Expression Profiling of Archival Tumors for Long-term Health Studies

Levi Waldron1,3, Shuji Ogino2,5, Yujin Hoshida8, Kaori Shima2, Amy E. McCart Reed9, Peter T. Simpson9, Yoshifumi Baba2, Katsuhiko Nosho2, Nicola Segataa, Ana Cristina Vargas5, Margaret C. Cummings9,11, Sunil R. Lakhani9,10,11, Gregory J. Kirkner6, Edward Giovannucci4,6, John Quackenbush1, Todd R. Golub7, Charles S. Fuchs2,6, Giovanni Parmigiani1,3, and Curtis Huttenhower3

Abstract

**Purpose:** More than 20 million archival tissue samples are stored annually in the United States as formalin-fixed, paraffin-embedded (FFPE) blocks, but RNA degradation during fixation and storage has prevented their use for transcriptional profiling. New and highly sensitive assays for whole-transcriptome microarray analysis of FFPE tissues are now available, but resulting data include noise and variability for which previous expression array methods are inadequate.

**Experimental Design:** We present the two largest whole-genome expression studies from FFPE tissues to date, comprising 1,003 colorectal cancer (CRC) and 168 breast cancer samples, combined with a meta-analysis of 14 new and published FFPE microarray datasets. We develop and validate quality control (QC) methods through technical replication, independent samples, comparison to results from fresh-frozen tissue, and recovery of expected associations between gene expression and protein abundance.

**Results:** Archival tissues from large, multicenter studies showed a much wider range of transcriptional data quality relative to smaller or frozen tissue studies and required stringent QC for subsequent analysis. We developed novel methods for such QC of archival tissue expression profiles based on sample dynamic range and per-study median profile. This enabled validated identification of gene signatures of microsatellite instability and additional features of CRC, and improved recovery of associations between gene expression and protein abundance of MLH1, FASN, CDX2, MGMT, and SIRT1 in CRC tumors.

**Conclusions:** These methods for large-scale QC of FFPE expression profiles enable study of the cancer transcriptome in relation to extensive clinicopathological information, tumor molecular biomarkers, and long-term lifestyle and outcome data. *Clin Cancer Res; 18(22); 6136–46. ©2012 AACR.*

Introduction

Formalin fixation and paraffin embedding (FFPE) is the tissue preservation method for virtually all routine histo-pathology tests (1), and excised tumors are routinely archived as FFPE blocks. These samples are preserved for decades and, as such, can be used for integrated analysis of environmental and host factors, molecular biomarkers, and tumor evolution, to decipher diseases at both the molecular and population levels (2, 3). The importance of quantitative measurement of RNA abundance from archival tissues has long been recognized (4), but technical challenges associated with extensive RNA fragmentation and cross-linking have limited their utility for translational applications. Initial assessments of whole-genome amplification and microarray hybridization from FFPE samples have shown that the resulting expression profiles can be replicated (5–10) and that they provide data comparable to that from fresh-frozen samples (6, 10–12). Validations of these technologies, such the Illumina DASL (6) and NuGEN Ovation (13), have focused on artificially degraded samples (6) or on small cohorts (5–7, 9, 10, 14), providing preliminary evidence that FFPE profiling will scale to transcriptome profiling of large cohort studies of cancer. Critically, standard measures of sample RNA quality have proven inadequate for predicting such expression data quality from FFPE tissues (12, 13), and specimen quality control is only one...
module of complete assay quality control. None of these recent technologies has as yet been tested on a large scale, and standard bioinformatic methods developed for the robust study of fresh-frozen tissues may not be appropriate in such clinical settings.

We have thus carried out the 2 largest whole-transcriptome studies to date using FFPE tissues and present them here accompanied by novel bioinformatic methods for quality control and validation of archival tissue expression profiles. To establish methods for quality control of archival tissue expression data, we first conducted a technical study comprising 168 samples and replicates from primary breast tumors, metastases, autopsy samples, and controls. We developed the use of interquartile range (IQR) of raw expression intensities as a surrogate quality measure for a range of sources of variation in sample data quality. We validated its utility for rejecting low-quality samples without the aid of technical replication, both for improving measurement reproducibility and for ranking lists of differentially expressed genes. Unlike other common quality metrics that use platform-specific control probes, IQR was universally available for all published datasets, enabling consistent data quality assessment in published FFPE expression profiles as compared with our novel data. This analysis showed that the smaller initial whole-genome expression studies of FFPE tissues are not individually representative of the range of clinical data quality to be expected in large studies with diverse sample sources, and that researchers should be prepared for a greater spectrum of data variability and higher rates of sample failure. Furthermore, we investigated the reproducibility of measurements of individual gene probes between replicates and assessed several indicators of probe utility. This resulted in an end-to-end quality control pipeline to improve the accuracy and reproducibility of whole-genome expression data analysis from FFPE tissues, validated by technical replication and applicable to all scenarios with or without technical replication.

