Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue

Steven J Potts¹, Joseph S Krueger¹, Nicholas D Landis¹, David A Eberhard², G David Young¹, Steven C Schmechel³ and Holger Lange¹

Quantitative clinical measurement of heterogeneity in immunohistochemistry staining would be useful in evaluating patient therapeutic response and in identifying underlying issues in histopathology laboratory quality control. A heterogeneity scoring approach (HetMap) was designed to visualize a individual patient's immunohistochemistry heterogeneity in the context of a patient population. HER2 semiquantitative analysis was combined with ecology diversity statistics to evaluate cell-level heterogeneity (consistency of protein expression within neighboring cells in a tumor nest) and tumor-level heterogeneity (differences of protein expression across a tumor as represented by a tissue section). This approach was evaluated on HER2 immunohistochemistry-stained breast cancer samples using 200 specimens across two different laboratories with three pathologists per laboratory, each outlining regions of tumor for scoring by automatic cell-based image analysis. HetMap was evaluated using three different scoring schemes: HER2 scoring according to American Society of Clinical Oncology and College of American Pathologists (ASCO/CAP) guidelines, H-score, and a new continuous HER2 score (HER2_{cont}). Two definitions of heterogeneity, cell-level and tumor-level, provided useful independent measures of heterogeneity. Cases where pathologists had disagreement over reads in the area of clinical importance (+1 and +2) had statistically significantly higher levels of tumor-level heterogeneity. Cell-level heterogeneity, reported either as an average or the maximum area of heterogeneity across a slide, had low levels of dependency on the pathologist choice of region, while tumor-level heterogeneity measurements had more dependence on the pathologist choice of regions. HetMap is a measure of heterogeneity, by which pathologists, oncologists, and drug development organizations can view cell-level and tumor-level heterogeneity for a patient for a given marker in the context of an entire patient cohort. Heterogeneity analysis can be used to identify tumors with differing degrees of heterogeneity, or to highlight slides that should be rechecked for QC issues. Tumor heterogeneity plays a significant role in discordant reads between pathologists.

Laboratory Investigation (2012) **92**, 1342–1357; doi:10.1038/labinvest.2012.91; published online 16 July 2012

KEYWORDS: breast cancer; digital pathology; HER2; immunohistochemistry; pathology; tumor heterogeneity

The companion diagnostic approach seeks to dictate therapeutic strategy based on a molecular description of a patient's disease, with drugs targeting the HER2 receptor one of the most widely adopted examples. There are well-established guidelines for selecting patients for anti-HER2 adjuvant therapies in breast cancer treatment. However, even with patient selection, many trastuzumab-treated patients do not benefit from therapy, as their disease progresses or becomes recurrent. For example, about 1/3 of breast cancer patients given Herceptin fail to respond (*de novo* resistance), and

about 1/5 of the responsive patients become refractory (acquired resistance).¹ The proportion of patients who are not responsive to therapy, even with the inclusion of a companion diagnostic to predict patient response, indicates that the current approaches to treatment strategy and patient selection may not be as robust as possible.

The current HER2 immunohistochemistry (IHC) score methodology does not account for heterogeneity. Since 2007, the American Society of Clinical Oncology and College of American Pathologists (ASCO/CAP) have recommended

¹Flagship Biosciences, Westminster, CO, USA; ²Department of Pathology and Lineberger Comprehensive Cancer Center, University of North Carolina Chapel Hill, NC, USA and ³Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA

Correspondence: Dr SJ Potts, PhD, Flagship Biosciences, 10955 Westmoor Dr., Suite 400, Westminster, CO 80021, USA.

E-mail: steve@flagshipbio.com

Received 30 October 2011; revised 24 March 2012; accepted 2 April 2012

specific guidelines for HER2 scoring.² These guidelines call for a consistent process of sample preparation and staining (IHC) or hybridization (FISH) approaches, as well as score reporting. The ASCO guidelines also suggest using the terminologies 'positive' (3+), 'equivocal' (2+), or 'negative' (+1 to 0) to define HER2 scoring. According to the immunohistochemical (IHC) scoring methodology, the difference between an 1+ and 2+ score is a description of a 'faint' (0/1+) compared to a 'weak-to-moderate' (2+) membrane staining in more than 10% of the tumor cells. In contrast, a 3+ score is described as a uniform intense membrane staining of >30% of tumor cells. Thus, this widely used scoring approach is semiquantitative, as it relies on a threshold percentage of positive cells to determine the score. Importantly, this overall score does not include any additional information about the percentages of tumor cells that score beyond the threshold levels. The ASCO/CAP guidelines for HER2 FISH scoring also relies on a stratified HER 2/CEP17 ratio for the score.

This lack of information about variability within the tumor, or between tumors with the same score, blinds clinicians to a potential readout that could represent a biology responsible for non-effective responses to therapy. It is intuitive that differential cell populations within or between tumors could contribute to clinical refractoriness to therapy and thereby affect patient outcomes. All potential factors within individual patients that contribute to a lack of response are not known, but cancer biologists have long hypothesized that such disparate populations within the tumor can be selected for outgrowth and emerge as a resistant tumor. This concept of tumor heterogeneity leading to drug resistance was debated as early in the 1950s as the 'Greenstein Hypothesis', and has become part of cancer biology doctrine.³ In more recent times, as more targeted therapies are being developed, the issue of tumor heterogeneity has re-emerged as a factor significant to clinical strategy. Thus, there is a need for clinical evaluation of tumor heterogeneity that is aligned with the emerging understanding of cancer biology.

Studies of intratumoral heterogeneity from the same site demonstrate that heterogeneity can affect prognosis in 2+ scored tissues.^{2,4} Another study found 16% of 3+ score cases exhibiting tumor heterogeneity.⁵ A recent case⁶ documented the personal significance of tumor heterogeneity, where a patient with invasive breast carcinoma demonstrated HER2 gene amplification on core biopsy, but relapsed while on adjuvant trastuzumab therapy after mastectomy, dying 15 months after diagnosis. Often, metastases harvested at autopsy demonstrated no evidence of HER2 gene amplification, but retrospective examination of the carcinoma in the patient's mastectomy specimen revealed only focal HER2 amplification within the tumor, localized to the region of the prior core biopsy site, highlighting the importance of both adequate sampling and awareness of heterogeneity issues. Another case was noted⁷ where a patient with breast cancer had areas of the tumor that were

3+ positive and negative for HER2/neu by IHC, adjacent to each other. These cases represent an underlying biology of tumor heterogeneity, which contributes to the clinical outcome.

The assessment of HER2 protein expression status in breast cancer provides a useful working example of tumor heterogeneity for future biomarker studies. There are substantial biological and clinical implications of intratumor clonal heterogeneity.^{8,9} This heterogeneity may reside within a single tumor (intratumoral), or between tumors at different sites (intertumoral). Consequently, researchers have attempted to identify the levels of clinically observed heterogeneity in multiple studies of HER2/neu in breast carcinoma, the results of which are summarized in Table 1. Eight different studies of HER2 heterogeneity between primary breast tumor and metastasis demonstrate the low discordance rates between these: 0 and 13%, with the majority of studies under 5% discordant. Thus, determining the disparity between primary tumor and metastases may not be of high clinical priority. However, one recent study found discordance rates of 14% between core needle and excisional biopsies, suggesting that tumor heterogeneity could contribute to misclassification utilizing needle biopsies.¹⁰ The ASCO/CAP guidelines define HER2 genetic heterogeneity in FISH testing as >5%, and noted that the incidence of intratumor heterogeneity by this definition ranged in the literature from 5 to 30%.¹¹

Accordingly, the ability to measure tumor heterogeneity may assist clinicians in verifying the predictive value of the HER2 score. It is critically important that the profession begin to develop improved approaches of reporting heterogeneity in samples. In the discipline of stereology, unbiased sampling is obtained by utilizing an entire tissue block, and randomly sampling both the sections and regions within a section to eliminate bias.¹² However, a heterogeneity measurement seeks to start with the entire population, and then sample in an unbiased manner to then determine a representative variation. In addition, in clinical trials, it is difficult and nearly impossible to obtain the blocks required for stereology sampling, so the industry is left with dealing with one or several tissue slides as the specimen from which to obtain heterogeneity assessments.

