Modern Pathology (2017) 30, 340–349

Quantitative and pathologist-read comparison of the heterogeneity of programmed death-ligand 1 (PD-L1) expression in non-small cell lung cancer

Jamaal A Rehman[1,4], Gang Han[2,4], Daniel E Carvajal-Hausdorf[1], Brad E Wasserman[1], Vasiliki Pelekanou[1], Nikita L Mani[1], Joseph McLaughlin[3], Kurt A Schalper[1,3] and David L Rimm[1,3]

[1]Department of Pathology, Yale University School of Medicine, New Haven, CT, USA; [2]Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M University, College Station, TX, USA and [3]Department of Medicine (Oncology), Yale University School of Medicine, New Haven, CT, USA

PD-L1 is expressed in a percentage of lung cancer patients and those patients show increased likelihood of response to PD-1 axis therapies. However, the methods and assays for the assessment of PD-L1 using immunohistochemistry are variable and PD-L1 expression appears to be highly heterogeneous. Here, we examine assay heterogeneity parameters toward the goal of determining variability of sampling and the variability due to pathologist-based reading of the immunohistochemistry slide. SP142, a rabbit monoclonal antibody, was used to detect PD-L1 by both chromogenic immunohistochemistry and quantitative immunofluorescence using a laboratory-derived test. Five pathologists scored the percentage of PD-L1 positivity in tumor- and stromal-immune cells of 35 resected non-small cell lung cancer cases, each represented on three separate blocks. An intraclass correlation coefficient of 94% agreement was seen among the pathologists for the assessment of PD-L1 in tumor cells, but only 27% agreement was seen in stromal/immune cell PD-L1 expression. The block-to-block reproducibility of each pathologist's score was 94% for tumor cells and 75% among stromal/immune cells. Lin's concordance correlation coefficient between pathologists' readings and the mean immunofluorescence score among blocks was 94% in tumor and 68% in stroma. Pathologists were highly concordant for PD-L1 tumor scoring, but not for stromal/immune cell scoring. Pathologist scores and immunofluorescence scores were concordant for tumor tissue, but not for stromal/immune cells. PD-L1 expression was similar among all the three blocks from each tumor, indicating that staining of one block is enough to represent the entire tumor and that the spatial distribution of heterogeneity of expression of PD-L1 is within the area represented in a single block. Future studies are needed to determine the minimum representative tumor area for PD-L1 assessment for response to therapy.
Modern Pathology (2017) **30,** 340–349; doi:10.1038/modpathol.2016.186; published online 11 November 2016

Last year, the Food and Drug Administration approved two second-line monoclonal IgG4 antibodies against PD-1 in advanced stage non-small cell lung cancer.[1] Pembrolizumab showed a 45.2% response rate in those patients whose tumors stained over 50% PD-L1 positive and this response was decreased in tumors with a lower ligand expression.[2] Similarly patients receiving Nivolumab had greater objective responses and tumor burden reductions for tumors expressing PD-L1, albeit defined by a different cut-point in a different assay.[3,4] Despite these findings, the predictive value of PD-L1 as a biomarker was questioned due to observations of response or benefit in patients with no evidence of PD-L1 expression.[5–7] One explanation for this observation could be that the tissue sample that tested negative for PD-L1 might have been from a region distinct from other untested areas of the tumor, which were positive.[6,7] Another explanation is that

Correspondence: Dr DL Rimm, MD, PhD, Yale Pathology Tissue Services, Department of Pathology, BML 116 Yale University School of Medicine, 310 Cedar Street, PO Box 208023, New Haven, CT 06520-8023, USA.
E-mail: david.rimm@yale.edu
[4]These authors contributed equally to this work.
Received 29 June 2016; revised 20 September 2016; accepted 25 September 2016; published online 11 November 2016

patients may respond to checkpoint inhibitors regardless of their tumors' PD-L1 expression.[8]

Previous work in our laboratory indicated discordance between different assays measuring PD-L1 among areas within similarly cut sections of the same tumor.[9] This difference could be related to tumor heterogeneity or variability of the assay, the antibody, or the assessment. Here we use a single rabbit monoclonal antibody SP142 (Spring Bioscience) and both quantitative immunofluorescence and conventional chromogenic immunohistochemistry to assess the PD-L1 expression in three separate blocks from 35 resected NSCLC cases. We evaluated the three-block concordance among readers for diaminobenzidine staining in both tumor and immune cells and then compared these results with QIF data of serial sections to define intra-block and inter-block heterogeneity in PD-L1 expression.

Materials and methods

Patient Cohort and Tissue Procurement

Thirty-five cases of untreated, non-small cell lung cancers resected in 2008–2009 were chosen based on tumor size and histology. The corresponding hematoxylin/eosin-stained slides of all 105 blocks were reviewed by a pathologist to verify the diagnosis and the presence of at least 1 cm^2 of tumor on each of three blocks. Only those tumors that were of sufficient size to be represented on three independent tissue blocks were selected for inclusion in the study. A consort diagram providing the overall outline of this project is described in Figure 1. About half of the cases were squamous cell carcinoma and the other half were adenocarcinoma. All the tissues were collected under the conditions of the Yale Human Investigation committee protocols (#9505008219 or #2003025173) to Dr Rimm stipulating signed consent or waiver of consent from all the patients. The clinical characteristics of this cohort are in Table 1.

PD-L1 Antibody Validation

SP142 (Spring Bioscience, Cat #: M4420), a rabbit monoclonal antibody clone of PD-L1, was used to stain whole-tissue sections of each of the 105 formalin-fixed paraffin-embedded blocks. Customized index tissue microarrays (YTMA 245 and 295) containing representative lung cases with variable PD-L1 expression were utilized for antibody titration and validation. Positive and negative control spots on these tissue microarrays included previously validated lung cases, which contained a range of PD-L1 expression. Control samples and reproducibility data are shown in Supplementary Figures 1 and 2. The antibody concentration needed to generate the optimal signal to noise was quantitatively determined on serial cuts of the index tissue microarrays by testing across two logs of antibody concentrations from 1:50 (1.54 μg/ml) to 1:5000 (0.0154 μg/ml). The use of 0.154 μg/ml (1:500 dilution) of SP142 for overnight incubation at 4 °C resulted in the highest signal-to-noise ratio of PD-L1 expression (Figure 2).

