Performance of three biomarker immunohistochemistry for intrinsic breast cancer subtyping in the AMBER consortium

Emma H Allott[1,2], Stephanie M Cohen[1,3], Joseph Geradts[4], Xuezheng Sun[2], Thaer Khoury[5], Wiam Bshara[5], Gary R Zirpoli[6], C Ryan Miller[1,3,7], Helena Hwang[8], Leigh B Thorne[7], Siobhan O'Connor[7], Chiu-Kit Tse[2], Mary B Bell[1], Zhiyuan Hu[1], Yan Li[1], Erin L Kirk[2], Traci N Bethea[9], Charles M Perou[1], Julie R Palmer[9], Christine B Ambrosone[6], Andrew F Olshan[1,2], Melissa A Troester[1,2]

[1]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC

[2]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC

[3]Translational Pathology Laboratory, University of North Carolina at Chapel Hill, Chapel Hill, NC

[4]Department of Pathology, Duke University, Durham, NC

[5]Departments of Pathology, Roswell Park Cancer Institute, Buffalo, NY

[6]Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY

[7]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC

[8]Department of Pathology, University of Texas Southwestern, Dallas, TX

[9]Slone Epidemiology Center, Boston University, Boston, MA, USA

Running title: Accuracy of protein and RNA-based breast cancer subtyping

Key words: automated digital pathology, breast cancer, estrogen receptor, HER2, immunohistochemistry, intrinsic subtype, molecular pathologic epidemiology, PAM50, progesterone receptor

Financial support: This work was supported by the National Cancer Institute [5P01CA151135-04 (AMBER Consortium; J. Geradts, T. Khoury, W. Bshara, G. Zirpoli, H. Hwang, C. M. Perou, J. R. Palmer, C. B. Ambrosone, A. F. Olshan, M. A. Troester), P50-CA058223 (SPORE in breast cancer; C. R. Miller, S. O'Connor, C. M. Perou, A. F. Olshan, M. A. Troester), U01-CA179715 (C.M. Perou, M.A. Troester), the University Cancer Research Fund, University of North Carolina at Chapel Hill (E. H. Allott, L. Thorne, C-K. Tse, M. E. Bell, E. L. Kirk, C. M. Perou, A. F. Olshan, M. A. Troester), and the Breast Cancer Research Foundation (C. B. Ambrosone, C. M. Perou). The Translational Pathology Laboratory (S. M. Cohen, C. R. Miller) is supported in part by grants from the National Cancer Institute (3P30CA016086) and the University of North Carolina at Chapel Hill University Cancer Research Fund.

Corresponding author: Dr. Melissa A. Troester, Department of Epidemiology, CB 7435, University of North Carolina at Chapel Hill, 135 Dauer Drive, Chapel Hill, NC 27599, USA. Phone: +1 919-966-7408; Fax: +1 919-966-2089; E-mail: troester@unc.edu

Conflict of interest statement: C. M. Perou is an equity stock holder and Board of Director Member of BioClassifier LLC. C. M. Perou is also listed an inventor on a patent application on the PAM50 molecular assay. None of the other authors declare any conflicts of interest.

Word count: 3,996; **Figures:** 1; **Tables:** 5

Abstract

Background: Classification of breast cancer into intrinsic subtypes has clinical and epidemiologic importance. To examine accuracy of immunohistochemistry (IHC)-based methods for identifying intrinsic subtypes, a three-biomarker IHC panel was compared to the clinical record and RNA-based intrinsic (PAM50) subtypes.

Methods: Automated scoring of estrogen receptor (ER), progesterone receptor (PR) and HER2 was performed on IHC-stained tissue microarrays (TMAs) comprising 1,920 cases from the African American Breast Cancer Epidemiology and Risk (AMBER) consortium. Multiple cores (1-6/case) were collapsed to classify cases, and automated scoring was compared to the clinical record and to RNA-based subtyping.

Results: Automated analysis of the three-biomarker IHC panel produced high agreement with the clinical record (93% for ER and HER2, and 88% for PR). Cases with low tumor cellularity and smaller core size had reduced agreement with the clinical record. IHC-based definitions had high agreement with the clinical record regardless of hormone receptor positivity threshold (1% vs. 10%), but a 10% threshold produced highest agreement with RNA-based intrinsic subtypes. Using a 10% threshold, IHC-based definitions identified the basal-like intrinsic subtype with high sensitivity (86%), while sensitivity was lower for luminal A, luminal B and HER2-enriched subtypes (76%, 40% and 37%, respectively).

Conclusion: Three-biomarker IHC-based subtyping has reasonable accuracy for distinguishing basal-like from non-basal-like, while additional biomarkers are required for accurate classification of luminal A, luminal B and HER2-enriched cancers.

Impact: Epidemiologic studies relying on three-biomarker IHC status for subtype classification should use caution when distinguishing luminal A from luminal B and when interpreting findings for HER2-enriched cancers.

Introduction

Breast cancer is a heterogeneous disease, comprised of distinct tumor subtypes [1]. While some epidemiologic studies have used additional markers or assays to define etiologic subtypes [2-4], the majority relied on estrogen receptor (ER), progesterone receptor (PR) and HER2 to classify breast cancers as hormone receptor (HR) positive, HER2-positive or triple negative [5-8]. These studies have identified heterogeneity in breast cancer etiology, but some lack of agreement between studies may be attributable to discordant subtype classification. Few studies have systematically compared immunohistochemical (IHC), clinical record and RNA-based intrinsic subtypes to estimate classification accuracy, particularly using tissue microarrays (TMAs) [9, 10].