As pathology evolves into a more digital and quantitative discipline, the challenge of quantifying tumor heterogeneity comes more clearly into focus. Whole slide imaging and quantification techniques for the evaluation of IHC biomarkers facilitate an approach for measuring tumor heterogeneity. The ability to distinguish and score individual cells across the whole tissue provides sufficient content to assess reliably diversity of a biomarker within the sample. Combined with a mathematical approach to describe a measure of variation within the sample, a heterogeneity index can be created. In this report, a novel, functional approach is described that assigns a numerical value to HER2 score diversity within a tumor sample, and thus serves to quantify

Table 1 Studies of tumor heterogeneity in HER2/neu

Tumor type and stage	Analytical approach and tissue extraction technique	Result
<i>Evaluation of heterogeneity on different samples from the same tumor site taken at the same time</i>		
38 Invasive breast carcinomas grade 2 tumors evenly split on FISH gene amplification (4)	Evaluated IHC and FISH on additional slides of the same tumor	For the 2+ cases with no FISH gene amplification, 72% had a 1+ IHC score on at least one additional slide, 22% remained 2+ and 6% had one slide scored 1+. For the 2+ cases with FISH gene amplification, 55% had a 3+ IHC score on at least one additional slide, 40% remained 2+ in IHC, and 5% had a slide scored 1+. The authors noted that significant intratumoral heterogeneity accounts for many breast cancers with 2+ HER2 protein expression
44 Breast carcinomas (all grades) and 5 normal breast tissues (10)	1-mm TMA cores in triplicate, evaluating seven proteins by IHC: HER2, ER, PR, E-cad, EGFR, p53, and MIB-1	Intratumoral heterogeneity was seen with ER, PR, HER2, p53, and MIB-1, but not with E-cad and EGFR. Results indicate that core needle biopsies are problematic as indicators of status for the entire tumor
21 Breast cancer tumors (11)	HER2, topoisomerase II α , c-myc, and cyclinD1 were evaluated with q-PCR in macroscopically and microscopically separate areas of individual tumors	HER2 mRNA expression by q-PCR had much lower levels of heterogeneity than the other genes, with heterogeneity occurring in 36% of amplified cases. C-myc and cyclinD1 exhibited heterogeneity in 100% and 83% of cases, respectively
48 Cases of grade 3 breast cancer were analyzed (5)	Multiple sites of morphologically similar tumor were analyzed with HER2 FISH	Intratumoral heterogeneity for HER2/neu gene amplification was demonstrated in only 5 (16%) of 31 cases, where morphologically similar areas of a single tumor were analyzed
<i>Heterogeneity between needle core biopsy and excisional biopsy</i>		
100 Patients with needle core biopsies and subsequent excisional biopsy samples (12)	HER2 FISH and IHC	The concordance rate between FISH results determined on the needle core biopsy and subsequent excisional biopsy of the same tumor was 86%. Of the 15 patients who received neoadjuvant chemotherapy, 93 and 87% had no change in HER2/neu status as determined by IHC or FISH, respectively, in the excisional biopsy specimen when compared with that determined on the prior core biopsy sample
<i>Frequency of heterogeneity in a test population, determined by FISH</i>		
Single institution results of all testing in calendar year 2006, total of 742 consecutive cases of breast carcinoma (13)	Reported on clinical results as run by CAP guidelines	Genotypic heterogeneity, defined as >5% but <50% of the tumor cells demonstrating HER2 gene amplification, was observed in 5% (40/7242) of the cases
<i>Heterogeneity comparison between DCIS and cancer</i>		
Multiple foci from 23 breast tumors with DCIS only and 20 cases with synchronous DCIS and infiltrating cancer (14)	Multiple foci extracted and microsatellite markers by PCR for allelic losses in individual foci for loci in chromosomes 6q, 9p, 11q, 13q, 16q, 17q, and 17p	Patterns of allelic losses were generally conserved in the synchronous infiltrating tumors, supporting the paradigm that infiltrating tumors are clonally derived from the <i>in situ</i> lesions. However, in 8 (40%) of the 20 DCIS cases with invasive cancer, heterogeneous patterns of allelic loss at one or more loci were observed
<i>Concordance between primary site and metastasis</i>		
44 Breast cancer patients with asynchronous metastasis or recurrence (15)	Samples from primary breast cancers and metastatic lesions were analyzed for p53, ER, PR, and HER2 IHC and HER2 FISH	Discordance rates between primary and secondary tumor were 4.5% for HER2 with IHC, with FISH results consistent with IHC

Table 1 Continued

Tumor type and stage	Analytical approach and tissue extraction technique	Result
<i>Concordance between primary site and metastasis</i>		
21 Breast cancer patients with metastasis (16)	Samples from primary breast cancers and metastatic lesions were analyzed with IHC for HER2, p53, ER, and PR	Discordance rate between primary and secondary tumors was 0% for HER2 and p53. Expression levels in breast cancer cells were almost unchanged as the disease progressed, regardless of hormone receptor status
47 Breast cancer patients with metastasis along with literature review of other similar studies (17)	Samples from primary breast cancers and lymph node metastatic lesions were analyzed with IHC and CISH/FISH for HER2	No cases of drastic changes in HER2 expression between primary and lymph node metastasis were reported. Authors conclude that breast cancer lymph node metastases generally overexpress HER2 to the same extent as the corresponding primary, including distant metastases
58 Breast cancer patients with metastasis (18)	Samples from primary breast cancers and metastatic lesions were analyzed by IHC and FISH for HER2	Discordance rate between primary and metastatic tumors was 14% (8 of 58 patients), with the majority (7) positive in metastasis and negative in primary. FISH results were concordant with IHC for the data set
789 Breast cancer patients with metastasis (19)	Samples from primary breast cancers and recurrent tumors were analyzed by IHC for ER, PR, and HER2, and by FISH for HER2	Discordance rate for ER, PR, and HER2 was 18.4%, 40.3%, and 13.6%, respectively. Patients with concordance have significantly better post-recurrence survival than discordant cases
205 Breast cancer cases from 20 institutions with matching primary and recurrent tumor (20)	Samples from primary and recurrent were analyzed for ER, PR, and HER2 by IHC	Discordance rates for ER, PR, and HER2 were 10.2%, 24.8%, and 2.9%, respectively, with no significant difference in locoregional or distant recurrence. The switch in receptor status led to a change in the subsequent treatment plan for 17.5% of the patients
107 Patients with primary breast cancer and at least one distant metastatic lesion (21)	HER2 levels were analyzed by IHC and FISH	Discordance rate of 6% with IHC, with all six discordant cases showing greater HER2 overexpression in the metastatic tumor. By FISH the discordant rate was 5%, and the discordant cases were split between under- and overexpression in the metastatic tumor
Breast cancer primary sites and matched metastatic lymph nodes (22)	HER2, p53, bcl-2, topoisomerase II α , HSP27, and HSP70 were evaluated by IHC	Disconcordance rates were 2% for HER2, 6% for p53, 15% for bcl-2, 19% for topoisomerase II α , 24% for HSP27, and 30% for HSP70
<i>Heterogeneity before and after treatment</i>		
39 Patients with locally advanced breast cancers who received neoadjuvant chemotherapy and 60 breast cancer patients who did not receive neoadjuvant chemotherapy (23)	IHC for HER2 was performed on paraffin sections of the core biopsy before treatment and the excised specimen following chemotherapy	HER2 IHC scores decreased in 28.5% (15/39) of patients receiving neoadjuvant chemotherapy compared to 11.7% (7/60) of patients in the control ($P < 0.013$). HER2/neu IHC status changed from strongly positive to negative (3+ to 0) in 5 of 39 (12.5%) in the study group and in 2 of 60 (3.3%) in the control group ($P = 0.104$)

heterogeneity. This output can be included with other digital pathology-based measurements of IHC biomarkers to provide a more contextual value to the numerical score. Two definitions are introduced to further assist with describing heterogeneity cell-level and tumor-level heterogeneity (Figure 1). Cell-level heterogeneity (Het_{cell}) is the variability of cells within a nest of tumor, and tumor-level

heterogeneity (Het_{tumor}) is the variability of nests of cells across an entire tumor. There is only one score per slide for Het_{tumor}, but as each nest or sampled region in a tumor has its own Het_{cell} score, it is challenging to combine these into a single measure for a given slide. Thus, several approaches are examined to aggregate measures of cell-level heterogeneity across a slide.

