Validation of Quantitative Digital Pathology Analyses

Auranuch Lorsakul[*,1], Joerg Bredno[*,2], Jim Martin[1], Shawn Wang[2], Kien Nguyen[1], Faith Ough[3], June Clements[3], and Solange Romagnoli[4]

[1] Roche Tissue Diagnostics / Ventana Medical Systems, Inc., Imaging & Algorithms, Mountain View, CA, USA
[2] Formerly of Roche Tissue Diagnostics / Ventana Medical Systems, Inc., Imaging & Algorithms, Mountain View, CA, USA
[3] Roche Tissue Diagnostics / Ventana Medical Systems, Inc., Medical and Scientific Affairs, Tucson, AZ, USA
[4] Roche Diagnostics, Biomarker Department, Basel, Switzerland

* Contributed equally

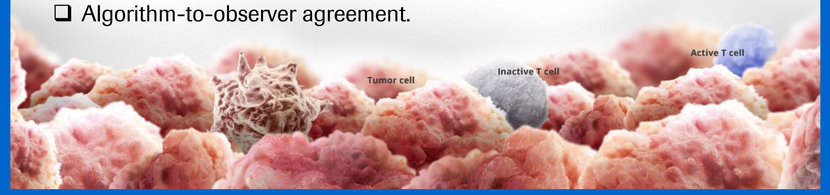
1 – Background

Introduction

- **Digital Pathology** algorithms quantify the content of a whole slide or selected field-of-view (FOV) with respect to number of cells for one or more phenotypes in Immunohistochemistry (IHC).
- For assessing the apparent **immune response** to cancer, a count and area density of immune cells e.g., T-lymphocytes can be readily generated.
- Automated analyses require stringent **validation** to establish and assure the accuracy of cell counts.

Objectives

- We **compared** automatically generated cell counts to ground truth counts obtained from expert pathologists in a framework that collects the following data:
  - ❑ Inter-observer agreement.
  - ❑ Section-to-section agreement using aligned and registered FOVs.
  - ❑ Algorithm-to-observer agreement.



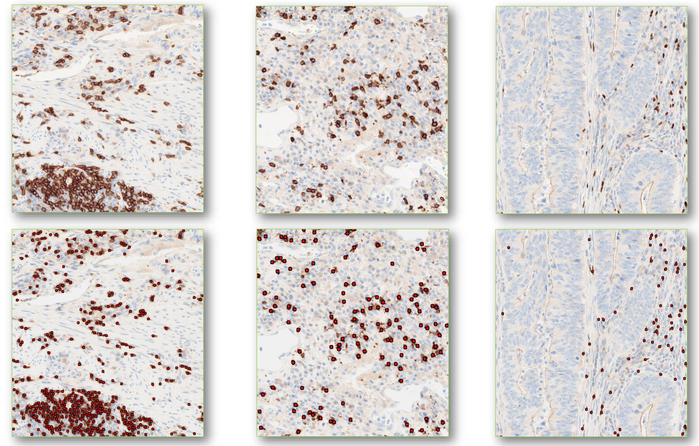
2 – Methods

2.1 Ground Truth Collection

- An easy-to-use graphic user interface (GUI) tool was used to facilitate the potentially fatiguing ground truth (GT) effort by the pathologists.
- To facilitate and avoid biasing the manual GT effort, some perturbed algorithm results (including random false positives and false negatives) were preloaded.
- Example studies are presented for the assessment of tumor cells and T-lymphocytes in a patient from a patient cohort with **Stage II colorectal cancer**.
  - ❑ The 4-μm tissue sections were stained for CD3 (anti-CD3 2GV6) and CD8 (anti-CD8 SP238/57) on consecutive tissue sections.
  - ❑ Two pathologists selected FOVs from a set of 119 slides stained with CD3 and 119 slides stained with CD8.
  - ❑ On each slide, a pathologist selected 3 FOVs that represent tumor with **high** immune infiltrate, tumor with **low** immune infiltrate, and the **invasive** margin, respectively.
  - ❑ The pathologists marked every T-cell in these FOVs.
  - ❑ On 10 consecutive slide pairs, both pathologists provided the cell count in 3 FOVS to determine inter-observer variability.

