

Supplementary Table S5. Intraobserver reproducibility of the manual H-score assessment*One pathologist (S.W.C.), two independent readings 2 weeks apart (n = 544 cores)***(A) Agreement of the continuous H-score**

Metric	Reading 1	Reading 2	Agreement (95% CI)
H-score, mean $\pm$ SD	113.6 $\pm$ 91.3	117.1 $\pm$ 93.2	-
H-score, median	80.0	95.0	-
ICC (A,1), single measurement, absolute agreement			0.955 (0.95–0.96)
ICC (A,k), mean of two readings			0.977 (0.97–0.98)
Pearson correlation			0.956
Spearman rank correlation			0.927
Bias (reading 1 - reading 2)			-3.6 (SD 27.5)
Mean absolute difference			20.0
95% limits of agreement			-57.5 to +50.3

(B) Between-reading difference stratified by expression level

H-score range (mean of the two readings)	n	Bias (reading 1 - reading 2)	SD of differences
0–19	71	+1.90	8.17
20–59	99	+7.78	25.73
60–119	187	-11.04	26.15
120–199	55	-5.36	42.71
200–300	132	-3.64	26.19

CI, confidence interval; ICC, intraclass correlation coefficient; SD, standard deviation.

^aThe same pathologist visually estimated the H-score (0-300) for all cores on two occasions separated by 2 weeks, blinded to the first reading. The mean of the two readings was used as the final manual H-score in all analyses.

^bAnalysis was restricted to the 544 cores in which at least 100 tumor cells were detected by both detection algorithms. Because reproducibility of the manual H-score does not depend on survival outcome, all such cores were retained; the 528-core final analytic cohort is a subset of these, additionally requiring complete survival data (16 of the 544 cores lacked survival follow-up). ICC, two-way random-effects, absolute-agreement model. Bias, mean difference between readings; limits of agreement by the Bland-Altman method. The two readings differed by a small but statistically significant systematic offset (paired t-test, $p = .003$; mean difference -3.6 H-score points, i.e. 1.2% of the 0-300 scale). ^cBetween-reading variability was smallest at low H-scores and largest in the intermediate range, where visual estimation of the proportion of moderately stained nuclei is most demanding.