Finally, we showed the utility and reproducibility of expression measurements in archival tissues from long-term health studies, and the need for these quality control methods for such studies. This demonstration included 1,003 colorectal cancer samples from dozens of hospitals across the United States and surgeries spanning several decades, drawn from participants in the Nurses’ Health Study (NHS; ref. 15) and Health Professionals Follow-up Study (HPFS; ref. 16). These are long-term epidemiological studies of 122,000 and 52,000 participants, respectively, who were recruited prospectively beginning in 1976 and 1986. We investigated associations of mRNA expression with promoter CpG island methylation and protein abundance in these samples, a correspondence that, when applicable, was improved by our quality control (QC) measures. Likewise the utility of these QC methods were confirmed for differential expression analysis (e.g., for transcripts segregating colon and rectal tumors) and for verifying transcripts associated with microsatellite instability (MSI) in fresh-frozen colorectal cancer (CRC) tumors. Thus, this study establishes validated quality control measures at the levels of study, sample, and probe to improve the efficacy of archival tissue gene expression profiling.

Materials and Methods

We present a FFPE gene expression quality control methodology validated on 2 large, novel gene expression profiling studies using archival, clinical tissues. We conducted a study of 168 primary breast tumors and metastases to lymph node, liver, chest wall, lung, and spleen, as well as positive and negative controls, including 44 technical replicates, to assess quality control methodology and probe-level reproducibility [referred to as the breast cancer/autopsy (BC/A) dataset]. In the second study, we profiled 1,003 tumors from colorectal cancer patients drawn from the NHS and HPFS (the CRC dataset), allowing investigation of disease subtypes and of association between gene expression and patient phenotypes and tumor characteristics. We also analyzed 17 publicly available datasets from 12 independent studies for comparison of data quality characteristics.

The CRC and BC/A datasets are available from the Gene Expression Omnibus under accession numbers GSE32651 and GSE32490.

Patient samples

We used 2 U.S. Nationwide prospective cohort studies, NHS (15) and HPFS (16). Cohort participants have received a questionnaire every two years to update information on weight, dietary and lifestyle factors, and to identify new cases of cancer. The National Death Index was used to identify unreported cases of lethal cancer. Tissue collection for the NHS/HPFS study was approved by the Brigham and Women’s Hospital and Harvard School of Public Health Institutional Review Boards. Informed consent was obtained to analyze tumor tissue.
The BC/A study included several types of tissues and control samples, detailed in the Supplemental Methods. Tissue collection for the BC/A study was approved by Human Research Ethics Committees from The University of Queensland, The Royal Brisbane & Women’s Hospital, and the Uniting Healthcare Trust.

Published data were obtained from the Gene Expression Omnibus or ArrayExpress. Patient cohorts considered in this study are summarized in Table 1.

Details of sample preparation of assay methods are provided in the Supplementary Methods.

NHS/HPFS cohort metadata

Clinicopathological and epidemiological data for the NHS/HPFS cohorts were extracted by Statistical Analysis System script in accordance with NHS and HPFS Cohort program review procedures. Immunohistochemistry (IHC), DNA methylation, and MSI methods are described in the Supplementary Methods, and representative IHC stains for CCND1 and SIRT1 are shown in Supplementary Fig. S1.

Per-probe reproducibility in the BC/A study

The sample QC pipeline described below was applied to the BC/A samples, retaining only 90 of the 168 samples and 9 matched pairs of replicate samples from primary breast tumors and lymph node metastases. In cases with more than 2 replicates available, 2 replicates were selected randomly. The sets of replicates were separated and quantile normalized independently, and 1 set was used to calculate standard deviation of each probe across the 9 samples (or other measures of probe activity shown in Supplementary Fig. S7). Similarity of expression measurements between technical replicates, within quintiles of probes with similar measures of probe activity, was compared by Spearman correlation, Euclidian distance, and Manhattan distance.

FFPE gene expression quality control process

Sample quality control. Overtly failed samples, with all-zero expression values, were first removed from all analysis. We subsequently conducted the following method for identifying low-quality data:

1. Construct a median pseudochip from remaining samples by calculating the median intensity of each probe across all samples in a study.
2. Calculate Spearman rank correlation of each sample to this pseudochip; plot these values against each sample’s IQRs.
3. Fit a Loess smoothing curve to a moving average of window width 7.
4. Identify the point of maximum downward inflection (greatest magnitude of negative second derivative) of this smoothing curve.
5. Reject samples if their IQR falls below this point or if their correlation to the median pseudochip is below the smoothing curve by more than 1.5 times the IQR of the residuals.

