Introducing a Novel and Robust Technique for Determining Lymph Node Status in Colorectal Cancer

John Hogan, MB, Conor Judge, MB,* Michael O'Callaghan, MB,* Amir Aziz, MB,* Cormac O'Connor, MB,* John Burke, PhD,* Colum Dunne, PhD,† Stewart Walsh, Msc, MB,*† Matthew Kalady, PhD,‡ and J. Calvin Coffey, PhD*†*

Objective: This study aims to harness the potential of public gene expression repositories, to develop gene expression profiles that could accurately determine nodal status in colorectal cancer.

Background: Currently, techniques that determine lymph node positivity (before resection) have poor sensitivity and specificity. The ability to determine lymph node status, based on preoperative biopsies, would greatly assist in planning treatment in colorectal cancer. This is particularly relevant in polyp-detected cancers.

Methods: Public gene expression repositories were screened for experiments comparing metastatic and nonmetastatic colorectal cancer. A customized graphic user interface was developed to extract genes dysregulated across most identified studies (ie, consensus profiles). The utility of consensus profiles was tested by determining whether classifiers could be derived that determined nodal positivity or negativity. Consensus profiles-derived classifiers were tested on separate Affymetrix- and Illumina-based experiments, and collated outputs were compiled in summary receiver operator curve characteristic format, with area under the curve (AUC) reflecting accuracy. The association between classification and oncologic outcome was determined using an additional, independent data set. Final validation was conducted using the Ingenuity network-linkage environment.

Results: Four consensus profiles were generated from which classifiers were derived that accurately determined node positive and negative status (pooled AUC were 0.79 ± 0.04 and 0.8 ± 0.03 for nodal positivity and negativity, respectively). Overall AUC ranged from 0.73 to 0.86, demonstrating high accuracy across consensus profile type, classification technique, and array platform used. As consensus profile enabled classification of nodal status, survival outcomes could be compared for those predicted node negative or positive. Patterns of disease-free and overall survival were identical to those observed for standard histopathologic nodal status. Genes contained within consensus profiles were strongly linked to the metastatic process and included (among others) FYN, WNT5A, COL8A1, BMP, and SMAD family members.

Conclusions: Microarray expression data available in public gene expression repositories can be harnessed to generate consensus profiles. The latter are a source of classifiers that have prognostic and predictive properties.

Keywords: colorectal cancer, consensus profiles, lymph node metastasis, public gene expression repositories, survival

(*Ann Surg* 2013;00:1–9)

From the *Department of General Surgery, Graduate Entry Medical School, University Hospital Limerick, Limerick, Ireland; †4i Centre for Interventions in Infection, Inflammation and Immunity, Graduate Entry Medical School, University of Limerick, Limerick, Ireland; and ‡Department of Colorectal Surgery, Digestive Diseases Institute, The Cleveland Clinic, Ohio, North America.

Dr. J. Hogan, Dr. C. Judge, and Dr. M. O'Callaghan are joint first authors.

Disclosure: The authors declare no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.annalsofsurgery.com).

Reprints: J. Calvin Coffey, PhD, Department of General Surgery, Graduate Entry Medical School, University Hospital Limerick, Limerick, Ireland. E-mail: calvin.coffey@ul.ie.

Copyright © 2013 by Lippincott Williams & Wilkins

ISSN: 0003-4932/13/00000-0001

DOI: 10.1097/SLA.0000000000000289

Colorectal cancer is a significant cause of mortality, with more than 40,000 new cases diagnosed annually in the United Kingdom contributing to more than 16,000 deaths.[1] The presence of lymph node metastasis is an important prognostic feature after curative resection in colorectal cancer and is strongly associated with disease recurrence.[2–4] Advances in neoadjuvant strategies have created a pressing need for accurate identification of high-risk (node positive) tumors before surgical resection.[5,6] With the introduction of screening programs, the dilemma increasingly posed by polyp-detected cancers means that mechanisms must be sought that determine lymph node status, and thereby guide the need for surgical resection.

Preoperative clinical staging (ie, the combination of endoscopic and radiologic findings) is used to determine tumor, nodal, and metastasis status in both colon and rectal cancer.[7,8] Unfortunately, the overall accuracy of imaging modalities in determining nodal stage is extremely variable and oftentimes limited. Significant discrepancy exists between the reported sensitivities and specificities of computed tomography (CT) in identifying lymph node positivity.[9–11] In particular, poor specificity and a high false-positive rate preclude CT evaluation in preoperative nodal classification in colon cancer. Magnetic resonance imaging and endoanal ultrasonography, though useful in assessing local tumor burden, are limited in discriminating between malignant and benign lymph nodes.[12,13] These limitations generate a compelling argument for development of an accurate method of preoperative nodal staging. Ideally, the discriminatory accuracy of 1 such test should surpass that of radiologic staging and closely correlate with surgical staging (ie, the gold standard in determining lymph node status).

Over the past decade, gene expression profiles emerged with predictive properties capable of identifying biologic states. For example, the 5 luminal subtypes of breast cancer can be characterized by their gene expression profiles.[14] However, a number of issues have precluded their widespread application in the clinical setting. Chief among these is disparity of findings across almost all studies, which prompts questions related to reproducibility, reliability, and correlation of data.[15] The lack of concordance between gene lists arises because of (*a*) studies including too few samples, (*b*) a lack of stringency in bioinformatic workflows, (*c*) variations relating to technique and microarray platform processing, and (*d*) lack of concordance between probe sets.[16–18]

A number of developments recently converged that could collectively address the earlier-mentioned clinicobioinformatic issues. First, the Microarray Quality Control consortium demonstrated interplatform correlation with significant concordance between gene lists derived from separate laboratories, using (*a*) different microarray platforms and (*b*) defined optimal algorithms for identifying genes for inclusion in classifier data sets.[19,20] Second, microarray data deposited in public gene expression repositories (PGER) such as Gene Expression Omnibus (GEO), CIBEX, Array Express, and the Stanford database, mean that investigators are no longer hampered by the limitations of small numbers of gene expression profiles.[21–24] In PGER, data are imported according to strict reporting criteria as

Copyright © 2013 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

set out in Minimum Information About a Microarray Experiment.[25] Microarray data (corresponding to gene expression) are clinically annotated, enabling the investigator correlate with various tumor- (eg, stage) and patient- (eg, sex, age, etc) related characteristics. Surprisingly, few studies to date have attempted to harness the potential inherent in these repositories, with a view to developing expression profiles that could accurately determine biologic states. More importantly, no study has applied the bioinformatics rigor proposed by Microarray Quality Control with the data available in PGER in the determination of lymph node status in colorectal malignancy.

The first objective of this study was to generate gene lists that were common to most experiments that (*a*) compare early- with late-stage colorectal cancer and (*b*) are available in PGER. A second aim was to test the utility of these lists in the prediction of histopathologic lymph node status. To facilitate this, software was developed to generate "consensus profiles," that is, gene expression profiles comprising genes commonly altered across most experiments in PGER. Consensus profiles provided a resource of classifiers highly accurate in determining nodal status (ie, predictive) and oncologic outcome (ie, prognostic).