The advent of automated methods to analyze digital pathology data has begun to support high-throughput IHC-based breast cancer subtyping in large epidemiologic studies [11]. Simultaneously, RNA-based methods have become more readily available for application in formalin-fixed paraffin-embedded (FFPE) tissues [12]. In the present study, we carried out IHC staining for ER, PR and HER2 using TMAs containing multiple cores per case for each of 1,920 cases in the African American Breast Cancer Epidemiology and Risk (AMBER) consortium. Herein, we describe 1) automated digital quantification of this three biomarker IHC panel, 2) methods for core-to-case collapsing, and 3) IHC-based agreement with clinical scoring in the medical record. Finally, treating PAM50 as the reference standard for the purposes of the present analysis, we evaluate 4) agreement between three biomarker IHC-based subtype classification and RNA-based PAM50 intrinsic subtyping. While earlier studies employed a 10% threshold [13-15], current clinical guidelines recommend a 1% threshold for defining ER and PR positivity status [16], and we evaluate the impact of this revised threshold on agreement between IHC-based and RNA-based intrinsic subtyping.

Materials and Methods

Study population and tissue microarray construction

The African American Breast Cancer Epidemiology and Risk (AMBER) consortium is comprised of observational studies of breast cancer in African American women [17]. The present study includes 1,920 cases within the AMBER consortium for which paraffin-embedded tissue was available in tissue microarrays (TMAs). Cases were from the Carolina Breast Cancer Study Phase 3 (CBCS; n=599), the Black Women's Health Study (BWHS; n=199), and the Women's Circle of Health Study (WCHS; n=309). In addition, we included 645 non-AA cases from CBCS and 168 non-AA cases from WCHS (Table 1) as these cases were present on the same TMAs and were relevant to our evaluation of subtype classification rates. Clinical ER, PR and HER2 status were abstracted from medical records. Tumor grade was centrally assigned for CBCS cases by a single pathologist (JG) using the Nottingham breast cancer grading system [18]. Each study was approved by Institutional Review Boards at participating hospitals and academic institutions.

Paraffin-embedded tumor blocks were requested from participating pathology laboratories for each case. Study pathologists (JG, HH, TK) marked hematoxylin & eosin (H&E)-stained slides to indicate areas enriched for invasive breast cancer for coring, and TMA construction and sectioning were carried out at the Translational Pathology Lab (TPL), University of North Carolina at Chapel Hill (UNC) for CBCS and at Roswell Park Cancer Institute (RPCI) for BWHS & WCHS. TMA blocks included between one and four tumor cores per case measuring 1.0 mm (CBCS), or between one and six tumor cores per case measuring 0.6 mm in diameter (BWHS and WCHS; Table 1). For CBCS, top and bottom sections from blocks containing a total of 4783 cores were stained with H&E and examined by study pathologists (JG or LT) for presence of tumor cells, and 503 cores (11%) lacking sufficient tumor cellularity (<50 tumor nuclei per core) were excluded from analyses. For TMA blocks not manually evaluated for tumor cellularity (BWHS and WCHS), digital analysis of IHC was used to eliminate cores

lacking sufficient tumor cellularity (<50 tumor nuclei per core) and, of a total of 2060 cores, 290 (14%) were excluded due to low tumor cellularity.

Immunohistochemistry and RNA-based assays

TMA blocks were cut into 5 μM serial sections and IHC staining protocols were optimized under pathologist supervision to achieve the best possible agreement with the clinical record. Final conditions were independently reviewed by multiple pathologists. Detailed IHC methods are described in Supplementary Materials and Methods.

Nanostring assays were used to measure the PAM50 gene signature on 535 cases from CBCS. Two cores from the same tumor block used for TMA construction were randomly sampled and pooled for analysis. The areas surrounding the holes left by the cores were examined by a study pathologist (JG or LT) to confirm the presence of tumor cells in the cores used for RNA extraction. RNA was isolated using the RNeasy FFPE kit from Qiagen and Nanostring analyses were performed in the Rapid Adoption Molecular (RAM) laboratory at UNC. Nanostring probe sets included 200 genes of which 50 were the PAM50 genes [19] and five were housekeeping genes. Gene expression was median-centered and samples were standardized to zero mean and unit variance. The NanoStringNorm package in Bioconductor was used to eliminate samples that did not have sufficient nanostring data quality (n=43) and the PAM50 predictor was performed as previously described [19] to categorize breast tumors into five intrinsic subtypes (luminal A, luminal B, HER2-enriched, basal-like, normal-like). Tumors classified as normal-like may result from extensive normal epithelial or stromal content in the tumor [20], so we excluded normal-like tumors (n=29) from our analysis. We also excluded 16 cases with equivocal HER2 IHC status, as we could not assess IHC-based subtype, giving rise to 447 cases with both IHC and intrinsic subtyping data.

Automated digital quantification of a three biomarker IHC-based panel

Automated digital image analysis of IHC staining (Figure 1) was performed using a Genie classifier and the Nuclear v9 algorithm (for ER and PR) or Membrane v9 algorithm (for HER2; Aperio Technologies, Vista, CA), and is described in more detail in Supplementary Materials and Methods.

To validate automated analyses, study pathologists (JG, HH, TK) carried out manual reviews of IHC staining within a training set of TMAs. For ER and PR, individual cores were classified as negative (<1% positive), borderline (≥1% and <10%) or positive (≥10%). For HER2, individual cores were classified as negative (0/1+), positive (3+) or equivocal (2+) [21]. Using a 10% threshold to define ER and PR status, agreement between automated and manual scoring was 89% for ER and 91% for PR (Supplementary Table S1). After excluding cores with equivocal HER2 (2+) status in either the TMA or clinical record (n=63), HER2 status agreement between automated and manual scoring was 93%. As expected, agreement was slightly lower with additional categories [ER and PR status as negative (<1%), borderline (1-10%) or positive (≥10%), HER2 status as negative (0/1+), equivocal (2+) or positive (3+), Supplementary Table S1].