Table 1. Original (NHS/HPFS and BC/A) and public microarray datasets considered in this study

<table>
<thead>
<tr>
<th>Study</th>
<th>Accession ID</th>
<th>Storage method</th>
<th>Sample type</th>
<th>Platform</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHS/HPFS</td>
<td>GSE32651</td>
<td>FFPE</td>
<td>CRC</td>
<td>Illumina DASL HumanRef-8 v3</td>
<td>1,003</td>
</tr>
<tr>
<td>BC/A</td>
<td>GSE32490</td>
<td>FFPE</td>
<td>Breast LN</td>
<td>Illumina DASL</td>
<td>168</td>
</tr>
<tr>
<td>Sadi and colleagues (18)</td>
<td>GSE23386</td>
<td>FF/FFPE</td>
<td>Breast</td>
<td>Illumina HumanRef-8 v3</td>
<td>25/25</td>
</tr>
<tr>
<td>Jiang and colleagues (19)</td>
<td>GSE10950</td>
<td>FF</td>
<td>CRC</td>
<td>Illumina HumanRef-8 v2</td>
<td>48</td>
</tr>
<tr>
<td>April and colleagues (8)</td>
<td>GSE17558/GSE17572</td>
<td>FF/FFPE</td>
<td>Lung</td>
<td>Illumina HumanRef-8 v3</td>
<td>8/8</td>
</tr>
<tr>
<td>Mittempergher and colleagues (12)</td>
<td>E-TABM-1081</td>
<td>FF/FFPE</td>
<td>Breast</td>
<td>Illumina HumanRef-8 v3</td>
<td>47/46</td>
</tr>
<tr>
<td>Budczies and colleagues (23)</td>
<td>GSE11001</td>
<td>FFPE</td>
<td>Breast</td>
<td>Affymetrix HG-U133-plus2</td>
<td>30</td>
</tr>
<tr>
<td>Collisson and colleagues (25)</td>
<td>GSE17891</td>
<td>FFPE</td>
<td>Pancreas</td>
<td>Affymetrix HG-U133-plus2</td>
<td>47</td>
</tr>
<tr>
<td>Hall and colleagues (24)</td>
<td>GSE27388</td>
<td>FFPE</td>
<td>Cervix</td>
<td>Affymetrix Human Exon 1.0 ST</td>
<td>28</td>
</tr>
<tr>
<td>Abdueva and colleagues (22)</td>
<td>GSE19249</td>
<td>FFPE</td>
<td>Multiple</td>
<td>Affymetrix HG-U133A/</td>
<td>15/8</td>
</tr>
<tr>
<td>Williams and colleagues (1)</td>
<td>GSE19246</td>
<td>FFPE</td>
<td>Lymphoma</td>
<td>Affymetrix HG-U133-plus2</td>
<td>38</td>
</tr>
<tr>
<td>Waddell and colleagues (10)</td>
<td>GSE21921</td>
<td>FFPE</td>
<td>Breast</td>
<td>Illumina DASL HumanRef-8 v3</td>
<td>85</td>
</tr>
<tr>
<td>Villanueva and colleagues (20)</td>
<td>GSE19977</td>
<td>FFPE</td>
<td>Liver</td>
<td>Illumina DASL HumanRef-8 v3</td>
<td>164</td>
</tr>
<tr>
<td>Minguez and colleagues (21)</td>
<td>GSE20017</td>
<td>FFPE</td>
<td>Hepatocellular carcinoma</td>
<td>Illumina DASL HumanRef-8 v3</td>
<td>135</td>
</tr>
</tbody>
</table>

NOTE: GSE accession IDs refer to the Gene Expression Omnibus, E-TABM IDs refer to the ArrayExpress database. Abbreviation: FF, fresh-frozen.
This methodology is provided in the ffpe Bioconductor package. This methodology, together with the steps below, was validated by assessment on 44 BC/A technical replicates (Fig. 1), behavior in 12 publicly available datasets (Fig. 2), and performance on 1,003 CRC samples (Figs. 3 and 5).

**Expression data preprocessing.** Probes present in fewer than 10% of NHS/HPFS samples remaining after sample quality control (nominal $P < 0.01$) were removed (4,476 probes) before normalization. Two alternative methods of data transformation and normalization were considered: log2 transformation followed by quantile normalization, and variance stabilizing transformation followed by Robust Spline Normalization (17). Optional imputation of missing expression values ($P > 0.01$) was conducted by $k$-nearest neighbors using the impute R package with default settings.