Figure 1 Definitions of cell-level (above) and tumor-level heterogeneity (below). Slide-level heterogeneity is a sampling substitute for tumor-level heterogeneity. The below figure also illustrates some contributions of anatomic heterogeneity, as parts of the lesser stained areas are ductal carcinoma *in situ* (DCIS).

Numerical Indices of Tumor Cell Diversity

Diversity measurement is a well-established field in the ecological sciences, and numerous approaches to quantifying the variability of species have been utilized in this discipline. Ecologists will describe diversity in terms of richness and evenness, and each can be ranked differently depending on the weighting of these concepts. For example, one area might have only two species, each covering half the area. The second area might have six different species, with one dominant species covering 95% of the area, and the other five each only covering 1%. Defined in terms of richness, the second area with eight different species would be considered more diverse. Defined in terms of evenness of distribution, the first area would be more diverse as it avoids having one type dominating over all others. Two commonly used diversity indices are the Shannon index¹³ and the Simpson index,¹⁴ for measuring plant and animal species diversity. The Shannon index of diversity is defined as:

$$\text{Shannon} = \sum_{i=1}^N p_i \times \ln p_i$$

where N is the number of biological types and p_i the proportional abundance of the i th type. This index, ranging in theory from 0 to infinity, estimates the average uncertainty in predicting to which species type a randomly selected subunit of area belongs. The Simpson index is defined as:

$$\text{Simpson} = \sum_{i=1}^N p_i \times p_i$$

Producing values from 0 to 1, Simpson's index defines the probability that two randomly selected equal-sized subunits of terrain belong to different species. A recent evaluation of tumor heterogeneity pioneered the use of both Shannon and Simpson indices in evaluating 8q24 copy number gain in both CD24+ and CD44+ cell populations in ductal carcinoma *in situ* and invasive regions of tumors.¹⁵ Copy numbers at each of three levels were considered as separate 'species' and the indices applied to deliver a measure of heterogeneity within each sample. Two distinct tumor subtypes of high and low diversity of 8q24 copy number, as measured by the Shannon index, and the group with lower diversity contained fewer samples of HER2+ tumors. There was no difference between diversity of the luminal A tumors and the normal cells, although basal-like tumors tended to have higher diversity scores. In this study, few qualitative differences were seen between Shannon and Simpson indices, although the data set were small. The Shannon index tends to blur distinctions of species richness and evenness, while the Simpson index can be dominated by the most abundant species in the population.

The disadvantage of both Shannon and Simpson indices is that they do not account for taxonomic distance between species. In the world of clinical anatomic pathology, most cells are binned and scored as one of three or four classes. In HER2 scoring methodology, pathologists (or pathologist-trained computer programs) score cells as populations of either 0+, 1+, 2+, or 3+ intensity. Consider two regions: Region A with ten 0+ cells and ten 3+ cells, and region B with ten 1+ cells and ten 2+ cells. Clearly, Region A has a higher level of heterogeneity than Region B, but Shannon and Simpson indices would score these as equal heterogeneity. To overcome this problem, an ecological diversity approach known as Rao's quadratic entropy (QE)¹⁶ was used. A distance matrix is incorporated in the diversity index, where, for example, a difference between a 0+ and 3+ cell would be weighted a '3', and a 1+ to 2+ would be weighted a '1'. When all weights are the same, the scoring schemes tend to be equivalent to those mentioned previously.

The equation is as follows:

$$\text{QE} = \sum_{i>j=1}^N d_{ij} p_i p_j \quad \bar{D} = \begin{array}{c|cccc} & 3+ & 2+ & 1+ & 0+ \\ \hline 3+ & 0 & 1 & 2 & 3 \\ 2+ & 1 & 0 & 1 & 2 \\ 1+ & 2 & 1 & 0 & 1 \\ 0+ & 3 & 2 & 1 & 0 \end{array}$$

where N is the total number of species, p_i and p_j are the proportions of the i th and j th species, respectively, in the sampling unit, and d_{ij} is a member of the symmetric taxonomic distance matrix $\bar{D}$ ($d_{ij} = d_{ji}$ and $d_{ii} = 0$). The values of $\bar{D}$ can be adjusted to match differences in classes in a particularly biological application, and the matrix shown above was utilized in this study for HER2-expressing cells. As an example of the flexibility of this approach, studies in plant ecology¹⁷ have utilized the following matrix, where $d_{ij} = 1$ if both species belong to the same genus, $d_{ij} = 2$ if both species belong to the same family but different genera, $d_{ij} = 3$ if both species belong to the same order but different families, $d_{ij} = 3.5$ if both species belong to the same subclass but different orders, $d_{ij} = 4$ if both species belong to the same class but different subclasses, $d_{ij} = 4.5$ if both species are both angiosperms but are from different classes, and finally $d_{ij} = 5$ otherwise.

The differences in diversity index scores are illustrated in Table 2, where several different distributions of samples in a given region are assumed. The QE weighs distances between species and will generally range from 0 (entirely homogenous population) to 1.5 (split evenly between extremes), although the upper range depends on the distance matrix used. Using a distance matrix also minimizes minor changes between cells classed into difference adjacent categories. Minimizing minor cell classification changes is important, as many researchers have noted the relative nature of IHC, and the difficulties associated with using IHC for quantitative analyses.¹⁸ One can further increase the numeric value between two classes in the distance matrix to make changes from one class to an adjacent class far less important than a change across several

classes (eg from 0+ to 1+ to 2+ to 3+ distance be changed from 0,1,2,3 to 0,1,4,9 respectively). The appropriate distance matrix should be discussed within the context of a particular protein and the therapeutic goals for prognosis decisions.

In this study, a simple approach to heterogeneity is sought that can be utilized in the clinic as anatomic pathology is practiced today, to ensure immediate assistance to improving clinical trials practice. The constraints include: (1) dealing with a single slide; (2) working in brightfield IHC (there are 16 brightfield FDA clearances for protein expression in tissue and none in fluorescence); (3) utilizing image analysis scoring approaches that have already been cleared for use in clinical practice and are familiar to practicing pathologists; and (4) delivering a scoring system that is easily communicated between pathologists and oncologists. A measure of heterogeneity (HetMap) was developed that incorporates both cell- and tumor-level heterogeneity measures. The approach was evaluated on HER2 IHC-stained breast cancer samples, using 200 specimens across two different laboratories, with three pathologists at each laboratory outlining 10–25 regions of tumor for scoring by automatic image analysis.

MATERIALS AND METHODS

Two slide sets prepared and scored by two different clinical laboratories were used for this study. One slide set of 100 breast carcinomas was selected with an equal distribution of slides scored from 0+ to 3+, and a second slide set of 100 breast carcinomas was taken from routine operation with a distribution of slides representative of the target population. The tissues were formalin-fixed, paraffin-embedded breast tissue specimens immunohistochemically stained using Dako *in vitro* diagnostic FDA-approved HerceptTest (Dako, Carpinteria, CA, USA). All slides were scanned on an Aperio ScanScope, and three board-certified pathologists for each slide set manually drew between 10 and 20 regions of interest of tumor on the slides using the Aperio ImageScope interface. The pathologists were asked to draw representative regions of tumor on each slide for routine HER2 scoring using automatic image analysis. A total of 8549 tumor regions on 100 slides were drawn electronically by three pathologists on the first slide set and 6002 tumor regions on 100 slides were drawn by three other pathologists on the second slide set. The Aperio HER2 membrane algorithm was adjusted and run on these regions of interest to identify cells and classify them as 0+, 1+, 2+, or 3+ staining (<http://www.aperio.com>). The algorithm was adjusted by consensus of the pathologists before the study on control slides, and then used with a fixed parameter set for each of the slide sets. HER2 IHC protein expression status was classified per cell following ASCO/CAP guidelines. Data from the automated image analysis scoring of cells for each region were exported to tab-delimited file format and then analyzed with R.¹⁹

Table 2 Example diversity indices and their scores for various hypothetical regions

3+ Cells	2+ Cells	1+ Cells	0+ Cells	Shannon	Simpson	QE
0	0	0	100	0.00	0.00	0.00
0	0	1	99	0.06	0.02	0.02
99	1	0	0	0.06	0.02	0.02
1	0	0	99	0.06	0.02	0.06
3	3	3	91	0.40	0.17	0.33
0	0	9	91	0.30	0.17	0.16
9	0	0	91	0.30	0.17	0.49
25	25	25	25	1.39	0.76	1.25
50	0	0	50	0.69	0.51	1.50
90	0	0	10	0.33	0.18	0.54
0	50	50	0	0.69	0.51	0.50
50	0	50	0	0.69	0.51	1.00
49	1	49	1	0.79	0.52	1.02
4900	100	4900	100	0.79	0.52	1.02

HER2 Scoring

In this study, two different scoring schemes for HER2 were compared to the classic ASCO/CAP scoring methodology. All scores were calculated for each region and for each slide (sum of all cells in all regions of a slide).

