE gene and pathway detection. We also performed simulation to assess the accuracy of detecting problematic studies. All implementation was written by *R* statistical language (43). An *R* package, 'MetaQC' is publicly available online at CRAN (http://cran.r-project.org/)

## RESULTS

### Quality assessment in four examples

Table 3 lists summary information of studies used in four examples: 7 brain cancer studies, 9 prostate cancer studies, 8 IPF studies, and 17 MDD studies. The different QC measures and SMR scores were obtained as described in details in the Methods section and are summarized in Table 4. Together with PCA biplots in Figure 2, studies were categorized into three sets: definite exclusion cases with poor quality, definite inclusion cases with good quality and borderline cases (refer to 'Materials and Methods' section for details in PCA biplots). These recommendations were then verified by consulting Table 3 for potential causes or interpretations of the problematic studies.

In the first brain cancer example, Dreyfuss *et al*. (21) previously combined four studies for meta-analysis, of which three were used in our evaluation. Figure 2A shows PCA biplot of the brain cancer result and Table 4 shows the detailed QC measures and SMR scores. The first two PCs in Figure 2A explained ~92% of total variance, and all scores were highly correlated with the first PC. The scores marked with asterisks in Table 4 indicated non-statistical significance ($P > 0.05/\#$ of studies), meaning that these studies were candidate of problematic studies, based on the specific QC measure, and including them might have an adverse effect on meta-analysis. The Yamanaka study (study 7 in Figure 2A) was clearly below statistical threshold and had low values in all QC measures; it is viewed as a definite exclusion case that should be excluded from the meta-analysis. On the other hand, the top five studies in Table 4 performed very well for all criteria, indicating that they are definite inclusion cases for meta-analysis. The Paugh study (study 6 in Figure 2A), was however a borderline case. The QC measures were mostly low and just passed the statistical significance. Interestingly, when scrutinizing the causes of poor quality of Yamanaka and Paugh studies, Yamanaka used a different platform (Agilent) and both studies were of smaller sample size ($n = 29$ for Yamanaka and $n = 42$ for Paugh). We thus recommend exclusion of both studies from meta-analysis.

In the second example, we applied the QC assessment to nine prostate cancer studies comparing normal and primary cancer patients (see Table 3 for summary information). QC results were shown in Figure 2B and Table 4. Compared to brain cancer studies, these prostate cancer studies were mostly performed in earlier years with older array platforms. Although the first two PCs also captured high percentage of variance (93%), the studies were more scattered in the biplot and even good performing studies had quite different performance when judged by different QC criteria. For example, Varambally and Wallace had better scores in IQC and EQC but not in CQC and AQC while Welsh, Lapointe and Singh, had better performance in CQC and AQC but not IQC and EQC. Yu had performed the best in all criteria. In considering sample size, array platform and QC measures, we regarded the bottom three studies: Nanni, Tomlins and Dhanasekaran as definite exclusion cases and marked Singh as a borderline case. The worse performance of prostate cancer studies compared to brain cancer shown here reflects the fact that many prostate cancer studies were performed using cDNA arrays or earlier platforms and that the cancer is a heterogeneous disease (28). Here, the overall scatter of SMR values suggests a limited potential for meta-analysis.

As a third example, we evaluated eight IPF studies which identified signature genes of IPF patients compared to normal. IPF is one of the most lethal chronic lung disease, and its mean survival is only 3–5 years regardless of treatment (29). Table 3 shows data summary, Figure 2C demonstrates the PCA biplot and