*Strong* and *permissive* QC thresholds. For the "permissive" QC threshold we removed 15 of 1,003 samples where complete hybridization failure occurred (IQR of zero). The “strict” QC threshold was determined as described under sample quality control, which resulted in removal of an additional 193 samples.

**Probe quality control.** We assessed probe QC methods including standard deviation across samples, fraction of samples in which the probe was detected (nominal $P < 0.01$), mean expression, and coefficient of variance. On the basis of the performance in BC/A technical replicates (Fig. 4), our final QC pipeline removes probes below median variance across each dataset.

**Published data used for analysis and validation**

We identified seven studies with publicly available Illumina WG-DASL raw data (refs. 6, 10, 12, 18–21; Table 1), four example datasets using fresh-frozen tissues assayed by Illumina BeadArray, and five datasets of FFPE tissues assayed by Affymetrix GeneChip (1, 22–25). Data quality for each study was summarized by the distribution of IQRs of raw log2 intensities for each sample (Fig. 2).

**Methods used during analysis and validation**

Previously published CRC gene signatures were obtained from the geneSigDB database (26), and 461 genes appearing in 2 or more published gene signatures were identified for the analysis shown in Fig. 3. DASL microarray probes present in fewer than 50% of the NHS/HPFS CRC cohort samples ($P < 0.01$) were eliminated. Duplicate probes for a gene were averaged, leaving 330 of the 461 genes identified from the literature for differential expression analysis. Quality control was conducted as described above to identify samples passing strict QC (795), passing permissive QC (988), and poor samples only (193). Concordance was calculated as previously described (27), but repeated for multiple random splits of the samples. Further details of generation of the Concordance at the Top (CAT) boxplot are provided in the Supplementary Methods. For the investigation of previously reported gene signatures of MSI (28), only probes present in fewer than 10% of samples ($P < 0.01$) were discarded.
Results

A quality control pipeline for archival tissue gene expression microarrays

We developed an end-to-end QC methodology for whole-genome expression studies of archival tissues and validated it using 2 large novel sets of FFPE clinical samples. Our first dataset (referred to as BC/A) included 168 profiles of primary breast tumors, metastases, and control samples, including 44 technical replicates. The second dataset (referred to as NHS/HPFS) comprised 1,003 CRC patient tumor samples from two long-term epidemiological studies, the NHS (15) and the HPFS (16). These 2 datasets, generated in distant facilities, each showed a range of data quality substantially beyond that of previous smaller-scale FFPE studies, and our proposed quality control measures for samples and for microarray probes correctly identified (i) the most reproducible technical replicates and (ii) differential gene expression reproducibly segregating with tumor phenotype. These results emphasize the critical importance of stringent quality control, and the risk of high sample failure rate, when profiling the transcriptome through archival samples. These expression profiling quality control methods are available through the ffpe Bioconductor package.

Interquartile range as a general quality control metric

We investigated the dynamic range of expression intensities as a QC metric for microarray data from archival tissues. Dynamic range was summarized by the IQR of each array's raw expression values. Microarray experimental designs rarely include technical replication for all samples, so instead we generate a median "pseudochip" reference sample, constructed from the median intensity of each probe across all samples. The median pseudochip represents a study-typical sample under the assumption that the expression profiles represent similar cell types, so caution is necessary when using this method for QC of profiles originating from very different cell types. The combination of low IQR and low correlation to the median pseudochip enabled the identification of irreproducible expression profiles similarly to what could be achieved by technical replication in the BC/A cohort (Fig. 1).

IQR of raw log2 expression intensities correlated significantly with control probes for the DASL platform, including oligo annealing controls, array hybridization controls, and detection P values for each feature (29). IQR correlated most strongly to oligo annealing control probes ($r = 0.84; n = 1,003; P < 2.2e-16; Supplementary Fig. S2). Assessment of these various control probes can sometimes provide insight into mechanisms of individual sample failure,
Supplementary Fig. S5), emphasizing the importance of strict QC for reproducible biological inference.

Alternative probe normalization provided modest benefit beyond quantile normalization. We observed similar results for other CRC tumor phenotypes (Supplementary Fig. S5), emphasizing the importance of strict QC for reproducible biological inference.