Fluorescent and Chromogenic Immunohistochemistry Staining

Whole-tissue sections with respective internal control tissue microarray slides were deparaffinized overnight at 60 °C in a standard laboratory convection oven followed by placement in xylenes twice (20 min each), followed by 100% ethanol twice (1 min each), then 70% ethanol (1 min), and finally a streaming tap water rinse (5 min). Tris-EDTA antigen retrieval buffer was prepared using 1.48 g of EDTA (J.T. Baker, Cat #: 8993-01) dissolved in 4 liters of deionized water, and using 1 M sodium hydroxide dropwise to bring the solution to pH 8. The slides and buffer were then placed in a PT Module (Lab Vision), which heated the buffer to 97 °C for 10 min. Afterwards, the slides were rinsed under a stream of tap water for 10 min before being placed in a methanol/hydrogen peroxide solution (0.75% hydrogen peroxide in methanol) for 30 min. After gently shaking all the slides in double-distilled water for 5 min, the samples were transferred to an autostainer (Thermo Scientific/Lab Vision) and blocked for 30 min at room temperature with 0.3% bovine serum albumin/tris-buffered saline and Tween.

For quantitative immunofluorescence, primary antibody and cytokeratin cocktail, SP142 (0.154 μg/ml (1:500)) and mouse monoclonal antihuman cytokeratin antibody (1:100; Dako, Cat #: M3515, clone AE1/AE3) were diluted in 0.3% bovine serum albumin/tris-buffered saline and Tween. This cocktail was then applied to all the slides, which were incubated overnight at 4 °C. The secondary antibody cocktail was prepared with a goat anti-mouse antibody, Alexa Fluor 546 (Life Technologies, Cat #: A11003), which was diluted 1:100 in an anti-rabbit horse radish peroxidase-labeled polymer reagent (Dako, Cat #: K4003), and applied to all the slides for 1 h at room temperature. Cyanine 5 Tyramide reagent (PerkinElmer, Cat #: FP1117) was then diluted 1:50 in an amplification diluent (PerkinElmer, Cat #: 1050) and then added to the batch slides for 10 min. All the slides were coverslipped using ProLong Gold reagent with 4′,6-diamidino-2-phenylindole (Life Technologies, Cat #: P36931) for nuclear staining.

The serial cuts of whole-tissue sections and control tissue microarrays used for quantitative immunofluorescence were then used for chromogenic immunohistochemistry. This primary antibody cocktail contained only SP142 (0.154 μg/ml (1:500)) diluted in 0.3% bovine serum albumin/tris-buffered

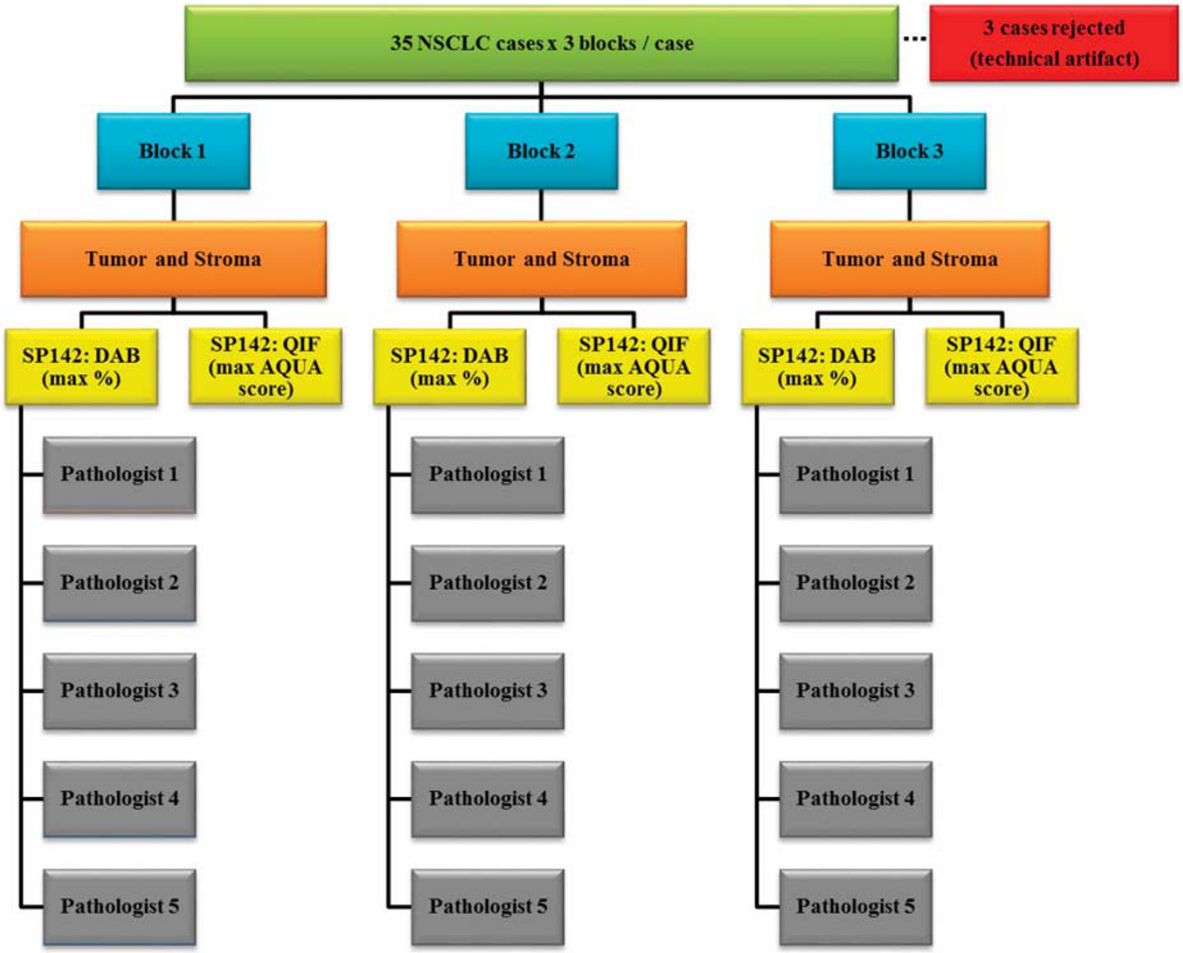

Figure 1 Consort diagram. This study included resections of 35 non-small cell lung cancer tumors. Three quantitative immunofluorescence cases were rejected due to the technical artifact of antibody trapping.