METHODS

Overview of Workflow

The following is an overview of the work-flow process (Fig. 1).

Phase 1—generation of consensus profiles: Public gene expression repositories were searched for experiments pertaining to metastatic and nonmetastatic colorectal cancer. Raw data were downloaded, and genes significantly dysregulated between the 2 disease states were retained and imported into "Consensus Profile Developer" (CPD). Consensus Profile Developer generated a consensus list of dysregulated genes common to most experiments. The resultant 4 lists were referred to as consensus profiles.

Phase 2—testing the utility of consensus profiles in the determination of nodal status: A PERL script was used to extract expression data for genes in each consensus profile from separate and independent Affymetrix- and Illumina-based experiments. In both, staging status was conducted on postoperative resection specimens (ie, surgical staging). The extracted expression data were subjected to a rigorous and standardized bioinformatic workflow

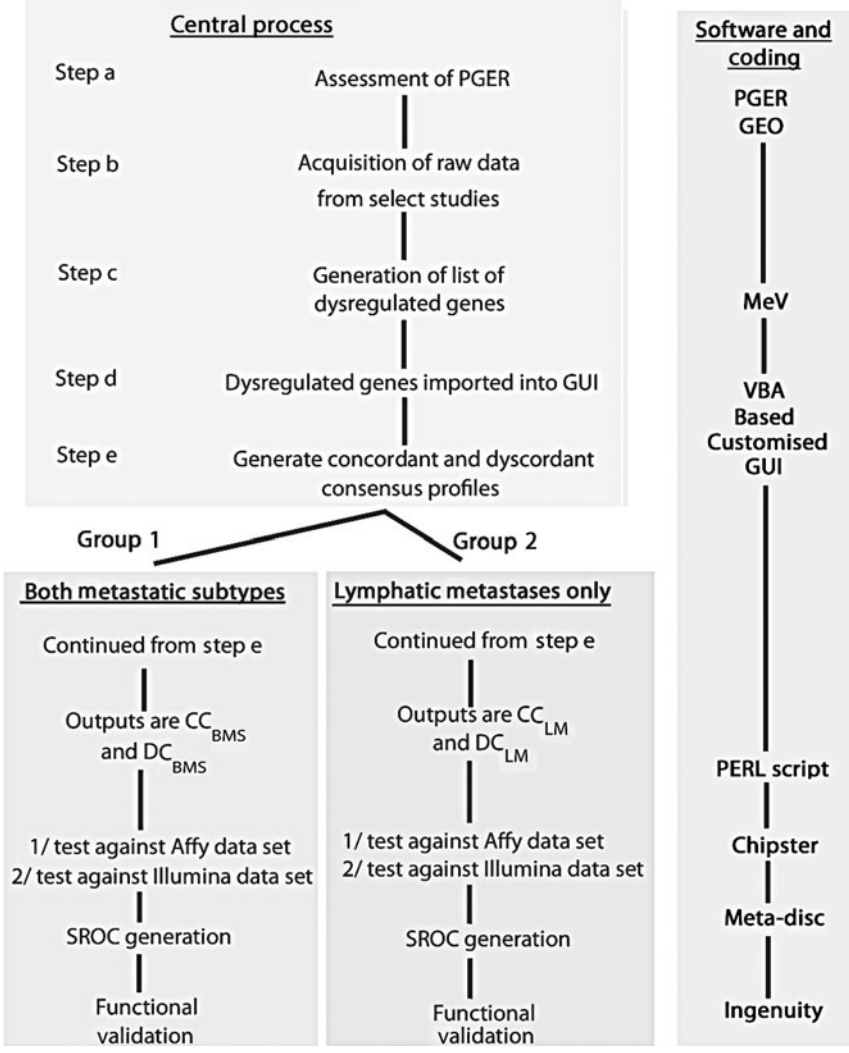

FIGURE 1. Flowchart summarizing key stages in experimental workflow. CC_{BMS} and DC_{BMS} are concordant and discordant consensus profiles generated when both lymphatic and metastatic expression data sets are included. CC_{LM} and DC_{LM} are concordant and discordant consensus profiles generated when only lymphatic metastatic data sets are included. *Step a*, All PGER relating to colorectal cancer were examined and filtered to retain only those 7 that were fully Minimum Information About a Microarray Experiment compliant. All were located within GEO. *Steps b and c*, Raw expression data were used in MultiExperiment Viewer, to generate 7 lists of dysregulated genes, which were then imported into a customized graphic user interface. *Steps d and e*, The latter outputted all 4 consensus profiles. A PERL script was used to extract expression data for genes in each consensus profile from independent Affymetrix- and Illumina-based experiments. These were imported to Chipster and processed to retain only those genes that were significantly differentially expressed between node negative (stage 2) and node positive (stage 3) colorectal cancer. The discriminant properties of the 4 classifier data sets were assessed using an array of classification techniques. Sensitivities/specificity profiles were established using SROC and the area under the curve, in meta-disc. Data sets were then further functionally analyzed in Ingenuity. Affy indicates Affymetrix; GUI, graphic user interface; OS, overall survival; VBA, Visual Basic Algorithm.

© 2013 Lippincott Williams & Wilkins

Copyright © 2013 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

generating and testing classifiers of nodal status. The oncologic outcome of patients classified as either node positive or negative was determined and compared with histopathologic correlates.

Appraisal and Filtration of Public Gene Expression Repository Data

In *step a* (Fig. 1), all PGER experiments comparing conditions associated with early primary colorectal cancer (ie, colorectal cancer without metastases) with late stage (ie, colorectal cancer with lymphatic and/or hepatic metastases) were identified. The specific origin of genomic data is provided in Supplementary Digital Content Table 2 at http://links.lww.com/SLA/A484. In 6 experiments, transcriptomic data derived from primary tumors provided the standard against which comparisons were made (in 1 experiment, transcriptomic data were derived from a primary tumor cell line). This approach was adopted, as the identification of expression differences between extreme stages of cancer could be used to predict intermediate stages.[25] Online archives containing relevant experiments included GEO, Array Express (European), CIBEX, and the Stanford database. In each, a search was conducted using "Colorectal [OR] Colon [+] Cancer [OR] Carcinoma" search terms. Experiments comparing data derived from early-stage conditions (ie, nonmetastatic) with samples derived from late-stage conditions (ie, metastatic) were retained (Supplementary Digital Content Table 1A at http://links.lww.com/SLA/A484). According to the American Joint Committee on Cancer, stages 1 and 2 (ie, early) of colorectal cancer are nonmetastatic, whereas stages 3 and 4 (ie, late) are metastatic. Although, theoretically, early- and late-stage properties may overlap during tumor progression, current classification/staging systems do not cater for this. Minimum Information About a Microarray Experiment compliance was used an index of completeness of the reporting process.[26] A 6-point scoring system was developed, incorporating the main Minimum Information About a Microarray Experiment features (Supplementary Digital Content Table 1B at http://links.lww.com/SLA/A484). Each of 12 suitable experiments was scored, and only those satisfying all criteria (ie, score of 6) were retained for further analysis (Supplementary Digital Content Table 1C at http://links.lww.com/SLA/A484). At completion of step a (Fig. 1), 7 experiments were retained, comprising 332 microarray samples generating 17,506 instances of differential gene expression (Supplementary Digital Content Table 2 at http://links.lww.com/SLA/A484).