Figure 3. Concordance at the Top boxplot showing reproducibility of identification of genes differentially expressed between colon and rectum, with varying quality control and data preprocessing approaches. Even under ideal conditions, the identification of any "top n" gene list differentially expressed with respect to a phenotype is affected by random sampling variation. This plot quantifies how much our quality-control measures improve such lists over randomly selected gene lists as a baseline. We identified 335 genes associated with CRC by inclusion in at least 2 CRC gene signatures from the geneSigDB database (26). We ranked these genes by fold-change between colon and rectum tumor location in 2 equal, independent subsets of the CRC cohort. The fractional overlap (concordance) of the top n genes in each list was calculated as a function of n, and the process repeated for 100 random splits of the samples to obtain the distribution of concordances shown in the boxplot. The diagonal dashed line indicates the baseline concordance of random gene lists. The distance above this line thus indicates the reproducibility of differentially expressed gene identification for poor samples only (rejected by our QC), permissive QC (samples with IQR > 0), strict QC (threshold automatically determined as in Fig. 1), and probe normalization (Illumina VST-RSN; ref. 17) and imputation (30) schemes. IQR-based quality control removes samples that independently show no reproducibility and improves reproducibility from the remaining samples, whereas alternative probe normalization provided modest benefit beyond quantile normalization. We observed similar results for other CRC tumor phenotypes (Supplementary Fig. S5), emphasizing the importance of strict QC for reproducible biological inference.

whereas IQR provides a general metric for expression data quality. For example, some sample failures were related rather than source tissue, as shown by low hybridization of sample-independent control probes (in particular the whole-chip failures on plates 4 and 5, Supplementary Fig. S3, also see Supplementary Fig. S2). Other sample failures, however, were not predicted by any control probes (in particular, those on plate 1, Supplementary Fig. S3). Furthermore, whereas control probes and associated probe detection calls are frequently unavailable for published datasets, IQR is widely available, making it a more general metric also for inter-study data comparison. For BC/A cohort samples with more than 2 replicates, low IQR (below 1 on the log2 scale in these data) was also indicative of low correlation to the median pseudochip of all replicates (Supplementary Fig. S4). The rate of QC rejection further depended significantly on the sample type in this cohort, with rejection rates of 15/50 for the first batch of matched primary tumors and lymph node metastases, 1/42 for 2 subsequent batches of select high-quality samples, and 14/18 for autopsy samples (control samples excluded, χ² test, P < 0.001, χ² = 36, df = 2). Dynamic range as measured by IQR thus provides a superset of the QC information in existing measures.

Expression data quality analysis of 1,003 colorectal cancer, 168 breast cancer, and 763 publicly available FFPE samples

We applied this QC methodology to microarrays from more than 1,900 FFPE and fresh-frozen tissue samples from 14 independent studies to assess the effects of sample type, source, and dataset size, and microarray platform on data quality. These included our CRC and BC/A samples described above, and we additionally identified 12 published studies including raw, unnormalized expression profiling of FFPE or FF tissues (Table 1). These studies used the Illumina WG-DASL platform (6, 10, 12, 18-21) as well as Affymetrix platforms with NuGEN-based sample preparation (1, 22-25). For comparison, we also included four datasets using the Illumina BeadArray platform for fresh-frozen tissues (6, 12, 18, 19). Relative to the large body of microarray expression studies using fresh-frozen tissues, very few have yet reported whole-genome profiling of archival FFPE tissues, and we believe that this meta-analysis contains all such publicly available datasets at this time.

Existing FFPE expression datasets were uniformly smaller than our CRC and BC/A clinical cohort datasets, exhibited less within-study variation of dynamic range, but showed large between-study variation (Fig. 2). This indicates that small, carefully controlled gene expression studies of FFPE tissues may not have captured the range of quality issues to be expected in larger studies. This suggests that assessments of data reproducibility based on any one such study may be overestimated, potentially because of factors including relatively homogeneous sample processing and preservation, to more consistent storage than is typical across multiple institutions, or to differences in tissue types or sample handling protocols. This difference is greater than could be remedied even by strict quality control of these larger population studies. Note that differences in absolute IQR...
Between Affymetrix and Illumina studies are not indicative of an overall difference in data quality between the platforms. We thus recommend that QC by dynamic range assessment be routinely applied to new FFPE expression profiling data, both for within-study quality control and for comparison to previous studies.