saline and Tween. After an overnight incubation at 4 °C, the slides were transferred to the in-house autostainer and only incubated with the anti-rabbit horse radish peroxidase-labeled polymer reagent for 1 h at room temperature, then incubated for 7 min at room temperature with diaminobenzidine chromogen (Dako, Cat #: K3468) diluted 1:50 in diaminobenzidine substrate buffer, then counterstained with hematoxylin (Dako, Cat #: S3301). This was followed with dehydration washes and coverslipping.

Scoring and Measurement of Fluorescence

The fluorophores 4′,6-diamidino-2-phenylindole and Cyanine 5 Tyramide were used during staining to visualize antibody target intensities in user-designated compartments within the tissue, such as tumor and stroma as previously described using the AQUA (Genoptix) method of quantitative immunofluorescence. Immunofluorescence scores are a reflection of PD-L1 antibody signal in either tumor or stromal compartments, and are calculated by dividing the PD-L1 compartment pixel intensities by the area within the respective compartment.[10] Scores were normalized to the exposure time and bit depth at which the images were captured, allowing scores collected at different exposure times to be comparable.

Whole slide tissue sections may represent 30–800 fields of view, where each field of view is about 0.5 mm^2. Owing to the time required for reading each field of view on current devices, it is impractical to read them all. To determine the number of fields of view that need to be measured to represent the entire section, a pilot study was performed that included all fields of view on all the three blocks of six cases representing 18 whole-tissue sections. A model was constructed to define the number of fields of view required to achieve a 95% likelihood that the average and max scores in the collected sample represented the average and max scores on the whole slide. Calculations showed that 29–70 fields of view, selected randomly and depending on the total number of fields of view, would be sufficient for a 95% chance of concordance between the sampled fields of view and the whole slide. To be sure we did

Table 1 Cohort clinical characteristics

Characteristic	*Number of patients*	*Percentage of patients*
All patients	35	100%
Age at diagnosis		
< 70	14	40%
≥ 70	21	60%
Sex		
Male	15	43%
Female	20	57%
Histology		
Adenocarcinoma	17	49%
Squamous cell	18	51%
Stage		
I	15	43%
II	14	40%
III–IV	6	17%
Tumor size, centimeters		
< 2	5	14%
2–5	27	77%
> 5	3	9%
Lymph node status		
Negative	20	57%
Positive	13	37%
NA	2	6%

not miss hot spots, rather than random selection, we subjectively selected fields from a low-resolution scan from the brightest to least bright even if the signal was dim and likely to represent noise. Subsequently, 29–70 fields of view were selected as dictated by the model based on the total number of fields of view on each slide. Once all the fields of view were captured using a high-resolution scan, those areas with < 2% tumor, normal lung tissue, and technical artifacts (damaged tissue, bubbles, or trapped antibody signal) were excluded from the analysis. Of 35 cases in the cohort used for this experiment, the quantitative immunofluorescence data from three cases required exclusion due to nonspecific trapping of antibody or other quality control issues preventing accurate scoring. Figure 3 shows an example of PD-L1 staining of whole-tissue sections for the three blocks from the same case. The serially cut sections were stained using quantitative immunofluorescence and chromogenic immunohistochemistry (diaminobenzidine) and a heat map of the quantitative immunofluorescence scores is shown below the diaminobenzidine images.

Scoring of Chromogenic Immunohistochemistry

Five pathologists (DEC, BEW, KAS, VP, and DLR) scored all whole-tissue sections by indicating the