Importation of Data and Consensus Profile Generation

As part of *steps b and c*, MultiExperiment Viewer (MeV) was used to generate a list of differentially expressed genes for each study retained in step a. Raw data were downloaded, imported to MeV, and normalized (robust multilevel and quartile normalization were used for Affymetrix and single-channel microarray platforms, respectively).[27] Experimental conditions were compared using an unpaired *t* test without adjustment for multiple sampling and using a $P < 0.05$ cutoff (ie, with a view to maximizing the capture of genes and thus possible overlaps, see hereafter).

To generate consensus profiles, a novel graphic user interface CPD was generated using Visual Basic (Windows Office 7) (see supplementary materials for instructions, usage, and code). This enabled a cross-comparison of all 7 data sets in Excel and outputted 2 types of consensus profile (see hereafter). Within each profile, genes were rank ordered according to frequency and direction of differential expression. As CPD entries included experiment number, gene symbol, and up- or downregulated status only, CPD is not condition-dependant and can thus be applied broadly.

"Consensus" was deemed to be present when dysregulation (defined as the presence of a significant difference in gene expression on MeV analysis) was evident across most experiments. Genes similar in expression (ie, not dysregulated) across the majority of experiments could also provide a type of consensus profile; however, this was not analyzed in the present study. Thereafter, 2 subtypes of consensus were possible; a consensus could be in the same direction when a gene was either significantly up- *or* downregulated across most experiments. This was arbitrarily termed a *concordant consensus* and abbreviated to "CC." A *discordant consensus* ("DC") arose when a gene was significantly dyrsregulated across most experiments, but where expression changes occurred in either direction (ie, up- *and* downregulated included). Consensus Profile Developer outputted both consensus types.

Two groups of concordant (ie, CC) and discordant (ie, DC) consensus profiles were then generated using CPD (Fig. 1). Group-1 consensus profiles were derived from all 7 experiments (ie, experiments involving hepatic or lymphatic metastatic disease as comparators). Group-2 consensus profiles were derived solely from experiments that compared early primary colorectal cancer with late-stage cancer involving lymphatic (*but not hepatic*) metastatic disease. The following notations were generated to clarify the grouping and types of consensus profile:

CC_{LM} and DC_{LM}: concordant (CC) and discordant (DC) profiles where comparators were associated with lymphatic metastatic disease (LM) only

CC_{BMS} and DC_{BMS}: concordant (CC) and discordant (DC) profiles where comparators were associated with both metastatic disease types (BMS)

Consensus Profile Assessment Through Prediction of Nodal Status

The discriminatory properties of all profiles were assessed by determining the capacity to differentiate node negative and node positive colorectal cancer in 2 independent experiments. Nodal status is the single strongest determinant of hepatic metastasis development, and thus is intimately related to overall survival in colorectal cancer.

GSE31595 is a file in GEO containing expression data from 37 Affymetrix Human Genome U133 Plus 2.0 microarrays. The Affymetrix Human Genome U133 Plus 2.0 is a commonly used microarray platform. Of the 37 samples, 20 and 17 were derived from stages 2 (node negative) and 3 (node positive) tumors, respectively, as per surgical staging on resected specimens. A PERL script was generated to extract expression data for each consensus profile, from GSE31595 (see supplementary methods for code). These were imported to Chiptser, renormalized, and 2 group testing was conducted using EmpiricalBayes (with a Benjamini-Hochberg *P*-value adjustment for multiple sampling). In this manner, the consensus profiles were filtered to include only those genes differentially expressed between node negative and node positive disease. This generated 4 *classifier data sets* (ie, 1 for each consensus profile type), using the R packages limma and LPE (Supplementary Digital Content Table 3 at http://links.lww.com/SLA/A484).[28] Clustering via sample and gene expression was evaluated for each data set, using R packages ape and amap.[28]

As classification results vary according to the data set and technique used, classification was conducted separately using the K-nearest neighbor (knn), linear discriminant analysis (lda), discriminant analysis (slda), and using svm, rpart, lvq, naiveBayes, and bagging techniques. This generated 9 confusion matrices for each consensus profile classifier (ie, 36 confusion matrices in total) (data not included). For each classifier, the optimal 3 classification outputs were combined, then analyzed using the Meta-analytical

© 2013 Lippincott Williams & Wilkins

Copyright © 2013 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Integration of Diagnostic Accuracy Studies command, after which the area-under-the-curve (AUC) values were calculated using Meta-Disc version 1.4 (Universidad Complutense, Madrid, Spain).

As GSE31595 was generated using an Affymetrix hgu133 plus2 Array platform, the earlier-mentioned process was repeated on an independent data set generated using the Illumina 6v2 platform. The latter comprised 158 colorectal tumor samples (Illumina 6v2 platform). Staging of these had been conducted on postoperative resection specimens. Methodology detailing RNA extraction, quality control, and hybridization is available in Supplementary Digital Content Methodology Section 2 at http://links.lww.com/SLA/A484. In this, arbitrarily termed the Illumina experiment, importation and comparison workflow processes were identical to those described earlier and the end product consisted of 4 *classifier data sets* (1 for each consensus profile classifier) (Supplementary Digital Content Table 4 at http://links.lww.com/SLA/A484). The earlier-mentioned classification process was repeated, generating 36 confusion matrices (ie, 9 confusion matrices for each consensus profile type) (data not included). The optimal 3 outputs were again pooled to generate summary receiver operating characteristic curves as described earlier.

Correlation Between Classification With Consensus Profiles and Oncologic Outcomes

GSE17536 is a GEO-based data set comprising 177 samples annotated with disease-free and overall survival data. As this data set had not been used in either consensus profile generation or validation, it was fully independant. GSE17536 was downloaded, and samples contained were classified as either lymph node positive or negative, using the consensus profile-based approach described earlier. After samples were classified as lymph node positive or negative, Kaplan-Meier estimates were plotted and compared between groups (for both disease-free and overall survival), using a log-rank analysis. Kaplan-Meier estimates were also generated for patients classified as either node negative or positive on routine histopathologic (ie, surgical) assessment. Both sets of KM estimates (ie, from transcriptomic and histopathologic classification approaches) were then compared (Supplementary Digital Content Figure 1 at http://links.lww.com/SLA/A484).

Functional Annotation

Each consensus data set containing HUGO Gene Nomenclature Committee gene identifiers and frequencies was imported into Ingenuity (ILP) for functional annotation.[29] Ingenuity permits a cross-referencing of genes against published literature to identify linkages. A detailed description of the associated methodology is included in the supplementary materials. Within the Ingenuity environment, the "functional analysis" tool was used to determine the relationship between consensus profile genes and the gastrointestinal metastatic process. For each consensus profile, the topmost 5 disease functions were filtered sequentially to retain only genes associated with gastrointestinal cancer metastasis. All genes contained within the consensus profiles, and not associated with gastrointestinal metastases, were excluded. The final 4 outputs were combined to generate a "customized pathway," including only those genes most strongly associated with gastrointestinal metastases.