Core-to-case collapsing

We tested two core-to-case collapsing methods to define a single ER, PR and HER2 status for each case. The first assigned case-level status using a tumor cellularity-weighted approach. The weighted average of percent positivity was calculated by summing the product of percent positivity and core weight across all cores per case. Core weight was defined as the number of tumor nuclei in a given core divided by the total number of tumor nuclei across all cores for that case. Thresholds for ER, PR (≥1% and ≥10%) and HER2 (≥10%) positivity were subsequently applied to define a dichotomous status for each of these three markers. Tumor cells staining positive at any intensity were considered positive for ER and PR, while only tumor cells staining at the 3+ level were considered positive for HER2. Equivocal (2+) HER2 status was defined as <10% of tumor cells staining at the 3+ level and ≥10% of tumor cells staining at the 2+ level, while negative HER2 status included all other cases. The 10% threshold for HER2 was optimized to agree with manual scoring performed according to the then-current HER2 guidelines,

indicating that a 30% threshold should be used for HER2 [21]. The second core-to-case collapsing method classified the case as ER, PR or HER2 positive if any core was positive for ER, PR or HER2, respectively. Core-level ER and PR positivity was defined considering both ≥1% and ≥10% thresholds, while HER2 positivity was defined as ≥10% of cells staining at the 3+ intensity level.

IHC-based agreement with clinical scoring in the medical record

Kappa statistics were used to examine case-level agreement between TMA and clinical ER, PR and HER2 status in the medical record [22]. Given that the majority of cases (76%) were diagnosed prior to ASCO guidelines recommending the use of a 1% threshold for classification of HR positivity [16], our primary analysis of case-level agreement between TMA and clinical ER, PR and HER2 status was conducted using a 10% threshold to define HR positivity status. Analyses were stratified by core-to-case collapsing method in addition to core diameter and median cellularity, in order to determine the impact of these factors on agreement with the clinical record. We conducted sensitivity analyses using a 1% threshold for ER and PR, and excluding cases with only one core.

Agreement between IHC-based and RNA-based subtyping

Using case-level ER, PR and HER2 status, we defined IHC-based subtypes as follows: ER or PR positive, HER2 negative (luminal A); ER or PR positive, HER2 positive (luminal B); ER and PR negative, HER2 positive (HER2-enriched); ER and PR and HER2 negative (basal-like). We also defined IHC-based subtypes using a PR cut point of 20% [23]; ER positive or PR>20%, HER2 negative (luminal A); ER or PR positive, HER2 positive *or* ER or PR≤20%, HER2 negative (luminal B). In the absence of Ki67 data, tumor grade may distinguish luminal A and HER2-negative luminal B cancers [24], and we explored combined grade as follows; ER positive or PR>20%, combined grade I or II, HER2 negative (luminal A); ER or PR positive, HER2 positive *or* ER or PR≤20% or combined grade III, HER2 negative (luminal B). Dichotomizing combined grade as I vs. II/III reduced subtype classification accuracy, and mitotic grade offered no classification advantages over combined grade and therefore these results are not

presented. Finally, we defined HER2-enriched cases as ER negative, HER2 positive, regardless of PR status. Cases with equivocal HER2 status remained unclassified and were excluded from subsequent analyses. We examined the impact of using 1% and 10% thresholds to define ER and PR status on the sensitivity [(true positives/(true positives + false negatives)], specificity [true negatives/(true negatives + false positives)], and accuracy [(true positives + true negatives)/total] of TMA IHC subtyping with respect to intrinsic subtyping for a subset of cases (n=447).

Statistical analyses were conducted using SAS version 9.3 (SAS Institute, Cary, NC) and STATA version 13.0 (Stata Corp., College Station, TX).

Results

IHC-based agreement with the clinical record: impact of core-to-case collapsing method

Using the weighted core-to-case collapsing method to define HR positivity status, agreement with the clinical record was 93% for ER and 88% for PR (using a ≥10% threshold; Table 2). These results were not impacted by a 1% positivity threshold (data not shown), nor was agreement substantially altered when using the any positive core-to-case collapsing method (Table 2). However, the any positive method with a 1% threshold resulted in reduced agreement with the clinical record (90% agreement for ER, 85% for PR). As such, we chose to proceed with the weighted method and 10% threshold to maximize agreement with clinical laboratory data for defining ER and PR positivity in our population.

For HER2, agreement rates were similar for weighted and any positive methods (95% and 93%, respectively; Table 3). In addition, while both methods had high specificity (95% and 98% for any positive and weighted methods, respectively), the any positive method had higher sensitivity for identifying HER2 positive cases (82% vs. 75%, respectively). However, the any positive method produced a higher number of equivocal (2+) cases, relative to the weighted method (Table 3). A hybrid of both core-to-case collapsing methods, defining HER2 status based on being positive by either the weighted or any positive method, maximized sensitivity (78%) and minimized the number of equivocal cases, while maintaining similar levels of agreement with the clinical record (Table 3). Moreover, the number of equivocal cases using the hybrid method (n=74; 4%) was similar to that of the clinical record (n=50; 3%). As such, we chose to proceed with this hybrid method to define HER2 status.

For all three biomarkers, agreement between automated analysis of TMAs and the clinical record was similar to agreement rates reported between automated analysis of TMAs and manual review (Supplementary Table S1).

IHC-based agreement with the clinical record: impact of TMA characteristics

Core cellularity and diameter impacted agreement rates between TMAs and the clinical record. Using a 10% threshold to define HR positivity, cases with high tumor cellularity had higher agreement with the clinical record for ER and PR, as shown in Table 4. Similarly, cases with larger cores had higher ER agreement and PR agreement (Table 4). However, neither core cellularity nor diameter substantially impacted agreement with the clinical record for HER2, with similar agreement rates for tumors with high and low cellularity, and for 1.0 mm and 0.6 mm cores (Table 4). Using a 1% threshold to define ER and PR status did not alter agreement rates for cores with high cellularity and 1.0 mm diameter, but improved agreement for cores with low cellularity and 0.6 mm diameter (data not shown). Excluding cases with only one core (7% of cases) did not impact our findings (data not shown).