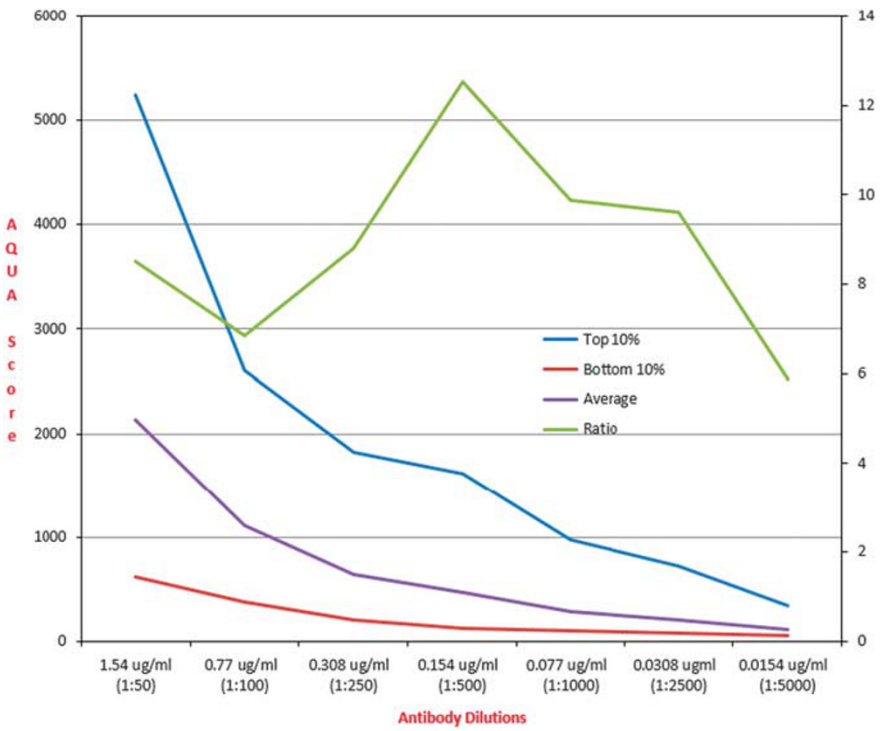

Figure 2 Illustration of quantitative assessment of optimal titration. Quantitative assessment of optimal antibody titer is achieved by plotting the average of the top 10% of scores on the test tissue microarray (blue line) and the average of the bottom 10% of the scores (red line) for each antibody concentration tested (*x* axis). The optimal titration is the maximal signal to noise (shown on the right side *y* axis) plotted (orange line) to show a peak at 0.154 μg/ml.

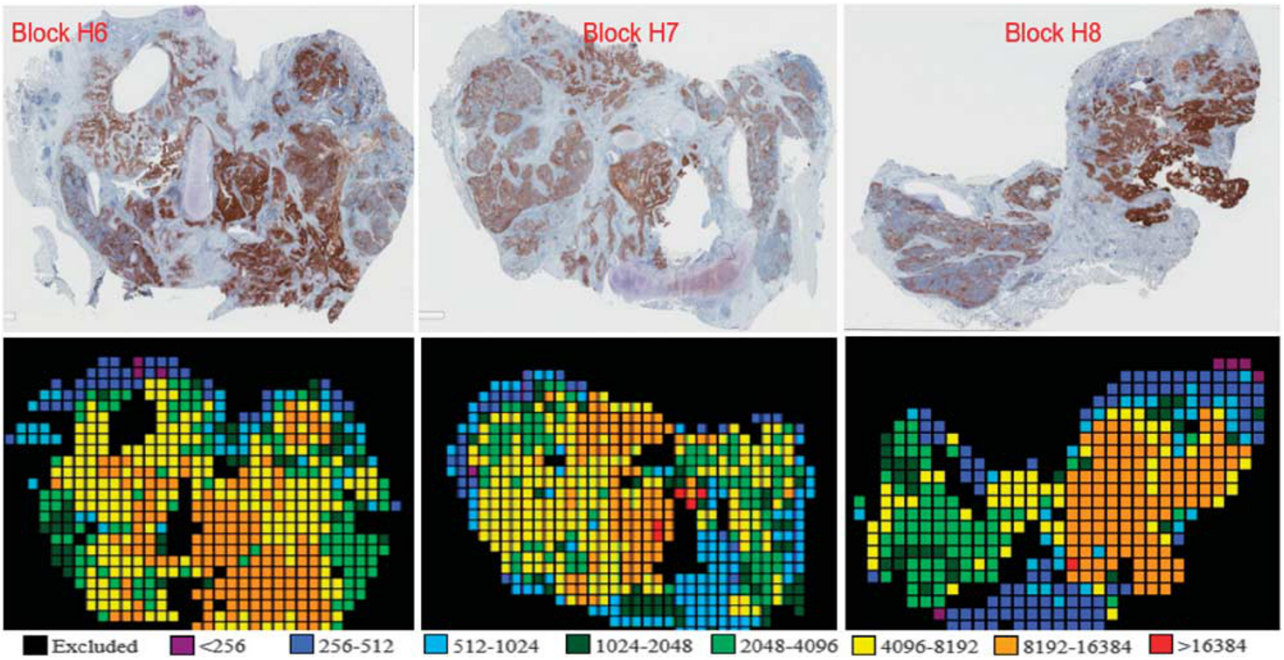

Figure 3 Images and heatmaps. Whole-tissue sections cut from three separate blocks from the same case. The top three panels indicate PD-L1 diaminobenzidine staining among all the three blocks. The bottom three panels show PD-L1 quantitative immunofluorescence staining of serial sections of corresponding blocks. The heatmaps are based on quantitative immunofluorescence data, which generated a quantitative immunofluorescence score as an arbitrary unit of fluorescence for each field of view within the tumor. The quantitative immunofluorescence score scale is presented below the heatmaps.

percentage of predominantly membranous PD-L1 staining of tumor cells and stromal or immune cells with perceptible PD-L1 signal at any intensity. The readers were not instructed to utilize certain percentage ranges or designations used in the clinical trials or other studies;[2–4,8,9,11–13] rather, each pathologist recorded his/her reading on the basis of a single, numerical, raw staining percentage of cells expressing PD-L1 at any intensity. All the 35 cases were adequately stained and passed quality control testing.

Statistical Analysis

The intraclass correlation coefficient applied to chromogenic immunohistochemistry data to evaluate the correlation between pathologists and blocks. The concordance between pathologists was evaluated using their readings of PD-L1 staining percentage in tumor and stroma. The heterogeneity between blocks was evaluated for each pathologist separately and also pooled into a single score.

The intraclass correlation coefficient was also used to assess PD-L1 heterogeneity as measured using quantitative immunofluorescence. For each case, both mean and maximum immunofluorescence score values in tumor and stroma of all three blocks were assessed. A mixed-effects model implementing 'analysis of variance' was used to quantify percentage of variance in the field of view values between and within blocks while adjusting random effects from the samples. The concordance between quantitative immunofluorescence data using immunofluorescence scores and chromogenic immunohistochemistry data using pathologists' maximum readings of PD-L1 staining percentages was then assessed using the Lin's concordance correlation coefficient and linear regression. The *P*-values were determined for tumor and stroma as a means of evaluating significant differences between blocks. Ultimately, block heterogeneity in this analysis was excluded by taking only the highest mean block immunofluorescence score, highest maximum block immunofluorescence score, and highest percentage staining from each pathologist among all the three blocks per case. Statistical analyses were performed using Statistical Analysis System software, or SAS, version 9.4 (SAS) and GraphPad Prism v6.0 (GraphPad Software).

