Published in final edited form as:
Clin Cancer Res. 2012 November 15; 18(22): 6136–6146. doi:10.1158/1078-0432.CCR-12-1915.

Expression profiling of archival tumors for long-term health studies

Levi Waldron[*], **Shuji Ogino**[*], **Yujin Hoshida**, **Kaori Shima**, **Amy E McCart Reed**, **Peter T Simpson**, **Yoshifumi Baba**, **Katsuhiko Nosho**, **Nicola Segata**, **Ana Cristina Vargas**, **Margaret Cummings**, **Sunil R Lakhani**, **Gregory J. Kirkner**, **Edward Giovannucci**, **John Quackenbush**, **Todd R. Golub**[*], **Charles S. Fuchs**[*], **Giovanni Parmigiani**[*], and **Curtis Huttenhower**[*]
Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA (SO, KS, YB, KN, CSF)

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA (LW, GP, JQ)

Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA (LW, NS, GP, CH)

Departments of Epidemiology and Nutrition, Harvard School of Public Health, Boston, MA, USA (EG)

Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA (SO)

Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA (GJK, EG, CSF)

The University of Queensland, UQ Centre for Clinical Research, Herston, Brisbane, QLD, Australia (AEMR, PTS, ACV, MC, SRL)

Pathology Queensland: The Royal Brisbane & Women's Hospital, Brisbane, QLD, Australia (MC, SRL)

The University of Queensland, School of Medicine, Herston, Brisbane, QLD, Australia (SRL)

Broad Institute of Harvard and MIT, Cambridge, MA, USA (YH, TG)

Abstract

Purpose—Over 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,

Corresponding Author: Curtis Huttenhower, 655 Huntington Avenue, Boston, MA 02115, Phone: 617.432.4912, chuttenh@hsph.harvard.edu.
[*]equal contribution

Conflicts of Interest: The authors declare no conflicts of interest.

comparison to results from fresh-frozen tissue, and recovery of expected associations between gene expression and protein abundance.

Results—Archival tissues from large, multi-center studies demonstrated 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.

Keywords

FFPE; formalin; gene expression profiling

Introduction

Formalin fixation and paraffin embedding (FFPE) is the tissue preservation method for virtually all routine histopathology tests(1), and excised tumors are routinely archived as FFPE blocks. These samples are preserved for decades and, as such, can be utilized 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 demonstrated 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(5, 6), and specimen quality control is only one component 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 two 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 performed 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 utilize platform-specific control probes, IQR was universally available for all published datasets, enabling consistent data quality assessment in published FFPE expression profiles as compared to 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 demonstrated 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)(15) and Health Professionals Follow-up Study (HPFS)(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 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 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 two large, novel gene expression profiling studies using archival, clinical tissues. We performed 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 BC/A dataset). In the second study, we profiled 1,003 tumors from colorectal cancer patients drawn from the Nurses' Health Study and Health Professionals Follow-up Study (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 GEO under accession numbers GSE32651 and GSE32490.

Patient Samples

We utilized two U.S. Nationwide prospective cohort studies, the Nurses' Health Study (NHS)(15) and the Health Professionals Follow-up Study (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 Supplemental Methods.

NHS/HPFS Cohort Metadata—Clinicopathological and epidemiological data for the NHS/HPFS cohorts were extracted by SAS script in accordance with NHS and HPFS Cohort program review procedures. Immunohistochemistry (IHC), DNA methylation and microsatellite instability (MSI) methods are described in the Supplemental Methods, and representative IHC stains for *CCND1* and *SIRT1* are shown in Supplemental Figure 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 nine matched pairs of replicate samples from primary breast tumors and lymph node metastases. In cases with more than two replicates available, two replicates were selected randomly. The sets of replicates were separated and quantile normalized independently, and one set was used to calculate standard deviation of each probe across the nine samples (or other measures of probe activity shown in Supplemental Figure 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 performed 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 Interquartile Range (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.

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 (Figure 1), behavior in 12 publicly available datasets (Figure 2), and performance on 1,003 CRC samples (Figures 3 and 5).

Expression data pre-processing—Probes present in fewer than 10% of NHS/HPFS samples remaining after sample quality control (nominal $p<0.01$) were removed (4,476 probes) prior to normalization. Two alternative methods of data transformation and

normalization were considered: $\log_2$ transformation followed by quantile normalization, and Variance Stabilizing Transformation followed by Robust Spline Normalization(28). Optional imputation of missing expression values ($p>0.01$) was performed 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 (Interquartile Range 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. Based on performance in BC/A technical replicates (Figure 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(6, 10, 12, 18–21) (Table 1), four example datasets employing 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 interquartile ranges of raw log2 intensities for each sample (Figure 2).

Methods used during analysis and validation

Previously published CRC gene signatures were obtained from the geneSigDB database(27), and 461 genes appearing in two or more published gene signatures were identified for the analysis shown in Figure 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 performed 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 (17), but repeated for multiple random splits of the samples. Further details of generation of the CAT-boxplot are provided in the Supplementary Methods. For the investigation of previously reported gene signatures of Microsatellite Instability(31), 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 quality control (QC) methodology for whole-genome expression studies of archival tissues and validated it using two large novel sets of FFPE clinical samples. Our first dataset (referred to as breast cancer/autopsy, 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 colorectal cancer (CRC) patient tumor samples from two long-term epidemiological studies, the Nurses' Health Study(15) and the Health Professionals Follow-up Study(16). These two datasets, generated in distant facilities, each demonstrated 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 A) the most reproducible technical replicates and B) 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 interquartile range (IQR) of each array's raw gene 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 unreproducibile expression profiles similarly to what could be achieved by technical replication in the BC/A cohort (Figure 1).