Agreement between IHC-based and RNA-based PAM50 intrinsic subtyping

For the subset of CBCS cases with intrinsic subtyping data (n=447), agreement between three biomarker IHC and RNA-based subtypes is shown in Table 5. Given recent changes in clinical thresholds for ER and PR positivity [16], both 1% and 10% thresholds were considered to assess agreement. While IHC-based definitions of luminal A (ER or PR positive and HER2 negative) and luminal B (ER or PR positive and HER2 positive) resulted in high sensitivity for identifying luminal A tumors, specificity was low for both 1% and 10% thresholds. Conversely, luminal B specificity was high but sensitivity was low, regardless of threshold (Supplementary Table S2). Work by Prat and colleagues [25], reflected in St. Gallen guidelines [23], indicated that cases with negative HER2 status but low (≤20%) PR positivity should be classified as luminal B, as an additional means of distinguishing luminal A from luminal B. Using this definition, the 10% ER threshold yielded greatest accuracy (Table 5). St Gallen guidelines also suggest that Ki67 expression may distinguish luminal A and HER2-negative luminal B disease [23]. Given that we lacked Ki67 data, we explored tumor grade as a surrogate [24]. Overall, gains in sensitivity were offset by losses in specificity and *vice versa*, and accuracy was substantively unchanged (~70-80%) regardless of grade (Supplementary Table S3).

HER2-enriched cases identified by RNA-based PAM50 analysis were difficult to accurately identify by IHC. Defining HER2-enriched cases as ER and PR negative, HER2 positive resulted in low sensitivity, although specificity was high (Supplementary Table S2). We found that defining HER2-enriched cases as ER negative and HER2 positive, regardless of PR status, resulted in slightly improved agreement with the intrinsic HER2-enriched subtype (Table 5). Using a 10% threshold, we found that 26% of HER2-enriched cases identified by RNA-based intrinsic subtyping were classified as luminal B by IHC, while the use of a 1% threshold gave rise to 43% of HER2-enriched cases identified by RNA-based intrinsic subtyping which were classified as luminal B by IHC (Table 5). Finally, utilization of clinical HER2 status (which includes ISH data) to identify HER2-enriched cases did not improve on agreement rates observed using TMA HER2 status (data not shown).

Triple negative status (ER, PR and HER2 negative) using the three biomarker IHC panel had relatively high accuracy for identifying the basal-like intrinsic subtype (Table 5). At the 10% threshold, we found that 10% of basal-like cases identified by RNA-based intrinsic subtyping were classified as luminal A or B by IHC (Table 5). However, rates of discordant classification between IHC and RNA-based intrinsic subtyping were higher at the 1% threshold, with 27% of basal-like cases identified by RNA-based intrinsic subtyping classified as luminal A or B by IHC (Table 5). Including normal-like intrinsic cases (n=26 with complete IHC data) in our comparisons between three biomarker IHC and RNA-based subtypes did not alter our findings (data not shown).

Discussion

In this consortium comprising three large epidemiologic studies, we found that classification of ER, PR and HER2 status using automated digital pathology resulted in high agreement with the clinical record for all three biomarkers. Agreement rates were highest when using a tumor cellularity-weighted method for core-to-case collapsing and in cases with high tumor cellularity. Furthermore, we report that this three biomarker IHC-based panel conferred high sensitivity and specificity for identifying tumors classified as basal-like using RNA-based intrinsic subtyping. However, our findings highlight the need for further optimization of methods to identify HER2-enriched breast cancers and to discriminate between luminal A and B cancers, likely through inclusion of new biomarkers like Ki67 and/or other biomarkers uniquely present in luminal B or HER2-enriched tumors. As such, we recommend that epidemiologic studies relying on ER, PR and HER2 IHC status for subtype classification use caution when distinguishing luminal A from luminal B cancers and when interpreting findings for HER2-enriched cancers.

The advent of digital pathology carries potential to advance molecular epidemiologic pathology research by providing a platform to obtain high-throughput and high-resolution quantitative data. Previous studies based on tumor-bank series have shown that inter-laboratory agreement rates for ER and PR range from 80-90% [26-28]. Our findings from this consortium, which drew from multiple clinical centers across the United States, showed that agreement rates were on the high end of this range for both biomarkers, thereby providing validation for digital IHC subtyping methods in epidemiologic research. Furthermore, our agreement rates for HER2 exceeded previously published inter-laboratory agreement rates of ~80% for IHC HER2 staining [29]. These findings may reflect improvements over time in IHC methods, improved efforts to ensure antigen stability, and other methodological strengths of our work.

While technical factors have been demonstrated to play a role [27, 30, 31], tumor biology and molecular factors may also drive discordance in biomarker classification. We and others [26, 28] have shown that PR agreement is often lower than that of ER, potentially due to more heterogeneous

expression within a given tumor, or intratumoral heterogeneity. Intratumoral heterogeneity would be expected to pose a greater classification problem for cores representing a smaller sample of the tumor (as reflected by lower tumor cellularity and/or smaller core diameter); indeed, our findings suggest that agreement rates were lowest for PR. This explanation seems less likely to explain HER2 discordance, as neither core size nor cellularity was associated with agreement between TMAs and the clinical record. As such, intratumoral heterogeneity, in addition to technical factors including TMA characteristics, may contribute to the level of agreement between central and clinical biomarker classification.

Recent clinical guidelines have lowered the threshold used to define ER and PR positivity from 10% to 1%, given evidence that cases with borderline (≥1-<10%) ER and PR positivity derive a benefit from endocrine therapy [16]. However, subtype distribution within these borderline cases, comprising approximately 6% of total breast cancer cases, is not well understood. One study indicated that approximately one quarter of ER borderline cases are luminal, while up to half are basal-like [32]. This is in general agreement with our findings that 25% of ER borderlines were luminal, and 38% were basal-like. Conversely, another study reported higher prevalence of luminals (44%) and lower prevalence of basal-likes (18%) among ER or PR borderline cases [33], suggesting that biological characteristics of the borderline group may vary across study populations or according to technical methods. We report that IHC-based surrogate definitions of PAM50 intrinsic subtypes were impacted by the threshold for defining ER and PR positivity, with a 1% threshold resulting in a greater proportion of basal-like cases classified as luminal A and B by IHC, and of HER2-enriched cases classified as luminal B by IHC, relative to the 10% threshold. Given the heterogeneity of borderline breast cancers, additional studies should refine classification for this group of tumors. Nonetheless, these findings demonstrate that changing clinical guidelines for HR positivity may adversely affect the accuracy of intrinsic subtype classification using a three biomarker IHC panel, potentially impacting the stability of epidemiologic findings over time.