Results

Pathologist Concordance

Five pathologists interpreted the diaminobenzidine staining of SP142 among three blocks of 35 cases. The staining percentages were recorded for tumor and stroma sections per block in raw whole number percentages from 0 to 100%. Tumor cells and immune cells exhibiting predominantly membranous staining were considered 'positive.' Figure 4a shows the distribution of the scores from each pathologist on

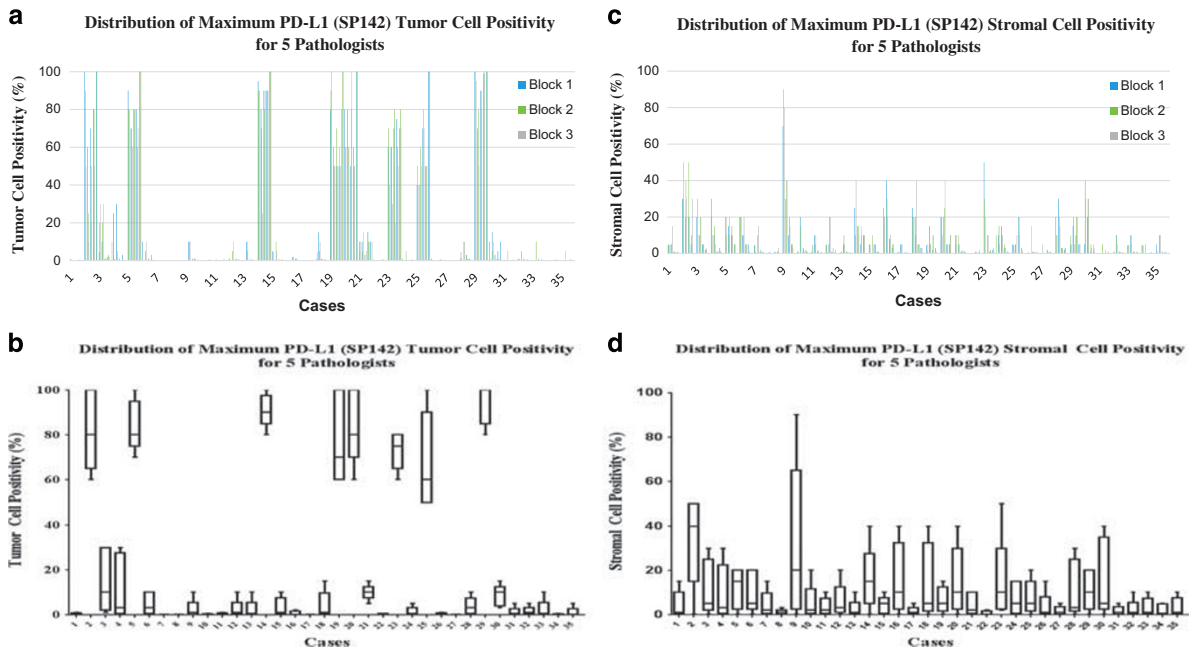

Figure 4 Distribution of maximum PD-L1 score among five pathologists. (**a**) A histogram of all chromogenic immunohistochemistry data for tumor: the raw percentage of staining assigned by each of the five pathologists for each of the three blocks, per case (15 bars per case, color coded by block as shown in the inset). (**b**) The distribution of the single maximum score provided by each of the five pathologists among all three blocks per case from tumor regions (five data points per case). Each boxplot represents 25th%, median, and 75th% readings, with the whiskers denoting minimum and maximum percentages of these five data points. The *y* axis labels the maximum reading among the three blocks. (**c**) Shows a histogram of all chromogenic immunohistochemistry data for stroma: the raw percentage of staining assigned by each of the five pathologists for each of the three blocks, per case (15 bars per case, color coded by block as shown in the inset). (**d**) The distribution of the single maximum score provided by each of the five pathologists among all the three blocks per case from stromal regions (five data points per case). Each boxplot represents 25th%, median, and 75th% readings, with the whiskers denoting minimum and maximum percentages of these five data points. The *y* axis labels the maximum reading among the three blocks.

Table 2a PD-L1 heterogeneity summary chromogenic immunohistochemistry (diaminobenzidine): programmed death-ligand 1 heterogeneity among pathologists and blocks

Table 2a	*Intraclass correlation coefficient among pathologists*	*Intraclass correlation coefficient among blocks*
Tumor	94%	94%
Stroma	27%	75%

Table 2b PD-L1 heterogeneity summary quantitative immunofluorescence: programmed death-ligand 1 heterogeneity among blocks

Table 2b	*Intraclass correlation coefficient among blocks (mean quantitative immunofluorescence score per block)*	*Intraclass correlation coefficient among blocks (maximum quantitative immunofluorescence score per block)*
Tumor	95%	88%
Stroma	88%	79%

each of the three blocks from each case. Overall, good concordance is evident both between blocks from the same cases and between pathologists. Figure 4b shows the box and whisker plot distributions for the maximum score (of the three blocks) for all five pathologists illustrating the overall variance between pathologists for each case. On the basis of these semi-quantitative readings from pathologists, the intraclass correlation coefficient was calculated between pathologists and between blocks for each case. The intraclass correlation coefficient between pathologists for tumor PD-L1 expression showed excellent concordance at 94% using the single maximum percentage per pathologist in all the three blocks per case (Table 2a). Figure 4b illustrates a generally bimodal distribution of PD-L1 expression where eight of the cases are 'high' PD-L1 expressers (all reads showing >50% staining) compared with the remainder of cases showing low or negative expression (all reads <25%).

Table 2c PD-L1 heterogeneity summary quantitative immunofluorescence: variance of fields of view among 3 blocks and within a block

Table 2c	*Variance of fields of view among 3 blocks*	*Variance of fields of view within a block*
Tumor	9%	91%
Stroma	4%	96%

The reading of the stromal scores were much less concordant. Figure 4c shows the distribution of the stromal scores illustrating the broad variation both between pathologists, and to a lesser extent between blocks. Figure 4d shows the high levels of variance and the absence of biomodality seen for the tumor cell scoring. The intraclass correlation coefficient among each pathologist's single maximum percentage score for stromal-immune cell staining was 27%, indicating substantial discordance (Table 2a).