**Table 4.** Results of quality control (QC) measuresand SMRs (SMR)

| Number | Study | IQC | EQC | CQCg | CQCp | AQCg | AQCp | SMR |
|---|---|---|---|---|---|---|---|---|
| Brain cancer studies | | | | | | | | |
| 1 | Sun | 4.96 | 5.00 | 307.65 | 251.33 | 152.83 | 108.37 | 1.58 |
| 2 | Freije | 5.42 | 5.00 | 239.31 | 158.73 | 118.06 | 81.62 | 2.75 |
| 3 | Petalidis | 4.11 | 3.16 | 274.25 | 171.48 | 111.27 | 101.10 | 3.33 |
| 4 | Phillips | 4.52 | 5.00 | 242.36 | 146.71 | 106.59 | 69.19 | 3.58 |
| 5 | Gravendeel | 6.64 | 4.70 | 98.37 | 107.06 | 47.67 | 63.89 | 4.33 |
| 6 | Paugh | 1.51* | 5.00 | 5.00 | 3.60 | 2.31 | 9.84 | 5.42 |
| 7 | Yamanaka | 0.10* | 0.78* | 1.69* | 2.26 | 1.58* | 0.71* | 7.00 |
| Prostate cancer studies | | | | | | | | |
| 1 | Welsh | 5.04 | 2.12* | 68.59 | 101.31 | 26.46 | 54.66 | 2.00 |
| 2 | Yu | 7.74 | 3.23 | 52.43 | 64.14 | 19.95 | 38.30 | 2.17 |
| 3 | Lapointe | 4.06 | 2.28 | 26.36 | 59.42 | 7.00 | 33.90 | 3.50 |
| 4 | Varambally | 4.68 | 4.70 | 15.38 | 21.15 | 4.18 | 13.21 | 4.50 |
| 5 | Singh | 2.14* | 2.05* | 19.60 | 28.74 | 4.61 | 24.17 | 4.83 |
| 6 | Wallace | 7.95 | 4.22 | 0.00* | 28.70 | 0.00* | 2.13* | 5.67 |
| 7 | Nanni | 1.92* | 1.92* | 2.22* | 6.01 | 2.00* | 13.61 | 6.67 |
| 8 | Tomlins | 2.67 | 0.52* | 3.76 | 3.65 | 1.19* | 6.12 | 7.17 |
| 9 | Dhanasekaran | 0.01* | 0.63* | 0.01* | 0.23* | 0.04* | 0.10* | 8.50 |
| IPF studies | | | | | | | | |
| 1 | KangA | 6.64 | 5.00 | 307.65 | 146.87 | 96.71 | 90.88 | 1.58 |
| 2 | KangB | 5.57 | 5.00 | 273.67 | 114.30 | 84.37 | 69.74 | 2.42 |
| 3 | Konishi | 6.89 | 5.00 | 58.19 | 42.70 | 25.50 | 57.20 | 2.92 |
| 4 | Yang | 4.34 | 5.00 | 41.70 | 56.35 | 14.20 | 29.43 | 3.92 |
| 5 | Pardo | 4.07 | 2.08* | 25.14 | 38.84 | 20.60 | 25.05 | 5.17 |
| 6 | Vuga | 2.28 | 5.00 | 1.37* | 26.25 | 1.77* | 18.01 | 5.58 |
| 7 | Larsson | 1.79* | 5.00 | 0.59* | 1.88* | 0.52* | 3.21 | 6.58 |
| 8 | Emblom | 0.03* | 1.12* | 0.83* | 0.57* | 0.43* | 1.98* | 7.83 |
| MDD studies | | | | | | | | |
| 1 | MD2_ACC_F[a] | 9.80 | 5.00 | 19.22 | 40.01 | 6.17 | 27.47 | 3.58 |
| 2 | MD2_DLPFC_M[a] | 3.22 | 5.00 | 56.70 | 41.02 | 9.37 | 33.27 | 3.58 |
| 3 | MD2_DLPFC_F[a] | 3.05 | 1.12* | 52.05 | 62.94 | 14.32 | 46.79 | 5.17 |
| 4 | MD2_ACC_M[a] | 3.76 | 3.05 | 24.59 | 33.78 | 3.80 | 17.16 | 5.67 |
| 5 | MD1_ACC[a] | 3.41 | 5.00 | 10.59 | 19.28 | 0.39* | 10.21 | 6.92 |
| 6 | NY_oFC_M[a] | 11.56 | 3.74 | 0.12* | 18.09 | 0.40* | 13.32 | 7.67 |
| 7 | Kemether[b] | 8.01 | 1.91* | 12.21 | 8.92 | 9.79 | 1.63* | 8.83 |
| 8 | MD3_AMY[a] | 0.96* | 5.00 | 3.23 | 12.03 | 1.54* | 7.05 | 8.92 |
| 9 | NY_DLPFC_M[a] | 4.05 | 5.00 | 1.63* | 14.82 | 0.30* | 6.61 | 8.92 |
| 10 | MD3_ACC[a] | 1.37* | 4.70 | 8.70 | 15.65 | 1.80* | 4.06 | 9.17 |
| 11 | MD1_AMY[a] | 3.09 | 2.97 | 1.49* | 17.14 | 0.39* | 16.76 | 9.33 |
| 12 | KatoB[b] | 11.54 | 5.00 | 0.00* | 1.46* | 0.45* | 2.00* | 9.92 |
| 13 | Sokolov[b] | 4.07 | 0.30* | 0.46* | 1.40* | 0.60* | 6.85 | 11.00 |
| 14 | SklarB[b] | 0.73* | 5.00* | 0.00* | 9.71 | 0.00* | 8.80 | 11.58 |
| 15 | Feinberg[b] | 0.35* | 5.00 | 0.32* | 2.41* | 0.17* | 0.77* | 13.08 |
| 16 | AltarC[b] | 0.69* | 0.08* | 0.00* | 15.95 | 0.00* | 0.91* | 14.67 |
| 17 | SklarA[b] | 1.20* | 1.93* | 0.00* | 1.01* | 0.00* | 2.41* | 15.00 |

[a]Data from Dr Etienne Sibille's lab.
[b]Data from Stanley Foundation, suspected worse quality in the tissue collection and processing.
*$p$-value not significant after Bonferroni correction (i.e. $p > 0.05/\#$ of studies).

Table 4 lists the details of QC scores. Interestingly, although these eight data sets are mostly from very different microarray platforms, at least five of them performed very well in quality assessment for meta-analysis. Of the three worst QC studies, Emblom utilized a custom cDNA array platform which might be the origin of the weaker performance. Vuga and Larsson both have small sample sizes ($n = 7$ for Vuga and $n = 12$ for Larsson) which might be the reasons of low QC scores. The two top studies, KangA and KangB, are unpublished data from Dr Kaminski's lab with large well-characterized cohorts from the Lung Tissue Resource Consortium (LTRC; www.ltrcpublic.com). In this case, it is adequate to remove the three low quality studies and perform meta-analysis of the remaining five.

In the final example, we applied QC evaluation to 17 MDD studies that compare normal and MDD patients. These 17 studies were obtained from post-mortem brain tissues of various brain regions. These datasets are heterogeneous and of small sample size, and are typically considered of weak disease signal, hence highlighting the need for upfront meta-QC for inclusion in meta-analysis. The details of each data set were in Table 3. The QC results were shown in Figure 2D and Table 4. In Figure 2D, noticeably many studies scattered near the origin because of overall weak signal. From Table 4, the top five studies were considered as definite inclusion studies and the bottom five studies were definite exclusion studies. Other studies were borderline cases with varying performance in different QC measures. Most CQCg and AQCg scores