RESULTS

Consensus Profile Generation

Four consensus profiles were generated using CPD, as detailed in Methods. The content of each is described in the following.

Group 1: Concordant and discordant consensus profiles derived from expression data where data were derived from experiments based on both metastatic subtypes (CC_{BMS} and DC_{BMS}).

CC_{BMS} comprised 58 genes concordantly dysregulated across most experiments (23 and 35 up- and downregulated, respectively). Supplementary Digital Content Table 5 at http://links.lww.com/SLA/A484 lists the functions, direction, and frequency of dysregulated genes. No gene was upregulated in greater than 4 experiments. Although no gene was downregulated in more than 5 experiments, prostaglandin D2 synthase and bone morphogenetic protein type 2 were significantly downregulated in 5 experiments. DC_{BMS} comprised 155 genes discordantly dysregulated across most of the 7 experiments. Supplementary Digital Content Table 6 at http://links.lww.com/SLA/A484 demonstrates the frequencies and distribution of up- and downregulated genes in DC_{BMS}. Adipocyte differentiation-related protein was significantly up- and downregulated across 5 and 1 experiment, respectively. As adipocyte differentiation-related protein was significantly dysregulated in both directions it was included in DC_{BMS} and not in CC_{BMS}.

Group 2: Concordant and discordant consensus profiles derived from expression data where data were derived from experiments based on lymphatic metastases (CCLM and DCLM).

CC_{LM} comprised 84 genes (44 and 40 up- and downregulated, respectively). Supplementary Digital Content Table 7 at http://links.lww.com/SLA/A484 lists the functions, direction, and frequency of dysregulated genes in CC_{LM}. WSB2, a member of the WD subfamily of proteins, was upregulated across all 4 experiments. Tetratricopeptide repeat domain 1 was downregulated across all 4 experiments included. One hundred ninety-seven genes were discordantly dysregulated across most experiments (ie, DC_{LM}). Supplementary Digital Content Table 8 at http://links.lww.com/SLA/A484 demonstrates the distribution of up- and downregulated genes and frequencies in DC_{LM}. Although adipocyte differentiation-related protein was the only gene upregulated across 3 experiments, polymerase III polypeptide G was the only gene downregulated across 3 experiments.

Consensus Profile Testing

To test the utility of consensus profiles in discriminating tumor properties, expression data for each gene in all 4 consensus profiles were extracted from separate Affymetrix- and Illumina-based experiments. Consensus profiles together with new expression data sets were imported to Chipster and classifiers derived that predicted nodal status in colorectal cancer. Classifiers were generated on the basis of 2 group testing with *P*-value adjustment for multiple sampling (Supplementary Digital Content Tables 3 and 4 at http://links.lww.com/SLA/A484). This retained only those genes that were significantly differentially expressed, between node negative and node positive colorectal tumors. As there were 4 consensus profiles to begin with, this generated 8 classifiers (ie, 4 each for Affymetrix- and Illumina-based experiments) (Supplementary Digital Content Tables 3 and 4 at http://links.lww.com/SLA/A484). The discriminant properties of each of the 8 classifiers were then assessed in 2 manners. First, classification results were pooled to determine false-negative, false-positive, true-negative, and true-positive ranges. Second, pooled classification results permitted summary receiver operator curve development and AUC comparisons.

Testing of Data Sets Against Independent Experiments: Confusion Matrices—Pooled Analysis

Classification outputs varied according to test used. To illustrate this, an array of confusion matrices is included in Supplementary

© 2013 Lippincott Williams & Wilkins

Copyright © 2013 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Digital Content Table 9 at http://links.lww.com/SLA/A484. Accuracy in determining lymph node negativity was first assessed. When all classification outcomes were pooled, false-negative rates ranged from 13% to 46% (24% ± 11%), with the lowest associated with the classifier from CC_{LM} (13%). The mean false-positive rate was 22% ± 10%, and the lowest false-positive rate was associated with the CC_{BMS}-derived classifier at 3% (in the Illumina-based experiment). The mean true-positive rate was 79% ± 9.3%, and the highest occurred with the classifier from CC_{BMS} at 97% (also observed in the Illumina-based experiment). Similarly, a pooled analysis of classification outputs demonstrated that true-negative percentages ranged from 65% to 97%, with the highest again associated with the CC_{BMS}-derived classifier at 97% (in the Illumina-based experiment). Accuracy in determining lymph node positivity was next assessed. False-negative rates ranged from 3% to 35% (22% ± 10%), with the lowest associated with the CC_{BMS} classifier. The mean false-positive rate was 31% ± 5.6%, and the lowest false-positive rate was associated with the classifier from CC_{LM} (ie, 25%). For true-positive rates, the mean was 69% ± 5.6%, and the highest rate was also associated with the CC_{LM} classifier (ie, 75%). True-negative percentages ranged from 65% to 97% (mean 79% ± 9.3%).

Testing of Data Sets Against Independent Experiments: Area Under the Summary Receiver Operating Characteristic Curve—Pooled Analysis

The sensitivity/specificity of each classifier data set is further summarized in Figure 2, demonstrating the AUC for each test. Values of AUC represent the sensitivity/specificity of the pooled classification results for each of 8 classifiers. Hence, each AUC value refers to a particular test and classifier. Two tests were used, (*a*) the detection of lymph node positive status and (*b*) the detection of lymph node negative status, in determining the clinical utility associated with each classifier (and thus, each consensus profile). This workflow generated an array of AUC results that was then compared between the following categories: (*a*) test type, (*b*) Affymetrix versus Illumina, (*c*) concordant versus discordant consensus profile, and (*d*) group 1 (where data were derived from experiments involving both metastatic subtypes) versus group 2 (where data were derived from experiments involving from lymphatic metastases only). An exhaustive within-category comparison of each parameter is beyond the scope of the article.

The mean pooled AUC for all results was 0.802 ± 0.04. Mean pooled AUC for all tests related to lymph node positivity alone was 0.79 ± 0.04. Mean pooled AUC for all tests relating to lymph node negativity was similar at 0.80 ± 0.03 ($P = 0.96$). Thus, classifiers were equally sensitive and specific in determining rates of lymph node positivity and negativity. Pooled AUC values for Affymetrix- and Illumina-based tests were 0.75 ± 0.01 and 0.83 ± 0.01, respectively ($P < 0.001$), indicating that classifiers had greater discriminatory capability in testing with Illumina-derived data. All AUC results were pooled for group 1 and 2 related tests. To recap, group 1 classifiers were generated from consensus profiles based on both metastatic subtypes, whereas group 2 classifiers were generated from experiments referring to lymphatic metastases only. Mean AUC did not differ between the 2 (0.79 ± 0.03 vs 0.80 ± 0.04, respectively; $P = 0.76$). Similarly, there were no AUC differences when classifiers from concordant profiles were compared with those from discordant profiles (0.79 ± 0.05 and 0.80 ± 0.02; $P = 0.5$).