IQR of raw $\log_2$ 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(18). IQR correlated most strongly to oligo annealing control probes ($r=0.84$, $n=1,003$, $p<2.2e{-}16$, Supplemental Figure S2). Assessment of these various control probes can sometimes provide insight into mechanisms of individual sample failure, whereas IQR provides a general metric for expression data quality. For example, some sample failures were related to assay 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, Supplemental Figure S3, also see Supplemental Figure S2). Other sample failures, however, were not predicted by any control probes (in particular, those on plate 1, Supplemental Figure 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 two replicates, low IQR (below 1 on the $\log_2$ scale in these data) was also indicative of low correlation to the median pseudochip of all replicates (Supplemental Figure 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 two subsequent batches of select high-quality samples, and 14/18 for autopsy samples (control samples excluded, Chi-square test, $p < 0.001$, $\chi 2 = 36$, $df=2$). 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 over 1,900 FFPE and fresh-frozen tissue samples from 12 independent studies in order 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 employed the Illumina WG-DASL platform(6, 10, 12, 19–22) as well as Affymetrix platforms with NuGEN-based sample preparation(1, 23–26). For comparison, we also included four datasets using the Illumina Bead Array platform for fresh-frozen tissues(6, 12, 19, 20). 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 (Figure 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 over-estimated, potentially due to 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 demonstrate that these QC procedures improve biological (as well as technical) analyses, we modified the Concordance at the Top (CAT) plot method(27) to assess reproducible detection of differential gene expression with respect to CRC pathology phenotypes (Figure 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 two or more previous CRC studies, using the geneSigDB database(28), and used two 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 (Figure 3, results for other CRC tumor phenotypes in Supplemental Figure S5). In addition to evaluating our sample QC method, we considered three different sample normalization strategies: $\log_2$ + quantile normalization, Illumina-specific Variance Stabilizing Transformation + Robust Spline Normalization preprocessing(29), 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, 9, 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 employed diverse methods without experimental validation, including no probe removal(24), selection of probes with high concordance with matched fresh-frozen tissues(12), supervised phenotypic association(19, 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 one 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 (Supplemental Figure S6). Higher variance probes showed better reproducibility as assessed by Spearman correlation (Figure 4), as did probes at the high end of each of all of these measures (Supplemental Figure 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 (Supplemental Figure S8), so we recommend standard deviation for filtering probes. As expected, all filtering methods demonstrated 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 employed, for example, by Mittempherger *et al.*(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 two 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 *et al.(2)*). We investigated the associations between transcript expression and CpG island methylation and protein abundance for 18 transcript – methylation/protein marker pairs (Supplemental Table S1), in addition to 41 gene transcripts previously reported to be differentially expressed in fresh-frozen CRC tumors with ahigh degree of microsatellite instability (MSI-high)(32).

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 eight biomarkers (FDR < 0.2, Welch’s t-test; see Supplemental 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 DMR0 (differentially methylated region) hypomethylation was associated with IGF2 transcript abundance. The direction of association was biologically consistent for each of these eight 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, three of the eight markers were identified only by using strict quality control (Supplemental Figure 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 up-regulated transcripts (92%) and 15 of 16 down-regulated transcripts (94%, Supplemental Figure S10). In 36 of 41 transcripts (88%), the strength of the expected association was improved by the proposed strict sample QC as compared to permissive sample QC ($p < 3 \times 10^{-6}$, chi-square 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 (Figure 5A). We noted that rare cases of stronger associations with permissive QC than with strict QC occurred only in the most highly expressed transcripts (Supplemental Figure 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 (Figure 5B), in keeping with the findings from technical replication in Figure 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 Interquartile Range(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 of the two 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 over 1,000 archival clinical samples of diverse storage age and origin from participants in two long-term prospective health studies(15, 16). The ability to validate expression of mRNA transcripts differential with respect to tissue of origin, epigenetics, and microsatellite instability (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, multi-center 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 to 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. While few examples yet exist of prediction models being validated between FF and FFPE tissues(1), 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. Related platforms remain relatively untested compared to microarray assays(33), 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(34) 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(35, 36). However, this study also highlighted the risks involved in studying the human transcriptome using archival samples, due to 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 towards utilization of clinically available FFPE tissues will represent a major shift in the translational and population study of gene expression.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Aditi Hazra, Lorelei Mucci, and Walter Willett, Benjamin Haibe-Kains, Sibylle Cocciardi, Georgia Chenevix-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 CA55075 (to W.C. Willett), P50 CA127003 (to CF), R01 CA151993 (to SO), by the National Science Foundation grant NSF DBI-1053486 (to CH), 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. PTS and ACV are recipients of fellowships from the National Breast Cancer Foundation, Australia and the Ludwiq 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.