Stringent quality control methods improve reproducibility of differentially expressed gene lists

To show that these QC procedures improve biological (as well as technical) analyses, we modified the CAT plot method (27) to assess reproducible detection of differential gene expression with respect to CRC pathology phenotypes.
(Fig. 3). This enabled us to assess quantitatively the degree with which associations between gene expression and important pathological types of CRC could be reproducibly identified with microarray data from these archival tissues. We selected 330 genes published in 2 or more previous CRC studies, using the geneSigDB database (26), and used 2 equal, independent, subsets of our NHS/HPFS CRC samples to rank these genes for differential expression between colonic and rectal tumors. The overlap (concordance) in the top n genes of each list was calculated for 100 random splits of the samples (Fig. 3, results for other CRC tumor phenotypes in Supplementary Fig. S5). In addition to evaluating our sample QC method, we considered 3 different sample normalization strategies: $\log_2 + $ quantile normalization, Illumina-specific Variance Stabilizing Transformation + Robust Spline Normalization preprocessing (17), and $\log_2 + $ quantile with $k$-nearest-neighbors imputation (30) of expression values undetected ($P > 0.05$) by Illumina BeadStudio. All of these normalizations are well established but more complex alternatives to simple $\log_2 + $ quantile preprocessing. Quality control was the most important factor improving concordance of independently generated gene lists. Differences between all normalization methods were small relative to the differences induced by QC, underscoring the importance of both sample and probe QC in archival tissue gene expression relative to within-chip or within-dataset variability. Samples passing a "permissive QC" involving removing only samples where no hybridization occurred, but failing our "strict QC" IQR threshold, showed no independent ability to generate reproducible differentially expressed gene lists associated with tumor phenotype.

Variance as a quality control metric for individual probes

We assessed the reproducibility of measurements by individual probes across multiple samples, which has not been examined in FFPE samples. Previous investigations of whole-sample expression reproducibility from FFPE tissues have reported high correlation between replicate profiles (5–7, 10, 14); however, this gives no indication of the reproducibility of individual probes across multiple samples, nor of biological validity as examined above. While it has become standard to remove uninformative probes, FFPE studies to date have used diverse methods without experimental validation, including no probe removal (23), selection of probes with high concordance with matched fresh-frozen tissues (18), supervised phenotypic association (31), or variability across samples (7, 12).

We thus evaluated several measures for identifying probes with poor reproducibility: standard deviation across all samples, fraction of samples in which the probe was detected, mean expression, and coefficient of variance. In our set of 44 BC/A technical replicates, we assessed probes in 1 set of replicates and then calculated their resulting Spearman correlation across independently normalized pairs. We also considered Euclidian and Manhattan distance as measures of probe reproducibility, but found that these tended to favor probes with saturated intensities at the upper limit of detection (Supplementary Fig. S6). Higher variance probes showed better reproducibility as assessed by Spearman correlation (Fig. 4), as did probes at the high end of each of these measures (Supplementary Fig. S7). However, probes at the upper end of mean intensity or fraction of samples in which the probe was detected also contained invariant probes at their saturation intensity (Supplementary Fig. S8), so we recommend standard deviation for filtering probes. As expected, all filtering methods showed a trade-off between the number of probes retained and probe reliability. For general differential expression analyses, we suggest retaining probes with variance above the dataset median, although for purposes such as unsupervised clustering, a stricter filter such as the quintile with greatest variance may be beneficial, as used, for example, by Mittempergher and colleagues (12).

Validation of mRNA quantitation associated with tumor phenotype in long-term health study archival specimens

These results informed an end-to-end FFPE expression quality control pipeline, which we applied to the study of 1,003 CRC patients from 2 long-term prospective health studies. These samples presented the opportunity to investigate molecular cancer phenotypes in the context of long-term health and lifestyle patterns. They also highlighted the challenge of working with archival samples collected over decades from dozens of centers. These samples have been extensively studied for protein abundance, methylation, mutation, and genomic instability [see examples in Ogino and colleagues (2)]. We investigated the associations between transcript expression and CpG island methylation and protein abundance for 18 transcript–methylation/protein marker pairs (Supplementary Table S1), in addition to 41 gene transcripts previously reported to be differentially expressed in fresh-frozen CRC tumors with a high degree of MSI (28).

