

# Observer variability in the quantitative assessment of tissue-based biomarkers

**Histological Image Analysis Workshop  
Pathology Informatics 2011, Pittsburgh, PA**

**Marios A Gavrielides, Ph.D.**

Division of Imaging and Applied Mathematics (DIAM),  
Office of Science and Engineering Laboratories  
Center for Devices and Radiological Health  
U.S. Food and Drug Administration

**Brandon Gallas, Ph.D (DIAM), Stephen M Hewitt, M.D.,Ph.D., (NCI/NIH)**

# Overview

- **Introduction**

- Tissue-based cancer biomarkers
- Assessment of biomarker expression with immunohistochemistry (IHC)
- Variability involved in IHC evaluation

- **Observer study**

- Examining the use of computer-aided assessment
  - HER2 IHC evaluation

- **Related ongoing research at DIAM**

# Overview-Cancer biomarkers

- **Biological markers (biomarkers) can identify characteristics linked to tumor behavior**
  - can lead to improved clinical decisions
  - specific to individual patients
- **Uses on all aspects of cancer:**
  - diagnosis, staging, prognosis, treatment selection, monitoring treatment

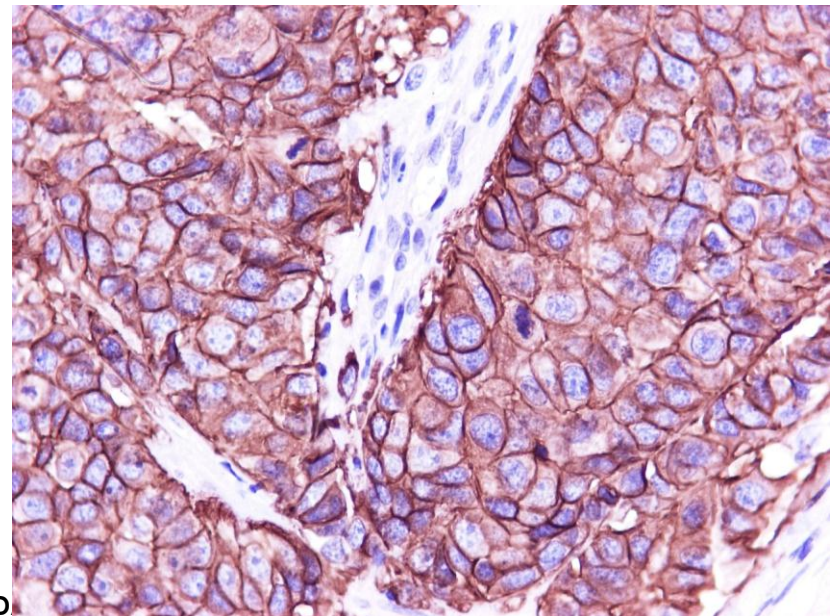
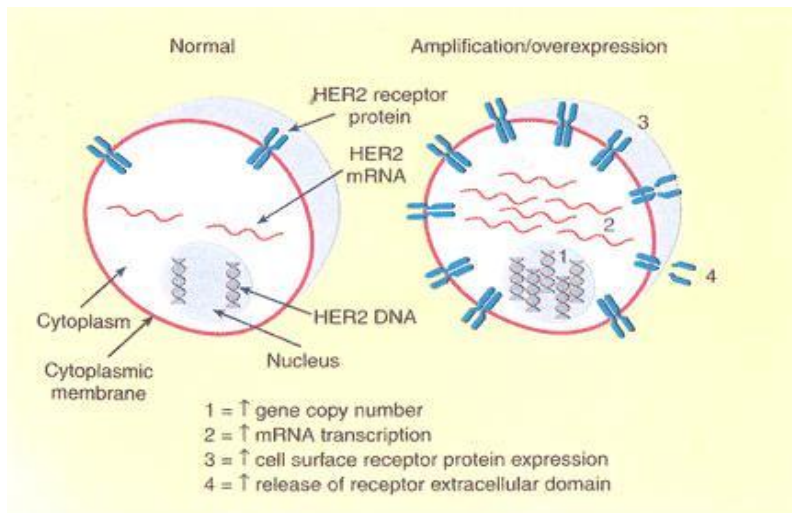
# Biomarkers for breast cancer

- **HER2/neu (human epidermal growth receptor)**
  - Tissue-based biomarker
  - over-expressed in 20-25% of breast cancer pat.
  - good responders to adjuvant trastuzumab (Herceptin, Genentech, CA)
  - shown to reduce:
    - risk of recurrence by ~50% and
    - mortality by ~30% (Wolff, *Arch Pathol Lab Med*, 2007)

# Immunohistochemistry (IHC)

- **HER2 and other tissue-based biomarkers are assessed with IHC**

- makes it possible to detect *antigens* in tissue
- Multi-step procedure resulting in paraffin-embedded stained tissue sections



# Immunohistochemistry (IHC)

- **Increasing levels of measurement**

- **accuracy**

- binary: negative/ positive (i.e. estrogen receptor)
- semi-quantitative (ordinal): 0, 1+, 2+, 3+ (i.e. HER2)
  - related to clinical follow-up decisions
- quantitative: continuous scale
  - related to number of receptor

# Immunohistochemistry (IHC)

• IHC has become a major part of surgical pathology practice:

- it can identify a wide number of antigens
- results can be viewed using only a light microscope
- slides retain properties for a long time
- inexpensive

# IHC is limited by lack of reproducibility

- **Inter- and intra-laboratory**

- In > 2000 patients in Canada >40% FN rate in determining ER-positive patients [1,2]

- **Inter- and intra-observer variability**

- Hsu et. al reported complete agreement in 48% of HER2 cases (22 out of 46, 5 observers) [3]
- Distinguishing 2+ from 3+ showed agreement in only 13 (59%) of 22 positive cases

[1] K. Hede, "Breast Cancer Testing Scandal Shines Spotlight on Black Box of Clinical Laboratory Testing," *Journal of the National Cancer Institute* vol. 100, 2008.

[2] D. C. Allred, "Commentary: Hormone Receptor Testing in Breast Cancer: A Distress Signal from Canada," *The Oncologist*, vol. 13, pp. 1134-1136, 2008.

[3] Hsu C-Y, Ho DM-T, Yang C-F, Lai C-R, Yu I-T, Chiang H. InterObserver Reproducibility of Her-2/neu Protein Overexpression in Invasive Breast Carcinoma Using the DAKO HercepTest. *American Journal of Clinical Pathology*. 2002;118(5):693-698.

# Variability in biomarker assessment

- **Hinders the clinical utility of biomarkers**
  - clinicians must trust the test
- **Reduces the statistical power of studies for biomarker discovery**
- **Reduces the statistical power of clinical trials for drug efficacy**
  - increases the size and cost of clinical trials
  - delays the adoption of new targeted therapies
- **What are the sources of variability?**

# Immunohistochemistry:

## Multiple sources of variability

**Tissue preparation**

**Tissue labeling**

**Tissue slide  
interpretation**

• IHC results can be affected greatly by:

- tissue section thickness
- choice of fixatives,
- delay in fixation
- over-fixation
- inadequate tissue dehydration prior to paraffin embedding

M. Werner et al, "Effect of Formalin Tissue Fixation and Processing on Immunohistochemistry," *The American Journal of Surgical Pathology*, vol. 24, pp. 1016-1019, 2000.