References

1. Williams PM, Li R, Johnson NA, Wright G, Heath J-D, Gascoyne RD. A Novel Method of Amplification of FFPET-Derived RNA Enables Accurate Disease Classification with Microarrays. The Journal of Molecular Diagnostics. 2010; 12:680–686. [PubMed: 20688907]
2. Ogino S, Chan A, Fuchs C, Giovannucci E. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. Gut. 2011; 60:397–808. [PubMed: 21036793]
3. Ogino S, Galon J, Fuchs C, Dranoff G. Cancer immunology--analysis of host and tumor factors for personalized medicine. Nature reviews Clinical oncology. 2011; 8:711–720.
4. Lewis F, Maughan NJ, Smith V, Hillan K, Quirke P. Unlocking the archive – gene expression in paraffin-embedded tissue. The Journal of Pathology. 2001; 195:66–71. [PubMed: 11568892]
5. Reinholz MM, Eckel-Passow JE, Anderson SK, Asmann YW, Zschunke MA, Oberg AL, et al. Expression profiling of formalin-fixed paraffin-embedded primary breast tumors using cancer-specific and whole genome gene panels on the DASL platform. BMC Medical Genomics. 2010; 3:60. [PubMed: 21172013]
6. April C, Klotzle B, Royce T, Wickham-Garcia E, Boyaniwsky T, Izzo J, et al. Whole-Genome Gene Expression Profiling of Formalin-Fixed, Paraffin-Embedded Tissue Samples. PLoS ONE. 2009; 4:e8162. e. [PubMed: 19997620]
7. Ton CC, Vartanian N, Chai X, Lin MG, Yuan X, Malone KE, et al. Gene expression array testing of FFPE archival breast tumor samples: an optimized protocol for WG-DASL sample preparation. Breast Cancer Research and Treatment. 2010
8. Pillai R, Deeter R, Rigl CT, Nystrom JS, Miller MH, Buturovic L, et al. Validation and reproducibility of a microarray-based gene expression test for tumor identification in formalin-fixed, paraffin-embedded specimens. The Journal of Molecular Diagnostics: JMD. 2011; 13:48–56. [PubMed: 21227394]
9. Grenert JP, Smith A, Ruan W, Pillai R, Wu AHB. Gene expression profiling from formalin-fixed, paraffin-embedded tissue for tumor diagnosis. Clinica chimica acta; international journal of clinical chemistry. 2011
10. Waddell N, Cocciardi S, Johnson J, Healey S, Marsh A, Riley J, et al. Gene expression profiling of formalin-fixed, paraffin-embedded familial breast tumours using the whole genome-DASL assay. The Journal of Pathology. 2010; 221:452–461. [PubMed: 20593485]
11. Roberts L, Bowers J, Sensinger K, Lisowski A, Getts R, Anderson MG. Identification of methods for use of formalin-fixed, paraffin-embedded tissue samples in RNA expression profiling. Genomics. 2009; 94:341–348. [PubMed: 19660539]
12. Mittempergher L, de Ronde JJ, Nieuwland M, Kerkhoven RM, Simon I, Th. Rutgers EJ, et al. Gene Expression Profiles from Formalin Fixed Paraffin Embedded Breast Cancer Tissue Are Largely Comparable to Fresh Frozen Matched Tissue. PLoS ONE. 2011; 6:e17163. e. [PubMed: 21347257]
13. Lassmann S, Kreutz C, Schoepflin A, Hopt U, Timmer J, Werner M. A novel approach for reliable microarray analysis of microdissected tumor cells from formalin-fixed and paraffin-embedded colorectal cancer resection specimens. Journal of Molecular Medicine. 2008; 87:211–224. [PubMed: 19066834]
14. Coudry RA, Meireles SI, Stoyanova R, Cooper HS, Carpino A, Wang X, et al. Successful Application of Microarray Technology to Microdissected Formalin-Fixed, Paraffin-Embedded Tissue. J Mol Diagn. 2007; 9:70–79. [PubMed: 17251338]
15. Colditz GA, Hankinson SE. The Nurses' Health Study: lifestyle and health among women. Nat Rev Cancer. 2005; 5:388–396. [PubMed: 15864280]

16. Morikawa T, Kuchiba A, Yamauchi M, Meyerhardt JA, Shima K, Nosho K, et al. Association of CTNNB1 (beta-catenin) alterations, body mass index, and physical activity with survival in patients with colorectal cancer. JAMA. 2011; 305:1685–1694. [PubMed: 21521850]

17. Irizarry R, Warren D, Spencer F, Kim I, Biswal S, Frank B, et al. Multiple-laboratory comparison of microarray platforms. Nature methods. 2005; 2:345–395. [PubMed: 15846361]

18. Illumina, I. Illumina, Inc. 2010.

19. Sadi AM, Wang D-Y, Youngson BJ, Miller N, Boerner S, Done SJ, et al. Clinical relevance of DNA microarray analyses using archival formalin-fixed paraffin-embedded breast cancer specimens. BMC Cancer. 2011; 11:253. [PubMed: 21679412]

20. Jiang X, Tan J, Li J, Kivimae S, Yang X, Zhuang L, et al. DACT3 is an epigenetic regulator of Wnt/beta-catenin signaling in colorectal cancer and is a therapeutic target of histone modifications. Cancer Cell. 2008; 13:529–541. [PubMed: 18538736]

21. Villanueva A, Hoshida Y, Battiston C, Tovar V, Sia D, Alsinet C, et al. Combining Clinical, Pathology, and Gene Expression Data to Predict Recurrence of Hepatocellular Carcinoma. Gastroenterology. 2011; 140:1501–1512. [PubMed: 21320499]

22. Minguez B, Hoshida Y, Villanueva A, Toffanin S, Cabellos L, Thung S, et al. Gene-expression signature of vascular invasion in hepatocellular carcinoma. J Hepatol. 2011

23. Abdueva D, Wing M, Schaub B, Triche T, Davicioni E. Quantitative expression profiling in formalin-fixed paraffin-embedded samples by affymetrix microarrays. The Journal of Molecular Diagnostics: JMD. 2010; 12:409–417. [PubMed: 20522636]