We considered "strict QC" as proposed here and "permissive QC" rejecting only clear failures where no hybridization occurred. While numerous factors can influence the relationship between mRNA transcript expression and protein abundance, we observed statistically significant correlations between mRNA transcript abundance and the corresponding molecular or protein change for 8 biomarkers (FDR $< 0.2$, Welch’s t test; see Supplementary Table S1): hypermethylation of CHFR and MGMT was associated with decreased corresponding transcript abundances; abundance of the MLH1, FASN, CDX2, MGMT, and SIRT1 proteins were positively associated with abundance of their gene transcripts; IGF2 DM1R0 (differentially methylated region) hypomethylation was associated with IGF2 transcript abundance. The direction of association was biologically consistent for each of these 8 molecular markers, and in each case the association was stronger with strict QC than with permissive QC. Critically for accurate experimental follow-up in new studies, 3 of the 8 markers were identified only by using strict quality control (Supplementary Fig. S9).
In spite of large variations in expression data quality, these quality control steps also allowed us to reproduce previously reported associations between MSI and 23 of 25 upregulated transcripts (92%) and 15 of 16 downregulated transcripts (94%, Supplementary Fig. S10). In 36 of 41 transcripts (88%), the strength of the expected association was improved by the proposed strict sample QC as compared with permissive sample QC \( P < 3 \times 10^{-6}, \chi^2 \) test). Correspondingly, in whole-genome discovery, these previously reported transcripts indeed tended to be differentially expressed with respect to MSI in the NHS/HPFS samples, and this tendency improved with strict QC (Fig. 5A). We noted that rare cases of stronger associations with permissive QC than strict QC occurred only in the most highly expressed transcripts (Supplementary Fig. S11), suggesting that for these transcripts, some detectable signal remains even in poor-quality samples. However, even among these most highly expressed transcripts, strict QC still improved the expected association for a majority of probes (11 of 16 probes in the top 80th percentile of intensities, and 4 of 7 probes in the top 90th percentile). Furthermore, the differential expression of genes with high-variance probes were more likely to be validated than those with low-variance probes (Fig. 5B), in keeping with the findings from technical replication in Fig. 4. In conjunction with the results above, this indicates that strict IQR-based sample quality control and variance-based probe QC enable both better reproducibility of archival tissue expression data and more accurate associations with phenotype.

Discussion

We established and validated quality control metrics for expression profiling of FFPE tissues at the level of study, sample, and individual gene probe. We propose IQR as a summary metric for study and sample quality assessment, which enabled a comparison of archival tissue microarray quality from 14 studies spanning six different platforms and both major RNA labeling and amplification technologies (Illumina DASL and NuGEN Ovation). These metrics proved to be critical to the effective analysis of gene expression in diverse archival samples, and they provide experimentally validated quality control methods to enable such analyses for clinical microarray data. Specifically, we applied these methods to a novel microarray study of more than 1,000 archival clinical samples of diverse storage age and origin from participants in 2 long-term prospective health studies (15, 16). The ability to validate expression of mRNA transcripts differential with respect to tissue of origin, epigenetics, and MSI were established and substantially improved by the application of strict quality measures introduced here, in spite of those measures resulting in the removal of approximately 20% of unrecoverable archival samples. Meta-analysis of variation in expression data quality in published studies emphasized that these smaller studies, with relatively homogeneous sample sources, are not representative of the greater sample quality variability to be found in larger, multicenter or population studies.

It is important to emphasize that gene expression measurements from archival tissues present greater levels of noise and of complete sample failure than corresponding measurements from high-quality frozen tissues. However, these technical considerations need not impede diagnostic or prognostic biomarker development from FFPE tissues when proper care is taken. The detection of differentially expressed genes is one of many diverse applications of whole-genome expression profiling from either FFPE or FF tissues, which can range from multivariate prognostic model development to discovery of gene coexpression networks. Initial studies have shown coordinated changes in transcript abundance through the FFPE process compared with FF tissues, evidenced by lower reproducibility between FFPE and FF tissues than between replicate FFPE tissues (6). This is not a problem for clinical biomarkers and predictive models both developed and applied in FFPE tissues, but should be taken into consideration when such models are applied across FFPE and FF tissues or when studying coexpression. Few examples yet exist of prediction models being validated between FF and FFPE tissues (1), and any such validation is likely to be gene, tissue, and platform-specific and should not be assumed to generalize. Predictive models focusing exclusively on archival tissue gene expression profiling are thus a promising area of specific focus in the future.

As with many analyses of tumor tissues, it is important to consider sample-specific features such as tissue heterogeneity, inflammatory cell content, and necrosis when applying these QC measures in any given dataset. In the diverse datasets considered here, the combination of both a low sample quality score (such as IQR) and a low correlation to a study-specific "typical" profile together provided strong evidence of low quality expression data, as well as deriving a study-specific quality rejection threshold. In most studies, this will also incorporate information on "typical" cellularity or necrosis, but low correlation to the median profile may also occur if the study includes very different samples (e.g., from completely different tissues). In such cases, it may be desirable to stratify quality analysis within multiple subsets of more homogeneous, directly comparable sample groups.