# Immunohistochemistry: Sources of variability

**Tissue preparation**

**Tissue labeling**

**Tissue slide  
interpretation**

- Choice of antigen retrieval and staining methods major source of inter-laboratory variability

**M. Mengel et al , "Inter-laboratory and inter-observer reproducibility of immunohistochemical assessment of the Ki-67 labelling index in a large multi-centre trial," *The Journal of Pathology*, vol. 198, pp. 292-299, 2002.**

# Immunohistochemistry: Sources of variability

**Tissue preparation**

**Tissue labeling**

**Tissue slide  
interpretation**

- **Several efforts have sought to:**
  - develop standardized assay methodologies [1, 2]
  - develop objective methods of measurement [3]
  - provide external staining standards and proficiency tests [4].
- **Quality control/standardization in IHC is an ongoing process.**

[1] T. J. O'Leary, "Standardization in immunohistochemistry," *Applied Imm Mol Morph*, vol. 9, p. 3, 2001.

[2] R. Leake, et al "Immunohistochemical detection of steroid receptors in breast cancer: a working protocol," *British Medical Journal*, vol. 53, p. 634, 2000.

[3] S. S. Cross, "Observer accuracy in estimating proportions in images: implications for the semiquantitative assessment of staining reactions and a proposal for a new system," *J Clin Pathol*, vol. 54, p. 385, 2001.

[4] A. Rhodes et al "Immunohistochemical demonstration of oestrogen and progesterone receptors: correlation of standards achieved on in house tumours with that achieved on external quality assessment material in over 150 laboratories from 26 countries," *J Clin Pathol*, vol. 53, p. 292, 2000.

# Immunohistochemistry: Procedure

Tissue preparation

Tissue labeling

Tissue slide  
interpretation



Optical (light)  
microscopy

Digital (virtual) microscopy)



- The interpretation of IHC staining by pathologists is one of the most important sources of variability in the assessment of biomarkers

# Biomarker staining interpretation in pathology

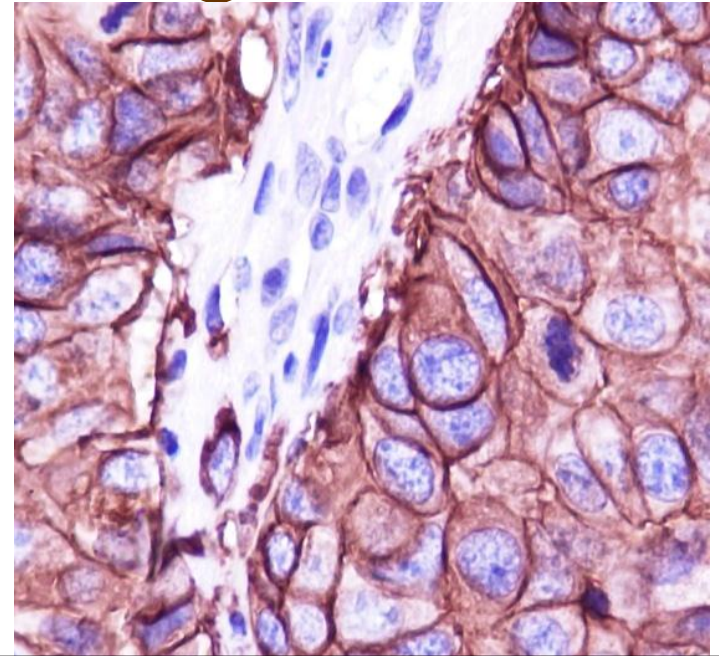
- **Generally: Human perception varies among individuals**
  - Can be affected by training, experience, physical differences, fatigue
  - Observer variability is known to exist in other fields like Radiology

# Biomarker staining interpretation in pathology

- **Observer variability specific to pathology:**
  - Evaluating different regions of a slide (tumor heterogeneity)
  - Using different cut-offs to determine stained cell positivity
  - Using different approaches to *combine* region scores into a single slide score
  - Use of different microscopes, illumination sources, reading conditions
  - Subjective criteria/guidelines for assessment of biomarker expression

# Interpretation of Her-2/neu using IHC: membranous staining

- Evaluation based on color stain assessment:
  - membrane staining completeness
  - membrane staining intensity



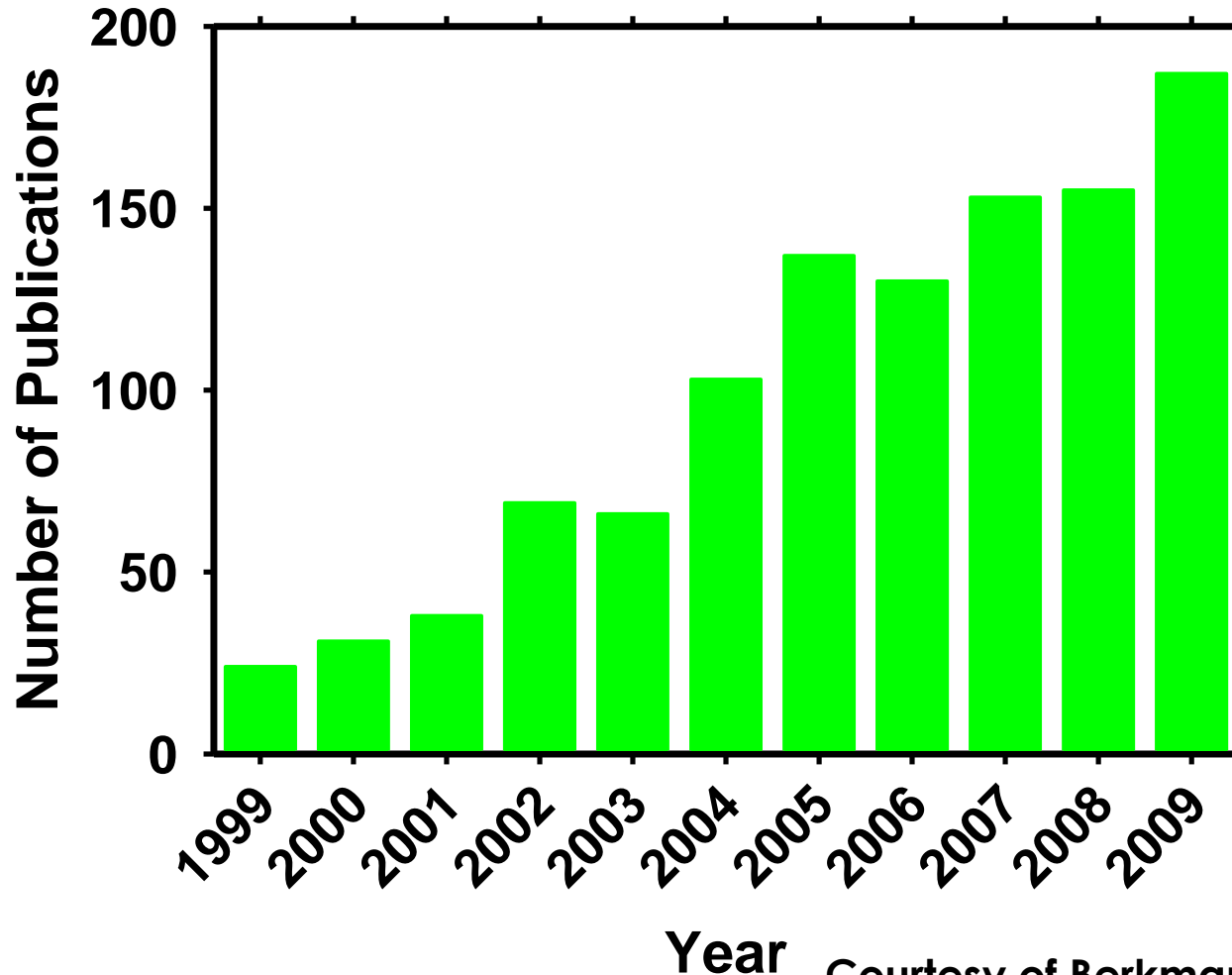
## Subjective criteria

SCORE	STAINING PATTERN
0	No staining observed or membrane staining <10% of tumor cells
1+	Faint partial membrane staining in >10%
2+	Weak to moderate complete membrane staining in >10% of cells
3+	Moderate to strong complete membrane staining in >30% of tumor cells