24. Budczies J, Weichert W, Noske A, Muller BM, Weller C, Wittenberger T, et al. Genome-wide gene expression profiling of formalin-fixed paraffin-embedded breast cancer core biopsies using microarrays. J Histochem Cytochem. 2011; 59:146–157. [PubMed: 21339180]

25. Hall JS, Leong HS, Armenoult LSC, Newton GE, Valentine HR, Irlam JJ, et al. Exon-array profiling unlocks clinically and biologically relevant gene signatures from formalin-fixed paraffin-embedded tumour samples. Br J Cancer. 2011; 104:971–981. [PubMed: 21407225]

26. Collisson EA, Sadanandam A, Olson P, Gibb WJ, Truitt M, Gu S, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat Med. 2011; 17:500–503. [PubMed: 21460848]

27. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, et al. Multiple-laboratory comparison of microarray platforms. Nat Meth. 2005; 2:345–350.

28. Culhane AıC, Schwarzl T, Sultana R, Picard KC, Picard SC, Lu TH, et al. GeneSigDB–a curated database of gene expression signatures. Nucleic acids research. 2010; 38 gkp1015+-gkp+

29. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics (Oxford, England). 2008; 24:1547–1548.

30. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17:520–525. [PubMed: 11395428]

31. Conway C, Mitra A, Jewell R, Randerson-Moor J, Lobo S, Nsengimana J, et al. Gene expression profiling of paraffin-embedded primary melanoma using the DASL assay identifies increased osteopontin expression as predictive of reduced relapse-free survival. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research. 2009; 15:6939–6946. [PubMed: 19887478]

32. Kim H, Nam SW, Rhee H, Shan Li L, Ju Kang H, Hye Koh K, et al. Different gene expression profiles between microsatellite instability-high and microsatellite stable colorectal carcinomas. Oncogene. 2004; 23:6218–6225. [PubMed: 15208663]

33. Nagalakshmi U, Waern K, Snyder M. RNA-Seq: a method for comprehensive transcriptome analysis, Chapter 4. Curr Protoc Mol Biol. 2010; 11:1–3. Unit 4.

34. Koscielny S. Why Most Gene Expression Signatures of Tumors Have Not Been Useful in the Clinic. Science Translational Medicine. 2010; 2 14ps2-ps2.

35. Giovannucci E, Rimm EB, Ascherio A, Stampfer MJ, Colditz GA, Willett WC. Alcohol, lowmethionine--low-folate diets, and risk of colon cancer in men. Journal of the National Cancer Institute. 1995; 87:265–273. [PubMed: 7707417]

36. Giovannucci E, Stampfer MJ, Colditz GA, Hunter DJ, Fuchs C, Rosner BA, et al. Multivitamin use, folate, and colon cancer in women in the Nurses' Health Study. Annals of Internal Medicine. 1998; 129:517–524. [PubMed: 9758570]

Statement of Translational Relevance

Formalin fixation and paraffin embedding (FFPE) has been the tissue preservation method of choice for nearly a century. Gene expression profiling of such tissues is challenging due to RNA degradation, but they are the only samples available for biomarker discovery in large cohorts and with long-term clinical follow-up. They are also an avenue by which molecular biomarkers can be integrated directly into standard clinical practice. We present the first large-scale transcription profiling studies of FFPE tissues, including 1,003 colorectal cancer samples, 168 breast cancer samples, and a meta-analysis of 14 new and published datasets. We show that while FFPE samples of diverse origin require novel and extensive quality control methods when used for gene expression profiling, they can provide valid and reproducible biomarkers. Strict quality control methods, in fact, proved more important to the discovery of reproducible biomarkers from these data than any other aspect of data processing.