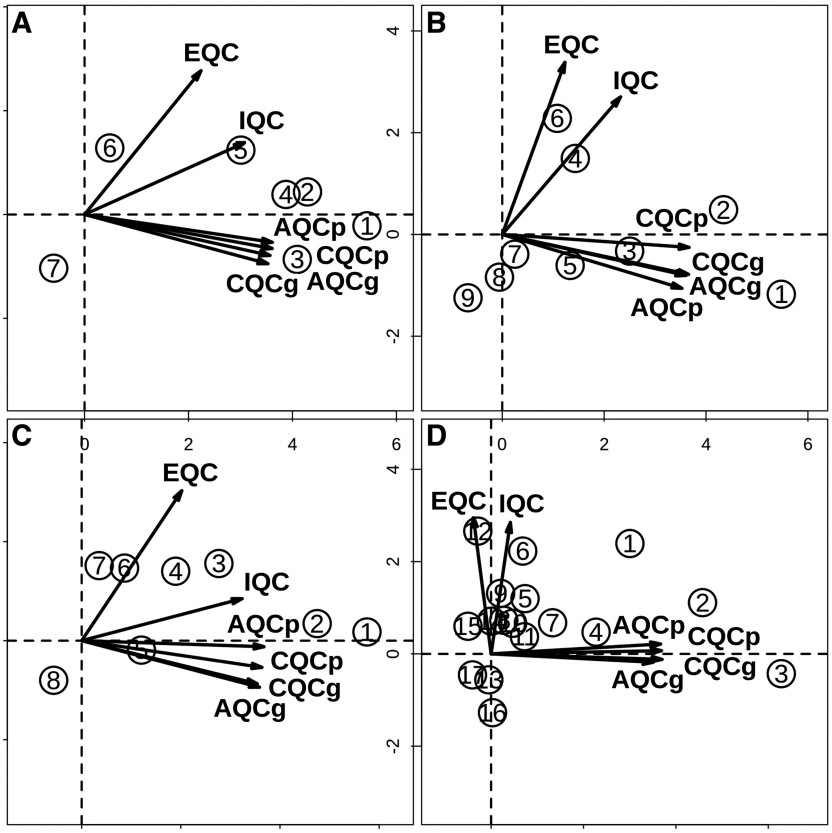


**Figure 2.** PCA biplots of QC measures in four examples. Each circled number represents the overall rank by SMR score of a study. Smaller numbers correspond to higher quality studies. (**A**) Seven brain cancer studies. (**B**) Nine prostate cancer studies. (**C**) Eight IPF studies. (**D**) Seventeen MDD studies.

were significantly lower than other examples since each individual MDD study contained weak signal and the DE gene and pathway detections were relatively less reproducible, which actually argued strong needs for meta-analysis. We note that the bottom six studies in Table 4 were all from Stanley Medical Research Institute Tissue Bank. This separation is not reflecting differences in subject cohorts, but rather lower quality as reflected by overall low evaluation criteria in the latter studies. This may reflect technical issues relating to tissue collection and processing, such as uneven postmortem interval of brain collection and low brain pH in the Stanley Tissue Bank cohorts (30,31). These results suggest that including the bottom low quality studies in a meta-analysis may weaken overall results for technical rather than biological reasons.

### Impacts on DE gene and pathway detection

To evaluate the impact of our MetaQC evaluation on the identification of biological effects, we investigated the marginal impact of a meta-analysis on DE gene and enriched pathway detection when we sequentially included studies from high to low quality into meta-analysis, as measured by SMR scores. We hypothesized that including an additional informative study to the meta-analysis would provide increased statistical power to detect more DE genes and enriched pathways while adding a lower quality study would deteriorate the performance, as manifested by fewer or stable numbers of detected DE genes (Figure 3) or biological pathways (Figure 4). Figures 3A and 4A show the number of DE genes and enriched pathways detected <0.5% FDR (false discovery rate; the ratio of falsely rejected null hypotheses among all rejected hypotheses in multiple testing), respectively, when seven brain cancer studies were added sequentially in the meta-analyses in the order of SMR score. Interestingly, the number of detected DE genes and pathways dropped significantly when including the two suspect problematic studies: Paugh and Yamanaka. The result supported the recommendation provided by MetaQC. This simple incremental analysis also argues for the necessity of adequate inclusion/exclusion criteria in meta-analysis.

The results for prostate cancer (Figures 3B and 4B) and IPF examples (Figures 3C and 4C) demonstrated a slightly different situation. The number of DE genes under FDR = 0.1% (a very stringent FDR is used here as both examples detect many DE genes) continued to increase as more studies were added while the number of detected pathways decreased when the fifth and the sixth studies were added in prostate cancer and IPF, respectively. In Supplementary Figure S1B, we found that Wallace, Singh and Tomlins generally had stronger DE evidence than other studies. The increased number of detected DE genes in Figure 3B might have been caused