ASCO/CAP HER2 Score Determination

HER2 scoring was calculated according to 2010 CLIA/CAP guidelines (2). Regions containing all cells with no staining or if cell membrane staining is observed in <10% of the tumor cells received a score of 0 ('negative'). Regions containing more than 10% of cells with a faint perceptible membrane staining received a score of 1+ ('negative'). Regions containing more than 10% of cells with weak-to-moderate complete membrane staining received a score of 2+ ('weakly positive'). Regions containing more than 30% of cells with a strong complete membrane staining received a score of 3+ ('positive').

H-Score Determination

A classic H-score was calculated as follows:

$$\text{H-score} = [(3 + \% \text{cells}) \times 3 + (2 + \% \text{cells}) \times 2 + (1 + \% \text{cells}) \times 1]$$

Continuous HER2 Score Determination

A continuous HER2 score (HER2_{cont}) was developed and then computed the value for each region and each slide (sum of all cells across all regions). The HER2_{cont} provides the same scoring report structure as the classic HER2 score, by rounding the HER2_{cont} score graduation to the nearest integer ([0.0, ..., 0.5): 0, [0.5, ..., 1.5): 1+, [1.5, ..., 2.5): 2, and [2.5, ..., 3.0]: 3+). To provide the threshold graduations for the HER2 score, two components were used for measurement of the HER2 score, which were intuitive to pathologists:

(a) The percentage of cells that is responsible for the actual HER2 score (percentage of 3+, 2+, and 1+ cells when the HER2 score is 1; percentage of 3+ and 2+ cells when the HER2 score is a 2+; and percentage of 3+ cells when the HER2 score is a 3+) contributes to the HER2_{cont} score. As the percentage of those cells increases beyond the critical threshold of the actual score, so does the HER2_{cont} in a linear manner from where the critical threshold was passed.

(b) The percentage of cells that is critical for the HER2 scoring to the next higher score (percentage of 3+, 2+, and 1+ cells, when the HER2 score is 0; percentage of 3+ and 2+ cells, when the HER2 score is a 1+; and percentage of 3+ cells, when the HER2 score is a 2+) also contributes to the HER2_{cont} score. As the percentage of those cells increases to reach the critical threshold to the next higher score, so does the HER2_{cont} in a linear manner (0%: -0.5; 1%: -0.4; 2%: -0.3; 3%: -0.2; 4%: -0.1; 5%: 0.0; 6%: 0.1; 7%: 0.2; 8%:

0.3; 9%: 0.4; and 10%: 0.5 = critical threshold for next higher score).

The two measurement components of the HER2_{cont} score complement each other as the HER2 score moves from one score to the next higher score. Because an HER2 score is assigned to each individual cell, histogram data can be generated, which tallies the quantum (0, 1+, 2+, 3+) scores for each region. The HER2_{cont} scoring approach examines the components of this histogram and assigns a threshold value based on a weighted mean of the data values of this histogram. Thus, this score fully represents the two components of data captured in a histogram: (1) the percent of total cells with each quantum score; and (2) consideration of the population profile of these cells. For example, a classic HER2 score of 2+ can be represented by an HER2_{cont} score ranging from 1.5 to 2.4. The specific value of the HER2_{cont} score consists, in part, of the sum of the quantum score multiplied by the number of cells given to each score, and divided by the number of cells. However, the percentages of these cells that contributed most to this mean value are weighted to determine a threshold value of the score. To understand how this approach works, we will use an example of a region that contained 10 cells, of which 4 cells had a score of 3+, 3 cells had a score of 2+, and the remaining 3 cells had a score of 1+. Intuitively, we could deduce the mean of this score as $(3 + 3 + 3 + 3 + 2 + 2 + 2 + 1 + 1 + 1)/10 = 21/10 = 2.1$. However, in actuality, the percentages of cells that contributed to the HER2_{cont} score are weighted to yield the final score. To understand how this weighting process contributes to the final HER2_{cont} score, we can examine the components of a score of 1.5, right at the threshold: If back-calculated, we would find that there are exactly 10% of 2+ or 3+ cells. In contrast, a score of 1.9 must result from either 82% of 2+ or 3+ cells, or 4% of 3+ cells. Finally, a score of 2.4 must result from 9% of 3+ cells, moving the score very close to the next higher 3+ score. Thus, the HER2_{cont} score better captures the effects of intra-regional variability, which contribute to the overall profile of cells within a given summary HER2 score for that region.

The HER2_{cont} was designed to capture the deficiency in a classic HER2 scoring approach, which has been in part addressed by the new HER2 scoring according to the ASCO/CAP guidelines, using the 30% of 3+ cells threshold for a 3+ score. The newer ASCO/CAP threshold approach to defining a 3+ score is based on a more stringent requirement for HER2 positivity, based on the now widely held understanding that the cells with the most HER2 expression are most responsive to trastuzumab. To understand how this decision was critical to predicting trastuzumab sensitivity, we can use the demonstration of how a classic H-scoring approach can result in an identical score for two dissimilar population profiles. For example, if a tumor had 30% of the cells being 3+ with the remaining cells being 0+, the H-score would be 90. If another tumor had 10% of the cells 3+, 60% of the cells 1+, and remaining cells 0+, the

H-score would also be 90. However, these tumors clearly have a different molecular profile. The HER2_{cont} score does not facilitate this scenario, and thus represents a similar approach to the ASCO/CAP HER2 scoring paradigm because the graduation of the continuous 3+ HER2 score only depends on the percentage of 3+ cells. For example, the new HER2 score can be obtained by applying a threshold of 2.61 (equals 30% of 3+ cells) to get a 3+ HER2 score (instead of 2.5 by rounding):

- If $3+ \%cells \geq 10\%$, then $region = 2.5 +$
b. $(3+ \%cells - 10\%)/90\% \times 0.5$
- If $(3+ \%cells + 2+ \%cells) \geq 10\%$, then $region = 1.5 +$
MAXIMUM of
a. $(3+ \%cells)/10\%$
b. $(3+ \%cells + 2+ \%cells - 10\%)/90\% \times 0.5$
- If $(3+ \%cells + 2+ \%cells + 1+ \%cells) \geq 10\%$, then $region = 0.5 +$
MAXIMUM of
a. $(3+ \%cells + 2+ \%cells)/10\%$
b. $(3+ \%cells + 2+ \%cells + 1+ \%cells - 10\%)/90\% \times 0.5$
- Else $region = 0.0 +$
a. $(3+ \%cells + 2+ \%cells + 1+ \%cells)/10\% \times 0.5$

Heterogeneity Scoring

The goal was to develop two measures of heterogeneity, one representing cell-level heterogeneity (Het_{cell}) and one representing slide-level heterogeneity (Het_{slide}), as a surrogate for tumor-level heterogeneity (Figure 2). By reporting both the cell- and tumor-level heterogeneity, the pathologist can communicate a view of both microvariation (cell-level) and macrovariation (tumor-level) in the tumor. HetMap is then a graph of the entire patient population, displaying cell-based heterogeneity within individual regions on one axis, and the slide-level heterogeneity on the other axis.