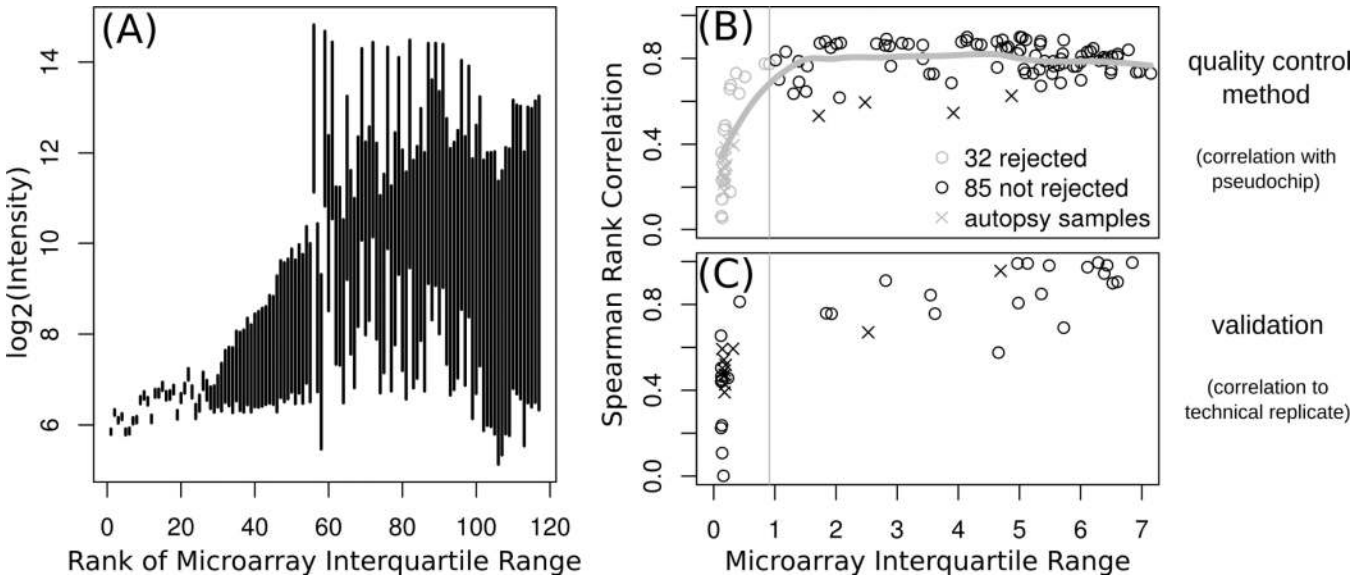


Figure 1. Novel quality control methodology for gene expression data from archival tissues, validated using a technical study of primary and metastatic breast tumors and of autopsy tissue samples

(A) Raw gene expression intensities from studies employing FFPE tissues often possess highly variable dynamic ranges, which can serve as a reliable quality control measure. Lines indicate the 25th to 75th percentile of log-scaled transcript levels (interquartile range, IQR) for 117 samples passing preliminary quality control (10% of features present at $p<0.01$), sorted from lowest to highest IQR. **(B)** In order to detect poor-quality samples, each sample IQR is compared with its Spearman correlation to the study median. A combination of low resulting correlation with low IQR identifies poor quality data, which we specifically threshold at the point of maximum downward inflection of a Loess smoothing line. Autopsy samples (labelled "x") are disproportionately of low quality. **(C)** The BC/A study included 44 technical replicates; we applied this IQR criterion to the samples, here showing the minimum IQR of each replicate pair compared with Spearman correlation of the replicates. Low reproducibility is seen between replicates falling below our IQR-based quality control threshold.

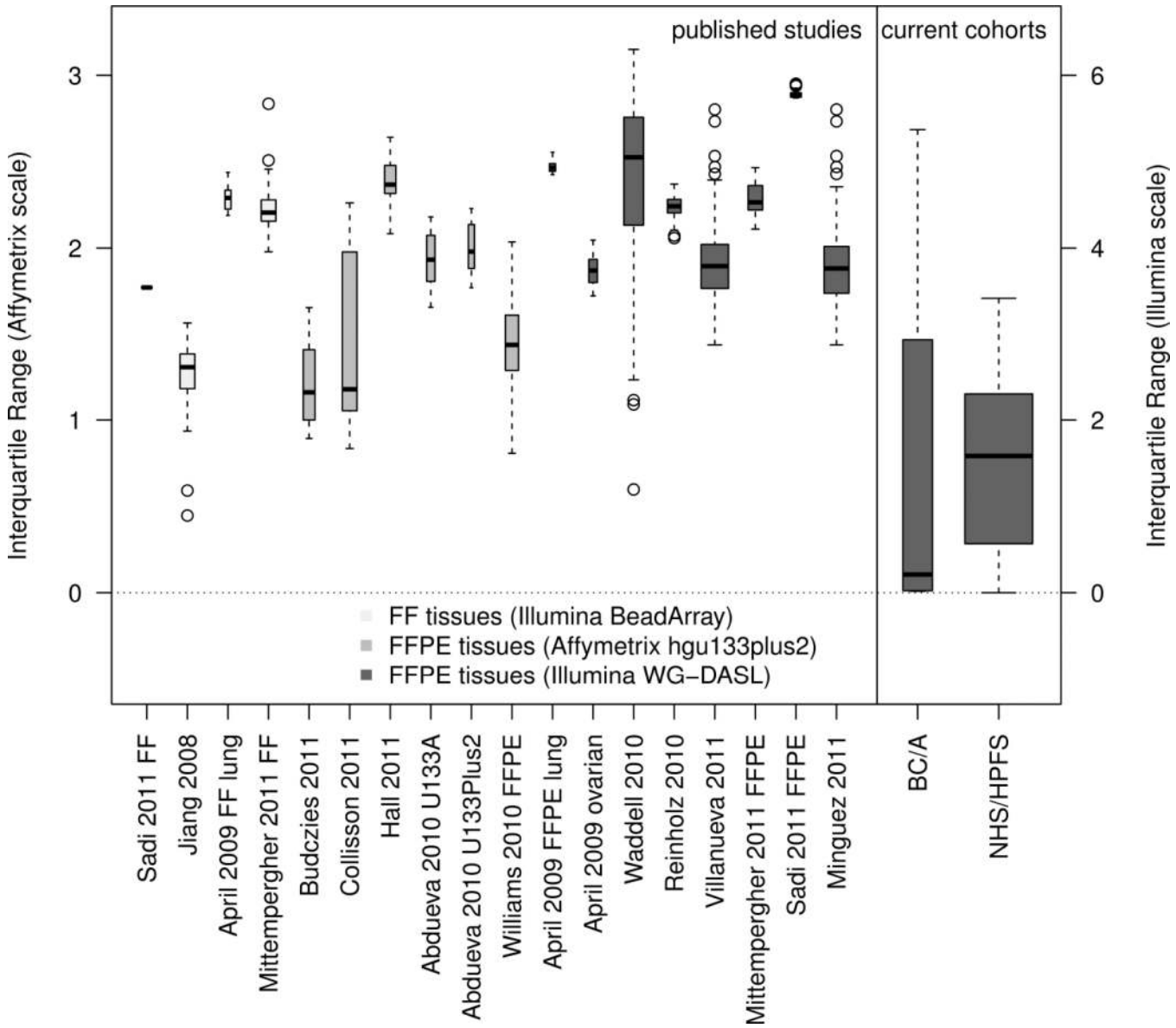


Figure 2. Population-scale gene expression studies of archival tissue samples possess wide dynamic ranges that must be quality controlled