# IHC interpretation in pathology: Computer-aided assessment

- ***Computer-aided assessment*** of IHC could make the task more objective, quantitative, reproducible, automated, more efficient
  - by-product of digital pathology
  - Enabled by technological advances in whole slide imaging (WSI)
    - WSI scanning of ~1 min/slide
    - autofocus algorithms, z-axis focus
    - High resolution
    - Automated feeding of multiple slides for high throughput

# PubMed search, “Computer-aided detection” OR “Computer-aided diagnosis”



Courtesy of Berkman Sahiner, FDA

# Computer-aided diagnosis

- **In 2008, radiologists used CAD in 74% of screening exams\***
  - **Medicare part B physician/supplier procedure summary master files**
    - **5,827,326 screening mammograms**
    - **4,305,595 with CAD**

\* VM Rao et al., “How widely is computer-aided detection used in screening and diagnostic mammography?” *Journal of the American College of Radiology*, 2010, 7:802-805.

# Computer-aided assessment of IHC

- **Commercial software available**
  - FDA cleared for IHC assessment
- **Still under-examined area**
  - Interaction of pathologists with computer aids
- **We have developed a computer aid for the assessment of HER2**
- **Examined benefit when used by observers**

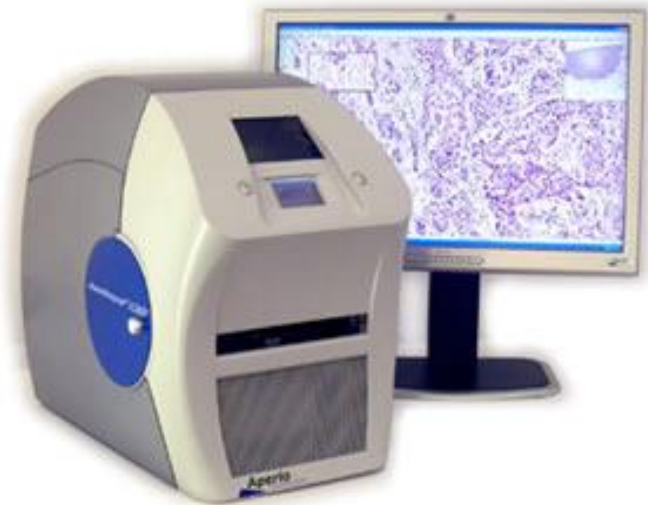
# Development of an automated method for the quantitative assessment of HER2

- Automated extraction of continuous measures of:
  - **membrane staining intensity** and
  - **membrane staining completeness**
- Measures can be provided to observer or used to classify slide

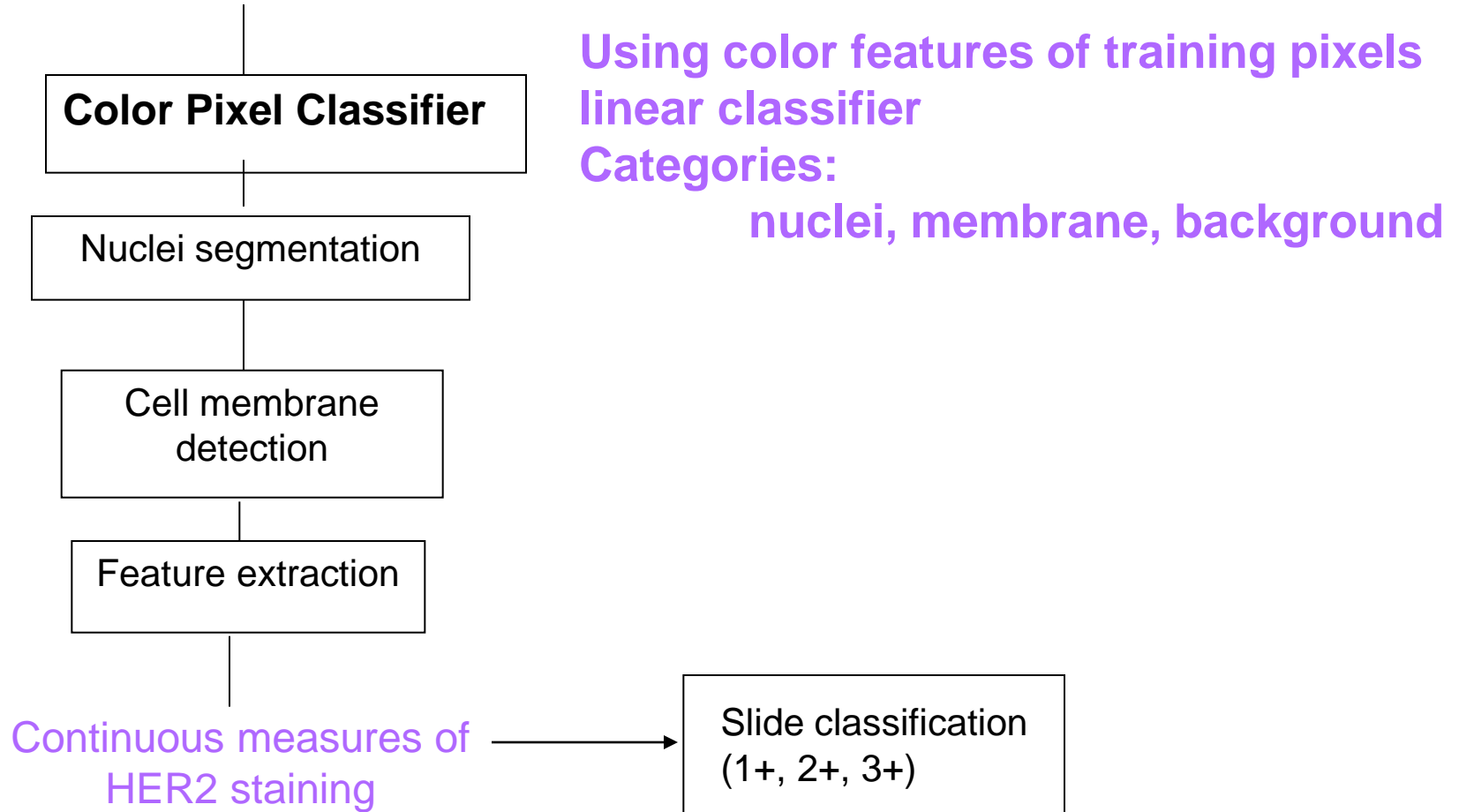
H Masmoudi, S Hewitt, K J Myers, N Petrick, and M A Gavrielides, "Automated quantitative assessment of HER-2/neu immunohistochemical expression in breast cancer", *IEEE Transactions on Medical Imaging*, vol. 28, n.6, pp.916-925, 2009.