Examining the Affymetrix test context alone, AUC was significantly greater for classifiers generated from discordant versus concordant consensus profiles (0.79 ± 0.02 vs 0.75 ± 0.02; $P = 0.027$). In the Illumina setting, the reverse occurred, with classifiers from concordant consensus profiles associated with a greater (albeit nonsignificantly) AUC at 0.83 ± 0.01 versus 0.81 ± 0.01 ($P = 0.095$). In the Affymetrix setting, AUC from group 1 and 2 classifiers were compared, with no significant difference emerging (0.77 ± 0.01 and 0.78 ± 0.05, respectively; $P = 0.46$). Similar observations were apparent for the Illumina context (0.84 ± 0.02 vs 0.81 ± 0.1; $P = 0.5$). There were no significant differences in AUC returned when determining rates of lymph node positivity was compared with determining lymph node negativity; for Affymetrix: 0.77 ± 0.03 versus 0.76 ± 0.03, respectively ($P = 0.69$); and for Illumina: 0.82 ± 0.01 versus 0.82 ± 0.02, respectively ($P = 0.752$).

Finally, individual AUC results were compared for classifiers within each experimental context. Within the Affymetrix setting, the DC_{BMS} classifier was most accurate at identifying rates of node negativity and positivity. In the Illumina context, the CC_{LM} classifier had the highest AUC in relation to both node negative and node positive tumors (0.84 and 0.86, respectively). CC_{LM}-classifier AUC values were third highest in the Affymetrix/GSE31959 study, whereas the DC_{BMS}-classifier AUC values were lowest in the Illumina study.

Correlation Between Classification With Consensus Profiles and Oncologic Outcomes

Samples from GSE17536 were classified as node positive or negative, using the consensus profile-based approach described earlier. When Kaplan-Meier estimates were compared between groups, both disease-free and overall survival were significantly and markedly reduced in patients whose samples were classified as node positive (Supplementary Digital Content Figure 1 at http://links.lww.com/SLA/A484). For the purposes of comparison, disease-free and overall survival were also plotted for patients classified as lymph node negative or positive based on standard histopathologic classification. As can be seen from Supplementary Digital Content Figure 1 at http://links.lww.com/SLA/A484, the overall survival patterns for transcriptomic and histopathologic nodal status designation were remarkably similar.

As a final test, the ability of consensus profiles to determine risk of recurrence was assessed in the GSE31595 experimental context. Primary tumor samples in GSE31595 are divided into recurrent and nonrecurrent, on the basis of whether they were followed by recurrence (the type of recurrence and time to recurrence are not documented in GSE31595). Again, classifiers could be developed that were predictive of recurrent and nonrecurrent status. Although the overall AUC was 0.77, specificity and false-positive rates were 85.7% and 14.3%, respectively.

Functional Analysis of Data Sets

Within the Ingenuity environment, all consensus profile components were cross-referenced against published literature to identify those genes most strongly associated with gastrointestinal metastasis to date. The results, termed "outputs," were compared in terms of networks, molecular and physiologic functions, disease, canonical pathways, and transcription factors associated. Notch was associated with both CC_{BMS} and CC_{LM}, whereas HOXa9 was associated with both DC_{BMS} and DC_{LM}. The top 2 diseases and molecular functions associated with each are summarized in Supplementary Digital Content Table 10 at http://links.lww.com/SLA/A484. In all except the CC_{BMS} profile, "cancer" was the strongest associated disease state, with genetic abnormalities second. The association with cancer was strongest for both CC_{BMS} and DC_{BMS} classifiers. The profile of molecular functions was similar between sets, and again the associations with DC_{BMS} and CC_{BMS} were strongest (Supplementary Digital Content Table 10 at http://links.lww.com/SLA/A484).

When all 4 consensus profiles were analyzed for the 5 topmost cancer-associated functions, "metastasis" was associated with each.