Dynamic ranges of expression intensities from 20 datasets in 14 independent studies are shown; these studies are summarized in Table 1. Two y-axes are used, for the Illumina and Affymetrix platforms, as these technologies provide intensities on different scales that should not be directly compared and do not indicate a between-platform quality difference. Box width is proportional to the square root of study sample size; box limits indicate *per-study interquartile range* of the *individual, per-sample interquartile ranges* of intensities for samples in that study. Thus short boxes indicate relatively uniform sample quality and dynamic range, while taller boxes indicate studies containing a wide range of data quality. Smaller, focused studies are often of atypically high quality and do not capture the range of sample qualities observed in population-scale studies of diverse clinical samples, as seen in these two large studies of breast cancer and autopsy samples (BC/A) and colorectal tumors from the NHS/HPFS prospective cohorts (CRC).

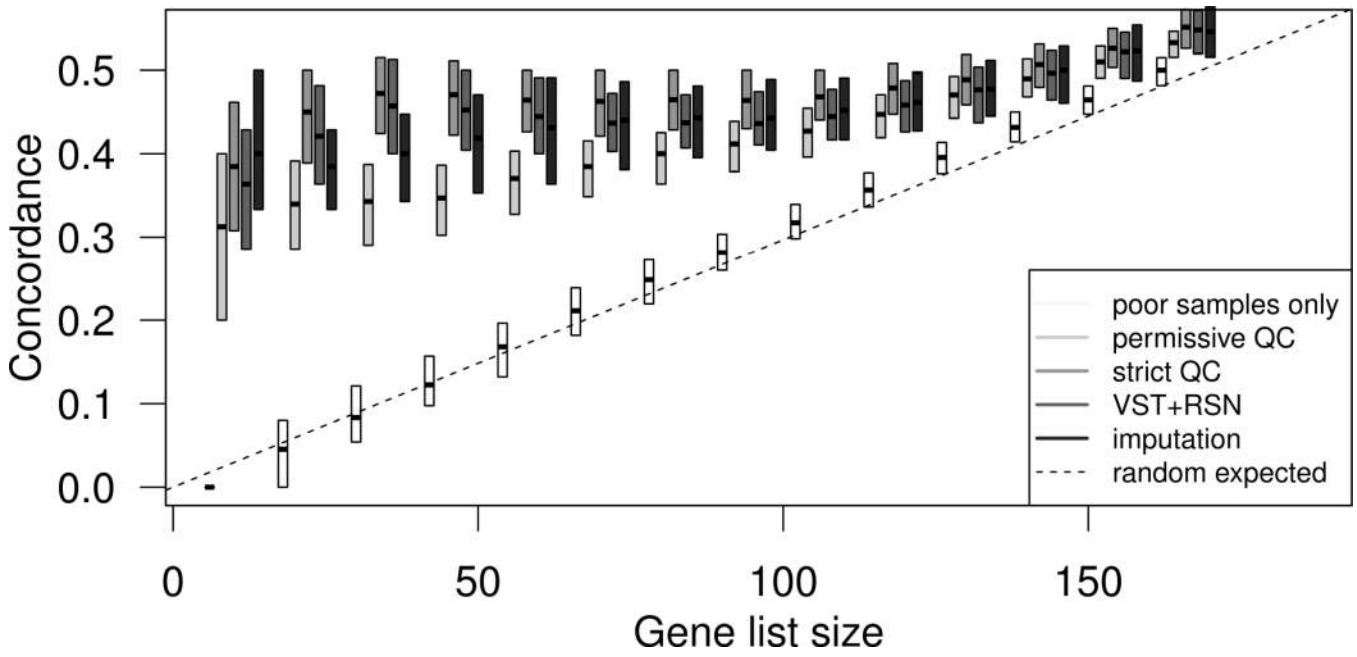


Figure 3. Concordance At the Top boxplot demonstrating reproducibility of identification of genes differentially expressed between colon and rectum, with varying quality control and data pre-processing approaches

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 330 genes associated with CRC by inclusion in at least two CRC gene signatures from the geneSigDB database(28). We ranked these genes by fold-change between colon and rectum tumor location in two 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 Figure 1), and probe normalization (Illumina VST+RSN(29) 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 (Supplemental Figure S5), emphasizing the importance of strict QC for reproducible biological inference.