# Materials

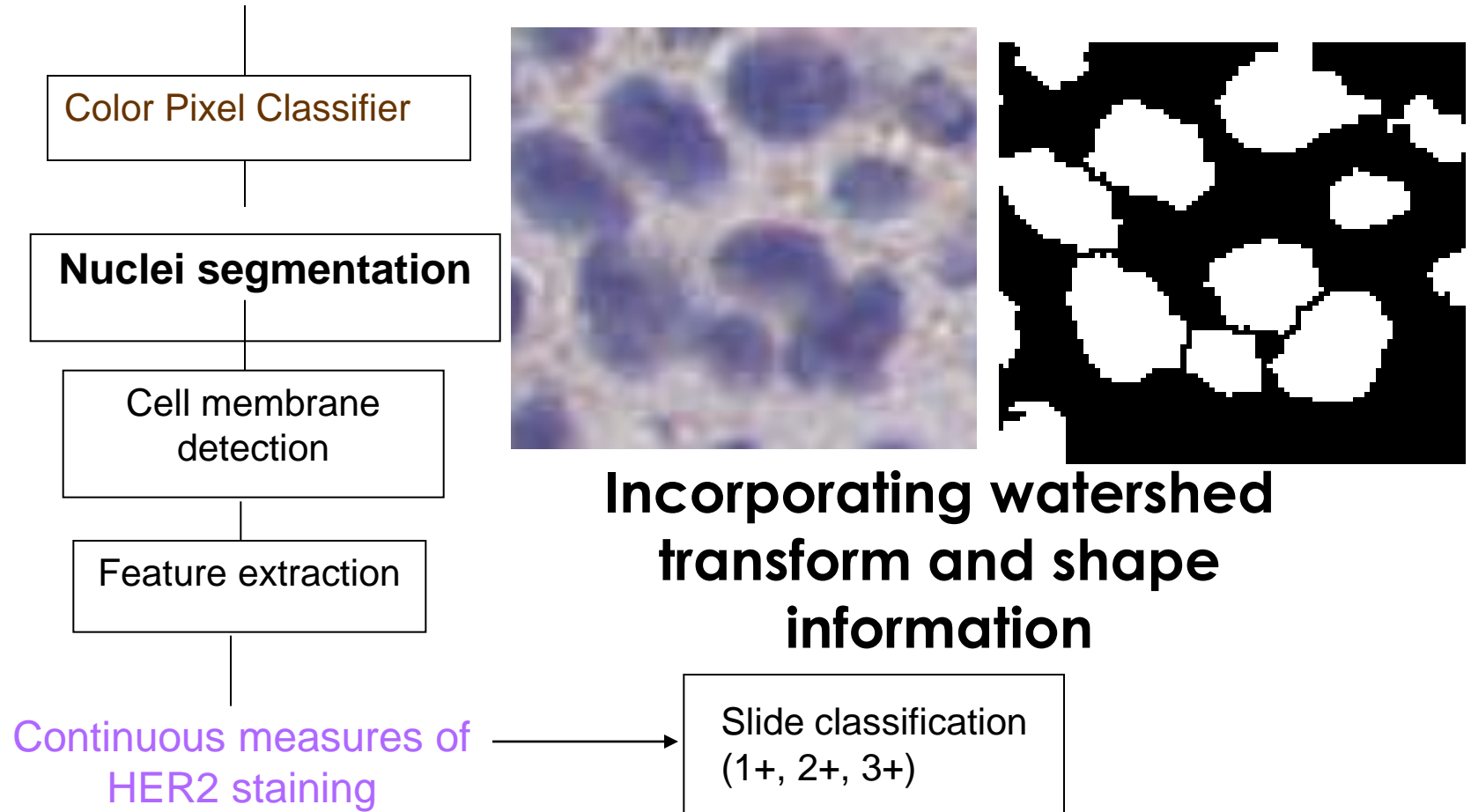
- **77 breast cancer tissue slides stained for HER2**
  - HER2 scores: 26 (1+), 27 (2+) , 24 (3+)
  - Truth from archives of the Department of Pathology, University of California, Irvine
- **Whole slide scanning**
  - Aperio Scanscope T2 system
  - Stephen Hewitt, NCI
- **Multiple images (~5) containing invasive cancer cells were extracted from each slide**



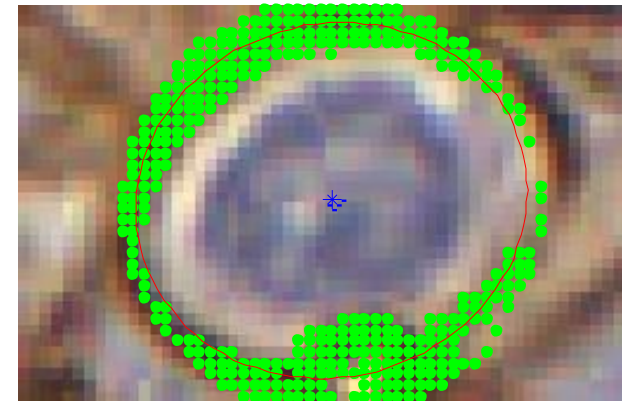
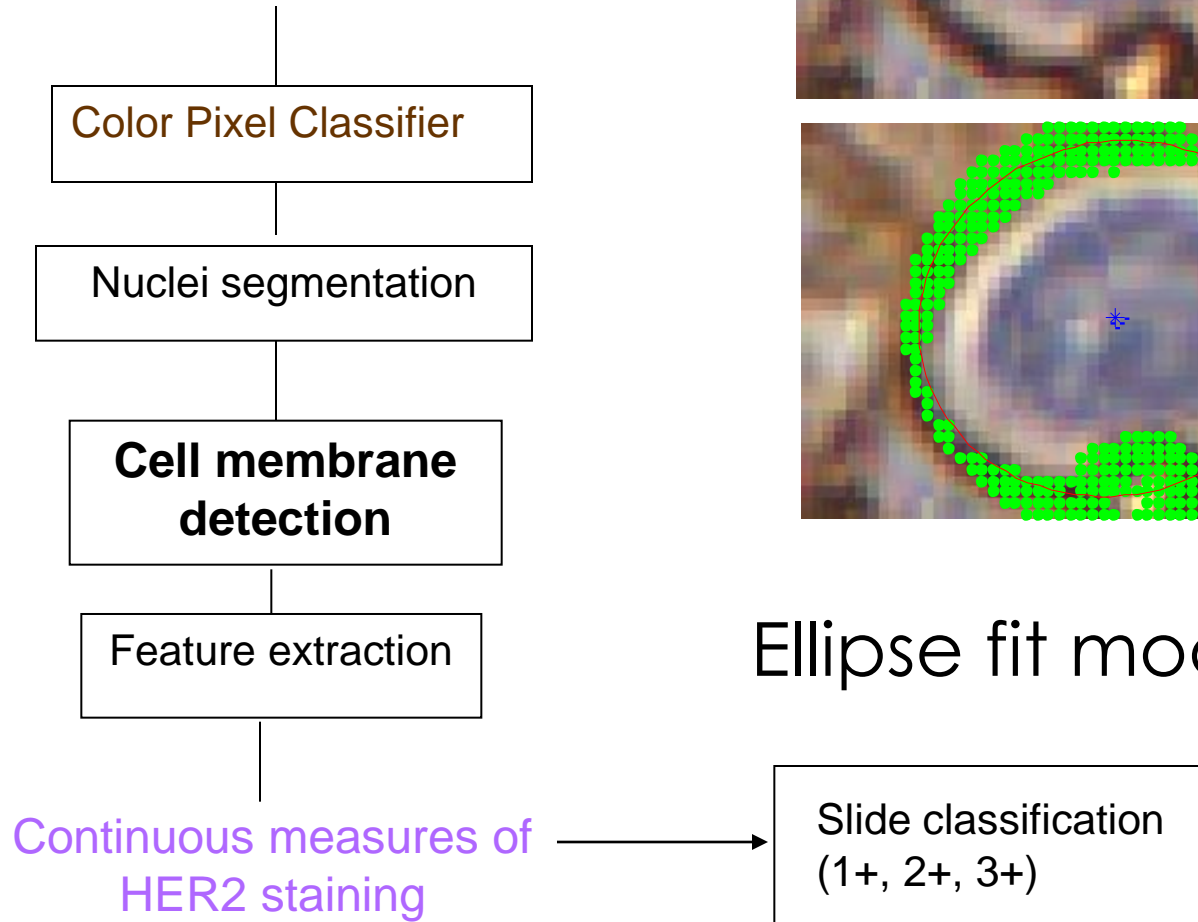
# Algorithm Overview