© 2013 Lippincott Williams & Wilkins

Copyright © 2013 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

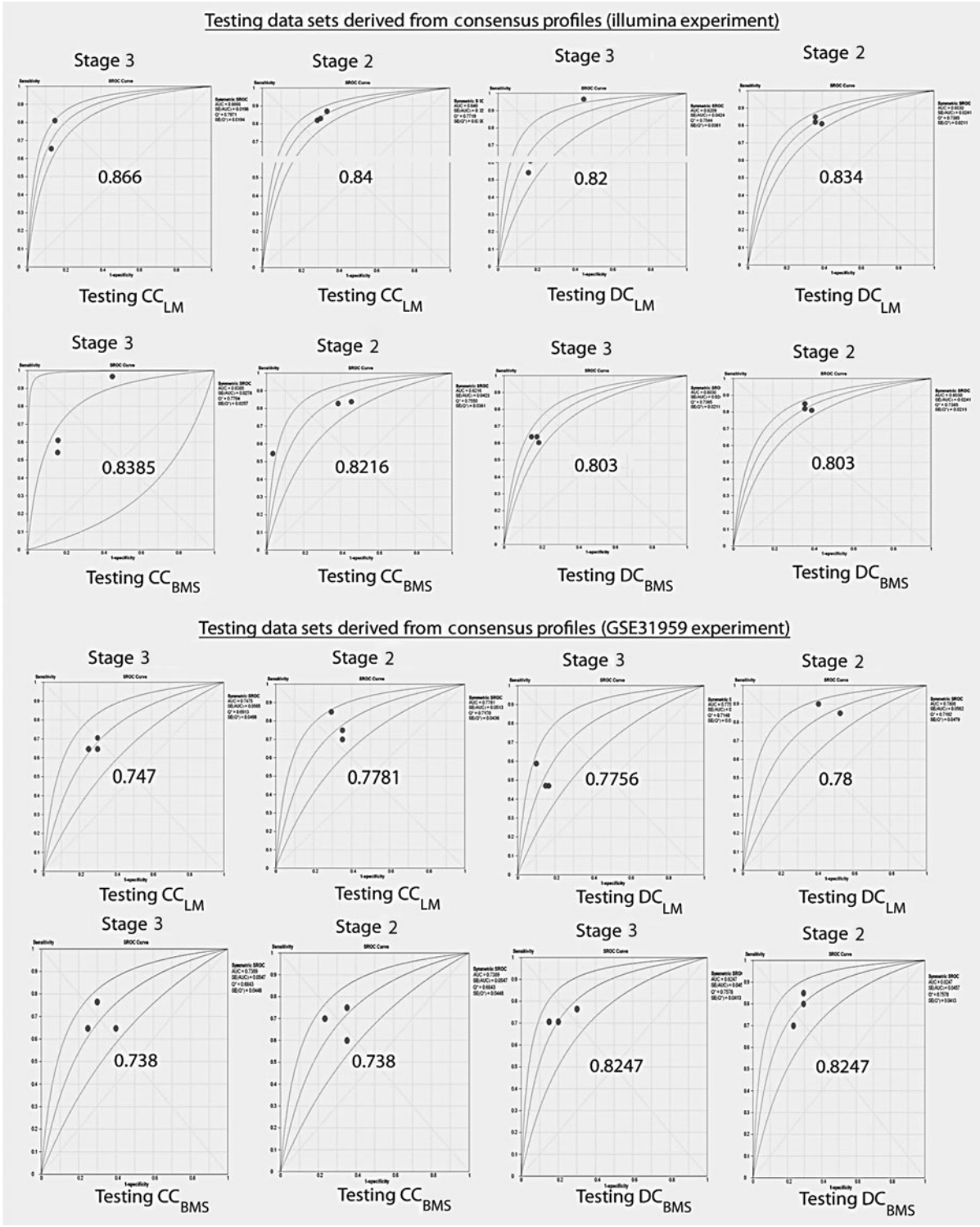

FIGURE 2. Summary receiver operator curves were generated using the DerSimonian and Laird procedure for random effects in Meta-Disc. The top 2 rows refer to the Illumina-based experiment. The first row graphs were generated from classifiers of the consensus profiles examining lymphatic metastases (ie, CC_{LM} and DC_{LM}). The second row graphs were generated from classifiers of consensus profiles examining both metastatic subtypes (ie, CC_{BMS} and DC_{BMS}). The bottom 2 rows follow an identical pattern and relate to the Affymetrix-based experiment. The central figures in each line graph are the respective AUC values. The first and third columns of line graphs relate to sensitivities/specificities associated with stage 3 status, whereas the second and fourth columns refer to testing stage 2 status.

 © 2013 Lippincott Williams & Wilkins

Copyright © 2013 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

This process highlighted specific molecules that were (*a*) present in each consensus profile and (*b*) strongly associated with colorectal cancer metastasis. These were compiled in what is referred to as "a network pathway" for metastatic colorectal cancer (Fig. 3), that is, a compilation of those genes most strongly associated with metastasis in published literature. Finally, an "IPA My Pathway Report" was generated from this network pathway, which highlights established biologic therapies known to target the associated molecular pathways. For example, FYN (a src kinase targeted by Dasatinib) and COL18A1 (a collagen subtype targeted by collagenase clostridium histolyticum) were highlighted using this approach (Fig. 3).

DISCUSSION

The present study screened all known PGER for experiments that (*a*) satisfied particular reporting criteria and (*b*) compared early- (ie, primary nonmetastatic) with late-stage (ie, metastatic) colorectal cancer. From experiments undergoing a stringent filtration process, genes that were common across all were collated from consensus profiles (of which there were 4 types). These were gene lists from which small-set classifiers were derived that determined lymph node status. When tested against independent surgical staging, high levels of sensitivity and specificity persisted. When classification was correlated with oncologic outcomes, patterns identical to those observed for histopathologic-based prognosis were apparent. Hence, consensus profiles are a resource of genetic classifiers that (*a*) accurately predict lymph node status and (*b*) prognosticate oncologic outcome.

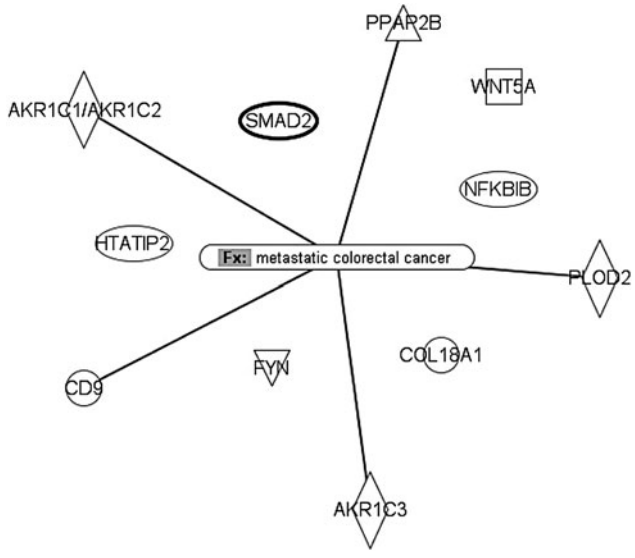

FIGURE 3. Network of molecules that (1) are associated with gastrointestinal metastases, (2) contained within the Ingenuity Knowledge database, and (3) overlap with the 4 consensus sets. Lines indicate molecules with known links to colorectal metastasis in particular. FYN and COL18A1 are targets for which therapeutic modalities are available (ie, dasatinib and clostridium collagenase histolyticum, respectively). AKR1C1-3 indicates aldo-ketose reductase family 1 member C1-3; CD9, CD9 molecule (previously CD9 antigen); COL18A1, collagen type 18 alpha; FYN, FYN oncogene related to SRC, FGR, YES; NFKBIB, nuclear factor of kappa polypeptide gene enhancer in B-cells inhibitor; PLOD2, procollagen-lysine, 2-oxoglutarate 5-dioxygenase; PPAP2B, phosphatidic acid phosphatase type 2B; SMAD2, mothers against DPP homologs; WNT5A, wingless-type MMTV integrations site family member 5A.

Omics-based technologies (including microarray and gene expression) developed considerably over the past decade but have yet to be broadly adopted in the clinical setting. This is due mainly to the disparity that persists between studies aiming to characterize similar biologic states.[15] Disparity arises because of (*a*) differences in array platforms, (*b*) effects of covariates, (*c*) feature selection, (*d*) batch effects, and (*e*) varying classification stringency coupled with failure to collate multiple classification results.[30,31] The present study departed from the normal approach by commencing with all available public gene expression data to identify genes commonly dysregulated (ie, consensus profiles) and use these as a platform from which to generate accurate classifiers. In so doing, batch, covariate, and replicate effects were reduced. The study used a stringent bioinformatic pipeline to generate classifiers of nodal status. *P*-value adjustment, collation of multiple classification techniques, minimized feature selection, and the depiction of classifier accuracy in formats recommended by the Microarray Quality Control consortium were all used to maximize bioinformatic stringency and prevent an overfitting of relations. The accuracy of the classifiers generated was reflected in similar results across independent experiments involving the 2 most commonly used microarray platforms (ie, Affymetrix and Illumina).[32] Accuracy was further reflected in correlations with oncologic outcome parameters (ie, disease-free and overall survival), which changed in a manner identical to that seen with standard histopathologic-based prognostication.

In almost all the GEO-based experiments used in this study, transcriptomic data were obtained from biopsy-derived RNA. As the consensus profiles and classifiers developed herein comprised a limited number of genes, they are suited to reverse transcription polymerase chain reaction–based analyses of tumor biopsies, and hence to preoperative staging. The high discriminatory accuracy associated with these classifiers should thus be contextualized by comparison with accuracy levels for current preoperative staging modalities. Bipat et al[29] performed a meta-analysis comparing accuracy of CT, magnetic resonance imaging, and endoanal ultrasonography in identifying lymph node positivity in rectal cancer. The sensitivities reported were 67%, 55%, and 66%, respectively, and specificities were 78%, 74% and 76%, respectively. Fluorine-18 2-fluoro-2-deoxy-D-glucose (18F-FDG) positron emission tomography is a specific preoperative diagnostic tool (87.9%) in determining lymph node status but is limited by poor sensitivity (42.9%).[11,33] The converse is true when CT is applied in determining lymph node status in colon cancer (exclusive of rectal cancer), with high reported sensitivity (83%) but low specificity (38%).[34] Although not directly compared with preoperative radiologic modalities for nodal staging in the present study, the classifiers generated demonstrated both high sensitivity and specificity (AUC ranging from 0.73 to 0.86). These findings prompt a prospective comparison of nodal staging using transcriptomic and radiologic staging modalities.