While it has been proposed that additional biomarkers such as Ki67 are required to accurately identify luminal A and luminal B cancers, these data are not commonly available in population-based

studies. In the absence of Ki67, tumor grade and quantitative PR expression may be useful in distinguishing luminal A and luminal B cancers [23, 24]. However, we found that maximal accuracy was ~80% for luminal A cases and ~75% for luminal B cases with the inclusion of these variables, suggesting that grade and PR thresholds are inadequate to accurately distinguish luminal subtypes. IHC is the most commonly-used method for clinical classification of HER2-positive tumors, but even with inclusion of clinically ISH-positive cases in our HER2-enriched group, the sensitivity of the three biomarker IHC panel was low for identifying RNA-based HER2-enriched cases. It is well known that amplification or over-expression of an oncogenic pathway can occur via multiple molecular events, and a single marker cannot always accurately detect defects in a pathway. This molecular heterogeneity may contribute to discordance between the three biomarker IHC-based panel and the multi-gene RNA-based panel to identify HER2 pathway enrichment. Thus, both molecular and intratumoral heterogeneity may add to technical factors in producing discordance between central and clinical analyses of tumor subtype. These data underscore that population-based studies with three biomarker IHC-based data should avoid making etiologic inferences about luminal A, luminal B and HER2-enriched intrinsic subtypes until better IHC-based subtyping methods can be identified for these subtypes. In contrast, high sensitivity and specificity of the three biomarker IHC panel for identifying basal-like breast cancers in our study suggests that additional biomarkers such as CK5/6 and EGFR may not be needed to classify basal-like cancers in epidemiologic studies. In sum, the current data suggest that a three biomarker IHC panel is able to distinguish between basal-like and luminal, two key etiologic subtypes [34, 35], but that finer resolution of intrinsic subtyping in non-basal-like cancers requires additional markers.

Our findings should be considered in the context of strengths and weaknesses of this study. First, although cores without invasive carcinoma were excluded based upon manual review of top and bottom H&E-stained sections by study pathologists, it is possible that benign epithelium or ductal carcinoma in situ (DCIS) was counted as invasive tumor in automated IHC quantification. However, high levels of agreement with the clinical record, exceeding that of previous studies, mitigate concerns about this

potential source of bias. Second, while the present analysis offers insights into IHC-based surrogates for intrinsic subtyping within the context of a consortium of epidemiologic studies, it is important to note that this study represents an incomplete sample of AMBER, specifically representing those for whom we had TMAs. Because this sample does not represent the entire population of AMBER, it cannot be used to make inferences about the distribution of subtypes within the AMBER source population. Third, we had incomplete data on whether tumor blocks used clinically were the same blocks provided to our study. Thus, the biospecimens may have differed leading to a downward bias in the estimation of agreement between TMAs and the clinical record. Strengths of this study include validation of automated staining guided by multiple pathologists, availability of a validated RNA-based multi-gene assay for molecular classification of tumor subtype, and a large, diverse sample population representing African American and Caucasian women.

In conclusion, we report that automated digital analysis of IHC staining for ER, PR and HER2 on TMAs resulted in high agreement with the clinical record, and high sensitivity and specificity for identifying basal-like breast cancer cases classified by RNA-based intrinsic subtyping. However, we report reduced sensitivity and specificity for identifying luminal A, luminal B, and HER2-enriched cancers using IHC-based panels, indicating that additional biomarkers are required for IHC-based classification of these intrinsic subtypes.

References

1. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000; 406:747-52.
2. Ma H, Wang Y, Sullivan-Halley J, Weiss L, Marchbanks PA, Spirtas R, et al. Use of four biomarkers to evaluate the risk of breast cancer subtypes in the women's contraceptive and reproductive experiences study. Cancer Res 2010; 70:575-87.
3. Millikan RC, Newman B, Tse CK, Moorman PG, Conway K, Dressler LG, et al. Epidemiology of basal-like breast cancer. Breast Cancer Res Treat 2008; 109:123-39.
4. Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, Milne RL, et al. Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. J Natl Cancer Inst 2011; 103:250-63.
5. Phipps AI, Chlebowski RT, Prentice R, McTiernan A, Wactawski-Wende J, Kuller LH, et al. Reproductive history and oral contraceptive use in relation to risk of triple-negative breast cancer. J Natl Cancer Inst 2011; 103:470-7.
6. Gaudet MM, Press MF, Haile RW, Lynch CF, Glaser SL, Schildkraut J, et al. Risk factors by molecular subtypes of breast cancer across a population-based study of women 56 years or younger. Breast Cancer Res Treat 2011; 130:587-97.
7. Anderson K, Thompson PA, Wertheim BC, Martin L, Komenaka IK, Bondy M, et al. Family history of breast and ovarian cancer and triple negative subtype in hispanic/latina women. Springerplus 2014; 3:727.
8. Kwan ML, Kushi LH, Weltzien E, Maring B, Kutner SE, Fulton RS, et al. Epidemiology of breast cancer subtypes in two prospective cohort studies of breast cancer survivors. Breast Cancer Res 2009; 11:R31.

9. Bastien RR, Rodriguez-Lescure A, Ebbert MT, Prat A, Munarriz B, Rowe L, et al. PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. BMC Med Genomics 2012; 5:44.

10. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. Clin Cancer Res 2010; 16:5222-32.

11. Howat WJ, Blows, F.M., Provenzano, E., Brook, M.N., Morris, L., Gazinska, P., Johnson, N. Performance of automated scoring of ER, PR, HER2, CK5/6 and EGFR in breast cancer tissue microarrays in the Breast Cancer Association Consortium. The Journal of Pathology: Clinical Research 2015; 1:18-32.

12. Sweeney C, Bernard PS, Factor RE, Kwan ML, Habel LA, Quesenberry CP, Jr., et al. Intrinsic subtypes from PAM50 gene expression assay in a population-based breast cancer cohort: differences by age, race, and tumor characteristics. Cancer Epidemiol Biomarkers Prev 2014; 23:714-24.