Heterogeneity Between Tissue Blocks

To estimate the heterogeneity of expression of PD-L1 in both the tumor cells and the stromal cells, the intraclass correlation coefficient was calculated between blocks for each pathologist. On average, pathologists scored tumor sections of all three blocks per case quite similarly (intraclass correlation coefficient = 94%), but their stromal sections shared a less substantial correlation (intraclass correlation coefficient = 75%; Table 2a).

Quantitative Measurement Of Tumor And Stromal PD-L1 Expression

Unlike the pathologists' estimate of percentage of cells positive at any intensity, the automated quantitative immunofluorescence method combines both the area of expression with the intensity of staining to generate a score that is more similar to a concentration than a percentage. Figure 5 shows the immunofluorescence score range for all the fields of view from all three blocks for each case, plotted from low to high, then color coded for the average of all pathologists' scores for each case. The generally continuous nature of the distribution is illustrated as is the general agreement with pathologist scores. Calculation of the block-to-block heterogeneity for tumor cell expression of PD-L1 showed an intraclass correlation coefficient of 95% using an immunofluorescence score that represents the mean score from all fields of view from each block. The intraclass correlation coefficient between blocks for average stromal scores was 88%. Intraclass correlation coefficients for the maximum immunofluorescence score of all the fields of view for tumor and stromal PD-L1 was 88% and 79%, respectively (Table 2b).

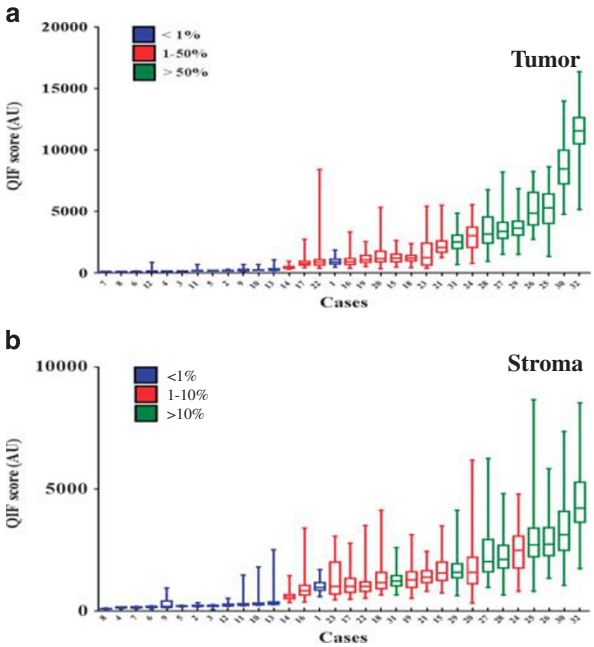

Figure 5 Quantitative immunofluorescence *vs* chromogenic immunohistochemistry. The quantitative immunofluorescence score is shown as a box and whisker plot for representative fields of view from each case. Each box represents the 25th%, median, and 75th% quantitative immunofluorescence score of the respective case. Whiskers represent the minimum and maximum score. The *x* axis indicates all cases organized by their median quantitative immunofluorescence score values. Cases are also color coded by their PD-L1 diaminobenzidine staining percentages: those in blue, stained $<1\%$; those in red, stained 1–50%; and those in green, stained $>50\%$. (**a**) The scores for the PD-L1 expression in the tumor. (**b**) The scores for the PD-L1 expression in the stroma.

Quantification allows for assessment of heterogeneity of PD-L1 expression using a linear mixed-effects regression model both between blocks for a given case and between fields of view on a given slide. Using measurements from all the fields of view measured, we found that the variance between blocks is quite small (9% for tumor and 4% for stroma), compared with the variance between fields of view on a single slide (91% for tumor and 96% for stroma, Table 2c).

Finally, the quantitative information can be compared with the reads by the pathologist. The Lin's concordance correlation coefficient and linear regression were used to assess the concordance between the scoring by pathologists and quantitative immunofluorescence data from serial sections. Pathologists' single maximum percentage score among all three blocks was compared with the single maximum immunofluorescence score among all the three blocks and the largest mean immunofluorescence score among all the three blocks. After standardizing these variables and averaging all five pathologists' concordance with the highest mean immunofluorescence score among all three blocks, there was a 94% concordance in tumor and 68%

Table 2d PD-L1 heterogeneity summary chromogenic immunohistochemistry (diaminobenzidine) *vs* quantitative immunofluorescence: concordance in tumor and stroma

Table 2d	*5 Pathologists* vs *highest mean quantitative immunofluorescence score (among all 3 blocks)*	*5 Pathologists* vs *single maximum quantitative immunofluorescence score (among all 3 blocks)*
Tumor	94%	92%
Stroma	68%	70%