$$\text{HetMap} = \text{Het}_{\text{cell}} \text{ and } \text{Het}_{\text{slide}}$$

where

$$\text{Het}_{\text{cell}} = \sum_{N_{\text{regions}}}^{i=1} \text{QE}_{\text{region}_i} / N_{\text{regions}}$$

$$\text{Het}_{\text{slide}} = \text{QE}_{\text{slide}}$$

To compute Het_{cell}, one first computes the individual quantitative entropy scores for all sampled regions on a slide. Then, these scores need to be combined into a meaningful representation per slide. This could either be done by presenting the average values or the maximum or minimum values. As each approach (average, maximum, minimum) potentially conveys different information content, it is not clear *a priori* as to which is the best approach

Figure 2 Diagrammatic illustration of the HetScore. The Het score is derived from analysis of the HER2 score for each individual cell (HER2_{cell}—color-coded circles represent different HER2 values assigned to each cell) or the HER2 score from each individual region (HER2_{region}—the overall score of cells within an outlined and shaded region). The quadratic entropy (QE) value of an outlined region (QE_{region}) is determined from the individual cellular HER2 scores measured within a region. Het_{cell} is defined as the mean of the QE value of all the regions. The QE value of the summary HER2 scores for all regions across the whole slide (QE_{slide}) is defined as Het_{slide}. The QE value ranges from 0 to 1.5, with 0 being entirely homogeneous and 1.5 displaying the maximum heterogeneity possible.

for determining this value. Thus, the values for all three approaches for reporting Het_{cell} were determined, and the advantages and disadvantages of each are discussed. Het_{slide} is defined as QE_{slide}, the quantitative entropy scores of the population of HER2 scores for each sampled region on a slide.

Variability Analysis by Lab and Pathologist

Once calculations were performed on each slide with regions drawn by each pathologist, these values could then be compared to determine the effect of each pathologist selecting his or her own regions, as well as the variability between labs. Each pathologist's annotations of regions on a slide were analyzed independently, and then the same digital slide compared between the three pathologists to analyze dependence on the regions of interest chosen by the pathologists. The HER2 score, H-score, HER2_{cont}, and Het_{cell} were calculated for each region and the HER2 score, H-score, HER2_{cont}, Het_{cell}, and Het_{tumor} were calculated across all regions for all 600 slides (2 labs × 3 pathologists × 100 slides). Across all the regions on a single slide, the mean, standard deviation, maximum, and minimum for regions scores for HER2, H-score, HER2_{cont}, and Het_{cell} were also computed. Once these computations were made, the assessment of the differences between pathologists and their choice of regions could be evaluated. This was done based on computing the standard deviation and coefficients of variation (CV) (standard deviation/mean) for the values computed by each pathologist's analysis.

RESULTS

Comparison of HER2 Scoring Approaches

As might be anticipated, the H-scores of the regions were nonlinear with the HER2 scores, with the HER2_{cont} score

being a closer representation of the CAP/CLIA guidelines (Figure 3). The nonlinearity is due to the heavier weighting of higher intensity staining over lower intensity staining to calculate the score, which disproportionately biases higher HER2 scored slides towards having a higher H score. For example, a tumor with 30% of 1+ cells receives the same H-score as a tumor with 10% of 3+ cells, but the first one would be a 1+ HER2 score and the second a 3+ HER2 score. These two slides, with the same H-score, have a quite different meaning to a pathologist and clinician determining the therapeutic strategy. Thus, the H-score, which was designed as a standard scoring scheme to provide continuous scores, is not well suited for the scoring HER2-stained slides.

In contrast to the H-score, the HER2_{cont} score remains linear with HER2 scores determined using the CAP/CLIA guidelines. Linearity of HER2_{cont} with the classic HER2 score is preserved because of the use of two components in calculating the score: the percentage of cells that is responsible for the actual HER2 score, and the percentage of cells that is critical for the HER2 scoring to the next higher score. The first component of the measurement confirms that the critical threshold is passed, whereas the second component measurement demonstrates how the different cell populations are moving the value towards the next score level. The H-score attempts to accomplish this by a score-weighted value. In contrast, the second component HER2_{cont} score assigns an unweighted, score-independent value to the effect of cell variability on the HER2 score. As such, the HER2_{cont} score remains linear with the ASCO/CAP-derived score.

Relationship of Heterogeneity with HER2 Score

The QE_{region} generally increases with the classic HER2 and HER2_{cont} scores (Figure 4). This observation is consistent with the attributes of tumors with increasing HER2 scores. The increasing variability observed in tumors with 3+ scores is a result of increasing complexity in the tumor: a 1+ or 0 tumor has little to no staining and thus is very homogeneous; and a 2+ or 3+ tumor has intensely staining cells. These more diverse components inherently permit more variability than a 0 or 1+ tumor.

The lack of information about what factors contributed to the determination of the HER2 score using the classic method results in a less complex plot comparing QE_{region} vs classic HER2 than the comparison of the QE_{region} vs HER2_{cont}. The plot comparing the QE_{region} vs HER2_{cont} score demonstrates a similar pattern, with QE scores increasing as the HER2_{cont} score increases towards a threshold. However, non-intuitively, the QE score decreases dramatically at the very highly scored 3+ tumors. This is due to the maximum staining intensity of any single cell being 3+. As there is no higher scoring population (+4), all the high staining cells are by definition 3+. Thus, the QE_{region} is actually decreased as the percentage of +3 intensity cells increases; as the amount of tumor with highly staining cell types increases, the variability decreases. At the extreme values of HER2_{cont}, which have the most homogeneity (highest and lowest percent of +3 intensity cells), QE_{region} would be expected to actually decrease, as is observed.

Three approaches to measuring an aggregate cell-heterogeneity score, the average, minimum, and maximum QE_{region} scores, found on a whole slide are shown in Figure 4.

Figure 3 (a) H-score and (b) HER2_{cont} vs HER2 scores. Left column: Lab A, 8500 tumor regions across 100 slides; right column: Lab B, 6000 tumor regions across 100 slides.

Figure 4 Cell-level heterogeneity vs (a) HER2 and (b) HER2_{cont} scores. Top set of four graphs measures cell-level heterogeneity as the average (QE_{cell}) values and bottom set of four graphs as the maximum (QE_{cell}) values.

Utilizing the maximum QE_{region} demonstrates the strongest correlation between HER2 score or HER2_{cont} and QE_{region}. This again is, in part, due to a larger dynamic range of HER2 scores enabled by an increasing complexity with the presence of + 3 stained cells within the region. Utilizing the maximum score, by definition, captures the maximum entropy observed, enabling clearer visualization. As expected, using the minimum value has the opposite effect. As might be anticipated, the range of values for the average was lower than that of the maximum score, as the use of the average smoothed the variability within the sample by decreasing the dynamic range of the QE score reported. Thus, utilizing the maximum QE_{region} value from a single whole slide to serve as a reporting measure for cell heterogeneity because of it having the greatest dynamic range, and potentially representing the most relevant aspects of cell biology leading to therapy resistance is proposed.

Figure 5 displays these three different cell-level heterogeneity (QE_{region}) approaches vs the slide heterogeneity (Het_{slide}) scores for each lab. As the data demonstrate, there is almost no correlation (all correlation values are $R^2 < 0.07$) between cell- and tumor-level heterogeneity (QE_{region} vs Het_{slide}). This is true whether the minimum, maximum, and average QE_{region} is used. This indicates that the variability

between regions within a tumor is independent of variability between neighboring cells. This concept is consistent with the accepted model of cell-autonomous, genetic regulation of HER2 expression.

A final representation of ‘unblinded’ data in Figure 6 displays a patient cohort visualization approach called HetMap. HetMap includes pseudocoloring to identify the HER2 score for each slide. HetMap is thus a graphical representation of the type of heterogeneity as it relates to a HER2 score. As demonstrated in Figures 4, the QE increases with the HER2 score, due to increased complexity of the sample with more 3 + staining cells. This representation shows a clear stratification of the HER2 score and cell heterogeneity is seen, as tumors with 3 + and 2 + scores demonstrate significantly more cell heterogeneity than tumors with 0 and 1 + scores. Again, there is no relationship between HER2 score and Het_{slide}, consistent with the data shown in Figure 5. However, the inclusion of the HER2 score in this graph indicates that there are specific subsets of patients within each score who either have a large or small degree of heterogeneity across the whole tumor. The current understanding of tumor heterogeneity on clinical response suggests that this measure has potential clinical value, as discussed later.