# Algorithm Overview

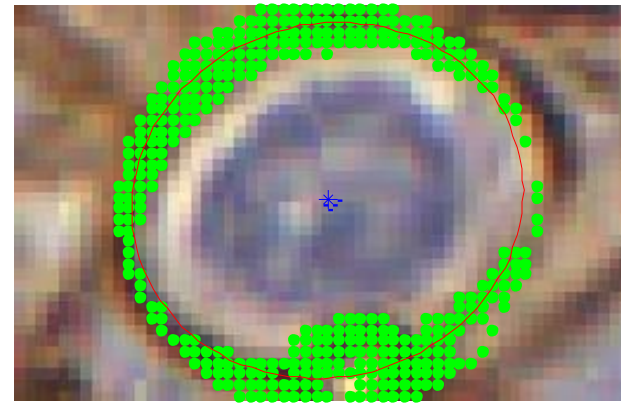


# Algorithm Overview



Ellipse fit model

## Assume an ellipse model on the membrane

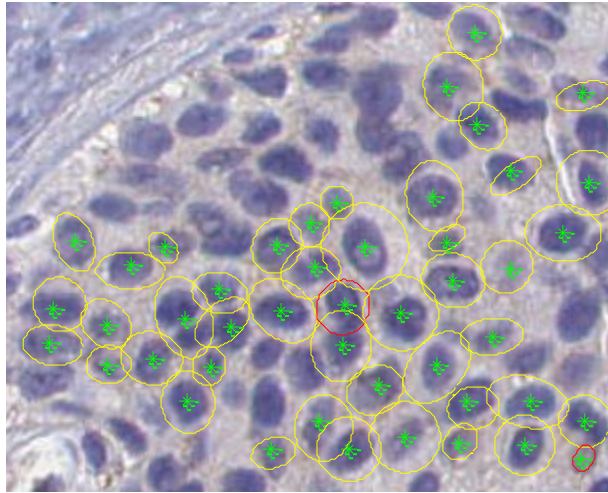


Two features are extracted for each detected membrane:

**Membrane closing (completeness):** the percentage of ellipse pixels overlapping with membrane pixels

**Membrane Intensity:** the average staining intensity (average membrane color pixel classifier output) of the fitted ellipse pixels.

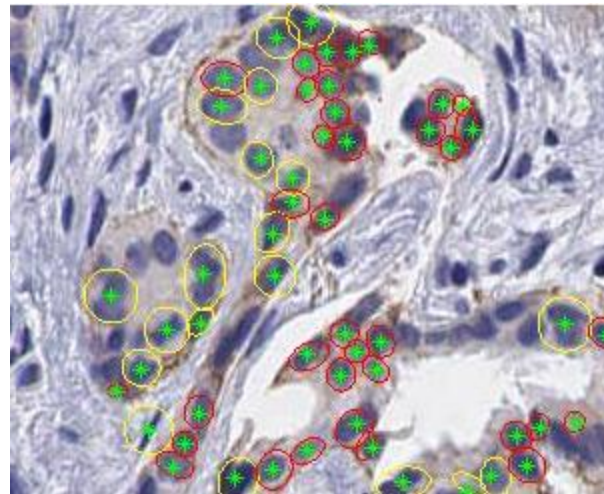
# Examples



1+

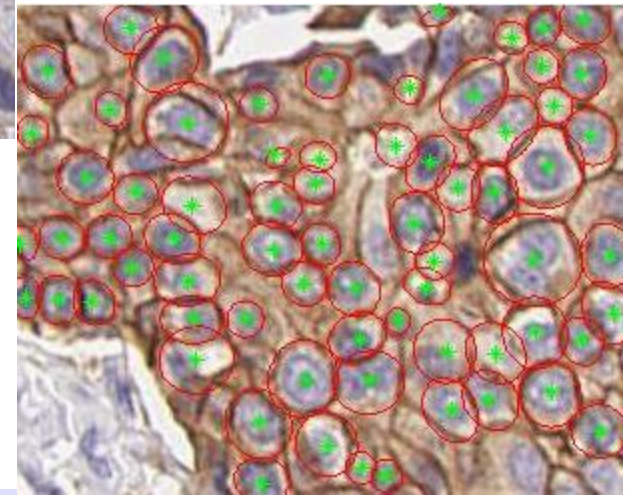
Yellow ellipses  $\rightarrow$  membrane completeness  $<0.5$