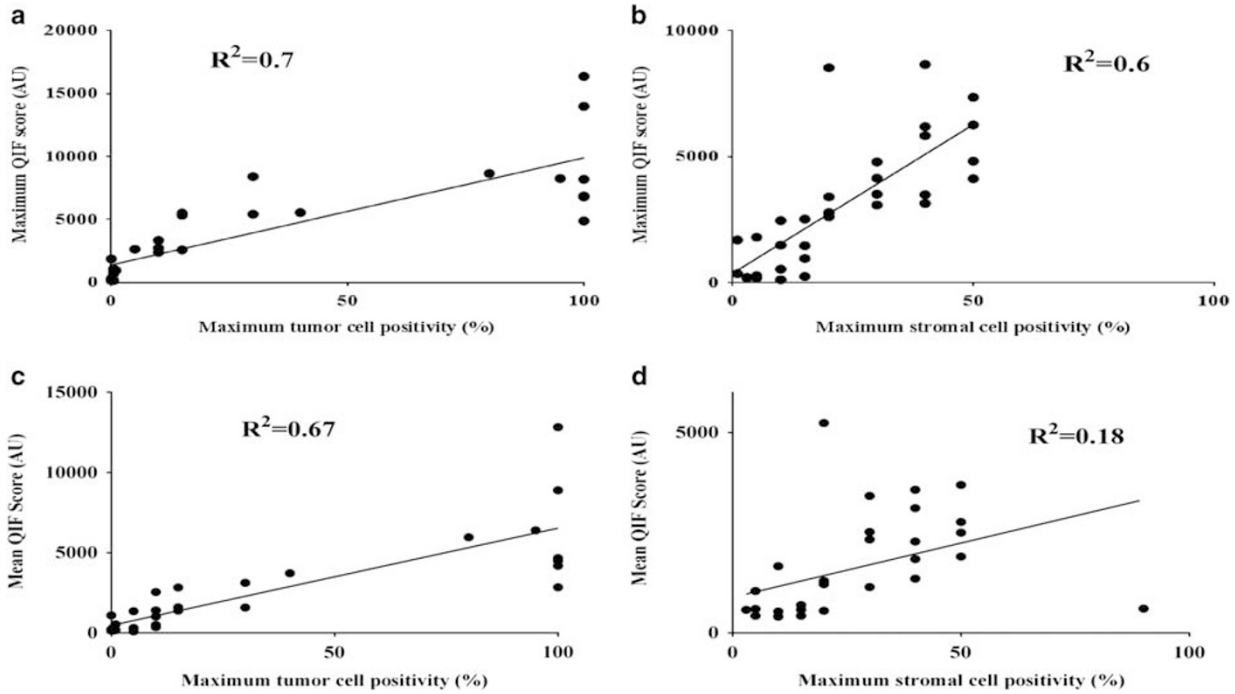

Figure 6 Regressions of maximum and mean quantitative immunofluorescence score *vs* maximum pathologists' score. The maximum percentage PD-L1 staining among all the pathologists was regressed with the maximum quantitative immunofluorescence score among all the three blocks, in both (**a**) tumor and (**b**) stromal regions. Also, the maximum percentage PD-L1 staining among all the pathologists was regressed with the highest average quantitative immunofluorescence score of a block among three blocks, in both (**c**) tumor and (**d**) stromal regions.

concordance in stromal regions of the blocks. Calculating the same concordance using the maximum immunofluorescence score among all the three blocks (rather than the mean immunofluorescence score) revealed a 92% concordance among pathologists' scoring in tumor and a 70% concordance in stroma. Table 2d summarizes these findings.

Regression analysis was also used to compare pathologist scores with quantitative measurements. Figures 6a and b are regression analyses between the single maximum immunofluorescence scores per case and the highest percentage of PD-L1 staining score by any pathologist, per case. Consistent with the interpretation of the Lin's concordance correlation coefficient values, the r-squared was greater in tumor ($r^2 = 0.7$) than in stroma ($r^2 = 0.6$). Figures 6c and d are regression analyses between the highest mean immunofluorescence score of any block, with the highest percentage of PD-L1 staining score by any pathologist, per case. These data indicated that not only was the r-squared greater in tumor regions ($r^2 = 0.67$) than in stroma ($r^2 = 0.18$), but also that the maximum immunofluorescence score among all the three blocks is more correlated to pathologists' maximum reading than is the highest mean immunofluorescence score among all the three blocks.

Discussion

Perhaps the most significant and promising finding in this work is that when pathologists score tumor cell percentages, they are highly concordant. This may be important as the first few PD-1 axis drugs that have been, or are about to be approved, use different cut-points. Similarly, the high concordance between different blocks from the same case suggests that a

single block may be representative of the larger tumor. However, more concerning is the lack of concordance in the estimation of stromal or immune cell scores. This may be due to the relatively low levels of immune cell expression and the challenge of concordance when estimating low frequency events.

Although our cohort is small, the observations in this work are generally concordant with that previously described with respect to the distribution of expression in the total population. For example, Garon *et al*[2] characterized the prevalence of PD-L1 expressers in their patient population showing that 23.2% of patients' tumors stained >50%, 37.6% stained 1–50%, and 39.2% staining <1%. Our analysis of 35 patients revealed similar results showing that 25% of patients' tumors stained >50%, 34% stained 1–50%, and 41% stained <1%.

Stromal measurements among pathologists resulted in an intraclass correlation coefficient of 27%, indicating prominent discordance (Table 2a). When comparing pathologists' stromal score with quantitative immunofluorescence data, there was a 68% concordance with using the highest mean immunofluorescence score and 70% concordance with using the maximum immunofluorescence score among all the three blocks per case. Thus, not only are pathologists relatively discordant in their abilities to score stromal-immune cell PD-L1 expression, but they are less concordant with quantitative methods than their readings for tumor samples. These findings raise questions about the ability of pathologists to score stromal-immune cells concordantly and accurately, in light of studies using Atezolizumab (MPDL3280A) with SP142 to explore the relationship between PD-L1 expression on immune cells and response to the drug.[5,11,14,15] This may be due to the lack of consensus on the exact method for reading stromal cells or it may be a function of the inherent challenge of scoring of scarce events.

A second key finding of this work is the quantitative assessment of heterogeneity of expression of PD-L1. The mixed-effects model suggests that well over 90% of the heterogeneity that we see is presented in a single slide and that the variance between different regions of the tumor (different blocks) is not substantial. Specifically, variance of fields of view between each of the three blocks was only 9% for tumor and 4% for stroma, in stark contrast to the variance between fields of view within a given block being 91% in tumor and 96% in stroma. Coupled with pathologists' interpretation of diaminobenzidine staining, these results indicated that PD-L1 expression is heterogeneous within fields of view of the same whole-tissue section (at the millimeter level), rather than from block-to-block (at the centimeter level). This lack of inter-block heterogeneity indicates that a single block is representative of PD-L1 expression of the entire tumor. However, the minimal representative area on a block required to predict response to therapy remains to be determined.