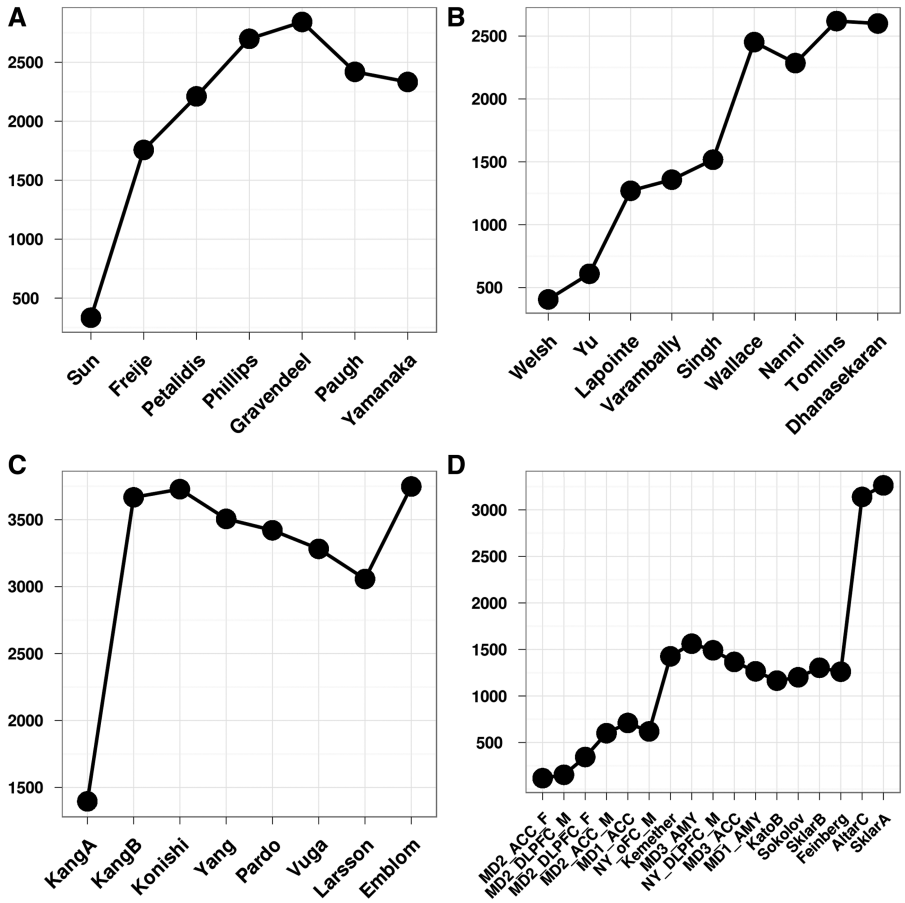


**Figure 3.** Marginal impacts on meta-analysis for DE genes detection. *X*-axis represents each study included cumulatively to a series of meta-analyses. The order of addition follows the SMR score in the Table 4. *Y*-axis represents the number of DE genes detected. (**A**) Brain cancer example under FDR = 0.5% threshold. (**B**) Prostate cancer example under FDR = 0.1% threshold. (**C**) IPF example under FDR = 0.1% threshold. (**D**) MDD example under *P*-value = 0.01 threshold.

by this bias although the pathway result in Figure 4B did not show increased finding. The prostate cancer example demonstrated a case that pure AQCg or CQCg method focusing on commonality of DE gene detection was not effective enough when studies were heterogeneous. In the IPF example, similar observations were found. Inclusion of Emblom greatly increased the number of DE genes (Figure 3C) but decreased the number of detected pathways (Figure 4C). This may be a result of the large number of DE genes detected by Emblom (Supplementary Figure S1C).

Figures 3D and 4D shows the biological impact evaluation result of MDD. In contrast to previous examples, MDD studies are characterized by weak overall signals. We, therefore, applied a liberal DE gene detection criterion at unadjusted *P*-value = 1%. The pathway identification, however, had strong enough signal and we applied usual FDR = 5% threshold. Despite the liberal threshold, the numbers of detected DE genes were still smaller than other examples. The number of detected DE genes increased moderately as more studies were included, and plateaued after inclusion of the low SMR score studies, except for the Kemether and AltarC studies, in which cases the DE genes increased significantly after their inclusion (Figure 3D). We highlight here that the Kemether and AltarC studies were considered problematic studies from MetaQC (Table 4). Their inclusion actually caused significant drop in the number of identified pathways (Figure 4D), suggesting that the large increase in DE genes may not be disease-related but instead may be related to technical specificities of the latter two studies. Again, Supplementary Figure S1D showed a large number of DE genes in these two studies compared to others. From the four examples in Figure 3A–D and Figure 4A–D, we conclude that the biological impact judged by the number of detected DE genes can be misleading and that the number of detected enriched pathways may represent a better assessment criterion.

## Simulations

To further validate the QC result of our proposed method, we investigated a simple yet insightful simulation scheme. In each simulation of a given example, a study is randomly selected from another example and added as a known outlier for MetaQC re-evaluation. For example, a prostate cancer study is randomly selected as a known outlier and added to the seven brain cancer studies (Figure 5A) and the MetaQC evaluation is performed.

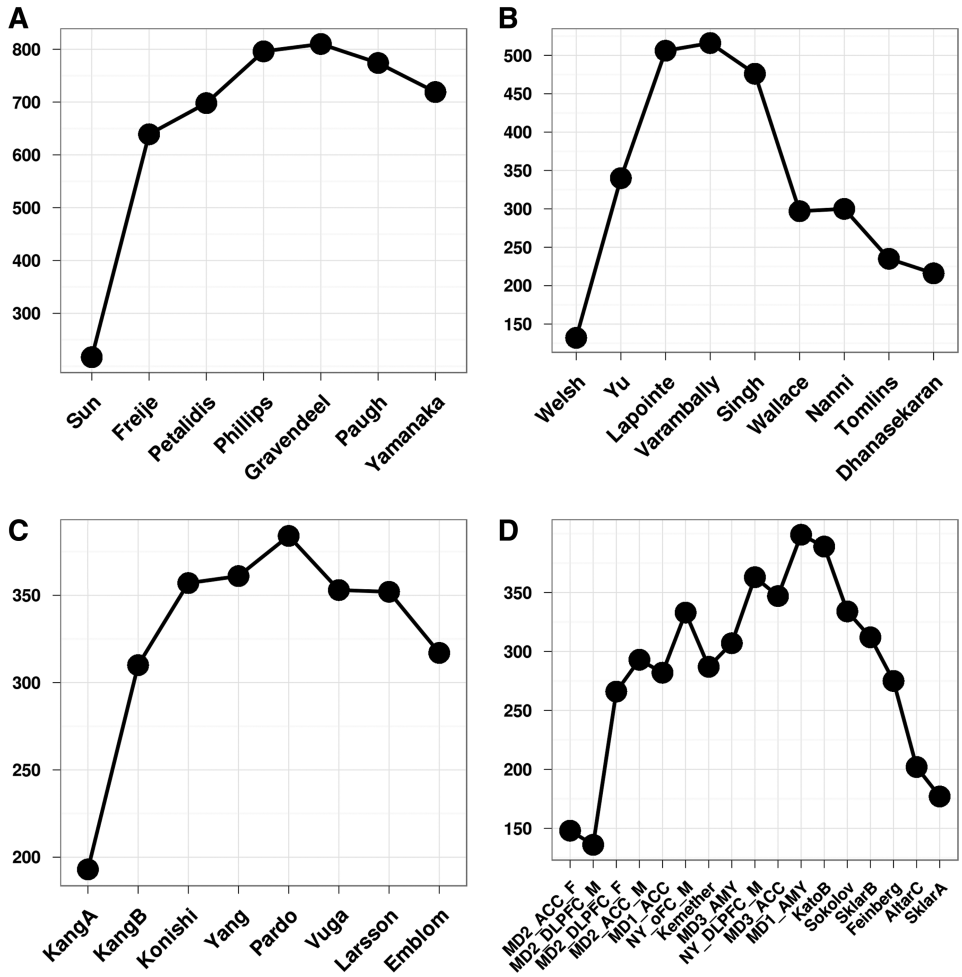


**Figure 4.** Marginal impacts on meta-analysis for enriched pathway detection. Similar to Figure 3, *X*-axis represents each study included cumulatively to a series of meta-analyses. The order of addition follows the SMR score in the Table 4. *Y*-axis represents the number of pathways detected. All examples are shown under FDR = 5% threshold. (**A**) Brain cancer example. (**B**) Prostate cancer example. (**C**) IPF example. (**D**) MDD example.