Red ellipses  $\rightarrow$  membrane completeness  $>0.5$

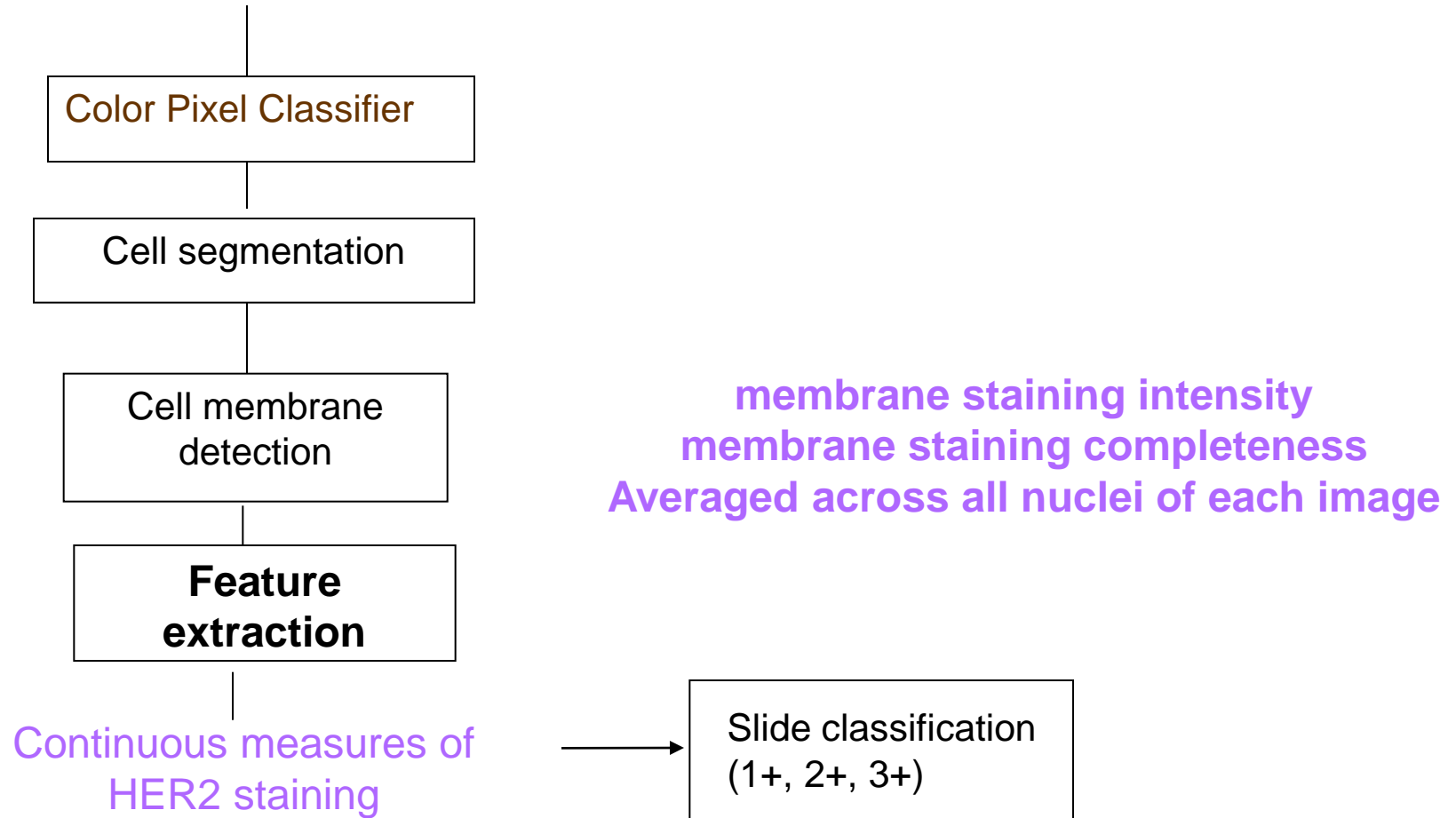


2+

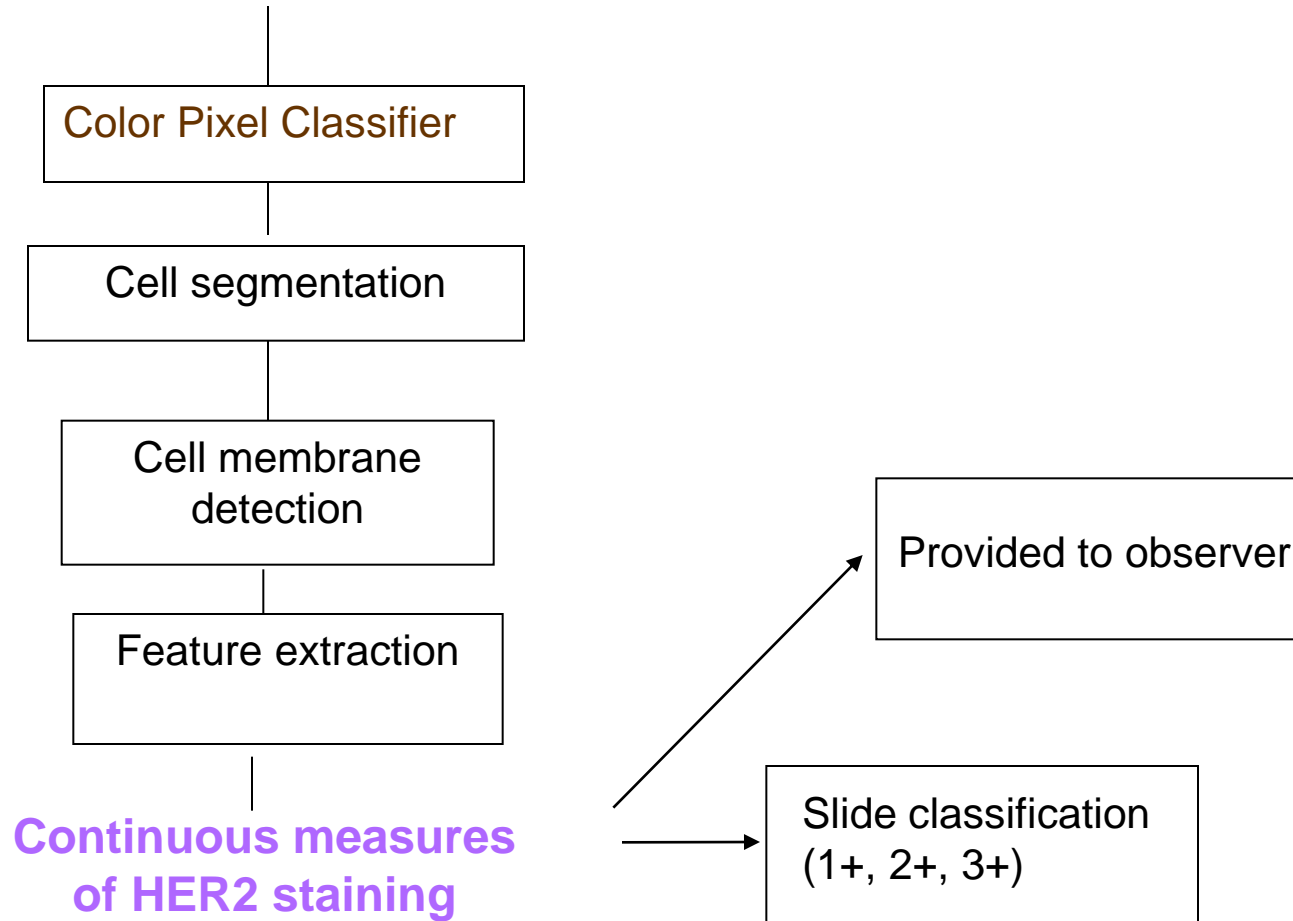
3+



# Algorithm Overview



# Algorithm Overview



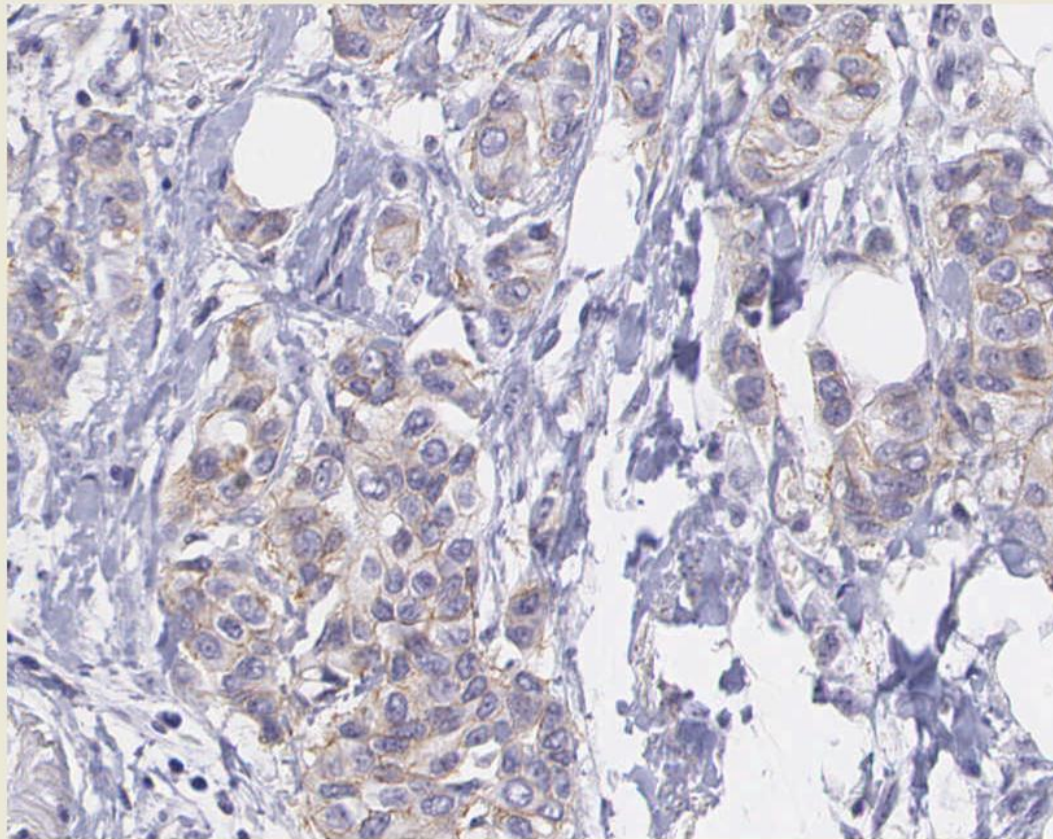
# Results

- 13 slides (out of 77) were used for:
  - pixel classification training
  - development of nuclei segmentation algorithm
- Remaining 64 slides (22 1+ , 22 2+ and 20 3+)
  - K-fold cross validation (*slide* classification)
- Performance assessment metric:
  - percent correct agreement with pathologist scores
  - Overall: 83% (82% 1+, 78% 2+, 88% 3+)
- **Could this computer aid provide benefit to pathologists?**