Figure 5 Tumor-level heterogeneity (Het_{slide}) is not correlated with any of the three measures of cell-level heterogeneity: (a) minimum, (b) maximum, and (c) average.

Figure 6 An illustration of HetMap showing each slide in the context of the patient population, including the IHC score, the tumor-level heterogeneity, and one of three measures of cell-level heterogeneity—(a) minimum, (b) maximum, and (c) average.

Effect of Pathologist Choice of Regions on the HetMap Score

The standard deviation, average, and CV (standard deviation/average $\times 100$) values were determined for the three sets of regions drawn by the three pathologists for each slide. These values for slide-level heterogeneity (Het_{slide}), the HER2 continuous score ($HER2_{cont}$), and the three approaches to cell-level heterogeneity, $Max(QE_{region})$, $Min(QE_{region})$,

and $Ave(QE_{region})$ are displayed for the two laboratories in Table 3. The variation of HER2 as a continuous score between pathologists drawing their own regions was low in both labs (8 and 9% CVs), validating this scoring methodology. This was lower than the ASCO/CAP recommended concordance rate of 95%, but far better than published concordance rates for HER2 IHC tests around 20%.^{20,21}

Table 3 Comparison of the variability introduced by the pathologist selecting regions of interest

	Slide-level heterogeneity	HER2 continuous	Cell-level heterogeneity (three measures)		
	Het _{slide}	HER2 _{cont}	Max (QE _{region})	Min (QE _{region})	Ave (QE _{region})
Lab A, s.d.	0.13	0.10	0.10	0.08	0.06
Lab B, s.d.	0.09	0.11	0.09	0.07	0.06
Lab A, average	0.41	1.2	0.64	0.13	0.35
Lab B, average	0.31	1.1	0.57	0.15	0.35
Lab A, CV	30%	8%	16%	60%	17%
Lab B, CV	25%	9%	15%	45%	16%

Figure 7 Tumor heterogeneity plays a role in pathologist discordant reading. The concordant slides (1 + / 1 +) had significantly lower tumor heterogeneity ($P < 0.01$) than 0 + / 1 + or 1 + / 2 + discordant slides.

The CV values for cell-level heterogeneity (Max(QE_{region}), and the Ave(QE_{region})) ranged higher ((16 and 15% CVs) and (17 and 16% CVs), respectively). Measures of cell-level heterogeneity based on a minimum (60 and 45% CVs) were much higher than those based on a maximum. The Min(QE_{region}) determination showed the most disparity in CV values between pathologists, reinforcing the idea that this particular assessment is subject to more noise interference due to the nature of its determination, and should not be used. Tumor-level heterogeneity (Het_{slide}) was far more impacted by pathologists' choice of regions (30 and 25% CVs) than the other measurements. It is likely that the number of regions sampled may not be sufficient to make determinations of tumor-level heterogeneity. Relying on a methodology that samples all the tumor on a slide is becoming possible and practical, and may be required for this type of analysis.

Tumor-Level Heterogeneity Impacts Pathologist Concordance Rates

Figure 7 shows the distribution of tumor-level heterogeneity vs slides that had either concordant or discordant reads

between the three pathologists in a laboratory. Discordant reads between 1 + and 2 + are particularly troubling for clinical operation, as a 1 + will not be reflexed for FISH HER2 testing and investigated further for possible treatment options. There was a significant difference ($P < 0.01$) between the concordant 1 + slides and those discordant either with 0 + or 2 +.

DISCUSSION

Although molecular-targeted therapies such as trastuzumab are based on a strong biological rationale for the target, a significant percent of patients who express the target are refractory to the therapy. Markers of resistance are not intuitive or well established, and often not related to the target itself. For example, ErbB2 mutations that lead to trastuzumab resistance are not found in the clinic, and resistance is a result of amplification of many other genes besides HER2.²² HER2-expressing cells may overcome anti-HER2 therapy, although a variety of other survival or compensatory signaling mechanisms that currently cannot be identified clinically.²³ This is consistent with the well-accepted idea that HER2 + breast tumors are not a single biological system, but are comprised

of several different subtypes of HER2+ and HER2- cells within a tumor. Cancer biologists have long hypothesized that minor populations of cells within a bulk of the tumor are key to understanding mechanisms that underlie resistance to trastuzumab treatment. Thus, a persistent problem that confounds ideas about trastuzumab resistance is the concept of tumor heterogeneity. Therefore, for the continued development of the best strategy for trastuzumab and other therapies, methodologies are needed that can describe heterogeneity in the clinical context in relation to the targeted therapy.

As shown in the literature review summarized in Table 1, there are different ways of defining heterogeneity in the context of HER2 expression. First, there is the most classical measure of heterogeneity between different samples from the same tumor site taken at the same time. This can be within a single sample, or between multiple samples, such as that seen between needle core biopsy and excisional biopsy. Second, heterogeneity can be measured between different subtypes of the same disease within the same patient, such as a comparison of between ductal carcinoma *in situ* and invasive disease, or between primary site and metastasis. Finally, heterogeneity can be measured between different tumor samples within a population. This can also include measures of heterogeneity in a patient population before and after a treatment. Regardless of the type of heterogeneity, its measure is determined by the variability for a biomarker score with an entire population.

In this report, heterogeneity is defined as the measure of the variability samples within a population found in a single breast tumor. The first assumption made with this approach is that the tissue section analyzed is indeed representative of the whole biopsy. The second assumption made is that the biopsy is, in fact, representative of the tumor. Although these assumptions are inherent in and necessary for the current approaches to clinical biomarker evaluation, to date there has been no methodology to address the potential effects of these assumptions on the measurement. To account for the potential effects of heterogeneity on a HER2 IHC score, a methodology must be used that can simultaneously incorporate a measurement of heterogeneity with evaluation of a biomarker score.

Classical IHC scoring approaches, which allow visualization of the tissue, do not sufficiently address heterogeneity. Within a tumor, HER2 heterogeneity may manifest as variability in HER2 expression between neighboring cells, or as variability in HER2 expression between different parts of the tumor section. Normally, pathologists will take an average expression across a slide by selecting several random regions to score and combined the scores from each region for a final score. Although this method scores individual cells and different regions of the tumor, it does not consider either component separately in determining the score. Furthermore, there is no consideration in these assessments for variation within the sample that may have impacted the

score. In this way, the current paradigm of assessing HER2 is not aligned with the fundamental understanding of tumor heterogeneity in cancer biology. With the current approach, potentially valuable information about the variability between areas sampled is lost.

To create a method that included these important aspects of tumor biology, a method of scoring heterogeneity was developed that could be incorporated into existing workflow and be readily implemented in the clinic called HetMap. To be a biologically relevant measure, the measure of heterogeneity needs to capture both cell- and tumor-level variation. To be a valid measurement, it needs to be independent of the pathologist's choice of regions, as well as the laboratory or antibody manufacturer. To be a useful measurement, it needs to be amenable to immediate use in a clinical environment. With this in mind, an approach was developed that meets these requirements, first by developing methods to score HER2, then cell- and tumor-level heterogeneity separately, and finally assembling these components into the graphical output of HetMap.

First, to better capture cell- and tumor-level heterogeneity, a continuous HER2 scoring method was developed. As designed, the HER2_{cont} scoring allows for a better benchmark when observing variability using standard of deviation. In contrast to the classic H-score methodology, the HER2_{cont} score remains linear with the ASCO/CAP HER2 scores. Perhaps most significantly, the HER2_{cont} also ensures that the graduation of the HER2 score remains meaningful to the pathologists, in contrast to the classic HER2 score, which is devoid of information about the factors that determined the score. Thus, the HER2_{cont} score enables better understanding of the variability that contributed to the overall score. Although the HER2_{cont} proved a useful approach for analyzing the data, its use outside of this context is not recommended, given the regulatory difficulty of making any changes to accepted scoring schemes in clinical environments. In this context, the HER2_{cont} could be utilized as a complement to the current reporting approach to HER2 analysis by IHC, to convey meaningful contextual information to the clinician about the reported HER2 score, or to identify near-threshold scores that may require re-evaluation.