Importantly, consensus profiles varied (though not significantly) in accuracy depending on the question asked (ie, lymph node positive or negative) and on the platform used. Thus, in deciding which type consensus profile one should use, the question, platform, and context should first be determined. In the Affymetrix-based experimental platform, classifiers derived from discordant consensus profiles more accurately predicted rates of nodal positivity and negativity compared with classifiers derived from concordant profiles. The reverse occurred for the Illumina platform in which concordant classifiers were most accurate (though not significantly so). These differences may reflect the number of arrays included in both platform types. The optimal classifier in determining lymph node positive tumors was derived from a concordant consensus profile. The latter

© 2013 Lippincott Williams & Wilkins

Copyright © 2013 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

was based on studies examining lymphatic metastases (ie, the CC_{LM} classifier had an AUC of 0.86 in determining lymph node positive tumors). The optimal classifier in determining lymph node negative tumors was derived from a discordant consensus profile. The latter was generated from studies including both metastatic subtypes (ie, the DC_{BMS} classifier had an AUC of 0.82 in determining lymph node negative tumors).

In testing the classification properties of consensus profiles, classification was tested against 2 independent data sets (ie, Affymetrix- and Illumina-based). In each of the latter, nodal status was confirmed on postresection specimens (ie, surgical staging). Hence, this study indirectly tested transcriptomic classification of nodal status with surgical classification. A close correlation was observed, in so far as areas under the curve ranged from 0.73 to 0.86. It is interesting to note that the mean false-positive rate in the present study was 22% ± 10%, which closely correlates with established false-negative rates associated with histologic determination of nodal status in colorectal cancer.[35–37] The false-positive rates observed in the present study could be explained by a combination of (*a*) false-negative rates associated with the histopathologic component of surgical staging, (*b*) genetic heterogeneity, and (*c*) differential epidemiologic effects. The data presented prompt the question as to whether transcriptomic classification could equate with or outperform surgical approaches to determining nodal status. Answering this would require a direct head-to-head comparison of both in terms of their concordance in classification and correlation with definitive oncologic outcomes (eg, disease-free and overall survival).

A penultimate validation of the approach used herein was conducted by comparing the oncologic outcomes of patients classified as node positive or node negative. Patients, samples, and outcome data were derived from an independent GEO-based study that was not used in either consensus profile generation or validation as described earlier. Both the disease-free and overall survival of patients classified as node positive (using the consensus profile–based approach) were significantly poorer than those classified as node negative. In addition, a comparison of prognostication between transcriptomic and histopathologic classification approaches was conducted. Remarkably similar patterns of prognostication were observed between both groups (Supplementary Digital Content Figure 1 at http://links.lww.com/SLA/A484). Although GEO archives do not generally contain information regarding time to recurrence, some experiments (eg, GSE31595) do compare primary tumors associated with recurrence (categorized as "recurrent" tumors) with primary tumors not associated with recurrence (designated as "nonrecurrent" in the GSE31595 experiment). The data here demonstrate that classifiers derived from consensus profiles were accurate in differentiating both types of primary tumor (the AUC, specificity, and false-positive rates were 0.77, 85%, and 14%, respectively). Hence, classifiers derived from consensus profiles also had both prognostic and predictive properties.

The final component of experimentation involved a functional annotation of consensus profiles in which these were cross-checked against the Ingenuity Knowledge database. The latter cross-references against all published literature and thus surpasses other hypergeometric linkage tools (eg, Onto-Express, MAPPFinder, GoMiner, DAVID, EASE, GeneMerge, FuncAssociate) in generating and contextualizing annotations.[38] In this manner, only genes significantly associated with colorectal metastases were retained to form a "network pathway" for colorectal metastasis. The network pathway identified numerous established targets (including the src tyrosine kinase FYN, which is targeted by dasatinib).[39] Other targets were also identified for which therapeutic targets are available, but which have yet to be characterized in the context of colorectal metastases. For example, COL8A1 is a collagen subtype targeted by clostridium collagenase histolyticum.[40,41] Interestingly, clostridium histolyticum is associated with colorectal metastases formation. The generation of a metastases network pathway from consensus profile components thus provides a fertile ground for investigation of novel therapeutic targets in metastasis formation.

The retrospective nature of this study means that inaccuracies could arise at a number of points. Probe sequences continue to evolve with successive iterations of the human genome, and it is feasible that probes thought firstly to map to 1 gene may actually map with an alternative gene.[42] Ideally, future comparisons should be directly made in a prospective manner between transcriptomic and surgical staging and should be cross-referenced against oncologic outcome and against a comprehensive molecular characterization of the tumors examined. Notwithstanding this, the consistency in AUC values observed across comparisons in the current study reflects a methodologic robustness. In general, classifiers from concordant or discordant consensus profiles had a similar predictive accuracy. Classifiers derived from profiles based on lymphatic metastases had similar predictive accuracy to those derived from profiles based on both metastatic subtypes. In addition, the overall mean AUC associated with lymph node negative and positive testing were similar. These similarities, coupled with the prognostic associations, indicate that approaches exploiting consensus profiles are robust.

CONCLUSIONS

Microarray data that were (*a*) available in PGER and (*b*) compared early- and late-stage colorectal cancer were screened. Genes that were differentially expressed across most experiments were used to populate consensus profiles. From the later, small-set classifiers were derived that accurately predicted node positive and negative status. Classifier-based prognostication was near identical with histopathologic-based prognostication. These findings prompt a direct comparison of transcriptomic and surgical staging in predicting lymph node status and in prognosticating oncologic outcome.

ACKNOWLEDGMENTS

The author contributions are as follows: J. Hogan—data acquisition, statistical analysis, and preparation of manuscript; C. Judge—generated PERL script; M. O'Callaghan—generated CPD; C. O'Connor—statistical analysis; A. Aziz—data acquisition; J. Burke—manuscript editing; C. Dunne—Manuscript editing; S. Walsh—manuscript editing; M. Kalady—furnished data for Illumina study; J. C. Coffey—concept generation, bioinformatic analysis, and manuscript preparation.

REFERENCES

1. Canna K, Hilmy M, McMillan DC, et al. The relationship between tumour proliferative activity, the systemic inflammatory response and survival in patients undergoing curative resection for colorectal cancer. *Colorectal Dis.* 2008;10:663–667.
2. Kang J, Hur H, Soh Min B, et al. Prognostic impact of inferior mesenteric artery lymph node metastasis in colorectal cancer. *Ann Surg Oncol.* 2011;18:704–710.
3. Kim TH, Chang HJ, Kim DY, et al. Pathologic nodal classification is the most discriminating prognostic factor for disease-free survival in rectal cancer patients treated with preoperative chemoradiotherapy and curative resection. *Int J Radiat Oncol Biol Phys.* 2010;77:1158–1165.
4. Hashiguchi Y, Hase K, Ueno H, et al. Prognostic significance of the number of lymph nodes examined in colon cancer surgery: clinical application beyond simple measurement. *Ann Surg.* 2010;251:872–881.
5. Sarmiento R, Longo R, Gasparini G. Antiangiogenic therapy of colorectal cancer: state of the art, challenges and new approaches. *Int J Biol Markers.* 2012;27:286–294.
6. Blake H, Dighe S, Koh MD, et al. Accuracy of multidetector computed tomography in identifying poor prognostic factors in colonic cancer. *Br J Surg.* 2010;97:1407–1415.