An additional emerging technology that will support such studies is expression profiling by RNA-sequencing, which has the advantage of sequencing all short cDNA fragments, without a priori selection of oligonucleotide transcripts that may have been fragmented during preservation and storage (32). Related platforms remain relatively untested compared with microarray assays, but they are at best also dependent on PCR amplification and sample history and cannot be expected to abrogate these issues. Quality control and awareness of the technical variability of clinical samples will remain crucial for sequencing-based biomarkers, and we anticipate that our quality control process and the dynamic range of summarized expression intensities will continue to provide a valuable assessment of expression data quality.

Opening the vast archives of FFPE tissues to high-throughput expression profiling is critical to the development of
clinically relevant biomarkers and to the genomic study of cancer in relation to health and lifestyle. Virtually all important molecular pathologic tests make use of FFPE tissues (1), and the current lack of clinically significant gene expression biomarkers (33) is due in part to inability to make full use of these tissues. The use of FFPE tissues in gene expression studies will not only increase potential sample size and follow-up time, but also have direct relevance to the tissues actually used in clinical pathology. A new breadth of studies of environmental interactions with gene expression for human disease populations will also become possible by making use of archival tissues from long-term, prospective health studies, for example the investigation of transcriptional mechanisms mediating epidemiologically established cancer risk factors such as that dietary B-vitamin intake (34, 35). However, this study also highlighted the risks involved in studying the human transcriptome using archival samples, because of potentially high rates of sample failure. This risk is best assessed through pilot studies of the actual samples at hand and comparisons with published data, and should be considered during early study planning stages. With due care to such issues, the move toward usage of clinically available FFPE tissues will represent a major shift in the translational and population study of gene expression.

Disclosure of Potential Conflicts of Interest

P.T. Simpson has honoraria from speakers bureau and has received funding from Illumina to attend a conference (travel and accommodation costs) and present a seminar about the WG-DASL assay. C. Huttenhower has a commercial research grant from DANAONE and honoraria from speakers bureau from BioGen. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: S. Ogino, J. Quackenbush, C.S. Fuchs, G. Parmigiani, C. Huttenhower

References


14. Coudry RA, Meereles St, Stoyanova R, Cooper HS, Carpino A, Wang X, et al. Successful application of microarray technology to

Development of methodology: L. Waldron, C.S. Fuchs, G. Parmigiani, C. Huttenhower

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): S. Ogino, Y. Hoshida, K. Shima, A.E.M. Reed, P.T. Simpson, Y. Baba, K. Nosho, A.C. Vargas, M.C. Cummings, S.R. Lakhan, G.J. Kirker, T.R. Golub, C.S. Fuchs

Analysis and interpretation of data (e.g., statistical analysis, bioinformatics, computational analysis): L. Waldron, S. Ogino, A.E.M. Reed, P.T. Simpson, N. Segata, E. Giovannucci, J. Quackenbush, C.S. Fuchs, G. Parmigiani, C. Huttenhower

Writing, review, and/or revision of the manuscript: L. Waldron, S. Ogino, A.E.M. Reed, P.T. Simpson, S.R. Lakhan, G.J. Kirker, E. Giovannucci, J. Quackenbush, C.S. Fuchs, G. Parmigiani, C. Huttenhower

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): S. Ogino, A.E.M. Reed, G.J. Kirker, C.S. Fuchs

Study supervision: S. Ogino, C.S. Fuchs, G. Parmigiani, C. Huttenhower

Acknowledgments

The authors thank Aditi Hazra, Lorelei Mucci, and Walter Willett, Benjamin Haibe-Kains, Sibylle Cocciardi, Georgia Chenexis-Trench, and Brian Fritz for their contributions to this work.

Grant Support

This work was supported by U.S. National Institute of Health (NIH) grants P01 CA087969 (to S.E. Hankinson), P01 CA50575 (to W.C. Willett), PS0 CA127003 (to C.S. Fuchs), and RO1 CA151993 (to S. Ogino), by the National Science Foundation grant NSF DBI-1053446 (to C. Huttenhower), and by grants from DFCI Friends, the Bennett Family Fund, the Entertainment Industry Foundation through National Colorectal Cancer Research Alliance, and the Wesley Research Institute, Australia. P.T. Simpson and A.C. Vargas are recipients of fellowships from the National Breast Cancer Foundation, Australia and the Ludwig Institute of Cancer Research, respectively. The content is solely the responsibility of the authors and does not represent the official views of any funders. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received June 13, 2012; revised August 21, 2012; accepted August 22, 2012; published OnlineFirst November 7, 2012.
Expression Profiling of Archival Tumors for Long-term Health Studies

Levi Waldron, Shuji Ogino, Yujin Hoshida, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-12-1915

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2012/11/15/1078-0432.CCR-12-1915.DC1

Cited articles
This article cites 30 articles, 7 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/18/22/6136.full.html#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.