To report directly heterogeneity of a biomarker within a tumor, population diversity approaches used in ecology were used. The measure of taxonomical diversity with Rao's QE proved to be a simple and versatile complement to HER2 scoring. In combining the cell-heterogeneity scores for each region into a single score for a slide, three alternative approaches using the minimum, maximum, or average QE_{region} score were presented. The maximum QE_{region} demonstrated the clearest trend between an increase in HER2 score and an increasing QE_{region}. It is hypothesized that this could, in part, due to a larger dynamic range of HER2 scores enabled by an increasing complexity with the presence of +3 stained cells within the region. However, it also suggests that there is a causative relationship between HER2 expression and het-

erogeneity. This could be for several hypothetical reasons pertaining to the complexities of HER2 biology, but a possible simple explanation points to well-known factors that are related to HER2 gene amplification: HER2 expression is regulated by the amplification of a 'hot spot' on chromosome 17q, in a so-called 'firestorm pattern'. As such, HER2 amplification also results in the amplification of multiple neighboring gene loci for other known oncogenes, through both known and unknown mechanisms. For this reason, HER2 expression is clinically well correlated with measures of disease progression, drug resistance, and recurrence.²² As discussed earlier, these same measures are well correlated with tumor heterogeneity. As such, the linear relationship between HER2 score and QE_{region} may, in fact, be causal and is in line with our current understanding of HER2 biology.

Although most useful with the $HER2_{\text{cont}}$ scoring method, the relationship between HER2 score and variability were demonstrated with either HER2 scoring method. One can imagine biological reasons for favoring each approach, depending on the biomarker of interest. Reporting maximum QE_{region} may be most beneficial for biomarkers whose overexpression is predictive, such as HER2. Although less useful for HER2, minimum cell-level QE may be important when examining the downregulation of a biomarker for susceptibility or resistance, like the tumor suppressor gene p53. The average QE_{region} may be most valuable for examining the levels of a biomarker that is more homogeneously expressed, but with varied expression depending on the phenotypic state of the cell. An example of this might be the widely monitored cell adhesion protein E-cadherin. In any case, there must be congruency between the biology of the biomarker and the variability of its expression within a tumor. Ultimately, it will require a comparison to clinical data to determine if the QE output is prognostic or predictive, and which of the three approaches will be most useful.

Throughout all the data, the results were highly reproducible between the two laboratories, even though the two labs had different specimens, different pathologists, and different standardization procedures (although both complying with CLIA/CAP guidelines for HER2 testing). This is encouraging both for clinical practice and for these measures of heterogeneity. These data require further investigation, but the initial findings that the lowest CV values are found with the $HER2_{\text{cont}}$ and the maximum QE_{region} outputs suggest that the pathologists are fairly consistently drawn to areas with higher staining (they may contain the highest HER2 score and maximum heterogeneity regions). This idea is congruent with the long-held observation that there is more continuity between pathologists in HER2 3+ scored slides compared to 2+ stained slides.²⁰ These data also initially suggest that pathologist do not systematically select areas with lower staining (containing perhaps lower levels of heterogeneity), demonstrated as weaker correlations between minimum QE_{region} score and HER2 score, and the increased CV

values seen with these measures. Finally, the data suggest that pathologists tend to more systematically include both low and high levels of staining in the same regions, demonstrated as modest CV values seen with these measures.

To demonstrate a working HetMap model in the widely used HER2 scoring platform, the data generated from these approaches was integrated into a graphical output. Perhaps counter-intuitively, there was little correlation between any of the cell-level HER2 heterogeneity measures and the tumor-level HER2 heterogeneity measures. A possible simple explanation is that HER2 expression is a cell-autonomous event, which is determined primarily by genetic amplification. For this reason, HER2 FISH scores correlate well with HER2 IHC scores, and both methods are acceptable for clinical use. Thus, HER2 expression is not influenced by known factors that affect the tumor cell microenvironment and heterogeneity within the tumors, such as autocrine and paracrine loops, hypoxia, inflammatory cells, and extracellular matrix components. The observation that cell- and slide-level heterogeneity did not correlate with each other, despite the clear correlation between all three approaches to cell-level heterogeneity and HER2 score, supports this idea. In contrast, other biomarkers represented a phenotype more dependent on these physiological/environmental factors, such as the proliferation marker Ki67, may show a correlation between cell- and tumor-level heterogeneity.

The HetMap not only provides a general measure of tumor heterogeneity, it has the potential to identify at-risk patients who may benefit from a different course of therapy. For example, as Figure 6 shows, there are subsets of tumors that are on the far right of the HetMap, which have significant intratumor heterogeneity. Although this cannot be determined from this data set, our current understanding of tumor heterogeneity would suggest that these outliers have a poorer prognosis and may be likely have *de novo* resistance to trastuzumab, despite having HER2 expression. In contrast, tumors on the far left side of HetMap who have low heterogeneity and express HER2 may benefit the most from trastuzumab therapy. Also, a correlation between HER2-expressing tumors with modest heterogeneity, which is indicative of a tumor subset that becomes refractory (acquired resistance), may be found. However, in the absence of patient outcome data, these scenarios can only be speculative. Despite this, the HER2 HetMap ties discrete subgroups of patients who have extreme or little heterogeneity with their HER2 score, thereby facilitating progress toward applying principles of tumor biology into clinical practice.

ASCO/CAP has not yet provided guidance on a potential diagnostic approach to consider HER2 IHC heterogeneity. In contrast, ASCO/CAP has provided specific guidance on assessments of HER2 FISH heterogeneity: These guidelines define 'heterogeneous amplification' as the presence of between 5 and 50% cells with HER2/CEP17 ratios of more than 2.20. A survey of 1329 consecutive breast cancer cases demonstrated that when utilizing the previously used criteria,

23.2% of cases demonstrated heterogeneity, of which 81.6% were not amplified and 15.5% were equivocal by the standard criteria. In contrast, the new ASCO/CAP criteria classified only 6.5% of cases as heterogeneous, of which only 8% were not amplified and 79% were equivocal by standard criteria. Thus, this new approach of defining a 'heterogeneous amplification' patient population better captures patients who truly have amplified HER2 in a non-uniform manner without introducing false positives.²⁴ Similarly, the TEAM (Tamoxifen vs Exemestane Adjuvant Multicentre) pathology study, which included 6461 eligible cases, demonstrated that 33.5% of the cases exhibited FISH heterogeneity defined by the newer ASCO/CAP criteria. These aforementioned studies show that carcinomas with HER2 genetic heterogeneity can still have an overall negative HER2 amplification status, despite still containing a significant number of tumor cells with HER2 gene amplification.

Interestingly, there was no prognostic value found in comparing the 'heterogeneous amplification' population to outcome in this study.²⁵ Although the lack of prognostic value is unexpected, a recent study found that tumors classified as 'heterogeneous' amplification lacked specific clinicopathological characteristics that would normally correlate with prognosis, suggesting that other more important factors may be critical to patient response. However, these same authors found that HER2 genetic heterogeneity according to the ASCO/CAP definition is most often present in breast carcinomas with an equivocal (2+) HER2 score.²⁶ This correlates with our observation that patients who received an equivocal (2+) HER2 score also tended to have the highest heterogeneity as well.

This study is the first to show the impact of tumor-level heterogeneity on pathologist discordant reading. There was no significant difference between cell-level heterogeneity and pathologist discordance, as might be expected. Tumor-level heterogeneity presents a much larger challenge to accurate and precise reading by pathologists, as the location for reading on a slide can impact the scoring. This study relied on pathologists to identify manually regions of interest on the whole slide image, and measured the variability introduced by this human sampling approach. There are several potential approaches to eliminating this sampling bias. One could increase the number of regions, but the time required already for drawing regions manually is beyond economic viability in most laboratories. One could computationally deconvolve each IHC image, displaying only the hematoxylin to the pathologist, to reduce the selection bias towards highly staining DAB regions.²⁷ Alternatively, it may be better to overlay a grid on an digital image of the sample, similar to stereology approaches, and then ask the pathologist to draw the closest 'tumor' nest within a set distance of randomly sampled grid points to be used as the data set. Another improvement could be the use of pattern recognition or special staining followed by pattern recognition on a consecutive section. This approach could remove bias by either using a

central slide without the biomarker of interest on it or having the pathologist select tumor nests with only the hematoxylin-stained and eosin-stained section. This method of aligning serial sections to be treated as identical sections could also increase the sampling size, and could automate the choice of regions. One can take the use of aligned serial sections further, but including not just one IHC marker but several to then provide a classification of cells based on multiple markers. As pattern recognition methods improve and whole slide imaging becomes more standardized, this may facilitate the identification of the optimal heterogeneity approaches.