13. Viale G, Regan MM, Maiorano E, Mastropasqua MG, Dell'Orto P, Rasmussen BB, et al. Prognostic and predictive value of centrally reviewed expression of estrogen and progesterone receptors in a randomized trial comparing letrozole and tamoxifen adjuvant therapy for postmenopausal early breast cancer: BIG 1-98. J Clin Oncol 2007; 25:3846-52.

14. Dowsett M, Allred C, Knox J, Quinn E, Salter J, Wale C, et al. Relationship between quantitative estrogen and progesterone receptor expression and human epidermal growth factor receptor 2 (HER-2) status with recurrence in the Arimidex, Tamoxifen, Alone or in Combination trial. J Clin Oncol 2008; 26:1059-65.

15. Regan MM, Viale G, Mastropasqua MG, Maiorano E, Golouh R, Carbone A, et al. Re-evaluating adjuvant breast cancer trials: assessing hormone receptor status by immunohistochemical versus extraction assays. J Natl Cancer Inst 2006; 98:1571-81.

16. Hammond ME, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, et al. American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. J Clin Oncol 2010; 28:2784-95.

17. Palmer JR, Ambrosone CB, Olshan AF. A collaborative study of the etiology of breast cancer subtypes in African American women: the AMBER consortium. Cancer Causes Control 2014; 25:309-19.

18 Elston CW, Ellis IO. Pathological prognostic features in breast cancer. The value of histological grade in breast cancer: experience from a large study with long-term follow up. Histopathology 1991; 19:403-10.

19. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009; 27:1160-7.

20. Elloumi F, Hu Z, Li Y, Parker JS, Gulley ML, Amos KD, et al. Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. BMC Med Genomics 2011; 4:54.

21. Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. J Clin Oncol 2007; 25:118-45.

22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74.

23. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thurlimann B, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann Oncol 2013; 24:2206-23.

24. Goldhirsch A, Wood WC, Coates AS, Gelber RD, Thurlimann B, Senn HJ, et al. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St. Gallen International

Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann Oncol 2011; 22:1736-47.

25. Prat A, Cheang MC, Martin M, Parker JS, Carrasco E, Caballero R, et al. Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal A breast cancer. J Clin Oncol 2013; 31:203-9.

26. Badve SS, Baehner FL, Gray RP, Childs BH, Maddala T, Liu ML, et al. Estrogen- and progesterone-receptor status in ECOG 2197: comparison of immunohistochemistry by local and central laboratories and quantitative reverse transcription polymerase chain reaction by central laboratory. J Clin Oncol 2008; 26:2473-81.

27. Collins LC, Marotti JD, Baer HJ, Tamimi RM. Comparison of estrogen receptor results from pathology reports with results from central laboratory testing. J Natl Cancer Inst 2008; 100:218-21.

28. Ma H, Wang Y, Sullivan-Halley J, Weiss L, Burkman RT, Simon MS, et al. Breast cancer receptor status: do results from a centralized pathology laboratory agree with SEER registry reports? Cancer Epidemiol Biomarkers Prev 2009; 18:2214-20.

29. Press MF, Sauter G, Bernstein L, Villalobos IE, Mirlacher M, Zhou JY, et al. Diagnostic evaluation of HER-2 as a molecular target: an assessment of accuracy and reproducibility of laboratory testing in large, prospective, randomized clinical trials. Clin Cancer Res 2005; 11:6598-607.

30. Li S, Yang X, Zhang Y, Fan L, Zhang F, Chen L, et al. Assessment accuracy of core needle biopsy for hormone receptors in breast cancer: a meta-analysis. Breast Cancer Res Treat 2012; 135:325-34.

31. McCullough AE, Dell'orto P, Reinholz MM, Gelber RD, Dueck AC, Russo L, et al. Central pathology laboratory review of HER2 and ER in early breast cancer: an ALTTO trial [BIG 2-06/NCCTG N063D (Alliance)] ring study. Breast Cancer Res Treat 2014; 143:485-92.

32. Iwamoto T, Booser D, Valero V, Murray JL, Koenig K, Esteva FJ, et al. Estrogen receptor (ER) mRNA and ER-related gene expression in breast cancers that are 1% to 10% ER-positive by immunohistochemistry. J Clin Oncol 2012; 30:729-34.

33. Cheang MC, Martin M, Nielsen TO, Prat A, Voduc D, Rodriguez-Lescure A, et al. Defining Breast Cancer Intrinsic Subtypes by Quantitative Receptor Expression. Oncologist 2015.

34. Anderson WF, Rosenberg PS, Prat A, Perou CM, Sherman ME. How many etiological subtypes of breast cancer: two, three, four, or more? J Natl Cancer Inst 2014; 106.

35. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 2014; 158:929-44.

Table 1: Design and characteristics of tissue microarrays across participating studies

Study	Cases	TMAs	Cores per case*	Core diameter (mm)	Tumor cellularity*, median (IQR)
CBCS	1,244	63	1-4	1.0	4747 (2422-8403)
BWHS	199	10	1-3	0.6	1050 (493-1928)
WCHS	477	16	1-6	0.6	1290 (684-2277)

BWHS, Black Women's Health Study; CBCS, Carolina Breast Cancer Study; WCHS, Women's Circle of Health Study

*numbers are reported for estrogen receptor (ER), with similar numbers for progesterone receptor (PR) and HER2

Table 2: Case-level agreement between automated scoring of central tissue microarrays and clinical estrogen receptor (ER) and progesterone receptor (PR) status: impact of core-to-case collapsing method