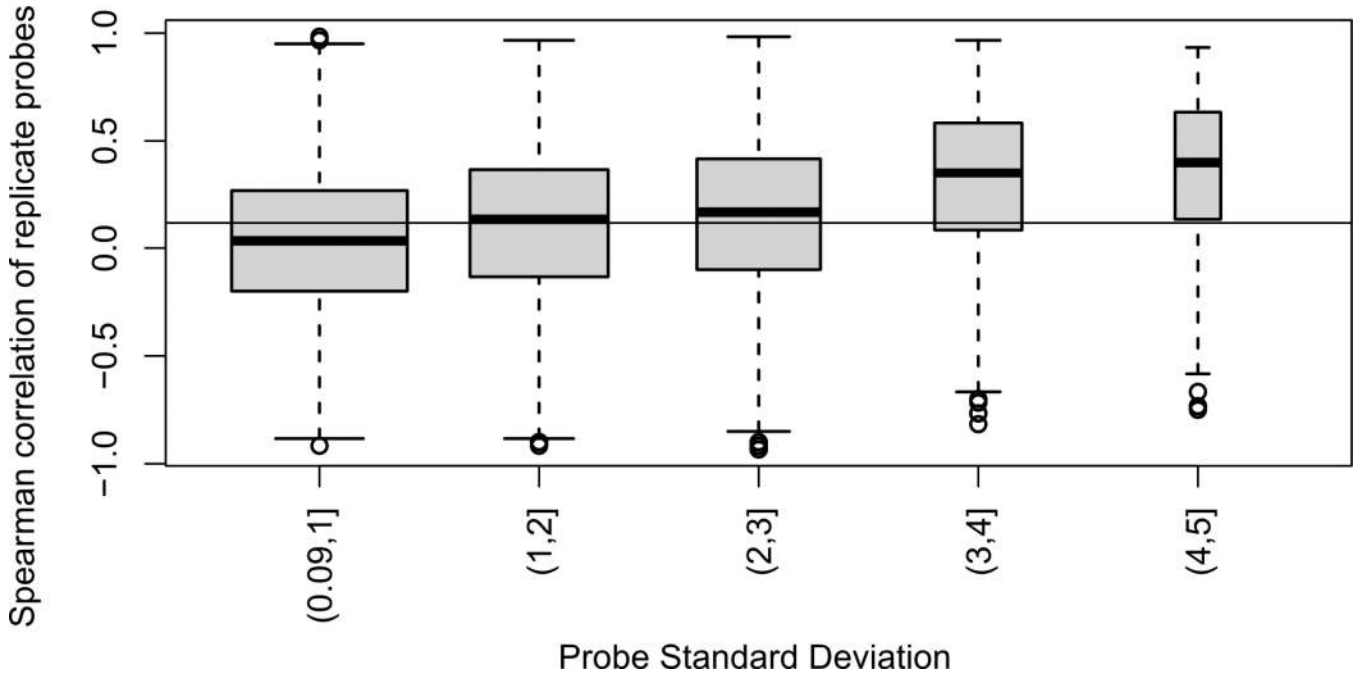
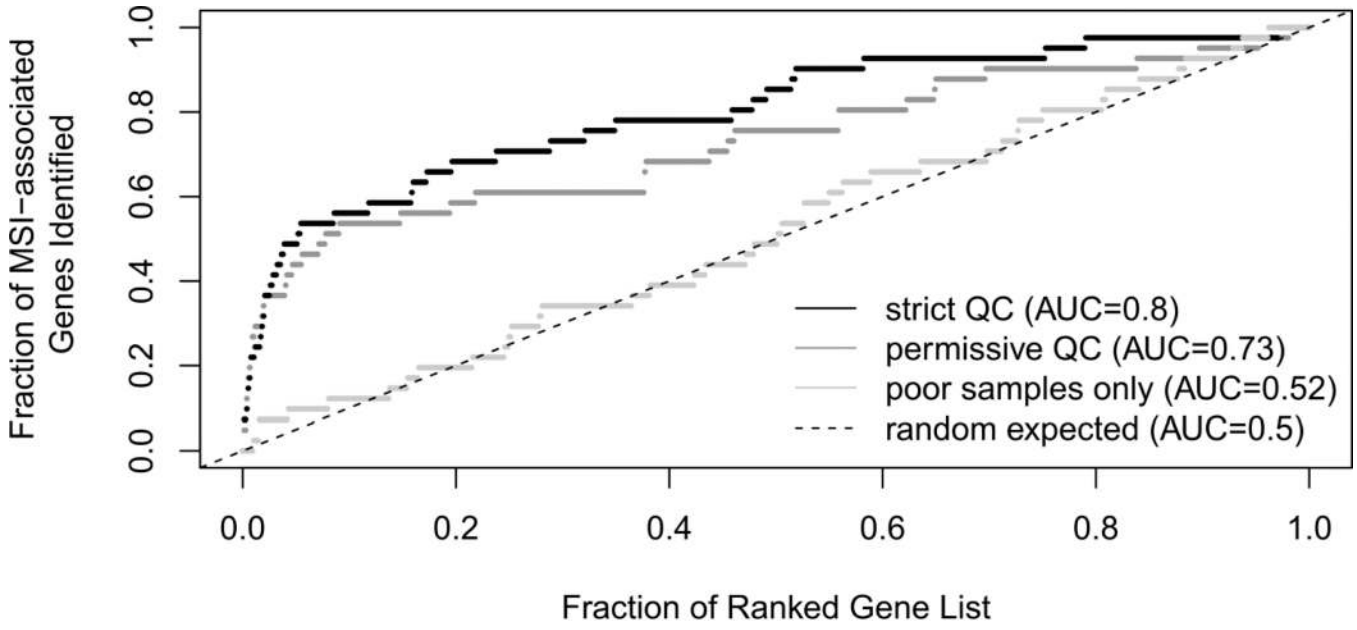


Figure 4. Reproducibility of individual probe intensity measurements can be assessed by dynamic range across samples

Nine pairs of replicate samples passing quality control in the BC/A study were quantile normalized independently, and one set of replicates was used to bin probes by standard deviation across samples. Correlation of these values was then assessed in the second set of replicates. Probes with higher standard deviation of expression values showed correspondingly higher Spearman correlation between technical replicates, indicating that removal of low variance probes prior to biological analyses can improve measurement reproducibility.