The simulations were repeated through all prostate cancer studies and the changes of SMR scores were recorded and compared. In Figure 5A, the scores of 1-SMR in seven brain cancer studies were plotted in the first columns (labeled as 'NA'). In the following nine simulations, a prostate cancer studies was added to the seven brain cancer studies and the scores of 1-SMR were recalculated. The added outlier study was plotted by an asterisk symbol.

Interestingly, the result showed that the added prostate cancer studies consistently generated small scores of 1-SMR similar to Yamanaka study and were always detected as a definite exclusion case. Although prostate cancer studies might share certain intrinsic biological mechanisms with brain cancers, they seemed to served well as control studies that further verified exclusion of Yamanaka study. The addition of a random irrelevant study as the 'null' study seems to provide an alternative objective and practical threshold to decide the exclusion of studies. The quality order of the brain cancers also did not change in general by the added prostate cancers.

For the second simulation in Figure 5B, a brain cancer was added as an outlier study to nine prostate cancer studies in each simulation. The results showed that the added brain cancer study had scores of 1-SMR better than Nanni, Tomlins and Dhanasekaran. In Figure 5C, we added brain cancer studies as outliers into the eight IPF studies. The result showed similar pattern that argued to exclude Embolm and Larsson studies. Figure 5D showed the result that one of seven brain studies were added to 17 MDD studies sequentially. The result suggested that the top six studies had good quality (curves always above the added outlier studies), the bottom 2–3 studies had problematic quality (curves always below the added outlier studies), and the middle studies were the borderline cases. These simulation results demonstrated the effectiveness and robustness of the MetaQC assessment to screen out outlier studies. The added outlier studies can serve as an alternative baseline control to confidently argue exclusion of problematic studies for meta-analysis.

## DISCUSSION

As more high-throughput genomic datasets are generated and stored in public domain, the statistical and informatic

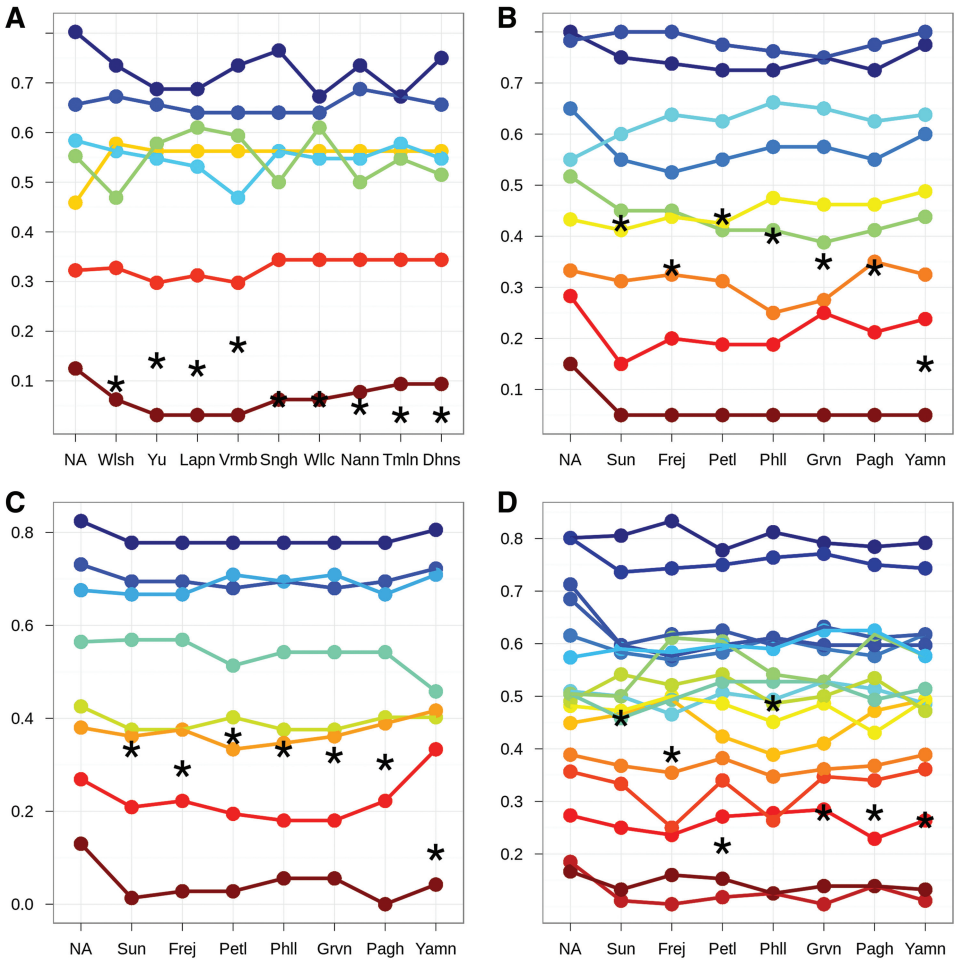


**Figure 5.** Simulations showing effects of adding an irrelevant 'spike-in' study. Y-axis represents the value of 1-SMR. Greater *Y*-axis values correspond to better quality studies judged by MetaQC. *X*-axis represents the addition of an irrelevant 'spike-in' study. A set of irrelvant studies were obtained from one of the other three examples. 'NA' represents the original quality result without the spike-in simulation which is the same result as Table 4. A black asterisk represents the 1-SMR of the added irrelevant study. (**A**) Brain cancer studies with a spiked-in prostate cancer study. (**B**) Prostate cancer studies with a spiked-in brain cancer study. (**C**) IPF studies with a spiked-in brain cancer study. (**D**) MDD studies with a spiked-in brain cancer study.