		Clinical ER or PR status in the medical record*						
Central TMA ER or PR status*		**ER**			**PR**			
	Weighted core-to-case collapsing method[a]		**Negative**	**Positive**	**Total**		**Negative**	**Positive**	**Total**
		Negative	427	100	527	**Negative**	565	108	673
		Positive	15	1,095	1,110	**Positive**	74	803	877
		Total	442	1,195	1,637	**Total**	639	911	1,550
		Agreement	93%			**Agreement**	88%		
		Kappa (95% CI)	0.83 (0.80-0.86)			**Kappa (95% CI)**	0.76 (0.73-0.79)		
			Negative	**Positive**	**Total**		**Negative**	**Positive**	**Total**
	Any positive core-to-case collapsing method[b]	**Negative**	419	70	489	**Negative**	529	83	612
		Positive	23	1,125	1,148	**Positive**	110	828	938
		Total	442	1,195	1,637	**Total**	639	911	1,550
		Agreement	94%			**Agreement**	88%		
		Kappa (95% CI)	0.86 (0.83-0.88)			**Kappa (95% CI)**	0.74 (0.71-0.78)		

TMA=tissue microarray

[a]Weighted method: Case level positivity is weighted by the tumor cellularity of each contributing core

[b]Any positive method: Case is positive if any core is positive

*using a 10% threshold to define ER and PR positivity

Table 3: Case-level agreement between automated scoring of central tissue microarrays and clinical HER2 status: impact of core-to-case collapsing method

		Clinical HER2 status in the medical record			
Central TMA HER2 status	**Weighted core-to-case collapsing method[a]**		**Negative**	**Positive**	**Total**
		Negative	1,312	53	1,365
		Positive	31	163	194
		Total	1,343	216	1,559
		Agreement	95%		
		Kappa (95% CI)	0.76 (0.72-0.81)		
			Negative	**Positive**	**Total**
	Any positive core-to-case collapsing method[b]	**Negative**	1,181	37	1,218
		Positive	67	173	240
		Total	1,248	210	1,458
		Agreement	93%		
		Kappa (95% CI)	0.73 (0.68-0.78)		
			Negative	**Positive**	**Total**
	Hybrid core-to-case collapsing method[c]	**Negative**	1,300	48	1,348
		Positive	69	174	243
		Total	1,369	222	1,591
		Agreement	93%		
		Kappa (95% CI)	0.71 (0.66-0.76)		

TMA=tissue microarray

[a]Weighted method: Case level positivity is weighted by the tumor cellularity of each contributing core

[b]Any positive method: Case is positive if any core is positive

[c]Hybrid method: Case is positive if classified as positive by either the weighted or any positive method

[a]n=106 (6%), [b]n=207 (12%), and [c]n=74 (4%) cases with equivocal (2+) central HER2 status

n=50 (3%) cases with equivocal (2+) clinical HER2 status in the medical record

Table 4: Case-level agreement between automated scoring of central TMAs and clinical estrogen receptor (ER), progesterone receptor (PR) and HER2 status: impact of core cellularity and diameter

Central TMA ER, PR and HER2 status*		Clinical ER, PR and HER2 status in the medical record*											
		ER			PR				HER2				
	Cellularity ≥median		Negative	Positive	Total		Negative	Positive	Total		0/1+	3+	Total
		Negative	271	34	305	Negative	336	32	368	0/1+	639	24	663
		Positive	6	509	515	Positive	31	376	407	3+	30	104	134
		Total	277	543	820	Total	367	408	775	Total	669	128	797
		Agreement	95%			Agreement	92%			Agreement	93%		
		Kappa (95% CI)	0.89 (0.86-0.93)			Kappa (95% CI)	0.84 (0.80-0.88)			Kappa (95% CI)	0.75 (0.69-0.82)		
	Cellularity <median		Negative	Positive	Total		Negative	Positive	Total		0/1+	3+	Total
		Negative	156	66	222	Negative	229	76	305	0/1+	661	24	685
		Positive	9	586	595	Positive	43	427	470	3+	39	70	109
		Total	165	652	817	Total	272	503	775	Total	700	94	794
		Agreement	91%			Agreement	85%			Agreement	92%		
		Kappa (95% CI)	0.75 (0.70-0.80)			Kappa (95% CI)	0.67 (0.62-0.73)			Kappa (95% CI)	0.64 (0.56-0.73)		
	1.0 mm cores		Negative	Positive	Total		Negative	Positive	Total		0/1+	3+	Total
		Negative	350	53	403	Negative	470	54	524	0/1+	933	29	962
		Positive	9	788	797	Positive	59	601	660	3+	45	115	160
		Total	359	841	1,200	Total	529	655	1,184	Total	978	144	1,122
		Agreement	95%			Agreement	90%			Agreement	93%		
		Kappa (95% CI)	0.88 (0.85-0.91)			Kappa (95% CI)	0.81 (0.77-0.84)			Kappa (95% CI)	0.72 (0.66-0.78)		
	0.6 mm cores		Negative	Positive	Total		Negative	Positive	Total		0/1+	3+	Total
		Negative	77	47	124	Negative	95	54	149	0/1+	367	19	386
		Positive	6	307	313	Positive	15	202	217	3+	24	59	83
		Total	83	354	437	Total	110	256	366	Total	391	78	469
		Agreement	88%			Agreement	81%			Agreement	91%		
		Kappa (95% CI)	0.67 (0.59-0.75)			Kappa (95% CI)	0.59 (0.51-0.68)			Kappa (95% CI)	0.68 (0.59-0.77)		

*using a 10% threshold to define ER and PR positivity for ER and PR analysis and excluding HER2 2+ cases from HER2 analyses

Table 5: Agreement between three biomarker IHC-based and intrinsic subtypes: impact of using a 10% versus 1% threshold to define estrogen receptor (ER) and progesterone receptor (PR) positivity