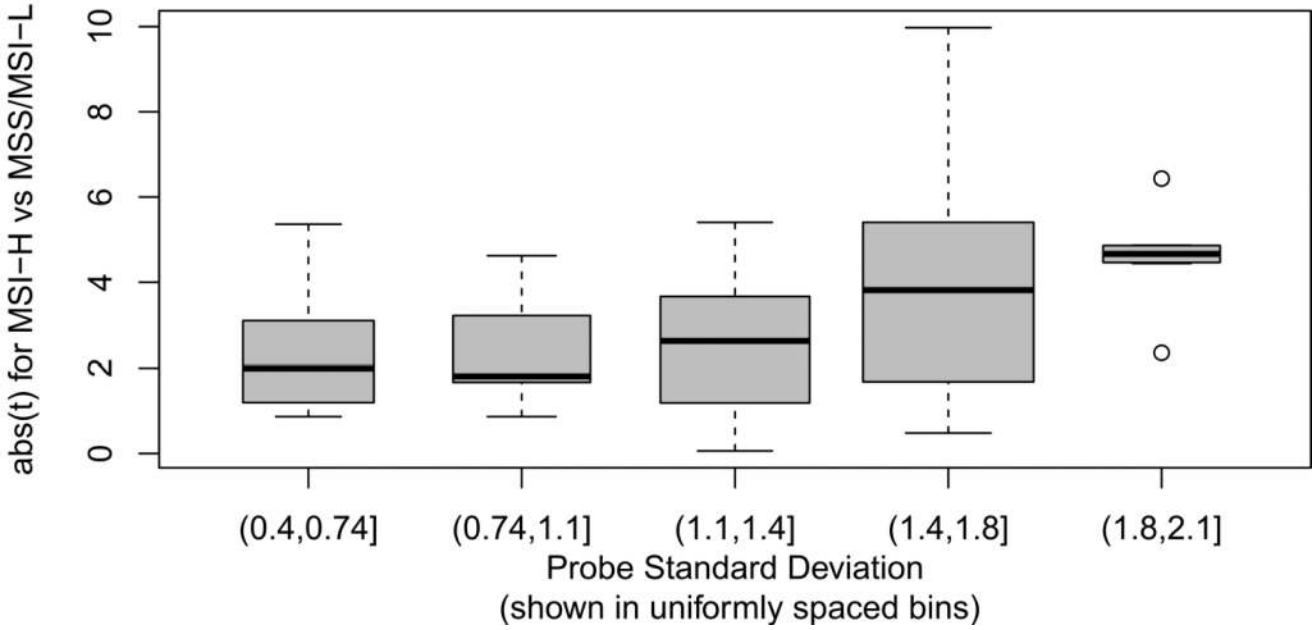


Figure 5. Sample- and probe-level quality control methods improve the accuracy and reproducibility of genes differentially expressed with respect to the microsatellite instability phenotype

(A) Identification of previously established microsatellite instability (MSI) associated genes in whole-genome differential expression analysis improves with strict IQR-based QC. Top differentially expressed genes in the NHS/HPFS samples are well-identified independently of QC, but the detection of more moderately differentially expressed genes is improved by strict QC. **(B)** Identification (by t-statistic) of 43 published MSI-High and MSI-Low associated genes(32) improves as probe standard deviation is used for quality control in NHS/HPFS samples. High-variance probes and higher dynamic range samples thus not only

show better technical reproducibility, but are also more likely to provide differential expression concordant with independent, fresh-frozen tissues.

Table 1

Original (Nurses' Health Study and Health Professionals Follow-Up Study - NHS/HPFS and breast cancer / autopsy - BC/A) and public microarray datasets considered in this study. GSE accession IDs refer to the Gene Expression Omnibus, E-TABM IDs refer to the ArrayExpress database. FF=fresh-frozen, FFPE=formalin-fixed, paraffin-embedded tissues.

Study	Accession ID	Storage method	Sample type	Platform	Sample size
NHS/HPFS	GSE32651	FFPE	CRC	Illumina DASL HumanRef-8 v3	1,003
BC/A	GSE32490	FFPE	Breast LN mets autopsy	Illumina DASL HumanRef-8 v3 v4 HT12	168
Sadi et al.(19)	GSE23386	FF/FFPE	breast	Illumina HumanRef-8 v3 Illumina DASL HumanRef-8 v3	25/25
Jiang et al.(20)	GSE10950	FF	CRC	Illumina HumanRef-8 v2	48
April et al. (6)	GSE17558 / GSE17572	FF/FFPE	lung ovarian	Illumina DASL HumanRef-8 v3	8/8 16
Mittempergher et al.(12)	E-TABM-1081	FF/FFPE	breast	Illumina DASL HumanRef-8 v3	47/46
Budczies et al.(24)	GSE11001	FFPE	breast	Affymetrix HG-U133-plus2	30
Collisson et al.(26)	GSE17891	FFPE	pancreas	Affymetrix HG-U133-plus2	47
Hall et al.(25)	GSE27388	FFPE	cervix	Affymetrix Human Exon 1.0 ST	28
Abdueva et al.(23)	GSE19249	FFPE	multiple	Affymetrix HG-U133A / Affymetrix HG-U133Plus2	15/8
Williams et al.(1)	GSE19246	FFPE	lymphoma	Affymetrix HG-U133-plus2	38
Waddell et al.(10)	GSE21921	FFPE	breast	Illumina DASL HumanRef-8 v3	85
Villanueva et al.(21)	GSE19977	FFPE	liver	Illumina DASL HumanRef-8 v3	164
Minguez et al.(22)	GSE20017	FFPE	hepatocellular carcinoma	Illumina DASL HumanRef-8 v3	135