infrastructure to retrieve information in the huge amount of data has become an essential component in biomedical research. Meta-analysis to combine information across multiple studies provides increased statistical power, allows to distinguish artifacts from single studies from true biological effects and thus generates more accurate results. In the literature, many meta-analysis methods have been developed and applied for genomic applications, but the quality control and objective inclusion/exclusion criteria have been largely overlooked. Hence there is a critical need for systematic quality assessment, as the inclusion of studies with variable information content will greatly affect the outcome of meta-analyses. In this article, we proposed the MetaQC evaluation tool to provide quantitative quality control and to assist selection of studies into microarray meta-analysis. Six QC measures were developed (Table 1). A PCA biplot and a SMR score were used to recommend the final decision (Figure 2). In the evaluation, we examined the impact of DE gene detection and pathway identification when studies were sequentially added in the meta-analysis by SMR score (Figures 3 and 4). Confirming our hypothesis, the result showed general adverse effects of adding problematic studies into meta-analysis. These adverse effects were shown more clearly in pathway analysis than in DE gene detection. Simulations by 'spike-in' a known outlying study into the meta-analysis found further validation of the effectiveness of MetaQC (Figure 5). The 'spiked-in' studies generally serve well as good negative controls to suggest filtering threshold. In conclusion, the proposed MetaQC evaluation system provides excellent quality evaluation for selecting studies into meta-analysis.

The 'MetaQC' package in *R* has been published in CRAN library (http://cran.r-project.org/). By its nature of dealing with multiple high-throughput experimental datasets, demand of computing is extensive but is affordable in the current *R* package using a regular computing machine (Intel Core 2 Duo Processor CPU and 4GB memory). Computing of IQC, AQC and CQC generally took 5–20 min for all examples except that the larger

MDD example needed 40 min to calculate IQC. EQC was the most demanding task due to permutation analyses in the algorithm. It took 3–7 h for prostate, brain and IPF examples and needed ~130 h for the MDD example. Rewriting the *R* code using more efficient C language or adopting parallel computing will solve the computing bottleneck if application to larger data sets is needed.

One has to note that MetaQC is not meant as a fully automated decision tool. Any attempt of such automation overlooks the complexity and heterogeneity involved in the high-throughput experiments and is likely to fail. The users are recommended to use the PCA biplot visualization tool, SMR scores and spike-in thresholds to obtain a first-step quality summary. Review of technical, clinical and biological information of the studies (such as sample size, platform, tissue collection, experimental protocols or demographics) help validate and understand the causes of problematic studies that should be excluded. In our evaluations, we tested four examples that covered different situations one might encounter in a genomic meta-analysis. The brain cancer example and IPF example are prototypes of strong signal and generally homogeneous studies. On the other hand, the prostate cancer example had strong signal but heterogeneous studies. Finally, the MDD example highlights the necessity of robust meta-QC for studies with overall weak signal. Weak signals may come from disease heterogeneity, complex biological disease mechanisms and use of post-mortem brain tissues, and will thus apply to other cases. Results from these different examples are consistent and support the validity of the MetaQC method.

The current MetaQC method is mainly developed for microarray meta-analysis. One future direction is to extend and tailor the QC measures and visualization and summarization tools developed in this article to other types of genomic meta-analysis, such as genome-wide association studies (GWAS) or the increasingly popular sequencing based data. In addition to meta-analysis of one type of genomic data, integrative analysis to combine information from multiple types of genomic data (e.g. combining gene expression, genotyping, copy number variation, methylation and miRNA for a given cohort of patients) has drawn increasing attention. We can foresee in the near future that multiple patient cohorts recruited in different medical centers may all be analyzed by the list of aforementioned high-throughput techniques independently. Meta-analysis of multi-dimensional data sets will bring new challenges and its quality evaluation or heterogeneity assessment will be another future research direction.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR online: Supplementary figure S1.

## FUNDING

National Institutes of Health (NIH) (RC2HL101715 to D.D.K., N.K.and G.C.T.; MH077159, MH084060, MH085111 and MH084053 to E.S. and G.C.T.). Funding for open access charge: NIH.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207.
2. Parkinson,H., Sarkans,U., Shojatalab,M., Abeygunawardena,N., Contrino,S., Coulson,R., Farne,A., Garcia Lara,G., Holloway,E., Kapushesky,M. *et al.* (2005) ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **33**, D553.
3. Sherlock,G., Hernandez-Boussard,T., Kasarskis,A., Binkley,G., Matese,J.C., Dwight,S.S., Kaloper,M., Weng,S., Jin,H., Ball,C.A. *et al.* (2001) The Stanford microarray database. *Nucleic Acids Res.*, **29**, 152.
4. Draghici,S., Khatri,P., Eklund,A.C. and Szallasi,Z. (2006) Reliability and reproducibility issues in DNA microarray measurements. *TRENDS Genet.*, **22**, 101–109.
5. Ein-Dor,L., Kela,I., Getz,G., Givol,D. and Domany,E. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171.
6. Ioannidis,J.P.A., Allison,D.B., Ball,C.A., Coulibaly,I., Cui,X., Culhane,A.C., Falchi,M., Furlanello,C., Game,L., Jurman,G. *et al.* (2008) Repeatability of published microarray gene expression analyses. *Nat. Genet.*, **41**, 149–155.
7. Tan,P.K., Downey,T.J., Spitznagel,E.L. Jr, Xu,P., Fu,D., Dimitrov,D.S., Lempicki,R.A., Raaka,B.M. and Cam,M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676.
8. Fisher,R.A. (1948) Question 14: Combining independent tests of significance. *Am. Statistician*, **2**, 30–30J.
9. Rhodes,D.R., Barrette,T.R., Rubin,M.A., Ghosh,D. and Chinnaiyan,A.M. (2002) Meta-analysis of microarrays. *Cancer Res.*, **62**, 4427.
10. Stouffer,S.A., DeVinney,L.C. and Suchmen,E.A. (1949) *The American Soldier: Adjustment During Army Life*. Princeton University Press, Princeton, NJ.
11. Ghosh,D., Barette,T.R., Rhodes,D. and Chinnaiyan,A.M. (2003) Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Functional Amp; Integrative Genomics*, **3**, 180–188.
12. Choi,J.K., Yu,U., Kim,S. and Yoo,O.J. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, i84.
13. Stevens,J.R. and Doerge,R.W. (2005) Combining affymetrix microarray results. *BMC Bioinformatics*, **6**, 57.
14. Jung,Y.Y., Oh,M.S., Shin,D.W., Kang,S. and Oh,H.S. (2006) Identifying differentially expressed genes in meta-analysis via Bayesian model-based clustering. *Biometrical J.*, **48**, 435–450.
15. Tseng,G.C., Oh,M.K., Rohlin,L., Liao,J.C. and Wong,W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549.
16. Breitling,R. and Herzyk,P. (2005) Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J. Bioinformatics Computational Biol.*, **3**, 1171–1190.
17. Hong,F., Breitling,R., McEntee,C.W., Wittner,B.S., Nemhauser,J.L. and Chory,J. (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**, 2825.
18. Lu,S., Li,J., Song,C., Shen,K. and Tseng,G.C. (2010) Biomarker detection in the integration of multiple multi-class genomic studies. *Bioinformatics*, **26**, 333.
19. Owen,A.B. (2007) Pearson's test in a large scale multiple meta-analysis.