Although the results of our data are encouraging, there are several limitations to our study. First, patient outcome data including their treatment and time to progression were not collected for this comparative study. This study would be much more valuable if we had the criterion standard of response to immune therapy for every case, but as the drugs are only recently released, this is not possible. The second major limitation is the relatively small sample size. However, future larger studies are in process and even with this sample size, some very compelling conclusions could be drawn. Third, no statistical method has been found to quantify the concordance between diaminobenzidine and quantitative immunofluorescence data using all the field of view values. The use of current single maximum or highest of means in the three blocks of field of view values takes into account only part of the quantitative immunofluorescence information. Extensions of Lin's concordance correlation coefficient to handle multivariate data will lead to improved interpretation of the concordance between pathologists' percentage scores and immunofluorescence scores. Finally, this study only used one commercially available antibody and the method was not that prescribed in the investigational-use-only studies. As such, this study provides no information related to the concordance of the Food and Drug Administration's approved or submitted PD-L1 assays that do, or will populate the drug labels.

In summary, with the Food and Drug Administration's approval of three monoclonal antibodies that target the PD-1 axis in lung cancer, high response rates and impressive duration in selected populations suggest that a companion diagnostic assay is inevitable for this class of therapy. Here we show some key characteristics related to the companion diagnostic test, including (1) pathologists are more concordant in scoring tumor than immune cells or stromal cells; (2) pathologists are concordant with quantitative measurement for tumor cells PD-L1 but less so for immune cell PD-L1, and (3) the heterogeneity seen in PD-L1 expression is represented within the block, rather than between blocks, as shown by the assessment of variance. These data suggest that pathologists can characterize PD-L1 expression in tumor using the conventional immunohistochemistry test. Future studies may be done to compare tests or compare the efficacy of this test with other methods of prediction of response to PD-1 axis therapies.

Acknowledgments

This work was supported by the Yale SPORE in Lung Cancer P50CA196530, the Yale Cancer Center

Support Grant, P30CA016359, the Breast Cancer Research Foundation, and a Sponsored Research Agreement from Genoptix.

Disclosure/conflict of interest

DLR has served as a paid consultant or advisor to Genoptix/Novartis, Applied Cellular Diagnostics, BMS, Amgen, Optrascan, Biocept, PerkinElmer, and Metamark Genetics. The remaining authors declare no conflict of interest.

References

1 Teixido C, Karachaliou N, Gonzalez-Cao M, *et al.* Assays for predicting and monitoring responses to lung cancer immunotherapy. Cancer Biol Med 2015;12: 87–95.
2 Garon EB, Rizvi NA, Hui R, *et al.* Pembrolizumab for the treatment of non-small cell lung cancer. N Engl J Med 2015;372:2018–2028.
3 Borghaei H, Paz-Ares L, Horn L, *et al.* Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med 2015;373:1627–1639.
4 Rizvi NA, Mazieres J, Planchard D, *et al.* Activity and safety of nivolumab, an anti-programmed death-protein 1 immune checkpoint inhibitor, for patients with advanced, refractory squamous non-small-cell lung cancer (CheckMate 063): a phase 2, single-arm trial. Lancet Oncol 2015;16:257–265.
5 Kerr KM, Tsao MS, Nicholson AG, *et al.* Programmed death-ligand 1 immunohistochemistry in lung cancer: in what state is this art? J Thorac Oncol 2015;10: 985–989.
6 Mansfield AS, Murphy SJ, Peikert T, *et al.* Heterogeneity of programmed cell death-ligand 1 expression in multifocal lung cancer. Clin Cancer Res 2015;22: 2177–2182.
7 Sheng J, Fang W, Yu J, *et al.* Expression of programmed death ligand-1 on tumor cells varies pre and post chemotherapy in non-small cell lung cancer. Sci Rep 2016;6:20090.
8 Brahmer J, Reckamp KL, Baas P, *et al.* Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. N Engl J Med 2015;373: 123–135.
9 McLaughlin J, Han G, Schalper KA, *et al.* Quantitative assessment of the heterogeneity of programmed death-ligand 1 expression in non-small-cell lung cancer. JAMA Oncol 2016;2:46–54.
10 Camp RL, Chung GG, Rimm DL. Automated subcellular localization and quantification of protein expression in tissue microarrays. Nat Med 2002;8: 1323–1327.
11 Herbst RS, Soria JC, Kowanetz M, *et al.* Predictive correlates of response to the anti-programmed death-ligand 1 antibody MPDL3280A in cancer patients. Nature 2014;515:563–567.
12 Paz-Ares L, Horn L, Borghaei H, *et al.* Phase III, randomized trial (CheckMate 057) of nivolumab (NIVO) versus docetaxel (DOC) in advanced non-squamous cell (non-SQ) non-small cell lung cancer (non-small cell lung cancer) in ASCO Annual Meeting. J Clin Oncol 2015;33 (suppl; abstr LBA109).
13 Rizvi NA, Brahmer JR, Ou S-HI, *et al.* Safety and clinical activity of MEDI4736, an anti-programmed cell death-ligand 1 (programmed death-ligand 1) antibody, in patients with non-small cell lung cancer (non-small cell lung cancer) in ASCO Annual Meeting. J Clin Oncol 2015;33 (suppl; abstr 8032).
14 Galon J, Pages F, Marincola FM, *et al.* The immune score as a new possible approach for the classification of cancer. J Transl Med 2012;10:1.
15 Spira AI, Keunchil P, Mazieres J *et al.* Efficacy, safety and predictive biomarker results from a randomized phase II study comparing MPDL3280A vs docetaxel in 2 L/3 L non-small cell lung cancer (POPLAR) in ASCO Annual Meeting. J Clin Oncol 2015;33 (Suppl, abstr 8010).

Supplementary Information accompanies the paper on Modern Pathology website (http://www.nature.com/modpathol)