The HetMap conveys the cell-level, tumor-level, and IHC expression score for each patient in the context of the entire patient population, and further clinical studies to determine its utility are anticipated. Unfortunately for the development of new companion diagnostics, HER2 may actually represent a best-case scenario for assessing heterogeneity. Several of the studies listed in Table 1 also included other biomarkers, and generally consistently found HER2 to have substantially lower heterogeneity. In particular, multiple studies find higher heterogeneity in the estrogen and progesterone markers. One study found HER2 to have far lower levels of heterogeneity than other biomarkers studied, with c-myc and cyclinD1 exhibiting heterogeneity in 83 and 100% of samples, respectively.²⁸ In regards to tumor type, high levels of heterogeneity is well known in gastric carcinoma,²⁹ and one study found rates of intratumoral heterogeneity of KRAS mutations as high as 35–47% in colorectal tumors.³⁰ HER2 expression is generally far more heterogeneous in gastric carcinoma than in breast tumors.³¹ The generally low levels of HER2 heterogeneity between primary and metastatic site cannot be assumed to be seen in other therapeutic targets under development. Given that new therapeutic agents and their corresponding companion diagnostics will present greater heterogeneity challenges than HER2, it only increases the urgency to move anatomic pathology in the direction of measuring and reporting IHC heterogeneity.

DISCLOSURE/CONFLICT OF INTEREST

Steven J Potts, Joseph S Krueger, G David Young, Trevor A Johnson, Holger Lange, Nicholas D Landis, and Frank Voelker are employees of Flagship Biosciences, and David A Eberhard is the Chief Medical Pathologist for Flagship Biosciences. The techniques described in this paper are produced by Flagship Biosciences.

1. Esteva FJ, Yu D, Hung MC, *et al*. Molecular predictors of response to trastuzumab and lapatinib in breast cancer. *Nat Rev Clin Oncol* 2010;7: 98–107.
2. Wolff AC, Hammond ME, Schwartz JN, *et al*. American society of clinical oncology/college of American pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Arch Pathol Lab Med* 2007;131:18–43.
3. Weinhouse S. Discussion of Doctor Greenstein's paper. *Cancer Res* 1956;16:654–657.
4. Lewis JT, Ketterling RP, Halling KC, *et al*. Analysis of intratumoral heterogeneity and amplification status in breast carcinomas with equivocal (2+) HER-2 immunostaining. *Am J Clin Pathol* 2005;124: 273–281.

5. Shin SJ, Hyjek E, Early E, *et al*. Intratumoral heterogeneity of her-2/neu in invasive mammary carcinomas using fluorescence *in-situ* hybridization and tissue microarray. *Int J Surg Pathol* 2006;14:279–284.
6. Wu JM, Halushka MK, Argani P. Intratumoral heterogeneity of HER-2 gene amplification and protein overexpression in breast cancer. *Hum Pathol* 2010;41:914–917.
7. Davila E, Amazon K. The clinical importance of the heterogeneity of HER2 neu. *Case Rep Oncol* 2010;3:268–271.
8. Polyak K, Shipitsin M, Campbell-Marrotta L, *et al*. Breast tumor heterogeneity: causes and consequences. *Breast Cancer Res* 2009;11: S1–S18.
9. Marusyk A, Polyak K. Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta* 2010;1805:105–117.
10. D'Alfonso T, Liu YF, Monni S, *et al*. Accurately assessing her-2/neu status in needle core biopsies of breast cancer patients in the era of neoadjuvant therapy: emerging questions and considerations addressed. *Am J Surg Pathol* 2010;34:575–581.
11. Vance GH, Barry TS, Bloom KJ, *et al*. Genetic heterogeneity in HER2 testing in breast cancer: panel summary and guidelines. *Arch Pathol Lab Med* 2009;133:611–612.
12. Howard V, Reed MG. *Unbiased Stereology*, 2nd edn. Garland Science/BIOS Scientific Publishers: New York, 2004.
13. Shannon CE, Weaver W. *The Mathematical Theory of Communication*. University of Illinois Press: Urbana, IL, 1949.
14. Simpson EH. Measurement of diversity. *Nature* 1949;163:688.
15. Park SY, Gonen M, Kim HJ, *et al*. Cellular and genetic diversity in the progression of *in situ* human breast carcinomas to an invasive phenotype. *J Clin Invest* 2010;120:636–644.
16. Rao CR. Diversity and dissimilarity coefficients: a unified approach. *Theoret Popul Biol* 1982;21:24–43.
17. Pavoine S, Dolédec S. The apportionment of quadratic entropy: a useful alternative for partitioning diversity in ecological data. *Environ Ecol Stat* 2005;12:125–138.
18. McCabe A, Dolled-Filhart M, Camp RL, *et al*. Automated quantitative analysis (AQUA) of *in situ* protein expression, antibody concentration, and prognosis. *J Natl Cancer Inst* 2005;97:1808–1815.
19. Ihaka RG, R. R: a language for data analysis and graphics. *J Comp Graph Stat* 1996;5:299–314.
20. Hameed O, Adams AL, Baker AC, *et al*. Using a higher cutoff for the percentage of HER2+ cells decreases interobserver variability in the interpretation of HER2 immunohistochemical analysis. *Am J Clin Pathol* 2008;130:425–427.
21. Perez EA, Suman VJ, Davidson NE, *et al*. HER2 testing by local, central, and reference laboratories in specimens from the North Central Cancer Treatment Group N9831 intergroup adjuvant trial. *J Clin Oncol* 2006;24:3032–3038.
22. Theillet C. What do we learn from HER2-positive breast cancer genomic profiles? *Breast Cancer Res* 2010;12:107.
23. Esteva FJ, Guo H, Zhang S, *et al*. PTEN, PIK3CA, p-AKT, and p-p70S6K status: association with trastuzumab response and survival in patients with HER2-positive metastatic breast cancer. *Am J Pathol* 2010;177: 1647–1656.
24. Allison KH, Dintzis SM, Schmidt RA. Frequency of HER2 heterogeneity by fluorescence *in situ* hybridization according to CAP expert panel recommendations: time for a new look at how to report heterogeneity. *Am J Clin Pathol* 2011;136:864–871.
25. Bartlett AI, Stanczyński J, Robson T, *et al*. Heterogeneous HER2 gene amplification: impact on patient outcome and a clinically relevant definition. *Am J Clin Pathol* 2011;136:266–274.
26. Ohlschlegel C, Zahel K, Kradolfer D, *et al*. HER2 genetic heterogeneity in breast carcinoma. *J Clin Pathol* 2011;64:1112–1116.
27. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. *Anal Quant Cytol Histol* 2001;23:291–299.
28. Glockner S, Buurman H, Kleeberger W, *et al*. Marked intratumoral heterogeneity of c-myc and cyclinD1 but not of c-erbB2 amplification in breast cancer. *Lab Invest* 2002;82:1419–1426.
29. de Aretxabala X, Yonemura Y, Sugiyama K, *et al*. Gastric cancer heterogeneity. *Cancer* 1989;63:791–798.
30. Baldus SE, Schaefer KL, Engers R, *et al*. Prevalence and heterogeneity of KRAS, BRAF, and PIK3CA mutations in primary colorectal adenocarcinomas and their corresponding metastases. *Clin Cancer Res* 2010;16:790–799.
31. Grabsch H, Sivakumar S, Gray S, *et al*. HER2 expression in gastric cancer: rare, heterogeneous and of no prognostic value—conclusions from 924 cases of two independent series. *Cell Oncol* 2010;32:57–65.