20. Shen,K. and Tseng,G.C. (2010) Meta-analysis for pathway enrichment analysis when combining multiple microarray studies. *Bioinformatics*, **26**, 1316–1323.
21. Dreyfuss,J.M., Johnson,M.D. and Park,P.J. (2009) Meta-analysis of glioblastoma multiforme versus anaplastic astrocytoma identifies robust gene markers. *Mol. Cancer*, **8**, 71.
22. Grutzmann,R., Boriss,H., Ammerpohl,O., Luttges,J., Kalthoff,H., Schackert,H.K., Kloppel,G., Saeger,H.D. and Pilarsky,C. (2005) Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, **24**, 5079–5088.
23. Mulligan,M.K., Ponomarev,I., Hitzemann,R.J., Belknap,J.K., Tabakoff,B., Harris,R.A., Crabbe,J.C., Blednov,Y.A., Grahame,N.J., Phillips,T.J. *et al.* (2006) Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis. *Proc. Natl Acad. Sci. USA*, **103**, 6368.
24. Rhodes,D.R., Yu,J., Shanker,K., Deshpande,N., Varambally,R., Ghosh,D., Barrette,T., Pandey,A. and Chinnaiyan,A.M. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA*, **101**, 9309.
25. Smith,D.D., Sætrom,P., Snøve,O., Lundberg,C., Rivas,G.E., Glackin,C. and Larson,G.P. (2008) Meta-analysis of breast cancer microarray studies in conjunction with conserved cis-elements suggest patterns for coordinate regulation. *BMC Bioinformatics*, **9**, 63.
26. Wirapati,P., Sotiriou,C., Kunkel,S., Farmer,P., Pradervand,S., Haibe-Kains,B., Desmedt,C., Ignatiadis,M. and Sengstag,T. (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.*, **10**, R65.
27. Eysenck,H. (1994) Systematic reviews: meta-analysis and its problems. *BMJ*, **309**, 789.
28. Sboner,A., Demichelis,F., Calza,S., Pawitan,Y., Setlur,S.R., Hoshida,Y., Perner,S., Adami,H.O., Fall,K., Mucci,L.A. *et al.* (2010) Molecular sampling of prostate cancer: a dilemma for predicting disease progression. *BMC Med. Genomics*, **3**, 8.
29. Kim,D.S., Collard,H.R. and King,T.E. Jr (2006) *Proceedings of American Thoraic Society*, **3**, 285.
30. Li,J.Z., Vawter,M.P., Walsh,D.M., Tomita,H., Evans,S.J., Choudary,P.V., Lopez,J.F., Avelar,A., Shokoohi,V., Chung,T. *et al.* (2004) Systematic changes in gene expression in postmortem human brains associated with tissue pH and terminal medical conditions. *Hum. Mol. Genet.*, **13**, 609–616.
31. Atz,M., Walsh,D., Cartagena,P., Li,J., Evans,S., Choudary,P., Overman,K., Stein,R., Tomita,H., Potkin,S. *et al.* (2007) Methodological considerations for gene expression profiling of human brain. *J. Neurosci. Methods*, **163**, 295–309.
32. Garrett-Mayer,E., Parmigiani,G., Zhong,X., Cope,L. and Gabrielson,E. (2008) Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics*, **9**, 333–354.
33. Parmigiani,G., Garrett-Mayer,E.S., Anbazhagan,R. and Gabrielson,E. (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin. Cancer Res.*, **10**, 2922–2927.
34. Mann,H.B. and Whitney,D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.
35. Cox,D.R. and Hinkley,D.V. (1974) *Theoretical Statistics*. Chapman and Hall, London.
36. North,B., Curtis,D. and Sham,P. (2003) A note on the calculation of empirical P values from Monte Carlo procedures. *Am. J. Hum. Genet.*, **72**, 498.
37. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545.
38. Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.
39. Kendall,M.G. and Stuart,A. (1973) *The Advanced Theory of Statistics*. Charles Griffin & Co. Ltd, London, *paragraph*, 33.
40. Jolliffe,I. (2002) *Principal component analysis*. Springer Series in Statistics, Hoboken, NJ.
41. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.
42. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B*, **57**, 289–300.
43. R Development Core Team. (2010) Vienna, Austria.
44. Freije,W.A., Castro-Vargas,F.E., Fang,Z., Horvath,S., Cloughesy,T., Liau,L.M., Mischel,P.S. and Nelson,S.F. (2004) Gene expression profiling of gliomas strongly predicts survival. *Cancer Res.*, **64**, 6503.
45. Phillips,H.S., Kharbanda,S., Chen,R., Forrest,W.F., Soriano,R.H., Wu,T.D., Misra,A., Nigro,J.M., Colman,H., Soroceanu,L. *et al.* (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*, **9**, 157–173.
46. Sun,L., Hui,A.M., Su,Q., Vortmeyer,A., Kotliarov,Y., Pastorino,S., Passaniti,A., Menon,J., Walling,J., Bailey,R. *et al.* (2006) Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell*, **9**, 287–300.
47. Yamanaka,R., Arao,T., Yajima,N., Tsuchiya,N., Homma,J., Tanaka,R., Sano,M., Oide,A., Sekijima,M. and Nishio,K. (2006) Identification of expressed genes characterizing long-term survival in malignant glioma patients. *Oncogene*, **25**, 5994–6002.
48. Petalidis,L.P., Oulas,A., Backlund,M., Wayland,M.T., Liu,L., Plant,K., Happerfield,L., Freeman,T.C., Poirazi,P. and Collins,V.P. (2008) Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. *Mol. Cancer Therap.*, **7**, 1013.
49. Gravendeel,L.A.M., Kouwenhoven,M., Gevaert,O., de Rooi,J.J., Stubbs,A.P., Duijm,J.E., Daemen,A., Bleeker,F.E., Bralten,L.B.C., Kloosterhof,N.K. *et al.* (2009) Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Res.*, **69**, 9065.
50. Paugh,B.S., Qu,C., Jones,C., Liu,Z., Adamowicz-Brice,M., Zhang,J., Bax,D.A., Coyle,B., Barrow,J., Hargrave,D. *et al.* (2010) Integrated molecular genetic profiling of pediatric high-grade gliomas reveals key differences with the adult disease. *J. Clin. Oncol.*, **28**, 3061.
51. Dhanasekaran,S.M., Barrette,T.R., Ghosh,D., Shah,R., Varambally,S., Kurachi,K., Pienta,K.J., Rubin,M.A. and Chinnaiyan,A.M. (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**, 822–826.
52. Welsh,J.B., Sapinoso,L.M., Su,A.I., Kern,S.G., Wang-Rodriguez,J., Moskaluk,C.A., Frierson,H.F. and Hampton,G.M. (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.*, **61**, 5974.
53. Singh,D., Febbo,P.G., Ross,K., Jackson,D.G., Manola,J., Ladd,C., Tamayo,P., Renshaw,A.A., D'Amico,A.V., Richie,J.P. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.
54. Lapointe,J., Li,C., Higgins,J.P., Van De Rijn,M., Bair,E., Montgomery,K., Ferrari,M., Egevad,L., Rayford,W., Bergerheim,U. *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 811.
55. Yu,Y.P., Landsittel,D., Jing,L., Nelson,J., Ren,B., Liu,L., McDonald,C., Thomas,R., Dhir,R., Finkelstein,S. *et al.* (2004) Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.*, **22**, 2790.
56. Varambally,S., Yu,J., Laxman,B., Rhodes,D.R., Mehra,R., Tomlins,S.A., Shah,R.B., Chandran,U., Monzon,F.A., Becich,M.J. *et al.* (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell*, **8**, 393–406.
57. Nanni,S., Priolo,C., Grasselli,A., D'Eletto,M., Merola,R., Moretti,F., Gallucci,M., De Carli,P., Sentinelli,S., Cianciulli,A.M. *et al.* (2006) Epithelial-restricted gene profile of primary cultures