 © 2013 Lippincott Williams & Wilkins

Copyright © 2013 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

7. Dighe S, Swift I, Magil L, et al. Accuracy of radiological staging in identifying high-risk colon cancer patients suitable for neoadjuvant chemotherapy: a multicentre experience. *Colorectal Dis*. 2012;14:438–444.
8. Junginger T, Hermanek P, Oberholzer K, et al. Rectal carcinoma: is too much neoadjuvant therapy performed? Proposals for a more selective MRI based indication. *Zentralbl Chir.* 2006;131:275–284.
9. Kwak JY, Kim JS, Kim HJ, et al. Diagnositc value of FDG-PET/CT for lymph node metastasis of colorectal cancer. *World J Surg.* 2012;36:1898–1905.
10. Dighe S, Purkayastha S, Swift L, et al. Diagnostic precision of CT in local staging of colon cancers: a meta-analysis. *Clin Radiol*. 2010;65:708–719.
11. Lim M, Hussain Z, Howe A, et al. The oncological outcome after right hemicolectomy and accuracy of CT scan as a preoperative tool for staging in right sided colonic cancers. *Colorectal Dis*. 2012;15:536–543.
12. Iafrate F, Laghi A, Paolantonio P, et al. Preoperative staging of rectal cancer with MR imaging: correlation with surgical and histopathologic findings. *Radiography*. 2006;26:701–714.
13. Yeung JM, Ferris NJ, Lynch AC, et al. Preoperative staging of rectal cancer. *Future Oncol*. 2009;5:1295–1306.
14. Van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 2002;415:530–536.
15. Carole LY, Berndt ML. Review of the literature examining the correlation among DNA microarray technologies. *Environ Mol Mutagen.* 2007;48:380–394.
16. Barrier A, Boelle PY, Roser F, et al. Stage II colon cancer prognosis prediction by tumor gene expression profiling. *J Clin Oncol*. 2006;24:4685–4691.
17. Barrier A, Roser F, Boelle PY, et al. Prognosis of stage II colon cancer by non-neoplastic mucosa gene expression profiling. *Oncogene*. 2007;26:2642–2648.
18. Tillinghast GW. Microarrays in the clinic. *Nat Biotechnol*. 2011;57:688–690.
19. Shi L, Shi L, Reid L, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 2006;24:1151–1161.
20. Edgar R, Domrachev M, Lash A. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2001;30:207–210.
21. Ikeo K, Isji-i J, Tamura T, et al. CIBEX: center for information biology gene expression database. *C R Biol*. 2003;326:1079–1082.
22. Brazma A, Parkinson H, Sarkans U, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 2001;31:68–71.
23. Sherlock G, Hernandez-Boussard T, Kasarskis A, et al. The Stanford Microarray Database. *Nucleic Acids Res.* 2003;29:152–155.
24. Jorissen RN, Gibbs P, Christie M, et al. Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. *Clin Cancer Res.* 2009;15:7642–7651.
25. Brazma A, Hingamp P, Quackenbush J, et al. Minimum Information About a Microarray Experiment (MIAME)–toward standards for microarray data. *Nat Genet.* 2001;29:365–371.
26. Groene J, Mansmann U, Meister R, et al. Transcriptional census of 36 microdissected colorectal cancers yields a gene signature to distinguish UICC II and III. *Int J Cancer*. 2006;119:1829–1836.
27. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3.
28. O'Connell MJ, Lavery I, Yothers G, et al. Relationship between tumor gene expression and recurrence in four independent studies of patients with stage II/III colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin. *J Clin Oncol.* 2010;28:3937–3944.
29. Bipat S, Glas AS, Slors FJ, et al. Rectal cancer: local staging and assessment of lymph node involvement with endoluminal US, CT, and MR imaging—a meta-analysis. *Radiology*. 2004;232:773–783.
30. Mao S, Wang C, Dong G. Evaluation of inter-laboratory and cross-platform concordance of DNA microarrays through discriminating genes and classifier transferability. *J Bioinform Comput Biol*. 2009;7:157–173.
31. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet.* 2010;11:733–739.
32. Barnes M, Freudenberg J, Thompson S, et al. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression. *Nucleic Acids Res.* 2005;33:5915–5923.
33. Lu YY, Chen JH, Ding JH, et al. A systematic review and meta-analysis of pretherapeutic lymph node staging of colorectal cancer by 18F-FDG PET or PET/CT. *Nucl Med Commun*. 2012;33:1127–1133.
34. Morson BC, Whiteway JE, Jones EA. Histopathology and prognosis of malignant colorectal polyps treated by endoscopic polypectomy. *Gut*. 1984;25:437–444.
35. Smith J, Hwang H, Wiseman KW, et al. Ex vivo sentinel node mapping in colon cancer: improving the accuracy of pathological staging. *Am J Surg.* 2006;191:665–668.
36. Stojadinovic A, Nissan A, Protic M, et al. Prospective randomized study comparing sentinel lymph node evaluation with standard pathologic evaluation for the staging of colon carcinoma: results from the United States Military Cancer Institute Clinical Trials Group Study GI-01. *Ann Surg.* 2007;245:846–857.
37. Kelder W, Van Der Berg A, Van Der Leij J, et al. RT-PCR and immunohistochemical evaluation of sentinel nodes after in vivo mapping with patent blue V in colon cancer patients. *Scand J Gastroenterol*. 2006;41:1073–1078.
38. Calvano S, Xiao W, Richards D, et al. A network-based analysis of systemic inflammation in humans. *Nature*. 2005;437:1032–1037.
39. Baillat G, Siret C, Delamarre E, et al. Early adhesion induces interaction of FAK and Fyn in lipid domains and activates raft-independent Akt signaling in SW480 colon cancer cells. *Biochim Biophys Acta*. 2008;1783:2323–2331.
40. Lourenco G, Cardossa-Filho C, Goncales N, et al. A high risk of occurrence of sporadic breast cancer in individuals with the 104NN polymorphism of the COL18A1 gene. *Breast Cancer Res Treat*. 2006;100:335–338.
41. Hwang K, Chung J, Jung M, et al. COL18A1 gene for the prognostic marker of breast cancer according to the analysis of the DNA copy number variation by array CGH. *J Breast Cancer*. 2010;13:37–45.
42. Salzberg SL. Genome re-annotation: a wiki solution. *Genome Biol*. 2007;8:102.

© *2013 Lippincott Williams & Wilkins*

Copyright © 2013 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.