# IHC assessment of HER2 with *unaided* and *computer-aided* digital microscopy: observer study

- **Goal: examine whether computer-aided information can benefit pathologist performance**
- **Reader study performed in our Display lab:**
- **14 readers**
  - 7 pathologists (range in experience) from FDA, NCI/NIH
  - 7 novices (DIAM scientists, no experience in pathology)
- **Two reading modes:**
  - Unaided assessment of HER2 expression
  - Computer-aided assessment of HER2 expression

# HER2 Assessment: unaided

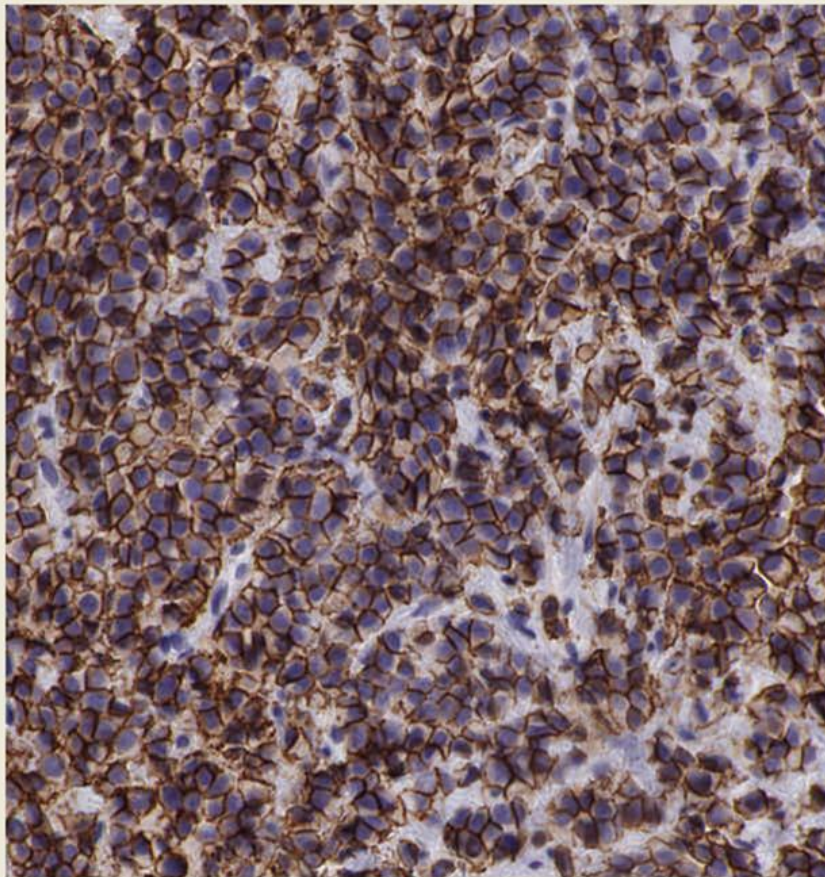


ROI 1/500

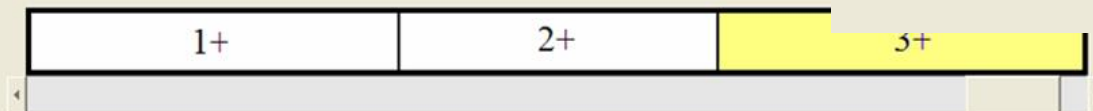
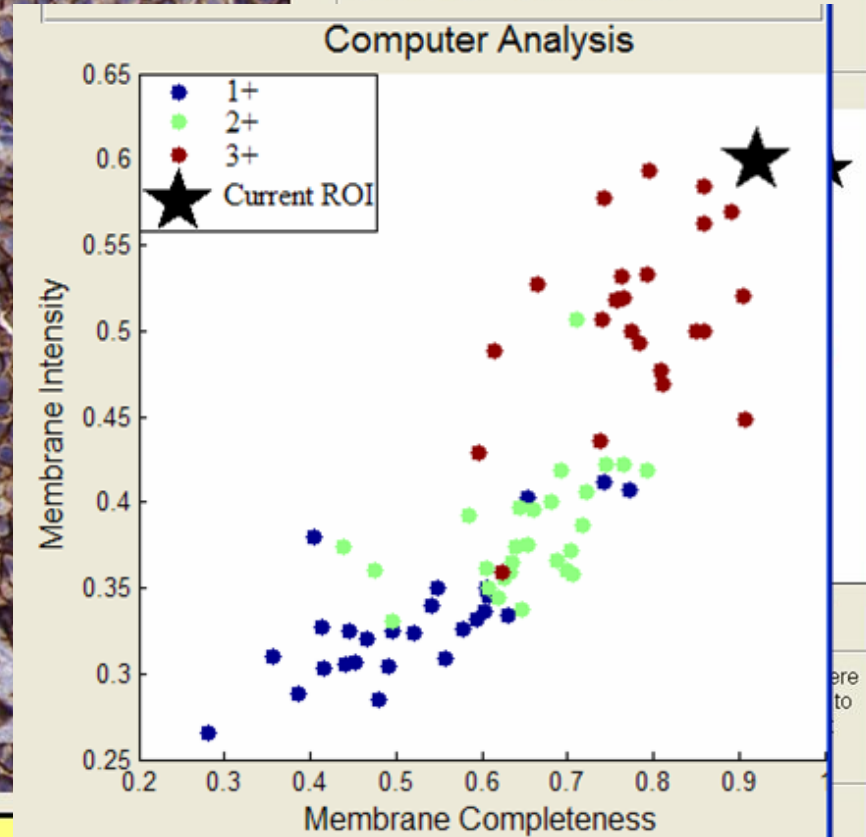
1+	2+	3+
----	----	----

Score	Rating	Pause
Score	Rating	Pause

# HER2 Assessment: computer-aided



ROI 1/500



Score

3

Rating

97

Pause

Next

# IHC assessment of HER2 with *unaided* and *computer-aided* digital microscopy: observer study

- **3-step Training**

- Presentation about study, HER2, guidelines for scoring, instructions for using computer-aid
  - Use the whole range! (not all 3s are equal)
- Scoring session with feedback (30 training im.)
  - If score differed significantly from score of expert, score again
- Practice session with unaided and computer-aided (30 training images)

# IHC assessment of HER2 with *unaided* and *computer-aided* digital microscopy: reader study

- **Main study**

- 335 images from 64 HER2 slides,
- Randomized: case order, reading mode
- Read in 2 sessions, > 1 month apart
- 241 read in 2 reading modes (inter-observer variability)
- 94 (47 aided, 47 unaided) read in same reading mode (intra-observer variability)
- All images read on same calibrated monitor, same reading conditions
- Reading time: 1 ½ - 2 ½ hours/session

# IHC assessment of HER2 with *unaided* and *computer-aided* digital microscopy: observer study

- **Analysis:**

- No truth available for HER2 expression
  - **Quantify observer variability with Agreement Analysis**
- Intra-reader agreement
  - *How well does a reader agree with self?*
- Inter-reader agreement within a group
  - *How well do like readers agree?*
- Inter-reader agreement across groups
  - How well do readers from two groups agree?