	IHC subtype		PAM50 Intrinsic subtype							
1% ER positivity threshold		**IHC definition**	**Luminal A**	**Luminal B**	**HER2-enriched**	**Basal-like**	**Total**	**Sensitivity**	**Specificity**	**Accuracy**
	Luminal A	ER+, PR>20%, HER2-	122 (76)	49 (57)	10 (14)	8 (6)	189	76%	77%	77%
	Luminal B	ER+, PR≥1%, HER2+ *or* ER+, PR≤20%, HER2-	38 (24)	36 (42)	30 (43)	28 (21)	132	42%	73%	67%
	HER2-enriched	ER-, HER2+	0 (0)	0 (0)	19 (27)	5 (4)	24	27%	99%	86%
	Basal-like	ER-, PR<1%, HER2-	0 (0)	1 (1)	11 (16)	90 (69)	102	69%	96%	88%
	Total		160	86	70	131	447			
10% ER positivity threshold		**IHC definition**	**Luminal A**	**Luminal B**	**HER2-enriched**	**Basal-like**	**Total**	**Sensitivity**	**Specificity**	**Accuracy**
	Luminal A	ER+, PR>20%, HER2-	122 (76)	49 (57)	10 (14)	8 (6)	189	76%	77%	77%
	Luminal B	ER+, PR≥10%, HER2+ *or* ER+, PR≤20%, HER2-	37 (23)	34 (40)	18 (26)	5 (4)	94	40%	84%	75%
	HER2-enriched	ER-, HER2+	1 (1)	0 (0)	26 (37)	6 (5)	33	37%	98%	89%
	Basal-like	ER-, PR<10%, HER2-	0 (0)	3 (3)	16 (23)	112 (86)	131	86%	94%	91%
	Total		160	86	70	131	447			

Figure legend

Figure 1: Immunohistochemical staining and automated scoring of estrogen receptor (ER), progesterone receptor (PR), and HER2 on tissue microarrays.

Legend: Representative slides staining positive for ER (A), PR (D), and HER2 (G) are shown alongside higher magnification insets from these same cores (B&C, E&F and H&I, respectively). For hormone receptors, negative nuclei are highlighted in blue, low positive in yellow, medium positive in orange and strongly positive in red (C for ER; F for PR). The nuclear algorithm was trained to recognize epithelial cell and to exclude stromal cell regions (shown in gray in C, F) for hormone receptor analyses. The algorithm used for HER2 analysis evaluated DAB staining intensity in membrane regions (yellow represents 1+, orange represents 2+, and red represents 3+ membrane intensity), and black lines highlight HER2 negative cells (I). Reduction in the number of stromal cells analyzed is a built-in function of the membrane algorithm (see non-segmented cells in the stromal area of Panel I). Scale bar (for core images) = 300 μm, scale bar (for magnified images) = 100 μm.

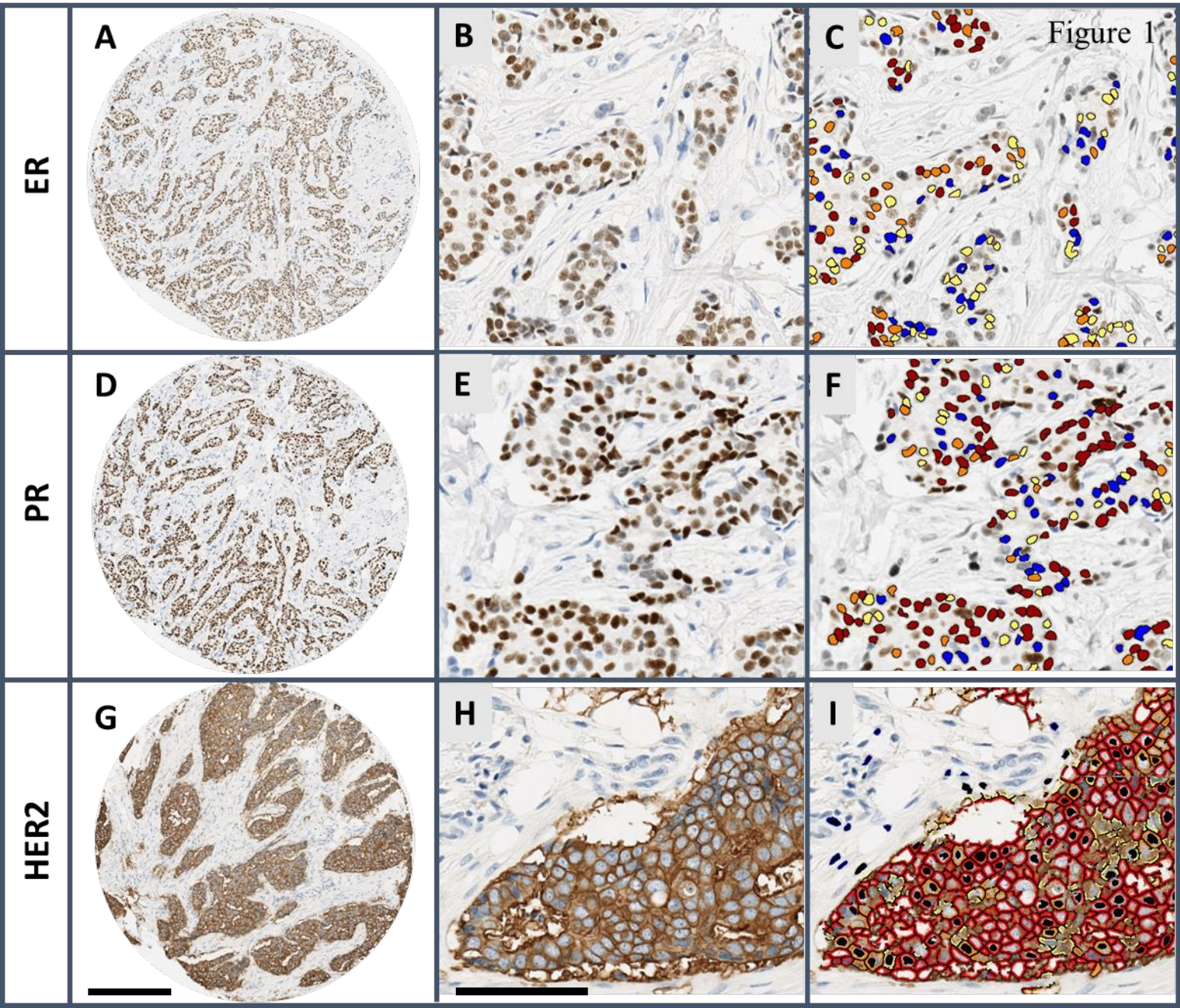

Figure 1