from human prostate tumors: a molecular approach to predict clinical behavior of prostate cancer. *Mol. Cancer Res.*, **4**, 79.
58. Tomlins,S.A., Mehra,R., Rhodes,D.R., Cao,X., Wang,L., Dhanasekaran,S.M., Kalyana-Sundaram,S., Wei,J.T., Rubin,M.A., Pienta,K.J. *et al.* (2006) Integrative molecular concept modeling of prostate cancer progression. *Nat. Genet.*, **39**, 41–51.
59. Wallace,T.A., Prueitt,R.L., Yi,M., Howe,T.M., Gillespie,J.W., Yfantis,H.G., Stephens,R.M., Caporaso,N.E., Loffredo,C.A. and Ambs,S. (2008) Tumor immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Res.*, **68**, 927.
60. Pardo,A., Gibson,K., Cisneros,J., Richards,T.J., Yang,Y., Becerril,C., Yousem,S., Herrera,I., Ruiz,V., Selman,M. *et al.* (2005) Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis. *PLoS Med.*, **2**, 891.
61. Yang,I.V., Burch,L.H., Steele,M.P., Savov,J.D., Hollingsworth,J.W., McElvania-Tekippe,E., Berman,K.G., Speer,M.C., Sporn,T.A., Brown,K.K. *et al.* (2007) Gene expression profiling of familial and sporadic cases of interstitial pneumonia. *Am. J. Resp.Crit. Care Med.*, **175**, 45–54.
62. Larsson,O., Diebold,D., Fan,D., Peterson,M., Nho,R.S., Bitterman,P.B. and Henke,C.A. (2008) Fibrotic myofibroblasts manifest genome-wide derangements of translational control. *PLoS One*, **3**, 3220.
63. Vuga,L.J., Ben-Yehudah,A., Kovkarova-Naumovski,E., Oriss,T., Gibson,K.F., Feghali-Bostwick,C. and Kaminski,N. (2009) WNT5A is a regulator of fibroblast proliferation and resistance to apoptosis. *Am. J. Resp. Cell Mol. Biol.*, **41**, 583–589.
64. Konishi,K., Gibson,K.F., Lindell,K.O., Richards,T.J., Zhang,Y., Dhir,R., Bisceglia,M., Gilbert,S., Yousem,S.A., Song,J.W. *et al.* (2009) Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. *Am. J. Resp. Crit. Care Med.*, **180**, 167.
65. Emblom-Callahan,M.C., Chhina,M.K., Shlobin,O.A., Ahmad,S., Reese,E.S., Iyer,E.P.R., Cox,D.N., Brenner,R., Burton,N.A., Grant,G.M. *et al.* (2010) Genomic phenotype of non-cultured pulmonary fibroblasts in idiopathic pulmonary fibrosis. *Genomics*, **96**, 134–145.