# Agreement metrics

- **Intra-class coefficient (ICC)**

- data are pooled to estimate the mean and variance
- describes how strongly units in the same group resemble each other > **group agreement (inter-observer)**

- **Kendall's tau**

- Consider a pair of readers ranking a pair of cases:
  - $C$  = # of Concordances: two readers rank a pair of cases in the same order
  - $D$  = # of Discordances: two readers rank a pair of cases in the opposite order
  - $T_1$ : Tie for reader 1 only
  - $T_2$ : Tie for reader 2 only
- suitable for both continuous and categorical data
- Takes ranking ties into account
  - > **pair-wise agreement (inter and intra-observer)**

$$K_{\tau} = \frac{C - D}{\sqrt{(C + D + T_1)(C + D + T_2)}}$$

# IHC assessment of HER2 with *unaided* and *computer-aided* digital microscopy: observer study

Observer group	Inter-observer (group) agreement from <i>continuous</i> data ( <i>Intraclass correlation coefficient</i> )		Inter-observer (group) agreement from <i>categorical</i> data ( <i>Intraclass correlation coefficient</i> )	
	Unaided	Computer-aided	Unaided	Computer-aided
<b>Overall</b>	0.81 (0.78-0.84)	0.92 (0.91-0.93)	0.72 (0.68-0.76)	0.83 (0.80-0.86)
<b>Pathologists</b>	0.80 (0.76-0.83)	0.91 (0.89-0.92)	0.71 (0.66-0.75)	0.82 (0.79-0.85)
<b>Novices</b>	0.83 (0.80-0.86)	0.93 (0.92-0.94)	0.74 (0.70-0.78)	0.85 (0.82-0.87)

# IHC assessment of HER2 with *unaided* and *computer-aided* digital microscopy: reader study

Reader group	Inter-reader (pairwise) agreement using <i>continuous</i> data (Kendall's $\tau_b$ )			Inter-reader (pairwise) agreement using <i>categorical</i> data (Kendall's $\tau_b$ )		
	Unaided	Computer-aided	Difference	Unaided	Computer-aided	Difference
<b>Overall</b>	0.62 (0.56-0.67)	0.76 (0.71-0.81)	<b>0.14</b> <b>(0.10-0.19)</b>	0.70 (0.64-0.76)	0.82 (0.76-0.87)	<b>0.12</b> <b>(0.06-0.17)</b>
<b>Pathologists</b>	0.61 (0.53-0.67)	0.75 (0.66-0.81)	<b>0.14</b> <b>(0.08-0.20)</b>	0.69 (0.61-0.76)	0.80 (0.72-0.89)	<b>0.11</b> <b>(0.03-0.19)</b>
<b>Novices</b>	0.63 (0.57-0.69)	0.78 (0.72-0.83)	<b>0.15</b> <b>(0.09-0.21)</b>	0.71 (0.63-0.77)	0.83 (0.77-0.89)	<b>0.12</b> <b>(0.06-0.19)</b>
<b>Pathol. vs. Novices</b>	0.62 (0.57-0.66)	0.76 (0.71-0.81)	<b>0.14</b> <b>(0.11-0.19)</b>	0.70 (0.65-0.75)	0.82 (0.76-0.86)	<b>0.12</b> <b>(0.06-0.16)</b>

# IHC assessment of HER2 with *unaided* and *computer-aided* digital microscopy: observer study

Reader group	Intra-reader agreement using <i>continuous</i> data (Kendall's $\tau_b$ )			Intra-reader agreement using <i>categorical</i> data (Kendall's $\tau_b$ )		
	Unaided	Computer-aided	Difference	Unaided	Computer-aided	Difference
<b>Overall</b>	0.71 (0.64-0.76)	0.81 (0.75-0.85)	<b>0.10</b> <b>(0.02-0.17)</b>	0.74 ( 0.66-0.80)	0.84 (0.77-0.89)	<b>0.10</b> <b>(0-0.19)</b>
<b>Pathologists</b>	0.72 (0.64-0.79)	0.80 (0.72-0.86)	0.08 (-0.01-0.18)	0.73 (0.63-0.82)	0.85 ( 0.76-0.92)	0.12 (-0.01-0.25)
<b>Novices</b>	0.70 (0.60-0.77)	0.81 (0.73-0.87)	<b>0.11</b> <b>(0.01-0.23)</b>	0.75 ( 0.62-0.84)	0.82 (0.75-0.89)	0.07 (-0.04-0.20)

# IHC assessment of HER2 with *unaided* and *computer-aided* digital microscopy: reader study

- **Inter- and intra-observer agreement was improved**
- **Novices performed comparably to pathologists**
  - some tasks in pathology can be done by non-physicians if properly trained
- **Score variability was much smaller for 3+ cases**
  - possible role of such computer-aided as “triage” software
- **Potential for computer-aided assessment to improve pathologist performance**

Marios A. Gavrielides, Brandon Gallas, Petra Lenz, Aldo Badano, and Stephen M Hewitt, “Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy”, *Archives of Pathology and Laboratory Medicine*, vol. 135, n. 2, pp. 233-242, 2011

# Discussion/Future work

- **On computer aid:**
  - Algorithm based on manually selected training pixels
    - Difficult to retrain on different datasets (different image properties)
    - Same issue exists with available “tunable” software
  - Have developed alternative method using color content
    - Allows practical, supervised re-training of the algorithm for slides with different color properties

**Brad Keller, Weijie Chen, Marios A Gavrielides, “Quantitative assessment and classification of tissue-based biomarker expression using color content analysis”, *Archives of Pathology and Laboratory Medicine*, (accepted July 2011)**

# Discussion/Future work

- **Study focused only on digital microscopy**
- **New study compares assessment with optical and digital microscopy**
  - Multiple biomarkers for breast cancer: Ki67, ER/PR, HER2
  - Analysis of whole section slides and WSIs (ROI in previous study)
  - Analysis of tissue microarray (TMA) slides and images
  - Analysis of IHC interpretation when using different staining antibodies
- Other projects focusing on:
  - color reproducibility/management
  - Inter-scanner variability

# Ongoing related projects

- **Will provide data and experience toward answering regulatory questions regarding the performance evaluation of digital pathology devices/software**
  - Technical evaluation of systems
  - Diagnostic performance of pathologists
- **Provide guidance to developers**

# In summary

- **Have presented ongoing research**
  - quantitative assessment of HER2
  - effect of computer aids on observer variability
- **New technologies such as WSI and computer aided assessment in tissue imaging have potential in improving diagnostic efficacy in pathology**
  - **Still need research to determine the role and limitations of such technologies**

# Acknowledgement

- Office of Women's Health for their support
- Collaborators at NCI/NIH (Dr. Stephen Hewitt)
- Collaborators at OIVD
- All the readers that participated in our reader studies (FDA/OIVD, NCI/NIH)
- Comments/questions/suggestions:
  - [marios.gavrielides@fda.hhs.gov](mailto:marios.gavrielides@fda.hhs.gov)