2.2 Validation of Image Analysis Algorithms

- Computer vision and machine learning algorithms automatically identified the presence and locations of CD3$^+$ (2GV6, SP162), CD8$^+$ (SP238, SP57) and FoxP3$^+$ (SP97) lymphocytes on 3,3'-diaminobenzidine (DAB) IHC stained tissue sections.

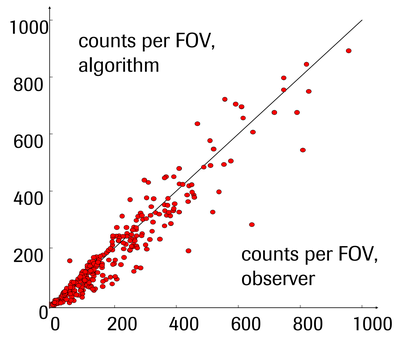


• CD3$^+$ • CD8$^+$ • FoxP3$^+$

3 – Results

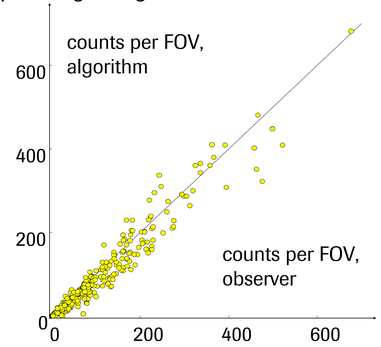
- **The algorithm-to-pathologist agreement was fully consistent with the pathologist-to-pathologist agreement.**
- A total of 60 FOVs was used for the inter-observer study. The two pathologists agreed with R^2=0.957 and R^2=0.925 for CD3 and CD8 cell counts, respectively.

CD3



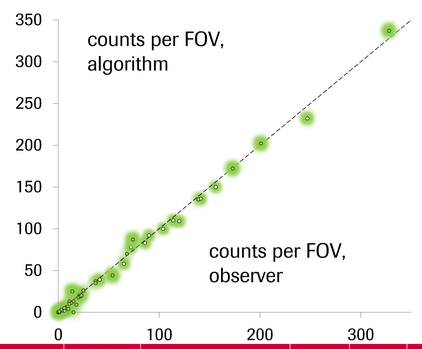
Experiment	# FOVs	Total Counts	Ratio	R^2	CCC
Inter-observer agreement	30	6,947 (obs. 1) 6,328 (obs. 2)	0.911	0.957	0.970
Section-to-section agreement	33	25,467 (slide set 1) 24,825 (slide set 2)	0.975	0.893	0.969
Algorithm-to-observer agreement	357	72,076 (observer) 66,179 (algorithm)	0.918	0.901	0.945

CD8



Experiment	# FOVs	Total Counts	Ratio	R^2	CCC
Inter-observer agreement	30	3,115 (obs. 1) 2,795 (obs. 2)	0.897	0.925	0.919
Section-to-section agreement	30	5,438 (slide set 1) 4,991 (slide set 2)	0.918	0.766	0.865
Algorithm-to-observer agreement	357	34,133 (observer) 30,438 (algorithm)	0.891	0.943	0.965

FoxP3



Experiment	# FOVs	Total Counts	Ratio	R^2	CCC
Section-to-section agreement	44	1,438 (slide set 1) 1,393 (slide set 2)	0.969	0.868	0.921
Algorithm-to-observer agreement	45	2,615 (observer) 2,535 (algorithm)	0.969	0.995	0.997

- A total of 714 manually counted FOVs was used for validation of the image analysis algorithm.
- Image analysis matched ground truth counts with R^2=0.901 and R^2=0.943 for CD3 and CD8, respectively.
- A total of 72,076 manual cell counts versus 66,179 automated (ratio 0.918), and 34,133 manual versus 30,438 automated (ratio 0.891) were used for CD3 and CD8, respectively.

4 - Conclusions

- A rigorous **validation** is required to relate algorithm-to-observer agreement to inter-observer agreement and section-to-section variability of cell counts.
- In this study, the section-to-section variability demonstrates a probable upper limit on **cell count** accuracies.